

An aerial photograph of a multi-lane highway interchange with several cars driving on it. The background shows a city skyline under a clear sky. Overlaid on the image are several semi-transparent circular nodes connected by thin white lines, forming a network. Each node contains a different icon: a signal tower, a smartphone, a laptop, a coffee cup, a truck, and a headset. A large, light-blue circular area in the center of the image contains the title text.

BUSINESS INTELLIGENCE, ANALYTICS, DATA SCIENCE, AND AI

**A MANAGERIAL
PERSPECTIVE**

FIFTH EDITION



**RAMESH
SHARDA**



**DURSUN
DELEN**



**EFFRAIM
TURBAN**

FIFTH EDITION

BUSINESS INTELLIGENCE, ANALYTICS, DATA SCIENCE, AND AI

A MANAGERIAL PERSPECTIVE

Ramesh Sharda

Oklahoma State University

Dursun Delen

Oklahoma State University

Efraim Turban

University of Hawaii

Content Production: Guneet Gulati
Product Management: Jenifer Niles
Product Marketing: Wayne Stevens
Rights and Permissions: Jenell Forschler

Please contact <https://support.pearson.com/getsupport/s/> with any queries on this content.

Cover Image by Jamesteohart/Shutterstock

Microsoft and/or its respective suppliers make no representations about the suitability of the information contained in the documents and related graphics published as part of the services for any purpose. All such documents and related graphics are provided “as is” without warranty of any kind. Microsoft and/or its respective suppliers hereby disclaim all warranties and conditions with regard to this information, including all warranties and conditions of merchantability, whether express, implied or statutory, fitness for a particular purpose, title and non-infringement. In no event shall Microsoft and/or its respective suppliers be liable for any special, indirect or consequential damages or any damages whatsoever resulting from loss of use, data or profits, whether in an action of contract, negligence or other tortious action, arising out of or in connection with the use or performance of information available from the services.

The documents and related graphics contained herein could include technical inaccuracies or typographical errors. Changes are periodically added to the information herein. Microsoft and/or its respective suppliers may make improvements and/or changes in the product(s) and/or the program(s) described herein at any time. Partial screen shots may be viewed in full within the software version specified.

Microsoft® and Windows®, and Microsoft Office® are registered trademarks of the Microsoft Corporation in the U.S.A. and other countries. This book is not sponsored or endorsed by or affiliated with the Microsoft Corporation.

Copyright © 2024, 2018, 2014 by Pearson Education, Inc., All rights reserved. Manufactured in the United States of America. This publication is protected by copyright, and permission should be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise. For information regarding permissions, request forms, and the appropriate contacts within the Pearson Education Global Rights and Permissions department, please visit www.pearsoned.com/permissions/.

Acknowledgments of third-party content appear on the appropriate page within the text.

PEARSON and MYLAB are exclusive trademarks owned by Pearson Education, Inc. or its affiliates in the U.S. and/or other countries.

Unless otherwise indicated herein, any third-party trademarks, logos, or icons that may appear in this work are the property of their respective owners, and any references to third-party trademarks, logos, icons, or other trade dress are for demonstrative or descriptive purposes only. Such references are not intended to imply any sponsorship, endorsement, authorization, or promotion of Pearson's products by the owners of such marks, or any relationship between the owner and Pearson Education, Inc., or its affiliates, authors, licensees, or distributors.

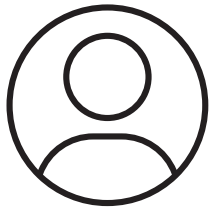
Library of Congress Cataloging-in-Publication Data

Business Intelligence, Analytics, Data Science, and AI: A Managerial Perspective
Library of Congress Cataloging in Publication Control Number: 2023012978

ScoutAutomatedPrintCode



ISBN-10: 0-13-793128-X
ISBN-13: 978-0-13-793128-6



Pearson's Commitment to Diversity, Equity, and Inclusion

Pearson is dedicated to creating bias-free content that reflects the diversity, depth, and breadth of all learners' lived experiences.

We embrace the many dimensions of diversity, including but not limited to race, ethnicity, gender, sex, sexual orientation, socioeconomic status, ability, age, and religious or political beliefs.

Education is a powerful force for equity and change in our world. It has the potential to deliver opportunities that improve lives and enable economic mobility. As we work with authors to create content for every product and service, we acknowledge our responsibility to demonstrate inclusivity and incorporate diverse scholarship so that everyone can achieve their potential through learning. As the world's leading learning company, we have a duty to help drive change and live up to our purpose to help more people create a better life for themselves and to create a better world.

Our ambition is to purposefully contribute to a world where:

- Everyone has an equitable and lifelong opportunity to succeed through learning.
- Our educational content accurately reflects the histories and lived experiences of the learners we serve.
- Our educational products and services are inclusive and represent the rich diversity of learners.
- Our educational content prompts deeper discussions with students and motivates them to expand their own learning (and worldview).

Accessibility

We are also committed to providing products that are fully accessible to all learners. As per Pearson's guidelines for accessible educational Web media, we test and retest the capabilities of our products against the highest standards for every release, following the WCAG guidelines in developing new products for copyright year 2022 and beyond.



You can learn more about Pearson's commitment to accessibility at <https://www.pearson.com/us/accessibility.html>



Contact Us

While we work hard to present unbiased, fully accessible content, we want to hear from you about any concerns or needs with this Pearson product so that we can investigate and address them.



Please contact us with concerns about any potential bias at <https://www.pearson.com/report-bias.html>



For accessibility-related issues, such as using assistive technology with Pearson products, alternative text requests, or accessibility documentation, email the Pearson Disability Support team at disability.support@pearson.com

This page intentionally left blank

BRIEF CONTENTS

Preface xxii
About the Authors xxix

PART I Introduction

- Chapter 1** An Overview of Business Intelligence, Analytics, Data Science, and AI 2
- Chapter 2** Artificial Intelligence: Concepts, Drivers, Major Technologies, and Business Applications 63

PART II Descriptive Analytics

- Chapter 3** Descriptive Analytics I: Nature of Data, Big Data, and Statistical Modeling 117
- Chapter 4** Descriptive Analytics II: Business Intelligence Data Warehousing and Visualization 185

PART III Predictive Analytics

- Chapter 5** Predictive Analytics I: Data Mining Process, Methods, and Algorithms 250
- Chapter 6** Predictive Analytics II: Text, Web, and Social Media Analytics 301
- Chapter 7** Deep Learning and Cognitive Computing 368

PART IV Prescriptive Analytics

- Chapter 8** Prescriptive Analytics: Optimization and Simulation 441

PART V Software and Trends

- Chapter 9** Landscape of Business Analytics Tools 494
- Chapter 10** AI-Based Trends in Analytics and Data Science 574
- Chapter 11** Ethical, Privacy, and Managerial Considerations in Analytics 629

Glossary 670
Index 680

Preface xxii

About the Authors xxix

PART I Introduction

Chapter 1 An Overview of Business Intelligence, Analytics, Data Science, and AI 2

- 1.1 Opening Vignette: Sports Analytics—An Exciting Frontier for Learning and Understanding Applications of Analytics 3**
- 1.2 Changing Business Environments and Evolving Needs for Decision Support and Analytics 17**
 - Decision-Making Process 18
 - Technologies for Data Analysis and Decision Support 18
- 1.3 Decision-Making Processes and Computerized Decision Support Framework 20**
 - Simon's Process: Intelligence, Design, and Choice 20
 - The Intelligence Phase: Problem (or Opportunity) Identification 21
 - **ANALYTICS IN ACTION 1.1** Making Elevators Go Faster! 22
 - The Design Phase 23
 - The Choice Phase 24
 - The Implementation Phase 24
- 1.4 Evolution of Computerized Decision Support to Analytics/Data Science 25**
- 1.5 A Framework for Business Intelligence 28**
 - Definitions of BI 28
 - A Brief History of BI 28
 - The Architecture of BI 29
 - The Origins and Drivers of BI 29
 - A Multimedia Exercise in Business Intelligence 30
 - Transaction Processing versus Analytic Processing 31
 - Appropriate Planning and Alignment with the Business Strategy 31
 - Real-Time, On-Demand BI Is Attainable 32
 - Developing or Acquiring BI Systems 33
 - Justification and Cost–Benefit Analysis 33
 - Security and Protection of Privacy 33
 - Integration of Systems and Applications 33
- 1.6 Analytics Overview 34**
 - Descriptive Analytics 35
 - Predictive Analytics 35

- ▶ **ANALYTICS IN ACTION 1.2** Silvaris Increases Business with Visual Analysis and Real-Time Reporting Capabilities 36
 - Prescriptive Analytics 36
- ▶ **ANALYTICS IN ACTION 1.3** How Big Will Be the Beef? Understanding Animals' Eating Behavior and Their Final Weight 37
 - The Modeling Process and Results 39
- ▶ **ANALYTICS IN ACTION 1.4** A Specialty Steel Bar Company Uses Analytics to Determine Available-to-Promise Dates 41
 - Analytics/Data Science/Machine Learning/AI? 42
- 1.7** Analytics Examples in Selected Domains 43
 - Analytics Applications in Healthcare—Humana Examples 43
 - Analytics in the Retail Value Chain 45
 - Gaming Industry Applications 48
 - ▶ **ANALYTICS IN ACTION 1.5** Gulfstream Park Casino Employs AI to make Optimal Decisions on Slot Machine Placement 49
 - COVID-19 Analytics Applications 52
 - Mask Mandates and COVID-19 Spread: Evidence from within State Variation Problem 54
 - Data 54
 - Results 55
- 1.8** Plan of the Book 58
- 1.9** Resources, Links, and the Teradata University for Academics Connection 59
 - Resources and Links 59
 - Vendors, Products, and Demos 59
 - Periodicals 59
 - The Teradata University for Academics Connection 59
 - The Book's Web Site 59
 - Chapter Highlights 60 • Key Terms 60 • Questions for Discussion 60 • Exercises 61 • References 61*

Chapter 2 Artificial Intelligence: Concepts, Drivers, Major Technologies, and Business Applications 63

- 2.1** Opening Vignette: Grant Thornton Employs Aisera Chatbot to Reduce IT Help Desk Burden 64
- 2.2** Introduction to Artificial Intelligence 65
 - Definitions 66
 - Major Characteristics of AI Machines 66
 - Major Elements of AI 66
 - AI Applications 66
 - Major Goals of AI 67
 - Drivers of AI 68
 - Three Flavors of AI Decisions 70
 - ▶ **TECHNOLOGY INSIGHT 2.1** Augmented Intelligence 71
- 2.3** Human and Computer Intelligence 71
 - What Is Intelligence? 71

	How Intelligent Is AI?	72
	Measuring AI	72
2.4	Major AI Technologies and Some Derivatives	74
2.5	AI Support for Decision Making	79
	► TECHNOLOGY INSIGHT 2.2 Schrage's Models for Using AI to Make Decisions	82
2.6	AI Applications in Various Business Functions	83
	► ANALYTICS IN ACTION 2.1 How EY, Deloitte, and PwC Are Using AI	83
	AI Applications in Financial Services	84
	AI in Marketing, Advertising, and CRM	86
2.7	Introduction to Robotics	88
	► ANALYTICS IN ACTION 2.2 AI/Analytics in Action	89
	What We Can Learn from This Analytics in Action 2.2?	91
2.8	Illustrative Applications of Robotics	93
	Autonomous Cars: Robots in Motion	100
2.9	Conversational AI—Chatbots	102
	What Is a Chatbot?	103
	Chatbot Evolution	103
	Components of Chatbots and the Process of Their Use	104
	Drivers and Benefits	105
	Representative Chatbots from around the World	105
2.10	Enterprise Chatbots	106
	The Interest of Enterprises in Chatbots	106
	Enterprise Chatbots: Marketing and Customer Experience	106
	Coca-Cola	106
	► ANALYTICS IN ACTION 2.3 WeChat's Super Chatbot	107
	Enterprise Chatbots: Financial Services	107
	Enterprise Chatbots: Service Industries	107
	Chatbot Platforms	108
	Knowledge for Enterprise Chatbots	109
	Virtual Personal Assistants	109
	If You Were Mark Zuckerberg, Facebook CEO	109
	Amazon's Alexa and Echo	109
	Apple's Siri	110
	Google Assistant	110
	Other Personal Assistants	111
	Chatbots as Professional Advisors (Robo Advisors)	111
	Robo Financial Advisors	111
	Evolution of Financial Robo Advisors	111
	<i>Chapter Highlights</i>	112 • <i>Key Terms</i>
	<i>Discussion</i>	113 • <i>Exercises</i>
	<i>References</i>	115

PART II Descriptive Analytics

Chapter 3 Descriptive Analytics I: Nature of Data, Big Data, and Statistical Modeling 117

- 3.1 Opening Vignette: SiriusXM Attracts and Engages a New Generation of Radio Consumers with Data-Driven Marketing 118
- 3.2 The Nature of Data in Analytics 121
- 3.3 A Simple Taxonomy of Data 125
- 3.4 The Art and Science of Data Preprocessing 127
 - **ANALYTICS IN ACTION 3.1** Improving Student Retention with Data-Driven Analytics 130
- 3.5 Definition of Big Data 136
 - The “V”s That Define Big Data 137
 - **TECHNOLOGY INSIGHTS 3.1** The Data Size Is Getting Big, Bigger, and Bigger 138
- 3.6 Fundamentals of Big Data Analytics 140
 - Business Problems Addressed by Big Data Analytics 143
- 3.7 Big Data Technologies 143
 - Hadoop 144
 - How Does Hadoop Work? 144
 - MapReduce 145
 - Why Use MapReduce? 146
 - Hadoop Technical Components 146
 - Hadoop: The Pros and Cons 147
 - **TECHNOLOGY INSIGHTS 3.2** A Few Demystifying Facts about Hadoop 148
 - Spark versus Hadoop 149
 - NoSQL 151
 - Data for Good 151
- 3.8 Big Data and Stream Analytics 153
 - Applications of Stream Analytics 154
 - e-Commerce 154
 - Telecommunications 155
 - Law Enforcement and Cybersecurity 155
 - Power Industry 155
 - Financial Services 156
 - Health Sciences 156
 - Government 156
- 3.9 Statistical Modeling for Business Analytics 157
 - Descriptive Statistics for Descriptive Analytics 158
 - Measures of Centrality Tendency (May Also Be Called Measures of Location or Centrality) 158
 - Arithmetic Mean 158

	Median	159
	Mode	159
	Measures of Dispersion	160
	Range	160
	Variance	160
	Standard Deviation	161
	Mean Absolute Deviation	161
	Quartiles and Interquartile Range	161
	Box-and-Whiskers Plot	161
	The Shape of a Distribution	162
	► TECHNOLOGY INSIGHT 3.3 How to Calculate Descriptive Statistics in Microsoft Excel	164
	► ANALYTICS IN ACTION 3.2 Town of Cary Uses Analytics to Analyze Data from Sensors, Assess Demand, and Detect Problems	167
3.10	Regression Modeling for Inferential Statistics	169
	How Do We Develop the Linear Regression Model?	170
	How Do We Know If the Model Is Good Enough?	171
	What Are the Most Important Assumptions in Linear Regression?	172
	Logistic Regression	173
	► ANALYTICS IN ACTION 3.3 Predicting NCAA Bowl Game Outcomes	174
	Time Series Forecasting	179
	<i>Chapter Highlights</i>	181 • <i>Key Terms</i> 182 • <i>Questions for Discussion</i> 182 • <i>Exercises</i> 183 • <i>References</i> 184

Chapter 4 Descriptive Analytics II: Business Intelligence Data Warehousing, and Visualization 185

4.1	Opening Vignette: Targeting Tax Fraud with Data Warehousing and Business Analytics	186
4.2	Business Intelligence and Data Warehousing	188
	What Is a Data Warehouse?	190
	A Historical Perspective to Data Warehousing	190
	► ANALYTICS IN ACTION 4.1 Data-Driven Customer Experience in Financial Services	192
	Characteristics of Data Warehousing	193
	Data Marts	194
	Operational Data Stores	194
	Enterprise Data Warehouses (EDW)	195
	Metadata	195
4.3	Data Warehousing Process	195
4.4	Data Warehousing Architectures	197
	Alternative Data Warehousing Architectures	199
	Which Architecture Is the Best?	202

4.5	Data Management and Warehouse Development	203
	Data Warehouse Development Approaches	204
	Additional Data Warehouse Development Considerations	206
	► TECHNOLOGY INSIGHTS 4.1 Hosted Data Warehouses	207
	Representation of Data in Data Warehouse	208
	Analysis of Data in Data Warehouse	209
	OLAP versus OLTP	209
	OLAP Operations	209
	Data Integration and the Extraction, Transformation, and Load (ETL) Processes	210
	Data Integration	211
	► ANALYTICS IN ACTION 4.2 AARP Transforms Its BI Infrastructure and Achieves a 347% ROI in Three Years	212
	Extraction, Transformation, and Load	214
4.6	Data Warehouse Administration, Security Issues, and Future Trends	217
	The Future of Data Warehousing	217
	► TECHNOLOGY INSIGHTS 4.2 Data Lakes	218
4.7	Business Reporting	223
4.8	Data Visualization	225
	A Brief History of Data Visualization	225
4.9	Different Types of Charts and Graphs	228
	Basic Charts and Graphs	228
	Specialized Charts and Graphs	229
	Which Chart or Graph Should You Use?	230
4.10	The Emergence of Visual Analytics	232
	► TECHNOLOGY INSIGHTS 4.3 Gartner Magic Quadrant for Business Intelligence and Analytics Platforms	233
	Visual Analytics	234
	► TECHNOLOGY INSIGHTS 4.4 Telling Great Stories with Data and Visualization	235
	High-Powered Visual Analytics Environments	236
4.11	Information Dashboards	238
	► ANALYTICS IN ACTION 4.3 Increasing the Efficiency of Social Media Campaign Reporting to Get to Insights Quicker	240
	Dashboard Design	241
	What to Look for in a Dashboard	242
	Best Practices in Dashboard Design	242
	Benchmark Key Performance Indicators with Industry Standards	242
	Wrap the Dashboard Metrics with Contextual Metadata	242
	Validate the Dashboard Design by a Usability Specialist	243
	Prioritize and Rank Alerts/Exceptions Streamed to the Dashboard	243
	Enrich the Dashboard with Business-User Comments	243

Present Information in Three Different Levels	243
Pick the Right Visual Construct Using Dashboard Design Principles	243
Provide for Guided Analytics	243
<i>Chapter Highlights</i>	244 • <i>Key Terms</i> 244 • <i>Questions for Discussion</i> 245 • <i>Exercises</i> 245 • <i>References</i> 247

PART III Predictive Analytics

Chapter 5 Predictive Analytics I: Data Mining Process, Methods, and Algorithms 250

5.1	Opening Vignette: Police Departments Are Using Predictive Analytics to Foresee and Fight Crime	251
5.2	Data Mining Concepts and Applications	254
	Definitions, Characteristics, and Benefits	256
	How Data Mining Works	257
	Data Mining versus Statistics	261
5.3	Data Mining Applications	261
5.4	Data Mining Process	264
	Step 1: Business Understanding	265
	Step 2: Data Understanding	265
	Step 3: Data Preparation	266
	Step 4: Model Building	266
	► ANALYTICS IN ACTION 5.1 Data Mining Helps in Cancer Research	266
	Step 5: Testing and Evaluation	269
	Step 6: Deployment	269
	Other Data Mining Standardized Processes and Methodologies	269
5.5	Data Mining Methods	272
	Classification	272
	Estimating the True Accuracy of Classification Models	273
	Cluster Analysis for Data Mining	280
	Association Rule Mining	282
5.6	Data Mining Software Tools	286
	► ANALYTICS IN ACTION 5.2 Data Mining Goes to Hollywood: Predicting Financial Success of Movies	287
5.7	Data Mining Privacy Issues, Myths, and Blunders	292
	► ANALYTICS IN ACTION 5.3 Predicting Customer Buying Patterns—the Target Story	293
	Data Mining Myths and Blunders	293
	<i>Chapter Highlights</i> 296 • <i>Key Terms</i> 297 • <i>Questions for Discussion</i> 297 • <i>Exercises</i> 297 • <i>References</i> 300	

Chapter 6 Predictive Analytics II: Text, Web, and Social Media Analytics 301

- 6.1 Opening Vignette: Machine versus Human on *Jeopardy!*: The Story of Watson 302
- 6.2 Text Analytics and Text Mining Overview 305
 - ▶ **TECHNOLOGY INSIGHTS 6.1** Text Mining Terminology 307
- 6.3 Natural Language Processing (NLP) 308
 - ▶ **ANALYTICS IN ACTION 6.1** Deliver Innovation by Understanding Customer Sentiments 310
- 6.4 Text Mining Applications 313
 - Marketing Applications 313
 - Security Applications 313
 - ▶ **ANALYTICS IN ACTION 6.2** Mining for Lies 314
 - Biomedical Applications 316
 - Academic Applications 317
- 6.5 Text Mining Process 318
 - Task 1: Establish the Corpus 319
 - Task 2: Create the Term–Document Matrix 319
 - Task 3: Extract the Knowledge 321
 - ▶ **ANALYTICS IN ACTION 6.3** Research Literature Survey with Text Mining 323
- 6.6 Sentiment Analysis and Topic Modeling 326
 - Sentiment Analysis 326
 - ▶ **ANALYTICS IN ACTION 6.4** Creating a Unique Digital Experience to Capture the Moments That Matter at Wimbledon 327
 - Sentiment Analysis Applications 330
 - Sentiment Analysis Process 332
 - Methods for Polarity Identification 334
 - Using a Lexicon 334
 - Using a Collection of Training Documents 335
 - ▶ **TECHNOLOGY INSIGHTS 6.2** Large Textual Data Sets for Predictive Text Mining and Sentiment Analysis 336
 - Identifying Semantic Orientation of Sentences and Phrases 336
 - Identifying Semantic Orientation of Documents 336
 - Topic Modeling 336
 - Latent Dirichlet Allocation 338
- 6.7 Web Mining Overview 339
 - Web Content and Web Structure Mining 342
- 6.8 Search Engines 344
 - Anatomy of a Search Engine 344
 - 1. Development Cycle 344
 - 2. Response Cycle 346
 - Search Engine Optimization 347

▶ TECHNOLOGY INSIGHTS 6.3	Top 15 Most Popular Search Engines	348
	Methods for Search Engine Optimization	348
6.9	Web Usage Mining (Web Analytics)	349
	Web Analytics Technologies	350
	Web Analytics Metrics	351
	Web Site Usability	351
	Traffic Sources	352
	Visitor Profiles	353
	Conversion Statistics	353
6.10	Social Analytics	355
	Social Network Analysis	355
	Social Network Analysis Metrics	356
	Connections	356
	Distributions	357
	Segmentation	357
	Social Media Analytics	357
	How Do People Use Social Media?	358
▶ ANALYTICS IN ACTION 6.5	Increasing the Efficiency of Social Media Campaigns	359
	Measuring the Social Media Impact	361
	Best Practices in Social Media Analytics	361
	<i>Chapter Highlights</i>	363 • <i>Key Terms</i>
	<i>for Discussion</i>	364 • <i>Exercises</i>
	<i>364</i> • <i>References</i>	365

Chapter 7 Deep Learning and Cognitive Computing 368

7.1	Opening Vignette: Fighting Fraud with Deep Learning and Artificial Intelligence	369
7.2	Introduction to Deep Learning	373
7.3	Basics of “Shallow” Neural Networks	376
▶ ANALYTICS IN ACTION 7.1	Gaming Companies Use Data Analytics to Score Points with Players	379
▶ TECHNOLOGY INSIGHTS 7.1	Elements of an Artificial Neural Network	381
7.4	Process of Developing Neural Network–Based Systems	383
	Learning Process in ANN	385
	Backpropagation for ANN Training	386
7.5	Illuminating the Black Box of Ann	389
▶ TECHNOLOGY INSIGHTS 7.2	Explanability and Transparency in Machine Learning Models	389
▶ ANALYTICS IN ACTION 7.2	Sensitivity Analysis Reveals Injury Severity Factors in Traffic Accidents	393
7.6	Deep Neural Networks	395
	Feedforward Multilayer Perceptron (Mlp)-Type Deep Networks	396
	Impact of Random Weights in Deep MLP	396
	More Hidden Layers versus More Neurons?	397

- ▶ **ANALYTICS IN ACTION 7.3** Georgia DOT Variable Speed Limit Analytics Help Solve Traffic Congestions 398
- 7.7** Convolutional Neural Networks 400
 - Convolution Function 401
 - Pooling 403
 - Image Processing Using Convolutional Networks 404
 - ▶ **ANALYTICS IN ACTION 7.4** From Image Recognition to Face Recognition 407
 - Text Processing Using Convolutional Networks 408
- 7.8** Recurrent Networks and Long Short-Term Memory Networks 411
 - LSTM Networks Applications 415
 - ChatGPT 416
 - What Is ChatGPT and How Does It Work? 416
 - Applications of ChatGPT 418
 - Limitations of ChatGPT 419
- 7.9** Computer Frameworks for Implementation of Deep Learning 420
 - Torch 421
 - Caffe 421
 - TensorFlow 422
 - Theano 422
 - Keras: An Application Programming Interface 422
 - ▶ **ANALYTICS IN ACTION 7.5** A Deep Learning Approach to Predicting Early Bounce-Backs to the Emergency Departments 423
- 7.10** Cognitive Computing 425
 - How Does Cognitive Computing Work? 426
 - How Does Cognitive Computing Differ from AI? 427
 - Cognitive Search 429
 - An Example of Cognitive Computing: IBM Watson 429
 - Chapter Highlights 434 • Key Terms 436 • Questions for Discussion 436 • Exercises 437 • References 438*

PART IV

Prescriptive Analytics

Chapter 8 Prescriptive Analytics: Optimization and Simulation 441

- 8.1** Opening Vignette: Balancing Delivery Routes, Production Schedules, and Inventory 442
- 8.2** Model-Based Decision-Making 444
 - ▶ **ANALYTICS IN ACTION 8.1** Canadian Football League Optimizes Game Schedule 445
 - Prescriptive Analytics Model Examples 446
 - Identification of the Problem and Environmental Analysis 446
 - ▶ **ANALYTICS IN ACTION 8.2** Refinery Modeling 447
 - Model Categories 448

8.3	Structure of Mathematical Models for Decision Support	450
	The Components of Decision Support Mathematical Models	450
	The Structure of Mathematical Models	451
8.4	Certainty, Uncertainty, and Risk	452
	Decision-Making under Certainty	452
	Decision-Making under Uncertainty	453
	Decision-Making under Risk (Risk Analysis)	453
8.5	Decision Modeling with Spreadsheets	453
	► ANALYTICS IN ACTION 8.3 Pennsylvania Adoption Exchange Uses Spreadsheet Model to Better Match Children with Families	454
	► ANALYTICS IN ACTION 8.4 Metro Meals on Wheels Treasure Valley Uses Excel to Find Optimal Delivery Routes	455
8.6	Mathematical Programming Optimization	458
	► ANALYTICS IN ACTION 8.5 Mixed-Integer Programming Model Helps the University of Tennessee Medical Center with Scheduling Physicians	458
	Linear Programming Model	459
	► TECHNOLOGY INSIGHTS 8.1 Linear Programming	460
	Modeling in LP: An Example	460
	Implementation	465
8.7	Multiple Goals, Sensitivity Analysis, What-If Analysis, and Goal Seeking	467
	Multiple Goals	467
	Sensitivity Analysis	468
	What-If Analysis	469
	Goal Seeking	470
8.8	Decision Analysis with Decision Tables and Decision Trees	471
	Decision Tables	471
	Decision Trees	472
8.9	Introduction to Simulation	473
	Major Characteristics of Simulation	473
	► ANALYTICS IN ACTION 8.6 Steel Tubing Manufacturer Uses a Simulation-Based Production Scheduling System	474
	Advantages of Simulation	475
	Disadvantages of Simulation	475
	The Methodology of Simulation	475
	Simulation Types	477
	Monte Carlo Simulation	478
	Discrete Event Simulation	478
	Conventional Simulation Inadequacies	478
	Visual Interactive Simulation	479
	Visual Interactive Models and DSS	479
	Simulation Software	479

- 8.10 Genetic Algorithms and Developing GA Applications 480**
 - Terminology of Genetic Algorithms 482
 - How Do Genetic Algorithms Work? 483
 - Genetic Algorithm Applications 485
 - Chapter Highlights 485 • Key Terms 486 • Questions for Discussion 486 • Exercises 487 • References 493*

PART V

Software and Trends

Chapter 9 Landscape of Business Analytics Tools 494

- 9.1 Opening Vignette: How Seagate Is Using Knime to Tackle the Digital Transformation 495**
- 9.2 Importance of Analytics Tools 497**
 - A Multidimensional Categorization of Analytics Tools 498
 - Popularity of the Analytics Tools 501
 - **ANALYTICS IN ACTIONS 9.1** Predictive Analytic and Data Mining Help Stop Terrorist Funding 504
- 9.3 Free and Open-Source Analytics' Programming Languages 505**
 - R 505
 - How to Get Start With R? 505
 - TUTORIAL 507
 - R for Analytics Application Tutorial—Predicting Employee Attrition 507
 - Data Visualization 510
 - Machine Learning for Predictive Modeling 512
 - Explainable AI 515
 - Rattle (for visual analytics programming in R) 516
 - Python 518
 - How to get started with Python? 520
 - TUTORIAL 521
 - Python for Analytics, Application Tutorial—Predicting Movie Financial Success 521
- 9.4 Free and Open-Source Analytics' Visual Tools 527**
 - KNIME 527
 - **ANALYTICS IN ACTION 9.2** Leveraging Predictive Analytics Prevents \$1.3 Million Worth of Medical Supply Waste 534
 - TUTORIAL 535
 - Predicting Customer Churn 535
 - Orange 541
 - TUTORIAL 542
 - Text Mining of Published Literature 542
 - Step 1. Gathering and cleaning the textual data 543
 - Step 2. Preprocessing the textual data 544

	Step 3. Performing the text analysis	546
	Step 4. Reporting of the obtained results	546
	Weka	547
	RapidMiner	549
	TUTORIAL	550
	Predicting Survival using the Titanic Data Set	550
9.5	Commercial Analytics Tools	552
	Alteryx	553
	IBM	555
	SAS	555
	JMP	558
	TUTORIAL	558
	Text Mining of Published Literature	558
	Step 1. Gathering and cleaning the textual data	558
	Step 2. Preprocessing the textual data	558
	Step 3. Performing the text analysis	560
	Step 4. Reporting of the obtained results	561
	A Comparison of JMP Pro and Orange	563
	Teradata	564
	Analytic Engines and Functions	564
	► ANALYTICS IN ACTION 9.3 Analyzing Disease Patterns from an Electronic Medical Records Data Warehouse	566
	TIBCO	567
	Other Analytics Tools	569
	<i>Chapter Highlights</i>	570 • <i>Key Terms</i> 571 • <i>Questions for Discussion</i> 572 • <i>Exercises</i> 572 • <i>References</i> 573

Chapter 10 AI-Based Trends in Analytics and Data Science 574

10.1	Application Vignette: Discover Foods Employs IoT and Machine Learning to Ensure Food Quality	575
10.2	Cloud-Based Analytics	576
	Data as a Service (DaaS)	579
	Desktop as a Service (DaaS)	580
	Software as a Service (SaaS)	580
	Platform as a Service (PaaS)	580
	Infrastructure as a Service (IaaS)	580
	Essential Technologies for Cloud Computing	581
	► ANALYTICS IN ACTION 10.1 Major West Coast Utility Uses Cloud-Mobile Technology to Provide Real-Time Incident Reporting	582
	Cloud Deployment Models	584
	Major Cloud Platform Providers in App Development and Deployment	584
	Analytics as a Service (AaaS)	585

	Representative Analytics as a Service Offerings	586
	▶ ANALYTICS IN ACTION 10.2 GO-JEK Employs Google Cloud Platform for Forecasting and Pricing	587
	Illustrative Analytics Applications Employing the Cloud Infrastructure	587
	▶ Using Azure IoT (Internet of Things), Stream Analytics, and Machine Learning to Improve Mobile Healthcare Services	587
	▶ Chime Enhances Customer Experience Using Snowflake	588
10.3	Location-Based Analytics	589
	Geospatial Analytics	589
	▶ ANALYTICS IN ACTION 10.3 Improving Home Loan Appraisal Process Using BI and Geographic Information Systems	591
	▶ ANALYTICS IN ACTION 10.4 Starbucks Exploits GIS and Analytics to Grow Worldwide	592
	A Multimedia Exercise in Analytics Employing Geospatial Analytics	593
	Real-Time Location Intelligence	594
	Analytics Applications for Consumers	595
10.4	Image Analytics/Alternative Data	596
	▶ ANALYTICS IN ACTION 10.5 Image Analysis Helps Estimate Plant Cover	597
	▶ ANALYTICS IN ACTION 10.6 How Unilever Used Image Analytics in Its Supply Chain to Examine Deforestation	598
10.5	IoT Essentials	599
	Definitions and Characteristics	599
	The IoT Ecosystem	600
	Structure of IoT Systems	601
	MAJOR BENEFITS AND DRIVERS OF IoT	603
	Major Benefits of IoT	603
	Major Drivers of IoT	603
	How IoT Works	604
	IoT and Decision Support	605
	Sensors and Their Role in IoT	605
	Brief Introduction to Sensor Technology	605
	How Sensors Work with IoT	605
	▶ ANALYTICS IN ACTION 10.7 Rockwell Automation Monitors Expensive Oil and Gas Exploration Assets to Predict Failures	605
	Sensor Applications and Radio-Frequency Identification (RFID) Sensors	606
	▶ TECHNOLOGY INSIGHTS 10.1 RFID Sensors	607
	Use of RFID and Smart Sensors in IoT	607
10.6	IoT Applications	608
	Smart Homes and Appliances	609
	Typical Components of Smart Homes	609
	Smart Appliances	610
	A Smart Home Is Where the Bot Is	611
	Barriers to Smart Home Adoption	611

	Smart Components in Smart Cities and Smart Factories	611
	Improving Transportation in a Smart City	612
	Autonomous (Self-Driving) Vehicles	613
	Implementation Issues in Autonomous Vehicles	613
	The Future of the IoT	613
10.7	5G Technologies and Impact on AI	614
10.8	Other Emerging AI Topics: Robotic Process Automation (RPA)	616
	▶ ANALYTICS IN ACTION 10.8 Monitoring Quarantined Persons in the State District	617
	▶ ANALYTICS IN ACTION 10.9 Walgreens' HR Shared Service Efficiency Increased by 73% by Blue Prism's Digital Workforce	617
10.9	Bioinformatics and Health Network Science	618
	▶ ANALYTICS IN ACTION 10.10 Analyzing the Genomics Data to Optimize Animals' Gut Health and Performance	618
NETWORK ANALYTICS 620		
	▶ ANALYTICS IN ACTION 10.11 Network Analytics for Predictive Modeling: Predicting Hospital Length of Stay using Comorbidity Networks	621
10.10	Other Recent Developments	622
	Web 3.0	622
	Metaverse and Digital Twins	623
	GPT-3/ChatGPT	624
	LaMDA	624
	Blenderbot	624
	<i>Chapter Highlights</i>	625 • <i>Key Terms</i> 626
	• <i>Exercises</i> 627 • <i>References</i> 627	

Chapter 11 Ethical, Privacy, and Managerial Considerations in Analytics 629

11.1	Opening Vignette: Lessons Learned from Analytics Journey in an Organization	630
11.2	Implementing Intelligent Systems: An Overview	633
	The Intelligent Systems Implementation Process	633
11.3	Successful Deployment of Intelligent Systems	634
	Top Management and Implementation	634
	System Development Implementation Issues	635
	Connectivity and Integration	635
	Security Protection	636
	Leveraging Intelligent Systems in Business	636
	Intelligent System Adoption	637
11.4	Implementing IoT and Managerial Considerations	637
	Major Implementation Issues	637
	Strategy for Turning Industrial IoT into Competitive Advantage	638
11.5	Legal, Privacy, and Ethical Issues	640
	Legal Issues	640

A Sample of AI Potential Legal Issues	640
Privacy Issues	642
Who Owns Our Private Data?	644
Ethics Issues	644
Ethical Issues of Intelligent Systems	645
Other Topics in Intelligent Systems Ethics	646
11.6 Ethical/Responsible/Trustworthy AI	646
▶ The Curious Case Of BlenderBot 3.0	647
The O’Neil Claim of Potential Analytics’ Dangers	650
11.7 Impacts of Intelligent Systems on Organizations	651
New Organizational Units and Their Management	651
Transforming Businesses and Increasing Competitive Advantage	652
▶ ANALYTICS IN ACTION 11.1 How 1-800-Flowers.com Uses Intelligent Systems for Competitive Advantage	652
Redesign of an Organization through the Use of Analytics	654
Intelligent Systems’ Impact on Managers’ Activities, Performance, and Job Satisfaction	654
Impact on Decision-Making	655
Industrial Restructuring	655
11.8 Impacts on Jobs and Work	656
An Overview	656
Are Intelligent Systems Going to Take Jobs—My Job?	657
AI Puts Many Jobs at Risk	658
▶ ANALYTICS IN ACTION 11.2 Administrative/Office Jobs That Robots Have Already Taken	658
Which Jobs Are Most in Danger? Which Ones Are Safe?	659
Some More Job Losses Observations	659
Intelligent Systems May Actually Add Jobs	659
Jobs and the Nature of Work Will Change	660
Conclusion: Let Us Be Optimistic!	662
11.9 Potential Dangers of AI	662
Position of AI Dystopia	662
The AI Utopia’s Position	663
11.10 Citizen Science and Citizen Data Scientists	664
Citizen Science	664
Citizen Data Scientist	664
Concluding Remarks	666
<i>Chapter Highlights</i>	666 • <i>Key Terms</i> 667 • <i>Questions for Discussion</i> 667 • <i>Exercises</i> 667 • <i>References</i> 668
Glossary	670
Index	680

Analytics, Data Science, and Artificial Intelligence (AI) have become the technology drivers of today. All major companies and organizations are creating new organizational units focused on analytics to help them become more effective and efficient in their operations. Decision makers are using more computerized tools to support their work, and even consumers are using analytics tools, either directly or indirectly, to make decisions on routine activities such as shopping, health/healthcare, travel, and entertainment.

The field of decision support systems, business intelligence, and business analytics (BI & BA) has evolved rapidly to become more focused on innovative applications for extracting knowledge and insight from data streams that were not even captured until recently, much less analyzed in significant ways. New applications turn up daily in all types of industries—healthcare, sports, travel, entertainment, supply-chain management, utilities, etc.

The term *analytics* has evolved into other terms such as data science and newer incarnations: artificial intelligence, deep learning, and the Internet of Things. With ChatGPT account registrations jumping to more than 100 million in just two months after its public launch, it is clear that AI technologies are becoming more mainstream.

Recently, AI has become a popular buzzword again with much promise for impacting business and personal lives. Our other book, *Analytics, Data Science, and Artificial Intelligence*, 11th ed, 2017 by Sharda, Delen, and Turban (referred hereafter as DSS11), introduced many of these current topics. Previous edition of this book, *Business Intelligence, Analytics, and Data Science: A Managerial Approach*, 4th ed, by Sharda, Dursun, and Turban, covered the core areas from a managerial perspective. Based on the convergence of all of these topical areas and the need for a managerial understanding, we have combined the two books into this new 5th edition, now called *Business Intelligence, Analytics, Data Science, and AI*, 5th ed.

This new edition integrates and updates the content of both books by streamlining the coverage and creating a more focused book that covers most of the important topics in AI, data science, and analytics. Thus, the *theme of this revised edition is analytics, data science, and AI for enterprise decision support*. This edition of the text provides a managerial perspective to a business analytics continuum beginning with *descriptive analytics* (e.g., the nature of data, statistical modeling, data visualization, and business intelligence), moving on to *predictive analytics* (e.g., data mining, text/web mining, social media mining), and then to *prescriptive analytics* (e.g., optimization and simulation).

In addition, AI, deep learning, robotics, chatbots, IoT, and Web/Internet-related enablers are introduced throughout the text. We highlight these technologies as emerging components of modern-day business analytics systems. AI technologies have a major impact on decision-making by enabling autonomous decisions and by supporting steps in the process of making decisions. AI and analytics support each other by creating a synergy that assists decision-making. We conclude with future trends, privacy, and managerial considerations. The book is supported by a Web site (pearsonhighered.com/sharda).

The purpose of this book is to introduce the reader to the technologies generally called *business analytics* or *data science*. We use these terms interchangeably. This book presents the fundamentals of the methods, methodologies, and techniques used to design and develop these systems. In addition, we introduce the essentials of AI both as it relates to analytics as well as a standalone discipline for decision support. This book presents the fundamentals of the techniques and the manner in which these systems are constructed and used.

We follow an EEE approach to introducing these topics: **Exposure, Experience, and Exploration**. The book primarily provides **exposure** to various analytics techniques and their applications. The idea is that a student will be inspired to learn from how other

organizations have employed analytics to make decisions or to gain a competitive edge. We believe that such **exposure** to what is being done with analytics and how it can be achieved is the key component of learning about analytics. In describing the techniques, we also introduce specific software tools that can be used for developing such applications. The book is not limited to any one software tool, so the students can **experience** these techniques using any number of available software tools. Specific suggestions are given in each chapter, but the student and the professor are able to use this book with many different software tools. Chapter 9 of the book includes an introduction to several software tools, but students can gain **experience** with these techniques in many different ways. Finally, we hope that this **exposure** and **experience** enable and motivate readers to **explore** the potential of these techniques in their own domain. To facilitate such **exploration**, we include activities and links to Web sites that contain team-oriented exercises where appropriate. Some chapters include links for material available at the Teradata University for Academics Web site.

- ***How to Use the New Edition? Use in a one-semester overview course*** on analytics/AI, which covers most, or all, of the topics/chapters included in the book. Such managerial coverage would provide an introductory overview of many of the topics and let the faculty member introduce any hands-on tools and exercises for selected topics to deepen coverage of a specific area.
For example, a general introductory class could introduce just the data management and visualization tools, besides providing managerial introduction to the other topics.
- ***Use it to teach two consecutive courses***
For example, one course could focus on the overall analytics coverage. It could cover selective sections of Chapters 1 and 3–9.
A second course could focus on artificial intelligence and emerging technologies as the enablers of modern-day analytics as a subsequent course to the first course. This second course could cover portions of Chapters 1, 2, 7, and 8–11, and introduce software tools for some of the techniques covered in these chapters.
- ***Use it to offer managerial-level exposure*** to applications and techniques as noted previously, but it also includes sufficient technical details in selected chapters to allow an instructor to focus on some technical methods and hands-on exercises.

WHAT'S NEW IN THE FIFTH EDITION?

Most of the specific improvements made in this fifth edition concentrate on four areas:

- Reorganization
- New chapters
- Content updates
- Sharper focus

Despite the many changes, we have preserved the comprehensiveness and user friendliness that have made the text a market leader. Finally, we present accurate and updated material that is not available in any other textbook.

As we noted above, this revision integrates and updates content from both books: DSS11 and BI5e. With the goal of improving the text, this edition marks a major reorganization of the text to reflect the focus on business analytics, and therefore, is organized around three major types of business analytics: *descriptive*, *predictive*, and *prescriptive*. The new edition has many timely additions and some of the dated content has been deleted. (We have retained some content because it helps to see how a particular topic's growth has evolved.) The following major specific changes have been made.

- ***New content and organization.*** The book is organized around the three types of analytics: descriptive, predictive, and prescriptive, a classification promoted by INFORMS. Besides updating content in previously included chapters, it includes several new chapters for this book. Some material appeared in our DSS11 book, but all of that has been updated as well for inclusion in this book. The book now includes 11 chapters organized into five sections:

Section 1 includes the introduction/overview of all topics in the analytics/data science/AI space, including a new chapter on overview of AI topics.

Section 2 covers data as the foundations of analytics and descriptive analytics.

Section 3 includes coverage of predictive analytics, including a new chapter on deep learning.

Section 4 introduces prescriptive analytics including optimization, simulation, and heuristic methods.

Section 5 is titled Software and Trends. It includes a new chapter (Chapter 9) covering the landscape of software tools and programming languages for analytics. Another largely new chapter includes coverage of current and future trends in AI and data science. The book concludes with ethical, privacy, and managerial considerations in analytics. More details follow.

- ***New chapters.*** The following chapters have been added to this book (as noted above, some of the material in these chapters has been updated from our DSS11):

Chapter 2. Artificial Intelligence, Robotics, and Smart Systems. It introduces various topics in AI. It includes coverage of topics that are relevant for analytics/data science but are typically not included in the traditional introductory courses—conversational AI, chatbots such as chatGPT, robo-advisors, and robotics. This chapter covers the essentials of AI, outlines its benefits, compares it with humans' intelligence, and describes the content of the field. Example applications in accounting, finance, human resource management, marketing and CRM, and production-operation management illustrate the benefits to business (95% new for this book).

Chapter 7. Deep Learning, Explainable AI, and Cognitive Computing This chapter covers the generation of machine learning technique, deep learning as well as the increasingly more popular AI topic, cognitive computing. It is an almost entirely new chapter (90% new material).

Chapter 9. The Landscape of Software Tools and Languages This is an entirely new chapter covering the landscape of tools for analytics/data science. It includes an introduction to R/Python, KNIME, JMP, and a number of other tools. Although it is not possible to cover all available tools in any one book, this chapter can help an instructor explore some options for which some guidance is now included in the book itself (100% new).

Chapter 10. New AI-Based Trends in Analytics and Data Science This chapter includes updates on some content from Chapter 8 of the last edition including coverage of geospatial analytics, Internet of Things, and a significant update of the material on cloud-based analytics. In addition, it includes AI-based trends in Analytics and Data Science, IoT, 5G, Robotic Process Automation, chatbots, metaverse, etc. (60% new).

- ***Revised Chapters.*** All the other chapters have been revised and updated as well. Here is a summary of the changes in these other chapters:

Chapter 1. An Overview of Business Intelligence, Analytics, and Data Science This chapter has been rewritten and significantly expanded. The opening vignette now includes sports analytics applications in international sports as well as

additional decision problems in sports analytics. It introduces the three types of analytics as proposed by INFORMS: descriptive, predictive, and prescriptive analytics. As noted earlier, this classification is used in guiding the complete reorganization of the book itself (earlier content but with a new figure). Then it includes several new examples of analytics in agriculture and gaming industry, besides updating the previous applications in healthcare and in the retail industry (about 40% new material).

Chapter 3. *Nature of Data, Big Data, and Statistical Modeling* This chapter updates the earlier content but also includes coverage of Big Data early in the book (using the material adapted from the DSS11 book). It also includes a section on “data for good.” It is restructured to include basic statistical methods; the data visualization is removed from this chapter. The application cases (Analytics in Actions) are reduced in number, some are kept, and new cases added (about 10% new material).

Chapter 4. *Business Intelligence, Data Warehousing, and Data Visualization* This chapter includes several new cases (about 15% new material).

Chapter 5. *Predictive Analytics I: Data Mining Process, Methods, and Algorithms* This is an old chapter with some new content organization/flow and some new figures (about 5% new material).

Chapter 6. *Predictive Analytics II: Text, Web, and Social Media Analytics* Includes some restructuring of content and new material on sentiment analysis and topic modeling (about 20% new).

Chapter 8. *Prescriptive Analytics* This chapter includes a new opening vignette and some new application cases. The coverage of simulation has been streamlined with new application cases. A new section on Genetic Algorithms has been added. Finally, it also includes several additional exercises for optimization and simulation modeling (about 30% new).

Chapter 11. *Ethical, Privacy, and Managerial Considerations in Analytics* This chapter updates the earlier content from Chapter 9 of the previous edition to include new issues related to responsible AI, additional examples of concerns about AI’s impact, and the like. It also adapts material from the DSS11 book to cover issues such as the protection of privacy, intellectual property, ethics, technical issues (e.g., integration and security), and administrative issues. We also cover the impact of these technologies on organizations and people and specifically deal with the impact on work and jobs. Special attention is given to possible unintended impacts of analytics and AI (robots). The chapter also includes a reflective narrative from an industry professional who has worked on many analytics projects for his company to identify broad lessons learned from an analytics journey in an organization (about 50% new).

NEW AND UPDATED FEATURES

- **Pearson eTextbook offered in a variety of purchase plans.**

The Pearson eTextbook engages students with a visually open design, ability to zoom in images, highlighting, note taking, and key vocabulary review. Using the Pearson app for Android or iPhone, students have access to their eTextbook anytime, anywhere, even offline. Instructors can also engage students through the eTextbook by sharing their notes and highlights with the students. <https://www.pearson.com/en-us/pearsonplus.html>

- **Full Color Design.** We are truly excited to have this book appear in full color. All of the new and previous figures are presented in color to enhance the many visualization examples and other material.

- **Software support.** The book does not promote any specific software tool. Based on the in-depth coverage of the content, case studies, and examples, the instructor and the students can choose to use their favorite analytics platforms, or can venture into

many commercial software products offered by industry leader companies such as SAS, IBM, TIBCO, Tableau, Teradata, Microsoft or opt for free/open-source workflow-type tools and high-level analytics languages such as KNIME, Python, and R.

Instructor Resources: www.pearsonhighered.com/sharda

A comprehensive and flexible technology-support package is available to enhance the teaching and learning experience. The following instructor and student supplements are available on the book's Web site, [pearsonhighered.com/sharda](http://www.pearsonhighered.com/sharda):

- **Test Bank and TestGen Software.** The Test Bank is a comprehensive collection of true/false, multiple-choice, fill-in-the-blank, and essay questions. The questions are rated by difficulty level, and the answers are referenced by book page number. The Test Bank is available in Microsoft Word and in TestGen. Pearson Education's test-generating software is available from www.pearsonhighered.com/irc. The software is PC/MAC compatible and preloaded with all the Test Item File questions. You can manually or randomly view test questions and drag-and-drop to create a test. You can add or modify test-bank questions as needed. Our TestGens are converted for use in a variety of Learning Management Systems.
- **PowerPoint Lectures.** PowerPoint Lecture slides are available for each chapter to help illuminate and build on key concepts in the text. Faculty can download the PowerPoint slides from [pearsonhighered.com/sharda](http://www.pearsonhighered.com/sharda) and share through their LMS.
- **Image Library.** All of the images from the textbook are included so that instructors can use them in their lectures or to create their own PowerPoint presentations based on the textbook content.

ACKNOWLEDGMENTS

Many individuals have provided suggestions and criticisms since the publication of the first edition of this book. Dozens of students participated in class testing of various chapters, software, and problems and assisted in collecting material. It is not possible to name everyone who participated in this project, but our thanks go to all of them. Certain individuals made significant contributions, and they deserve special recognition.

First, we appreciate the efforts of those individuals who provided formal reviews of the first through fourth editions (school affiliations as of the date of review):

Ann Aksut, Central Piedmont Community College
 Bay Arinze, Drexel University
 Andy Borchers, Lipscomb University
 Ranjit Bose, University of New Mexico
 Marty Crossland, Mid America Nazarene University
 Kurt Engemann, Iona College
 Badie Farah, Eastern Michigan University
 Gary Farrar, Columbia College
 Jerry Fjermestad, New Jersey Institute of Technology
 Christie M. Fuller, Louisiana Tech University
 Martin Grossman, Bridgewater State College
 Jahangir Karimi, University of Colorado, Denver
 Huei Lee, Eastern Michigan University
 Natalie Nazarenko, SUNY Fredonia
 Joo Eng Lee-Partridge, Central Connecticut State University
 Gregory Rose, Washington State University, Vancouver
 Khawaja Saeed, Wichita State University
 Kala Chand Seal, Loyola Marymount University

Joshua S. White, PhD, State University of New York Polytechnic Institute
 Roger Wilson, Fairmont State University
 Vincent Yu, Missouri University of Science and Technology
 Fan Zhao, Florida Gulf Coast University

We also appreciate the efforts of those individuals who provided formal reviews of this text and our other DSS11 book—*Analytics, Data Science, and Artificial Intelligence: Systems for Decision Support*. New York: Pearson Publishing Company, 11th Edition, Pearson Education, 2017.

Second, several individuals contributed material to the text or the supporting material. Susan Baskin and Yenny Yang of Teradata provided special help in identifying new Teradata content for the book and arranging permissions for the same. Dr. Dave Schrader contributed the opening vignette for the book. It will provide a great introduction to analytics. The original vignette included material developed by Dr. Ashish Gupta of Auburn University and Gary Wilkerson of the University of Tennessee–Chattanooga. The updated vignette includes material prepared by students who worked on sports analytics projects at various universities including Oklahoma State, Auburn, LSU, and Purdue. The students are credited under the sports analytics vignette. Besides Dr. Dave Schrader, we want to also acknowledge supervision by Dr. Miriam McGough at Oklahoma State and Mr Fawad Qureshi of Teradata-London. We also thank INFORMS for their permission to highlight content from *Interfaces*. Assistance from the following colleagues and friends is recognized for this edition of the book: Enes Eryarsoy, Sabanci University (help with Analytics With R section in Chapter 9); Nurettin Oner, University of Alabama–Birmingham (help with Orange and JMP sections in Chapter 9); Rosaria Silipo, Stefan Helfrich, Aline Bessa—KNIME (help with the KNIME section in Chapter 9); Behrooz Davazdaemami, University of Wisconsin-Whitewater (help with the Python section in Chapter 9 and with Deep Learning in Chapter 7); Manohar Lahoti, Land O'Lakes (new application stories in Chapter 1 and a reflection on the analytics journey in Chapter 11); Vishal Agrawal, AISera (assistance with application story from AISERA); Pankush Kalgotra, Auburn (new application stories on network data science); Milind Kopikare and the team at Great Learning for an IoT application story from India, Dr. Anand Rao at PwC (a figure on responsible AI); Rudra Ankaiyen (assistance with bibliography); and Asif Syed for an update on ethical issue in AI. Many other specific content contributors are recognized throughout the book.

We also recognize the following individuals for their assistance in developing earlier edition of this book and the companion DSS11 book: Pankush Kalgotra (now at Auburn), Behrooz Davazdahemami, Bhavana Baheti, Varnika Gottipati, Chakradhar Pathi, Prasoon Mathur, Rupesh Agarwal, Shubham Singh, Nan Liang, Jacob Pearson, Kinsey Clemmer, and Evan Murlette (all at Oklahoma State University when this edition was developed). We also acknowledge Jongswas Chongwatpol (NIDA, Thailand) for the material on a Thai credit risk case study. Their help for this edition is gratefully acknowledged. Dr. Brian LeClaire, retired CIO of Humana Corporation, led with contributions of several real-life healthcare case studies developed by his team at Humana. Abhishek Rathi of vCreaTek contributed his vision of analytics in the retail industry. Dr. Rick Wilson's excellent exercises for teaching and practicing linear programming skills in Excel are also gratefully acknowledged. Matt Turck agreed to let us adapt his IoT ecosystem material. Ramesh also recognizes the copyediting assistance provided by his daughter, Ruchy Sharda Sen. In addition, the following former PhD students and research colleagues of ours have provided content or advice and support for the book in many direct and indirect ways:

Asil Oztekin, Universality of Massachusetts–Lowell
 Enes Eryarsoy, Sabanci University
 Hamed Majidi Zolbanin, University of Dayton
 Behrooz Davazdahemami, University of Wisconsin–Whitewater
 Saeed Piri, University of Oregon

Amir Hassan Zadeh, Wright State University
 Supavich (Fone) Pengnate, North Dakota State University
 Christie Fuller, Boise State University
 Daniel Asamoah, Wright State University
 Selim Zaim, Istanbul Zaim University
 Nihat Kasap, Sabanci University
 Haluk Demirkan, University of Washington–Takoma

Third, for the previous editions, we acknowledge the contributions of Dave King (JDA Software Group, Inc.). Other major contributors to the previous editions include J. Aronson (University of Georgia), who was our coauthor, contributing to the data warehousing chapter; Jerry Wagner (University of Nebraska–Omaha), who provided PlannersLab software; Leila A. Halawi (Bethune-Cookman College), who provided material for the chapter on data warehousing; late Mike Goul (Arizona State University), whose contributions were included in Chapter 1; and late T. P. Liang (National Sun Yet-Sen University–Taiwan), who contributed material on neural networks in the previous editions. Judy Lang collaborated with all of us, provided editing, and guided us during the entire project in the first edition.

Fourth, several vendors cooperated by providing case studies and/or demonstration software for the previous editions: Acxiom (Little Rock, Arkansas); California Scientific Software (Nevada City, California); Cary Harwin of Catalyst Development (Yucca Valley, California); IBM (San Carlos, California); DS Group, Inc. (Greenwich, Connecticut); Gregory Piatetsky-Shapiro of KDnuggets.com; Gary Lynn of NeuroDimension Inc. (Gainesville, Florida); Palisade Software (Newfield, New York); Promised Land Technologies (New Haven, Connecticut); Salford Systems (La Jolla, California); Sense Networks (New York, New York); Gary Miner of StatSoft, Inc. (Tulsa, Oklahoma); Ward Systems Group, Inc. (Frederick, Maryland); Idea Fisher Systems, Inc. (Irving, California); and Wordtech Systems (Orinda, California).

Fifth, special thanks to the Teradata University for Academics and especially to Susan Baskin and Yenny Yang; Hugh Watson, who started Teradata University Network (TUN); and late Michael Goul, Barb Wixom, and Mary Gros for their encouragement to tie this book with TUN and for providing useful material for the book.

Finally, the Pearson team is to be commended: Jenifer Niles, who has worked with us on the new edition; and the production team, Guneet Gulati specifically and the team at Integra, led by Gowthaman Sadhanandham, all of whom helped turn the manuscript into a book.

We would like to thank all these individuals and corporations. Without their help, the creation of this book would not have been possible.

R.S.

D.D.

E.T.

Note that Web site URLs are dynamic. As this book went to press, we verified that all the cited Web sites were active and valid. Web sites to which we refer in the text sometimes change or are discontinued because companies change names, are bought or sold, merge, or fail. Sometimes Web sites are down for maintenance, repair, or redesign. Most organizations have dropped the initial “www” designation for their sites, but some still use it. If you have a problem connecting to a Web site that we mention, please be patient and simply run a Web search to try to identify the new site. Most times, the new site can be found quickly. We apologize in advance for this inconvenience.



ABOUT THE AUTHORS

Ramesh Sharda (MBA, PhD, University of Wisconsin–Madison) is the Vice Dean for Research and Graduate Programs, Watson/ConocoPhillips Chair, and a Regents Professor of Management Science and Information Systems in the Spears School of Business at Oklahoma State University (OSU). He cofounded and directed OSU's PhD in Business for the Executives Program. About 200 papers describing his research have been published in major journals, including *Operations Research*, *Management Science*, *Information Systems Research*, *Decision Support Systems*, and the *Journal of MIS*. He cofounded the AIS SIG on Decision Support Systems and Knowledge Management (SIGDSA). Dr. Sharda serves on several editorial boards, including those of *Decision Sciences Journal*, *Decision Support Systems*, and *ACM Data Base*. He has authored and edited several textbooks and research books and serves as the coeditor of several Springer book series (Integrated Series in Information Systems, Operations Research/Computer Science Interfaces, and Annals of Information Systems) with Springer. He served as the Executive Director of the Teradata University Network from 2013 to 2020. His current research interests are in decision support systems, business analytics, and technologies for managing information overload. Ramesh is a Fellow of INFORMS and AIS and was inducted into the Oklahoma Higher Education Hall of Fame in 2015. He was awarded a Fulbright Distinguished Chair at Aalto University of Finland for spring 2023.

Dursun Delen (PhD, Oklahoma State University) is the Spears Endowed Chair in Business Administration, Patterson Foundation Endowed Chair in Business Analytics, Director of Research for the Center for Health Systems Innovation, and Regents Professor of Management Science and Information Systems in the Spears School of Business at Oklahoma State University (OSU). Prior to his academic career, he worked for a privately owned research and consultancy company, Knowledge Based Systems Inc., in College Station, Texas, as a research scientist for 5 years, during which he led a number of decision support and other information systems–related research projects funded by several federal agencies including the Department of Defense (DoD), National Aeronautics and Space Administration (NASA), National Institute for Standards and Technology (NIST), Ballistic Missile Defense Organization (BMDO), and Department of Energy (DOE). Dr. Delen has published more than 200 peer-reviewed articles, some of which have appeared in major journals like *Decision Sciences*, *Decision Support Systems*, *Communications of the ACM*, *Computers and Operations Research*, *Computers in Industry*, *Journal of Production Operations Management*, *Artificial Intelligence in Medicine*, *International Journal of Medical Informatics*, *Expert Systems with Applications*, and *IEEE Wireless Communications*. He recently authored/coauthored 12 textbooks in the broad areas of business analytics, data mining, text mining, business intelligence, and decision support systems. He is often invited to national and international conferences for keynote addresses on topics related to data/text mining, business analytics, decision support systems, business intelligence, and knowledge management. He served as the General Cochair for the Fourth International Conference on Network Computing and Advanced Information Management (September 2–4, 2008, in Seoul, South Korea) and regularly chairs, tracks, and minitracks at various information systems and analytics conferences. He is currently serving as the Editor-in-Chief for the *Journal of Business Analytics*, *Journal of AI in Business* (part of Frontiers in AI family of journals), and as Senior Editor, Associate Editor, or Editorial Board Member for more than another dozen of academic journals. His research and teaching interests are in data and text mining, business analytics, decision support systems, data science, knowledge management, business intelligence, and enterprise modeling.

Efraim Turban (MBA, PhD, University of California, Berkeley) is a Visiting Scholar at the Pacific Institute for Information System Management, University of Hawaii. Prior to this, he was on the staff of several universities, including City University of Hong Kong; Lehigh University; Florida International University; California State University, Long Beach; Eastern Illinois University; and the University of Southern California. Dr. Turban is the author of more than 100 refereed papers published in leading journals, such as *Management Science*, *MIS Quarterly*, and *Decision Support Systems*. He is also the author of 20 books, including *Electronic Commerce: A Managerial Perspective* and *Information Technology for Management*. He is also a consultant to major corporations worldwide. Dr. Turban's current areas of interest are Web-based decision support systems, social commerce, and collaborative decision-making.

BUSINESS INTELLIGENCE, ANALYTICS, DATA SCIENCE, AND AI

A MANAGERIAL PERSPECTIVE

This book deals with a collection of computer technologies that support managerial work—essentially, decision making. These technologies have had a profound impact on corporate strategy, performance, and competitiveness. Collectively, these technologies are called *business intelligence*, *business analytics*, *data science*, and *artificial intelligence*. Although the evolution of the terms is discussed, these names are also used interchangeably. This book tells stories of how smart people are employing these techniques to improve performance, service, and relationships in business, government, and non-profit worlds.

An Overview of Business Intelligence, Analytics, Data Science, and AI

LEARNING OBJECTIVES

- Understand the need for computerized support of managerial decision making
- Recognize the evolution of such computerized support to the current state of analytics/data science and artificial intelligence
- Describe the business intelligence (BI) methodology and concepts
- Understand the different types of analytics and see applications in diverse domains

The business environment (climate) is constantly changing, and it is becoming more and more complex. Organizations, both private and public, are under pressures that force them to respond quickly to changing conditions and to be innovative in the way they operate. Such activities require organizations to be agile and to make frequent and quick strategic, tactical, and operational decisions, some of which are very complex. Making such decisions may require considerable amounts of relevant data, information, and knowledge. Processing these in the framework of the needed decisions must be done quickly, frequently in real time, and usually requires significant computerized support. As technologies are evolving, many decisions are being automated, leading to a major impact on knowledge work and workers in many ways.

This book is about using business analytics and artificial intelligence as computerized support for managerial decision making. It concentrates on the theoretical and conceptual foundations of decision support, as well as on the commercial tools and techniques that are available. This book presents the fundamentals of the techniques and the manner in which these systems are constructed and used. We follow an EEE approach to introducing these topics: **Exposure**, **Experience**, and **Exploration**. The book primarily provides exposure to various analytics techniques and their applications. The idea is that

a student will be inspired to learn from how other organizations have employed analytics to make decisions or to gain a competitive edge. We believe that such **exposure** to what is being done with analytics and how it can be achieved is the key component of learning about analytics. In describing the techniques, we also give examples of specific software tools that can be used for developing such applications. The book is not limited to any one software tool, so students can **experience** these techniques using any number of available software tools. We hope that this exposure and experience enable and motivate readers to explore the potential of these techniques in their own domain. To facilitate such **exploration**, we include hands-on exercises and links to other sites that include team-oriented exercises where appropriate.

This chapter introduces analytics and data science as well as an overview of the book. The chapter has the following sections:

- 1.1** Opening Vignette: Sports Analytics—An Exciting Frontier for Learning and Understanding Applications of Analytics 3
- 1.2** Changing Business Environments and Evolving Needs for Decision Support and Analytics 17
- 1.3** Decision-Making Processes and Computerized Decision Support Framework 20
- 1.4** Evolution of Computerized Decision Support to Analytics/Data Science 25
- 1.5** A Framework for Business Intelligence 28
- 1.6** Analytics Overview 34
- 1.7** Analytics Examples in Selected Domains 43
- 1.8** Plan of the Book 58
- 1.9** Resources, Links, and the Teradata University for Academics Connection 59

1.1 OPENING VIGNETTE: Sports Analytics—An Exciting Frontier for Learning and Understanding Applications of Analytics

The application of analytics to business problems is a key skill, one that you will learn in this book. Many of these techniques are now being applied to improve decision making in all aspects of sports, a very hot area called sports analytics. Sports analytics is the art and science of gathering data about athletes and teams to create insights that improve sports decisions, such as deciding which players to recruit, how much to pay them, who to play, how to train them, how to keep them healthy, and when they should be traded or retired. For teams, it involves business decisions such as ticket pricing, as well as opposition research, analysis of each competitor's strengths and weaknesses, and many game-day decisions.

The use of analytics for sports was popularized by the *Moneyball* book by Michael Lewis in 2003 and the movie starring Brad Pitt in 2011. It showcased Oakland A's general manager Billy Beane and his use of data and analytics to turn a losing team into a winner. In particular, he hired an analyst who used analytics to draft players able to get on base as opposed to players who excelled at traditional measures like runs batted in or stolen bases. These insights allowed them to draft prospects overlooked by other teams at reasonable starting salaries. It worked—they made it to the playoffs in 2002 and 2003. (Recently, women's athletics are also beginning to use analytics. For example, see this story: <https://theathletic.com/2547245/2021/04/28/the-rise-of-analytics-in-womens-football/>).

As an Industry, sports worldwide is a multibillion-dollar industry. A report by Statista.com (<https://www.statista.com/statistics/1087391/global-sports-market-size/>) showed the

size of the worldwide sports industry in 2018 itself to be almost \$471 billion. Estimates vary widely, but according to Rice University, a conservative estimate of this industry's current size is about \$500 billion. This estimate includes various professional athletic leagues as well as college sports. According to some reports, college sports in the United States represent a \$18 billion-dollar industry. Suffice it to say that sports is a major economic driver of activity in the United States and in many countries around the world.

Sports analytics is becoming a specialty within analytics. It is an important area because sports is a big business. In 2014, \$125M was spent on analytics. A recent report produced by Grand View Research (<https://www.grandviewresearch.com/industry-analysis/sports-analytics-market>) estimates that in 2020 the sports analytics industry had already grown to \$885m, and is expected to grow at a staggering growth rate of over 27% per year. Thus, sports analytics is not only a fun way to learn about analytics, but it is also a potential career option for many graduates of analytics programs.

Analytics are being used in all parts of sports. Figure 1.1 presents a simple way to understand potential analytics applications in sports. The top row of this figure points out the potential for the usual business office/administrative analytics like allocating budget dollars across multiple sports in colleges, or determining the mix of money spent on facilities vs. coaches and trainers vs. player salaries and benefits. The next two layers of analytics can be divided between the front office and back office (often called Business and Operations). Front-office business analytics include analyzing fan behavior ranging from predictive models for season ticket renewals and regular ticket sale pricing, to scoring tweets by fans regarding the team, athletes, coaches, and owners. This is very similar to traditional customer relationship management (CRM). For individual players, there is a focus on recruitment models and scouting analytics. Financial analysis is also a key area, where salary caps on rosters (for pros) or scholarship limits (colleges) are part of the equation.

Back-office uses include analytics to improve a team's operation as well as for health and safety of the players. Team analytics include strategies and tactics, competitive assessments, and optimal lineup choices under various on-field or on-court situations. Health/safety analytics focus on medical, strength and fitness as well as development, and predictive models for avoiding overtraining and injuries. Concussion research is a hot field, for example.

Finally, the bottom layer points out the potential for application of analytics at the league/conference level to optimize schedules and locations of games across a pool of teams, including tournament seeding.

The following representative examples illustrate how various sports organizations use data and analytics to improve sports operations, in the same way analytics have improved

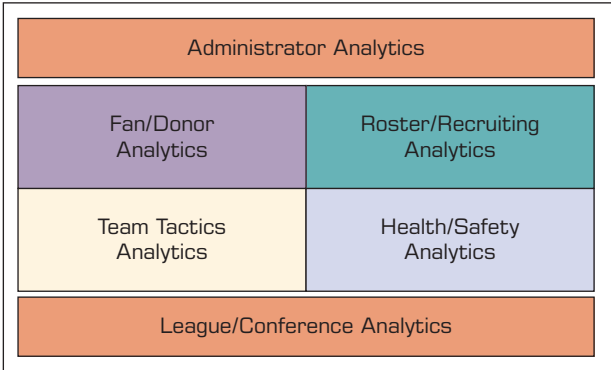


FIGURE 1.1 Six Areas of Sports Analytics.

traditional industry decision making. Almost all of these are based on actual projects conducted by students and researchers. In some cases, the names have been changed to protect identity of the stakeholders.

Example 1: The Business Office—Fan Analytics

Katie Ward works as a business analyst for a major pro baseball team, focusing on revenue. She analyzes ticket sales, both from season ticket holders as well as single-ticket buyers. Sample questions in her area of responsibility include why season ticket holders renew (or do not renew) their tickets, as well as what factors drive last-minute individual seat ticket purchases. Another question is how to price the tickets.

Some of the analytical techniques Katie uses include simple statistics on fan behavior like overall attendance and answers to survey questions about likelihood to purchase again. However, what fans say versus what they do can be different. Katie runs a survey of fans by ticket seat location (“tier”) and asks about their likelihood of renewing their season tickets. But when she compares what they say versus what they do, she discovers big differences (see Figure 1.2). She found that 69% of fans in Tier 1 seats who said on the survey that they would “probably not” renew actually did. This is an useful insight that leads to action—customers in the green cells are the most likely to renew tickets, so require fewer marketing touches and dollars to convert, for example, compared to customers in the blue cells who are “on the edge.”

However, many factors influence fan ticket purchase behavior, especially price, which drives more sophisticated statistics and data analysis. For both areas, but especially single-game tickets, Katie is driving the use of dynamic pricing—moving the business from simple static pricing by seat location tier to day-by-day up-and-down pricing of individual seats. This is a rich research area for many sports teams and has huge upside potential for revenue enhancement. For example, her pricing takes into account the team’s record, who they are playing, game dates and times, which star athletes play for each team, each fan’s history of renewing season tickets or buying single tickets, as well as factors like seat location, number of seats, and real-time information like traffic congestion historically at game time and even the weather. See Figure 1.3.

Which of these factors are important? How much? Given her extensive statistics background, Katie builds regression models to pick out key factors driving these historic behaviors and create predictive models to identify how to spend marketing resources to drive revenues. She builds churn models for season ticket holders to create segments of customers who will renew, won’t renew, or are fence-sitters, which then drives more refined marketing campaigns.

In addition, she does sentiment scoring on fan comments like tweets that help her segment fans into different loyalty segments. Other studies about single-game attendance drivers help the marketing department understand the impact of giveaways like bobble-heads or T-shirts, or suggestions on where to make spot TV ad buys.

Tier	Highly Likely	Likely	Maybe	Probably Not	Certainly Not
1	92	88	75	69	45
2	88	81	70	65	38
3	80	76	68	55	36
4	77	72	65	45	25
5	75	70	60	35	25

FIGURE 1.2 Season Ticket Renewals—Survey Scores.

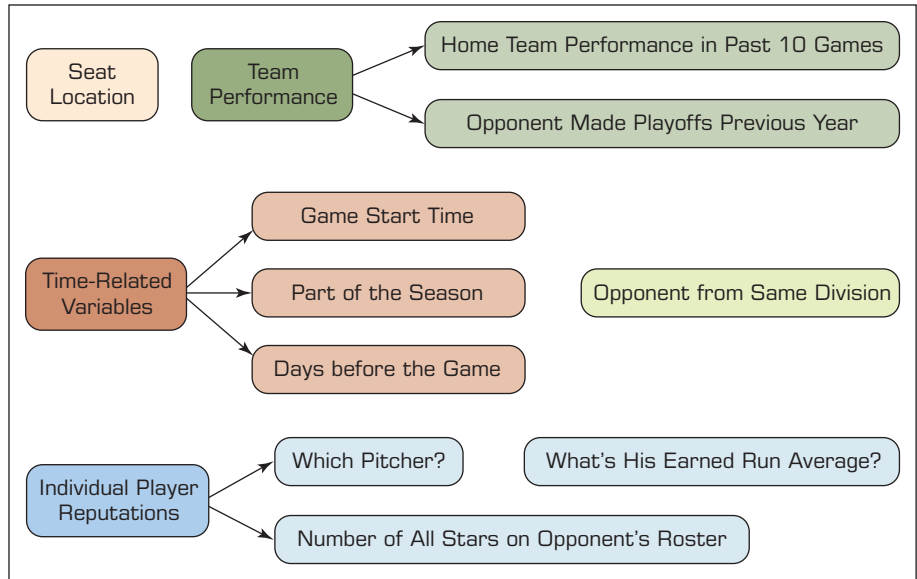


FIGURE 1.3 Dynamic Pricing Previous Work—Major League Baseball. Source: Adapted from Kemper, C., & Breuer, C. (2016). How efficient is dynamic pricing for sports events? Designing a dynamic pricing model for Bayern Munich. *International Journal of Sports Finance*, 11, 4–25.

Beyond revenues, there are many other analytical areas that Katie’s team works on, including merchandising, TV and radio broadcast revenues, inputs to the general manager on salary negotiations, draft analytics especially given salary caps, promotion effectiveness including advertising channels, and brand awareness, as well as partner analytics. She’s a very busy person!

Example 2: The College Football Coach—Play Tactics

Bob Breedlove is the football coach for a major college team. For him, it’s all about winning games. His areas of focus include recruiting the best high school players, developing them to fit his offense and defense systems, and getting maximum effort from them on game days. Sample questions in his area of responsibility include: Who do we recruit? What drills help develop their skills? How hard do I push our athletes? Where are opponents strong or weak, and how do we figure out their play tendencies?

Fortunately, his team has hired a new team operation expert, Dar Beranek, who specializes in helping the coaches make tactical decisions. She is working with a team of student interns who are creating opponent analytics. They used the coach’s annotated game film to build a cascaded decision tree model (Figure 1.4) to predict whether the next play will be a running play or passing play. This shows some tendencies they might want to exploit. For example (follow the right-side sequence), when they see a personnel formation that looks like a pass, and it’s third or fourth down with more than 5 yards to go, their opponent team passes 95.45% of the time—very predictable!

For the defensive coordinator, they have built heat maps (Figure 1.5) of each opponent’s passing offense, illustrating their tendencies to throw left or right and into which defensive coverage zones. Finally, they built some time series analytics (Figure 1.6) on explosive plays (defined as a gain of more than 16 yards for a passing play or more than 12 yards for a run play). For each play, they compare the outcome with their own defensive formations and the other team’s offensive formations,

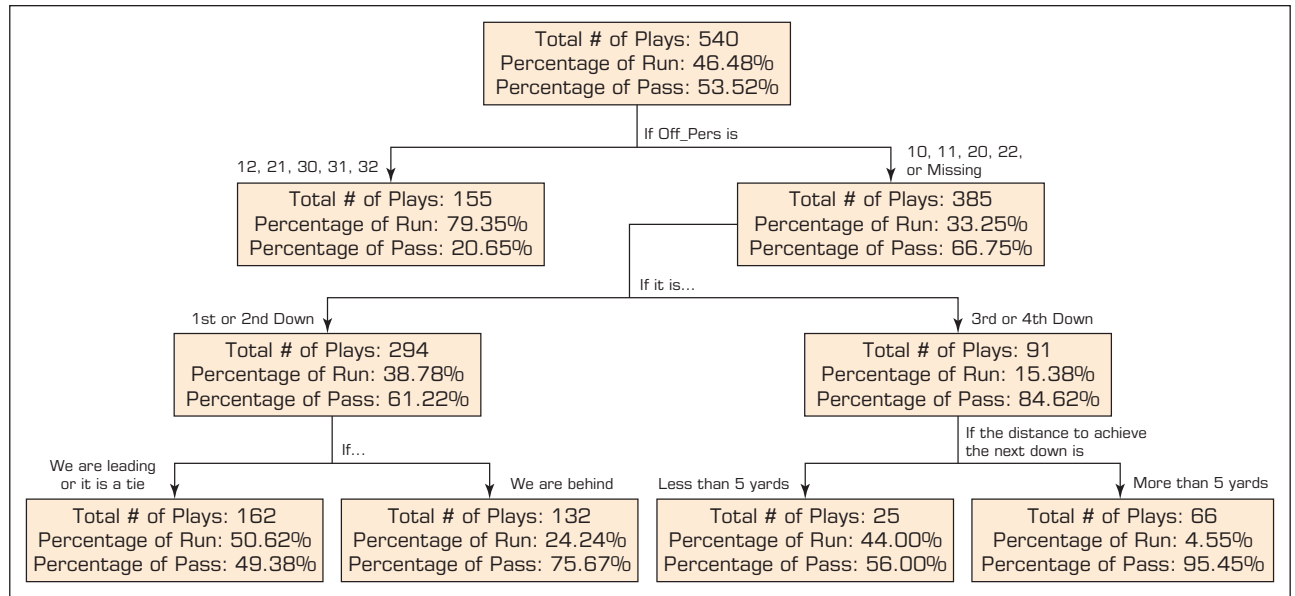


FIGURE 1.4 Cascaded Decision Tree for Run or Pass Plays.

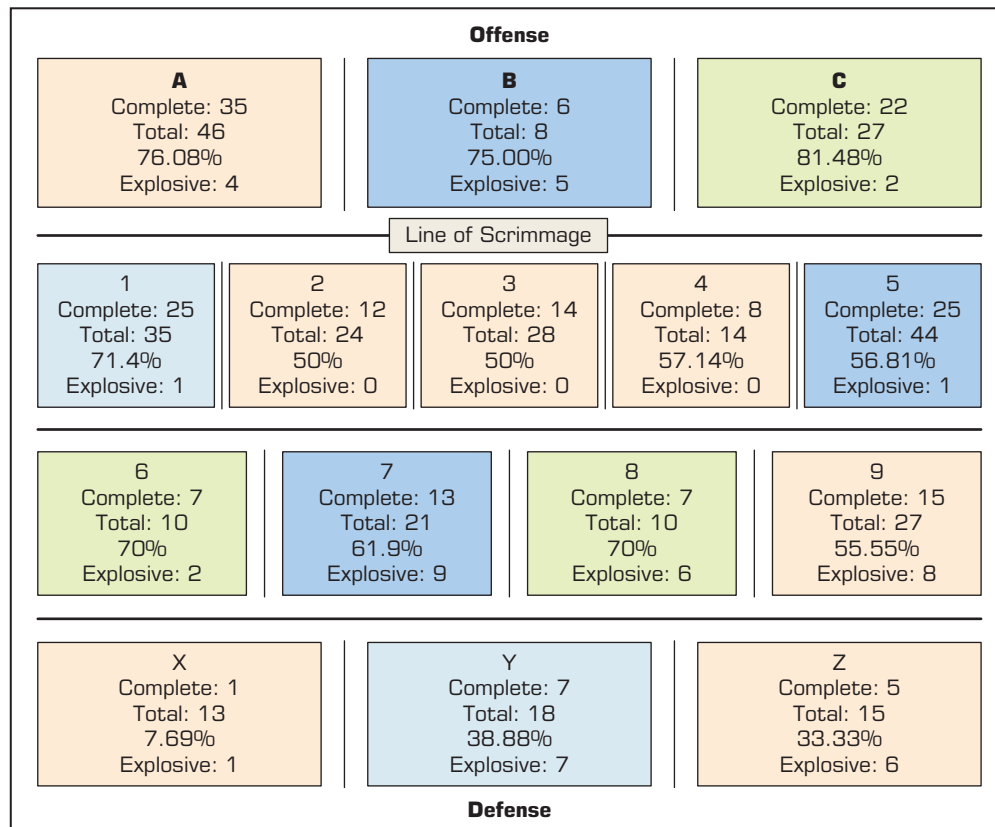


FIGURE 1.5 Heat Map Zone Analysis for Passing Plays.

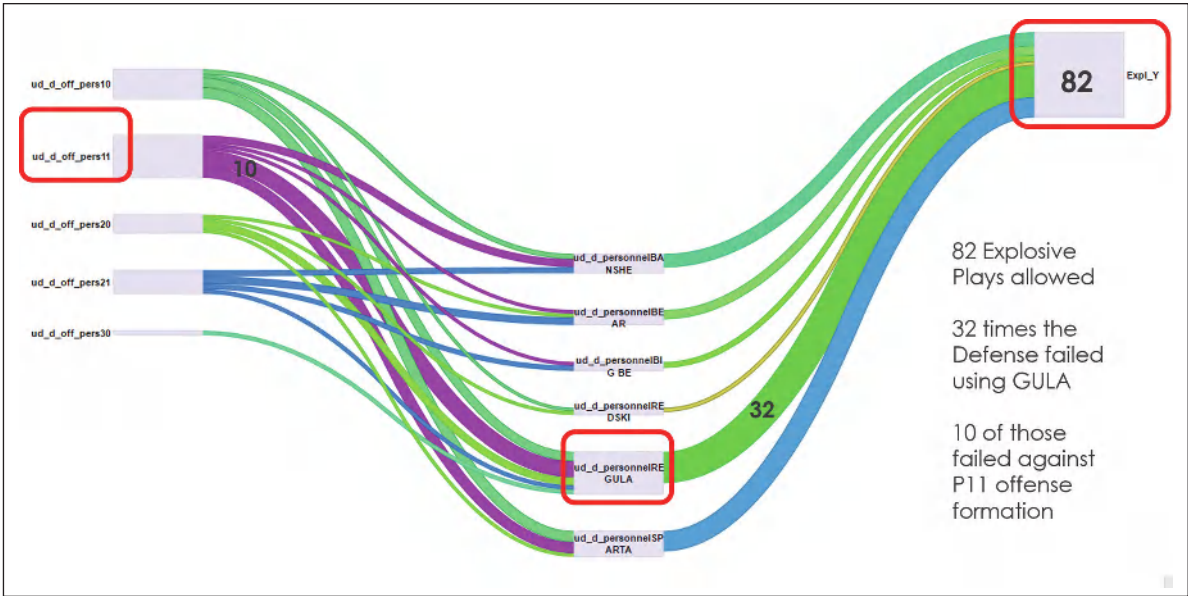


FIGURE 1.6 Sankey Diagram of Explosive Plays.

which helps Coach Breedlove react more quickly to formation shifts during a game. New work that Dar is fostering involves building better high school athlete recruiting models. For example, each year the team gives scholarships to three students who are wide receiver recruits. For Dar, picking out the best players goes beyond simple measures like how fast athletes run, how high they jump, or how long their arms are to newer criteria like how quickly they can rotate their heads to catch a pass, what kinds of reaction times they exhibit to multiple stimuli, and how accurately they run pass routes. Some of her ideas illustrating these concepts appear on the Teradata University for Academics Web site; look for the BSI Case of Precision Football. Figure 1.6 illustrates a special type of chart called a Sankey chart or Sankey diagram. A Sankey chart can be used to describe flow from one stage to another. The nodes at each stage may be connected to the nodes at the next stage. The thickness of the arrows connected from one stage to the next typically describe the proportion of the time that path is noted in the data. Thus, Figure 1.6 shows that (right hand side) the Defense gave up 82 explosive plays. Backtracking to the left, the coach can see that the defensive formation GULA gave up 32 of those explosive plays. Backtracking further to the first column, GULA failed against Offensive Personnel formation coded “11” ten of the 32 times. This information can lead to better practice planning and in-game formation matchup decisions.

Example 3: The High School Coach—Beat the Arch-Enemy

Coach Joe is focused on each of his upcoming opponents, but especially his arch-rival who regularly wins the state tournament and trounces all the teams in the conference. His offense and defense coordinators need the competitive edge that comes from knowing their play-calling tendencies. He gets connected with two team tactics experts, Spencer and Archis, who provide the coaches with deep insights, using six years of detailed play data on this opponent.

Archis starts by creating some statistics (Figure 1.7) about this opponent’s historical choice of formations and plays are problematic for Joe’s defense.

TOP Offensive Formations			TOP Offensive Plays		
FORMATION	AVG YARDS	% USED	PLAY	AVG YARDS	% USED
THUNDER	6.62	21%	24 ISO	6.09	6.45%
4 WIDE	12.44	18%	25 ISO	9.06	5.28%
LIGHTNING	9.71	12%	93	18.75	4.69%
WISHBONE	4.00	8%	JET 18	4.77	3.81%
SINGLE WING	9.17	5%	FB Dive	3.54	3.81%
OPEN LEFT	5.08	3%	14 SINGLE OMAHA	3.23	3.81%
ACE	8.90	3%	JET 19	5.50	3.52%
TROJAN RIGHT	7.33	3%	15 SINGLE OMAHA	7.73	3.23%
TRIPS RIGHT	-0.13	2%	JET 93	11.25	2.35%
TWINS SPLIT	0.29	2%	50 ALL SEAM	10.00	2.05%
			91	7.43	2.05%
			62 FLORIDA	9.57	2.05%
			JET 26 QB POWER READ	19.14	2.05%

FIGURE 1.7 Opponent Statistics about Formations and Plays that Give Up Explosive Yardage.

Using the detailed play by play data, they identify both offensive and defensive tendencies that they convert into static game play sheets. Figure 1.8 shows a sheet that the defensive coordinator can use when the Arch-Enemy has the ball. Which play is likely coming next depends on the field position, down and distance.

Spencer decided to build a new real-time app for the defensive coordinators in the press box. They have 5–8 seconds to communicate what defensive set to use. In this tool, checking and advancing the situational context (on the right side), the previous plays' resulting

DANGER_ZONE				
DOWN & DISTANCE	Pass%	Run%	Pass Avg Yards	Run Avg Yards
1st and Short	0%	0%	0.00	0.00
1st and Mid	0%	0%	0.00	0.00
1st and Long	5%	95%	0.00	3.78
2nd and Short	0	100%	0.00	23.33
2nd and Mid	0	100%	0.00	1.00
2nd and Long	25%	67%	7.00	2.38
3rd and Short	0%	0%	0.00	0.00
3rd and Mid	0	100%	0.00	5.00
3rd and Long	57%	43%	30.75	32.67

OWN_MID				
DOWN & DISTANCE	Pass%	Run%	Pass Avg Yards	Run Avg Yards
1st and Short	0%	0%	0.00	0.00
1st and Mid	0%	100%	0.00	11.00
1st and Long	16%	84%	26.00	4.77
2nd and Short	33%	67%	0.00	6.50
2nd and Mid	22%	78%	11.00	6.43
2nd and Long	39%	61%	13.00	9.18
3rd and Short	0%	100%	0.00	2.75
3rd and Mid	0%	100%	0.00	1.00
3rd and Long	67%	33%	4.00	-5.00

MID				
DOWN & DISTANCE	Pass%	Run%	Pass Avg Yards	Run Avg Yards
1st and Short	0%	0%	0.00	0.00
1st and Mid	0%	0%	0.00	0.00
1st and Long	40%	58%	14.19	11.87
2nd and Short	0%	100%	0.00	7.00
2nd and Mid	13%	88%	0.00	3.57
2nd and Long	38%	54%	2.80	6.43
3rd and Short	0%	100%	0.00	5.00
3rd and Mid	33%	67%	10.00	8.00
3rd and Long	75%	25%	14.33	6.00

OPP_MID				
DOWN & DISTANCE	Pass%	Run%	Pass Avg Yards	Run Avg Yards
1st and Short	0%	0%	0.00	0.00
1st and Mid	0%	100%	0.00	8.50
1st and Long	33%	65%	4.54	9.27
2nd and Short	33%	67%	16.50	2.25
2nd and Mid	0%	100%	0.00	10.50
2nd and Long	45%	55%	14.00	6.73
3rd and Short	33%	67%	7.00	3.00
3rd and Mid	0%	100%	0.00	4.00
3rd and Long	83%	17%	13.00	-4.00

RED_ZONE				
DOWN & DISTANCE	Pass%	Run%	Pass Avg Yards	Run Avg Yards
1st and Short	0%	100%	0.00	2.83
1st and Mid	20%	80%	0.00	2.00
1st and Long	33%	57%	3.57	5.17
2nd and Short	0%	100%	0.00	2.83
2nd and Mid	25%	75%	6.00	2.33
2nd and Long	38%	63%	6.67	-1.40
3rd and Short	20%	80%	15.00	5.50
3rd and Mid	100%	0%	5.00	0.00
3rd and Long	60%	40%	8.33	5.50

FIGURE 1.8 Coach Play Sheet Showing Opponent's Offensive Tendencies by Field Position, Down and Distance.

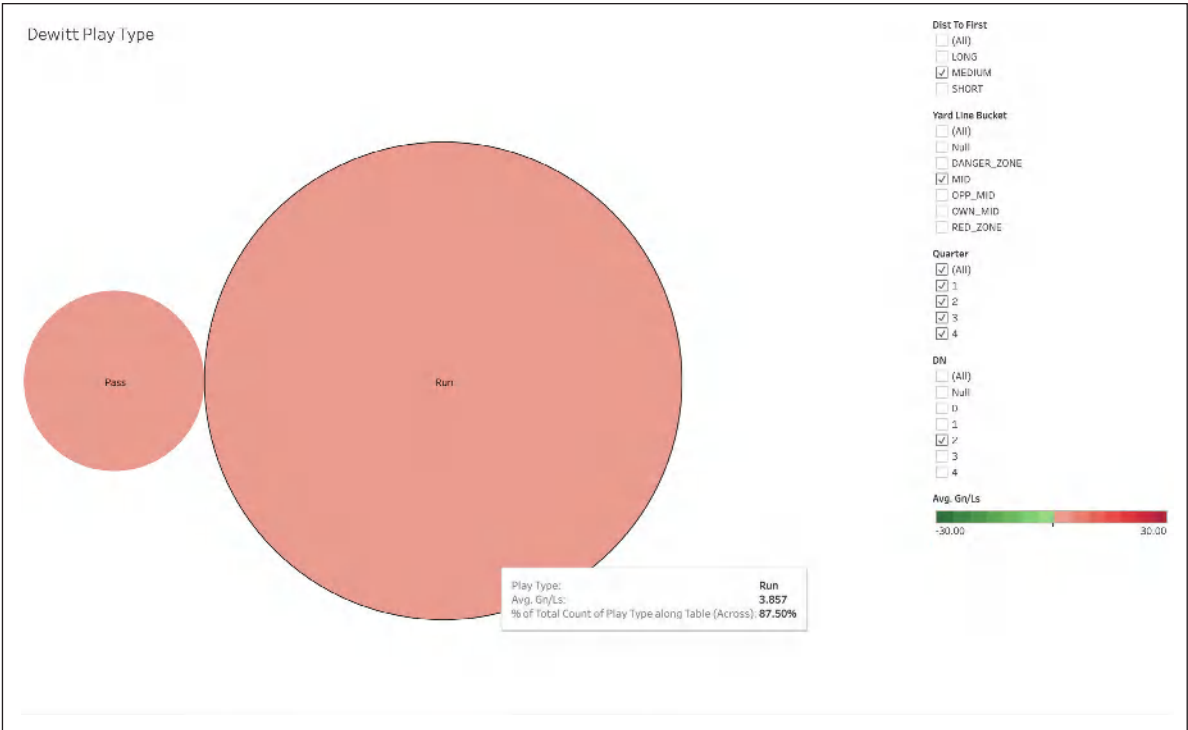


FIGURE 1.9 Defense Coordinator Visual Tool—What Play Is Coming Next?

new yard line, down and distance instantly brings up a visual that shows the percent of time the next play will be a run or pass (Figure 1.9). In this case, it’s probably a run.

With one more click on the Run Bubble, the defensive coordinator can “drill into” the details to see what kind of run it’s likely to be and which will yield the most yardage (Jet 18 in light blue gives up the most yards). The defensive coordinator can then call in the right defensive set to the sideline coach to thwart that play (Figure 1.10).

Finally, Joe’s analytics experts decide to use the four predictive play models (offense and defense play and outcome predictions for both teams) to build a Monte Carlo simulator and “play” a complete game multiple times under various tactical changes to see what the score outcomes would be. We introduce Monte Carlo simulation in Chapter 8.

In particular, for each play of the game, the offense chooses to run a randomly selected play from its offensive model based on the percent of times they have run it in past games, and gains (or loses) a certain number of yards based on how many yards that play gains on average against a particular defense, taken from the defense model statistics for the opponent. The simulator iterates through the 80 or so plays that each team plays during a game, alternating possessions as usual as in a real game. It considers time, kickoffs, punts, turnovers, yard lines, down and distances as base variables to get a score distribution for each of 5000 simulations, which they use to create the “baseline” model. Figure 1.11 shows that against their Arch-Enemy (scores in blue), Coach Joe’s team (distribution of scores in orange) on average will lose the game 30–10.

Next, the analysts took a look at adding variations on tactics that might reduce the spread, like increasing or decreasing Joe’s defense’s blitzing percent to surprise the offense. Different tactics for the offense included going for it unexpectedly on 4th downs, or doing onside kicks, or increasing time per play to eat up the clock and keep the ball out of the Arch-Enemy’s hands (don’t have the ball—can’t score on us!). Each of these can be looked at independently or in combination by adjusting the offense or defense

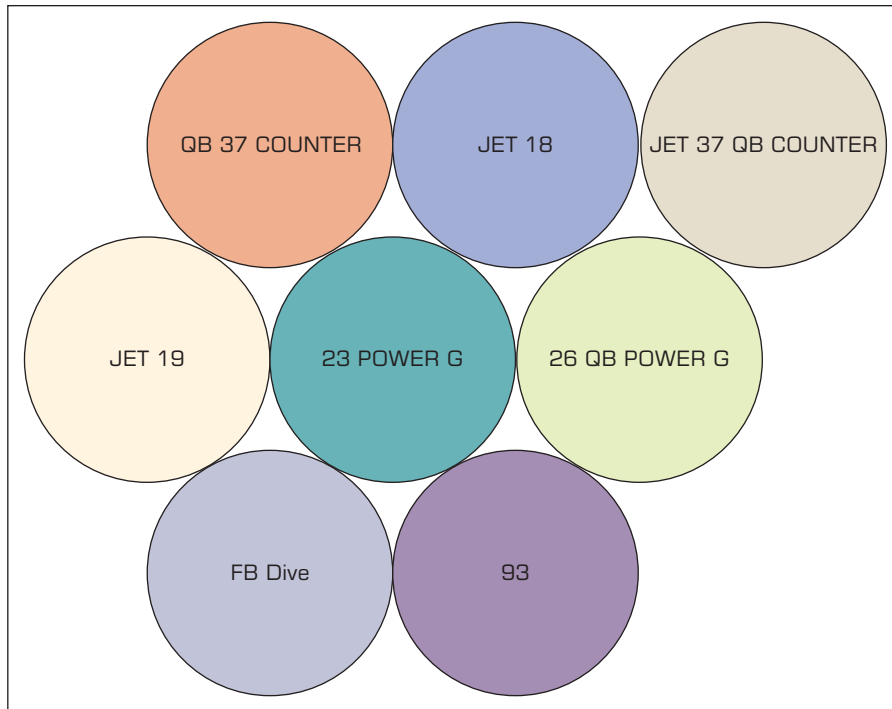


FIGURE 1.10 Interactive Tool Drilldown: Which Run Play Is Likely Coming and How Many Yards.

models for Joe's team, so the analysts can recommend the best play tactic changes to increase Joe's win likelihood. With four changes of tactics, the model still unfortunately predicts a loss but the score differential is reduced to 24–10. See Figure 1.12. We'll see when they play this fall how much analytics helped!

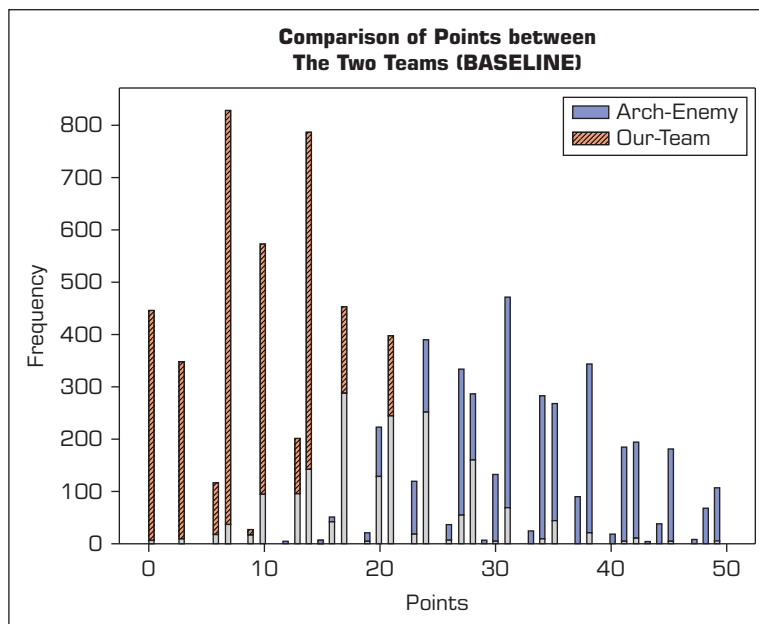


FIGURE 1.11 Monte Carlo Simulator Baseline Score Distributions—5000 Simulated Games.

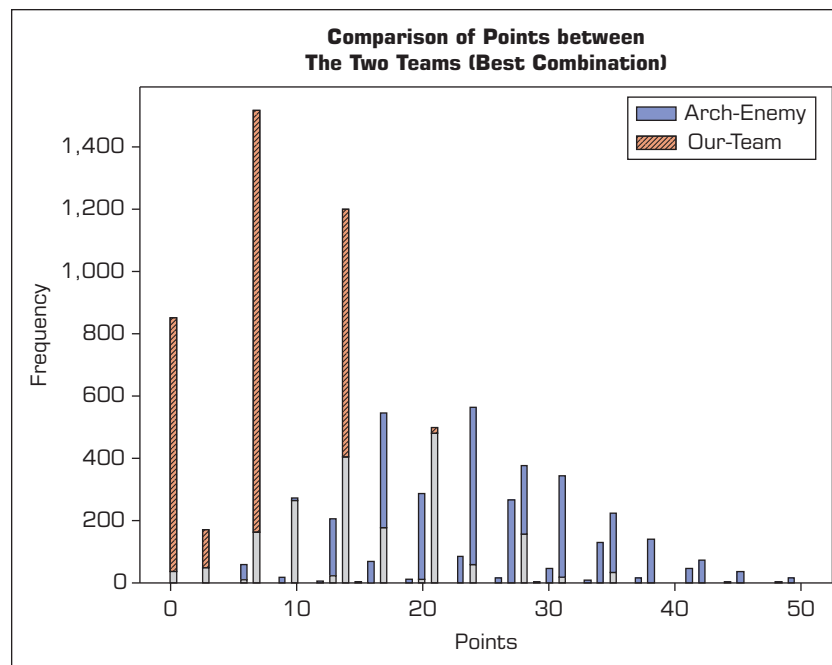


FIGURE 1.12 Monte Carlo Simulator Best Four Tactics Score Distributions—5000 Simulated Games.

Example 4: The Cricket Coach—Roster and Play Decisions

Krishna Kulkarni is a new cricket coach for the Bombay Desis, a team that's been struggling in the Cricket Supreme League. Fan attendance has dropped and he's under pressure to make some roster changes. He has hired some data analytics experts to get a better idea where his bowlers (pitchers) have weaknesses and strengths against the batters from other teams, notably his arch-rivals, the Delhi Drivers and the Bangalore Bees. He knows cricket is behind baseball in the use of sophisticated analytics. In particular, cricket is still just using box scores while other sports have detailed event data to analyze.

The data analysts proposed a project to scrape publicly available blog data by commentators about each pitch and hit, then use that to build deeper insights about bowler-batter matchup events. The first part of the project used text analytics to try to pull interesting extra data fields out of the free-form blog text.

Unlike baseball, cricket pitches can be bounced at various distances to the batter, called the Length. Figure 1.13 illustrates how a ball is pitched from a bowler to the batter's stumps. It ranges from "Short Length" to "Full Toss." There are also a variety of delivery types (such as fast vs. slow balls, curve balls, sinkers). Two more dimensions that can be pulled out of the text information involve where the ball is aimed, and how the batter swings, providing four new variables for analysis. All of this information is extracted from the recorded blog commentary of each play. It leads to additional fields of data about each play event, besides noting the game info (date, team, location, etc.) and of course, the names of bowlers and batsmen, the number of runs scored on that play, etc. Since the data is extracted from text mining analysis of the commentary blog, it is potentially incomplete. However, it is still better than what the analysts had before the project.

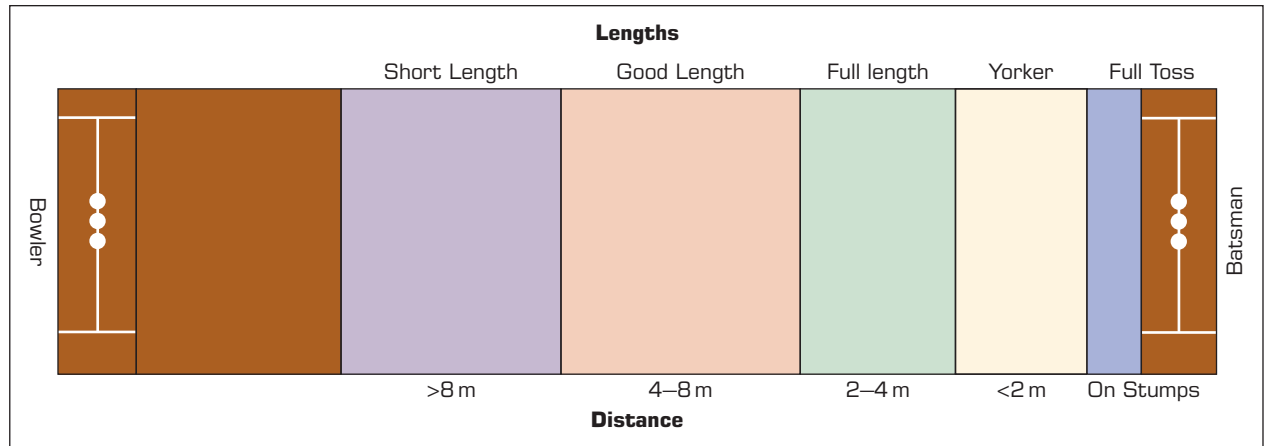


FIGURE 1.13 Delivery Lengths in a Cricket Bowling Play.

The text information was extracted to add these variables:

- **Delivery Length:** Short Length, Good Length, Fuller Length, Yorker Length, Full Toss
- **Delivery Type:** Googly, Arm Ball, Cutter, LegBreak, OffBreak, TopSpinner, Carrom Ball, Bouncer
- **Ball Placement:** On Stump, Away from Stump
- **Batsman Reaction:** Steps Out, Defense, On Toes

With the above dataset in place, the first analysis focused on identifying vulnerability of a batter against different types of bowling (length of delivery). Figure 1.14 shows this analysis. Note the numbers in each row do not add up to 100% because the data on delivery lengths may not be available for each play. This information assists a coach with both offense and defense. The data can be used by coaches to help the batsmen improve their skills at specific

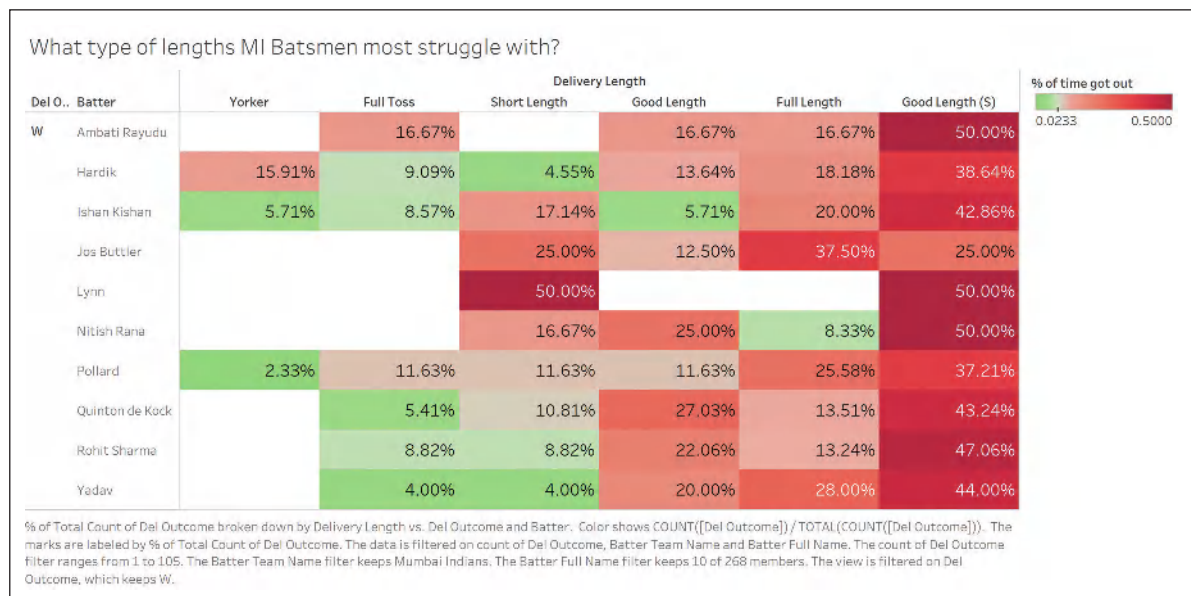


FIGURE 1.14 Different Players' Performance on Plays at Different Delivery Lengths.

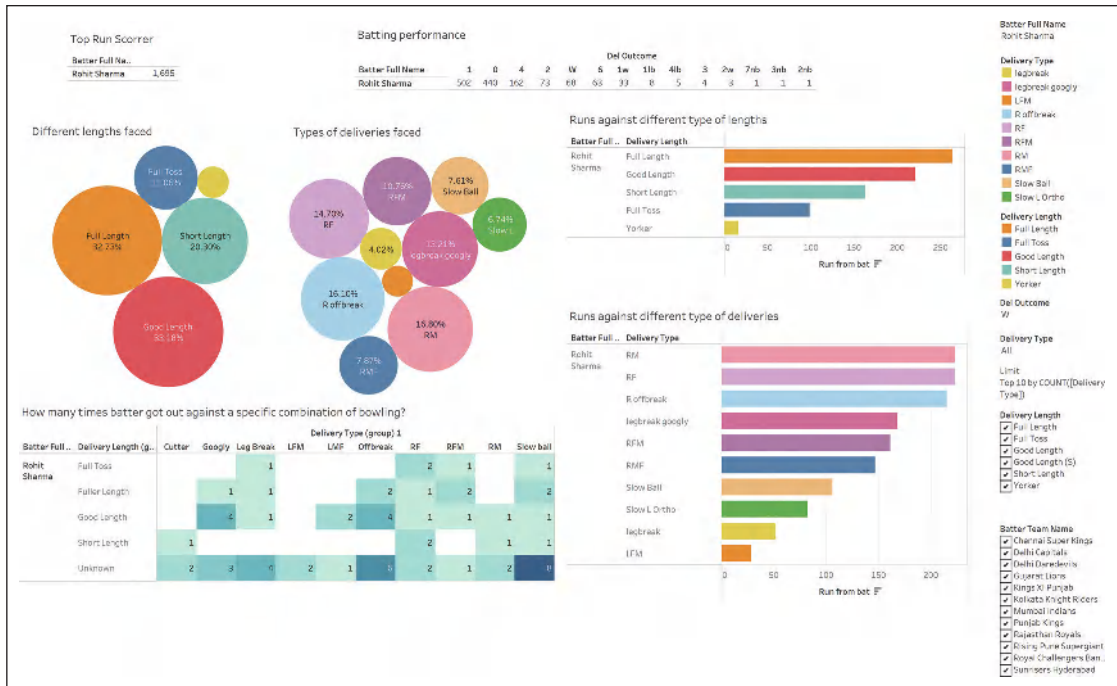


FIGURE 1.15 An Example of the Analytical Summary Dashboard of a Player's Performance.

pitch length weaknesses through better training, or help a bowler figure out a batter's weakness at some delivery lengths to be able to take advantage of this vulnerability in a game.

The balls are also delivered by bowlers using many different styles of bowling. Examples of these labels include bouncer, googly, legbreak, offbreak, etc. An online search will let you understand these styles if interested. See for example: <https://cricketmastery.com/types-of-bowling-in-cricket/>. The analysts developed a similar chart of which batter was more likely to fall victim to a particular type of ball delivery.

The data analytics team also put together a detailed analysis of a player's performance against various types of conditions. Figure 1.15 illustrates this analysis for a player who is very well known in the cricket world—Rohit Sharma. This chart displays the total runs scored by the player when playing for this team, his performance in terms of types of runs scored, or when he was out under different plays. It also includes two charts displaying different types of delivery lengths and delivery styles faced by him, runs against these different types of delivery lengths and delivery styles, and a cross-sectional analysis of his performance in different delivery length/style combinations. For example, the chart on the bottom right corner of this dashboard shows that Rohit does fairly well in facing Right balls—right fastbreak, fast, etc. However, he scores poorly when facing a legbreak style delivery. The chart in the bottom left corner shows that goodlength is the type of delivery where he struggled the most. Within goodlength deliveries, his biggest struggles were on balls delivered as googly, offbreak, or slowball.

The Bombay Desis scouts can work with the analysts to create these kinds of scouting reports for each prospect they are considering to help the coach recruit players to build a comprehensive team. Alternatively, such an analysis of the opposing team's batsmen can help plan to use bowlers who will most likely thwart hits and runs by a batter. A similar analysis can then be performed between each batter and a bowler to identify the best team combination (batsmen and bowlers) for a particular match.

Cricket is a multibillion-dollar industry around the world. As can be seen from this example, coaches and managers can use analytics to identify the best future players for their teams as well as plan their bowling and batting strategies to play against specific teams.

Example 5: The Commissioner—League Tournament Design

At the NCAA, a committee for Division III Wrestling sets the assignment of schools to regional tournaments to qualify for the national tournament. The assignment is done every three years, manually. The president of the Coaches Association asked a team of analysts to “do math” to automate the process (and avoid politics), which turned into a very interesting analytics project.

The research question is simply stated: assign N (e.g., 103) teams to six regions such that:

- Roughly the same number of teams are in each region ($103/6 = 17$ approximately),
- The average “power” of each region (based on summing the assigned teams’ power scores) is roughly equal, and
- The distance from each school to the center of each region is minimized, since these colleges do not have big travel budgets (most will drive).

Because there are many optimization techniques, we ran a friendly competition with five schools to come up with the best algorithm. A faculty member at the University of Cincinnati had been a collegiate wrestler so worked on formulating a predictive Power Model based on 24 factors from the past two years of match results for all competitions for all schools. Out of these factors, many were correlated and only four were predictive, and when back-tested had 70% accuracy. The four other faculty at Bentley, Notre Dame, Purdue, and Wright State tried variations on k -d trees, set partitioning, genetic algorithms, and a two-stage clustering algorithm to come up with the “best” approach. While the problem is simply stated, the algorithms all have strengths and weaknesses; details appear in (Carter et al., 2022). The 106 Division III teams are mostly in the north-east quadrant of the United States, with three exceptions which were taken out of the data set since they are outliers and would skew results. (Once the team made the optimal regional assignments, these teams could choose a region that minimized their expenses given flight costs).

A number of different algorithms were used to build the ideal schedule. The winner was genetic algorithm. The approach is to compute regional assignments like strands of DNA. Each strand is represented by a vector of 103 entries and each entry has a value from 1 to 6. The algorithm mates pairs of vectors and occasionally imposes mutations, making one change at a time by swapping two schools’ assignments. Each resulting “offspring” is then evaluated on the criteria—is it better or worse? The analysts equally weighted the three criteria (each region having an equal number of teams, each region’s power being roughly equal, and each region minimizing travel distance to the center of the region). The breeding experiment begins with 10 sets of different starting sets called partitions.

Over time the “gene pool” of partitions improves as they advantageously inherit beneficial features from their parents or fortunately get them from mutation, and as poor solutions are removed from the gene pool through natural selection. This process does not always reach a global optimum. Indeed, genetic algorithms are typically applied in situations where no algorithm for achieving a global optimum is known. But the random mutations create opportunities for exiting local optima by introducing a solution that is not necessarily “near” any of the other solutions in the pool.

In this analysis, 10 candidate solutions were used for each round of breeding, and the process continued for 10,000 generations. By that time, evolution stopped yielding substantive changes in the objective function. (Rapid improvement took place in the first 500 generations, slowing until about 4,000 generations, after which it plateaued.) The authors choose the best-scoring partition from the gene pool when the algorithm terminates. Figure 1.16 displays one sample output.

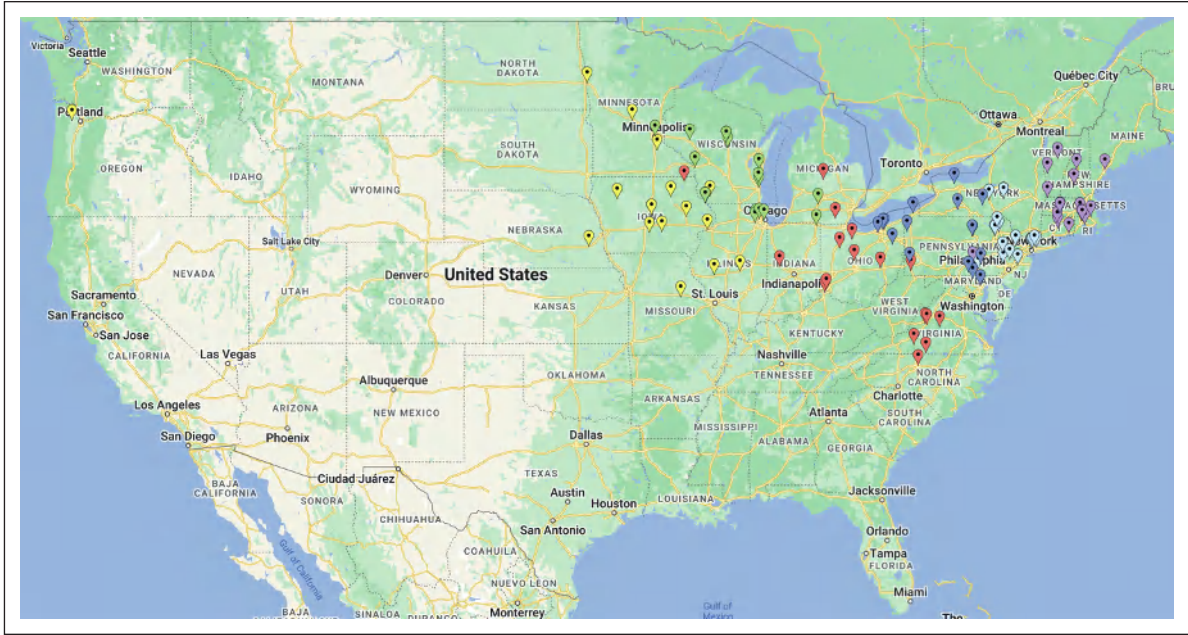


FIGURE 1.16 Genetic Algorithm School Assignments to Six Regions (color-coded).

The good news is that the algorithm runs in 15 minutes with some Excel plug-ins and the results can then be mapped using Tableau. The Commissioner was impressed and the tournament committee is now eager to try it out!

► QUESTIONS ABOUT THESE EXAMPLES

1. What are three factors that might be part of a predictive model for season ticket renewals?
2. What are three factors that football teams can use to do opponent analysis?
3. What other cricket player analytics can you imagine to assist in making better coaching decisions?
4. What other analytics uses can you envision for your favorite sport?

WHAT CAN WE LEARN FROM THESE VIGNETTES?

Beyond the front-office business analysts, the coaches, trainers, and performance experts, there are many other people in sports who use data, ranging from golf groundskeepers who measure soil and turf conditions for PGA tournaments, to baseball and basketball referees who are rated on the correct and incorrect calls they make. In fact, it's hard to find an area of sports that is *not* being impacted by the availability of more data, especially from sensors.

Skills you will learn in this book for business analytics will apply to sports. If you want to dig deeper into this area, we encourage you to look at the Sports Analytics section of the Teradata University for Academics Website, <https://learning-academics.teradata.com/>, a free resource for students and faculty. On this Web site, you will find descriptions of what to read, compilations of places where you can find publicly available data sets for analysis, as well as examples of student projects in sports analytics and interviews of sports professionals who use data and analytics to do their jobs. Good luck learning analytics!

Source and Credits: Sports examples were contributed by Dr. Dave Schrader, who retired after 24 years in advanced development and marketing at Teradata. He has remained on the

Board of Advisors of the Teradata University for Academics, where he spends his retirement helping students and faculty learn more about sports analytics by doing projects.

The football visuals (Figures 1.4–1.6) were constructed by Peter Liang and Jacob Pearson, graduate students at Oklahoma State University, as part of a student project in the spring of 2016. The second football analytics story is based on a spring 2021 project with students from Auburn, LSU, and Purdue. The visuals (Figures 1.7–1.12) are from Purdue undergraduate data science students Spencer Prentiss and Archis Dhar. Special thanks to coaches Mike Durnin, Joe Brya, and Scott Farley for giving students access to live data and providing feedback on what analytics are useful, as well as weekly mentoring tips from experts Hayden Ellis and Emmett Clifford. Cricket analytics visuals (Figures 1.13–1.15) are from a project conducted by Rupom Bhattacharjee, Zach Miller, Raunak Sengupta, and Sushma Reddy, students in the MS-BAnDS program at Oklahoma State University. The cricket project was completed under the supervision of Dr. Miriam McGaugh, Fawad Qureshi of Teradata-London, and Dr. Dave Schrader. The NCAA regional tournament project was a multi-school collaborative effort. Thanks to coach Jon McGovern of the University of Dubuque for suggesting the project, JP Williams of the NCAA for supporting it, and several faculty members. Kudos to Prof. Andrew Harrison at the University of Cincinnati who built the Power Model. Professors Nathan Carter at Bentley University, Scott Nestler at Notre Dame, Matt Lanham at Purdue, and Amir Hassan Zadeh at Wright State University each provided a unique analytic approach for solving the tournament problem. Figure 1.16 was provided by Prof. Nathan Carter.

1.2 CHANGING BUSINESS ENVIRONMENTS AND EVOLVING NEEDS FOR DECISION SUPPORT AND ANALYTICS

The opening vignette illustrates how an entire industry can employ analytics to develop reports on what is happening, predict what is likely to happen, and then also make decisions to make the best use of the situation at hand. These steps require an organization to collect and analyze vast stores of data. From traditional uses in payroll and bookkeeping functions, computerized systems have now penetrated complex managerial areas ranging from the design and management of automated factories to the application of analytical methods for the evaluation of proposed mergers and acquisitions. Nearly all executives know that information technology is vital to their business and extensively use information technologies.

Decision making is one of the most important activities in organizations of all kind—probably the most important one. Decision making leads to the success or failure of organizations and how well they perform. Making decisions is getting difficult due to internal and external factors. The rewards of making appropriate decisions can be very high and so can the loss of inappropriate ones.

Unfortunately, it is not simple to make decisions. To begin with, there are several types of decisions, each of which requires a different decision-making approach. For example, De Smet et al. (2017) of McKinsey & Company management consultants classify organizational decision into the following four groups:

- Big-bet, high-risk decisions
- Cross-cutting decisions, which are repetitive but high risk that require group work (Chapter 11)
- Ad hoc decisions that arise episodically
- Delegated decisions to individuals or small groups

Therefore, it is necessary first to understand the nature of decision making. For a comprehensive discussion, see De Smet et al. (2017).

Modern business is full of uncertainties and rapid changes. To deal with these, organizational decision makers need to deal with ever-increasing and changing data. This book is about the technologies that can assist decision makers in their jobs.

Decision-Making Process

For years, managers considered decision making purely an art—a talent acquired over a long period through experience (i.e., learning by trial and error) and by using intuition. Management was considered an art because a variety of individual styles could be used in approaching and successfully solving the same types of managerial problems. These styles were often based on creativity, judgment, intuition, and experience rather than on systematic quantitative methods grounded in a scientific approach. However, recent research suggests that companies with top managers who are more focused on persistent work tend to outperform those with leaders whose main strengths are interpersonal communication skills. It is more important to emphasize methodical, thoughtful, analytical decision making rather than flashiness and interpersonal communication skills.

We will see several times in this book how an entire industry can employ analytics to develop reports on what is happening, predict what is likely to happen, and then make decisions to make the best use of the situation at hand. These steps require an organization to collect and analyze vast stores of data. In general, the amount of data doubles every two years. From traditional uses in payroll and bookkeeping functions, computerized systems are now used for complex managerial areas ranging from the design and management of automated factories to the application of analytical methods for the evaluation of proposed mergers and acquisitions. Nearly all executives know that information technology is vital to their business and extensively use these technologies.

Computer applications have moved from transaction processing and monitoring activities to problem analysis and solution applications, and much of the activity is done with cloud-based technologies, in many cases accessed through mobile devices. Analytics and BI tools such as data warehousing, data mining, online analytical processing (OLAP), dashboards, and the use of the cloud-based systems for decision support are the cornerstones of today's modern management. Managers must have high-speed, networked information systems (wired or wireless) to assist them with their most important task: making decisions. In many cases, such decisions are routinely being automated, eliminating the need for any managerial intervention.

Technologies for Data Analysis and Decision Support

Besides the obvious growth in hardware, software, and network capacities, some developments have clearly contributed to facilitating growth of decision support and analytics in a number of ways, including the following:

- **Group communication and collaboration.** Many decisions are made today by groups whose members may be in different locations. Groups can collaborate and communicate readily by using collaboration tools as well as the ubiquitous smartphones. COVID-19 pandemic resulted in millions of workers having to work remotely and has ushered in a new era of collaboration through technologies. Even under normal times, assembling a group of decision makers, especially experts, in one place can be costly. Information systems can improve the collaboration process of a group and enable its members to be at different locations (saving travel costs). Collaboration is especially important along the supply chain, where partners—all the way from vendors to customers—must share information. More critically, such supply chain collaboration permits manufacturers to know about the changing patterns of demand in near real time and thus react to marketplace changes faster.

- **Improved data management.** Many decisions involve complex computations. Data for these can be stored in different databases anywhere in the organization and even possibly outside the organization. The data may include text, sound, graphics, and video, and these can be in different languages. Many times, it is necessary to transmit data quickly from distant locations. Systems today can search, store, and transmit needed data quickly, economically, securely, and transparently.
- **Managing giant data warehouses and Big Data.** Large data warehouses (DWs), like the ones operated by Walmart, contain humongous amounts of data. Special methods, including parallel computing, Hadoop/Spark, and new cloud-based systems are available to organize, search, and mine the data. The costs related to data storage and mining are declining rapidly. Technologies that fall under the broad category of Big Data have enabled massive data coming from a variety of sources and in many different forms, which allows a very different view into organizational performance that was not possible in the past.
- **Analytical support.** With more data and analysis technologies, more alternatives can be evaluated, forecasts can be improved, risk analysis can be performed quickly, and the views of experts (some of whom may be in remote locations) can be collected quickly and at a reduced cost. Expertise can even be derived directly from analytical systems. With such tools, decision makers can perform complex simulations, check many possible scenarios, and assess diverse impacts quickly and economically. This, of course, is the focus of several chapters in the book.
- **Overcoming cognitive limits in processing and storing information.** According to Simon (1977), the human mind has only a limited ability to process and store information. People sometimes find it difficult to recall and use information in an error-free fashion due to their cognitive limits. The term *cognitive limits* indicates that an individual's problem-solving capability is limited when a wide range of diverse information and knowledge is required. Computerized systems enable people to overcome their cognitive limits by quickly accessing and processing vast amounts of stored information.
- **Knowledge management.** Organizations have gathered vast stores of information about their own operations, customers, internal procedures, employee interactions, and so forth, through the unstructured and structured communications taking place among the various stakeholders. Knowledge management systems have become sources of formal and informal support for decision making to managers, although they may not be called *KMS* anymore. Technologies such as text analytics are making it possible to generate value from such knowledge stores.
- **Anywhere, anytime support.** Using mobile technologies, managers can access information anytime and from anyplace, analyze and interpret it, and communicate with those involved. This perhaps is the biggest change that has occurred in the last few years. The speed at which information needs to be processed and converted into decisions has truly changed expectations for both consumers and businesses. These and other capabilities have been driving the use of computerized decision support since the late 1960s, but especially since the mid-1990s. The growth of mobile technologies, social media platforms, and analytical tools has enabled a different level of information systems (IS) support for managers. This growth in providing data-driven support for any decision extends to not just the managers but also to consumers. We will first study an overview of technologies that have been broadly referred to as BI. From there we will broaden our horizons to introduce various types of analytics.
- **Innovation and artificial intelligence.** Because of the complexities in the decision-making process discussed earlier and the environment surrounding the process, a more innovative approach is frequently needed. A major facilitation of innovation is provided by AI. Almost every step in the decision-making process can be influenced by AI. AI is also integrated with analytics, creating synergy in making decisions.

► SECTION 1.2 REVIEW QUESTIONS

1. What are some of the key system-oriented trends that have fostered IS-supported decision making to a new level?
2. List some capabilities of information systems that can facilitate managerial decision making.
3. How can a computer help overcome the cognitive limits of humans?

1.3 DECISION-MAKING PROCESSES AND COMPUTERIZED DECISION SUPPORT FRAMEWORK

In this section, we focus on some classical decision-making fundamentals and the decision-making process. These two concepts will help us ground much of what we will learn in terms of analytics, data science, and artificial intelligence.

Decision making is a process of choosing among two or more alternative courses of action for the purpose of attaining one or more goals. According to Simon (1977), managerial decision making is synonymous with the entire management process. Consider the important managerial function of planning. Planning involves a series of decisions: What should be done? When? Where? Why? How? By whom? Managers set goals, or plan; hence, planning implies decision making. Other managerial functions, such as organizing and controlling, also involve decision making.

Simon's Process: Intelligence, Design, and Choice

It is advisable to follow a systematic decision-making process. Simon (1977) said that this involves three major phases: intelligence, design, and choice. He later added a fourth phase: implementation. Monitoring can be considered a fifth phase—a form of feedback. However, we view monitoring as the *intelligence phase* applied to the *implementation phase*. Simon's model is the most concise and yet complete characterization of rational decision making. A conceptual picture of the decision-making process is shown in Figure 1.17. It is also illustrated as a decision support approach using modeling.

There is a continuous flow of activity from intelligence to design to choice (see the solid lines in Figure 1.17), but at any phase, there may be a return to a previous phase (feedback). Modeling is an essential part of this process. The seemingly chaotic nature of following a haphazard path from problem discovery to solution via decision making can be explained by these feedback loops.

The decision-making process starts with the **intelligence phase**; in this phase, the decision maker examines reality and identifies and defines the problem. *Problem ownership* is established as well. In the **design phase**, a model that represents the system is constructed. This is done by making assumptions that simplify reality and by writing down the relationships among all the variables. The model is then validated, and criteria are determined in a principle of choice for evaluation of the alternative courses of action that are identified. Often, the process of model development identifies alternative solutions and vice versa.

The **choice phase** includes the selection of a proposed solution to the model (not necessarily to the problem it represents). This solution is tested to determine its viability. When the proposed solution seems reasonable, we are ready for the last phase: implementation of the decision (not necessarily of a system). Successful implementation results in solving the real problem. Failure leads to a return to an earlier phase of the process. In fact, we can return to an earlier phase during any of the latter three phases. The decision-making situations described in the opening vignette follow Simon's four-phase model, as do almost all other decision-making situations.

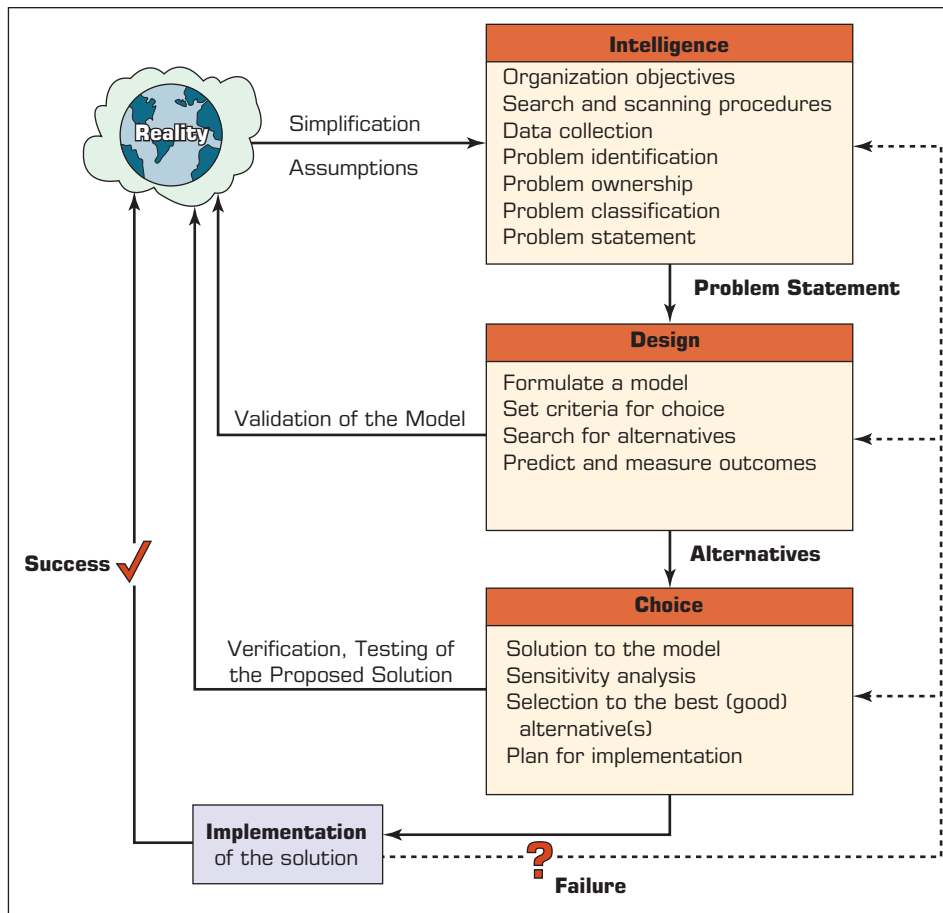


FIGURE 1.17 The Decision-Making/Modeling Process.

The Intelligence Phase: Problem (or Opportunity) Identification

The intelligence phase begins with the identification of organizational goals and objectives related to an issue of concern (e.g., inventory management, job selection, lack of or incorrect Web presence) and determination of whether they are being met. Problems occur because of dissatisfaction with the status quo. Dissatisfaction is the result of a difference between what people desire (or expect) and what is occurring. In this first phase, a decision maker attempts to determine whether a problem exists, identify its symptoms, determine its magnitude, and explicitly define it. Often, what is described as a problem (e.g., excessive costs) may be only a symptom (i.e., measure) of a problem (e.g., improper inventory levels). Because real-world problems are usually complicated by many interrelated factors, it is sometimes difficult to distinguish between the symptoms and the real problem. New opportunities and problems certainly may be uncovered while investigating the causes of symptoms.

The existence of a problem can be determined by monitoring and analyzing the organization's productivity level. The measurement of productivity and the construction of a model are based on real data. The collection of data and the estimation of future data are among the most difficult steps in the analysis.

ISSUES IN DATA COLLECTION

The following are some issues that may arise during data collection and estimation and thus plague decision makers:

- Data are not available. As a result, the model is made with and relies on potentially inaccurate estimates.
- Obtaining data may be expensive.
- Data may not be accurate or precise enough.
- Data estimation is often subjective.
- Data may be insecure.
- Important data that influence the results may be qualitative (soft).
- There may be too many data (i.e., information overload).
- Outcomes (or results) may occur over an extended period. As a result, revenues, expenses, and profits will be recorded at different points in time. To overcome this difficulty, a present-value approach can be used if the results are quantifiable.
- It is assumed that future data will be similar to historical data. If this is not the case, the nature of the change has to be predicted and included in the analysis.

When the preliminary investigation is completed, it is possible to determine whether a problem really exists, where it is located, and how significant it is. A key issue is whether an information system is reporting a problem or only the symptoms of a problem. For example, if reports indicate that sales are down, there is a problem, but the situation, no doubt, is symptomatic of a potentially bigger problem. It is critical to know the real problem. Sometimes it may be a problem of perception, incentive mismatch, or organizational processes rather than a poor decision model.

To illustrate why it is important to identify the problem correctly, we provide a classical example in Analytics in Action 1.1.

ANALYTICS IN ACTION 1.1 Making Elevators Go Faster!

This story has been reported in numerous places and has almost become a classic example to explain the need for problem identification. Ackoff (1960) described the problem of managing complaints about slow elevators in a tall hotel tower. After trying many solutions for reducing the complaint—staggering elevators to go to different floors, adding operators, and so on—the management determined that the real problem was not about the *actual* waiting time but rather the *perceived* waiting time. So the solution was to install full-length mirrors on elevator doors on each floor. The mirrors provided a distraction. By reducing the perceived waiting time, the problem went away. Baker and Cameron (1996) give several other examples of distractions, including lighting and displays, that organizations use to reduce perceived waiting time. If the real problem is identified as *perceived* waiting time, it can make a big difference in the proposed solutions and their costs. For example, full-length mirrors probably cost a whole lot less than adding an elevator!

Sources: Based on Baker, J., & Cameron M. (1996, September). The effects of the service environment on affect and consumer perception of waiting time: An integrative review and research propositions. *Journal of the Academy of Marketing Science*, 24, 338–349; Hesse, R., & Woolsey, G. (1975). *Applied Management Science: A Quick and Dirty Approach*. Chicago, IL: SRA Inc; Larson, R. C. (1987, November/December). Perspectives on queues: Social justice and the psychology of queuing. *Operations Research*, 35(6), 895–905.

PROBLEM CLASSIFICATION Problem classification is the conceptualization of a problem in an attempt to place it in a definable category, possibly leading to a standard solution approach. An important approach classifies problems according to the degree of structuredness evident in them. This ranges from totally structured (i.e., programmed) to totally unstructured (i.e., unprogrammed).

PROBLEM DECOMPOSITION Many complex problems can be divided into subproblems. Solving the simpler subproblems may help in solving a complex problem. Also, seemingly poorly structured problems sometimes have highly structured subproblems. Just as a semistructured problem results when some phases of decision making are structured whereas other phases are unstructured, and when some subproblems of a decision-making problem are structured with others unstructured, the problem itself is semistructured. As a decision support system is developed and the decision maker and development staff learn more about the problem, it gains structure.

PROBLEM OWNERSHIP In the intelligence phase, it is important to establish problem ownership. A problem exists in an organization only if someone or some group takes the responsibility for attacking it and if the organization has the ability to solve it. The assignment of authority to solve the problem is called *problem ownership*. For example, a manager may feel that they have a problem because interest rates are too high. Because interest rate levels are determined at the national and international levels and most managers can do nothing about them, high interest rates are the problem of the government, not a problem for a specific company to solve. The problem that companies actually face is how to operate in a high interest-rate environment. For an individual company, the interest rate level should be handled as an uncontrollable (environmental) factor to be predicted.

When problem ownership is not established, either someone is not doing their job or the problem at hand has yet to be identified as belonging to anyone. It is then important for someone to either volunteer to own it or assign it to someone.

The intelligence phase ends with a formal problem statement. This helps us develop a plan for addressing the problem.

The Design Phase

The design phase involves finding or developing and analyzing possible courses of action. These include understanding the problem and testing solutions for feasibility. A model of the decision-making problem is constructed, tested, and validated. Let us first define a model.

MODELS A major characteristic of computerized decision support and many BI tools (notably those of business analytics) is the inclusion of at least one model. The basic idea is to perform the analysis on a model of reality rather than on the real system. A *model* is a simplified representation or abstraction of reality. It is usually simplified because reality is too complex to describe exactly and because much of the complexity is actually irrelevant in solving a specific problem.

Modeling involves conceptualizing a problem and abstracting it to quantitative and/or qualitative form. For a mathematical model, the variables are identified and their mutual relationships are established. Simplifications are made, whenever necessary, through assumptions. For example, a relationship between two variables may be assumed to be linear even though in reality there may be some nonlinear effects. A proper balance between the level of model simplification and the representation of reality must be obtained because of the cost–benefit trade-off. A simpler model leads to lower development costs,

easier manipulation, and a faster solution but is less representative of the real problem and can produce inaccurate results. However, a simpler model generally requires fewer data, or the data are aggregated and easier to obtain.

The Choice Phase

Choice is the critical act of decision making. The choice phase is the one in which the actual decision and the commitment to follow a certain course of action are made. The boundary between the design and choice phases is often unclear because certain activities can be performed during both of them and because the decision maker can return frequently from choice activities to design activities (e.g., generate new alternatives while performing an evaluation of existing ones). The choice phase includes the search for, evaluation of, and recommendation of an appropriate solution to a model. A solution to a model is a specific set of values for the decision variables in a selected alternative. Choices can be evaluated as to their viability and profitability.

Each alternative must be evaluated. If an alternative has multiple goals, they must all be examined and balanced against each other. Sensitivity analysis is used to determine the robustness of any given alternative; slight changes in the parameters should ideally lead to slight or no changes in the alternative chosen. What-if analysis is used to explore major changes in the parameters. Goal seeking helps a manager determine values of the decision variables to meet a specific objective. These topics are addressed in Chapter 8.

The Implementation Phase

In *The Prince*, Machiavelli astutely noted some 500 years ago that there was “nothing more difficult to carry out, nor more doubtful of success, nor more dangerous to handle, than to initiate a new order of things.” The implementation of a proposed solution to a problem is, in effect, the initiation of a new order of things or the introduction of change. And change must be managed. User expectations must be managed as part of change management.

The definition of *implementation* is somewhat complicated because implementation is a long, involved process with vague boundaries. Simplistically, the **implementation phase** involves putting a recommended solution to work, not necessarily implementing a computer system. Many generic implementation issues, such as resistance to change, degree of support of top management, and user training, are important in dealing with information system–supported decision making. Indeed, many previous technology–related waves (e.g., business process reengineering [BPR] and knowledge management) have faced mixed results mainly because of change management challenges and issues. Management of change is almost an entire discipline in itself, so we recognize its importance and encourage readers to focus on it independently. Implementation also includes a thorough understanding of project management. The importance of project management goes far beyond analytics, so the last few years have witnessed a major growth in certification programs for project managers. A very popular certification now is the Project Management Professional (PMP). See pmi.org for more details.

Implementation must also involve collecting and analyzing data to learn from the previous decisions and improve the next decision. Although analysis of data is usually conducted to identify the problem and/or the solution, analytics should also be employed in the feedback process. This is especially true for any public policy decisions. We need to be sure that the data being used for problem identification is valid. Sometimes people find this out only after the implementation phase.

SECTION 1.3 REVIEW QUESTIONS

1. List and briefly describe Simon's four phases of decision making.
2. What is the difference between a problem and its symptoms?
3. Can you think of other situations where the real problem was different from what was apparent at the surface?
4. Which phase is likely to involve data collection and model building?
5. COVID-19 pandemic presented a major problem to be addressed at various levels—from policy makers, healthcare providers, business owners, workers, and of course, the general public. Take any one of these groups and employ the phases of decision making from that group's perspective to understand how a problem would have been identified, solutions developed, and a choice made to address it.

1.4 EVOLUTION OF COMPUTERIZED DECISION SUPPORT TO ANALYTICS/DATA SCIENCE

The timeline in Figure 1.18 shows the terminology used to describe analytics since the 1970s. During the 1970s, the primary focus of information systems support for decision making focused on providing structured, periodic reports that a manager could use for decision making (or ignore them). Businesses began to create routine reports to inform decision makers (managers) about what had happened in the previous period (e.g., day, week, month, quarter). Although it was useful to know what had happened in the past, managers needed more than this: They needed a variety of reports at different levels of granularity to better understand and address changing needs and challenges of the business. These were usually called management information systems (MIS). In the early 1970s, Scott-Morton first articulated the major concepts of decision support system (DSS). He defined DSSs as “interactive computer-based systems, which help decision makers utilize *data* and *models* to solve unstructured problems” (Gorry and Scott-Morton, 1971). The following is another classic DSS definition, provided by Keen and Scott-Morton (1978):

Decision support systems couple the intellectual resources of individuals with the capabilities of the computer to improve the quality of decisions. It is a computer-based support system for management decision makers who deal with semistructured problems.

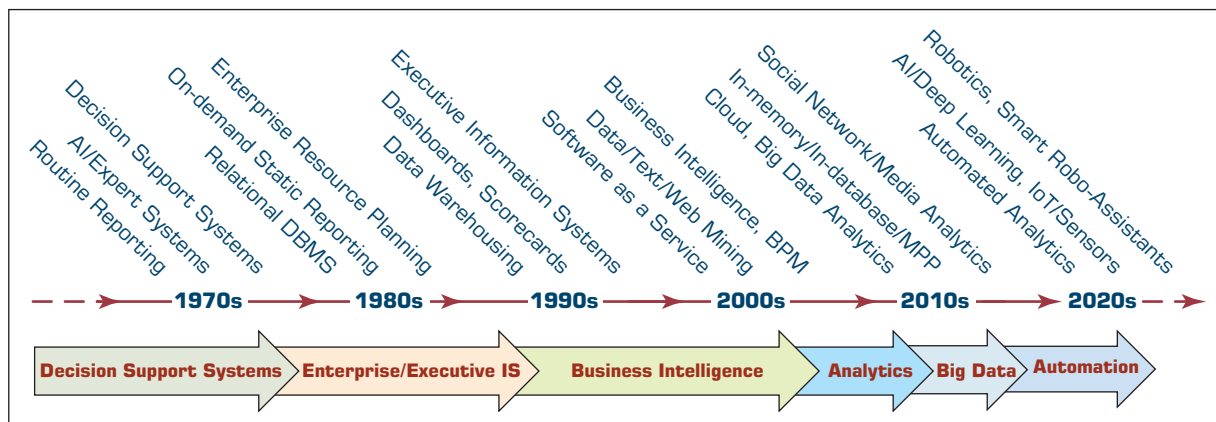


FIGURE 1.18 Evolution of Decision Support, Business Intelligence, Analytics, and AI.

Note that the term *decision support system*, like *management information system* and several other terms in the field of IT, is a content-free expression (i.e., it means different things to different people). Therefore, there is no universally accepted definition of DSS.

During the early days of analytics, data were often obtained from the domain experts using manual processes (i.e., interviews and surveys) to build mathematical or knowledge-based models to solve constrained optimization problems. The idea was to do the best with limited resources. Such decision support models were typically called operations research (OR). The problems that were too complex to solve optimally (using linear or nonlinear mathematical programming techniques) were tackled using heuristic methods such as simulation models. (We will introduce these as prescriptive analytics later in this chapter and in a bit more detail in Chapter 8.)

In the late 1970s and early 1980s, in addition to the mature OR models that were being used in many industries and government systems, a new and exciting line of models had emerged: rule-based expert systems (ES). These systems promised to capture experts' knowledge in a format that computers could process (via a collection of if-then-else rules or heuristics) so that these could be used for consultation much the same way that one would use domain experts to identify a structured problem and to prescribe the most probable solution. ESs allowed scarce expertise to be made available where and when needed, using an "intelligent" DSS. However, other technologies soon took over.

The 1980s saw a significant change in the way organizations captured business-related data. The old practice had been to have multiple disjointed information systems tailored to capture transactional data of different organizational units or functions (e.g., accounting, marketing and sales, finance, manufacturing). In the 1980s, these systems were integrated as enterprise-level information systems that we now commonly call enterprise resource planning (ERP) systems. The old mostly sequential and nonstandardized data representation schemas were replaced by relational database management (RDBM) systems. These systems made it possible to improve the capture and storage of data, as well as the relationships between organizational data fields while significantly reducing the replication of information. The need for RDBM and ERP systems emerged when data integrity and consistency became an issue, significantly hindering the effectiveness of business practices. With ERP, all the data from every corner of the enterprise is collected and integrated into a consistent schema so that every part of the organization has access to the single version of the truth when and where needed. In addition to the emergence of ERP systems, or perhaps because of these systems, business reporting became an on-demand, as-needed business practice. Decision makers could decide when they needed to or wanted to create specialized reports to investigate organizational problems and opportunities.

In the 1990s, the need for more versatile reporting led to the development of executive information systems (EISs; DSSs designed and developed specifically for executives and their decision-making needs). These systems were designed as graphical dashboards and scorecards so that they could serve as visually appealing displays while focusing on the most important factors for decision makers to keep track of the key performance indicators. To make this highly versatile reporting possible while keeping the transactional integrity of the business information systems intact, it was necessary to create a middle tier known as a data warehouse (DW) as a repository to specifically support business reporting and decision making. In a very short time, most large to medium-sized businesses adopted data warehousing as their platform for enterprise-wide decision making. The dashboards and scorecards got their data from a DW, and by doing so, they were not hindering the efficiency of the business transaction systems mostly referred to as (ERP) systems.

In the 2000s, the DW-driven DSSs began to be called BI systems. As the amount of longitudinal data accumulated in the DWs increased, so did the capabilities of hardware and software to keep up with the rapidly changing and evolving needs of the decision

makers. Because of the globalized competitive marketplace, decision makers needed current information in a very digestible format to address business problems and to take advantage of market opportunities in a timely manner. Because the data in a DW is updated periodically, it does not reflect the latest information. To elevate this information latency problem, DW vendors developed a system to update the data more frequently, which led to the terms *real-time data warehousing* and, more realistically, *right-time data warehousing*, which differs from the former by adopting a data-refreshing policy based on the needed freshness of the data items (i.e., not all data items need to be refreshed in real time). DWs are very large and feature rich, and it became necessary to “mine” the corporate data to “discover” new and useful knowledge nuggets to improve business processes and practices, hence the terms *data mining* and *text mining*. With the increasing volumes and varieties of data, the needs for more storage and more processing power emerged. Although large corporations had the means to tackle this problem, small to medium-sized companies needed more financially manageable business models. This need led to service-oriented architecture and software and infrastructure-as-a-service analytics business models. Smaller companies, therefore, gained access to analytics capabilities on an as-needed basis and paid only for what they used, as opposed to investing in financially prohibitive hardware and software resources.

In the 2010s, we saw yet another paradigm shift in the way that data is captured and used. Largely because of the widespread use of the Internet, new data generation mediums have emerged. Of all the new data sources (e.g., radio-frequency identification [RFID] tags, digital energy meters, clickstream Web logs, smart home devices, wearable health monitoring equipment), perhaps the most interesting and challenging is social networking/social media. This unstructured data is rich in information content, but analysis of such data sources poses significant challenges to computational systems, from both software and hardware perspectives. Recently, the term *Big Data* has been coined to highlight the challenges that these new data streams have brought on us. Many advancements in both hardware (e.g., massively parallel processing with very large computational memory and highly parallel multiprocessor computing systems) and software/algorithms (e.g., Hadoop with MapReduce and NoSQL) have been developed to address the challenges of Big Data.

The last few years and the upcoming decade are bringing massive growth in many exciting dimensions. For example, streaming analytics and the sensor technologies have enabled the IoT. Artificial intelligence is changing the shape of BI by enabling new ways of analyzing images through deep learning, not just traditional visualization of data. Deep learning and AI are also helping grow voice recognition and speech synthesis, leading to new interfaces in interacting with technologies. Almost half of U.S. households already have a smart speaker—such as Amazon Echo or Google Home—and interact with data and systems using voice interfaces. Growth in video interfaces will eventually enable gesture-based interaction with systems. Recent buzz about chatGPT and other AI advances has increased public's interest as well. All of these are being enabled due to massive cloud-based data storage and amazingly fast processing capabilities. And more is yet to come.

It's hard to predict what the next decade will bring and what the new analytics-related terms will be. The time between new paradigm shifts in information systems and particularly in analytics has been shrinking, and this trend will continue for the foreseeable future. Even though analytics is not new, the explosion in its popularity is very new. Thanks to the recent explosion in Big Data, ways to collect and store this data, and intuitive software tools, data-driven insights are more accessible to business professionals than ever before. Therefore, in the midst of global competition, there is a huge opportunity to make better managerial decisions by using data and analytics to increase revenue while decreasing costs by building better products, improving customer experience, and catching fraud before it happens, improving customer engagement through targeting and

customization all with the power of analytics and data. More and more companies are now preparing their employees with the know-how of business analytics to drive effectiveness and efficiency in their day-to-day decision-making processes.

The recent COVID-19 pandemic has had a dual edge impact on analytics and data science. On the one hand, it demonstrated the need for data driven decision-making and use of models in projecting spread and impacts of the disease. On the other hand, the recent macro economic impact led many companies to conserve their resources, focus on the short-term survival, and thus cut down on their investments in building their analytics capabilities. We believe it is a temporary phenomenon. When the economic activity gets back to normal, the focus on analytics and data science will also return.

The next section focuses on a framework for BI. Although most people would agree that BI has evolved into analytics and data science, many vendors and researchers still use that term. So, Section 1.5 pays homage to that history by specifically focusing on what has been called BI. Following the next section, we introduce analytics and will use that as the label for classifying all related concepts.

► SECTION 1.4 REVIEW QUESTIONS

1. List three of the terms that have been predecessors of analytics.
2. What was the primary difference between the systems called MIS, DSS, and Executive Support Systems?
3. Did DSS evolve into BI or vice versa?

1.5 A FRAMEWORK FOR BUSINESS INTELLIGENCE

The decision support concepts presented in Sections 1.2 through 1.4 have been implemented incrementally, under different names, by many vendors that have created tools and methodologies for decision support. As noted in Section 1.4, as the enterprise-wide systems grew, managers were able to access user-friendly reports that enabled them to make decisions quickly. These systems, which were generally called EISs, then began to offer additional visualization, alerts, and performance measurement capabilities. By 2006, the major *commercial* products and services appeared under the term *business intelligence* (BI).

Definitions of BI

Business intelligence (BI) is an umbrella term that combines architectures, tools, databases, analytical tools, applications, and methodologies. It is, like DSS, a content-free expression, so it means different things to different people. Part of the confusion about BI lies in the flurry of acronyms and buzzwords that are associated with it (e.g., business performance management [BPM]). BI's major objective is to enable interactive access (sometimes in real time) to data, to enable manipulation of data, and to give business managers and analysts the ability to conduct appropriate analyses. By analyzing historical and current data, situations, and performances, decision makers get valuable insights that enable them to make more informed and better decisions. The process of BI is based on the *transformation* of data to information, then to decisions, and finally to actions.

A Brief History of BI

The term *BI* was coined by the Gartner Group in the mid-1990s. However, as the history in the previous section points out, the concept is much older; it has its roots in the MIS reporting systems of the 1970s. During that period, reporting systems were static, were two dimensional, and had no analytical capabilities. In the early 1980s, the concept of EISs emerged. This concept expanded the computerized support to top-level managers

and executives. Some of the capabilities introduced were dynamic multidimensional (ad hoc or on-demand) reporting, forecasting and prediction, trend analysis, drill-down to details, status access, and critical success factors. These features appeared in dozens of commercial products until the mid-1990s. Then the same capabilities and some new ones appeared under the name BI. Today, a good BI-based enterprise information system contains all the information executives need. So, the original concept of EIS was transformed into BI. By 2005, BI systems started to include *artificial intelligence* capabilities as well as powerful analytical capabilities. Figure 1.19 illustrates the various tools and techniques that may be included in a BI system. It illustrates the evolution of BI as well. The tools shown in Figure 1.19 provide the capabilities of BI. The most sophisticated BI products include most of these capabilities; others specialize in only some of them.

The Architecture of BI

A BI system has four major components: a *DW*, with its source data; *business analytics*, a collection of tools for manipulating, mining, and analyzing the data in the DW; *BPM* for monitoring and analyzing performance; and a *user interface* (e.g., a **dashboard**). The relationship among these components is illustrated in Figure 1.20.

The Origins and Drivers of BI

Where did modern approaches to data warehousing and BI come from? What are their roots, and how do those roots affect the way organizations are managing these initiatives today? Today's investments in information technology are under increased scrutiny in terms of their bottom-line impact and potential. The same is true of DW and the BI applications that make these initiatives possible.

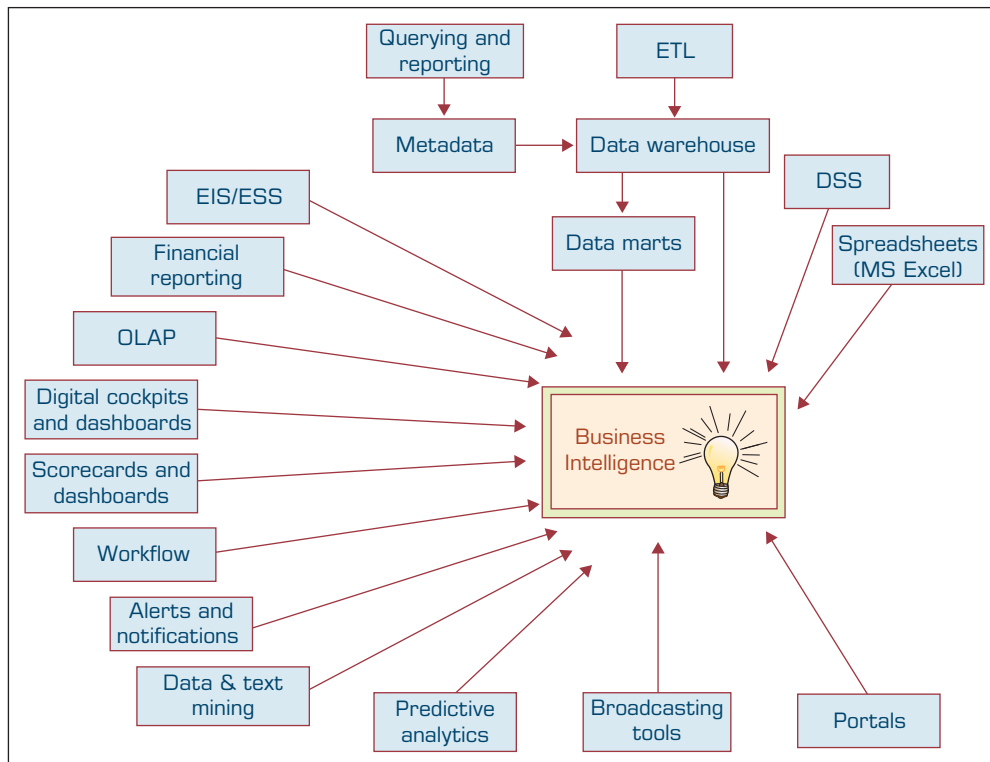


FIGURE 1.19 Evolution of Business Intelligence (BI).

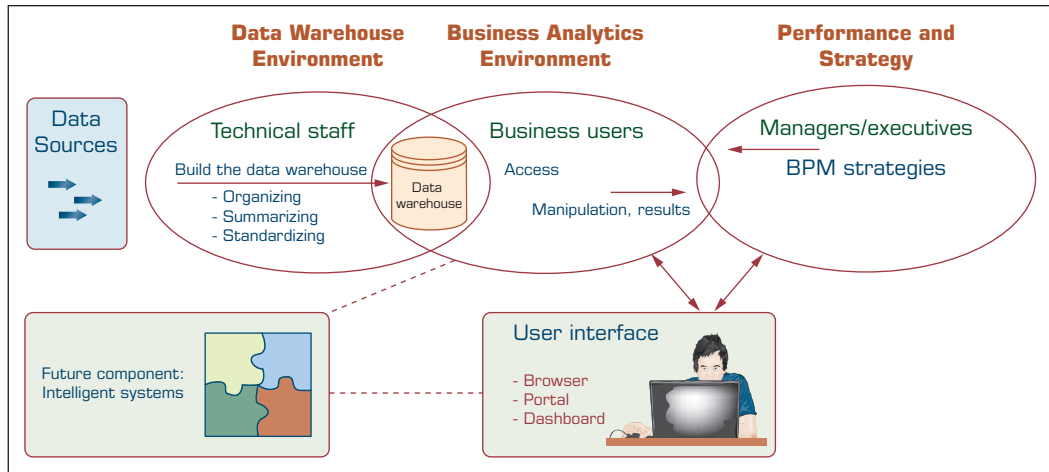


FIGURE 1.20 A High-Level Architecture of BI.

Organizations are being compelled to capture, understand, and harness their data to support decision making to improve business operations. Legislation and regulation (e.g., the Sarbanes-Oxley Act of 2002) now require business leaders to document their business processes and to sign off on the legitimacy of the information they rely on and report to stakeholders. Moreover, business cycle times are now extremely compressed; faster, more informed, and better decision-making is, therefore, a competitive imperative. Managers need the *right information* at the *right time* and in the *right place*. This is the mantra for modern approaches to BI.

Organizations have to work smart. Paying careful attention to the management of BI initiatives is a necessary aspect of doing business. It is no surprise, then, that organizations are increasingly championing BI and under its new incarnation as analytics.

A Multimedia Exercise in Business Intelligence

Teradata University for Academics (TUA) includes videos (similar to the television show *CSI*) to illustrate concepts of analytics in different industries. These are called “BSI Videos (Business Scenario Investigations).” Not only are these entertaining, but they also provide the class with some questions for discussion. For starters, please go to <https://learning-academics.teradata.com> and click on the BSI videos tab. Find the case-of-the-misconnecting-passengers or go to www.youtube.com/watch?v=NXEL5F4_aKA. Watch the video that appears on YouTube. Essentially, you have to assume the role of a customer service center professional. An incoming flight is running late, and several passengers are likely to miss their connecting flights. There are seats on one outgoing flight that can accommodate two of the four passengers. Which two passengers should be given priority? You are given information about customers’ profiles and relationships with the airline. Your decisions might change as you learn more about those customers’ profiles.

Watch the video, pause it as appropriate, and answer the questions on which passengers should be given priority. Then resume the video to get more information. After the video is complete, you can see the slides related to this video and how the analysis was prepared on a slide set at www.slideshare.net/teradata/bsi-how-we-did-it-the-case-of-the-misconnecting-passengers. This multimedia excursion provides an example of how additional available information through an enterprise DW can assist in decision making.

Although some people equate DSS with BI, these systems are not, at present, the same. It is interesting to note that some people believe that DSS is a part of BI—one of its analytical tools. Others think that BI is a special case of DSS that deals mostly with

reporting, communication, and collaboration (a form of data-oriented DSS). Another explanation (Watson, 2005) is that BI is a result of a continuous revolution, and as such, DSS is one of BI's original elements. Further, as noted in the next section onward, in many circles BI has been subsumed by the new terms *analytics* or *data science*.

Transaction Processing versus Analytic Processing

To illustrate the major characteristics of BI, first we will show what BI is not—namely, transaction processing. We're all familiar with the information systems that support our transactions, like ATM withdrawals, bank deposits, cash register scans at the grocery store, and so on. These *transaction processing* systems are constantly involved in handling updates to what we might call *operational databases*. For example, in an ATM withdrawal transaction, we need to reduce our bank balance accordingly; a bank deposit adds to an account; and a grocery store purchase is likely reflected in the store's calculation of total sales for the day, and it should reflect an appropriate reduction in the store's inventory for the items we bought, and so on. These **online transaction processing (OLTP)** systems handle a company's routine ongoing business. In contrast, a DW is typically a distinct system that provides storage for data that will be used for *analysis*. The intent of that analysis is to give management the ability to scour data for information about the business, and it can be used to provide tactical or operational decision support, whereby, for example, line personnel can make quicker and/or more informed decisions. We will provide a more technical definition of DW in Chapter 3, but suffice it to say that DWs are intended to work with informational data used for **online analytical processing (OLAP)** systems.

Most operational data in enterprise resources planning (ERP) systems—and in its complementary siblings like *supply chain management* (SCM) or *CRM*—are stored in an OLTP system, which is a type of computer processing where the computer responds immediately to user requests. Each request is considered to be a *transaction*, which is a computerized record of a discrete event, such as the receipt of inventory or a customer order. In other words, a transaction requires a set of two or more database updates that must be completed in an all-or-nothing fashion.

The very design that makes an OLTP system efficient for transaction processing makes it inefficient for end-user ad hoc reports, queries, and analysis. In the 1980s, many business users referred to their mainframes as “black holes” because all the information went into them, but none ever came back. All requests for reports had to be programmed by the IT staff, whereas only “pre-canned” reports could be generated on a scheduled basis, and ad hoc real-time querying was virtually impossible. Although the client/server-based ERP systems of the 1990s were somewhat more report-friendly, it has still been a far cry from a desired usability by regular, nontechnical, end users for things such as operational reporting, interactive analysis, and so on. To resolve these issues, the notions of DW and BI were created.

DWs contain a wide variety of data that present a coherent picture of business conditions at a single point in time. The idea was to create a database infrastructure that was always online and contained all the information from the OLTP systems, including historical data, but reorganized and structured in such a way that it was fast and efficient for querying, analysis, and decision support. Separating the OLTP from analysis and decision support enables the benefits of BI that were described earlier.

Appropriate Planning and Alignment with the Business Strategy

First and foremost, the fundamental reasons for investing in BI must be aligned with the company's business strategy. BI cannot simply be a technical exercise for the information systems department. It has to serve as a way to change the manner in which the company

conducts business by improving its business processes and transforming decision-making processes to be more data driven. Many BI consultants and practitioners involved in successful BI initiatives advise that a framework for planning is a necessary precondition. One framework, developed by Gartner, Inc. (2004), decomposes planning and execution into *business*, *organization*, *functionality*, and *infrastructure* components. At the business and organizational levels, strategic and operational objectives must be defined while considering the available organizational skills to achieve those objectives. Issues of organizational culture surrounding BI initiatives and building enthusiasm for those initiatives and procedures for the intra-organizational sharing of BI best practices must be considered by upper management—with plans in place to prepare the organization for change. One of the first steps in that process is to assess the IS organization, the skill sets of the potential classes of users, and whether the culture is amenable to change. From this assessment, and assuming there is justification and the need to move ahead, a company can prepare a detailed action plan. Another critical issue for BI implementation success is the integration of several BI projects (most enterprises use several BI projects) among themselves and with the other IT systems in the organization and its business partners.

If the company's strategy is properly aligned with the reasons for DW and BI initiatives, and if the company's IS organization is or can be made capable of playing its role in such a project, and if the requisite user community is in place and has the proper motivation, it is wise to start BI and establish a BI Competency Center within the company. The center could serve some or all of the following functions (Gartner, 2004):

- The center can demonstrate how BI is clearly linked to strategy and execution of strategy.
- A center can serve to encourage interaction between the potential business user communities and the IS organization.
- The center can serve as a repository and disseminator of best BI practices between and among the different lines of business.
- Standards of excellence in BI practices can be advocated and encouraged throughout the company.
- The IS organization can learn a great deal through interaction with the user communities, such as knowledge about the variety of types of analytical tools that are needed.
- The business user community and IS organization can better understand why the DW platform must be flexible enough to provide for changing business requirements.
- It can help important stakeholders like high-level executives see how BI can play an important role.

Another important success factor of BI is its ability to facilitate a real-time, on-demand agile environment, introduced next.

Real-Time, On-Demand BI Is Attainable

The demand for instant, on-demand access to dispersed information has grown as the need to close the gap between the operational data and strategic objectives has become more pressing. As a result, a category of products called *real-time BI applications* has emerged. The introduction of new data-generating technologies, such as RFID and other sensors is only accelerating this growth and the subsequent need for real-time BI. Traditional BI systems use a large volume of *static* data that has been extracted, cleansed, and loaded into a *DW* to produce reports and analyses. However, the need is not just reporting because users need business monitoring, performance analysis, and an understanding of why things are happening. These can assist users, who need to know (virtually in real time) about changes in data or the availability of relevant reports, alerts, and

notifications regarding events and emerging trends in social media applications. In addition, business applications can be programmed to act on what these real-time BI systems discover. For example, an SCM application might automatically place an order for more “widgets” when real-time inventory falls below a certain threshold or when a CRM application automatically triggers a customer service representative and credit control clerk to check a customer who has placed an online order larger than \$10,000.

One approach to real-time BI uses the DW model of traditional BI systems. In this case, products from innovative BI platform providers provide a service-oriented, near-real-time solution that populates the DW much faster than the typical nightly *extract/transfer/load* batch update does. A second approach, commonly called *business activity management* (BAM), is adopted by pure-play BAM and/or hybrid BAM-middleware providers (such as Savvion, Iteration Software, Vitria, webMethods, Quantive, Tibco, or Vineyard Software). It bypasses the DW entirely and uses **Web services** or other monitoring means to discover key business events. These software monitors (or **intelligent agents**) can be placed on a separate server in the network or on the transactional application databases themselves, and they can use event- and process-based approaches to proactively and intelligently measure and monitor operational processes.

Developing or Acquiring BI Systems

Today, many vendors offer diversified tools, some of which are completely preprogrammed (called *shells*); all you have to do is insert your numbers. These tools can be purchased or leased. For a list of products, demos, white papers, and more current product information, see product directories at tdwi.org. Free user registration is required. Almost all BI applications are constructed with shells provided by vendors who may themselves create a custom solution for a client or work with another outsourcing provider. The issue that companies face is which alternative to select: purchase, lease, or build. Each of these alternatives has several options. One of the major criteria for making the decision is justification and cost–benefit analysis.

Justification and Cost–Benefit Analysis

As the number of potential BI applications increases, the need to justify and prioritize them arises. This is not an easy task due to the large number of intangible benefits. Both direct and intangible benefits need to be identified. Of course, this is where the knowledge of similar applications in other organizations and case studies is extremely useful. For example, The Data Warehousing Institute (tdwi.org) provides a wealth of information about products and innovative applications and implementations. Such information can be useful in estimating direct and indirect benefits.

Security and Protection of Privacy

This is an extremely important issue in the development of any computerized system, especially BI that contains data that may possess strategic value. Also, the privacy of employees and customers needs to be protected.

Integration of Systems and Applications

With the exception of some small applications, all BI applications must be integrated with other systems such as databases, legacy systems, enterprise systems (particularly ERP and CRM), e-commerce (sell side, buy side), and many more. In addition, BI applications are usually connected to the Internet and many times to information systems of business partners.

Furthermore, BI tools sometimes need to be integrated among themselves, creating synergy. The need for integration pushed software vendors to continuously add capabilities to their products. Customers who buy an all-in-one software package deal with only one vendor and do not have to deal with system connectivity. But, they may lose the advantage of creating systems composed from the “best-of-breed” components.

► SECTION 1.5 REVIEW QUESTIONS

1. Define *BI*.
2. List and describe the major components of BI.
3. Define *OLTP*.
4. Define *OLAP*.
5. List some of the implementation topics addressed by Gartner's report.
6. List some other success factors of BI.

1.6 ANALYTICS OVERVIEW

The word *analytics* has largely replaced the previous individual components of computerized decision support technologies that have been available under various labels in the past. Indeed, many practitioners and academics now use the word *analytics* in place of BI. Although many authors and consultants have defined it slightly differently, one can view **analytics** as the process of developing actionable decisions or recommendations for actions based on insights generated from historical data. According to the Institute for Operations Research and Management Science (INFORMS), analytics represents the combination of computer technology, management science techniques, and statistics to solve real problems. Of course, many other organizations have proposed their own interpretations and motivations for analytics. For example, SAS Institute Inc. proposed eight levels of analytics that begin with standardized reports from a computer system. These reports essentially provide a sense of what is happening with an organization. Additional technologies have enabled us to create more customized reports that can be generated on an ad hoc basis. The next extension of reporting takes us to OLAP-type queries that allow a user to dig deeper and determine specific sources of concern or opportunities. Technologies available today can also automatically issue alerts for a decision maker when performance warrants such alerts. At a consumer level, we see such alerts for weather or other issues. But similar alerts can also be generated in specific settings when sales fall above or below a certain level within a certain time period or when the inventory for a specific product is running low. All of these applications are made possible through analysis and queries on data being collected by an organization. The next level of analysis might entail statistical analysis to better understand patterns. These can then be taken a step further to develop forecasts or models for predicting how customers might respond to a specific marketing campaign or ongoing service/product offerings. When an organization has a good view of what is happening and what is likely to happen, it can also employ other techniques to make the best decisions under the circumstances.

This idea of looking at all the data to understand what is happening, what will happen, and how to make the best of it has also been encapsulated by INFORMS in proposing three levels of analytics. These three levels are identified (informatics.org/Community/Analytics) as descriptive, predictive, and prescriptive. Figure 1.21 presents a graphical view of these three levels of analytics. It suggests that these three are somewhat independent steps and one type of analytics applications leads to another. It also suggests that there is actually some overlap across these three types of analytics. In either case, the interconnected nature of different types of analytics applications is evident. We next introduce these three levels of analytics.

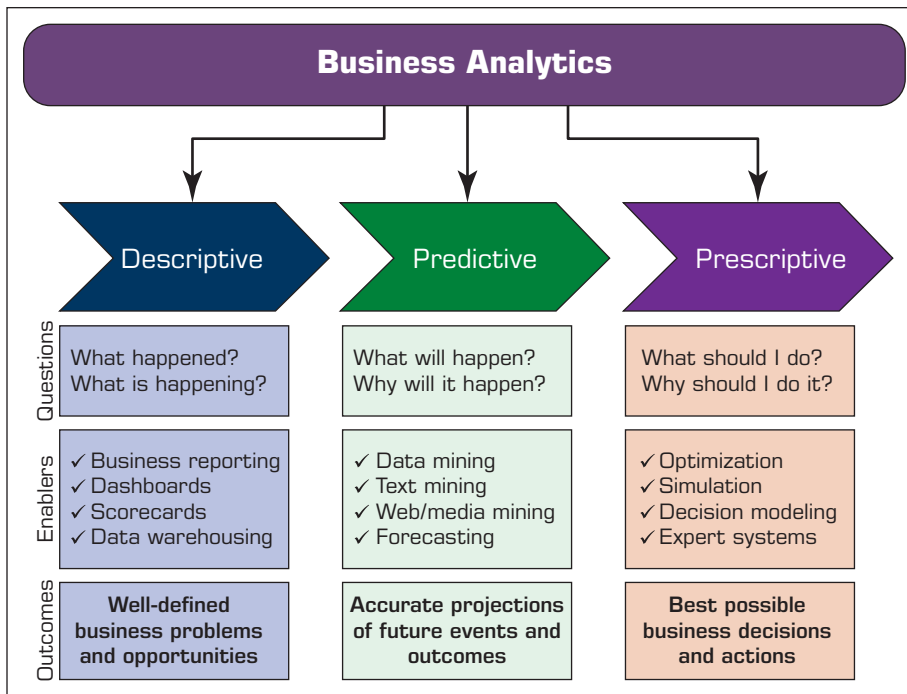


FIGURE 1.21 Three Types of Analytics.

Descriptive Analytics

Descriptive (or reporting) analytics refers to knowing what is happening in the organization and understanding some underlying trends and causes of such occurrences. First, this involves the consolidation of data sources and availability of all relevant data in a form that enables appropriate reporting and analysis. Usually, the development of this data infrastructure is part of DWs. From this data infrastructure we can develop appropriate reports, queries, alerts, and trends using various reporting tools and techniques.

A significant technology that has become a key player in this area is visualization. Using the latest visualization tools in the marketplace, we can now develop powerful insights in the operations of our organization. Chapters 4 and 5 focus on data and descriptive analytics. Analytics in Action 1.2 highlights one such classical application. Color renderings of visualizations discussed in these applications are available online.

Predictive Analytics

Predictive analytics aims to determine what is likely to happen in the future. This analysis is based on statistical techniques as well as other more recently developed techniques that fall under the general category of **data mining**. The goal of these techniques is to be able to predict if the customer is likely to switch to a competitor (“churn”), what the customer would likely buy next and how much, what promotions a customer would respond to, whether this customer is a creditworthy risk, and so forth. A number of techniques are used in developing predictive analytical applications, including various classification algorithms. For example, as described in Chapters 4 and 5, we can use classification techniques such as logistic regression, decision tree models, and neural networks to predict how well a motion picture will do at the box office. We can also use clustering algorithms for segmenting customers into different clusters to be able to target specific promotions to them. Finally, we can use association mining techniques to estimate relationships between different purchasing behaviors. That is, if a customer buys one product, what else

ANALYTICS IN ACTION 1.2 Silvaris Increases Business with Visual Analysis and Real-Time Reporting Capabilities

Silvaris Corporation was founded in 2000 by a team of forest industry professionals to provide technological advancement in the lumber and building material sector. Silvaris is the first e-commerce platform in the United States specifically for forest products and is headquartered in Seattle, Washington. It is a leading wholesale provider of industrial wood products and surplus building materials.

Silvaris sells its products and provides international logistics services to more than 3,500 customers. To manage various processes that are involved in a transaction, they created a proprietary online trading platform to track information flow related to transactions between traders, accounting, credit, and logistics. This allowed Silvaris to share its real-time information with its customers and partners. But due to the rapidly changing prices of materials, it became necessary for Silvaris to get a real-time view of data without moving data into a separate reporting format.

Silvaris started using Tableau because of its ability to connect with and interpret live data. Due to dashboards created by Tableau that are easy to understand and explain, Silvaris started using Tableau for reporting purposes. This helped Silvaris in pulling out information quickly from the data and identifying issues that impact their business. Silvaris succeeded in managing online versus offline orders with the help of reports generated by Tableau. Now, Silvaris keeps track of online orders placed by customers and knows when to send renew pushes to which customers to keep them purchasing online. Also, analysts of Silvaris can save time by generating dashboards instead of writing hundreds of pages of reports by using Tableau.

QUESTIONS FOR DISCUSSION

1. What was the challenge faced by Silvaris?
2. How did Silvaris solve its problem using data visualization with Tableau?

What We Can Learn from This Analytics in Action

Many industries need to analyze data in real time. Real-time analysis enables the analysts to identify issues that impact their business. Visualization is sometimes the best way to begin analyzing the live data streams. Tableau is one such data visualization tool that has the capability to analyze live data without bringing live data into a separate reporting format.

Sources: Tableau.com, “Silvaris Augments Proprietary Technology Platform with Tableau’s Real-Time Reporting Capabilities,” https://www.tableau.com/sites/default/files/case-studies/silvaris-business-dashboards_0.pdf (accessed July 2022); Silvaris.com, “Overview,” <http://www.silvaris.com/About/> (accessed July 2022).

is the customer likely to purchase? Such analysis can assist a retailer in recommending or promoting related products. For example, any product search on Amazon.com results in the retailer also suggesting other similar products that a customer may be interested in. We will study these techniques and their applications in Chapters 5 through 7. Analytics in Action 1.3 illustrates a comprehensive story of a large organization’s challenges of compiling the data from various sources, performing some descriptive analytics on the data and also conducting predictive analytics work on the data to make optimal decisions.

Prescriptive Analytics

The third category of analytics is termed **prescriptive analytics**. The goal of prescriptive analytics is to recognize what is going on as well as the likely forecast and make decisions to achieve the best performance possible. This group of techniques has historically been studied under the umbrella of OR or management sciences and is generally aimed at optimizing the performance of a system. The goal here is to provide a decision or a recommendation for a specific action. These recommendations can be in the form of a

ANALYTICS IN ACTION 1.3 How Big Will Be the Beef?

Understanding Animals' Eating Behavior and Their Final Weight

Land O'Lakes is a cooperative of hundreds of dairy farmers. It started in 1921, and is a major producer of dairy products. A subsidiary of Land O'Lakes, Purina Animal Nutrition LLC (www.purinamills.com) is a national organization serving producers, animal owners and their families through more than 4,700 local cooperatives, independent dealers, and other large retailers throughout the United States. Over the years, they've developed over 235 nutritional solutions for animals across 24 animal species. By using a talented group of PhD nutritionists and veterinarians, they have conducted more than 24,000 research studies in fields across the country and registered over 125 patents. To date, their research teams have analyzed over 1,500 nutrient combinations to drive product development and rigorous testing of new diet formulations, innovative milk replacers, and nutritional programs.

There was always a need to better understand and predict an animal's performance, feed intake and health issues, discover hidden relationships, patterns, and seasonality from all the data collected. However, it was difficult, if not impossible because they did not have processes/pipelines/strategies/teams in place to compile and analyze the results of these hundreds of studies. They had both structured and unstructured data—feed records on paper, old legacy data in tapes (OCR), spreadsheets, flat files, pdfs, word file summaries, data dumps from feed management systems, raw data from sensors, data from APIs, final summarized data (aggregated), and field trial data (observational). It was necessary to get all the research data (Big Data) at one place in a uniform format to be able to make use of these data sources to help make better predictions and then decisions on products.

We can categorize the challenges in two main buckets: data engineering challenges and data science challenges.

Data Engineering Challenges:

- **Scattered**—Datasets were scattered all over the place. Significant efforts were needed to bring the datasets at a central location.
- **Missing**—Many historical/legacy datasets had lots of rows and columns missing. The data collection processes evolved over the time, leading to changes in data fields, etc.
- **Variety**—As can be expected, they deal with the entire life cycle of animals, from farm to fork. This clearly leads to a large variety of attributes/datasets available for respective species. Examples include breeding, gestation, pregnancy, litter, body weights, locomotion, milk yield, carcass, performance, health observation, blood tissue and excretion related data. The major challenge here was to merge all the above datasets and set some common goals and metrics.
- **Granularity**—Some of the datasets were raw while other were aggregated and summarized so it was hard to bring them in a form where you can track all the activities by a particular animal.
- **Messy**—Datasets also included all kinds of special characters, identifiers, and delimiters. In some places, all the columns came were recorded as a single column. At other places everything was recorded as textual data.
- **Lack of Data Governance**—There were no governance processes in place to validate the data records.
- **No Real-time access**—Since the datasets were scattered it was difficult to automate the process and make the data pipelines. Specific programs had to be developed to be able to pull the data in a central repository.
- **No clear boundaries in the data flows**—There was no clear boundary in transactional layer (OLTP) versus analytical layer (OLAP) of the data flows.

To address the above data challenges the data science team implemented a precision agriculture solution called Purina Animal Nutrition Data Analytics (PANDA). PANDA is hosted on Microsoft Azure cloud. First, they architected, designed, and built the data pipelines to make a central repository of all data collected at the farm using Microsoft Azure Cloud services. One of the early key decisions was to identify what outcomes to achieve and monitor. The company

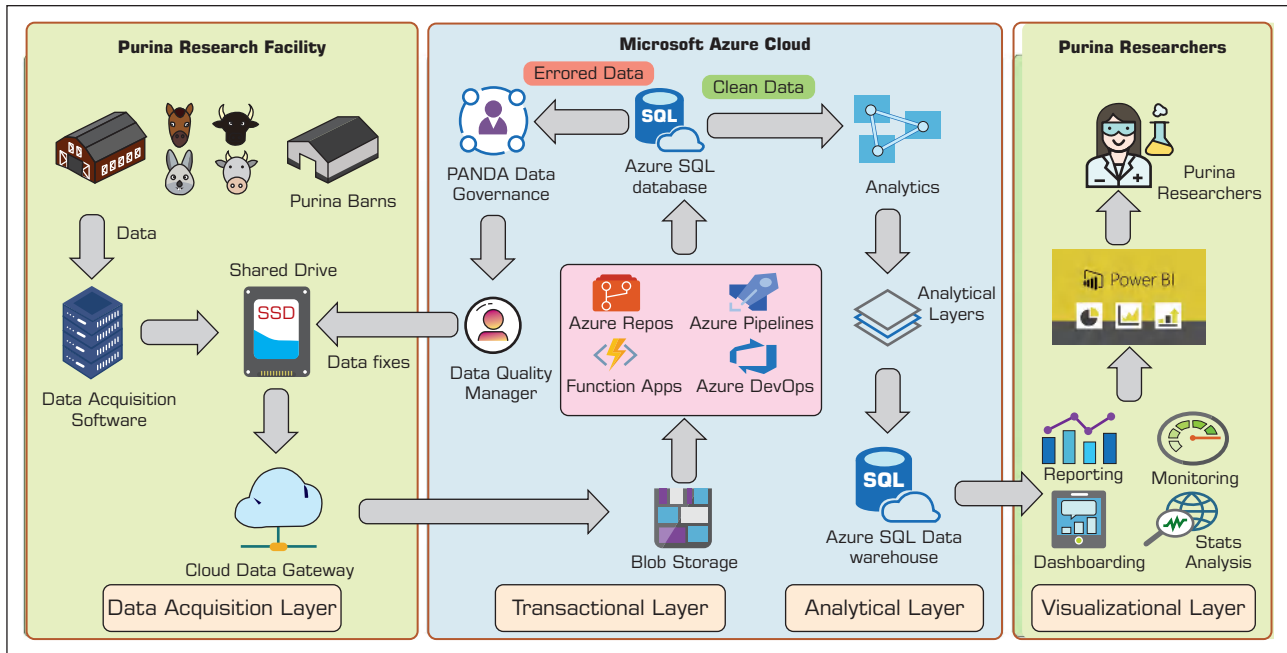


FIGURE 1.22 PANDA Platform's Architecture.

used the terminology of “North Star Metrics” by species. With the help of business and research teams, they identified the key measures (termed north star metrics) using the available attributes. This involved understanding the lowest possible granularity/detail. Working with the life cycle of living creatures requires understanding of not just the major events but also all the small details.

Figure 1.22 summarizes the technical architecture of the PANDA platform. The PANDA Platform has been designed with a layered approach and has four layers:

- Data Acquisition**—All the raw data, spreadsheets, flat files, pdfs, word file summaries, data dumps from feed management systems, raw data from sensors, data from APIs, final summarized data, and field trial data get dumped on the shared drive and later have put it on scheduler to copy the files from shared drive to blob storage on Azure. Thus, they have staged all their raw data on blob storage forming the staging layer.
- Transactional**—Transaction layer is the second layer. In this layer, all the unstructured and structured data brought in the form of database table objects. At some places, complex logic needed to bring them in right format while some raw files got easily converted into the table object. Thus, transactional layer solves the “scattered” challenge of the data and gives the uniformity.
- Analytical**—Analytical layer is the third layer. In this layer, they have applied all the complex logics and created multiple different summarized datasets. They are chasing after the calculations needed for generating the KPIs decided by the business. Statistical Models and Machine Learning models are also applied on some of the datasets. The model output results are also stored in the table for further use in visualization layer.
- Visualization**—Visualization layer is the last layer. In this layer, reports, summaries, dashboards, and statistical analysis summaries are developed for data-based decision making in real time.

The PANDA platform is now live and can be accessed in real time to track the animals. The system provides the summaries that are needed for the business and product development teams.

After successful implementation of project PANDA, they worked on many data science use cases, but we will focus on one just use case here—“**Predicting the final weight of cattle after finishing their first 60 days of feeding.**” At Purina research facility, beef cattle are fed multiple experimental diets to understand any significant performance improvement in their growth and increase in the feed efficiency. It takes almost 140–160 days to conclude if a diet has any statistically significant impact on the performance and feed efficiency. Can this test cycle be shortened? If they can predict the feed efficiency and performance of the cattle after 60 days, then beef research and product development team can test two to three times as many feed products than they are currently able to test.

Livestock producer's profits are generally linked to the weight of their animals. A better prediction of their livestock's weight will allow them to better plan their supply of the meat. Hence, a predictive model is an interesting approach to address the issue of final performance prediction for livestock producers.

Each data science project faces some unique challenges. The following is a partial list of issues the data science team encountered:

- Animal Science is very different compared to other domains. Very few data science folks are also domain experts in animal science. Thus, understanding the animal science terminologies, KPIs, and metrics was a major challenge that Data Intelligence team at Purina faced.
- There were many additional data preprocessing steps needed even after development of the PANDA platform to be able to prepare data in a format suitable for various models. This is not unusual, however, in most applications.
- The team also faced the usual challenge of communicating the results from models to the business users—more advanced machine learning models are hard to explain.

The data science team took the following approach to working on this project. This is an adaptation of the data mining process approaches (CRSP-DM or SEMMA) described in a later chapter:

Hypothesis Development—The team asked a lot of questions to research and product development teams to form multiple hypotheses for this problem. This step also helped them perform extensive data preprocessing steps to form the meaningful datasets.

Exploratory Data Analysis—The Data Intelligence team spent most of the time in this step. They plotted multiple graphs and conducted different statistical tests to explore the relationships between the attributes. They also confirmed the insights they got from the data with researchers.

Feature Engineering—The goal of the first two steps was to better understand the business problem and the data. An outcome of that process was to understand the important features to be used as inputs to the machine learning step.

Advanced Analytical Methods and evaluation—They started with the simplest model and applied advanced/complex models based on the interpretation from the simple models because simple models are highly interpretable. That way they can connect the dots and explain the journey better to the end users.

Recommendations—Using the results from first four steps, Data Intelligence team presented and communicated the insights, hidden patterns, discoveries, predictive models, trees, and the accuracy to product development and researchers.

The Modeling Process and Results

Multiple predictive algorithms were applied and evaluated on the data. (Many of these approaches are described in future chapters.)

- Ridge Regression
- Tree-based algorithm
- Random Forest
- Gradient Boosting Decision Tree
- Support Vector machines

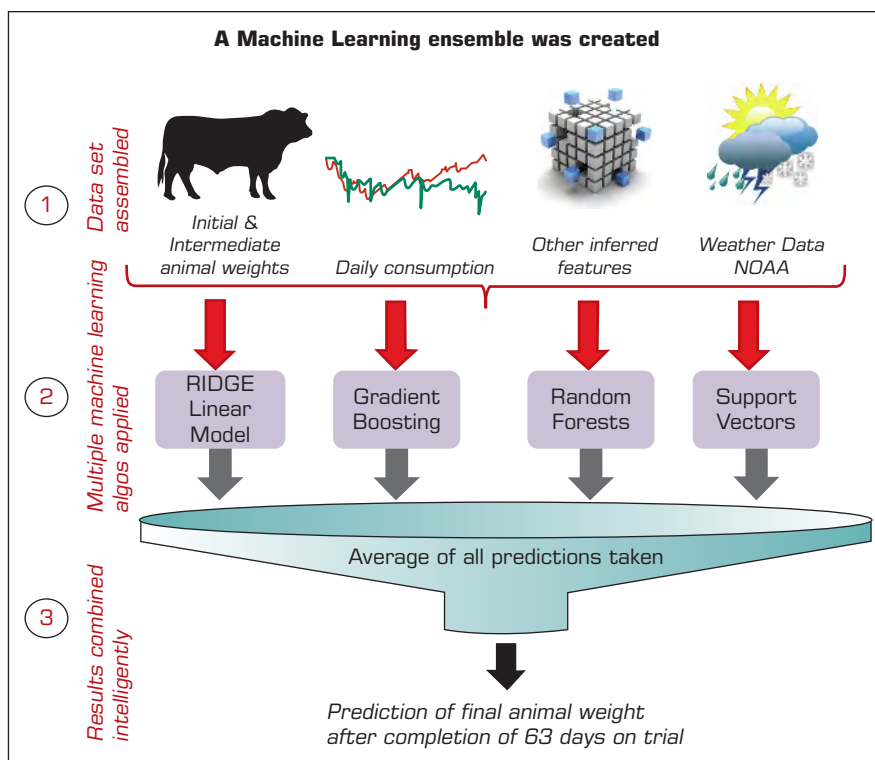


FIGURE 1.23 Analytics Model Development Process.

After multiple iterations of cross validation and model evaluation, no single model emerged as a clear winner. A common approach is to combine the predictions from an ensemble of models as this average of all different predicted weights from multiple models can be taken as the predicted final weight. The final weight of an animal after completion of 60 days on feed was predicted with a Mean Absolute Percentage Error (MAPE) of 2%. This was determined to be an excellent predictive performance. Figure 1.23 provides a graphical view of the process described above.

This project has had a significant impact at Purina:

- It spearheaded the business transformation from a silo-based environment for data analytics to a centralized Microsoft Azure cloud-based approach for compilation of all data with analytics and machine learning solutions built around it.
- It saved the company over \$700k in Capex costs by having a data agnostic platform approach.
- It streamlined the data pipelines and provided the right processes to manage 3 billion data points collected every year. The need for maintenance and support resources reduced by three times with this platform. Also, PANDA enabled proactive support than reactive. Process automation implemented in PANDA reduced the turnaround time for any additional features by five times. It increased the overall efficiency over 10–15 times because now research insights are available at their fingertips and became self-service for researchers and product development team.
- New Product Development pipeline is more efficient now because the cattle team can experiment with the performance of more candidate products, thus increasing the efficiency of the new product development process.
- Livestock producers can better plan their supply based on the final weights from the prediction model to get the maximum profit by selling the animals at the best possible prices.

What can we learn from this story?

- Most organizations still face issues related to data. Creating a data platform to be able to share all internal and most external data is an expensive but crucial step in beginning any data science journey.
- Constant communication between business domain experts and the data science teams is critical to success of any such projects. The team reported the importance of communicating the intermediate results with business users before moving to the next steps.
- Exploratory Data Analysis and Feature Engineering steps are the keys for the model development.
- An analytics platform can significantly impact the speed of product experimentation in the new product development process.

Contributed by Manobar Mohanlal Lahoti, Data Science and Digital Product Manager, Land O'Lakes Inc., Purina Animal Nutrition subsidiary.

specific yes/no decision for a problem, a specific amount (say, price for a specific item or airfare to charge), or a complete set of production plans. The decisions may be presented to a decision maker in a report or may be used directly in an automated decision rules system (e.g., in airline pricing systems). Thus, these types of analytics can also be termed **decision or normative analytics**. Analytics in Action 1.4 gives an example of such prescriptive analytic applications. We will learn about some aspects of prescriptive analytics in Chapter 8.

ANALYTICS IN ACTION 1.4 A Specialty Steel Bar Company Uses Analytics to Determine Available-to-Promise Dates

This application case is based on a project that we worked on a few years ago. A company that does not wish to disclose its name (or even the precise industry) was facing a major problem of making decisions on which inventory of raw materials to use to satisfy which customers. This company supplies custom configured steel bars to its customers. These bars may be cut into specific shapes or sizes and may have unique material and finishing requirements. The company procures raw materials from around the world and stores them in its warehouse. When a prospective customer calls the company to request a quote for the specialty bars meeting specific material requirements (composition, origin of the metal, quality, shapes, sizes, etc.), the salesperson usually has just a little bit of time to submit such a quote including the date when the product can be delivered and, of course, prices, and so on. It must make available-to-promise (ATP) decisions, which determine in real time the dates when it can promise delivery of products that customers requested during the quotation stage. Previously, a salesperson had to make such decisions by analyzing reports on available inventory of raw materials. Some of the available raw material may have already been committed to another customer's order. Thus the inventory in stock may not really be the free inventory available. On the other hand, there may be raw material that is expected to be delivered in the near future that could also be used for satisfying the order from this prospective customer. Finally, there might even be an opportunity to charge a premium for a new order by repurposing previously committed inventory to satisfy this new order while delaying an already committed order. Of course, such decisions should be based on the cost-benefit analyses of delaying a previous order. The system should thus be able to pull real-time data about inventory, committed orders, incoming raw material, production constraints, and so on.

To support these ATP decisions, a real-time DSS was developed to find an optimal assignment of the available inventory and to support additional what-if analysis. The DSS uses a suite of mixed-integer programming models that are solved using commercial software. The company has incorporated the DSS into its enterprise resource planning system to seamlessly facilitate its use of business analytics.

QUESTIONS FOR DISCUSSION

1. Why would reallocation of inventory from one customer to another be a major issue for discussion?
2. How could a DSS help make these decisions?

Source: Pajouh Foad, M., Xing, D., Hariharan, S., Zhou, Y., Balasundaram, B., Liu, T., & Sharda, R. (2013). Available-to-promise in practice: An application of analytics in the specialty steel bar products industry. *Interfaces*, 43(6), 503–517.

Analytics/Data Science/Machine Learning/AI?

Even as the concept of analytics is receiving more attention in industry and academic circles, another term has already been introduced and is becoming popular. The new term is *data science*. Thus, the practitioners of data science are data scientists. D. J. Patil of LinkedIn is sometimes credited with creating the term *data science*. There have been some attempts to describe the differences between data analysts and data scientists. One view is that *data analyst* is just another term for professionals who were doing BI in the form of data compilation, cleaning, reporting, and perhaps some visualization. Their skill sets included Excel, some SQL knowledge, and reporting. You would recognize those capabilities as descriptive or reporting analytics. In contrast, a data scientist is responsible for predictive analysis, statistical analysis, and more advanced analytical tools and algorithms. They may have a deeper knowledge of algorithms and may recognize them under various labels—data mining, knowledge discovery, or machine learning. Thus, machine learning has become a substitute term for predictive analytics in some cases. Some of these professionals may also need deeper programming knowledge to be able to write code for data cleaning/analysis in current Web-oriented languages such as Java or Python and statistical languages such as R. Many analytics professionals also need to build significant expertise in statistical modeling, experimentation, and analysis. Again, our readers should recognize that these fall under the predictive and prescriptive analytics umbrella. However, prescriptive analytics also includes more significant expertise in OR including optimization, simulation, decision analysis, and so on. Those who cover these fields are more likely to be called data scientists than analytics professionals.

Our view is that the distinction between analytics and data scientist is more of a degree of technical knowledge and skill sets than functions. It may also be more of a distinction across disciplines. Computer science, statistics, and applied mathematics programs appear to prefer the data science label, reserving the analytics label for more business-oriented professionals. And the recent industry move is to reuse a term that has been around a while—artificial intelligence as an umbrella term. As another example of this, applied physics professionals have proposed using *network science* as the term for describing analytics that relate to groups of people—social networks, supply chain networks, and so forth. See <http://barabasi.com/networksciencebook/> for an evolving textbook on this topic.

Aside from a clear difference in the skill sets of professionals who only have to do descriptive/reporting analytics versus those who engage in all three types of analytics, the distinction is fuzzy between these labels, at best. We observe that graduates of our analytics programs tend to be responsible for tasks which are more in line with data science professionals (as defined by some circles) than just reporting analytics. This book is clearly aimed at introducing the capabilities and functionality of all analytics (which include data science and AI), not just reporting analytics. From now on, we will use these terms interchangeably.

SECTION 1.6 REVIEW QUESTIONS

1. Define *analytics*.
2. What is descriptive analytics? What are the various tools that are employed in descriptive analytics?
3. How is descriptive analytics different from traditional reporting?
4. What is predictive analytics? How can organizations employ predictive analytics?
5. What is prescriptive analytics? What kinds of problems can be solved by prescriptive analytics?
6. Define modeling from the analytics perspective.
7. Is it a good idea to follow a hierarchy of descriptive and predictive analytics before applying prescriptive analytics?
8. How can analytics aid in objective decision making?

1.7 ANALYTICS EXAMPLES IN SELECTED DOMAINS

You will see examples of analytics applications throughout various chapters. That is one of the primary approaches (exposure) of this book. In this section, we highlight two application areas—healthcare and retail, where there have been the most reported applications and successes.

Analytics Applications in Healthcare—Humana Examples

Although healthcare analytics span a wide variety of applications from prevention to diagnosis to efficient operations and fraud prevention, we focus on some applications that have been developed at a major health insurance company, Humana. According to the company's Web site, "The company's strategy integrates care delivery, the member experience, and clinical and consumer insights to encourage engagement, behavior change, proactive clinical outreach and wellness..." Achieving these strategic goals includes significant investments in information technology in general, and analytics in particular. Dr. Brian LeClaire was the senior vice president and CIO until recently of Humana, a major health insurance provider in the United States. He has a PhD in MIS from Oklahoma State University. He has championed analytics as a competitive differentiator at Humana—including cosponsoring the creation of a center for excellence in analytics. He described the following projects as examples of Humana's analytics initiatives, led by Humana's Chief Clinical Analytics Officer, Vipin Gopal.

Example 1: Preventing Falls in a Senior Population—An Analytic Approach

Accidental falls are a major health risk for adults age 65 years and older with one-third experiencing a fall every year.¹ Falls are also the leading factor for both fatal and nonfatal injuries in older adults, with injurious falls increasing the risk of disability by up to 50%.² The costs of falls pose a significant strain on the U.S. healthcare system, with the direct costs of falls estimated at \$34 billion in 2013 alone. With the percent of seniors in the U.S. population on the rise, falls and associated costs are anticipated to increase. According to the Centers for Disease Control and Prevention (CDC), "Falls are a public health problem that is largely preventable."

Humana is the nation's second-largest provider of Medicare Advantage benefits with approximately 3.2 million members, most of whom are over 65. Keeping their senior members well and helping them live safely at their homes is a key business objective, of which prevention of falls is an important component. However, no rigorous methodology was available to identify individuals most likely to fall, for whom falls prevention efforts

would be beneficial. Unlike chronic medical conditions such as diabetes and cancer, a fall is not a well-defined medical condition. In addition, falls are usually underreported in claims data as physicians typically tend to code the consequence of a fall such as fractures and dislocations. Although many clinically administered assessments to identify fallers exist, they have limited reach and lack sufficient predictive power.³ As such, there is a need for a prospective and accurate method to identify individuals at greatest risk of falling, so that they can be proactively managed for fall prevention. The Humana analytics team undertook the development of a Falls Predictive Model (PM) in this context. It is the first comprehensive PM reported that utilizes administrative medical and pharmacy claims, clinical data, temporal clinical patterns, consumer information, and other data to identify individuals at high risk of falling over a time horizon.

Today, the Falls PM is central to Humana's ability to identify seniors who could benefit from fall mitigation interventions. An initial proof-of-concept with Humana consumers, representing the top 2% of highest risk of falling, demonstrated that the consumers had increased utilization of physical therapy services, indicating consumers are taking active steps to reduce their risk for falls. A second initiative utilizes the Falls PM to identify high-risk individuals for remote monitoring programs. Using the PM, Humana was able to identify 20,000 consumers at a high risk of falls, who benefited from this program. Identified consumers wear a device that detects falls and alerts a 24/7 service for immediate assistance.

This work was recognized by the Analytics Leadership Award by Indiana University Kelly School of Business in 2015, for innovative adoption of analytics in a business environment.

¹<http://www.cdc.gov/homeandrecreationalafety/falls/adultfalls.html>.

²Gill, T. M., Murphy, T. E., Gahbauer, E. A., et al. (2013). Association of injurious falls with disability outcomes and nursing home admissions in community living older persons. *American Journal of Epidemiology*, 178(3), 418–425.

³Gates, S., Smith, L. A., Fisher, J. D., et al. (2008). Systematic review of accuracy of screening instruments for predicting fall risk among independently living older adults. *Journal of Rehabilitation Research and Development*, 45(8), 1105–1116.

Contributors: Harpreet Singh, PhD; Vipin Gopal, PhD; Philip Painter, MD.

Example 2: Predictive Models to Identify the Highest Risk Membership in a Health Insurer

The 80/20 rule generally applies in healthcare, that is, roughly 20% of consumers account for 80% of healthcare resources due to their deteriorating health and chronic conditions. Health insurers like Humana have typically enrolled the highest-risk enrollees in clinical and disease management programs to help manage the chronic conditions the members have.

Identification of the right members is critical for this exercise, and in the recent years, PMs have been developed to identify enrollees with the high future risk. Many of these PMs were developed with heavy reliance on medical claims data, which results from the medical services that the enrollees use. Because of the lag that exists in submitting and processing claims data, there is a corresponding lag in identification of high-risk members for clinical program enrollment. This issue is especially relevant when new members join a health insurer, as they would not have a claims history with an insurer. A claims-based PM could take on average of 9–12 months after enrollment of new members to identify them for referral to clinical programs.

In the early part of this decade, Humana attracted large numbers of new members in its Medicare Advantage products and needed a better way to clinically manage this membership. As such, it became extremely important that a different analytic approach be developed to rapidly and accurately identify high-risk new members for clinical management, to keep this group healthy and costs down.

Humana's Clinical Analytics team developed the New Member Predictive Model (NMPM) that would quickly identify at-risk individuals soon after their new plan enrollments

with Humana, rather than waiting for sufficient claim history to become available for compiling clinical profiles and predicting future health risk. Designed to address the unique challenges associated with new members, NMPM developed a novel approach that leveraged and integrated broader data sets beyond medical claims data such as self-reported health risk assessment data and early indicators from pharmacy data, employed advanced data mining techniques for pattern discovery, and scored every MA consumer daily based on the most recent data Humana has to date. The model was deployed with a cross-functional team of analytics, IT, and operations to ensure seamless operational and business integration.

Ever since NMPM was implemented in January 2013, it has been rapidly identifying high-risk new members for enrollment in Humana's clinical programs. The positive outcomes achieved through this model have been highlighted in multiple senior leader communications from Humana. In the first quarter 2013 earnings release presentation to investors, Bruce Broussard, CEO of Humana, stated the significance of "improvement in new member PMs and clinical assessment processes," which resulted in 31,000 new members enrolled in clinical programs, compared to 4,000 in the same period a year earlier, a 675% increase. In addition to the increased volume of clinical program enrollments, outcome studies showed that the newly enrolled consumers identified by NMPM were also referred to clinical programs sooner, with over 50% of the referrals identified within the first three months after new MA plan enrollments. The consumers identified also participated at a higher rate and had longer tenure in the programs.

Contributors: Sandy Chiu, MS; Vipin Gopal, PhD.

These examples illustrate how an organization explores and implements analytics applications to meet its strategic goals. You will see several other examples of healthcare applications throughout various chapters in the book.

Analytics in the Retail Value Chain

The retail sector is where you would perhaps see the most applications of analytics. This is the domain where the volumes are large but the margins are usually thin. Customers' tastes and preferences change frequently. Physical and online stores face many challenges in succeeding. And market dominance at one time does not guarantee continued success. So investing in learning about your suppliers, customers, employees, and all the stakeholders that enable a retail value chain to succeed and using that information to make better decisions have been a goal of the analytics industry for a long time. Even casual readers of analytics probably know about Amazon's enormous investments in analytics to power their value chain. Similarly, Walmart, Target, and other major retailers have invested millions of dollars in analytics for their supply chains. Most of the analytics technology and service providers have a major presence in retail analytics. Coverage of even a small portion of those applications to achieve our exposure goal could fill a whole book. So this section just highlights a few potential applications. Most of these have been fielded by many retailers and are available through many technology providers, so in this section we will take a more general view rather than point to specific cases. This general view has been proposed by Abhishek Rathi, CEO of vCreaTek.com. vCreaTek, LLC is a boutique analytics software and service company that has offices in India, the United States, the United Arab Emirates (UAE), and Belgium. The company develops applications in multiple domains, but retail analytics is one of their key focus areas.

Figure 1.24 highlights selected components of a retail value chain. It starts with suppliers and concludes with customers, but illustrates many intermediate strategic and operational planning decision points where analytics—descriptive, predictive, or prescriptive—can play a role in making better data-driven decisions. Table 1.1 also illustrates some of the important areas of analytics applications, examples of key questions that can be answered through analytics, and of course, the potential business value derived from fielding such analytics. Some examples are discussed next.

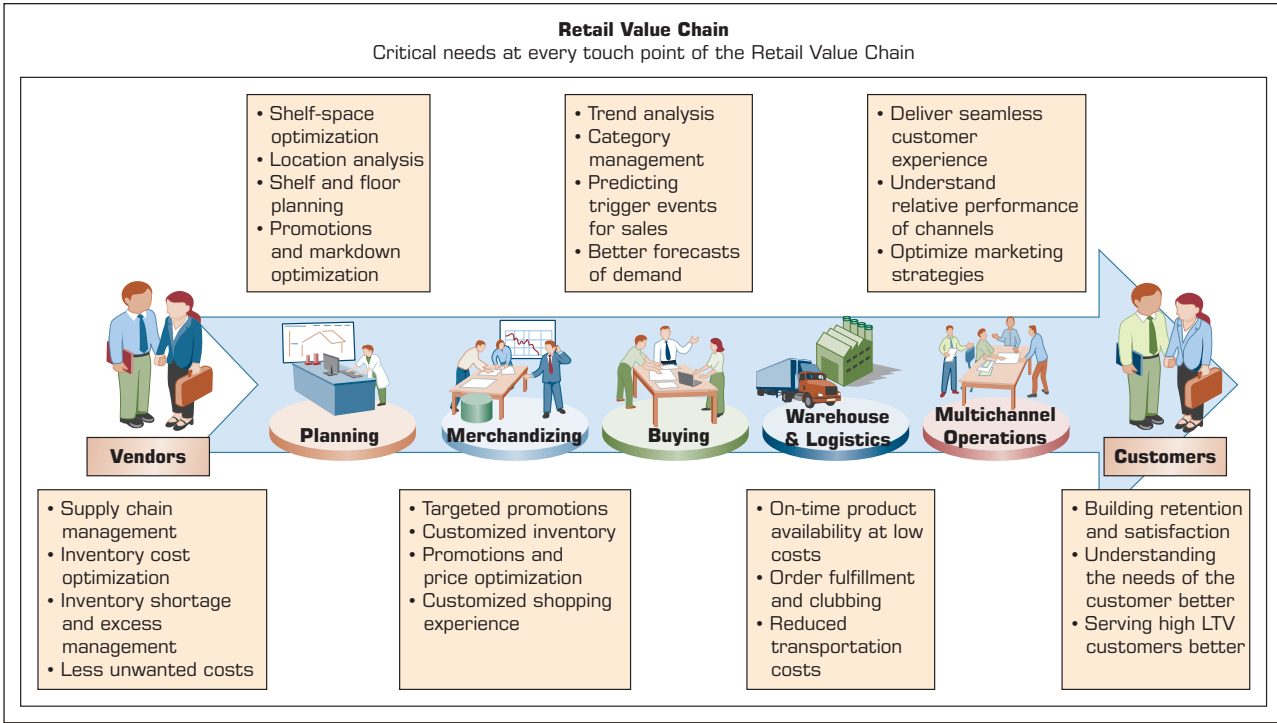


FIGURE 1.24 Example of Analytics Applications in a Retail Value Chain. Contributed by Abhishek Rathi, CEO, vCreaTek.com.

An online retail site usually knows its customer as soon as the customer signs in, and thus they can offer customized pages/offerings to enhance the experience. For any retail store, knowing its customer at the store entrance is still a huge challenge. By combining the video analytics and information/badge issued through their loyalty program, the store may be able to identify the customer at the entrance itself and thus enable an extra opportunity for a cross-sell or up-sell. Moreover, a personalized shopping experience can be provided with more customized engagement during the customer's time in the store.

Store retailers invest lots of money in attractive window displays, promotional events, customized graphics, store decorations, printed ads, and banners. They may also update the lighting and the store layout to make it inviting for disabled customers. To discern the effectiveness of these marketing methods, the team can use shopper analytics by observing closed-circuit television (CCTV) images to figure out the demographic details of the in-store foot traffic. The CCTV images can be analyzed using advanced algorithms to derive demographic details such as age, gender, and mood of the person browsing through the store.

Further, the customer's in-store movement data when combined with shelf layout and planogram can give more insight to the store manager to identify the hot-selling/profitable areas within the store. Moreover, the store manager can use this information to also plan the workforce allocation for those areas for peak periods.

Market basket analysis has commonly been used by the category managers to push the sale of the slowly moving Stock Keeping Units (SKU). By using advanced analytics of data available, the product affinity can be done at the lowest level of SKU to drive better ROIs on the bundle offers. Moreover, by using price elasticity techniques, the markdown or optimum price of the bundle offer can also be deduced, thus reducing any loss in the profit margin.

Thus, by using data analytics, a retailer can not only get information on its current operations but can also get further insight to increase the revenue and decrease the operational cost for higher profit. A fairly comprehensive list of current and potential retail

TABLE 1.1 Examples of Analytics Applications in the Retail Value Chain

Analytic Application	Business Question	Business Value
Inventory Optimization	1. Which products have high demand? 2. Which products are slow moving or becoming obsolete?	1. Forecast the consumption of fast-moving products and order them with sufficient inventory to avoid a stock-out scenario. 2. Perform fast inventory turnover of slow-moving products by combining them with one in high demand.
Price Elasticity	1. How much net margin do I have on the product? 2. How much discount can I give on this product?	1. Markdown prices for each product can be optimized to reduce the margin dollar loss. 2. Optimized price for the bundle of products is identified to save the margin dollar.
Market Basket Analysis	1. What products should I combine to create a bundle offer? 2. Should I combine products based on slow-moving and fast-moving characteristics? 3. Should I create a bundle from the same category or different category line?	1. The affinity analysis identifies the hidden correlations between the products, which can help in following values: a) Strategize the product bundle offering based on focus on inventory or margin. b) Increase cross-sell or up-sell by creating bundle from different categories or the same categories, respectively.
Shopper Insight	1. Which customer is buying what product at what location?	1. By customer segmentation, the business owner can create personalized offers resulting in better customer experience and retention of the customer.
Customer Churn Analysis	1. Who are the customers who will not return? 2. How much business will I lose? 3. How can I retain them? 4. What demography of customer is my loyal customer?	1. Businesses can identify the customer and product relationships that are not working and show high churn. Thus can have better focus on product quality and reason for that churn. 2. Based on the customer lifetime value (LTV), the business can do targeted marketing resulting in retention of the customer.
Channel Analysis	1. Which channel has lower customer acquisition cost? 2. Which channel has better customer retention? 3. Which channel is more profitable?	1. Marketing budget can be optimized based on insight for better return on investment.
New Store Analysis	1. What location should I open? 2. What and how much opening inventory should I keep?	1. Best practices of other locations and channels can be used to get a jump-start. 2. Comparison with competitor data can help to create a differentiator/USP factor to attract the new customers.
Store Layout	1. How should I do store layout for better topline? 2. How can I increase my in-store customer experience?	1. Understand the association of products to decide store layout and better alignment with customer needs. 2. Workforce deployment can be planned for better customer interactivity and thus satisfying customer experience.
Video Analytics	1. What demography is entering the store during the peak period of sales? 2. How can I identify a customer with high LTV at the store entrance so that a better personalized experience can be provided to this customer?	1. In-store promotions and events can be planned based on the demography of incoming traffic. 2. Targeted customer engagement and instant discount enhances the customer experience resulting in higher retention.

analytics applications that a major retailer such as Amazon could use is proposed by a blogger at analytikus. That list is available at <https://www.analytikus.com/post/2016/12/30/21-data-science-systems-used-by-amazon-to-operate-its-business>. As noted earlier, there are too many examples of these opportunities to list here, but you will see many examples of such applications throughout the book.

Gaming Industry Applications

Besides the sports industry analytics examples (starting with *Moneyball*) covered in the opening vignette, perhaps another early example of analytics that is commonly mentioned is in the casino industry. Gary Loveman was a faculty member in service management at Harvard Business School. His work in customer loyalty management led him to eventually working as the Chief Operating Officer of Harrah's Entertainment (which then emerged as Caesars Entertainment). Harrah's was the first major company in the gaming industry to establish a customer loyalty program, called Total Rewards program. The program gathered data on casino visitors—their travel, spending in casino, and other places, etc. Analytics of this data revealed that the repeat slot players were their most profitable customers, not the high rollers. This insight led to the company working to recognize and reward this customer segment. Loveman also added a rewards program for employees that tied the customer satisfaction to their bonuses and incentives. Many stories have been written on this transition of gaming industry to a highly data-driven industry. For example, see this article by Loveman (2003), and many industry stories on his successful career in the gaming industry. This is a tremendous example of analytics expertise leading to top leadership of an organization.

To summarize such loyalty programs, the key is to recognize a customer's relationship with your company. At the beginning stages, for example, a customer's revenue contribution to your company may be rather limited. As that relationship grows, that revenue increases. At some point, the customer begins to move away from your organization. By recognizing such a stage, it is then important for you to reenergize that relationship. By compiling data on the customers and by analyzing it, a company can not only attempt to recognize a customer's value, relationship stage, etc., but can also make customized offers that will keep that customer engaged with you. It all begins with a customer having a card/id/app that recognizes each customer. As Loveman described in his HBR paper (Loveman 2003), by recognizing a customer through the card reader, the machine pages a host to make sure this customer is recognized by name, offered appropriate incentives/drinks, etc. to let them feel special and stay at our casino rather than go somewhere else. This type of reward mechanism increases a casino's overall income. Recall the analysis had already shown that such repeat slot machine customers were their most profitable segment.

There are many companies now offering solutions to gaming industry for collecting and analyzing data using various data science and AI approaches. These applications cover a wide range of opportunities for employing analytics in these operations. Some examples include:

- Analyze slot floor performance
- Visualize to determine how location affects performance
- Measure and track slot ROI
- Optical bet recognition and table game analysis
- Track every bet, monitor every game, and rate every player
- Identifying addictive behavior and problem gamblers
- Identify money-laundering customers
- Fraud detection

Analytics in Action 1.5 highlights a specific casino operator's experience in using data science for understanding and improving their operations.

ANALYTICS IN ACTION 1.5 Gulfstream Park Casino Employs AI to make Optimal Decisions on Slot Machine Placement

Gulfstream Park Casino near Fort Lauderdale, FL, also includes sports. The slot area is somewhat unique in design, split between two floors. Due to the pandemic, the casino floor's layout changed for social distancing and a more spacious experience. The slot area features approximately 600 machines.

Lisa Siples, Director of Slots at Gulfstream, was challenged with deciding where to best position slot machines to maximize play—this led to countless hours of analyzing data, utilizing Tableau and Excel, and studying reports. The process was labor-intensive; it meant physically moving machines around, analyzing machine performance, and some guesswork—it was a constant cycle of trial and error. Sometimes, the necessary information proved elusive or took days or weeks to uncover. After analyzing multiple gaming analytics platforms, Gulfstream chose Gaminganalytics.ai. Below is a brief description of the Gaminganalytics.ai's platform:

Gaming Analytics Enterprise Intelligence Platform

In almost all the business scenarios, data are collected in multiple locations. Before any analytics can be performed, it is critical to develop a combined data model. The Gaming Analytics Enterprise Intelligence Platform (GA-EAIP) was built ground up to leverage this “connectedness of data” to seamlessly integrate data from multiple sources to derive relevant, contextual and actionable insights and make it available to all stakeholders of an enterprise through a simplified, natural and easy to access interface.

Starting with data within an enterprise, new, dynamic data points are discovered every day which needs to be considered to make optimal decisions. The underlying data model of a good enterprise should seamlessly mimic how data is connected, accessed, and consumed in the real world. The Graph Data Model is a natural fit for this purpose and is very good for doing complex traversals across large landscapes of data. New Insights are discovered by traversing through the edges that connect nodes.

Graphs represent data via relationships. Nodes represent an entity in the graph. Relationships (edges/links) connect nodes to each other. Properties describe a node or relationships. A **knowledge graph** is an insight layer of interconnected data enriched with semantics, so you can reason with the underlying data and use it for complex decision making. GA-EAIP is an early adopter of graph technology to build a robust, extensible platform that caters to this fast-evolving landscape.

Today, almost all human–computer interactions start with a Search (similar to “Google”). When anyone can type any question and get answers, the team at Gaming Analytics were intrigued by this simplistic idea of information access and made “Search” as the primary interface for all business users to know everything about their Casino's day-to-day operations.

Figure 1.25 illustrates the architecture of Gaming Analytics' platform.

The architecture of GA-EAIP can be broken down into as two important layers, data ingestion and information access. The platform is hosted in the cloud. It can be accessed through a Web interface and also can be integrated as “PAAS” into any host system. The platform is also available for On-Premise deployment.

The data ingestion layer accepts data from source systems. The data-schema and protocol is kept very simple to integrate with different enterprise gaming systems providers in the market (Bally, Konami, IGT, Aristocrat, etc.). Systems periodically push data into a secure shared location or in real time. Before data is ingested, the meta-data to identify facts/dimensions and their types are captured and stored in a meta-graph. These are further passed down to the platform's natural language processing (NLP) engine for training the model vocabulary (entities, sentiments, context, and intents). The meta-graph gets updated for every new data-point that gets integrated with the GA-EAIP. Storing and linking dimensions with their values in the meta-graph helps infer context of ambiguous questions by suggesting possible matches, allowing the user to narrow down to the exact intent, making the search more intuitive.

GA-EAIP leverages native graph and graph data science features heavily to do deep exploration of data. These discoveries are fed back to further enrich the knowledge graph. Scheduled supervised AI/ML algorithms run on top of the knowledge graph to churn predictions, recommendations, and forecasts.

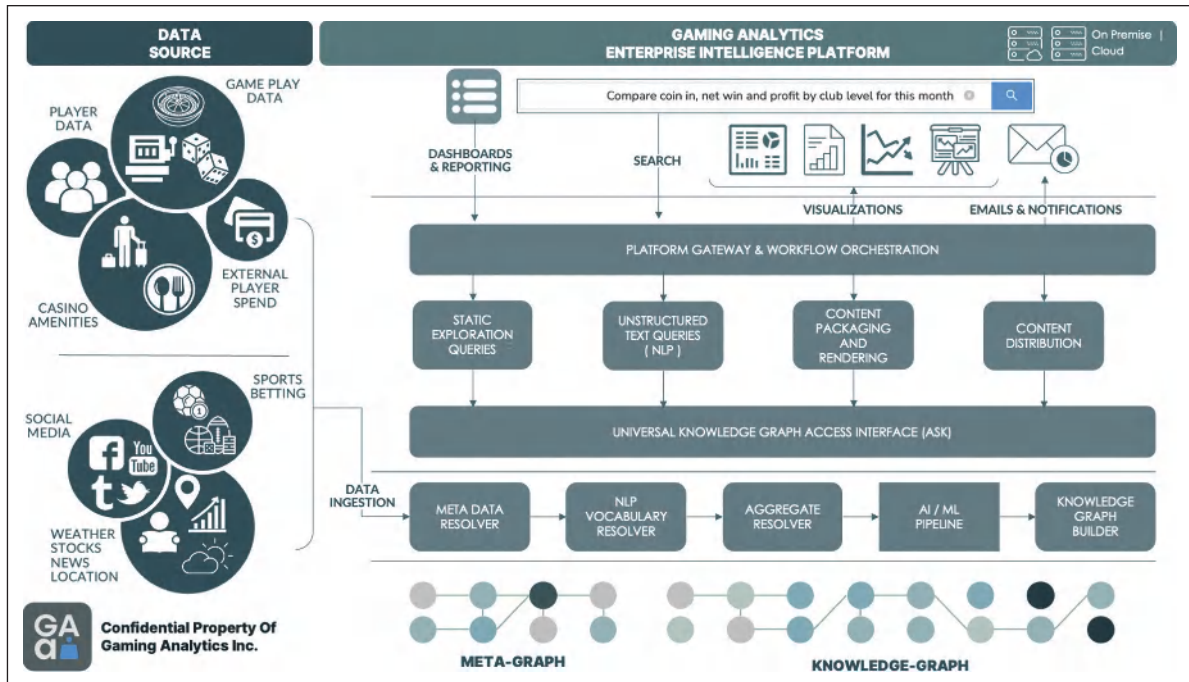


FIGURE 1.25 System Architecture of GamingAnalytics.ai's Platform.

At this point, the knowledge graph is fully equipped to serve exploratory queries, comparisons, insights, and also provide advanced recommendations.

Life Cycle of a User Request

While the knowledge graph continuously evolves, there should also be a dynamic mechanism that has the ability to interpret and map new user questions to pull the needed information. Traditional relational models won't scale since the schema and the data access layer (DAL) need periodic updates. To overcome this, GA-EAIP uses a language agnostic abstraction called "Ask," a DSL (Domain Specific Language) object, to interpret the unstructured text and generate native graph queries. This allows the knowledge graph and the search interface (NLP) to stay loosely coupled and evolve constantly. The query finally is executed against the database to fetch the required result.

The final step is to present the result to the user. Visualizations are the de-facto standard for rendering analytical information. The content packaging layer of the GA-EAIP also leverages the meta-graph to derive the best and optimal visualizations for the result. The platform provides a lot of tools for the end user to customize visualizations and create their specific, preferred story boards. The usage and customizations of these visualizations are fed back into the content packaging layer to learn (get trained) and recommend the most useful and relevant visualizations for a given context in future.

Figure 1.26 presents an example of the user typed query into Gaming Analytics, platform and the resultant query. Hundreds of related queries can be processed by their system. Examples may include the following:

- Which machines (no, location, etc.) performed in the top 10 this week?
- Which manufacturer's machines performed the best this month (top 10)?
- Which games performed the best this week?
- Who were the best players last week in terms of their revenue contribution to us?

Due to the unique design of Gulfstream's slot area—split between two floors—Siples was challenged with knowing which machines perform best on which floor. This led to manually moving machines and testing locations; the process was part guesswork and part data analysis. The data analysis was dependent on running the right report, which sometimes took days.

Best Performing Games Last 60 Days 📊 🔍

Apr 17, 2022 - Jun 15, 2022 | Rows: 100/100

Acid tags...

Game	Net Win PUPD	Net Win xHa	GrossWin PUPD	GrossWin xHa	Theo Win PUPD	Theo Win xHa	Coin In PUPD	Coin In xHa	Handle Puts PUPD	Handle Puts xHa	DOF	Theo Win
1 DRAGON LINK CASH-PANDA MAGIC	\$3,822.69	11.61	\$4,832.92	16.96	\$3,232.17	7.89	\$35,132.21	6.86	3,779	1.89	120	\$387,900
2 DRAGON LINK CASH-AUTUMN MOON	\$2,814.58	7.04	\$3,184.60	7.22	\$2,000.97	5.38	\$23,923.82	4.67	2,672	1.34	77	\$189,475
3 LIGHTNING CASH - MOON RACE	\$2,127.49	6.46	\$2,603.60	6.36	\$2,262.60	5.81	\$26,473.44	5.17	1,963	0.98	45	\$107,217
4 DRAGON LINK CASH-GOLDEN CENTURY	\$1,749.76	5.31	\$2,538.41	5.76	\$2,605.49	6.35	\$28,320.57	5.53	3,116	1.56	120	\$112,968
5 LIGHTNING CASH - TIKI RRE	\$1,452.55	4.41	\$1,763.64	3.86	\$916.30	1.99	\$9,069.97	1.77	833	0.42	77	\$62,856
6 DRAGON LINK CASH-HAPPY AND PROSPEROUS	\$1,189.16	3.61	\$1,863.88	4.23	\$2,171.61	5.29	\$23,604.48	4.61	2,778	1.38	166	\$168,316
7 BUFFALO LINK	\$1,131.72	3.44	\$1,610	3.67	\$1,187.08	2.89	\$18,095.69	3.53	3,619	1.81	180	\$113,674
8 DRAGON LINK-PANDA MAGIC	\$1,079.18	3.27	\$1,482.60	3.18	\$1,257.45	3.06	\$11,332.07	2.21	5,810	2.86	137	\$172,270
											13,358	\$8,577,420

Related Questions

🔍 best performing games last week | 🔍 best performing games | 🔍 best games last saturday | 🔍 top 20 performing game titles | 🔍 best games last sunday

FIGURE 1.26 Search Result from the Query—Which Games Performed the Best in Last 60 Days?

Gaming Analytics helped optimize slot floor design by:

- Identifying best machine by a vendor.
- Determining which machines perform best, broken down by player behavior and exact location.
- Improving machine section.
- Expediting machine performance review, leading to faster and more impactful floor changes.
- Saving labor and saving time by eliminating guesswork and unnecessary moving of machines.

The efficiencies resulting from the Gaming Analytics platform have boosted casino profits. The slots budget has improved as a result of Gaming Analytics. With the savings, funds can be reallocated to improve the guest experience and employee training. Lisa Siples also notes that they have saved over 2500 person-hours by not moving the machines to run their experiments. With payroll savings, the budget has been reallocated to incorporate specialized staff training, make employee incentives/merit rewards more robust, and invest in tools to increase efficiency on the slot floor. Chapter 11 will further introduce the labor impact of incorporating such analytics platforms in the casinos' operations. Thus, the benefits of employing this AI platform include improved slot floor design, greater efficiency, reports that take seconds instead of days or weeks, improved employee relationships, and more targeted marketing campaigns resulting in greater ROI.

What We Can Learn from This Analytics in Action?

There are several takeaways from this application case. First, we note that the casino industry was an early adopter of analytics technologies. Indeed, an early proponent of using such loyalty programs and the analytics moved from academia to become the CEO of a large multibillion-dollar enterprise. Second, newer technological platforms continue to spring up and offer new and unique solutions. Gaminganalytics.ai's platform employs a graph-database approach to building the internal data model of a data warehouse for the gaming industry. This approach gives their platform excellent versatility and adaptability to new situations. The application story also points out the benefits of being able to retrieve results of analytics using a simple natural language-based query system. It avoids a user having to learn any additional commands or navigation of multiple menus, etc. Finally, the story also suggests that the benefits of employing can be very significant for an organization.

Sources: Loveman, G. (2003). Diamonds in the data mine. *Harvard Business Review*, 81(5), 109–113. Available at <https://hbr.org/2003/05/diamonds-in-the-data-mine> (accessed June 2022); "Gulfstream boosts profits and enables better decisions with casino AI," https://gaminganalytics.ai/post_news/case-study-casino-ai-gulfstream-gaming-analytics/ (accessed June 2022). Figures provided by gaminganalytics.ai.

COVID-19 Analytics Applications

The COVID-19 pandemic has continued to ravage the health of communities and economies around the world. Although things seem to be a bit more stable as of this writing, new variants of COVID-19 vaccine may change the situation at any time. To manage a crisis and make effective decisions throughout the lifecycle of a disaster, information sharing and consumption play a critical role in information monitoring. In the current COVID-19 pandemic, information has been extensively utilized for critical tasks such as issuing public alerts, sharing case data, contact tracing, and planning for facilities and supplies needed to manage medical and other public health resources. Accurate and timely information sharing and consumption can help first responders and policy makers in making informed choices and effective policies and operational decisions during the pandemic.

Perhaps the first known application of analytics in public health relates to identifying a source of cholera epidemic in London in 1854. Reportedly, Dr. John Snow collected and plotted data (addresses) on a number of cholera deaths. His maps then pointed to a common source—a public water well. National Geographic provides a resource for the originally compiled data and producing a more modern map to identify that source of cholera epidemic (<https://www.nationalgeographic.org/activity/mapping-cholera-epidemic-1854/>).

COVID-19 pandemic wreaked havoc around the world in many ways. The impact on human life, economies, politics, social structures has been horrific. However, it provides an interesting example of the need for timely and accurate data collection, and performing analytics on the data to be able to inform decision makers on public policy and more. The pandemic continues to also illustrate how one can “spin” the analytics on the same data to help promote one’s political and personal beliefs.

The pandemic also provided a major opportunity for journals in virtually every discipline to identify and promote research in their domain as it relates to the pandemic. Thousands of papers were published or continue to be published that study the impact of the pandemic in their discipline or the potential contributions of their field to managing the pandemic. Not surprisingly, medicine and health were the primary opportunities in this space. Analytics/data science/AI were closely behind, because so much data have been collected and compiled. The US Center for Disease Control (CDC) was primarily responsible for compiling and reporting US data. Similar organizations around the world worked on their respective data collection and analysis. Much of this data was then also linked to the World Health Organization (WHO). The reader is encouraged to explore these data resources:

COVID Data Tracker: <https://covid.cdc.gov/covid-data-tracker/#datatracker-home>

Coronavirus Worldwide Data: <https://ourworldindata.org/coronavirus-source-data>

World Health Organization: <https://covid19.who.int/>

There are hundreds of data portals that provide access to data on COVID-19 from their own area, and in many cases also link to the national and international sources. In many cases, these data sources have also provided some capabilities to let the user visualize the data in many ways. For example, most US COVID-19 data sites present the number of cases, 7-day average rate, positivity rate (number of positive cases/number of tests performed), hospitalization counts, ICU bed-usage counts, death counts, etc., by neighborhood, community, state, and the country. Thus, the pandemic brought some ideas of descriptive analytics to the masses. In addition, a few early modelers began analyzing the data and started providing daily, weekly, monthly forecasts of the spread of the virus. These researchers at the University of Washington (<https://www.washington.edu/coronavirus/>) and the Johns Hopkins University (<https://coronavirus.jhu.edu/>) were

perhaps the earliest popular users of descriptive and predictive analytics. Many researchers then took these data resources and extend this analysis to prescriptive analytics by considering these forecasts and analyzed impacts of various mitigation strategies such as the lockdowns, limited lockdowns, mask mandates, vaccination requirements, testing mandates, etc.

With thousands of publications in COVID-19 analytics, it will take a book by itself to summarize the data science/AI publications in this space. However, a recent paper (Galetsi et al., 2022) reported a summary and classification of a sampling of the analytics/AI papers on COVID-19. They reviewed about 600 papers from a number of journals and identified the following major categories of analytics/AI papers:

1. Methods for detecting COVID-19 positive patients using various measures. The most common result of this work led to development of rapid antigen tests and PCR tests. Another application of analytics in this area is analysis of sewage data to predict early onset of COVID-19 positivity in a community even before the standard medical tests confirm it.
2. Methods/systems for monitoring current levels of infection and prediction of future spread by applying forecasting models that broadly make up predictive analytics. Early attempts also included prediction through other indirect indicators through GPS tracking of the movement of infected individuals and their interactions with other people. Also includes examples of analysis of Google searches of a topic in an area.
3. Methods/models for predicting mortality risk among COVID-19 infected individuals.
4. Many papers employed prescriptive analytics to assist with patient management in hospitals when the patient demand far outstripped the available supply of hospital beds, particularly ICU.
5. Analytical studies of existing medicine and their potential value in treating COVID-19 disease or its symptoms.
6. A category of studies also focused on improving the false negative or false positive rates of other detection methods.
7. Many studies also develop models to detect and prevent spread of misinformation regarding COVID-19 through social media analytics and many other approaches.
8. As mentioned earlier, a group of studies analyzed impacts of lockdown measures on individuals and societies.

The above categories present just illustrative examples of analytics application in COVID-19. As noted before, thousands of papers have been published in this area. Interested readers can find much information through an online search. One Web site that curates a very small sampling of papers/projects employing data science for COVID-19 is at <https://www.covid-datascience.com/>.

The following provides a brief example of the type of analysis that was commonly performed using various available data to persuade the general public to understand the impact of measures to prevent further spread of COVID-19.

Mask Mandates and COVID-19 Spread: Evidence from within State Variation

The efficacy of mask mandates in reducing the spread of COVID-19 was a hotly contested topic. This research provides new evidence by studying a US state where mask mandates vary by municipality, allowing us to compare the growth in reported COVID-19 cases for a group subject to mask mandates versus other communities. Our estimates suggest that the mean daily case growth drops 1.5 to 2.9 percentage points more following mask mandates for cities subject to the policy relative to non-mandate cities.

Problem

The use of masks and the implementation of mask mandates have been proposed as a possible method to slow the spread of the virus (Greenhalgh, Schmid, Czypionka, Bassler, & Gruer, 2020); however, the efficacy of mask mandates is contested, as mandates are hotly debated in public discourse and the academic literature has produced conflicting evidence. Multiple studies find evidence that COVID-19 growth rates dropped by 25% to 40% in areas that enacted mask mandates (Karaivanov, Lu, Shigeoka, Chen, & Pamplona, 2021). Further, some predictive models suggest that national mandates could have led to 19–47% less deaths in the early stages of the pandemic (Chernozhukov, Kasahara, & Schrimpf, 2020). The above is a representative list of studies rather than a comprehensive review, but evidence is gathering in support of mask mandates. Conversely, some studies (e.g., Bundgaard et al., 2020) find no significant difference in infection rates among mask wearers and non-wearers. This research studied community-level data from Oklahoma, where mask mandate policies vary by municipality within the state.

Data

We retrieved data on daily confirmed COVID-19 cases from the Oklahoma State Department of Health with data current as of October 2020. Population data for each city is retrieved from publicly available census data (Advameg, 2020). We follow prior work (Lauer et al., 2020) that defines the incubation period for COVID-19 as 14 days. This time period is a conservative estimate based on data and guidelines from the World Health Organization and represents the most extreme value for incubation period. In general, 14 days would likely represent a generation interval period from exposure through incubation and into an infectious period to themselves and posing a risk of transmission to others. Given this, we define total cases on day i as the total cases over the previous 14-day period.

We employed a difference-in-differences (DiD) (Abadie, 2005) approach to estimate the relation between mask mandates and reported COVID-19 cases. Oklahoma did not implement a mask mandate, but some cities in the state did. Therefore, the DiD method is a useful tool for this setting because we use it to compare changes in cases surrounding mask mandates for the treatment group (cities that enacted a mask mandate) with changes in cases for a group that did not receive the treatment (cities that did not enact a mask mandate). This approach is designed to isolate the treatment effect, that is, the effect of mask mandates on the spread of COVID-19. We estimate the DiD model as follows:

$$y = \beta_0 + \beta_{1p} + \beta_{2p} + \beta_{3p} * d + \varepsilon \quad (1)$$

where y is the daily percentage change in reported 14-day cumulated cases for each city over the sample period of May 1, 2020 to September 30, 2020, p is an indicator variable equal to one if the observation occurred after the city had mandated masks and zero otherwise, d is an indicator variable equal to one if the city had a mask mandate and zero otherwise, and ε represents the error term. The key coefficient is β_3 , which estimates the effect of mask mandates on the spread of COVID-19.

In our initial analysis, the control group, or untreated group, included all cities in the state that did not enact a mask mandate. In order to match the control group cities to the cities with mask mandates, we need to assign an event date or dates. Doing so allows us to compare the growth in reported COVID-19 cases before and after mask mandates for the cities that enacted the mandates to the growth in cases surrounding the policy date for cities that did not. The policy date we choose for non-mandated cities is July 17, 2020. We chose this date because the two most populous cities in the state, Oklahoma City and Tulsa, enacted mask mandates on July 16 and July 17, respectively. These two cities represent the vast majority of the population in Oklahoma that was subject to a local mask mandate (about 1.2 million of the 1.6 million people subject to mandates).

Results

Accurate identification of the effects of mask mandates requires that trends in the growth of COVID-19 were similar leading up to the policy enactment for both the mandated and non-mandated groups of cities. Figure 1.27 shows plots of reported cases for the two groups. Panel A plots total cases over time for both groups. The trends leading up to the mandate effective date (day 0) are quite similar, suggesting that the control group is a reasonable benchmark for the treatment group. We also plot cases per capita for each group in Panel B. As in Panel A, the cases per capita rapidly increase prior to day 0 for both mandated and non-mandated cities, though the pre-mandate rise is sharper for the mandated cities (there are differences between Panel A and Panel B because the control group has more people, about 2.4 million compared to about 1.6 million for the treatment group), which raises concerns about a violation of parallel trends in the pre-policy time period. In our regression analysis, we perform robustness tests where we control for past growth in cases in an attempt to address concerns that mask mandates are endogenously timed following a rise in cases, and these results are similar to those for the main model that does not include controls. Additionally, we are unaware of any other systematic events or interventions that occurred around the same time as the mask mandates that may have impacted case growth. Nonetheless, we recognize that it is difficult to completely rule out alternative stories, and we cannot definitively claim causality in the analysis that follows.

Figure 1.27 presents time-series plots of reported COVID-19 cases in Oklahoma. We plot cases for the group of cities with mask mandates and those without. Day 0 represents the mask mandate effective date; for the non-mandated cities, day 0 is July 17, 2020, which is the date by which the vast majority of the mandated population enacted the mask mandates. Panel A presents total daily cases, and Panel B presents cases per capita. The plots also illustrate the relation between mask mandates and the subsequent growth in reported COVID-19 cases. Regardless of whether cases are in per-capita terms or not, we see a notable drop in reported cases following mask mandates (day 0) for the mandated cities relative to the non-mandated cities. We also note that cases begin to

Panel A. Total daily reported COVID-19 cases

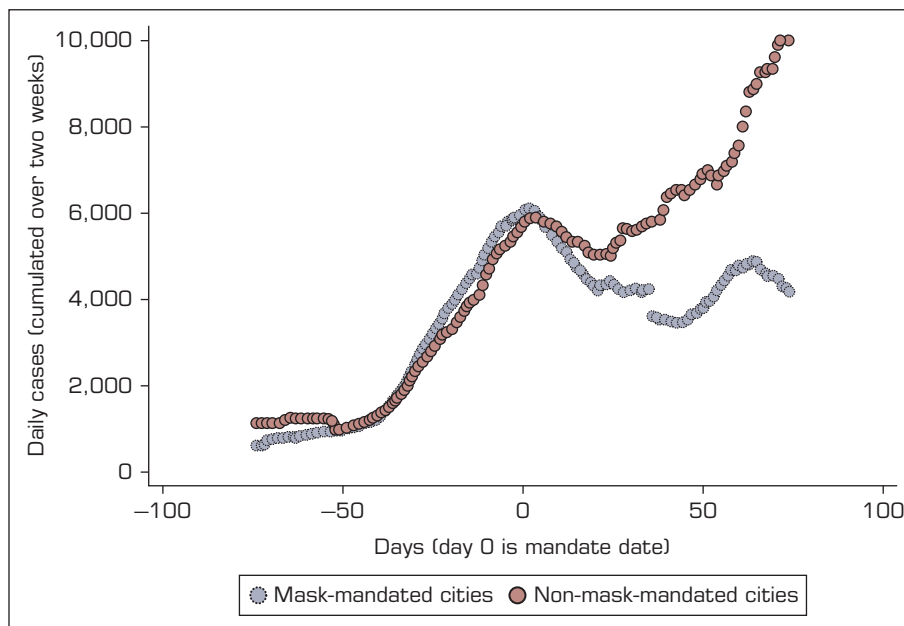


FIGURE 1.27 Time Series of Reported COVID-19 Cases in Oklahoma.

Panel B. Daily reported COVID-19 cases per capita

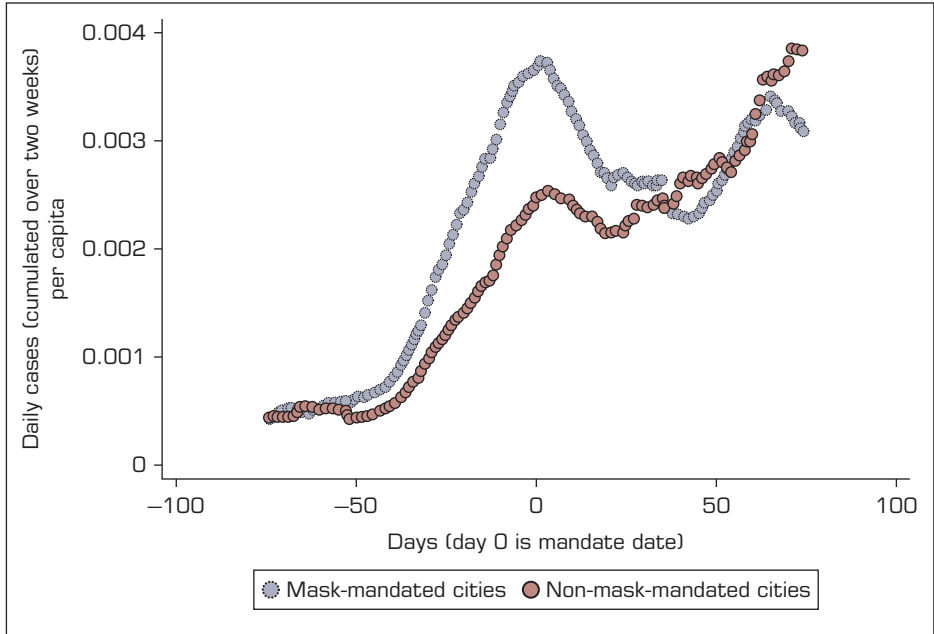


FIGURE 1.27 Time Series of Reported COVID-19 Cases in Oklahoma.
Reported COVID-19 cases in Oklahoma. (Panel A—Total daily cases; Panel B—Cases per capita).

rise again one to two months following the mandates. This period coincides with schools and universities reopening, which could cloud the mask mandate effects. We therefore conduct robustness tests for our regression analysis where we consider shorter windows, such as one month following the mandates, and we obtain qualitatively similar results.

We next present regression results of our DiD model. These results are reported in Table 1.2, which reports estimated slope coefficients and associated *t*-statistics, in parentheses, that are computed using heteroskedastic-robust standard errors that are clustered by city. We present results for models with no fixed effects, a model with date fixed effects, and a model with both date and city fixed effects. In some specifications, we augment the model with a control for past case growth. We also report results for an Ordinary Least Squares (OLS) model and a Weighted Least Squares (WLS) model that weights observations by population. To mitigate the effect of outliers, we winsorize (Hastings, Mosteller, Tukey, & Winsor, 1947) the daily percentage change in 14-day cumulated cases at the one percent tails.

This table presents alternative estimates of equation (1), which represents a DiD model. We present estimated slope coefficients and associated *t*-statistics, in parentheses, that are computed using heteroskedastic-robust standard errors that are clustered by city. We present estimates using OLS (M1, M3, M4, and M5) and WLS (M2), where the weights are each city's population. M3, M4, and M5 include past case growth as a control variable.

The regression results confirm that there is a significant, negative relation between mask mandates and subsequent growth in reported cases. The key variable in the regression model is the interaction term, which captures how the mask mandate impacts the growth in reported cases for mandated cities relative to the growth in cases over the same time period for non-mandated cities. The estimated coefficient on this variable is significantly negative in all of the specifications, suggesting that the growth in cases drops substantially more following the mandate policy for mandated cities relative to non-mandated cities. Not only is this effect statistically significant, but it is also practically significant, as the mean daily drop in case growth for mandated cities ranges from 1.5% to 2.9%.

TABLE 1.2 Difference-in-Differences Analysis of Mask Mandates in Oklahoma

	M1	M2	M3	M4	M5
	OLS	WLS	OLS	OLS	OLS
y = daily case cumulated in the past 14 days					
Mandate indicator	0.023 (7.16)	0.018 (5.00)	0.009 (3.78)	0.015 (6.43)	0.016 (8.07)
Mandate effective date indicator	0.002 (1.81)	−0.008 (−2.73)	−0.005 (−5.49)	0.025 (4.82)	0.030 (4.89)
Interaction	−0.029 (−6.30)	−0.028 (−6.89)	−0.015 (−3.66)	−0.023 (−6.65)	−0.026 (−6.91)
Past growth in cases			0.019 (42.48)	0.019 (41.31)	0.018 (37.26)
Date fixed effects				X	X
City fixed effects					X
Observations	104,067	104,067	64,815	64,815	64,815
R^2 (%)	0.0822	1.776	4.612	5.213	5.832

In summary, using a DiD approach, we find a significantly negative relation between mask mandates and the subsequent spread of COVID-19 reported cases. We note that our analysis relies on reported COVID-19, which likely substantially underestimates the true number of COVID-19 cases, since we do not observe undiagnosed cases. However, we do not have reason to believe that this measurement error impacts our inferences, as this error would have to relate to both time trends and cross-sectional differences across cities to compromise the DiD model. We do not have evidence of such systematic errors in reported cases.

Source: Luse, A., Eaton, G., Taylor, J., & Sharda, R. (2022). Mask mandates and COVID-19 spread: Evidence from within state variation. Working paper, 2022.

SECTION 1.7 REVIEW QUESTIONS

1. Why would a health insurance company invest in analytics beyond fraud detection? Why is it in their best interest to predict the likelihood of falls by patients?
2. What other applications similar to prediction of falls can you envision?
3. How would you convince a new health insurance customer to adopt healthier lifestyles (Humana Example 2)?
4. Identify at least three other opportunities for applying analytics in the retail value chain beyond those covered in this section.
5. Which retail stores that you know of employ some of the analytics applications identified in this section?
6. List three more queries you might want to run to learn about your casino's operational performance.
7. What types of COVID-19 analytics did you see reported in your area?

1.8 PLAN OF THE BOOK

The previous sections have given you an understanding of the need for information technology in decision making, the evolution of BI, and now into analytics and data science. In the last several sections, we have seen an overview of various types of analytics and their applications. Now we are ready for a more detailed managerial excursion into these topics, along with some deep hands-on experience in some of the technical topics. Figure 1.28 presents a plan on the rest of the book.

In this chapter, we have provided an introduction, definitions, and overview of DSS, BI, and analytics, including data science. We also gave you an overview of the analytics applications in diverse domains to have you appreciate the breadth and depth of the industry. Chapter 2 similarly introduces a segment that is currently growing and becoming popular—artificial intelligence. Chapters 3 and 4 cover descriptive analytics and data issues. Chapter 3 is also a place where we introduce Big Data and related technologies. Data clearly form the foundation for any analytics application. Thus, we cover an introduction to data warehousing issues, applications, and technologies. This section also covers business reporting and visualization technologies and applications.

The next section covers predictive analytics. Chapter 5 provides an introduction to predictive analytics applications. It includes many of the common data mining techniques: classification, clustering, association mining, and so forth. Chapter 6 focuses on text mining applications as well as Web analytics, including social media analytics, sentiment analysis, and other related topics. Chapter 7 covers deep learning, explainable AI, and cognitive computing. Chapter 8 covers prescriptive analytics—optimization and simulation. Chapter 9 introduces several analytics platforms so you

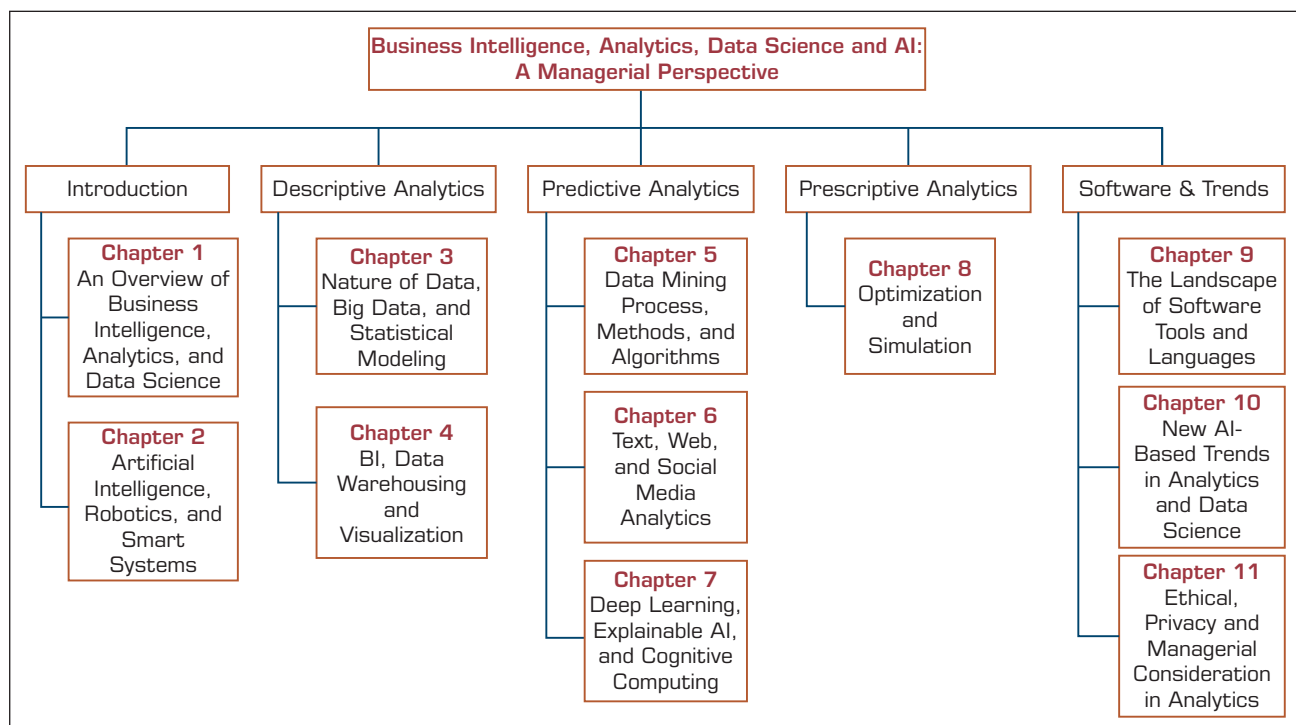


FIGURE 1.28 Plan of the Book.

are aware of the various tools and technologies to employ data science in your work. Chapter 10 includes a discussion of emerging trends, focusing on IoT, location-based analytics, and alternative data including satellite images. The ubiquity of wireless and GPS devices and other sensors is resulting in the creation of massive new databases and unique applications. A new breed of analytics companies is emerging to analyze these new databases and create a much better and deeper understanding of customers' behaviors and movements. It is leading to the automation of analytics and has also spanned a new area called the "Internet of Things." The chapter also covers cloud-based analytics. Finally, Chapter 11 attempts to integrate all the material covered in this book and concludes with a brief discussion of security/privacy dimensions of analytics.

1.9 RESOURCES, LINKS, AND THE TERADATA UNIVERSITY FOR ACADEMICS CONNECTION

The use of this chapter and most other chapters in this book can be enhanced by the tools described in the following sections.

Resources and Links

We recommend the following major resources and links:

- The Data Warehousing Institute (tdwi.org)
- Data Science Central (datasciencecentral.com)
- Microsoft Enterprise Consortium (<https://walton.uark.edu/enterprise/>)

Vendors, Products, and Demos

Most vendors provide software demos of their products and applications. Information about products, architecture, and software is available at the vendors' websites which are noted in the various chapters.

Periodicals

We recommend the following periodicals:

- Decision Support Systems* (www.journals.elsevier.com/decision-support-systems)
- CIO Insight* (cioinsight.com)

The Teradata University for Academics Connection

Several chapters in this book recommend or refer to the free resources provided by Teradata University for Academics (TUA). The TUA portal is divided into two major parts: one for students and one for faculty. We provide hands-on exercises, using software and other material (e.g., cases) available at TUA in several chapters. To explore this site, please go to <https://learning-academics.teradata.com/>. If you register as a student using your university address, you are automatically approved. If you register as a faculty member, the program administrators check to make sure a university Website does list you as a faculty member, and then you are approved.

The Book's Web Site

This book's Web site, pearsonhighered.com/sharda, contains supplemental material including online data files, powerpoints, etc.

Chapter Highlights

- The business environment is becoming more complex and is rapidly changing, making decision making more difficult.
- Businesses must respond and adapt to the changing environment rapidly by making faster and better decisions.
- The time frame for making decisions is shrinking, whereas the global nature of decision making is expanding, necessitating the development and use of computerized DSSs.
- DSSs use data, models, and sometimes knowledge management to find solutions for semistructured and some unstructured problems.
- BI methods utilize a central repository called a DW that enables efficient data mining, OLAP, BPM, and data visualization.
- BI architecture includes a DW, business analytics tools used by end users, and a user interface (such as a dashboard).
- Many organizations employ descriptive analytics to replace their traditional flat reporting with interactive reporting that provides insights, trends, and patterns in the transactional data.
- Predictive analytics enable organizations to establish predictive rules that drive the business outcomes through historical data analysis of the existing behavior of the customers.
- Prescriptive analytics help in building models that involve forecasting and optimization techniques based on the principles of OR and management science to help organizations to make better decisions.
- Big Data analytics focuses on unstructured, large data sets that may also include vastly different types of data for analysis.
- Analytics as a field is also known by industry-specific application names, such as sports analytics. It is also known by other related names such as data science or network science.
- Healthcare and retail chains are two areas where analytics applications abound, with much more to come.
- Gaming industry was one of the early adopters of BI and analytics technologies. New technologies are being employed in the gaming industry extensively.
- COVID-19 pandemic has also led to wide adoption of analytics/data science in understanding, predicting, and managing the pandemic issues.

Key Terms

analytics	decision or normative	online analytical	predictive analytics
business	analytics	processing	prescriptive analytics
intelligence (BI)	descriptive (or reporting)	(OLAP)	Web services
dashboard	analytics	online transaction	
data mining	intelligent agents	processing (OLTP)	

Questions for Discussion

1. Survey the literature from the past six months to find one application each for DSS, BI, and analytics. Summarize the applications on one page, and submit it with the exact sources.
2. Distinguish BI from DSS.
3. Compare and contrast predictive analytics with prescriptive and descriptive analytics. Use examples.
4. Discuss the major issues in implementing BI.

Exercises

Teradata University for Analytics and Other Hands-On Exercises

- Go to <https://learning-academics.teradata.com/>. Register for the site using your .edu address if you have not already previously registered. Log on and learn the content of the site. You may receive assignments related to this site. Prepare a list of 20 items on the site that you think could be beneficial to you.
- Go to the Teradata University for Analytics site. Explore the Sports Analytics page, and summarize at least two applications of analytics in any sport of your choice.
- Enter the Teradata University for Analytics site, and search for “Harrah’s.” Then select the case study “Harrah’s High Payoff from Customer Information.” Answer the following questions about this case:
 - What information does the data mining generate?
 - How is this information helpful to management in decision making? (Be specific.)
 - List the types of data that are mined.
 - Is this a DSS or BI application? Why?
- Go to <https://learning-academics.teradata.com/> and find the paper titled “Data Warehousing Supports Corporate Strategy at First American Corporation” (by Watson, Wixom, and Goodhue). Read the paper, and answer the following questions:
 - What were the drivers for the DW/BI project in the company?
 - What strategic advantages were realized?
 - What operational and tactical advantages were achieved?
 - What were the critical success factors for the implementation?
- Go to <http://analytics-magazine.org/issues/digital-editions>, and find the January/February 2013 edition titled “Work Social.” Read the article “Big Data, Analytics and Elections,” and answer the following questions:
 - What kinds of Big Data were analyzed in the article Coo? Comment on some of the sources of Big Data.
 - Explain the term *integrated system*. What is the other technical term that suits an *integrated system*?
 - What kinds of data analysis techniques are employed in the project? Comment on some initiatives that resulted from data analysis.
 - What are the different prediction problems answered by the models?
 - List some of the actionable decisions taken that were based on the prediction results.
 - Identify two applications of Big Data analytics that are not listed in the article
- Search the Internet for material regarding the work of managers and the role analytics plays. What kinds of references to consulting firms, academic departments, and programs do you find? What major areas are represented? Select five sites that cover one area, and report your findings.
- Explore the public areas of dssresources.com. Prepare a list of its major available resources. You might want to refer to this site as you work through the book.
- Go to microstrategy.com. Find information on the five styles of BI. Prepare a summary table for each style.
- Go to oracle.com, and click the Hyperion link under Applications. Determine what the company’s major products are. Relate these to the support technologies cited in this chapter.
- Go to the Teradata University for Analytics site. Look for BSI videos. Review the video of the “Case of Retail Tweeters.” Prepare a one-page summary of the problem, proposed solution, and the reported results. You can also find associated slides on slideshare.net.

References

- Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1), 1–19.
- Ackoff, R. L. (1960). Systems, organizations and interdisciplinary research. *General Systems Yearbook*, 5 (Part 1).
- Advameg. (2020). City-Data. Retrieved from <https://www.city-data.com/>
- Baker, J., & Cameron, M. (1996, September). The effects of the service environment on affect and consumer perception of waiting time: An integrative review and research propositions. *Journal of the Academy of Marketing Science*, 24, 338–349.
- Bundgaard, H., Bundgaard, J. S., Raaschou-Pedersen, D. E. T., Buchwald, C. V., Todsén, T., Norsk, J. B., ... Iversen, K. (2020). Effectiveness of adding a mask recommendation to other public health measures to prevent SARS-CoV-2 infection in Danish mask wearers. *Annals of Internal Medicine*, November 2020. DOI: 10.7326/m20-6817 %m 33205991
- Carter, N., Harrison, A., Iyengar, A., Lanham, M., Nestler, S., Schrader, D., & Zadeh, A. (2022). Clustering algorithms to increase fairness in collegiate wrestling. *Journal of Quantitative Analysis in Sports*, 18(2), 113–125, <https://doi.org/10.1515/jqas-2020-0101>

- Chernozhukov, V., Kasahara, H., & Schrimpf, P. (2020). Causal impact of masks, policies, behavior on early COVID-19 pandemic in the U.S. *Journal of Econometrics*. DOI: <https://doi.org/10.1016/j.jeconom.2020.09.003>
- De Smet, A. et al. (2017, June). Untangling your organization's decision making. *McKinsey Quarterly*.
- Galets, P., Katsaliaki, K., & Kumar, S. (2022). The medical and societal impact of big data analytics and artificial intelligence applications in combating pandemics: A review focused on COVID-19. *Social Science & Medicine*, 301. 114973. <https://doi.org/10.1016/j.socscimed.2022.114973>
- Gartner, Inc. (2004). Using business intelligence to gain a competitive edge. A special report.
- Gorry, G. A., & Scott-Morton, M. S. (1971). A framework for management information systems. *Sloan Management Review*, 13(1), 55–70.
- Greenhalgh, T., Schmid, M. B., Czypionka, T., Bassler, D., & Gruer, L. (2020). Face masks for the public during the COVID-19 crisis. *BMJ*, 369. doi: <https://doi.org/10.1136/bmj.m1435>
- Hastings, C., Mosteller, F., Tukey, J. W., & Winsor, C. P. (1947). Low moments for small samples: A comparative study of order statistics. *The Annals of Mathematical Statistics*, 18(3), 413–426.
- Karaivanov, A., En Lu, S., Shigeoka, H., Chen, C., & Pamplona, S. (2021). Face masks, public policies and slowing the spread of COVID-19: Evidence from Canada. *Journal of Health Economics*, 78. 102475, ISSN 0167–6296, <https://doi.org/10.1016/j.jhealeco.2021.102475>.
- Keen, P. G. W., & Scott-Morton, M. S. (1978). *Decision support systems: An organizational perspective*. Reading, MA: Addison-Wesley.
- Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., ... Lessler, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 172(9), 577–582.
- Loveman, Gary. (2003). Diamonds in the data mine. *Harvard Business Review*, 81(5), 109–113. <https://hbr.org/2003/05/diamonds-in-the-data-mine> (accessed June 2022).
- Simon, H. (1977). *The New Science of Management Decision*. Englewood Cliffs, NJ: Prentice Hall.
- Watson, H. (2005). Sorting out what's new in decision support. *Business Intelligence Journal*.

Artificial Intelligence

Concepts, Drivers, Major Technologies, and Business Applications

LEARNING OBJECTIVES

- Understand the concepts of artificial intelligence (AI)
- Become familiar with the drivers, capabilities, and benefits of AI
- Describe human and machine intelligence
- Describe the major AI technologies and some derivatives
- Discuss the manner in which AI supports decision making
- Describe AI applications in accounting, banking and financial services, and other business functions
- Understand basics of robotics, history, and applications
- Describe chatbots, the drivers, and capabilities of chatbots and their use
- Describe the use of chatbots as advisors
- Discuss the major issues related to the implementation of chatbots

Artificial intelligence (AI), which was a curiosity for generations, is rapidly developing into a major applied technology with many applications in a variety of fields. mission states that AI will be the most significant technology ever created by humans. AI appears in several shapes and has several definitions. In a crude way, it can be said that AI's aim is to make machines exhibit intelligence as close as possible to what people exhibit, hopefully for the benefit of humans. The latest developments in computing technologies drive AI to new levels and achievements. Companies are spending huge sums of money in AI technologies. In this chapter, we provide the essentials of AI, its major technologies, its support for decision making, and a sample of its applications in the major business functional areas. Although some topics and references may appear dated, we have kept these to provide a broader framework of AI rather than just the current buzzwords.

The chapter has the following sections:

- 2.1** Opening Vignette: Grant Thornton Employs Aisera Chatbot to Reduce IT Help Desk Burden 64
- 2.2** Introduction to Artificial Intelligence 65
- 2.3** Human and Computer Intelligence 71
- 2.4** Major AI Technologies and Some Derivatives 74
- 2.5** AI Support for Decision Making 79
- 2.6** AI Applications in Various Business Functions 83
- 2.7** Introduction to Robotics 88
- 2.8** Illustrative Applications of Robotics 93
- 2.9** Conversational AI—Chatbots 102
- 2.10** Enterprise Chatbots 106

2.1 OPENING VIGNETTE: Grant Thornton Employs Aisera Chatbot to Reduce IT Help Desk Burden

THE PROBLEM

Grant Thornton, a major accounting/consulting services firm was exploring how to manage their IT help desk operations efficiently. They rapidly needed to provide universal access to services under remote work conditions and began by adding more outsourced agents to address the inflated ticket volumes, but soon realized that this was not a sustainable option. Grant Thornton sought to give users effective self-service and automate the routine tasks that burdened the IT support team. As tickets continued to mount, Grant Thornton turned to Aisera to deploy a chatbot solution to deliver instant IT support at scale, leveraging their existing environment.

THE SOLUTION

Chatbots have been around for some time. Most early versions of chatbots employ basic natural language processing to understand the incoming query and process some basic rules to provide some canned answer from a Frequently Asked Questions (FAQ) list, usually with a link to more detailed information about the topic. This rarely provides a final resolution to the questioner, but may get them more focused on where to look for an answer. A new generation of chatbots employ conversational AI technologies to learn from the Q&A that has just taken place and adapt their future responses accordingly. Aisera's chatbot is a product in this genre. Their chatbot implementation begins with a visual work flow to connect tasks and actions in a particular setting. The company provides over 1,200 pre-built action workflows for most common and repetitive requests in key functional areas of an organization—IT, HR, or customer service applications. These can be employed to launch a prototype chatbot solution relatively quickly. The AI capability of the chatbot helps build future action work flows for other settings. Their solution titled Ticket AI brings this automation. Ticket AI allows agents to mark similar tickets and cases, knowledge articles, and ideal as well as next-best actions.

An example described on the Aisera site displays a worker from the company interacting with the chatbot to report a broken monitor. The chatbot proposes a substitute monitor. When the employee accepts this option, a monitor is ordered by the chatbot, a reference number is generated, and the interaction concludes. All of it is electronic, no human involvement is needed.

THE RESULTS

According to the managers from Grant Thornton, the benefits of Aisera were readily apparent with an issue auto-resolution rate of 75% and a 90% improvement in resolution time. The speed and accuracy of the resolutions led to an overall improvement in employee satisfaction of 85%. With the amount of turmoil going on in the world, employees should not have to worry about getting their work done and waiting for hours or even days to resolve their support requests. The decrease in mean-time-to-resolve meant less time was lost for employees to focus on their day-to-day work and agents were no longer burdened by cumbersome, manual tasks.

According to Aisera, this type of automation through chatbots leads to:

- Auto-resolve support requests across channels
- Automatically assign and route tickets and cases to the right agents
- Recommend next-best-actions for ticket and case resolution
- Automate knowledge resolution and ticket/case resolution notes
- Agent coaching with AI-driven recommendations

Sources: Compiled with permission from Aisera.com Website, <https://aisera.com/customers/grant-thornton/> (accessed July 2022), and <https://aisera.com/customers/grant-thornton/?uvideo=2mmguci779>, and conversations with Muddu Sudhakar, CEO/Founder, Aisera, Karthik S. J., and Vishal Agrawal, Aisera.

► SECTION 2.1 REVIEW QUESTIONS

1. Have you interacted with a Website where an online chatbot provides answers? What has been your experience in getting your questions resolved?
2. What type of learning will be needed by a chatbot to understand a customer's query? Will a system such as chatGPT replace such focused chatbots?
3. What other examples of chatbot applications can you find?

WHAT CAN WE LEARN FROM THIS VIGNETTE?

As we will see in this chapter and the book, AI is a very broad field with applications in many domains. Chatbots (introduced later in the chapter) provide a relatively commonly used technology for automating at least some commonly asked questions. Most older chatbots do just some word-matching to provide some general answers or links, usually sending most visitors to a human agent. The next generation of chatbots that can learn from their past interactions and also integrate back-office processes provide significantly more capability and ROI. The recent buzz about chatGPT is of interest as well.

2.2 INTRODUCTION TO ARTIFICIAL INTELLIGENCE

We would all like to see computerized decision making being simpler, easier to use, more intuitive, and less threatening. And indeed, efforts have been made over time to simplify and automate several tasks in the decision-making process. Just think of the day that refrigerators will be able to measure and evaluate their contents and place orders for goods that need replenishment. Such a day is not too far in the future, and the task will be supported by AI.

CIO Insight projected that by 2035, intelligent computer technologies will result in \$5–\$8.3 trillion in economic value (see cioinsight.com/blogs/how-ai-will-impact-the-global-economy.html). Among the technologies listed as intelligent ones are the IoT, advanced robotics, and self-driven vehicles, all introduced in this chapter. For a history of AI, see Zarkadakis (2016) and en.wikipedia.org/wiki/History_of_artificial_intelligence.

Definitions

Artificial intelligence has several definitions (for an overview see Marr, 2018); however, many experts agree that AI is concerned with two basic ideas: (1) the study of human thought processes (to understand what intelligence is) and (2) the representation and duplication of those thought processes in machines (e.g., computers, robots). That is, the machines are expected to have humanlike thought processes.

One well-publicized definition of AI is “the capabilities of a machine to imitate intelligent human behavior” (per *Merriam-Webster Dictionary*). The theoretical background of AI is based on logic, which is also used in several computer science innovations. Therefore, AI is considered a subfield of computer science. For the relationship between AI and logic, see plato.stanford.edu/entries/logic-ai.

A well-known early application of artificial intelligence was the chess program hosted at IBM’s supercomputer (Deep Blue). The system beat the famous world champion, Grand Master Garry Kasparov. A very recent (as of the time of this writing) example that has added to the buzz about AI is the ChatGPT system developed by OpenAI. It is able to generate text/responses to many types of queries. We introduce this a bit further in the last section of this chapter and also in several other chapters.

AI is an umbrella term for many techniques that share similar capabilities and characteristics. Spacey (2016) identifies 33 types of AI; see simplicable.com/new/types-of-artificial-intelligence.

Major Characteristics of AI Machines

There is an increasing trend to make computers “smarter.” For example, Web 3.0 supposes to enable computerized systems that exhibit significantly more intelligence than Web 2.0. Several applications are already based on multiple AI techniques. For example, the area of machine translation of languages is helping people who speak different languages to collaborate as well as to buy online products that are advertised in languages they do not speak. Similarly, machine translation can help people who know only their own language to converse with people speaking other languages and to make decisions jointly in real time.

Major Elements of AI

As described in Chapter 1, the landscape of AI is huge, including hundreds or more components. We illustrate the foundation and the major technologies in Figure 2.1. Notice that we divide them into two groups: Foundations, and Technologies and Applications. The major technologies will be defined later in this chapter and described throughout this book.

AI Applications

The technologies of AI are used in the creation of a large number of applications. In Section 2.6 we provide a sampler of applications in the major functional areas of business.

Example

Smart or intelligent applications include those that can help machines to answer customers’ questions asked in natural languages. Another area is that of knowledge-based systems which can provide advice, assist people to make decisions, and even make decisions on their own. For example, such systems can approve or reject buyers’ requests to purchase online (if the buyers are not preapproved or do not have an open line of credit). Other examples include the automatic generating of online purchasing orders and

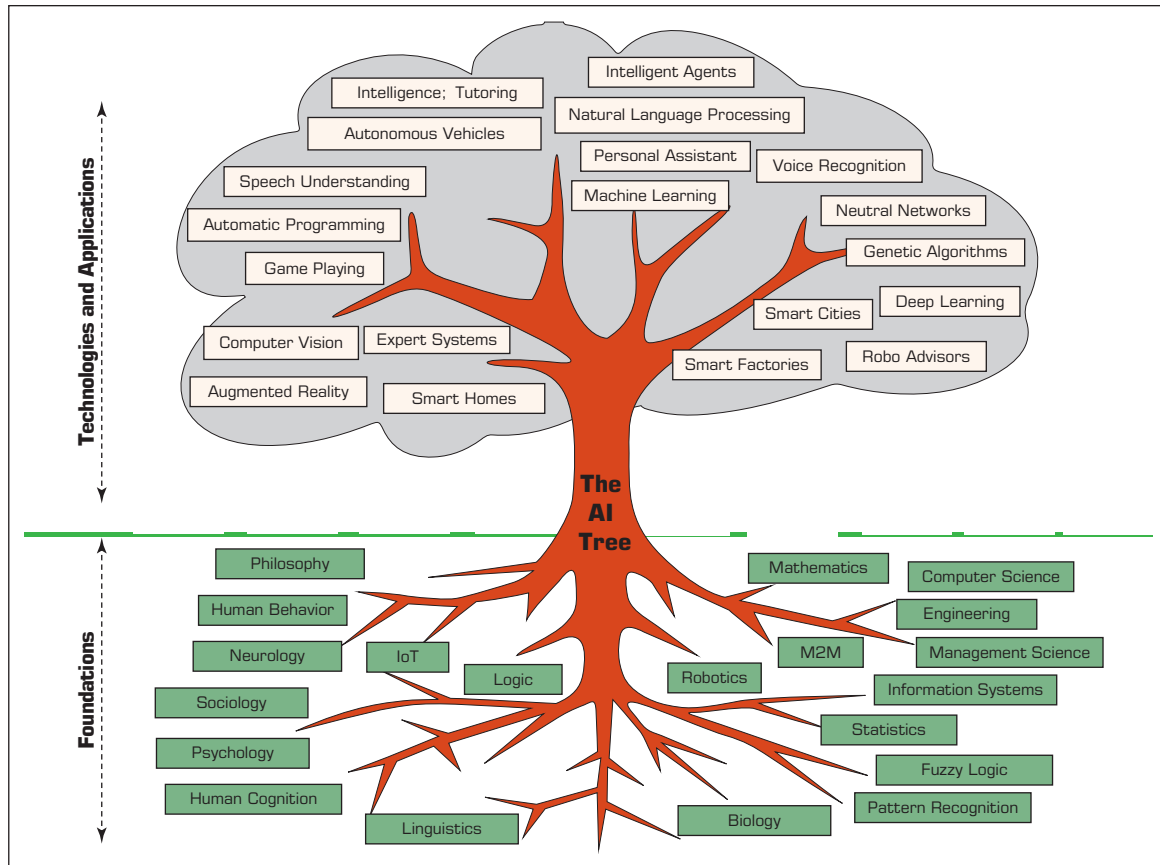


FIGURE 2.1 The Functionalities and Applications of Artificial Intelligence.

arranging fulfillment of orders placed online. Google, Facebook and hundreds of other companies are experimenting with projects that attempt to teach machines how to learn and support or even make autonomous decisions.

Example: Pitney Bowes Is Getting Smarter with AI

Pitney Bowes Inc. is a U.S.-based global business solutions provider in areas such as product shipments, location intelligence, customer engagement, and customer information management. The company powers billions of physical and digital transactions—annually across the connected and borderless world of commerce.

Today, at Pitney Bowes, shipping prices are determined automatically based on the dimensions, weight, and packaging of each package. The fee calculations create data that are fed into AI algorithms. The more data processed, the more accurate are the calculations (a machine-learning characteristic). The company estimates a 25% improvement in calculations achieved from their algorithms. This gives Pitney Bowes an accurate base for pricing, better customer satisfaction, and improved competitive advantage.

Major Goals of AI

The overall goal of AI is to create intelligent machines that are capable of executing a variety of tasks currently done by people. Ideally, AI machines should be able to reason, think abstractly, plan, solve problems, and learn.

Some specific goals are to:

- Perceive and properly react to changes in the environment that influence specific business processes and operations.
- Introduce creativity in business processes and decision making.

Drivers of AI

The use of AI has been driven by the following forces:

- People's interest in smart machines and artificial brains
- The low cost of AI applications versus the high cost of manual labor (doing the same work)
- The desire of large tech companies to capture competitive advantage and market share of the AI market and their willingness to invest billions of dollars in AI
- The pressure on management to increase productivity and speed
- The availability of quality data contributing to the progress of AI
- The increasing functionalities and reduced cost of computers in general
- The development of new technologies, particularly cloud computing

BENEFITS OF AI The major benefits of AI are as follows:

- AI has the ability to complete certain tasks much faster than humans.
- The consistency of the completed AI work can be much better than that of humans. AI machines do not make mistakes.
- AI systems allow for continuous improvement projects.
- AI can be used for predictive analysis via its capability of pattern recognition.
- AI can manage delays and blockages in business processes.
- AI machines do not stop to rest or sleep.
- AI machines can work autonomously or be assistants to humans.
- The functionalities of AI machines are ever increasing.
- AI machines can learn and improve their performance.
- AI machines can work in environments that are hazardous to people.
- AI machines can facilitate innovations by human (i.e., support research and development [R&D]).
- No emotional barriers interfere with AI work.
- AI excels in fraud detection and in security facilitations.
- AI improves industrial operations.
- AI optimizes knowledge work.
- AI increases speed and enables scale.
- AI helps with the integration and consolidating of business operations.
- AI applications can reduce risk.
- AI can free employees to work on more complex and productive jobs.
- AI improves customer care.
- AI can solve difficult problems that previously were unsolved (Kharpal, 2017).
- AI increases collaboration and speeds up learning.

These benefits help facilitate competitive advantages.

Note: Not all AI systems deliver all these benefits. Specific systems may deliver only some of them.

The capability of reducing costs and increasing productivity may result in large increases in profit (Violino, 2017). In addition to benefiting individual companies, AI can dramatically increase a country's economic growth, as it is doing in Singapore.

Examples of AI Benefits The following are typical benefits of AI in various areas of applications:

1. The International Swaps and Derivatives Association (ISDA) uses AI to eliminate tedious activities in contract procedures. For example, by using optical character recognition (OCR) integrated with AI, ISDA digitizes contracts and then defines, extracts, and archives the contracts.
2. AI is starting to revolutionize business recruitment by (1) conducting more efficient and fairer candidate screening, (2) making better matches of candidates to jobs, and (3) helping safeguard future talent pipelines for organizations.
3. AI is redefining management. According to Kolbjørnsrud et al. (2016), the following five practices result from the use of AI:

- It can perform routine administrative tasks.
- Managers can focus on the judgment portions of work.
- Intelligent machines are treated as colleagues (i.e., managers trust the advice generated by AI). In addition, there is people-machine collaboration.
- Managers concentrate on creative abilities that can be supported by AI machines.
- Managers are developing social skills, which are needed for better collaboration, leadership, and coaching.

4. Accenture Inc. developed AI-powered solutions using natural language processing (NLP) and image recognition to help blind people in India improve the way that they can experience the world around them. This enables them to have a better life, and those who work can work better, faster, and do jobs that are more challenging.
5. Ford Motor Credit uses machine learning to spot overlooked borrowers. In addition, it uses machine learning to help its underwriters better understand loan applicants. The program helps the productivity of both underwriters and overlooked applicants. Finally, the system predicts potential borrowers' creditworthiness, thus minimizing losses for Ford.
6. Alastair Cole uses data collected from several sources with IBM Watson to predict what customers are expecting from the company. The generated data are used for supporting more efficient business decisions.
7. Companies are building businesses around AI. There are many examples of start-ups or existing companies that are attempting to create new businesses.

SOME LIMITATIONS OF AI MACHINES The following are the major limitations of AI machines:

- They lack human touch and feel
- They lack attention to non-task surroundings
- They can lead people to rely on AI machines too much (e.g., people may stop to think on their own)
- They can be programmed to create destruction
- They may cause some people to lose their jobs (see Chapter 11)
- They can start to think by themselves, causing significant damage

Some of the limitations are diminishing with time. However, risks exist. Therefore, it is necessary to properly manage the development of AI and try to minimize risks.

What AI Can and Cannot Do The limitations just identified constrain the capabilities of commercial AI. For example, it could cost too much to be commercially used. This is important for two reasons: (1) executives need to know what AI can do economically and how companies can use it to benefit their business and (2) executives need to know what AI cannot economically do.

AI is already transforming Web search, retailing and banking services, logistics, online commerce, entertainment, and more. Hundreds of millions of people use AI on their smartphones and in other ways. However, according to Ng (2016), applications in these areas are based on how simple input is converted to simple output as a response; for example, in automatic loan approval, the input is the profile of the applicant and the output will be an approval or rejection.

Applications in these areas are normally fully automated. Automated tasks are usually repetitive and done by people with short periods of training. AI machines depend on data that may be difficult to get (e.g., belong to someone else) or inaccurate. A second barrier is the need for AI experts, who are difficult to find and/or expensive to hire. For other barriers, see Chapter 11.

Three Flavors of AI Decisions

A writer at TG Daily (2017) divided the capabilities of AI systems into three levels: assisted, autonomous, and augmented.

ASSISTED INTELLIGENCE This is equivalent mostly to weak AI, which works only in narrow domains. It requires clearly defined inputs and outputs. Examples are some monitoring systems and low-level virtual personal assistants. Such systems and assistants are used in our vehicles for giving us alerts. Similar systems can be used in many healthcare applications (e.g., monitoring, diagnosing).

AUTONOMOUS AI These systems are in the realm of the strong AI but in a very narrow domain. Eventually, a computer will take over many tasks, automating them completely. Machines act as experts and have absolute decision-making power. Pure robo-advisors are examples of such machines. Autonomous vehicles and robots that can fix themselves are also good examples.

AUGMENTED INTELLIGENCE Most of the existing AI applications, which are between assisted and autonomous, are referred to as **augmented intelligence** (or intelligence augmentation). Their technology can augment computer tasks to extend human cognitive abilities, resulting in high performance, as described in Technology Insight 2.1.

Padmanabhan (2018) and many others believe that at the moment, augmented AI is the best option to deal with practical problems and transform organizations to be “smarter.”

In contrast with autonomous AI, which describes machines with a wide range of cognitive abilities (e.g., driverless cars), augmented intelligence has only a few cognitive abilities.

TECHNOLOGY INSIGHT 2.1 Augmented Intelligence

The idea of combining the performance of people and machines is not new. In this section, we discuss combining (augmenting) human abilities with powerful machine intelligence—not replacing people, which autonomous AI does, but extending human cognitive abilities. The result is the ability of humans to solve more complex problems, as in the opening vignette to Chapter 1. Computers have provided data to help people solve problems for which no solution had been available. Padmanabhan (2018) specifies the following differences between traditional and augmented AI:

1. Augmented machines extend human thinking capabilities rather than replace human decision making. These machines facilitate creativity.
2. Augmentation excels in solving complex human and industry problems in specific domains in contrast with strong, general AI machines, which are still in development.
3. In contrast with a “black box” model of some AI and analytics, the augmented intelligence provides insights and recommendations, including explanations.
4. In addition, augmented technology can offer new solutions by combining existing and discovered information in contrast to assisted AI that identifies problems or symptoms and suggests predetermined known solutions.

SECTION 2.2 REVIEW QUESTIONS

1. Define AI.
2. What are the major aims and goals of AI?
3. List some characteristics of AI.
4. List some AI drivers.
5. List some benefits of AI applications.
6. List some AI limitations.
7. List the three flavors of AI and describe augmentation.

2.3 HUMAN AND COMPUTER INTELLIGENCE

AI usage is growing rapidly due to its increased capabilities. To understand AI, we need to first explore the meaning of intelligence.

What Is Intelligence?

Intelligence can be considered to be an umbrella term and is usually measured by an IQ test. However, some claim that there are several types of intelligence. For example, Dr. Howard Gardner of Harvard University proposed the following types of intelligence (also explained by Brualdi, 1996):

- Linguistic and verbal
- Logical
- Spatial
- Body/movement
- Musical
- Interpersonal
- Intrapersonal
- Naturalist

Thus, intelligence is not a simple concept.

Content of Intelligence Intelligence is composed of reasoning, learning, logic, problem-solving ability, perception, and linguistic ability.

Obviously, the concept of intelligence is not simple.

Capabilities of Intelligence To understand what artificial intelligence is, it is useful to first examine those abilities that are considered signs of human intelligence:

- Learning or understanding from experience
- Making sense out of ambiguous, incomplete, or even contradictory messages and information
- Responding quickly and successfully to a new situation (i.e., using the most correct responses)
- Understanding and inferring in a rational way, solving problems, and directing conduct effectively
- Applying knowledge to manipulate environments and situations
- Recognizing and judging the relative importance of different elements in a situation

AI attempts to provide some, hopefully all, of these capabilities, but in general, it is still not capable of matching human intelligence.

How Intelligent Is AI?

AI machines have demonstrated superiority over humans in playing complex games such as chess (beating the world champion), *Jeopardy!* (beating the best players), and Go (a complex Chinese game) whose top players were beaten by a computer using the well-known program, Google's DeepMind (see Hughes, 2016). Recently, ChatGPT is attracting headlines for generating text that is almost similar to that created by a human user. Despite these remarkable demonstrations (whose cost is extremely high), many AI applications still show significantly less intelligence than humans.

Comparing Human Intelligence with AI Several attempts have been made to compare human intelligence with AI. There is difficulty in doing so because it is a multidimensional situation. A comparison is presented in Table 2.1.

Measuring AI

The Turing Test is a well-known attempt to measure the intelligence level of AI machines.

TURING TEST: THE CLASSICAL MEASURE OF MACHINE INTELLIGENCE Alan Turing designed a test known as the **Turing Test** to determine whether a computer exhibits intelligent behavior. According to this test, a computer can be considered smart only when a human interviewer asking the same questions to both an unseen human and an unseen computer cannot determine which is which (see Figure 2.2). Note that this test is limited to a question-and-answer (Q&A) mode. To pass the Turing Test, a computer needs to be able to understand a human language (NLP), to possess human intelligence (e.g., have a knowledge base), to reason using its stored knowledge, and to be able to learn from its experiences (machine learning).

OTHER TESTS Over the years, there have been several other proposals of how to measure machine intelligence. For example, improvements in the Turing Test appear in several variants. Major U.S. universities (e.g., University of Illinois, Massachusetts Institute of Technology [MIT], Stanford University) are engaged in studying the IQ of AI.

In conclusion, it is difficult to measure the level of intelligence of humans as well as that of machines. Doing so depends on the circumstances and the metrics used.

TABLE 2.1 Artificial Intelligence versus Human Intelligence

Area	AI	Human
Execution	Very fast	Can be slow
Emotions	Not yet	Can be positive or negative
Computation speed	Very fast	Slow, may have trouble
Imagination	Only what is programmed for	Can expand existing knowledge
Answers to questions	What is in the program	Can be innovative
Flexibility	Rigid	Large, flexible
Foundation	A binary code	Five senses
Consistency	High	Variable, can be poor
Process	As modeled	Cognitive
Form	Numbers	Signals
Memory	Built in, or accessed in the cloud	Use of content and scheme memory
Brain	Independent	Connected to a body
Creativity	Uninspired	Truly creative
Durability	Permanent, but can get obsolete if not updated	Perishable, but can be updated
Duplication, documentation, and dissemination	Easy	Difficult
Cost	Usually low and declining	Maybe high and increasing
Consistency	Stable	Erratic at times
Reasoning process	Clear, visible	Difficult to trace at times
Perception	By rules and data	By patterns
Figure missing data	Usually cannot	Frequently can

Regardless of the determination of how intelligent a machine is, AI exhibits a large number of benefits as described earlier.

It is important to note that the capabilities of AI are increasing with time. Besides the Turing test mentioned above, many other stories appear reporting a computer beating human players against games such as chess, GO, etc. For a discussion of AI versus human intelligence, see Carney (2018).

► SECTION 2.3 REVIEW QUESTIONS

1. What is intelligence?
2. What are the major capabilities of human intelligence? Which are superior to that of AI machines?
3. How intelligent is AI?
4. How can we measure AI's intelligence?
5. What is the Turing Test and what are its limitations?
6. How can one measure the intelligence level of a vacuum cleaner?

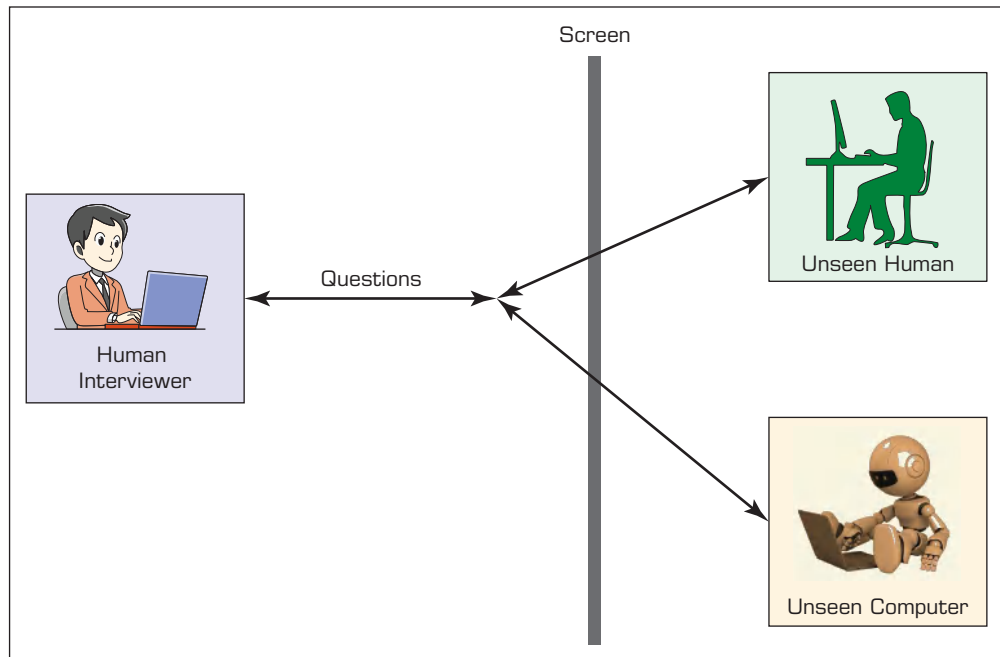


FIGURE 2.2 A Pictorial Representation of the Turing Test.

2.4 MAJOR AI TECHNOLOGIES AND SOME DERIVATIVES

The AI field is very broad; we can find AI technologies and applications in hundreds of disciplines ranging from medicine to sports. Press (2017) lists 10 top AI technologies similar to what is covered in this book. Press also provides the status of the life cycle (ecosystem phase) of the technologies. In this section, we present some major AI technologies and their derivatives as related to business. The selected list is illustrated in Figure 2.3.

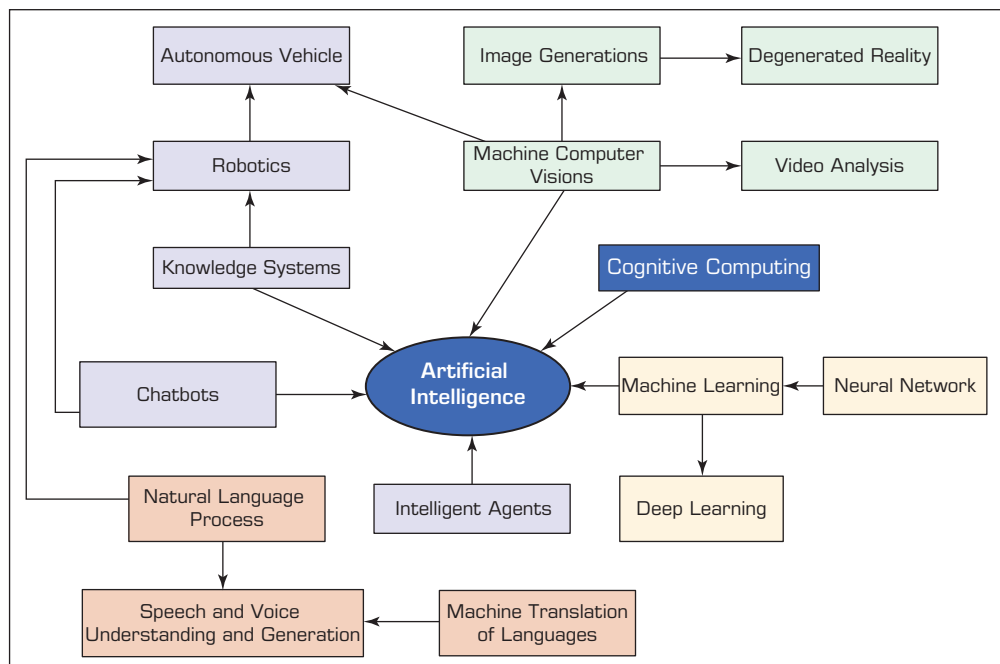


FIGURE 2.3 The Major AI Technologies.

INTELLIGENT AGENTS An **intelligent agent (IA)** is an autonomous, relatively small computer software program that observes and acts upon changes in its environment by running specific tasks autonomously. An IA directs an agent's activities to achieve specific goals related to the changes in the surrounding environment. Intelligent agents may have the ability to learn by using and expanding the knowledge embedded in them. Intelligent agents are effective tools for overcoming the most critical burden of the Internet information overload and making computers more viable decision support tools. Interest in using intelligent agents for business and e-commerce started in the academic world in the mid-1990s. However, only since 2014, when the capabilities of IA increased remarkably, have we started to see powerful applications in many areas of business, economics, government, and services.

Initially, intelligent agents were used mainly to support routine activities such as searching for products, getting recommendations, determining products' pricing, planning marketing, improving computer security, managing auctions, facilitating payments, and improving inventory management. However, these applications were very simple, using a low level of intelligence. Their major benefits were increasing speed, reducing costs, reducing errors, and improving customer service. Today's applications, as we will see throughout this chapter, are much more sophisticated.

Example 1: Virus Detection Program

A simple example of an intelligent software agent is a virus detection program. It resides in a computer, scans all incoming data, and removes found viruses automatically while learning to detect new virus types and detection methods.

Example 2

Allstate Business Insurance is using an intelligent agent to reduce call center traffic and provide help to human insurance agents during the rate-quoting process with business customers. In these cases, rate quotes can be fairly complicated. Using this system, agents can quickly answer questions posted by corporate customers, even if the agents are not fully familiar with the related issue.

Intelligent agents are also utilized in e-mail servers, news filtering and distribution, appointment handling, and automated information gathering.

MACHINE LEARNING At this time, AI systems do not have the same learning capabilities that humans have; rather, they have simplistic (but improving) **machine learning** (modeled after human learning methods). The machine-learning scientists try to teach computers to *identify patterns* and make connections by showing the machines a large volume of examples and related data. Machine learning also allows computer systems to monitor and sense their environmental activities so the machines can adjust their behavior to deal with changes in the environment. The technology can also be used to predict performance, to reconfigure programs based on changing conditions, and much more. Technically speaking, machine learning is a scientific discipline concerned with the design and development of algorithms that allow computers to learn based on data coming from **sensors**, databases, and other sources. This learning is then used for making predictions, recognizing patterns, and supporting decision makers.

The process of machine learning involves computer programs that learn as they face new situations. Such programs collect data and analyze them and then “train” themselves to arrive at conclusions. For example, by showing examples of situations to a machine-learning program, the program can find elements not easily visible without it. A well-known example is that of computers detecting credit card fraud.

According to Taylor (2016), the “increased computing power, coupled with other improvements including better algorithms and deep neural networks for image processing, and ultra-fast in-memory databases like SAP HANA, are the reasons why machine learning is one

of the hottest areas of development in enterprise software today.” Machine-learning applications are also expanding due to the availability of Big Data sources, especially those provided by the IoT (Chapter 10). There are several methods of machine learning. They range from neural networks to case-based reasoning. The major ones are presented in Chapters 5–7.

Deep Learning One subset, or refinement, of machine learning is called **deep learning**. This technology, which is discussed in Chapter 6, tries to mimic how the human brain works. Deep learning uses artificial neural technology and plays a major role in dealing with complex applications that regular machine learning and other AI technologies cannot handle. Deep learning (DL) delivers systems that not only think but also keep learning, enabling self-direction based on fresh data that flow in. DL can tackle previously unsolvable problems using its powerful learning algorithms.

For example, DL is a key technology in autonomous vehicles by helping to interpret road signs and road obstacles. DL is also playing critical roles in smartphones, robotics, tablets, smart homes, and smart cities (Chapter 10). DL is mostly useful in real-time interactive applications in the areas of machine vision, scene recognition, robotics, and speech and voice processing. The key is continuous learning. As long as new data arrive, learning occurs.

Example

Cargill Corp. offers conventional analytics, and DL-based analytics help farmers to do more profitable work. For example, farmers can produce better shrimp at lower cost. DL is used extensively in stock market analysis and predictions.

MACHINE AND COMPUTER VISION The definitions of **machine vision** vary because several different computer vision systems include different hardware and software as well as other components. Generally speaking, the classical definition is that the term *machine vision* includes “the technology and methods used to provide imaging-based automated inspection and analysis for applications such as robot guidance, process control, autonomous vehicles, and inspection.” Machine vision is an important tool for the optimization of production and robotic processes. A major part of machine vision is the industrial camera, which captures, stores, and archives visual information. This information is then presented to users or computer programs for analysis and eventually for automatic decision making or for support of human decision making. Machine vision can be confused with computer vision because sometimes the two are used as synonyms, but some users and researchers treat them as different entities. Machine vision is treated more as an engineering subfield, while computer vision belongs to the computer science area.

Computer Vision **Computer vision**, according to Wikipedia, “is an interdisciplinary field that deals with how computers can be made for gaining high-level understanding from digital images or videos. From the perspective of engineering, it seeks to automate tasks that the human visual system can do.” Computer vision acquires or processes, analyzes, and interprets digital images and produces meaningful information for making decisions. Image data can take several formats, such as photos or videos, and they can come from multi-dimensional sources (e.g., medical scanners). Scene and item recognitions are important elements in computer vision. The computer vision field plays a vital role in the domains of safety, security, health, and entertainment. Computer vision is considered a technology of AI, which enables robots and autonomous vehicles to see. Both computer vision and machine vision automate many human tasks (e.g., inspection). These tasks can deal with one image or a sequence of images. The major benefit of both technologies is lowering the costs of performing tasks, especially those that are repetitive and make the human eyes tired. The two technologies are also combined with *image processing* that facilitates complex applications, such as in visual quality control. Another view shows them as being interrelated based on image processing and sharing a variety of contributing fields.

An applied area of machine vision is **scene recognition**, which is performed by computer vision. Scene recognition enables recognition and interpretation of objects, scenery, and photos.

Example of Application

Significant illegal logging exists in many countries. To comply with the laws in the United States, Europe, and other countries, it is necessary to examine wood in the field. This requires expertise. According to the U.S. Department of Agriculture, “the urgent need for such field expertise, training and deploying humans to identify processed wood in the field [i.e., at ports, border crossings, weigh-stations, airports, and other points of entry for commerce] is prohibitively expensive and difficult logistically. The machine vision wood identification project (MV) has developed a prototype machine vision system for wood identification.” Similarly, AI computer vision combined with deep learning is used to identify illegal poachers of animals (see USC, 2018).

Another example of this application is *facial recognition* in several security applications, such as those used by the Chinese police that employ smart glasses to identify (via facial recognition) potential suspects. In 2018, the Chinese police identified a suspect who attended a pop concert. There were 60,000 people in the crowd. The person was recognized at the entrance gate where a camera took his picture; see the video at youtube.com/watch?v=Fq1SEqNT-7c. Indeed, facial recognition and its implications for surveillance is becoming a big privacy concern.

Video Analytics Applying computer vision techniques to videos enables the recognition of patterns (e.g., for detecting fraud) and identifying events. This is a derivative application of computer vision. Another example is one in which, by letting computers view TV shows, it is possible to train the computers to make predictions regarding human interactions and the success of advertising.

NATURAL LANGUAGE PROCESSING **Natural language processing (NLP)** is a technology that gives users the ability to communicate with a computer in their native language. The communication can be in written text and/or in voice (speech). This technology allows for a conversational type of interface in contrast with using a programming language that consists of computer jargon, syntax, and commands. NLP includes two subfields:

- *Natural language understanding* that investigates methods of enabling computers to comprehend instructions or queries provided in ordinary English or other human languages.
- *Natural language generation* that strives to have computers produce ordinary spoken language so that people can understand the computers more easily. For details and the history of NLP, see en.wikipedia.org/wiki/Natural_language_processing. NLP is related to voice-generated data as well as text and other communication forms.

Speech (Voice) Understanding **Speech (voice) understanding** is the recognition and understanding of spoken languages by a computer. Applications of this technology have become more popular. For instance, many companies have adopted this technology in their automated call centers. For an interesting application from almost ten years back that is still accessible, see cs.cmu.edu/~./listen.

Machine Translation of Languages Machine translation uses computer programs to translate words and sentences from one language to another. For example, Babel Fish Translation, available at babelfish.com, offers more than 25 different combinations of language translations. Similarly, Google’s Translate (translate.google.com) or the available

App can translate dozens of different languages. An interesting combination of vision and machine translation has been the ability of Google Translate app to view any text through the camera and present its translation in another desired language. Similarly, the app can also listen to speech and convert it into another language. Finally, users can post their status on Facebook in several languages.

Knowledge Sources and Acquisition for Intelligent Systems For many intelligent systems to work, it is necessary for them to have knowledge. The process of acquiring this knowledge is referred to as **knowledge acquisition**. This activity can be complex because it is necessary to make sure what knowledge is needed. It must fit the desired system. In addition, the sources of the knowledge need to be identified to ensure the feasibility of acquiring the knowledge. The specific methods of acquiring the knowledge need to be identified and if expert(s) are the source of knowledge, their cooperation must be ensured. In addition, the method of knowledge representation and reasoning from the collected knowledge must be taken into account, and knowledge must be validated and be consistent.

Given this information, it is easy to see that the process of knowledge acquisition (see Figure 2.4) can be very complex. It includes extracting and structuring knowledge. It has several methods (e.g., observing, interviewing, scenario building, and discussing), so specially trained knowledge engineers may be needed for knowledge acquisition and system building. In many cases, teams of experts with different skills are created for knowledge acquisition. Knowledge can be generated from data, and then experts may be used to verify it. The acquired knowledge needs to be organized in an activity referred to as *knowledge representation*.

Knowledge Representation Acquired knowledge needs to be organized and stored. There are several methods of doing this, depending on what the knowledge will be used for, how the reasoning from this knowledge will be done, how users will interact with the knowledge, and more. A simple way to represent knowledge is in the form of questions and matching answers (Q&A).

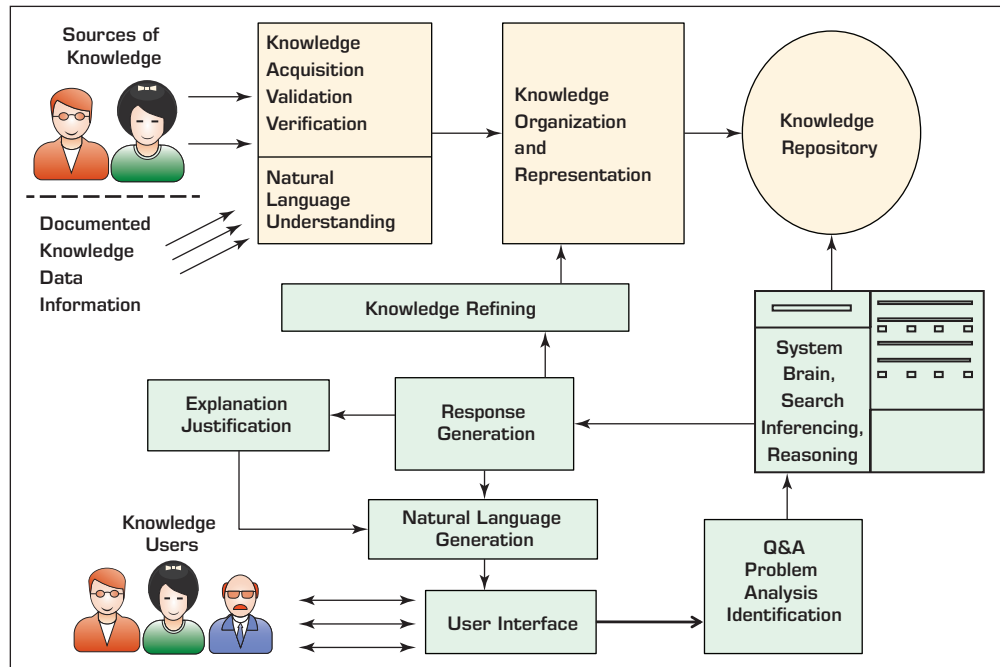


FIGURE 2.4 Automated Decision-Making Process.

Reasoning from Knowledge Perhaps the most important component in an intelligent system is its reasoning feature. This feature processes users' requests and provides answers (e.g., solutions, recommendations) to the user. The major difference among the various types of the intelligent technologies is the type of reasoning they use.

Cognitive Computing **Cognitive computing** is the application of knowledge derived from cognitive science (the study of the human brain) and computer science theories in order to simulate the human thought processes (an AI objective) so that computers can exhibit and/or support decision-making and problem-solving capabilities. To do so, computers must be able to use *self-learning algorithms*, pattern recognition, NLP, machine vision, and other AI technologies. IBM is a major proponent of the concept by developing technologies (e.g., Watson) that support people in making complex decisions. Cognitive computing systems learn to reason with purpose, and interact with people naturally.

Augmented Reality **Augmented reality** (AR) refers to the integration of digital information with the user environment in real time (mostly vision and sound). The technology provides people real-world interactive experience with the environment. Therefore, information may change the way people work, learn, play, buy, and connect. Sophisticated AI programs may include machine vision, scene recognition, and gesture recognition. AR is available on iPhones as ARKit. These AR systems use data captured by sensors (e.g., vision, sound, temperature) to augment and supplement real-world environments. For example, if a photo is taken of a house on a cellphone, there is immediate public availability of information about its configuration, ownership, and tax liabilities.

► SECTION 2.4 REVIEW QUESTIONS

1. Define *intelligent agents* and list some of their capabilities.
2. Prepare a list of applications of intelligent agents.
3. What is machine learning? How can it be used in business?
4. Define *deep learning*.
5. Define *robotics* and explain its importance for manufacturing and transportation.
6. What is NLP? What are its two major formats?
7. Describe machine translation of languages. Why it is important in business?
8. What are knowledge systems?
9. What is cognitive computing?
10. What is augmented reality?

2.5 AI SUPPORT FOR DECISION MAKING

Almost since the inception of AI, researchers have recognized the opportunity of using it for supporting the decision-making process and for completely automating decision making. Jeff Bezos, the-then CEO of Amazon.com, said in May 2017 that AI is in a golden age, and it is solving problems that were once in the realm of science fiction (Kharpal, 2017). Bezos also said that Amazon.com is using AI in literally hundreds of applications, and AI is really of amazing assistance. Amazon.com has been using AI, for example, for product recommendations for over 20 years. The company also uses AI for product pricing, and as Bezos said, to solve many difficult problems. And indeed, since its inception, AI has been related to problem solving and decision making. AI technologies allow people to make better decisions. The fact is that AI can:

- Solve complex problems that people have not been able to solve. (Note that solving problems frequently involves making decisions.)
- Make much faster decisions. For example, Amazon makes millions of pricing and recommendation decisions, each in a split second.
- Find relevant information, even in large data sources, very fast.
- Make complex calculations rapidly.
- Conduct complex comparisons and evaluations in real time.

In a nutshell, AI can drive some types of decisions many times faster and more consistently than humans can. For details, watch the video at [youtube.com/watch?v=Dr9jeRy9whQ/](https://www.youtube.com/watch?v=Dr9jeRy9whQ/). The nature of decision making, especially nonroutine ones, as noted in Chapter 1, is complex. We discussed in Chapter 1 the fact that there are several types of decisions and several managerial levels of making them, and we looked at the typical process of making decisions. Making decisions, many of which are used for problem solving, requires intelligence and expertise. AI's aim is to provide both. As a result, it is clear that using AI to facilitate decision making involves many opportunities, benefits, and variations. For example, AI can successfully support certain types of decision making and fully automate others.

In this section, we discuss some general issues of AI decision support. The section also distinguishes between *support of decision making* and *fully automating decision making*.

SOME ISSUES AND FACTORS IN USING AI IN DECISION MAKING Several issues determine the justification of using AI and its chance of success. These include:

- The nature of the decision. For example, routine decisions are more likely to be fully automated, especially if they are simple.
- The method of support, what technology(ies) is (are) used. Initially, automated decision supports were rule-based. Practically, expert systems were created to generate solutions to specific decision situations in well-defined domains. Another popular technology mentioned earlier was “recommender,” which appeared with e-commerce in the 1990s. Today, there is an increased use of machine learning and deep learning. A related technology is that of pattern recognition. Today, attention is also given to biometric types of recognition.

For example, research continues to develop an AI machine that will interview people at airports, asking one or two questions, and then determining whether they are telling the truth. Similar algorithms can be used to vet refugees and other types of immigrants.

- *Cost-benefit and risk analyses.* These are necessary for making large-scale decisions, but computing these values may not be simple with AI models due to difficulties in measuring costs, risks, and benefits. For example, as we cited earlier, researchers used 100 metrics to measure the intelligence level of vacuum cleaners.
- *Using business rules.* Many AI systems are based on business or other types of rules. The quality of automated decisions depends on the quality of these rules. Advanced AI systems can learn and improve business rules.
- *AI algorithms.* There is an explosion in the number of AI algorithms that are the basis for automated decisions and decision support. The quality of the decisions depends on the input of the algorithms, which may be affected by changes in the business environment.
- *Speed.* Decision automation is also dependent on the speed within which decisions need to be made. Some decisions cannot be automated because it takes too much time to get all the relevant input data. On the other hand, manual decisions may be too slow for certain circumstances.

AI SUPPORT OF THE DECISION-MAKING PROCESS Much AI support can be applied today to the various steps of the decision-making process. Fully automated decisions are common in routine situations and will be discussed in the next section. Here we follow the steps in the decision-making process described in Chapter 1.

Problem Identification AI systems are used extensively in problem identification typically in diagnosing equipment malfunction and medical problems, finding security breaches, estimating financial health, and so on. Several technologies are used. For example, sensor-collected data are used by AI algorithms. Performance levels of machines are compared to standards, and trend analysis can point to opportunities or troubles.

Generating or Finding Alternative Solutions Several AI technologies offer alternative solutions by matching problem characteristics with best practices or proven solutions stored in databases. Both expert systems and chatbots employ this approach. They can generate recommended solutions or provide several options from which to choose. AI tools such as case-based reasoning and neural computing are used for this purpose.

Selecting a Solution AI models are used to evaluate proposed solutions, for example, by predicting their future impact (predictive analysis), assessing their chance of success, or predicting a company's reply to action taken by a competitor.

Implementing the Solutions AI can be used to support the implementation of complex solutions. For example, it can be used to demonstrate the superiority of proposals and to assess resistance to changes.

AUTOMATED DECISION MAKING As the power of AI technologies increases, so does its ability to fully automate more and more complex decision-making situations.

Intelligent and Automated Decision Support As early as 1970, there were attempts to automate decision making. These attempts were typically done with the use of rule-based expert systems that provided recommended solutions to repetitive managerial problems. Examples of decisions made automatically include the following:

- Small loan approvals
- Initial screening of job applicants
- Simple restocking
- Prices of products and services (when and how to change them)
- Product recommendation (e.g., at Amazon.com)

The process of automated decision making is illustrated in Figure 2.4. The process starts with knowledge acquisition and creation of a knowledge repository. Users submit questions to the system brain, which generates a response and submits it to the users. In addition, the solutions are evaluated so that the knowledge repository and the reasoning from it can be improved. Complex situations are forwarded to humans' attention. This process is especially used in knowledge-based systems. Note that the process in Figure 2.4 for knowledge acquisition illustrates automatic decision making as well. Companies use automated decision making for both their external operations (e.g., sales) and internal operations (e.g., resource allocation, inventory management). Automated decisions can take several forms, as illustrated in Technology Insight 2.2.

CONCLUSION There is little doubt that AI can change the decision-making process for businesses. The nature of the change varies based on the circumstances. But, in general, we expect AI to have a major impact for making better, faster, and more efficient decisions.

SECTION 2.5 REVIEW QUESTIONS

1. Distinguish between fully automated and supported decision making.
2. List the benefits of AI for decision support.
3. What factors influence the use of AI for decision support?
4. Relate AI to the steps in the classical decision-making process.
5. What are the necessary conditions for AI to be able to automate decision making?
6. Describe Schrage's four models.

TECHNOLOGY INSIGHT 2.2 Schrage's Models for Using AI to Make Decisions

Schrage (2017) of MIT's Sloan School has proposed the following four models for AI to make autonomous business decisions:

1. *The Autonomous Advisor*. This is a data-driven management model that uses AI algorithms to generate best strategies and instructions on what to do and makes specific recommendations. However, only humans can approve the recommendations (e.g., proposed solutions).
Schrage provided an example in which an American retailing company replaced an entire merchandising department with an AI machine, ordering employees to obey directives from it. Obviously, resistance and resentment followed. To ensure compliance, the company had to install monitoring and auditing software.
2. *The Autonomous Outsource*. Here, the traditional business process outsourcing model is changed to a business process algorithm. To automate this activity, it is necessary to create crystal-clear rules and instructions. It is a complex scenario since it involves resource allocation. Correct predictability and reliability are essential.
3. *People-Machine Collaboration*. Assuming that algorithms can generate optimal decisions in this model, humans need to collaborate with the brilliant, but constrained, fully automated machines. To ensure such collaboration, it is necessary to train people to work with the AI machines (see the discussion in Chapter 11). This model is used by tech giants such as Netflix, Alibaba, and Google.
4. *Complete Machine Autonomy*. In this model, organizations fully automate entire processes. Management needs to completely trust AI models, a process that may take years. Schrage provides an example of a hedge fund that trades very frequently based on a machine's recommendations. The company uses machine learning to train the trading algorithms.

Implementing these four models requires appropriate management leadership and collaboration with data scientists. For suggestions of how to do so, consult Schrage (2017), who has written several related books. An interesting note is that some competition among companies will actually occur among data-driven autonomous algorithms and related business models.

QUESTIONS FOR DISCUSSION

1. Differentiate between the autonomous advisor and the people-machine collaboration models.
 2. In all four models, there are some degrees of people-machine interaction. Discuss.
 3. Why it is easier to use model 4 for investment decisions than, for example, marketing strategies?
 4. Why is it important for data scientists to work with top management in autonomous AI machines?
-

2.6 AI APPLICATIONS IN VARIOUS BUSINESS FUNCTIONS

Throughout this book, we provide many examples of AI applications in business, services, and government. In the following five subsections, we provide additional applications in the traditional areas of business: accounting; finance; human resource management; marketing, advertising, and CRM; and production-operation management.

AI IN ACCOUNTING: AN OVERVIEW Major users of AI are the big tax and accounting companies as illustrated in Analytics in Action 2.1.

ACCOUNTING APPLICATIONS IN SMALL FIRMS Small accounting firms also use AI. For example, Crowe Horwath of Chicago is using AI to solve complex billing problems in the healthcare industry. This helps its clients to deal with claims processing and reimbursements. The firm can now solve difficult problems that had previously resisted solutions. Many other applications are used with the support of AI, ranging from analyzing real estate contracts to risk analysis. It is only a question of time before even smaller firms will be able to utilize AI as well.

ANALYTICS IN ACTION 2.1 How EY, Deloitte, and PwC Are Using AI

The big accounting companies use AI to replace or support human activities in tasks such as tax preparation, auditing, strategy consulting, and accountancy services. They mostly use NLP, robotic process automation, text mining, and machine learning. However, they use different strategies as described by Zhou (2017):

- EY attempts to show quick, positive return on investment (ROI) on a small scale. The strategy concentrates on business value. EY uses AI, for example, to review legal documents related to leasing (e.g., to meet new government regulations).
- PricewaterhouseCoopers (PwC) favors small projects that can be completely functioning in four weeks. The objective is to demonstrate the value of AI to client companies. Once demonstrated to clients, the projects are refined. PwC demonstrates 70 to 80 such projects annually.
- Deloitte Touche Tohmatsu Limited, commonly referred to as Deloitte, builds cases that guide AI-based projects for both clients and internal use. The objective is to facilitate innovation. One successful area is the use of NLP for review of large contracts that may include hundreds of thousands of legal documents. The company reduced such review time from six months to less than a month, and it reduced the number of employees who had performed the review by more than 70%. Deloitte, like its competitors, is using AI to evaluate potential procurement synergies for merger and acquisition decisions. Such evaluation is a time-consuming task since it is necessary to check huge quantities of data (sometime millions of data lines). As a result, Deloitte can finish such evaluation in a week compared to the four to five months required earlier. Deloitte said that with AI, it is viewing data in ways never even contemplated before (Ovaska-Few, 2017).

All big accounting companies use AI to assist in generating reports and to conduct many other routine, high-volume tasks. AI has produced high-quality work, and its accuracy has become better and better with time.

QUESTIONS FOR DISCUSSION

1. What are the characteristics of the tasks for which AI is used?
2. Why do the big accounting firms use different implementation strategies?

Sources: Compiled from Chandi (2017), Zhou (2017), and Ovaska-Few (2017).

JOB OF ACCOUNTANTS AI and analytics will automate many routine tasks done today by accountants (see discussion in Chapter 11), some of whom may lose their jobs. On the other hand, accountants will need to manage AI-based accounting systems. Finally, accountants need to drive AI innovation in order to succeed or even survive

AI Applications in Financial Services

Financial services are much diversified, and so is AI usage in the area. One way to organize the AI activities is by major segments of services. In this section, we discuss only two segments: banking and insurance.

AI ACTIVITIES IN FINANCIAL SERVICES Singh (2017) observed the following activities that may be found across various types of financial services:

- Extreme personalization (e.g., using chatbots, personal assistants, and robo investment advisors) (Section 2.9)
- Shifting customer behavior both online and in brick-and-mortar branches
- Facilitating trust in digital identity
- Revolutionizing payments
- Sharing economic activities (e.g., person-to-person loans)
- Offering financial services 24/7 and globally (connecting the world)

AI IN BANKING: AN OVERVIEW The following are some examples of banking institutions use of AI:

- Banks are using AI machines to step up employee surveillance. This is important in preventing illegal activities such as those that occurred at Wells Fargo, the financial services and banking company. For details, see information-management.com/articles/banks-using-algorithms-to-step-up-employee-surveillance.
- Banks use applications for tax preparation. H&R Block is using IBM Watson to review tax returns. The program makes sure that individuals pay only what they owe. Using interactive conversations, the machine attempts to lower people's tax bills.
- Answering many queries in real time. For example, Rainbird Co. (rainbird.ai/) is an AI vendor that trains machines to answer customers' queries. Millions of customers' questions keep bank employees busy. Bots assist staff members to quickly find the appropriate answers to queries. This is especially important in banks where turn-overs of employees are high. Also, there is knowledge degrading overtime, due to frequent changes in policies and regulations.
- At Capital One and several other banks, customers can talk with Amazon's Alexa to pay credit card bills and check their accounts.
- Bank Danamon uses machine learning for fraud detection and anti-money-laundering activities. It also improves the customer experience.
- At HSBC, customers can converse with the virtual banking assistant, Olivia, to find information about their accounts and even learn about security. Olivia can learn from its experiences and become more useful.
- Santander Bank employs a virtual assistant (called Nina) that can transfer money, pay bills, and do more. Nina can also authenticate its customers via an AI-based voice recognition system. Luvo of RBS is a customer service and customer relationship management (CRM) bot that answers customers' queries.

- At Accenture, Collette is a virtual mortgage advisor that provides personalized advice.
- A robot named Nao can analyze facial expression and behavior of customers that enter the branches of certain banks and determine their nationality. Then the machine selects a matching language (Japanese, Chinese, or English) to interact with the customer.

It is worth noting that there appears to be an inherent bias of largely having a female assistant. However, search assistants such as iPhone's Siri or Google Assistant can allow one to select a male voice to respond to queries.

INSURANCE SERVICES Advancements in AI are improving several areas in the insurance industry, mostly in issuing policies and handling claims.

According to Hauari (2017), the major objectives of the AI support are to improve analysis results and enhance customer experience. Incoming claims are analyzed by AI, and, depending on their nature, are sent to appropriate available adjusters. The technologies used are NLP and text recognition. The AI software can help in data collection and analysis and in data mining old claims.

Agents previously spent considerable time asking routine questions from people submitting insurance claims. AI machines, according to Beauchamp (2016), provide speed, accuracy, and efficiency in performing this process. Then AI can facilitate the underwriting process.

Similarly, claims processing is streamlined with the help of AI. It reduces processing time (by up to 90%) and improves accuracy. Capabilities of machine-learning and other AI programs can be shared in seconds in multi-office configurations, including global settings.

AI IN HRM: AN OVERVIEW Savar (2017) points to the following reasons for AI to transform HRM, especially in recruiting: (1) reducing human bias, (2) increasing efficiency, productivity, and insight in evaluating candidates, and (3) improving relationships with current employees.

Wislow (2017) sees the use of AI as a continuation of automation that supports HRM and keeps changing it. Wislow suggests that such automation changes how HRM employees work and are engaged. This change also strengthens teamwork. Wislow divided the impact of AI into the following areas:

Recruitment (Talent Acquisition) One of the cumbersome tasks in HRM, especially in large organizations, is recruiting new employees. The fact is that many job positions are unfilled due to difficulties in finding the right employees. At the same time, many qualified people cannot find the right jobs.

Companies that help recruiters and job seekers, especially LinkedIn, are using AI algorithms to suggest matches to both recruiters and job seekers. Haines (2017) describes the process, noting that a key benefit of this process is the removal of unconscious biases and prejudices of humans.

AI Facilities Training The rapid technological developments make it necessary to train and retrain employees. AI methods can be used to facilitate learning. For example, chatbots can be used as a source of knowledge to answer learners' queries. Online courses are popular with employees. AI can be used to test progress, for example. In addition, AI can be used to personalize online teaching for individuals and to design group lectures.

AI Supports Performance Analysis (Evaluation) AI tools enable HR management to conduct performance analysis by breaking work into many small components and by measuring the performance of each employee and team on each component. The performance is compared to objectives, which are provided to employees and teams. AI also can track changes and progress by combining AI with analytical tools.

AI Use in Retention and Attrition Detection In order to keep employees from leaving, it is necessary for businesses to analyze and predict how to make workers happy. Machine learning can be used to detect reasons why employees leave companies by identifying influencing patterns.

AI IN ONBOARDING Once new employees are hired, the HR department needs help introducing them to the organizational culture and operating processes. Some new employees require much attention. AI helps HRM prepare customized onboarding paths that are best for the newcomers. Results showed that those employees supported by AI-based plans tend to stay longer in organizations (Wislow, 2017).

Using Chatbots for Supporting HRM The use of chatbots in HRM is increasing rapidly. Their ability to provide current information to employees anytime is a major reason. Dickson (2017) refers to the following chatbots: Mya, a recruiting assistant, and Job Bot, which supports the recruitment of hourly workers. This bot is also used as a plug-in to Craigslist.

AI in Marketing, Advertising, and CRM

Compared to other business areas, there are probably more applications of AI in marketing and advertising. For example, AI-based product recommendations have been in use by Amazon.com and other e-commerce companies for more than 20 years. Due to the large number of applications, we provide only a few examples here.

CUSTOMER EXPERIENCES AND CRM As described earlier, a major impact of AI technologies is changing customer experiences. A notable example is the use of conversational bots. Bots (e.g., Alexa) can provide information about products and companies and can provide advice and guidance (e.g., robo advisors for investment; see Section 2.9).

A well-known example of AI in CRM is Salesforce's Einstein.

Example: Salesforce's AI Einstein

Salesforce Einstein is an AI set of technologies (e.g., Einstein Vision for image recognition) that is used for enhancing customer interactions and supporting sales. For example, the system delivers dynamic sales dashboards to sales reps. It also tracks performance and manages teamwork by using sales analytics. The AI product also can provide predictions and recommendations. It supports Salesforce Customer Successful Platform and other Salesforce products.

Einstein's automatically prioritized sales leads make sales reps more productive when dealing with sales leads and potential opportunities. The sales reps also get insights about customers' sentiments, competitors' involvement, and other information.

OTHER USES OF AI IN MARKETING The following show the diversity of AI technologies used in marketing:

- It is used to mimic the expertise of in-store salespeople. In many physical stores, humans are not readily available to help customers who do not want to wait very long. Thus, shopping is made easier when bots provide guidance. A Japanese store already provides all services in a physical store by speaking robots.
- It provides lead generation. As seen in the case of Einstein, AI can help generate sales leads by analyzing customers' data. The program can generate predictions. Insights can be generated by intelligent analytics.
- It can increase customer loyalty using personalization. For example, some AI techniques can recognize regular customers (e.g., in banks). IBM Watson can learn about people from their tweets.

- It can improve the sales pipeline. Narayan (2018) provides a process of how companies can use AI and robots to do this. Specifically, robots convert unknown visitors into customers. Robots use three stages: (1) prepare a list of target customers in the database, (2) send information, ads, videos, and so on to prospects on the list created earlier, and (3) provide the company sales department with a list of leads that successfully convert potential customers to buyers.

AI IN MANUFACTURING To handle ever-increasing labor costs, changes in customers' requirements, increased global competition, and government regulations (Chapter 1), manufacturing companies are using elevated levels of automation and digitization. According to Bollard et al. (2017), companies need to be more agile, and react quicker and more effectively. They also need to be more efficient and improve customers' (organizations' and individuals') experiences. Companies are pressured to cut costs and increase quality and transparency. To achieve these goals, they need to automate processes and make use of AI and other cutting-edge technologies.

Companies have used AI in the form of robots for a long time. Actually, robotics have been used since around 1960 (e.g., Unimate in General Motors). However, the robots were each usually doing one simple task. Today, companies use intelligent robots for complex tasks, enabling make-to-order products and mass customization. In other words, many mental and cognitive tasks are being automated. These developments, involving AI and sensors, allow supporting or even automating production decisions in real time.

Example

When a sensor detects a defective product or a malfunction, the data are processed by an AI algorithm. An action then takes place instantly and automatically. For example, a defective item can be removed or replaced. AI can even make predictions about equipment failures before they occur. This real-time action saves a huge amount of money for manufacturers. (This process may involve the IoT; see Chapter 10.)

LOGISTICS AND TRANSPORTATION AI and intelligent robots are used extensively in corporate logistics and internal and external transportation, as well as in supply chain management. For example, Amazon.com is using over 50,000 robots to move items in its distribution centers (other e-commerce companies are doing the same). Soon, we will see driverless trucks and other autonomous vehicles all over the world.

Example: DHL Supply Chain

DHL is a global delivery company (competing with FedEx and UPS). It has a supply chain division that works with many business partners. AI and IoT are changing the manner by which the company, its partners, and even its competitors operate. DHL is developing innovative logistics and transportation business models, mostly with AI, IoT, and machine learning. These models also help DHL's customers gain a competitive advantage (and this is why the company cannot provide details in its reports).

Several of the IoT projects are linked to machine learning, specifically in the areas of sensors, communication, device management, security, and analysis. Machine learning in such cases assists in tailoring solutions to specific requirements.

Overall, DHL concentrates on the areas of supply chains (e.g., identifies inventories and controls them along the supply chain) and warehouse management. Machine learning and other AI algorithms enable more accurate procurement, production planning, and work coordination. Tagging and tracking items using Radio Frequency Identification (RFID) and Quick Response (QR) code allow for item tracking along the supply chain. Finally, AI facilitates predictive analytics, scheduling, and resource planning.

► SECTION 2.6 REVIEW QUESTIONS

1. What are the major reasons for using AI in accounting?
2. List some applications big accounting firms use.
3. It is said that financial services are more personalized with AI support. Explain.
4. What back-office activities in banks are facilitated by AI?
5. List the activities in recruiting and explain the support provided by AI to each.
6. What are the benefits rewarded to recruiters by AI?
7. How can AI be used to improve CRM?
8. Why use AI in manufacturing?

2.7 INTRODUCTION TO ROBOTICS

Every robotics scientist has their own view about the definition of robot. But a common notion of robot is a machine or a physical device or software that with the cooperation of AI can accomplish a responsibility autonomously. A robot can sense and affect the environment. Applications of robotics in our day-to-day lives have been increasing. This evolution and use of technologies are called the *fourth industrial revolution*. Applications of robotics in manufacturing, health, and information technology (IT) fields in the last decade have led to rapid development in changing the future of industries. Robots are moving from just performing preselected repetitive tasks (**automation**) and being unable to react to unforeseen circumstances (Ayres and Miller, 1981) to performing specialized tasks in healthcare, manufacturing, sports, financial services—virtually every industry. This capability of adaptation to new situations leads to **autonomy**, a sea change from previous generations of robots.

Although our imagination of a robot may be based on the R2D2 or C3-PO from the *Star Wars* movies, we have experienced robots in many other ways. Factories have been using robots for a long time for manufacturing. On the consumer side, an early application was Roomba, a robot that can clean floors on its own. Perhaps the best example of robots that we will all experience soon if not already is an autonomous (self-driving) car. *Tech Republic* called the self-driving car the first robot we will all learn to trust. We will dig a bit deeper into self-driving vehicles in the following pages. With the growth in machine learning, especially image recognition systems, applications of robots are increasing in virtually every industry. Robots can cut sausages into the right size pieces for pizza and can automatically determine that the right number and type of pepperoni pieces have been placed on a pizza before it is baked. Surgeries conducted by and with the assistance of robots are growing at a rapid pace. Section 2.8 provides many illustrative applications of robots. It also discusses self-driving cars as another category of robots.

Sensory systems, such as those for scene recognition and signal processing, when combined with other AI technologies, define a broad category of integrated, possibly complex, systems, generally called *robotics*. There are several definitions of robots, and they are changing over time. A classical definition is this: “A **robot** is an electromechanical device that is guided by a computer program to perform manual and/or mental tasks.” The Robotics Institute of America formally defines a robot as “a programmable multi-functional manipulator designed to move materials, parts, tools, or specialized devices through variable programmed motions for the performance of a variety of tasks.” This definition ignores the many mental tasks done by today’s robots.

An “intelligent” robot has some kind of sensory apparatus, such as a camera, that collects information about the robot’s surroundings and its operations. The collected data are interpreted by the robot’s “brain,” allowing it to respond to the changes in the environment.

Robots can be fully autonomous (programmed to do tasks completely on their own, even repair themselves), or can be remotely controlled by a human. Some robots known as *androids* resemble humans, but most industrial robots are not this type. Autonomous robots are equipped with AI intelligent agents. The more advanced smart robots are not only autonomous but also can learn from their environment, building their capabilities. Some robots today can learn complex tasks by watching what humans do. This leads to better human–robot collaboration. The Interactive Group at MIT is experimenting with this capability by teaching robots to make complex decisions.

Example: Walmart Is Using Robots to Properly Stock Shelves

The efficiency of Walmart stores depends on appropriately stocking their shelves. Using manual labor for checking what is going on is expensive and may be inaccurate. As of late 2017, robots were supporting the company's stocking decisions.

At Walmart, the 2-foot-tall robots use a camera/sensor to scan the shelves to look for misplaced, missing, or mispriced items. The collected information and the interpretation of problems are done by these self-moving robots. The results are transmitted to humans who take corrective actions. The robots carry out their tasks faster and frequently more accurately than humans. Preliminary results are significantly positive and are also expected to increase customer satisfaction. The robots will not cause employees to lose their jobs.

Robots are used extensively in e-commerce warehouses (e.g., tens of thousands are used by Amazon.com). They also are used in make-to-order manufacturing as well as in mass production (e.g., cars), lately of self-driven vehicles. A new generation of robots designed to work as advisors is described in a following section. These robots are already advising on topics such as investments, travel, healthcare, and legal issues. Robots can serve as front desk receptionists and even can be used as teachers and trainers. Analytics in Action 2.2 provides an example of a robot in action in a social setting.

ANALYTICS IN ACTION 2.2 AI/Analytics in Action

Robots Provide Emotional Support to Patients and Children

As discussed later in this chapter, robots have impacted industrial manufacturing and other physical activities. Now, with the research and evolution of AI, robotics can straddle the social world. For example, hospitals today try to give social and emotional support to patients and their families. This support is especially sensitive when offering treatment to children. Children in a hospital are in an unfamiliar environment with medical instruments attached to them, and in many cases, doctors may recommend movement restrictions. The pandemic brought its own restrictions due to quarantine and social distancing rules. This restriction leads to stress, anxiety, and depression in children and consequently in their family members. Hospitals try to provide childcare support specialist or companion pet therapies to reduce the trauma. These therapies prepare children and their parents for future treatment and provide them with temporary emotional support with their interactions. Due to the small number of such specialists, there is a gap between demand and supply for childcare specialists. Also, it is not possible to provide pet therapy at many centers due to the fear of allergies, dust, and bites that may cause the patient's condition to be aggravated. To fill these gaps, the use of social robots is being explored to resolve depression and anxiety among children. A study (Jeong et al., 2015) found that the physical presence of a robot is more effective concerning emotional response as compared to a virtual machine interaction in a pediatric hospital center.

Researchers have known for a long time (e.g., Goris et al., 2010) that more than 60% of human communication is not verbal but rather occurs through facial expressions. Thus, a

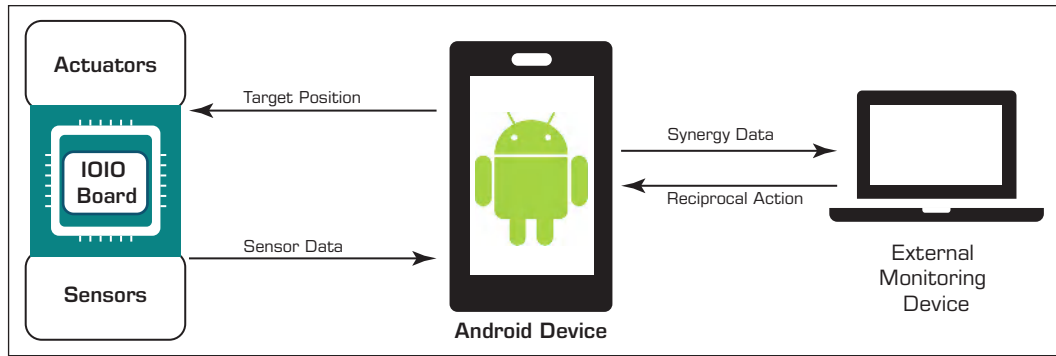


FIGURE 2.5 A General Schematic of a Huggable Robot.

social robot has to be able to provide emotional communication like a child specialist. One popular robot that is providing such support is Huggable. With the help of AI, Huggable is equipped to understand facial expressions, temperament, gestures, and human cleverness. It also uses touch for communication, which increases communication capabilities for those with and without sight. It is like a staff member added to the team of specialists who provide children some general emotional health assistance (Jeong and Logan, 2018).

Huggable looks like a teddy bear having a ringed arrangement. A furry soft body provides a childish look to it and hence is perceived as a friend by the children. With its mechanical arms, Huggable can perform specific actions quickly. Rather than sporting high-tech devices, a Huggable robot is composed of an Android device whose microphone, speaker, and camera are in its internal sensors, and a mobile phone that acts as the central nervous system. The Android device enables the communication between the internal sensors and teleoperation interface. Its segmental arm components enable an easy replacement of sensors and hence increase its reusability. These haptic sensors along with AI enable it to process the physical touch and use it expressively.

Sensors incorporated in a Huggable transmit physical touch and pressure data to the teleoperation device or external device via an IOIO board. The Android device receives the data from the external sensors and transmits them to the motors that are attached to the body of the robot. These motors enable the movement of the robot. The capacitors are placed at various parts of the robot, known as *pressure points*. These pressure points enable the robot to understand the pain of a child who is unable to express it verbally but may be able to touch the robot to convey the pain. The Android device interprets the physical touch and pressure sensor data in a meaningful way and responds effectively. The Android phone enables communication between the other devices while keeping the design minimalistic. The computing power of the robot and the Android device is good enough to allow real-time communication with a child. Figure 2.5 exhibits a schematic of the Huggable robot.

Huggable has been used with children undergoing treatment at Boston Children's Hospital. Reportedly, Aurora, a 10-year-old who had leukemia, was being treated at Dana-Farber/Boston Children's Cancer and Blood Disorder Center. According to Aurora's parents, "There were many activities to do at the hospital but the Huggable being there is great for kids." Beatrice, another child who visits the hospital frequently due to her chronic condition, misses her classes and friends and is unable to do anything that a typical child of her age would do. She was nervous and disliked the process of treatment, but during her interplay with the Huggable, she was more willing to take medicine as if it were the most natural activity to do. She recommended the robot to be a bit faster so that the next time she could play peek-a-boo correctly.

During these interactions with Huggable, children were seen hugging it, holding its hand, tickling it, giving it high-fives, and treating it as someone they need for

support. Children were polite with it and used expressions such as “no, thank you” and “one second, please.” In the end, when bidding it goodbye, one child hugged the Huggable, and another wished to play with it longer.

Another benefit of such emotional support robots is in the prevention of infections. Patients may have contagious diseases, but the robots are sterilized after each use to prevent infection from spreading. Thus, Huggable not only provides support to children but also can be a useful tool for reducing the spread of infectious diseases.

QUESTIONS FOR DISCUSSION

1. What characteristics would you expect to have in a robot that provides emotional support to patients?
2. Can you think of other applications where robots such as the Huggable can play a helpful role?
3. Visit the Website <https://www.universal-robots.com/case-stories/aurolab/> to learn about collaborative robots. How could such robots be useful in other settings?

What We Can Learn from This Analytics in Action 2.2?

As we have seen, AI is opening many interesting and unique applications. The stories about the Huggable introduce us to the idea of using robots for one of the most difficult aspects of work—to provide emotional support to patients, both children and adults. Combinations of technologies such as machine learning, voice synthesis, voice recognition, natural language processing, machine vision, automation, micromachines, and so on make it possible to combine these technologies to satisfy many needs. The applications can come entirely in virtual forms such as IBM Watson, which won the *Jeopardy!* game implementing industrial automation, producing self-driving cars, and even providing emotional support as noted in this Analytics in Action 2.2.

Sources: Broekens, J., Heerink, M., & Rosendal, H. (2009). Assistive social robots in elderly care: A review. *Gerontechnology*, 8, 94–103. doi: 10.4017/gt.2009.08.02.002.00; Fallon, S. (2015). YouTube video at <https://youtu.be/UaRCCA2rRR0> (accessed July 2022); “A Social Robot to Mitigate Stress, Anxiety, and Pain in Hospital Pediatric Care.” *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*; Jeong, S, & Logan, D. (2018, April 21–26).

HISTORY OF ROBOTICS Wikipedia includes an interesting history of robotics. Humans have been fascinated with the idea of machines serving us for a long time. The first idea of robotics was conceptualized in 320 BC when Aristotle, a Greek philosopher, stated, “If every tool, when ordered, or even of its own accord, could do the work that befits it, then there would be no need either of apprentices for the master workers” In 1495, Leonardo Da Vinci drafted strategies and images for a robot that looked like a human. Between 1700 and 1900, various automations were created, including an excellent automation structure built by Jacques De Vaucanson, who made one clockwork duck that could flap its wings, quack, and appear to eat and digest food.

Throughout the industrial revolution, robotics was triggered by the advances in steam power and electricity. As consumer demand increased, engineers strove to devise new methods to increase production by automation and create machines that can perform the tasks that were dangerous for a human to do. In 1893, “Steam Man,” a prototype for a humanoid robot, was proposed by Canadian professor George Moore. It was composed of steel and powered by a steam engine. It could walk autonomously at nearly nine miles per hour and could even pull relatively light loads. In 1898, Nikola Tesla exhibited a submarine prototype. These events led to the integration of robotics in manufacturing, space, defense, aerospace, medicine, education, and entertainment industries.

In 1913, the world's first moving conveyor belt assembly line was started by Henry Ford. With the aid of a conveyor belt, a car could be assembled in 93 minutes. Later in 1920, the term *robot* was coined by Karel Capek in his play *Rossum's Universal Robots*. Then a toy robot, Lilliput, was manufactured in Japan.

By the 1950s, innovators were creating machines that could handle dangerous, repetitive tasks for defense and industrial manufacturing. Since the robots were primarily designed for heavy-duty industries, they were required to pull, lift, move, and push the same way humans did. Thus, many robots were designed like a human arm. Examples include a spray-painting gadget for a position-controlling apparatus by W. L. V. Pollard in 1938. DeVilbiss Company acquired this robot and later became a leading supplier of the robotic arms in the United States.

In the mid-1950s, the first commercial robotic arm, Planetbot, was developed, and General Motors later used it in a manufacturing plant for the production of radiators. A total of eight Planetbots were sold. According to the company, it could perform nearly 25 movements and could be reset in minutes to perform another set of operations. However, Planetbot did not achieve the desired results due to the unusual behavior of the hydraulic fuel inside it.

George Devol and Joe Engelberger designed Unimate to automate the manufacturing of TV picture tubes. It weighed close to 4,000 pounds and was controlled by preprogrammed commands fed on a magnetic drum. Later this was used by General Motors Corporation for production to sequence and stack hot die-cast metal components. This arm with specific upgrades became one of the famous features in assembly lines. A total of 8,500 machines was sold, and half of them went to the automotive industries. Later Unimate was modified to perform spot welding, die casting, and machine tool stacking.

In the 1960s, Ralph Moshier and his team created two remotely operated robotic arms, Handyman and Man-mate. A Handyman was a two-arm electro-hydraulic robot, and the design of the Man-mate's arm was based on the human spine. The arms gave the robots the flexibility for artifact examination procedures. The fingers were designed in a way that they could grasp objects via a single command.

New mobile robots came into the picture. The first one, Shakey, was developed in 1963. It could move freely, avoiding obstacles in its path. A radio antenna was attached to its head. It had a vision sensor atop a central processing unit. Shakey was attached to two wheels, and its two sensors could sense obstacles. Using logic-based problem solving, it could recognize the shape of objects, move them, or go around them.

The space race started by Russia's Sputnik and embraced by the United States led to many technology advances leading to growth of robotics. In 1976, during NASA's mission to Mars, a Viking lander was created for the atmospheric conditions of Mars. Its arms opened out and created a tube to gather samples from the Mars surface. There were some technical issues during the mission, but the scientists were able to fix them remotely.

In 1986, the first LEGO-based educational products were put on the market by Honda. In 1994, Dante II, an eight-legged walking robot built by Carnegie Mellon University, collected the volcanic gas sample from Mount Spur.

Recently, Hanson Robotics, a Hong Kong-based company has introduced Grace, a robot (still under development) that is being promoted to support healthcare, patient well-being, and emotional support. It has reported been tried to provide support to elderly patients during the COVID-19 pandemic to address the loneliness issues that many people faced, particularly during the early stages of the pandemic.

Robotics expanded exponentially as more research and money were invested. Robotic applications and research spread to Japan, Korea, China, and European nations. Indeed, Asia is probably now a bigger user of robots as of this time. Robots have applications in the fields of social support, defense, toys and entertainment, healthcare, food,

and rescue. Many robots are now moving into next stages, going from deep-sea to interplanetary and extrasolar research. And as noted, self-driving cars have brought robots to the masses. We review several robot applications in the following section.

► SECTION 2.7 REVIEW QUESTIONS

1. Define *robot*.
2. Identify some of the key milestones in the history of manufacturing that have led to the current interest in robotics.
3. How would Shakey's capabilities compare to today's robots?
4. How have robots helped with space missions?

2.8 ILLUSTRATIVE APPLICATIONS OF ROBOTICS

This section highlights examples of robot applications in various industries. Each of these is presented as a mini-application case, with the discussion questions presented at the end of the section.

CHANGING PRECISION TECHNOLOGY A mobile production company in China, Changing Precision Technology switched to the use of robotic arms to produce parts for mobile phones. The company previously employed 650 workers to operate the factory. Now, robots perform most of its operations, and the company has reduced its workforce to 60, decreasing the human workforce by 90%. In the future, the company intends to drop its employee count to about 20. With the robots in place, the company not only has achieved an increase in production of 250% but also cut the defect levels from 25% to a mere 5%.

Compiled from Forrest, C. (2015). "Chinese Factory Replaces 90% of Humans with Robots, Production Soars." TechRepublic. <https://www.techrepublic.com/article/chinese-factory-replaces-90-of-humans-with-robots-production-soars/> (accessed July 2022); Javelosa, J., & Houser, K. (2017). "Production Soars for Chinese Factory Who Replaced 90% of Employees with Robots." Future Society. <https://futurism.com/2-production-soars-for-chinese-factory-who-replaced-90-of-employees-with-robots/> (accessed July 2022).

ADIDAS Adidas is a worldwide leading sportswear manufacturer. Keeping trends, innovation, and customization in mind, Adidas has started to automate factories such as Speedfactory in Ansbach, Germany, and Atlanta, Georgia. A conventional supply chain from the raw materials to final product takes around two months, but with automation, it takes just a few days or weeks. The implementation of robotics there was different from that of other manufacturing industries because the raw materials used in shoes manufactured by Adidas are soft textile materials. Adidas is working with the company Oechsler to implement the robotics in its supply chain. Adidas uses technologies such as additive manufacturing, robotic arms, and computerized knitting. At the Speedfactory, the robot that makes a part of a sneaker attaches a scannable QR code to the part. During quality check, if any part of the product turns out to be faulty, the robot that created it is thus traceable and repaired. Adidas has optimized this process, which offers the company the option to roll out a few thousands of customized shoes in the market and see how it performs and optimize the process accordingly. In the next few years, the company plans to roll out around 1 million pairs of the custom styles annually. In the long term, this strategy supports moving from manufacturing large stocks of inventory to creating the products on demand.

Compiled from "Adidas's High-Tech Factory Brings Production Back to Germany." (2017, January 14). *The Economist*. <https://www.economist.com/business/2017/01/14/adidass-high-tech-factory-brings-production-back-to-germany> (accessed July 2022).

BMW EMPLOYS COLLABORATIVE ROBOTS The increased use of AI and automation in industries has resulted in the development of robots. Yet, human cognitive capabilities are irreplaceable. The combination of robots and humans has been achieved using collaborative robots at a BMW manufacturing unit. By doing so, the company has maximized the efficiency of its production unit and modernized the work environment.

BMW's Spartanburg, South Carolina, plant has employed 60 collaborative robots that work side by side with its human workforce. These robots, for example, furnish the interior of BMW car doors with sound and moisture insulation. This sealing protects the electronic equipment that is fixed on the door and the vehicle as a whole from moisture. Previously, human workers performed this intensive task of fixing the foil with the adhesive beads by using a manual roller. With the use of cobots, a robot's arms perform this task with precision. Cobots run on low speed and stop immediately as soon as the sensors detect any obstacle in their way to maintain the safety of assembly-line workers.

At BMW's Dingolfing factory located in Germany, a lightweight cobot is ceiling mounted in the axle transmission assembly area to pick up bevel gears. These gears can weigh up to 5.5 kilos. The cobot fits the bevel gears accurately, avoiding damage to the gear wheels.

Compiled from Allinson, M. (2017, March 4). "BMW Shows Off Its Smart Factory Technologies at Its Plants Worldwide." BMW Press Release. Robotics and Automation. <https://roboticsandautomationnews.com/2017/03/04/bmw-shows-off-its-smart-factory-technologies-at-its-plants-worldwide/11696/> (accessed July 2022); "Innovative Human-Robot Cooperation in BMW Group Production." (2013, October 9). <https://www.press.bmwgroup.com/global/article/detail/T0209722EN/innovative-human-robot-cooperation-in-bmw-group-production?language=en> (accessed July 2022).

TEGA Tega is a social bot intended to provide extended support to preschoolers by engaging them via storytelling and offering help with vocabulary. Like Huggable, Tega is an Android-based robot and resembles an animation character. It has an external camera and onboard speakers and is designed to run for up to six hours before needing a recharge. Tega uses Android capabilities for expressive eyes, computation abilities, and physical movements. Children's response is fed to the Tega as a reward signal into a reinforcement learning algorithm. Tega uses a social controller, sensor processing, and motor control for moving its body and tilting and rotating left or right.

Tega is designed not only to tell stories but also to hold a conversation about the stories. With the help of an app on a tablet, Tega interacts with a child as a peer and teammate, not as an educator. Children communicate with the tablet, and Tega provides the feedback and reactions by watching the children's emotional states. Tega also offers help with vocabulary and understands a child's physical and emotional responses, enabling it to build a relationship with the child. The tests have shown that Tega can positively impact a child's interest in education, free thinking, and mental development. For more information, watch the video at <https://www.youtube.com/watch?v=16in922JTsw> (accessed July 2022).

Compiled from Westlund, J. K., et al. (2016). "Tega: A Social Robot." Video Presentation. *Proceedings of the Eleventh ACM/IEEE International Conference on Human Robot Interaction*; Park, H. W., et al. (2017). "Growing Growth Mindset with a Social Robot Peer." *Proceedings of the Twelfth ACM/IEEE International Conference on Human Robot Interaction*; Personal Robots Group. (2016). <https://www.youtube.com/watch?v=sF0tRCqyTO> (accessed July 2022).

SAN FRANCISCO BURGER EATERY Flipping burgers is considered a low-pay, mundane task that provides many people with employment at a low salary. Such jobs are likely to disappear over time because of robots. One such implementation of robotics in the food industry is at a burger restaurant in San Francisco. The burger-making machine is not a traditional robot sporting arms and legs that can move around and work as a human. Instead, it is a complete burger prep device that can work from prepping a burger for

cooking and bringing together a full meal. It blends the robotic power in bringing the right taste with the help of a Michelin-star chef's recipes and being friendly on the pocket. The restaurant has put in place two 14-foot-long machines that can make around 120 burgers per hour. Each machine has 350 sensors, 20 computers, and close to 7,000 parts.

Buns, onions, tomatoes, pickles, seasoning, and sauces are filled in transparent tubes over a conveyor belt. Once an order is placed via a mobile device, it takes close to five minutes to prepare the order. First, air pressure pushes a burger brioche roll from the transparent tube on the conveyor belt. Different components of the robot work one after the other to prepare the order, from slicing the roll in two halves, applying butter on the bun, shredding vegetables, and dropping the sauces. Also, a light specialized grip is placed on the patty to keep it intact and to bake it per the recipe. With the use of thermal sensors and an algorithm, the cooking time and temperature of the patty are determined, and once cooked, the patty is placed on the bun by a robotic arm. Workers receive a notification via an Apple watch when there is an issue with the machine regarding a malfunction on an order or the need for refills on supplies.

Compiled from "A Robot Cooks Burgers at Startup Restaurant Creator." (2018). TechCrunch. <https://techcrunch.com/video/a-robot-cooks-burgers-at-startup-restaurant-creator/> (accessed July 2022); Zimmeroff, L. (2018, June 21). "A Burger Joint Where Robots Make Your Food." <https://www.wsj.com/articles/a-burger-joint-where-robots-make-your-food-1529599213> (accessed July 2022).

SPYCE Using robots to make affordable foods is demonstrated by a fast-food restaurant operating in Boston that serves grain dishes and salad bowls. Spyce is a budget-friendly restaurant founded by MIT engineering graduates. Michael Farid created the robots that can cook. This restaurant employs few people with good pay and employs robots to do much of the fast-food work.

Orders are placed at a kiosk with touch screens. Once the order is confirmed, the mechanized systems start preparing the food. Ingredients are placed in refrigerated bins that are passed via transparent tubes and are collected using a mobile device that delivers the ingredients to the requested pot. A metal plate attached to the side of the robotic pot heats the food. A temperature of about 450 degrees Fahrenheit is maintained, and the food is tumbled for nearly two minutes and cooked. This resembles clothes being washed in a machine. Once the meal is ready, the robotic pot tilts and transfers the food to a bowl. After each cooking round, the robotic pot washes itself with a high-pressure hot water stream and then returns to its initial position, ready to cook the next meal. The customer name is also added to the bowl. The meal is then served by a human after any final changes. Spyce is also trying to put in place a robot that can cook pancakes.

Compiled from Coxworth, B. (2018, May 29). "Restaurant Keeps Its Prices Down—With a Robotic Kitchen." *New Atlas*. <https://newatlas.com/spyce-restaurant-robotic-kitchen/54818/> (accessed July 2022); Engel, J. (2018, May 3). "Spyce, MIT-Born Robotic Kitchen Startup, Launches Restaurant: Video." Xconomy. <https://www.xconomy.com/boston/2018/05/03/spyce-mit-born-robotic-kitchen-startup-launches-restaurant-video/> (accessed July 2022).

MAHINDRA & MAHINDRA LTD. As the population increases, the agricultural industry is expanding to keep up with demand. To keep increasing the food supply at a reasonable cost and to maintain quality, the Indian multinational firm Mahindra & Mahindra Ltd. is seeking to improve the process of harvesting tabletop grapes. The company is establishing a research and development center at Virginia Polytechnic Institute and University. It will work with other Mahindra centers situated in Finland, India, and Japan.

The grapes can be used for juice, wine, and tabletop grapes. The quality that must be maintained is vastly different for each of these. The ripeness and presentation of

tabletop grapes differ from the other two uses; hence, quality control is critical. Deciding which grapes are ready to pick is a labor-intensive approach, and one must ensure the maturity, consistency, and quality of grapes. Making this decision visually requires expert training, which is not easily scalable. Using robotic harvesting instead of human pickers is being explored. Robots can achieve these goals using sensors that will keep the quality in view while speeding the process.

Compiled from Rosencrance, L. (2018, May 31). "Tabletop Grapes to Get Picked by Robots in India, with Help from Virginia Tech." *Robotics Business Review*. <https://www.roboticsbusinessreview.com/agriculture/tabletop-grapes-picked-robots-india-virginia-tech/> (accessed July 2022).

ROBOTS IN THE DEFENSE INDUSTRY For obvious reasons, the military has invested in robotic applications for a long time. Robots can replace humans in places where risk of loss of human life is too great. Robots can also reach areas where humans may not be able to go due to extreme conditions—heat, water, and so forth. Besides the recent growth of drones in military applications, several specific robots have been developed over a long time. Some are highlighted in the list below.

MAARS MAARS (Modular Advanced Armed Robotic System) is an upgraded version of special weapons observation reconnaissance detection system (SWORDS) robots that were used by the U.S. military during the Iraq war. It is designed for reconnaissance, surveillance, and target acquisition and can have a 360-degree view. Depending on the circumstances, MAARS can drape much firepower into its tiny frame. A variety of ammunition such as tear gas, nonlethal lasers, and grenade launcher can be wrapped in it. MAARS is an army robot that can fight autonomously thus reducing risk to soldiers' lives while also protecting itself. This robot has seven types of sensors to track the heat signature of an enemy during the day and night. It uses night vision cameras to monitor enemy activities during the night. On command, MAARS fires at opponents. Its other uses include moving heavy loads from one place to another. It provides a range of options from nonlethal force such as warning of an attack. It can also form a two-sided communication system. The robot can also use less lethal weapons such as laughing gas, pepper spray, and smoke and start clusters to disperse crowds. The robot can be controlled from about one kilometer and is designed to increase or decrease speed, climb stairs, and walk on nonpaved roads using wheels rather than tracks.

SAFFiR (Shipboard Autonomous Firefighting Robot) Fire on a ship is one of the greatest risks to shipboard life. Shipboard fires have a different and crucial set of problems. Because of the confined space, there are challenges regarding smoke, gas, and limited ability to escape. Even though procedures like fire drills, onboard alarms, fire extinguishers, and other measures provide ways of dealing with fire on the sea, modern technology is in place to tackle this threat in a better way. A U.S. Navy team at the Office of Naval Research has developed SAFFiR. It is a 5 foot 10 inch tall robot. It is not designed to be completely autonomous. It has a humanoid robotic structure so that it can pass through confined aisles and other nooks and corners of a ship and climb ladders. The robot has been designed to work with the obstacles in the passageways in a ship. SAFFiR can use protective fire gear such as fire-protective coats, suppressants, and sensors that are designed for humans. Lightweight and low-friction linear actuators improve its efficiency and control. It is equipped with several sensors: regular camera, gas, and infrared camera for night vision and in black smoke. Its body is designed not only to be fire resistant but also to throw extinguishing grenades. It can work for around half an hour without needing a charge. SAFFiR can also balance itself on an uneven surface.

PEPPER Pepper is a semi-humanoid robot manufactured by SoftBank Robotics that can understand human emotions. A screen is located on its chest. It can identify frowning, tone of voice, smiling, and user actions such as the angle of a person's head and crossed fingers. This way Pepper can determine if a person's mood is good or bad. Pepper can walk autonomously, recognize individuals, and can even lift their mood through its conversation.

Pepper has a height of 120 cms (about 4 feet). It has three directional wheels attached, enabling it to move all around the place. It can tilt its head and move its arms and fingers and is equipped with two high-definition cameras to understand the environment. Because of its anticollision functionalities, Pepper reduces unexpected collisions and can recognize humans as well as obstacles nearby. It can also remember human faces and accepts smart-phone and card payments. Pepper supports commands in Japanese, English, and Chinese.

Pepper is deployed in service industries as well as homes. It has several advantages for effectively communicating with customers but has also been criticized at places for incompetence or security issues. The following examples provide information on its applications and drawbacks:

- Interacting with robots while shopping is changing the face of AI in commercial settings. Nestlé Japan, a leading coffee manufacturer, has employed Pepper to sell Nescafé machines to enhance customer experience. Pepper can explain the range of products Nestlé has to offer and recognize human responses using facial recognition and sounds. Using a series of questions and responses to them, the robot identifies a consumer's need and can recommend the appropriate product.
- Some hotels such as Courtyard by Marriott and Mandarin Oriental are employing Pepper to increase customer satisfaction and efficiency. The hotels use Pepper to increase customer engagement, guide guests toward activities that are taking place, and promote their reward programs. Another goal is to collect customer data and fine-tune the communication according to customer preferences. Pepper was deployed steps away from the entry at Disneyland theme park hotels, and it immediately increased customer interactions. Hotels use Pepper to converse with guests while they are checking in or out or to guide them to the spa, gym, and other amenities. It can also inform guests about campaigns and promotions and help staff members avoid the mundane task of enrolling guests in a loyalty program. Customer reactions are largely quite positive in regard to this.
- Fabio, a Pepper robot, was installed as a retail assistant at an upmarket food and wine store in England and Scotland. A week after implementing it, the store pulled the service because it was confusing customers, and they preferred the service from personal staff rather than Fabio. It provided generic answers on queries such as the shelf location of items. However, it failed to understand completely what the customer was requesting due to background noise. Fabio was provided another chance by placing it in a specific area that attracted only a few customers. Then they also complained about Fabio's inability to move around the supermarket and direct them to a specific section. Surprisingly, the staff at the market became accustomed to Fabio rather than considering it as a competitor.
- Pepper has several security concerns that were pointed out by Scandinavian researchers. According to them, it is easy to have unauthenticated root-level access to the bot. They also found the robot to be prone to brute force attack. Pepper's functions can be programmed using various application programming interfaces (APIs) through languages such as Python, Java, and C + +. This feature can cause it to provide access to all its sensors, making it not secure. An attacker can establish a connection and then use Pepper's mic, camera, and other features to spy on people and their conversations. This is an ongoing issue for many robots and smart speakers.

DA VINCI SURGICAL SYSTEM Over the last decade, the use of robotics has emerged in surgeries. One of the most famous robotic systems used in surgery is the Da Vinci system that has performed thousands of surgeries. According to surgeons, Da Vinci is the most ubiquitous robot used in more units than any other robot. It is designed to perform numerous nominally invasive operations and can perform simple as well as complex and delicate surgeries. The critical components of Da Vinci are the surgeon console, patient side cart, endowrist instruments, and vision system.

The surgeon console is where the surgeon operates the machine. It provides a high-definition, 3D image of the inside of the patient's body. The console has master controls that a surgeon can grasp by the robotic fingers and operate on the patient. The movements are accurate and in real time, and the surgeon is entirely in control and can prevent the robotic fingers from moving by themselves. The patient side cart is the location where the patient resides during the operation. It has either three or four arms attached that the surgeon controls using master controls, and each arm has certain fixed pivot points around which the arms move. The third component is the endowrist instruments, which are available while performing surgery. They have a total of seven degrees of freedom, and each instrument is designed for a specific purpose. Levers can be released quickly for a change of instruments. The last component is a vision system, which has a high-definition, 3D endoscope and image-processing device that provides real-life images of the patient's anatomy. A viewing monitor also helps the surgeon by providing a broad perspective during the process.

Patients who have surgery that used the Da Vinci system heal faster than those performed by traditional methods because the cuts by robotic arms are quite small and precise. A surgeon must undergo online and hands-on training and must perform at least five surgeries in front of a surgeon who is certified to use the Da Vinci system. This technology does increase the cost of the surgery, but its ability to ease pain while increasing precision makes it the future of such procedures.

Compiled from "Da Vinci Robotic Prostatectomy—A Modern Surgery Choice!" (2018). *Robotic Oncology*. <https://www.roboticoncology.com/da-vinci-robotic-prostatectomy/> (accessed July 2022); "The da Vinci® Surgical System." (2015, September). *Da Vinci Surgery*. <http://www.davincisurgery.com> (accessed July 2022).

AGROBOT The combination of sweetness loaded with multiple health benefits makes strawberries one of the world's most popular and consumed fruits. Close to 5 million tons of strawberries are harvested every year, an upward trend in the United States, Turkey, and Spain as top harvesters. AGROBOT, a company engaged in the business of agricultural robots, has developed a robot that can harvest strawberries at any place. Robots using 24 robotic manipulators built on a mobile platform work to identify superior quality strawberries.

Strawberries require a high degree of care because they are delicate compared to other fruits. Fruits such as apples, bananas, and mangoes ripen after being picked whereas strawberries are picked at their full maturity. Hence, harvesting strawberries has been an entirely manual process until recently. AGROBOT was developed in Spain; this robot performs automated processes except selecting the strawberries and packing them. To protect strawberries from being squeezed during picking, the robot cuts them with two razor-sharp blades and catches them in baskets lined with rubber rolls. Once full, the baskets are placed on a conveyor belt and passed to the packing station. Human operators can directly select and pack the berries.

AGROBOT is operated by one person, and a maximum of two people can ride on it. Robotic arms control the coordination between blades and basket. The robot has four main components: inductive sensors, ultrasonic sensors, a collision control system, and a camera system. Camera-based sensors view each fruit and analyze it for ripeness according to its form and color; once a berry is ripe, the robot cuts it from its branches with

precise movements. Each arm is fortified with two inductive sensors to stop at the end positions. The collision control system must be capable of responding to dust, temperature change, vibration, and shock; hence, an ultrasonic sensor is attached to the robot to prevent the arms from touching the ground. Each wheel is equipped with ultrasonic sensors to determine the distance between the strawberry and the robot's current position. These sensors also help in keeping the robot on track and preventing damage to the fruit. Signals received from the sensors are continuously transmitted to an automatic steering system to regulate the position of wheels.

Compiled from "Berry Picking at Its Best with Sensor Technology." Pepperl+Fuchs. <https://www.pepperl-fuchs.com/usa/en/27566.htm> (accessed July 2022); Intorobotics. <https://www.intorobotics.com/35-robots-in-agriculture/> (accessed July 2022).

VARIOUS CATEGORIES OF ROBOTS Robots perform a variety of functions. Depending on these, robots can be categorized into the following categories:

Preset Robots Preset robots are preprogrammed. They have been designed to perform the same task over time and can work 24 hours a day, 7 days a week without any breaks. Preset robots do not alter their behavior. Therefore, these robots have an incredibly low error rate and are suitable for wearisome work. They are frequently used in manufacturing sectors such as the mobile industry, vehicle manufacturing, material handling, and welding to save time and money. Preset robots deliver jobs in environments where it is hazardous for humans to work. Robots move heavy objects, perform assembly tasks, paint, inspect parts, and handle chemicals. A preset robot articulates according to the operation it performs. It can perform a significant role in the medical field because the tasks it performs must have high efficiency at a level comparable to human beings.

Collaborative Robots or Cobots Cobots are the robots that can collaborate with human workers, assisting them to achieve their goals. The use of cobots is trending in the market, and there is an excellent outlook for collaborative robots. According to the survey by MarketsandMarkets, the cobots market in 2020 will be worth around \$3.3 billion. There are various functions of collaborative robots. Depending on the usage, the collaborative robots are used. Collaborative robots have various applications in manufacturing as well as the medical industry.

Stand-Alone Robots Stand-alone robots are the robots that have a built-in AI system and work independently without much interference from humans. These robots perform tasks depending on the environment and adapt to changes in it. With the use of AI, a stand-alone robot learns to modify its behavior and excel in performing its assignment. Autonomous robots have household, military, education, and healthcare applications. They can walk like a human being, avoid obstacles, and provide social-emotional support. Some of these robots are used for domestic purposes as stand-alone vacuum cleaners, such as iRobot Roomba. Stand-alone robots are also used in hospitals to deliver medications, keep track of patients who are yet to receive them, and send this information to the nurses working on that shift and other shifts without chance of any error.

Remote-Controlled Robots Even though robots can perform stand-alone tasks, they do not have human brains; hence, many tasks require human supervision. These robots can be controlled via Wi-Fi, Internet, or satellite. Humans direct remote-controlled robots to perform complicated or dangerous tasks. The military uses these robots to detonate bombs or to act as soldiers around the clock on the battlefield. In the space program research field, their scope of use is extensive. Remote-controlled cobots are also used to perform marginally invasive surgeries.

Supplementary Robots Supplementary robots enhance the existing capabilities or replace capabilities that a human has lost or does not have. This type of robot can be directly attached to a human's body. It connects to a user's body and communicates with the robot's operator directly or when the operator grips the body. The robot can be controlled by a human body, and in some cases, even by thinking of a specific action. Its applications include serving as a robotic prosthetic arm or providing precision for the surgeons. Extensive research on building prosthetic limbs is being conducted.

Autonomous Cars: Robots in Motion

A robot that may eventually touch most people's lives is an autonomous (self-driving) car. Like many other technologies, self-driving cars have been at peak hype recently, but people also recognize their technical, behavioral, and regulatory challenges. Nevertheless, technology and processes are evolving to make the self-driving car a reality in the future, at least in specific settings if not all over the world. Early versions of self-driving cars were enabled by the radio antenna developed in 1925. In 1989, researchers at Carnegie Mellon used neural networks to control an autonomous vehicle. Since then, many technologies have come together to accelerate development of self-driving cars. These include:

- **Mobile phones:** With the help of low-powered computer processors and other accessories such as cameras, mobile phones have become ubiquitous. Many technologies developed for phones, such as location awareness and computer vision, are finding applications in cars.
- **Wireless Internet:** Connectivity has become much more feasible with the rise of 4G networks and Wi-Fi. As 5G deployments move forward, this may enable self-driving cars to allow their processors to communicate with each other in real time.
- **Computer centers in cars:** A number of new technologies are available in today's cars, such as rearview cameras and front and back sensors that help vehicles detect objects in the environment and alert the driver to them or even take necessary actions automatically. For example, adaptive cruise control automatically adjusts the speed of a car based upon the speed of the vehicle in front.
- **Maps:** Navigation maps on mobile phones or navigation systems in cars have made a driver's job easy with regard to navigation. These maps enable an autonomous vehicle to follow a specific path.
- **Deep learning:** With advances in deep learning, the ability to recognize an object is a key enabler of self-driving cars. For example, being able to distinguish a person from an object such as a tree, or whether the object is moving or stationary is critical in taking actions in a moving vehicle.

AUTONOMOUS VEHICLE DEVELOPMENT The heart of an autonomous vehicle system is a laser rangefinder (or light detection and ranging—lidar device), which is on the vehicle's roof. The lidar generates a 3D image of the car's surroundings and then combines it with high-resolution world maps to produce different data models for acting to avoid obstacles and follow traffic rules. In addition, many other cameras are mounted. For example, a camera positioned near a rearview mirror detects traffic lights and takes videos. Before making any navigation decisions, the vehicle filters all data collected from the sensor and camera and builds a map of its surroundings and then precisely locates itself in that map using GPS. This process is called *mapping and localization*.

The vehicle also consists of other sensors such as the four radar devices that are on the front and back bumpers. These devices allow the vehicle to see far distances so that they can make decisions beforehand and deal with fast-moving traffic. A wheel encoder determines the vehicle's location and maintains records of its movements. Algorithms such as neural networks, rule-based decision making, and a hybrid approach are used to

determine the vehicle's speed, direction, and position, and the collected data are used to direct the vehicle on the road to avoid obstacles.

Autonomous vehicles must rely on detailed maps of roads. Thus, before sending driverless cars on roads, engineers drive a route several times and collect data about its surroundings. When driverless vehicles are in operation, engineers compare the data acquired by them to the historical data.

Google's Waymo unit is one of the early pioneers of self-driving vehicles. However, Tesla became the first mass market car to offer near autonomous self-driving cars. To be sure, Tesla vehicles are available as electric vehicles (EV) first and self-driving second. The bigger focus thus far has been in building electric vehicles. Indeed, many traditional and new companies are racing EVs to the market—GM, Ford, Rivian, Lucid, Audi, BMW, etc. At the same time, many of these companies are also building autonomous self-driving capabilities either by themselves or in cooperation with other vendors to offer the self-driving capabilities in the near future.

ISSUES WITH SELF-DRIVING CARS Autonomous cars have been connected to a number of issues.

- **Challenges with technology:** There have been several challenges with the technology used in self-driven cars. Several software and mechanical hurdles are still to be overcome in order to roll out a fully **autonomous car**. For example, companies are still trying to figure out the amount of authority to be transferred when a human driver takes control from an automatic vehicle.
- **Environmental challenges:** Technology and mechanical capabilities cannot yet address many environmental factors affected by self-driving cars. For example, there are still concerns regarding their performance in bad weather. Likewise, several systems have not been tested in extreme conditions such as snow and hail. There are several tricky navigating situations on the road, such as when an animal jumps onto it.
- **Regulatory challenges:** All companies planning to become involved with self-driving cars need to address regulatory hurdles. There are still many unanswered questions about the regulation of autonomous driving. Several questions about liability include these: What will a license involve? Will new drivers be required to get traditional licenses even if they are not drivers? What about young people, or older people with disabilities? What will be required to operate these new vehicles? Governments need to work quickly to catch up with the booming technology. Considering that public safety is on the line, auto regulations should be some of the strictest regulations in the modern world.
- **Public trust issues:** Most people do not yet believe that an autonomous car can keep them safe. Trust and consumer acceptance are the crucial factors. For example, if there is a situation when an autonomous car is being forced to choose between the life of a passenger versus that of a pedestrian, what should be done? Although Tesla has come long ways in demonstrating the power and potential of self-driving cars, consumers still remain a bit skeptical about driverless cars. No technology can be perfect, but the question is which company will be able to best convince its customers to entrust their lives to them.

Advances similar to those for self-driving cars are being explored in other autonomous vehicles. For example, several companies have already launched trials of self-driving trucks. Autonomous trucks, if ever fully deployed, will have a massive disruptive effect on jobs in the transportation industry. Similarly, self-driving tractors are being tested. Finally, autonomous drones and aircrafts are also being developed. These developments will have a huge impact on future jobs while creating other new jobs in the process.

Self-driving vehicles have become part of this world of technology in spite of related technical and regulatory barriers. Autonomous vehicles are yet to achieve the knowledge capabilities of human drivers, but as the technology improves, more-reliable driving vehicles will become a reality. Like many technologies, the short-term impact may be cloudy, but the long-term impact is yet to be determined.

SECTION 2.8 REVIEW QUESTIONS

1. Identify applications of robots in agriculture.
2. How could a social support robot such as Pepper or MEDi be useful in healthcare?
3. Based on the illustrative applications of robots in this section, build a matrix where the rows are the robots' capabilities and the columns are industries. What similarities and differences do you observe across these robots?
4. Identify some key categories of robots.
5. Distinguish between a preset robot and a stand-alone robot. Give examples of each.
6. What are some of the key technology advancements that have enabled the growth of self-driving cars?
7. Give examples of regulatory issues in self-driving cars.
8. Conduct online research to identify the latest developments in autonomous car deployment. Give examples of positive and negative developments.

2.9 CONVERSATIONAL AI—CHATBOTS

As we saw in the last two sections, robots come in several shapes and types. One type that has become popular in recent years is the chatbot. A chatbot, which will be presented in this section, is a conversational robot that is used for chatting with people. (A “bot” is short for “robot.”) Depending on the purpose of the chat, which can be done in writing or by voice, bots can be in the form of intelligent agents that retrieve information or personal assistants that provide advice. In either case, chatbots are usually equipped with NLP that enables conversations in natural human languages rather than in a programmed computer language. Note that Google has rolled out many different voices to its **Google's Assistant**.

Advancement in artificial intelligence (AI) technologies and especially natural language processing (NLP), machine and deep learning and knowledge systems, coupled with the increased quality and functionalities of other intelligent systems, and mobile devices and their apps, have driven the development of chatbots (bots) for inexpensive and fast execution of many tasks related to communication, collaboration, and information retrieval. The use of chatbots in business is increasing rapidly, partly because of their fit with mobile systems and devices.

In the last two to three years, many thousands of bots have been placed into service worldwide by both organizations (private and public) and individuals. Many people refer to these phenomena as the *chatbot revolution*. They are extensively used, for example, in marketing; customer, government, and financial services; healthcare; and in manufacturing. Chatbots make communication more personal than faceless computers and excel in data gathering. Chatbots can stand alone or be parts of other knowledge systems. Obviously, no discussion of chatbot is complete without mentioning ChatGPT. We will cover more details of ChatGPT in Chapter 7.

In this section, we introduce a few examples of chatbots for communication and collaboration, virtual personal assistants (native products, such as Alexa), and chatbots that are used as professional advisors.

What Is a Chatbot?

Short for *chat robot*, a **chatbot**, also known as a “bot” or “robo,” is a computerized service that enables easy conversations between humans and humanlike computerized robots or image characters, sometimes over the Internet. The conversations can be in writing, and more and more are by voice and images. The conversations frequently involve short questions and answers and are executed in a natural language. More intelligent chatbots are equipped with NLPs, so the computer can understand unstructured dialog. Interactions also can occur by taking or uploading images. Some companies experiment with *learning chatbots*, which gain more knowledge with their accumulated experience. The ability of the computer to converse with a human is provided by a knowledge system (e.g., rule-based) and a natural language understanding capability. The service is often available on *messaging services* such as Facebook Messenger or WeChat, and on Twitter.

Chatbot Evolution

Chatbots originated decades ago. They were simple “expert systems” that enabled machines to answer questions posted by users. The first known such machine was Eliza (en.wikipedia.org/wiki/ELIZA, accessed July 2022). Eliza and similar machines were developed to work in Q&A mode. The machine evaluated each question, usually to be found in a bank of FAQs, and generated an answer matched to each question. Obviously, if the question was not in the FAQ collection, the machine provided irrelevant answers. In addition, because the power of the natural language understanding was limited, some questions were misunderstood and the answers were at times at best entertaining. Therefore, many companies opted to use live chats, some with inexpensive labor, organized as call centers around the globe.

Types of Bots Bots can be classified by their capabilities; three classes follow:

1. **Regular bots.** These are essentially conversational intelligent agents. They can do simple, usually repetitive, tasks for their owners, such as showing their bank’s debits, helping them to purchase goods online, and to sell or buy stocks online.
2. **Chatbots.** In this category, we include more capable bots, for example, those that can stimulate conversations with people. This is the major focus of this section.
3. **Intelligent bots.** These have a knowledge base that is improving with experience. That is, these bots can learn, for example, a customer’s preferences (e.g., like Alexa and some robo advisors).

A major limitation of the older types of bots was that updating their knowledge base was both slow and expensive. They were developed for specific narrow domains and/or specific users. It took many years to improve the supporting technology. NLP has become better and better. Knowledge bases are updated today in the “cloud” in a central location; the knowledge is shared by many users so the cost per user is reduced.

The stored knowledge is matched with questions asked by users. The answers by the machines have improved dramatically. Since 2000, we have seen more and more capable AI machines for Q&A dialogs. Around 2010, conversational AI machines were named chatbots and later were developed into virtual personal assistants, championed by Amazon’s Alexa.

Drivers of Chatbots The major drivers are as follows:

- Developers are creating powerful tools to build chatbots quickly and inexpensively with useful functionalities.
- The quality of chatbots is improving, so conversations are getting more useful to users.
- Demand for chatbots is growing due to their potential cost reduction and improved customer service and marketing services, which are provided 24/7.

- Use of chatbots allows rapid growth without the need to hire and train many customer service employees.
- Using chatbots, companies can utilize the messaging systems and related apps that are the darlings for consumers, especially younger ones.

Components of Chatbots and the Process of Their Use

The major components of chatbots are as follows:

- A person (client).
- A computer, avatar, or robot (the AI machine).
- A knowledge base that can be embedded in the machine or available and connected to the “cloud.”
- A human–computer interface that provides the dialog for written or voice modes.
- An NLP that enables the machine to understand natural language.

Advanced chatbots can also understand human gestures, cues, and voice variations.

Person–Machine Interaction Process The components just listed provide the framework for people–bot conversation. Figure 2.6 shows the conversation process.

- A person (left side of the figure) needs to find some information, or need some help.
- The person asks a related question from the bot by voice, texting, and so on.
- NLP translates the question to machine language.
- The chatbot transfers the question to cloud services.
- The cloud contains a knowledge base, business logic, and analytics (if appropriate) to craft a response to the question.
- The response is transferred to a natural language generation program and then to the person who asked the question in the preferred mode of dialog.

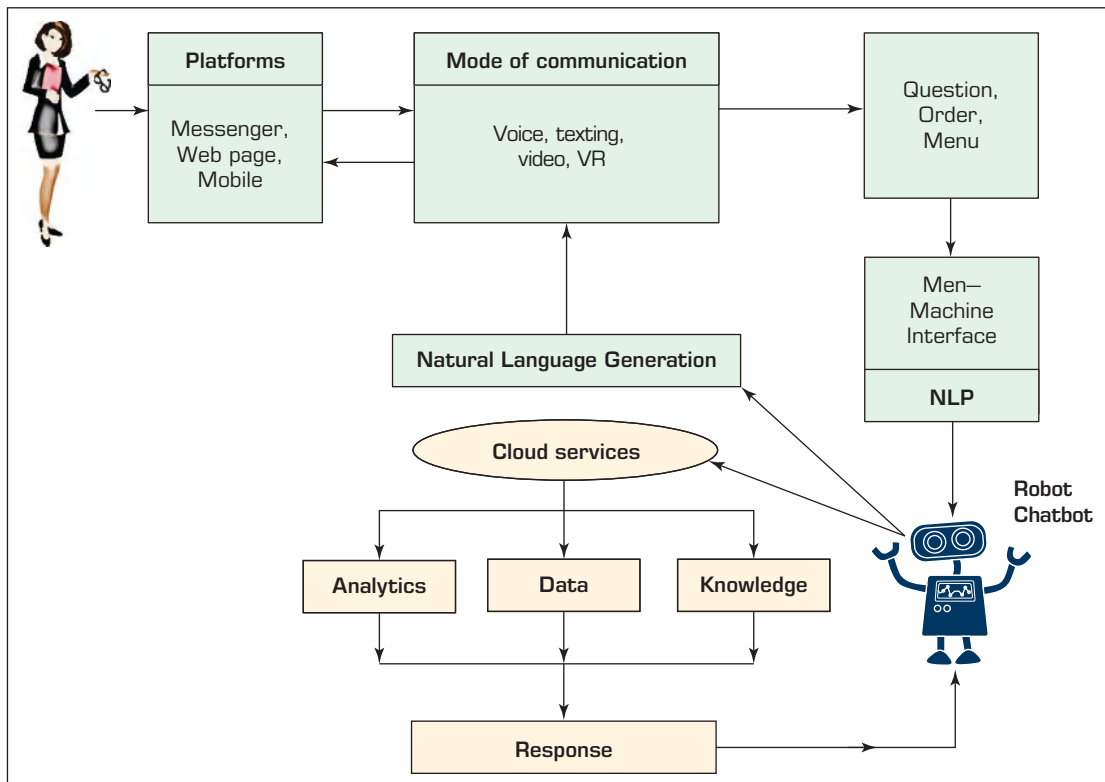


FIGURE 2.6 The Process of Chatting with Chatbots.

Drivers and Benefits

Chatbot use is driven by the following forces and benefits:

- The need to cut costs.
- The increasing capabilities of AI, especially NLP and voice technologies.
- The ability to converse in different languages (via machine translation).
- The increased quality and capability of captured knowledge.
- The push of devices by vendors (e.g., virtual personal assistants such as Alexa from Amazon and Google Assistant from Alphabet).
- Its use for providing superb and economic customer service and conducting market research.
- Its use for text and image recognition.
- Its use to facilitate shopping.
- Its support of decision making.

Chatbots and similar AI machines have been improved over time. Chatbots are beneficial to both users and organizations. For example, several hospitals employ robot receptionists to direct patients to their place of treatment. Zora Robotics created a robot named Nao to act as a chatting companion for people who are sick or at an advanced age. The bot acts, for example, as a form of therapy for those suffering from dementia. Grace, a robot developed by a Hong Kong-based company also provides similar support but can also perform limited medical diagnostics remotely.

Representative Chatbots from around the World

- **RoboCoke.** This is a party and music recommendation bot created for Coca-Cola in Hungary.
- **Kip.** This shopping helper is available on Slack (a messaging platform). Tell Kip what you want to buy, and Kip will find it and even buy it for you.
- **Walnut.** This chatbot can discover skills relevant to you and help you learn them. It analyzes a large set of data points to discover the skills.
- **Ride sharing by Taxi Bot.** If you are not sure whether Uber, Lyft, Grab, or Comfort DelGro is the cheapest service, you can ask this bot. In addition, you can get current promo codes.
- **ShopiiBot.** When you send a picture of a product to this bot, it will find similar ones in seconds. Alternatively, tell ShopiiBot what kind of product you are looking for at what price, and it will find the best one for you.
- **Concerning desired trips.** It can answer questions regarding events, restaurants, and attractions in major destinations.
- **BO.T.** The first Bolivian chatbot, it talks to you (in Spanish) and answers your questions about Bolivia, its culture, geography, society, and more.
- **Hazie.** This is your digital assistant that aims to close the gap between you and your next career move. Job seekers can converse directly with Hazie just as they do with a job placement agent or friends.
- **Green Card.** This Visabot product helps users to properly file requests for Green Cards in the United States.

► SECTION 2.9 REVIEW QUESTIONS

1. Define *chatbots* and describe their use.
2. List the major components of chatbots.
3. What are the major drivers of chatbot technology?
4. How do chatbots work?
5. Why are chatbots considered AI machines?

2.10 ENTERPRISE CHATBOTS

Chatbots play a major role in enterprises, both in external and internal applications. Some believe that chatbots can fundamentally change the way that business is done.

The Interest of Enterprises in Chatbots

The benefits of chatbots to enterprises are increasing rapidly, making dialog less expensive and more consistent. Chatbots can interact with customers and business partners more efficiently, are available anytime, and can be reached from anywhere. Businesses are clearly paying attention to the chatbot revolution.

Enterprise Chatbots: Marketing and Customer Experience

Chatbots can be very useful in providing marketing and customer service (e.g., Mah, 2016), obtaining sales leads, persuading customers to buy products and services, providing critical information to potential buyers, optimizing advertising campaigns (e.g., a bot named Baroj; see Radu, 2016), and much more. Customers want to do business on the app they are already in. For this reason, many bots are on Facebook Messenger, Snapchat, WhatsApp, Kik, and WeChat. Using voice and texting, it is possible to provide personalization as well as superb customer experience. Chatbots can enable vendors to improve personal relationships with customers.

In addition to the marketing areas, plenty of chatbots are in areas such as financial (e.g., banks) and HRM services as well as production and operation management for communication, collaboration, and other external and internal enterprise business processes. In general, enterprises use chatbots on messaging platforms to run marketing campaigns and to provide superb customer experience.

Improving the Customer Experience Enterprise chatbots create improved customer experience by providing a conversation platform for quick and 24/7 contact with enterprises. When customers benefit from the system, they are more inclined to buy and promote a specific brand. Chatbots can also supplement humans in providing improved customer experience.

Coca-Cola

Customers worldwide can chat with Coca-Cola bots via Facebook Messenger. The bots make users feel good with conversations that are increasingly becoming personalized. The bots collect customers' data, including their interests, problems, local dialect, and attitudes and then can target advertisements tailored to each user.

Why Use Messaging Services? So far, we have noted that enterprises are using messaging services such as Facebook Messenger, WeChat, Kik, Skype, and WhatsApp. The reason is that messaging is becoming the most widespread digital behavior. WeChat of China was the first to commercialize its service by offering "chat with business" capabilities.

Facebook's Chatbots Following the example of WeChat, Facebook launched users' conversations with businesses' chatbots on a large scale on Messenger, suggesting that users could message a business just the way they would message a friend. The service allows businesses to conduct text exchanges with users. In addition, the bots have a learning ability that enables them to accurately analyze people's input and provide correct responses. A major benefit of these bots for Facebook is their collection of data and creation of profiles of users.

ANALYTICS IN ACTION 2.3 WeChat's Super Chatbot

WeChat is a very large comprehensive messaging service in China and other countries with about 1.2 billion members in 2022. It pioneered the use of bots as far back as in 2013 (see mp.weixin.qq.com). Users can use the chatbot for activities such as the following:

- Hail a taxi.
- Order food to be delivered.
- Buy movie tickets and other items.
- Customize and order a pair of Nikes.
- Send an order to the nearest Starbucks.
- Track your daily fitness progress.
- Shop Burberry's latest collection.
- Book doctor appointments.
- Pay your water bill.
- Host a business conference call.
- Send voice messages, emoticons, and snapshots to friends.
- Send voice messages to communicate with businesses.
- Communicate and engage with customers.
- Provide a framework for teamwork and collaboration.
- Conduct market research.
- Get information and recommendations on products and services.
- Launch a start-up on WeChat (you can make your own bot on WeChat for this purpose).

Griffiths (2016) provided information concerning a Chinese online fashion flash sales company, Meici. The company used its WeChat account to gather information related to sales. Each time new users followed Meici's account, a welcome message instructed them on how to trigger resources. WeChat is available in English and other languages worldwide due to its usefulness. Facebook installed similar capabilities in 2015.

QUESTIONS FOR DISCUSSION

1. Find some recent activities that WeChat does.
2. What makes this chatbot so unique?
3. Compare the bot of WeChat to bots offered by Facebook.

Enterprise Chatbots: Financial Services

The second area in which enterprise bots are active is financial services. Here we briefly discuss their use in banking.

Example

POSB of Singapore has an AI-driven bot on Facebook Messenger. The bot was created with the help of Kasisto, Inc. of the United States. Using actual Q&A sessions, it took IT workers 11,000 hours to create the bot. Its knowledge base was tested and verified. The bot can learn to improve its performance. Known as POSB digi-bank virtual assistant, the service is accessed via Messenger. Customers save time rather than waiting for human customer service. In the future, the service will be available on other messaging platforms.

Enterprise Chatbots: Service Industries

Chatbots are used extensively in many services. We provide several examples in the following sections.

Healthcare Chatbots are extremely active in the healthcare area, helping millions of people worldwide (Larson, 2016). Here are a few examples:

- Robot receptionists direct patients to departments in hospitals. (Similar services are available at airports, hotels, universities, government offices, and private and other public organizations.)
- Several chatbots are chatty companions for people who are older or sick (e.g., Zora Robotics, Grace by Hanson Robotics).
- Chatbots are used in telemedicine; patients converse with doctors and healthcare professionals who are in different locations. For example, the Chinese company Baidu developed the Melody chatbot for this purpose.
- Chatbots can connect patients quickly and easily with information they need.

Education Chatbot tutors are used in several countries to teach subjects ranging from English (in Korea) to mathematics (in Russia). One thing is certain: The chatbot treats all students equally. Students like the chatbots in online education as well. Machine translation of languages will enable students to take online classes in languages other than their own. Finally, chatbots can be used as private tutors.

Government According to Lacheca (2017), chatbots are spreading in government as a new dialog tool for use by the public. The most popular use is in providing access to government information and answering government-related questions.

Travel and Hospitality Chatbots are working as tour guides in several countries (e.g., Norway). They are not only cheaper (or free) but also may know more than some human guides. Chatbots work as guides in several hotels in Japan. In hotels, they act as concierges, providing information and personalized recommendations (e.g., about restaurants). Chatbots can arrange reservations for hotel rooms, meals, and events. In busy hotels, there is frequently a wait for human concierges; chatbots are available on smartphones all the time. As with other computer services, the chatbots are fast, inexpensive, easy to reach, and always nice. They give excellent customer experience.

Chatbot Platforms

Chatbots inside Enterprises So far we have seen chatbots that are working in the external side of enterprises, mostly in customer care and marketing (e.g., the opening vignette). However, companies lately have started to use chatbots to automate tasks for supporting internal communication, collaboration, and business processes. According to Hunt (2017), “Enterprise and internal chatbots are revolutionizing the way companies do business.” Chatbots in enterprises can do many tasks and support decision-making activities. Chatbots can cut costs, increase productivity, assist working groups, and foster relationships with business partners.

Industry-Specific Bots As we have seen, bots can be specialists (e.g., for investment advice, customer service) or industry-specific experts (e.g., banking, airlines). An interesting bot for the waste industry is Alto (from Bio Hi Tech Global), which enables users to communicate intelligently with industrial equipment. This helps owners of the equipment make decisions that improve performance levels, smooth maintenance routines, and facilitate communication.

AISERA The opening case vignette described an application of a new generation of chatbot built using AISERA’s platform. This genre of technology is better able to learn from previous interactions between a human and the bot and adapt actions/responses accordingly. More importantly, the chatbot can also initiate other actions in the system. This type of integration is leading to a new category of applications termed as Robotic Process Automation (RPA). We will learn more about RPA in Chapter 10.

Knowledge for Enterprise Chatbots

Knowledge for chatbots depends on their tasks. Most marketing and customer care bots require proprietary knowledge, which is usually generated and maintained in-house. This knowledge is similar to that of ES; in many cases, enterprise chatbots operate very similarly to ES except that the interface occurs in a natural language and frequently by voice. On the other hand, chatbots that are used within the enterprise (e.g., to train employees or to provide advice on security or compliance with government regulations) may not be company specific. A company can buy this knowledge and modify it to fit local situations and its specific needs. Newer chatbots use machine learning to extract knowledge from data.

Virtual Personal Assistants

In the previous section, we introduced enterprise chatbots that can be used to conduct conversations. In marketing and sales, they can facilitate customer relationship management (CRM), execute searches for customers, provide information, and execute many specific tasks in organizations for their customers and employees. An emerging type of chatbot is designed as a virtual personal assistant for both individuals and organizations. Known as a **virtual personal assistant (VPA)**, this software agent helps people improve their work, assist in decision making, and facilitate their lifestyle. VPAs are basically extensions of intelligent software agents that interact with people. VPAs are chatbots whose major objective is to help people better perform certain tasks. At this time, millions of people are using Siri with their Apple products, Google Assistant, and Amazon's Alexa. The assistants' knowledge bases are usually universal, and they are maintained centrally in the "cloud," which makes them economical for a large number of users. Users can get assistance and advice from their virtual assistants anytime. In this section, we provide some interesting applications. The first set of applications involves virtual personal assistants, notably Amazon's Alexa and Apple's Siri and Google Assistant.

During the pandemic, the term virtual personal assistant was also used to describe a category of folks who worked remotely but provided services that a personal assistant can provide. Our focus in this section is not on that group, but rather digital assistants which are also termed smart speakers.

If You Were Mark Zuckerberg, Facebook CEO

While Siri and Alexa were in development, Zuckerberg decided to develop his own personal assistant to help him run his home and his work as the CEO of Facebook. He viewed this assistant as Jarvis from *Iron Man*. Zuckerberg trained the bot to recognize his voice and understand basic commands related to home appliances. The assistant can recognize the faces of visitors and monitor the movement of Zuckerberg's young daughter.

The essentials of this assistant can be seen in a 2:13-minute video at youtube.com/watch?v=vvimBPJ3XGQ (accessed July 2022) and one (5:01 minutes) at youtube.com/watch?v=vPoT2vdVkVc (accessed July 2022), with the narration by Morgan Freeman. Today, similar assistants are available for a minimal fee or even for free. The most well-known such assistant is Amazon's Alexa.

Amazon's Alexa and Echo

Of the several virtual personal assistants, the one considered the best in 2018 was Alexa. It was developed by Amazon to compete with Apple's Siri. Alexa works with a smart speaker such as Amazon's **Echo**. Google's smart speaker is a similar product—Google Home. Our goal here is not to compare the various products but to illustrate the general capabilities of this category of intelligent digital assistants.

Amazon's **Alexa** is a cloud-based virtual personal voice assistant or a smart speaker that can do many things such as:

- Answer questions in several domains.
- Control smartphone operations with voice commands.
- Provide real-time weather and traffic updates.
- Control smart home appliances and other devices by using itself as a home automation hub.
- Make to-do lists.
- Arrange music in Playbox.
- Set alarms.
- Play audio books.
- Control home automation devices, as well as home appliances (e.g., a microwave).
- Analyze shopping lists.
- Control a car's devices.
- Deliver proactive notification.
- Shop for its user.
- Make phone calls and send text messages.

Alexa has the ability to recognize different voices, so it can provide personalized responses. Also, it uses a mix of speech and touch to deliver news, hail an Uber, and play games. There are thousands of such applications enabled through Alexa's skills.

Alexa's Skills In addition to the standard (native) capabilities listed, people can use Alexa apps (referred to as *Skills*) to download customized capabilities to Alexa (via your smartphone). Skills are intended to teach Alexa something new.

The following are examples of Alexa's Skills (Apps):

- Call Uber and find the cost of a ride.
- Order a pizza.
- Order take-out meals.
- Obtain financial advice.
- Control home audio visual equipment.

These skills are provided by third-party vendors; they are required to activate invocation commands. There are tens of thousands of them. For example, a person can say, "Alexa, call Uber to pick me up at my office at 4:30 p.m."

Apple's Siri

Siri (short for Speech Interpretation and Recognition Interface) is an intelligent virtual personal assistant and knowledge navigator. It is a part of Apple's several operating systems. It can answer questions, make recommendations, and perform some actions by delegating requests to a set of *Web services* in the "cloud." The software can adapt itself to the user's individual language, search preferences with continuing use, and return personalized results. Siri is available for free to iPhone and iPad users.

Siri can be integrated into Apple's *Siri Remote*. Using CarPlay, Siri is available in some auto brands where it can be controlled by iPhone.

Google Assistant

Competition regarding virtual personal assistants is increasing with the improved capabilities of Google Assistant, which was developed as a competitor to Siri to fit Android smartphones. An interesting demonstration of it is available at youtube.com/watch?v=WTMbF0qYWVs (accessed July 2022); some advanced capabilities are illustrated in the video at youtube.com/watch?v=17rY2ogJQQs (accessed July 2022).

Other Personal Assistants

Several other companies have virtual personal assistants. For example, Microsoft Cortana is well known. In September 2016, Microsoft combined Cortana and Bing. Alexa and Cortana can now work together.

Chatbots as Professional Advisors (Robo Advisors)

The personal assistants described in this section can provide much information and rudimentary advice. A special category of virtual personal assistants is designed to provide personalized professional advice in specific domains. A major area for their activities is investment and portfolio management where robo advisors operate.

Robo Financial Advisors

It is known that the vast majority of “buy” and “sell” decisions of stock trading on the major exchanges, especially by financial institutions, are made by computers. However, computers can also manage an individual’s accounts in a personalized way.

Robo advisors are defined as online providers that offer automated, low-cost, personalized *investment advisory* services, usually through mobile platforms. These robo advisors use algorithms that allocate, deploy, rebalance, and trade investment products. Once enrolled for the robo service, individuals enter their *investment objectives* and preferences. Then, using advanced AI algorithms, the robo will offer *alternative personalized* investments for individuals to choose from funds or exchange-traded funds [ETFs]. By conducting a dialog with the robo advisor, an AI program will refine the investment portfolio. This is all done digitally without having to talk to a live person.

Evolution of Financial Robo Advisors

The pioneering emergence of Betterment Inc. in 2010 was followed by several other companies. Now virtually every major financial service firm offers robo adviser services. There is no question that robo advisors are game-changing phenomena for the wealth management business, even though their performance so far has not been much different from that of traditional, manual, and financial services.

Robo advising companies try to cut costs by using ETFs, whose commission fees are significantly lower than that of mutual funds. Annual fees vary as does the minimum amount of required assets. Premium services are more expensive since they offer the opportunity to consult human experts (advisors 2.0).

Example: Smart Assistant Shopping Bots

Shopping bots ask a few questions to understand what a customer needs and prefers. Then they recommend the best match for the customer. This makes customers feel they are receiving personalized service. The assistance simplifies the customer’s decision-making process. Smart assistants also offer advice on issues of concern to customers via Q&A conversations. Note that these bots are essentially recommendation systems and that users need to ask for advice whereas other recommendation systems (e.g., that of Amazon.com) provide advice even when users do not ask for it.

► SECTION 2.10 REVIEW QUESTIONS

1. Describe an intelligent virtual personal assistant.
2. Describe the capabilities of Amazon’s Alexa.
3. Relate Amazon’s Alexa to Echo.
4. Describe Echo Dot and Tap.

5. Describe Apple's Siri and Google's Assistant.
6. How is the knowledge of personal assistants maintained?
7. Explain the relationship between virtual personal assistants and chatbots.
8. Define *robo advisor*.
9. Explain how robo advisors work for investments.
10. Discuss some of the shortcomings of robo advisors for investments.

Chapter Highlights

- The aim of artificial intelligence is to make machines perform tasks intelligently, possibly like people do.
- A major reason for using AI is to cause work and decision making to be easier to perform. AI can be more capable (enable new applications and business models), more intuitive, and less threatening than other decision support applications.
- A major reason to use AI is to reduce cost and/or increase productivity.
- AI systems can work autonomously, saving time and money, and perform work consistently. They can also work in rural and remote areas where human expertise is rare or not available.
- AI can be used to improve all decision-making steps.
- Intelligent virtual systems can act as assistants to humans.
- AI systems are computer systems that exhibit low (but increasing) levels of intelligence.
- AI has several definitions and derivatives, and its importance is growing rapidly. The U.S. government postulated that AI will be a "critical driver of the U.S. economy" (Gaudin, 2016).
- The major technologies of AI are intelligent agents, machine learning, robotic systems, NLP and speech recognition, computer vision, and knowledge systems.
- Expert systems, recommendation systems, chatbots, and robo advisors are all based on knowledge transferred to machines.
- The major limitations of AI are the lack of human touch and feel, the fear that it will take jobs from people, and the possibility that it could be destructive.
- AI is not a match to humans in many cognitive tasks, but it can perform many manual tasks quicker and at a lower cost.
- There are several types of intelligence, so it is difficult to measure AI's capacity.
- In general, human intelligence is superior to that of machines. However, machines can beat people in complex games.
- Machine learning is currently the most useful AI technology. It attempts to learn from its experience to improve operations.
- Deep learning enables AI technologies to learn from each other, creating synergy in learning.
- Intelligent agents excel in performing simple tasks considerably faster and more consistently than humans (e.g., detecting viruses in computers).
- The major power of machine learning is a result of the machine's ability to learn from data and its manipulation.
- Deep learning can solve many difficult problems.
- Computer vision can provide understandings from images, including from videos.
- Robots are electromechanical computerized systems that can perform physical and mental tasks. When provided with sensory devices, they can become intelligent.
- Industrial automation brought the first wave of robots, but now the robots are becoming autonomous and finding applications in many areas.
- Robotic applications span industries such as agriculture, healthcare, and customer service.
- Social robots are emerging as well to provide care and emotional support to children, patients, and older adults.
- Collaborative robots are evolving quickly, leading to a category called *cobots*.
- Autonomous cars are probably the first category of robots to touch most consumers.
- Self-driving cars are challenging the limits of AI innovation and legal doctrines.
- Computers can understand human languages and can generate text or voice in human languages.
- Cognitive computing simulates the human thought process for solving problems and making decisions.

- Computers can be fully automated in simple manual and mental tasks using AI.
- Several types of decision making are fully automated using AI; other types can be supported.
- AI is used extensively in all functional business departments, reducing cost and increasing productivity, accuracy, and consistency. There is a tendency to increase the use of chatbots. They all support decision making well.
- AI is used extensively in accounting, automating simple transactions, helping deal with Big Data, finding fraudulent transactions, increasing security, and assisting in auditing and compliance.
- AI is used extensively in financial services to improve customer service, provide investment advice, increase security, and facilitate payments among other tasks. Notable applications are in banking and insurance.
- HRM is using AI to facilitate recruitment, enhance training, help onboarding, and streamline operations.
- There is considerable use of AI in marketing, sales, and advertising. AI is used to support product recommendation, help in search of products and services, facilitate Web site design, support pricing decisions, provide language translation in globe trade, assist in forecasting and predictions, and use chatbots for many marketing and customer service activities.
- AI has been used in manufacturing for decades. Now it is applied to support planning, supply chain coordination, logistics and transportation, and operation of intelligent factories.
- Chatbots can save organizations money, provide a 24/7 link with customers and/or business partners, and are consistent in what they say.
- An expert system was the first commercially applied AI product.
- We distinguish three major types of chatbots: enterprise, virtual personal assistants, and robo advisors.
- A relatively new application of knowledge systems is the virtual personal assistant. Major examples of such assistants are Amazon's Alexa, Apple's Siri, and Google's Assistant.
- Knowledge for virtual personal assistants is centrally maintained in the "cloud" and it is usually disseminated via a Q&A dialog.
- Chatbots contain a knowledge base and a natural language interface.
- Chatbots can facilitate online shopping by providing information and customer service.
- Robo advisors provide personalized online investment advice at a much lower cost than human advisors. So far, the quality seems to be comparable.

Key Terms

artificial intelligence (AI)	computer vision	robot	Alexa
augmented intelligence	deep learning	sensor	Echo
automation	intelligent agent (IA)	social robot	Google's Assistant
autonomous car	machine learning	scene recognition	robo advisors
autonomy	machine vision	speech (voice)	Siri
chatbots	natural language processing (NLP)	understanding	virtual personal assistant (VPA)
		Turing Test	

Questions for Discussion

1. Discuss the difficulties in measuring the intelligence of machines.
2. Discuss the process that generates the power of AI.
3. Discuss the differences between machine learning and deep learning.
4. Describe the difference between machine vision and computer vision.
5. How can a vacuum cleaner be as intelligent as a six-year-old child?
6. Why are NLP and machine vision so prevalent in industry?
7. Why are chatbots becoming very popular?
8. Discuss the advantages and disadvantages of the Turing Test.
9. Why is augmented reality related to AI?
10. Discuss the support that AI can provide to decision makers.
11. Discuss the benefits of automatic and autonomous decision making.

12. Why is general (strong) AI considered to be “the most significant technology ever created by humans”?
13. Why is the cost of labor increasing, whereas the cost of AI is declining?
14. If an artificial brain someday contains as many neurons as the human brain, will it be as smart as a human brain? (Students need to do extra research.)
15. Distinguish between single-task robots and intelligent ones.
16. Discuss why applications of natural language processing and computer vision are popular and have many uses.
17. Some people say that chatbots are inferior for chatting. Others disagree. Discuss.
18. Discuss the financial benefits of chatbots.
19. Discuss how IBM Watson will reach 1 billion people by 2018 and what the implications of that are.
20. Discuss the limitation of chatbots and how to overcome them.
21. Discuss what made ES popular for almost 30 years before their decline.
22. Summarize the difficulties in knowledge acquisition from experts.
23. Compare the ES knowledge-refining system with knowledge improvement in machine learning.
24. Discuss the difference of enterprises’ use of chatbots internally and externally.
25. Some people say that without a virtual personal assistant, a home cannot be smart. Why?
26. Compare Facebook Messenger virtual assistant project M with that of competitors.
27. Examine Alexa’s skill in ordering drinks from Starbucks.
28. Discuss the advantages of robo advisors over human advisors. What are the disadvantages?
29. Explain how marketers can reach more customers with bots.
30. Are robo advisors the future of finance? Debate; start with Demmissie (2017).

Exercises

1. Go to sitezeus.com and view the 2:07-minute video. Explain how the technology works as a decision helper.
2. Go to Investopedia and learn about investors’ tolerance. Then find out how AI can be used to contain this risk, and write a report.
3. In 2017, McKinsey & Company created a five-part video titled “Ask the AI Experts: What Advice Would You Give to Executives About AI?” View the video and summarize the advice given to the major issues discussed. (Note: This is a class project.)
4. Watch the McKinsey & Company video (3:06 minutes) on today’s drivers of AI at youtube.com/watch?v=yv0IG1D-OdU and identify the major AI drivers. Write a report.
5. Go to the Web site of the Association for the Advancement of Artificial Intelligence aaai.org/home.html and describe its content. Compare it to that of csail.mit.edu/.
6. Go to crosschx.com and find information about Olive. Explain how it works, what its limitations and advantages are, and which types of decisions it automates and which it only supports.
7. Go to waze.com and moovitapp.com and find their capabilities. Summarize the help they can provide users.
8. Find recent information on research that is aimed to measure artificial intelligence. Write a report.
9. Go to salesforce.com and find recent developments on AI Einstein. Why it is so popular?
10. Find the latest information on IBM Watson’s advising activities. Write a report.
11. Find information on the use of AI in iPhones. Explore the role of Edge AI. Write a report.
12. Explore the AI-related products and services of Nuance Inc. (nuance.com). Explore the Dragon voice recognition product.
13. Go to salesforce.com and investigate the capabilities of Gecko HRM. Relate it to Salesforce Einstein. Provide examples of two applications.
14. Enter McKinsey & Company and find in its Fifty Five “The Value AI Can Bring to Your Business” (mckinsey.com/featured-insights/artificial-intelligence/five-fifty-real-world-ai). Then look for “Real-World AI.” Find the banking section and dive more deeply into its content.
15. Find material on the impact of AI on advertising. Write a report.
16. Identify applications other than those discussed in this chapter where Pepper is being used for commercial and personal purposes.
17. Conduct online research to find at least one new robotics application in agriculture. Prepare a brief summary of your research: the problem addressed, technology summary, results achieved if any, and lessons learned.
18. Conduct online research to find at least one new robotics application in healthcare. Prepare a brief summary of your research: the problem addressed, technology summary, results achieved if any, and lessons learned.
19. Conduct online research to find at least one new robotics application in customer service. Prepare a brief summary of your research: the problem addressed, technology summary, results achieved if any, and lessons learned.
20. Conduct online research to find at least one new robotics application in an industry of your choice. Prepare a

brief summary of your research: the problem addressed, technology summary, results achieved if any, and lessons learned.

21. Conduct research to identify the most recent developments in self-driving cars.
22. Conduct research to learn and summarize any new investments and partnerships in self-driving cars.
23. Compare the chatbots of Facebook and WeChat. Which has more functionalities?
24. Enter nuance.com and find information about Dragon Medical Advisor. Describe its benefits. Write a report.
25. Enter chatbots.org/ and join a forum of your interest. Also explore research issues of your interest. Write a report.
26. There is intense competition between all major tech companies regarding their virtual personal assistants. New innovations and capabilities appear daily. Research the status of these assistants for Amazon, Apple, Microsoft, Google, and Samsung. Write a report.
27. Some people believe that chatbots will change how people interact with the Internet and browse online. Prepare a report regarding this.
28. Explain why is Amazon's Echo needed to work with Alexa? Read howtogeek.com/253719/do-i-need-an-amazon-echo-to-use-alexa/. Write a report.
29. Find out how Simon Property Group is using chatbots across over 200 shopping malls. Write about the benefits to different types of users and to the company.
30. Read recent information about enterprise bots. Write a report.
31. Enter gravityinvestments.com/digital-advice-platform-demo. Would you invest in this project? Research and write a report.
32. Enter visirule.co.uk and find all products it has for expert systems. List them and write a short report.
33. Research the role of chatbots in helping patients with dementia.
34. Find information on the Baidu's Melody chatbot and how it works with Baidu Doctor.
35. Pose a question related to a chatbot on quora.com. Summarize the answers received in a report.
36. Nina is an intelligent chatbot from Nuance Communication Inc. that works for Alexa Internet of Things (IoT), smart homes, and more. Find information and write a report about Nina's capabilities and benefits.
37. Microsoft partners with the government of Singapore to develop chatbots for e-services. Find out how this is done.
38. Study the Tommy Hilfiger Facebook Messenger bot. Find out how it is (and was) used in the company's marketing campaigns.
39. Two comprehensive building tools for chatbots are Botsify and Personality Forge (personalityforge.com). Compare the tools. Write a report.
40. It is time now to build your own bot. Consult with your instructor about which software to use. Have several bots constructed in your class and compare their capabilities. Use Microsoft's Azure if you have some programming experience.

References

- Allinson, M. (2017, March 4). "BMW Shows Off Its Smart Factory Technologies at Its Plants Worldwide." *Robotics and Automation*. <https://roboticsandautomationnews.com/2017/03/04/bmw-shows-off-its-smart-factory-technologies-at-its-plants-worldwide/11696/> (accessed July 2022).
- Ayres, R., & Miller, S. (1981, November). "The Impacts of Industrial Robots." Report CMU-RI-TR-81-7. Pittsburgh, PA: The Robotics Institute at Carnegie Mellon University.
- Bollard, A., et al. (2017). "The next-generation operating model for the digital world." *McKinsey & Company*, March 2017.
- Brualdi Timmins, Amy C. (1996) "Multiple Intelligences: Gardner's Theory," *Practical Assessment, Research, and Evaluation*: Vol. 5, Article 10. DOI: <https://doi.org/10.7275/7251-ea02> <https://scholarworks.umass.edu/pare/vol5/iss1/10> (accessed January 2023).
- Beauchamp, P. (2016). "Artificial Intelligence and the Insurance Industry: What You Need to Know." *The Huffington Post*, October 27, 2016.
- Coxworth, B. (2018, May 29). "Restaurant Keeps Its Prices Down – With a Robotic Kitchen." *New Atlas*. <https://newatlas.com/spyce-restaurant-robotic-kitchen/54818/> (accessed July 2022).
- Carney, P. (2018). "Pat Carney: Artificial Intelligence versus Human Intelligence." *Vancouver Sun*, April 7, 2018.
- Chandi, N. (2017). "How AI Is Reshaping the Accounting Industry." *Forbes.com*, July 20, 2017.
- Demmissie, L. (2017). "Robo Advisors: The Future of Finance." *Huffpost.com*. https://www.huffpost.com/entry/robo-advisors-the-future-of-finance_b_586fb1eee4b0eb9e49bfba9b (accessed February 2023).
- Dickson, B. (2017). "How Artificial Intelligence Optimizes Recruitment." *The Next Web*, June 3, 2017.
- Engel, J. (2018, May 3). "Spyce, MIT-Born Robotic Kitchen Startup, Launches Restaurant: Video." *Xconomy*. <https://www.xconomy.com/boston/2018/05/03/spyce-mit-born-robotic-kitchen-startup-launches-restaurant-video/> (accessed July 2022).
- Fallon, S. (2015). "A Blue Robotic Bear to Make Sick Kids Feel Less Blue." YouTube video at <https://youtu.be/UaRC-CA2rRR0> (accessed July 2022).
- Forrest, C. (2015). "Chinese Factory Replaces 90% of Humans with Robots, Production Soars." *TechRepublic*.

- <https://www.techrepublic.com/article/chinese-factory-replaces-90-of-humans-with-robots-production-soars/> (accessed July 2022).
- Goris, K., et al. (2010, September). "Mechanical Design of the Huggable Robot Probo." Robotics & Multibody Mechanics Research Group. Brussels, Belgium: Vrije Universiteit Brussels.
- Gaudin, S. (2016). "White House: A.I. Will Be Critical Driver of U.S. Economy." *Computerworld*, October 12, 2016.
- Griffiths, T. (2016). "Using Chatbots to Improve CRM Data: A WeChat Case Study." *Half a World*, November 16, 2016.
- Haines, D. (2017). "Is Artificial Intelligence Making It Easier and Quicker to Get a New Job?" *Huffington Post UK*, December 4, 2017.
- Huari, G. (2017). "Insurers Leverage AI to Unlock Legacy Claims Data." *Information Management*, July 3, 2017.
- Hughes, T. (2016). "Google DeepMind's Program Beat Human at Go." *USA Today*, January 27, 2016.
- Hunt, M. (2017). "Enterprise Chatbots and the Conversational Commerce Revolutionizing Business." *Entrepreneur*, July 3, 2017.
- Jeong, S., et al. (2015). "A Social Robot to Mitigate Stress, Anxiety, and Pain in Hospital Pediatric Care." *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*.
- Jeong, S., & Logan, D. (2018, April 21–26). "Huggable: The Impact of Embodiment on Promoting Socio-emotional Interactions for Young Pediatric Surgeons." MIT Media Lab, Cambridge, MA, CHI 2018, Montréal, QC, Canada.
- Kharpal, A. (2017). "A.I. Is in a 'Golden Age' and Solving Problems That Were Once in the Realm of Sci-Fi, Jeff Bezos Says." *CNBC News*, May 8, 2017.
- Kolbjørnsrud, V., Amico, R., & Thomas. (2016). "How Artificial Intelligence Will Redefine Management." *Harvard Business Review*, November 2, 2016.
- Lacheca, D. (2017). "Conversational AI Creates New Dialogues for Government." *eGovInnovation*, October 24, 2017.
- Larson, S. (2016). "Baidu Is Bringing AI Chatbots to Healthcare." *CNNTech*, October 11, 2016.
- Mah, P. (2016). "The State of Chatbots in Marketing." *CMOInnovation*, November 4, 2016.
- Marr, B. (2018). "The Key Definitions of Artificial Intelligence That Explain Its Importance." *Forbes*, February 14, 2018.
- Narayan, K. (2018). "Leverage Artificial Intelligence to Build your Sales Pipeline." *LinkedIn*, February 14, 2018.
- Ng, A. (2016). "What Artificial Intelligence Can and Can't Do Right Now." *Harvard Business Review*, November 9, 2016.
- Ovaska-Few, S. (2017). "How Artificial Intelligence Is Changing Accounting." *Journal of Accountancy*, October 9, 2017.
- Padmanabhan, G. (2018). "Industry-Specific Augmented Intelligence: A Catalysts for AI in the Enterprise." *Forbes*, January 4, 2018.
- Press, G. (2017). "Top 10 Hot Artificial Intelligence (AI) Technologies." *Forbes*, January 23, 2017.
- Radu, M. (2016). "How to Pay Less for Advertising? Use Baro—An Ad Robot for Campaigns Optimization." *150sec.com*, August 18, 2016.
- Savar, A. (2017). "3 Ways That A.I. Is Transforming HR and Recruiting." *INC.com*, June 26, 2017.
- Schrage, M. "4 Models for Using AI to Make Decisions." *Harvard Business Review*, January 27, 2017.
- Singh, H. "How Artificial Intelligence Will Transform Financial Services." *Information Management*, June 6, 2017.
- Spacey, J. (2016). "33 Types of Artificial Intelligence," October 30, 2016, <https://simplicable.com/IT/types-of-artificial-intelligence> (accessed January 2023).
- Staff. "Assisted, Augmented and Autonomous: The 3 Flavours of AI Decisions." *Software and Technology*, June 28, 2017. tgdaily.com/technology/assisted-augmented-and-autonomous-the-3-flavours-of-ai-decisions.
- Taylor, P. "Welcome to the Machine – Learning." *Forbes BrandVoice*, June 3, 2016. forbes.com/sites/sap/2016/06/03/welcome-to-the-machine-learning/#3175d50940fe (accessed July 2022).
- USC. "AI Computer Vision Breakthrough IDs Poachers in Less Than Half a Second." *Press Release*, February 8, 2018.
- Violino, B. "Most Firms Expect Rapid Returns on Artificial Intelligence Investments." *Information Management*, November 1, 2017.
- Wislow, E. "5 Ways to Use Artificial Intelligence (AI) in Human Resources." *Big Data Made Simple*, October 24, 2017. <https://wislow1.rssing.com/chan-75761987/article2.html> (accessed July 2022).
- Westlund, J. K., Lee, J. M., Plummer, J., Faridia, L., Gray, F., Berlin, J., Quintus-Bosz, M., Harmann, H., Hess, R., Dyer, M., dos Santos, S. Adalgeirsson, K., Gordon, S., Spaulding, G., Martinez, S. Das, M., Archie, M. Jeong, M., & Breazeal, C. (2016). "Tega: A Social Robot." Video Presentation. *Proceedings of the Eleventh ACM/IEEE International Conference on Human Robot Interaction*.
- Zimberoff, L. (2018, June 21). "A Burger Joint Where Robots Make Your Food." <https://www.wsj.com/articles/a-burger-joint-where-robots-make-your-food-1529599213> (accessed July 2022).
- Zarkadakis, G. *In Our Own Image: Savior or Destroyer? The History and Future of Artificial Intelligence*. New York, NY: Pegasus Books, 2016.
- Zhou, A. (2017). "EY, Deloitte and PwC Embrace Artificial Intelligence for Tax and Accounting." *Forbes.com*, November 14, 2017.

Descriptive Analytics I: Nature of Data, Big Data, and Statistical Modeling

LEARNING OBJECTIVES

- Understand the nature of data as it relates to business intelligence (BI) and analytics
- Learn the methods used to make real-world data analytics ready
- Learn what Big Data is and how it is changing the world of analytics
- Understand the motivation for and business drivers of Big Data analytics
- Become familiar with the wide range of enabling technologies for Big Data analytics
- Learn about Hadoop, Spark, MapReduce, and NoSQL as they relate to Big Data analytics
- Become familiar with the Data for Good concept
- Understand the need for and appreciate the capabilities of stream analytics
- Learn about the applications of stream analytics
- Describe statistical modeling and its relationship to business analytics
- Learn about descriptive and inferential statistics

In the age of Big Data and business analytics in which we are living, the importance of data is undeniable. The newly coined phrases like “data is the oil,” “data is the new bacon,” “data is the new currency,” and “data is the ruler” are further stressing the renewed importance of data. But what type of data are we talking about? Obviously, not just any data. The “garbage in garbage out—GIGO” concept/principle applies to today’s “Big Data” phenomenon more so than any data definition that we have had in the past. To live up to its promise, its value proposition, and its ability to turn into insight, data has to be carefully created/identified, collected, integrated, cleaned, transformed, and properly contextualized for use in accurate and timely decision making.

Data is the main theme of this chapter. Accordingly, the chapter starts with a description of the nature of data: what it is, what different types and forms it can come in, and how it can be preprocessed and made ready for analytics. The first few sections of the chapter are dedicated to a deep yet necessary understanding and processing of data. Following the fundamental concepts and the nature of data along with data preprocessing, the chapter continues with sections on big data, its definition, related concepts, tools and technologies. The last few sections describe the statistical methods used to prepare data as input to produce both descriptive and inferential measures.

- 3.1** Opening Vignette: SiriusXM Attracts and Engages a New Generation of Radio Consumers with Data-Driven Marketing 118
- 3.2** The Nature of Data in Analytics 121
- 3.3** A Simple Taxonomy of Data 125
- 3.4** The Art and Science of Data Preprocessing 127
- 3.5** Definition of Big Data 136
- 3.6** Fundamentals of Big Data Analytics 140
- 3.7** Big Data Technologies 143
- 3.8** Big Data and Stream Analytics 153
- 3.9** Statistical Modeling for Business Analytics 157
- 3.10** Regression Modeling for Inferential Statistics 169

3.1 OPENING VIGNETTE: SiriusXM Attracts and Engages a New Generation of Radio Consumers with Data-Driven Marketing

SiriusXM Radio is a satellite radio powerhouse, the largest radio company in the world with \$3.8 billion in annual revenues and a wide range of hugely popular music, sports, news, talk, and entertainment stations. The company, which began broadcasting in 2001 with 50,000 subscribers, grew to 18.8 million subscribers in 2009, and today has nearly 29 million.

Much of SiriusXM's growth to date is rooted in creative arrangements with automobile manufacturers; today, nearly 70% of new cars are SiriusXM-enabled. Yet the company's reach has extended far beyond car radios in the United States to a worldwide presence on the Internet, on smartphones and through other services and distribution channels, including SONOS, JetBlue, and Dish.

BUSINESS CHALLENGE

Despite these remarkable successes, over the past few years changing customer demographics, changing technology, and a changing competitive landscape have posed a new series of business challenges and opportunities for SiriusXM. Here are some notable ones:

- As its market penetration among new cars increased, the demographics of the buyers changed, skewing younger, with less discretionary income. How could SiriusXM reach this new demographic?
- As new cars became used cars and changed hands, how could SiriusXM identify, engage, and convert second owners to paying customers?
- With its acquisition of the connected vehicle business from Agero—the leading provider of telematics in the U.S. car market—SiriusXM gained the ability to deliver its service via both satellite and wireless networks. How could it successfully use this acquisition to capture new revenue streams?

PROPOSED SOLUTION: SHIFTING THE VISION TOWARD DATA-DRIVEN MARKETING

SiriusXM recognized that to address these challenges it would need to become a high-performance, data-driven marketing organization. The company began making that shift by establishing three fundamental tenets.

First, personalized interactions—not mass marketing—would rule the day. The company quickly understood that to conduct more personalized marketing, it would have to draw on past history and interactions, as well as on a keen understanding of the consumer's place in the subscription life cycle.

Second, to gain that understanding, information technology (IT) and its external technology partners would need the ability to deliver integrated data, advanced analytics, integrated marketing platforms, and multichannel delivery systems.

And third, the company could not achieve its business goals without an integrated and consistent point of view across the company. Most important, the technology and business sides of SiriusXM would have to become true partners to best address the challenges involved in becoming a high-performance marketing organization that draws on data-driven insights to speak directly with consumers in strikingly relevant ways.

Those data-driven insights, for example, would enable the company to differentiate between consumers, owners, drivers, listeners, and account holders. The insights would help SiriusXM understand what other vehicles and services are part of each household—and to create new opportunities for engagement. In addition, by constructing a coherent and reliable 360-degree view of all its consumers, SiriusXM could ensure that all messaging in all campaigns and interactions would be tailored, relevant, and consistent across all channels. The important bonus is that more tailored and effective marketing is typically more cost-efficient.

IMPLEMENTATION: CREATING AND FOLLOWING THE PATH TO HIGH-PERFORMANCE MARKETING

At the time of its decision to become a high-performance marketing company, SiriusXM was working with a third-party marketing platform that did not have the capacity to support SiriusXM's ambitions. The company then made an important, forward-thinking decision to bring its marketing capabilities in-house—and then carefully plotted out what it would need to do to make the transition successfully.

- Improve data cleanliness through improved high-level data management and governance. Although the company was understandably impatient to put ideas into action, data hygiene was a necessary first step to creating a reliable window into consumer behavior.
- Bring marketing analytics in-house and expand the data warehouse to enable scale and fully support integrated marketing analytics.
- Develop new segmentation and scoring models to run in-database, eliminating latency and data duplication.
- Extend the integrated data warehouse to include marketing data and scoring, leveraging in-database analytics.
- Adopt a marketing platform for campaign development.
- Bring all that capability together to deliver real-time offer management across all marketing channels: call center, mobile, Web, and in-app.

Completing those steps meant finding the right technology partner. SiriusXM chose Teradata because its strengths were a strong match for the project and company. Teradata offered the ability to:

- Consolidate data sources with an integrated data warehouse (IDW), advanced analytics, and powerful marketing applications.
- Solve data latency issues.

- Significantly reduce data movement across multiple databases and applications.
- Seamlessly interact with applications and modules for all of the marketing areas.
- Scale and perform at very high levels for running campaigns and analytics in-database.
- Conduct real-time communications with customers.
- Provide operational support, either via the cloud or on-premises.

This partnership has enabled SiriusXM to move smoothly and swiftly along its road map, and the company is now in the midst of a transformational, five-year process. After establishing its strong data governance process, SiriusXM began by implementing its Integrated Data Warehouse, which allowed the company to quickly and reliably operationalize insights throughout the organization.

Next, the company implemented Customer Interaction Manager—part of the Teradata Integrated Marketing Cloud, which enables real-time, dialog-based customer interaction across the full spectrum of digital and traditional communication channels. And, SiriusXM will incorporate the Teradata Digital Messaging Center.

Together, the suite of capabilities will allow SiriusXM to handle direct communications across multiple channels. This evolution will enable real-time offers, marketing messages and recommendations based on previous behavior.

In addition to streamlining how they execute and optimize outbound marketing activities, SiriusXM is also taking control of their internal marketing operations with the implementation of Marketing Resource Management, also part of the Teradata Integrated Marketing Cloud. The solution will allow SiriusXM to streamline workflow, optimize marketing resources, and drive efficiency through every penny of their marketing budget.

RESULTS: REAPING THE BENEFITS

As the company continues its evolution into a high-performance marketing organization, already SiriusXM is benefiting from its thoughtfully executed strategy. Household-level consumer insights and a complete view of marketing touch strategy with each consumer enable SiriusXM to create more targeted offers at the household, consumer, and device levels. By bringing the data and marketing analytics capabilities in-house, SiriusXM achieved the following:

- Campaign results in near real-time rather than four days, resulting in massive reductions in cycle times for campaigns and the analysts that support them.
- Closed-loop visibility allowing the analysts to support multistage dialogs and in-campaign modifications to increase campaign effectiveness.
- Real-time modeling and scoring to increase marketing intelligence and sharpen campaign offers and responses at the speed of their business.

Finally, SiriusXM's experience has reinforced the idea that high-performance marketing is a constantly evolving concept. The company has implemented both processes and the technology that give it the capacity for continued and flexible growth.

► QUESTIONS FOR THE OPENING VIGNETTE

1. What does SiriusXM do? In what type of market does it conduct its business?
2. What were the challenges? Comment on both technology and data-related challenges.
3. What were the proposed solutions?
4. How did they implement the proposed solutions? Did they face any implementation challenges?

5. What were the results and benefits? Were they worth the effort/investment?
6. Can you think of other companies facing similar challenges that can potentially benefit from similar data-driven marketing solutions?

WHAT WE CAN LEARN FROM THIS VIGNETTE

Striving to thrive in a fast-changing competitive industry, SiriusXM realized the need for a new and improved marketing infrastructure (one that relies on data and analytics) to effectively communicate the value proposition to its existing and potential customers. As is the case in any industry, in entertainment, success or mere survival depends on intelligently sensing the changing trends (likes and dislikes) and putting together the right messages and policies to win new customers while retaining the existing ones. The key is to create and manage successful marketing campaigns that resonate with the target population of customers and have a close feedback loop to adjust and modify the message to optimize the outcome. At the end, it was all about the precision in the way that they conducted business: being proactive about the changing nature of the clientele, creating and transmitting the right products and services in a timely manner using a fact-based/data-driven holistic marketing strategy. Source identification, source creation, access and collection, integration, cleaning, transformation, storage, and processing of relevant data played a critical role in SiriusXM's success in designing and implementing a marketing analytics strategy, as is the case in any analytically savvy successful company nowadays, regardless of the industry in which they are participating.

Sources: Teradata customer success story. SiriusXM attracts and engages a new generation of radio consumers. <http://assets.teradata.com/resourceCenter/downloads/CaseStudies/EB8597.pdf?processed=1> (accessed May 2022). Quinn, C. (2016). "Innovation – Data-Driven Marketing at SiriusXM," *Forbes Magazine*, <https://www.forbes.com/sites/teradata/2016/01/14/data-driven-marketing-at-siriusxm/?sb=1096319f5ce0> (accessed May 2022).

3.2 THE NATURE OF DATA IN ANALYTICS

Data is the main ingredient for any BI, data science, and business analytics initiative. In fact, it can be viewed as the raw material for what these popular decision technologies produce—information, insight, and **knowledge**. Without data none of these technologies could exist and be popularized—although traditionally we have built analytics models using expert knowledge and experience coupled with very little or no data at all; however, those were the old days, and now data is of the essence. Once perceived as a big challenge to collect, store, and manage, data nowadays is widely considered among the most valuable assets of an organization, with the potential to create invaluable insight to better understand customers, competitors, and the business processes.

Data can be small, or it can be very large. It can be structured (nicely organized for computers to process), or it can be unstructured (e.g., text that is created for humans and hence not readily understandable/consumable by computers). It can come in smaller batches continuously or it can pour in all at once as a large batch. These are some of the characteristics that define the inherent nature of today's data, which we often call Big Data. Even though these characteristics of data make it more challenging to process and consume, it also makes it more valuable because it enriches the data beyond its conventional limits, allowing for the discovery of new and novel knowledge. Traditional ways to manually collect data (either via surveys or via human-entered business transactions) mostly left their places to modern-day data collection mechanisms that use Internet and/or sensor/RFID-based computerized networks. These automated data collection systems are not only enabling us to collect more volumes of data but also enhancing the **data quality** and integrity. Figure 3.1 illustrates a typical analytics continuum—data to analytics to actionable information.

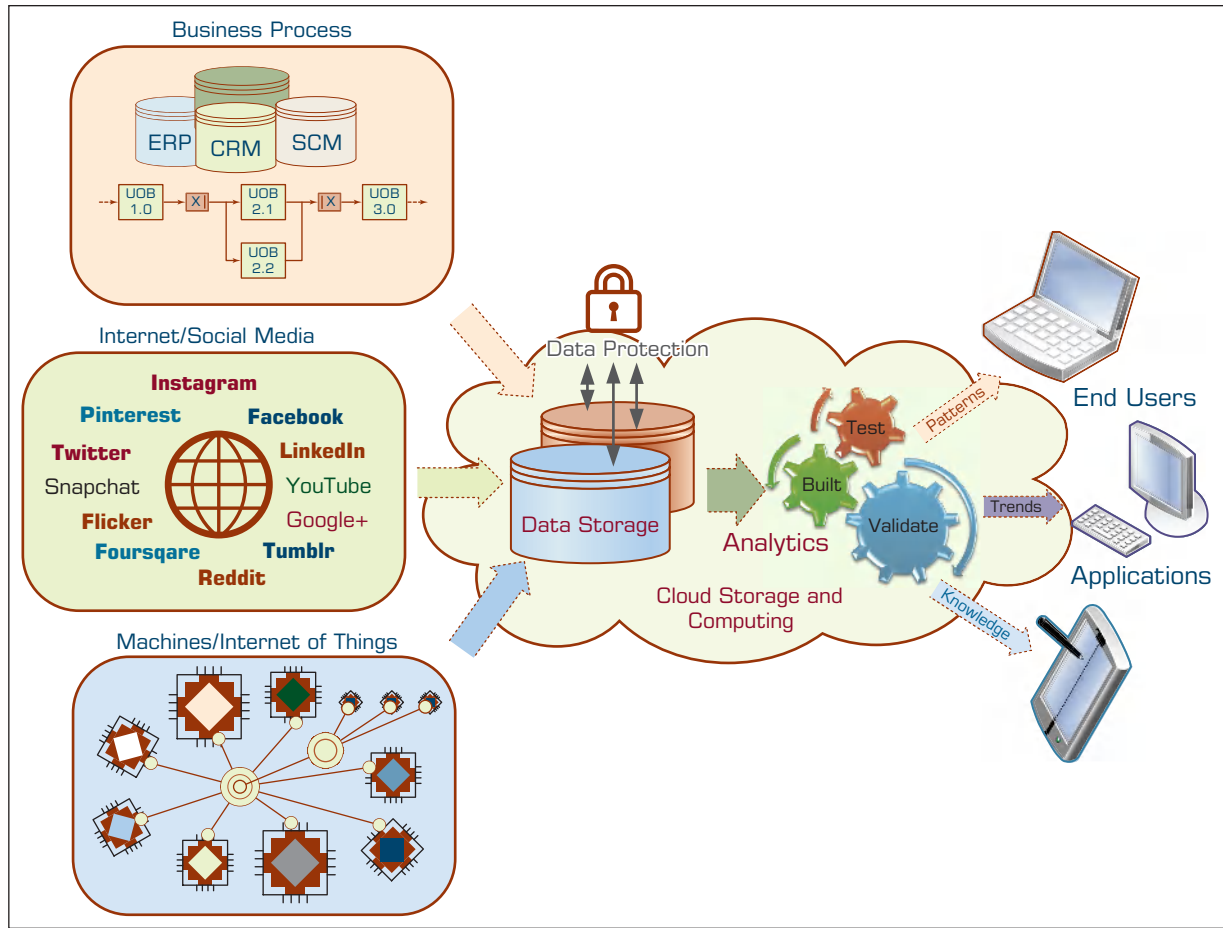


FIGURE 3.1 A Data to Knowledge Continuum.

Although its value proposition is undeniable, to live up its promise, the data must comply with some basic usability and quality metrics. Not all data are useful for all tasks, obviously. That is, data have to match with (have the coverage of the specifics for) the task for which it is intended to be used. Even for a specific task, the relevant data on hand need to comply with the quality and quantity requirements. Essentially, data have to be analytics ready. So what does it mean to make data analytics ready? In addition to its relevancy to the problem at hand and the quality/quantity requirements, it also has to have a certain data structure in place with key fields/variables with properly normalized values. Furthermore, there must be an organization-wide agreed-on definition for common variables and subject matters (sometimes also called high-level data management), such as how you define a customer (what characteristics of customers are used to produce a holistic enough representation to analytics) and where in the business process the customer-related information is captured, validated, stored, and updated.

Sometimes the representation of the data may depend on the type of analytics being employed. Predictive algorithms generally require a flat file with a target variable, so making data **analytics ready** for prediction means that data sets must be transformed into a flat-file format and made ready for ingestion into those predictive algorithms. It is also imperative to match the data to the needs and wants of a specific predictive algorithm and/or a software tool—for instance, neural network algorithms require all input variables to be numerically represented (even the nominal variables need to be converted

into pseudo binary numeric variables) and decision tree algorithms do not require such numerical transformation, easily and natively handling a mix of nominal and numeric variables.

Analytics projects that overlook data-related tasks (some of the most critical steps) often end up with the wrong answer for the right problem, and these unintentionally created, seemingly good, answers may lead to inaccurate and untimely decisions. Following are some of the characteristics (metrics) that define the readiness level of data for an analytics study (Delen, 2020; Kock, McQueen, & Corner, 1997).

- **Data source reliability** refers to the originality and appropriateness of the storage medium where the data is obtained—answering the question of “Do we have the right confidence and belief in this data source?” If at all possible, one should always look for the original source/creator of the data to eliminate/mitigate the possibilities of data misrepresentation and data transformation caused by the mishandling of the data as it moved from the source to destination through one or more steps and stops along the way. Every move of the data creates a chance to unintentionally drop or reformat data items, which limits the integrity and perhaps true accuracy of the data set.
- **Data content accuracy** means that data are correct and are a good match for the analytics problem—answering the question of “Do we have the right data for the job?” The data should represent what was intended or defined by the original source of the data. For example, the customer’s contact information recorded in a record within a database should be the same as what the patient said it was. Data accuracy will be covered in more detail in the following subsection.
- **Data accessibility** means that the data are easily and readily obtainable—answering the question of “Can we easily get to the data when we need to?” Access to data may be tricky, especially if the data is stored in more than one location and storage medium and need to be merged/transformed while accessing and obtaining it. As the traditional relational database management systems leave their place (or coexist with) a new generation of data storage mediums like data lakes and Hadoop infrastructure, the importance/criticality of data accessibility is also increasing.
- **Data security and data privacy** means that the data is secured to only allow those people who have the authority and the need to access it and to prevent anyone else from reaching it. Increasing popularity in educational degrees and certificate programs for Information Assurance is an evidence to the criticality and the increasing urgency of this data quality metric. Any organization that maintains health records for individual patients must have systems in place that not only safeguard the data from unauthorized access (which is mandated by federal laws like Health Insurance Portability and Accountability Act [HIPPA]) but also accurately identifies each patient to allow proper and timely access to records by authorized users (Annas, 2003).
- **Data richness** means that all the required data elements are included in the data set. In essence, richness (or comprehensiveness) means that the available variables portray a rich enough dimensionality of the underlying subject matter for an accurate and worthy analytics study. It also means that the information content is complete (or near complete) to build a predictive and/or prescriptive analytics model.
- **Data consistency** means that the data are accurately collected and combined/merged. Consistent data represent the dimensional information (variables of interest) coming from potentially disparate sources but pertaining to the same subject. If the data integration/merging is not done properly, some of the variables of different subjects may find themselves in the same record—having two different patient records mixed up—for instance, it may happen while merging the demographic and clinical test result data records.

- **Data currency/data timeliness** means that the data should be up-to-date (or as recent/new as it needs to be) for a given analytics model. It also means that the data is recorded at or near the time of the event or observation so that the time-delay-related misrepresentation (incorrectly remembering and encoding) of the data is prevented. Because accurate analytics rely on accurate and timely data, an essential characteristic of analytics-ready data is the timeliness of the creation and access to data elements.
- **Data granularity** requires that the variables and data values be defined at the lowest (or as low as required) level of detail for the intended use of the data. If the data is aggregated, it may not contain the level of detail needed for an analytics algorithm to learn how to discern different records/cases from one another. For example, in a medical setting, numerical values for laboratory results should be recorded to the appropriate decimal place as required for the meaningful interpretation of test results and proper use of those values within an analytics algorithm. Similarly, in the collection of demographic data, data elements should be defined at a granular level to determine the differences in outcomes of care among various subpopulations. One thing to remember is that the data that is aggregated cannot be disaggregated (without access to the original source), but it can easily be aggregated from its granular representation.
- **Data validity** is the term used to describe a match/mismatch between the actual and expected data values of a given variable. As part of data definition, the acceptable values or value ranges for each data element must be defined. For example, a valid data definition related to gender would include three values: male, female, and other (non-binary, transgender, and unknown).
- **Data relevancy** means that the variables in the data set are all relevant to the study being conducted. Relevancy is not a dichotomous measure (whether a variable is relevant or not); rather, it has a spectrum of relevancy from least relevant to most relevant. Based on the analytics algorithms being used, one may choose to include only the most relevant information (i.e., variables) or if the algorithm is capable enough to sort them out, may choose to include all the relevant ones, regardless of their relevancy level. One thing that analytics studies should avoid is to include totally irrelevant data into the model building, as this may contaminate the information for the algorithm, resulting in inaccurate and misleading results.

Although these are perhaps the most prevailing metrics to keep up with, the true data quality and excellent analytics readiness for a specific application domain would require different levels of emphasis paid on these metric dimensions and perhaps add more specific ones to this collection. The following section will dive into the nature of data from a taxonomical perspective to list and define different data types as they relate to different analytics projects.

► SECTION 3.2 REVIEW QUESTIONS

1. How do you describe the importance of data in analytics? Can we think of analytics without data?
2. Considering the new and broad definition of business analytics, what are the main inputs and outputs to the analytics continuum?
3. Where does the data for business analytics come from?
4. In your opinion, what are the top three data-related challenges for better analytics?
5. What are the most common metrics that make for analytics-ready data?

3.3 A SIMPLE TAXONOMY OF DATA

Data (**datum** in singular form) refers to a collection of facts usually obtained as the result of experiments, observations, transactions, or experiences. Data may consist of numbers, letters, words, images, voice recordings, and so on, as measurements of a set of variables (characteristics of the subject or event that we are interested in studying). Data are often viewed as the lowest level of abstraction from which information and then knowledge is derived.

At the highest level of abstraction, one can classify data as structured and unstructured (or semistructured). **Unstructured data**/semistructured data is composed of any combination of textual, imagery, voice, and Web content. Unstructured/semistructured data will be covered in more detail in the text mining and Web mining chapter. **Structured data** is what data mining algorithms use and can be classified as categorical or numeric. The categorical data can be subdivided into nominal or ordinal data, whereas numeric data can be subdivided into intervals or ratios. Figure 3.2 shows a simple **data taxonomy**.

- **Categorical data** represent the labels of multiple classes used to divide a variable into specific groups. Examples of categorical variables include race, sex, age group, and educational level. Although the latter two variables may also be considered in a numerical manner by using exact values for age and highest grade completed, it is often more informative to categorize such variables into a relatively small number of ordered classes. The categorical data may also be called discrete data, implying that it represents a finite number of values with no continuum between them. Even if the values used for the categorical (or discrete) variables are numeric, these numbers are nothing more than symbols and do not imply the possibility of calculating fractional values.
- **Nominal data** contain measurements of simple codes assigned to objects as labels, which are not measurements. For example, the variable *marital status* can be generally categorized as (1) single, (2) married, and (3) divorced. Nominal data can be represented with binomial values having two possible values (e.g., yes/no, true/false, good/bad), or multinomial values having three or more possible values (e.g., brown/green/blue, white/black/Latinx/Asian, single/married/divorced/legal-unions).

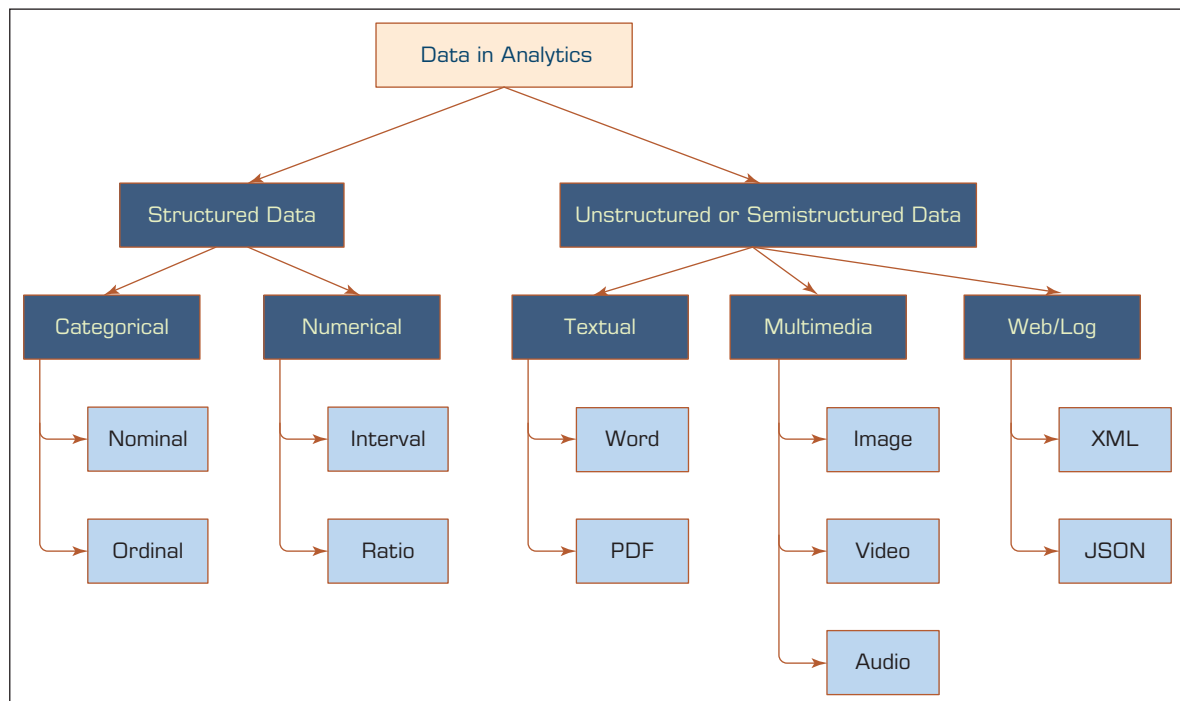


FIGURE 3.2 A Simple Taxonomy of Data.

- **Ordinal data** contain codes assigned to objects or events as labels that also represent the rank order among them. For example, the variable *credit score* can be generally categorized as (1) low, (2) medium, or (3) high. Similar ordered relationships can be seen in variables such as age group (i.e., child, young, middle-aged, older adults) and educational level (i.e., high school, college, graduate school). Some predictive analytic algorithms, such as *ordinal multiple logistic regression*, take into account this additional rank-order information to build a better classification model.
- **Numeric data** represent the numeric values of specific variables. Examples of numerically valued variables include age, number of children, total household income (in U.S. dollars), travel distance (in miles), and temperature (in Fahrenheit degrees). Numeric values representing a variable can be integer (taking only whole numbers) or real (taking also the fractional number). The numeric data may also be called continuous data, implying that the variable contains continuous measures on a specific scale that allows insertion of interim values. Unlike a discrete variable, which represents finite, countable data, a continuous variable represents scalable measurements, and it is possible for the data to contain an infinite number of fractional values.
- **Interval data** are variables that can be measured on interval scales. A common example of interval scale measurement is temperature on the Celsius scale. In this particular scale, the unit of measurement is 1/100 of the difference between the melting temperature and the boiling temperature of water in atmospheric pressure; that is, there is not an absolute zero value.
- **Ratio data** include measurement variables commonly found in the physical sciences and engineering. Mass, length, time, plane angle, energy, and electric charge are examples of physical measures that are ratio scales. The scale type takes its name from the fact that measurement is the estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind. Informally, the distinguishing feature of a ratio scale is the possession of a nonarbitrary zero value. For example, the Kelvin temperature scale has a nonarbitrary zero point of absolute zero, which is equal to -273.15 degrees Celsius. This zero point is nonarbitrary because the particles that comprise matter at this temperature have zero kinetic energy.

Other data types, including textual, spatial, imagery, video, and audio/voice, need to be converted into some form of categorical or numeric representation before they can be processed by analytics methods (data mining algorithms; Delen, 2020). Data can also be classified as static or dynamic (i.e., temporal or time series).

Some predictive analytics (i.e., data mining) methods and machine-learning algorithms are very selective about the type of data that they can handle. Providing them with incompatible data types may lead to incorrect models or (more often) halt the model development process. For example, some data mining methods need all the variables (both input as well as output) represented as numerically valued variables (e.g., neural networks, support vector machines, logistic regression). The nominal or ordinal variables are converted into numeric representations using some type of *1-of-N* pseudo variables (e.g., a categorical variable with three unique values can be transformed into three pseudo variables with binary values—1 or 0). Because this process may increase the number of variables, one should be cautious about the effect of such representations, especially for the categorical variables that have large numbers of unique values.

Similarly, some predictive analytics methods, such as ID3 (a classic decision tree algorithm) and rough sets (a relatively new rule induction algorithm), need all the variables represented as categorically valued variables. Early versions of these methods required the user to discretize numeric variables into categorical representations before they could be processed by the algorithm. The good news is that most implementations

of these algorithms in widely available software tools accept a mix of numeric and nominal variables and internally make the necessary conversions before processing the data.

Data comes in many different variable types and representation schemas. Business analytics tools are continuously improving in their ability to help data scientists in the daunting task of data transformation and data representation so that the data requirements of specific predictive models and algorithms can be properly executed.

► SECTION 3.3 REVIEW QUESTIONS

1. What is data? How does data differ from information and knowledge?
2. What are the main categories of data? What types of data can we use for BI and analytics?
3. Can we use the same data representation for all analytics models? Why, or why not?
4. What is a *1-of-N* data representation? Why and where is it used in analytics?

3.4 THE ART AND SCIENCE OF DATA PREPROCESSING

Data in its original form (i.e., the real-world data) is not usually ready to be used in analytics tasks. It is often dirty, misaligned, overly complex, and inaccurate. A tedious and time-demanding process (so-called **data preprocessing**) is necessary to convert the raw real-world data into a well-refined form for analytics algorithms (Kotsiantis, Kanellopoulos, & Pintelas, 2006). Many analytics professionals would testify that the time spent on data preprocessing (which is perhaps the least enjoyable phase in the whole process) is significantly longer than the time spent on the rest of the analytics tasks (the fun of analytics model building and assessment). Figure 3.3 shows the main steps in the data preprocessing endeavor.

In the first phase of data preprocessing, the relevant data is collected from the identified sources, the necessary records and variables are selected (based on an intimate understanding of the data, the unnecessary information is filtered out), and the records coming from multiple data sources are integrated/merged (again, using the intimate understanding of the data, the synonyms and homonyms are able to be handled properly). Such an integration or merging process is nowadays called data blending.

Data blending is a critical part of the most popular job of the 21st century (i.e., data science), including data source blending, data type blending, database blending, time blending, and tool blending (Silipo & Rudnitckaia, 2022). In order to help with wide range of data blending requests, in November 2016, Silipo and her colleagues at KNIME started a blog post series with an intriguing title “Will they Blend?”. In this series, each blog post has been dedicated to a data blending challenge and offered a potential solution. These blog posts were then combined into a book under the same title, *Will They Blend?* Currently, the third edition of the e-book is in existence and can be accessed at (<https://www.knime.com/knimepress/will-they-blend>). The current edition of the e-book contains 32 chapters describing data blending techniques for more than 50 data sources and external tools, from SQL and NoSQL databases to cloud resources, from SharePoint and SAP to Web services and social media, from R and Python scripts to text and images, from MS Word to Web crawling. If you are curious to know whether your data source is covered in the book, just take a look at the topic Index, more than likely, it will be there.

In the second phase of data preprocessing, the data is cleaned (this step is also known as data scrubbing). Data in its original/raw/real-world form is usually dirty (Delen, 2020). In this step, the values in the data set are identified and dealt with. In some cases,

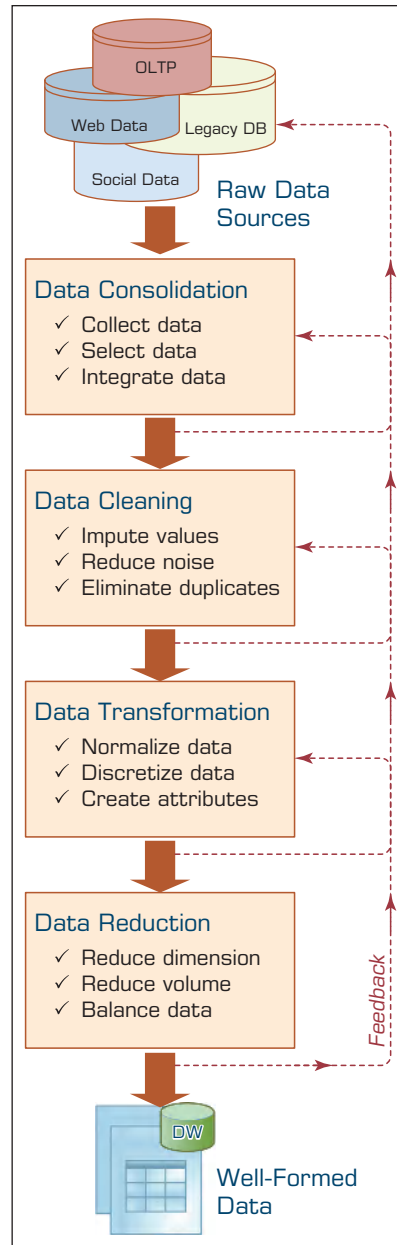


FIGURE 3.3 Data Preprocessing Steps.

missing values are an anomaly in the data set, in which case they need to be imputed (filled with a most probable value) or ignored; in other cases, the missing values are a natural part of the data set (e.g., the *household income* field is often left unanswered by people who are in the top income tier). In this step, the analyst should also identify noisy values in the data (i.e., the outliers) and smooth them out. In addition, inconsistencies (unusual values within a variable) in the data should be handled using domain knowledge and/or expert opinion.

In the third phase of data preprocessing, the data is transformed for better processing. For instance, in many cases the data is normalized between a certain minimum and maximum for all variables to mitigate the potential bias of one variable (having large numeric values, such as for household income) dominating other variables (such as *number of dependents* or *years in service*, which may potentially be more important) having smaller values. Another

transformation that takes place is discretization and/or aggregation. In some cases, the numeric variables are converted to categorical values (e.g., low, medium, high); in other cases, a nominal variable's unique value range is reduced to a smaller set using concept hierarchies (e.g., as opposed to using the individual states with 50 different values, one may choose to use several regions for a variable that shows location) to have a data set that is more amenable to computer processing. Still, in other cases one might choose to create new variables based on the existing ones to magnify the information found in a collection of variables in the data set. For instance, in an organ transplantation data set one might choose to use a single variable showing the blood-type match (1: match, 0: no-match) as opposed to separate multinomial values for the blood type of both the donor and the recipient. Such simplification may increase the information content while reducing the complexity of the relationships in the data.

The final phase of data preprocessing is data reduction. Even though data scientists (i.e., analytics professionals) like to have large data sets, too much data may also be a problem. In the simplest sense, one can visualize the data commonly used in predictive analytics projects as a flat file consisting of two dimensions: variables (the number of columns) and cases/records (the number of rows). In some cases (e.g., image processing and genome projects with complex microarray data), the number of variables can be rather large, and the analyst must reduce the number down to a manageable size. Because the variables are treated as different dimensions that describe the phenomenon from different perspectives, in predictive analytics and data mining this process is commonly called **dimensional reduction** (or **variable selection**). Even though there is not a single best way to accomplish this task, one can use the findings from previously published literature; consult domain experts; run appropriate statistical tests (e.g., principal component analysis or independent component analysis); and, more preferably, use a combination of these techniques to successfully reduce the dimensions in the data into a more manageable and most relevant subset.

With respect to the other dimension (i.e., the number of cases), some data sets may include millions or billions of records. Even though computing power is increasing exponentially, processing such a large number of records may not be practical or feasible. In such cases, one may need to sample a subset of the data for analysis. The underlying assumption of sampling is that the subset of the data will contain all relevant patterns of the complete data set. In a homogeneous data set, such an assumption may hold well, but real-world data is hardly ever homogeneous. The analyst should be extremely careful in selecting a subset of the data that reflects the essence of the complete data set and is not specific to a subgroup or subcategory. The data is usually sorted on some variable, and taking a section of the data from the top or bottom may lead to a biased data set on specific values of the indexed variable; therefore, always try to randomly select the records on the sample set. For skewed data, straightforward random sampling may not be sufficient, and stratified sampling (a proportional representation of different subgroups in the data is represented in the sample data set) may be required. Speaking of skewed data: it is a good practice to balance the highly skewed data by either oversampling the less represented or undersampling the more represented classes. Research has shown that balanced data sets tend to produce better prediction models than unbalanced ones (Thammasiri et al., 2014).

The essence of data preprocessing is summarized in Table 3.1, which maps the main phases (along with their problem descriptions) to a representative list of tasks and algorithms.

It is almost impossible to underestimate the value proposition of data preprocessing. It is one of those time-demanding activities where investment of time and effort pays off without a perceivable limit for diminishing returns. That is, the more resources you invest in it, the more you will gain at the end. Analytics in Action 3.1 illustrates an interesting study where raw, readily available academic data within an educational organization is used to develop predictive models to better understand attrition and improve freshmen student retention in a large higher education institution. As the Analytics in Action clearly states, each and every data preprocessing task described in Table 3.1 was critical to a successful execution of the underlying analytics project, especially the task that related to the balancing of the data set.

TABLE 3.1 A Summary of Data Preprocessing Tasks and Potential Methods

Main Task	Subtasks	Popular Methods
Data consolidation	Access and collect the data	SQL queries, software agents, Web services.
	Select and filter the data	Domain expertise, SQL queries, statistical tests.
	Integrate and unify the data	SQL queries, domain expertise, ontology-driven data mapping.
Data cleaning	Handle missing values in the data	Fill in missing values (imputations) with most appropriate values (mean, median, min/max, mode, etc.); recode the missing values with a constant such as “ML”; remove the record of the missing value; do nothing.
	Identify and reduce noise in the data	Identify the outliers in data with simple statistical techniques (such as averages and standard deviations) or with cluster analysis; once identified, either remove the outliers or smooth them by using binning, regression, or simple averages.
	Find and eliminate erroneous data	Identify the erroneous values in data (other than outliers), such as odd values, inconsistent class labels, odd distributions; once identified, use domain expertise to correct the values or remove the records holding the erroneous values.
Data transformation	Normalize the data	Reduce the range of values in each numerically valued variable to a standard range (e.g., 0 to 1 or −1 to +1) by using a variety of normalization or scaling techniques.
	Discretize or aggregate the data	If needed, convert the numeric variables into discrete representations using range- or frequency-based binning techniques; for categorical variables, reduce the number of values by applying proper concept hierarchies.
	Construct new attributes	Derive new and more informative variables from the existing ones using a wide range of mathematical functions (as simple as addition and multiplication or as complex as a hybrid combination of log transformations).
Data reduction	Reduce number of attributes	Principal component analysis, independent component analysis, chi-square testing, correlation analysis, and decision tree induction.
	Reduce number of records	Random sampling, stratified sampling, expert-knowledge-driven purposeful sampling.
	Balance skewed data	Oversample the less represented or undersample the more represented classes.

ANALYTICS IN ACTION 3.1 Improving Student Retention with Data-Driven Analytics

Student attrition has become one of the most challenging problems for decision makers in academic institutions. Despite all the programs and services that are put in place to help retain students, according to the U.S. Department of Education, Center for Educational Statistics (nces.ed.gov), only about half of those who enter higher education actually earn a bachelor’s degree. Enrolment management and the retention of students has become a top priority for administrators of colleges and universities in the United States and other countries around the world. High dropout of students usually results in overall financial loss, lower graduation rates, and inferior school reputation in the eyes of all stakeholders. The legislators and policy makers who oversee higher education and allocate funds, the parents who pay for their children’s education to prepare them for a better future, and the students who make college choices look for evidence of institutional quality and reputation to guide their decision-making processes.

Proposed Solution

To improve student retention, one should try to understand the nontrivial reasons behind the attrition. To be successful, one should also be able to accurately identify those students that are at risk of dropping out. So far, the vast majority of student attrition research has been devoted

to understanding this complex, yet crucial, social phenomenon. Even though these qualitative, behavioral, and survey-based studies revealed invaluable insight by developing and testing a wide range of theories, they do not provide the much-needed instruments to accurately predict (and potentially improve) student attrition. The project summarized in this case study proposed a quantitative research approach where the historical institutional data from student databases could be used to develop models that are capable of predicting as well as explaining the institution-specific nature of the attrition problem. The proposed analytics approach is shown in Figure 3.4.

Although the concept is relatively new to higher education, for more than a decade now, similar problems in the field of marketing management have been studied using predictive data analytics techniques under the name of “churn analysis,” where the purpose has been to identify among the current customers to answer the question, “Who among our current customers are more likely to stop buying our products or services?” so that some kind of mediation or intervention process can be executed to retain them. Retaining existing customers is crucial because as we all know, and as the related research has shown time and time again, acquiring a new customer costs on an order of magnitude more effort, time, and money than trying to keep the one that you already have.

Data Is of the Essence

The data for this research project came from a single institution (a comprehensive public university located in the Midwest region of the United States) with an average enrollment of 23,000 students, of which roughly 80% are the residents of the same state and roughly 19% of the students are listed as non-white classification. There is no significant difference among genders in the enrollment numbers. The average first-year student retention rate for the institution was about 80%, and the average six-year graduation rate was about 60%.

The study used five years of institutional data, which entailed to 16,000+ students enrolled as first-year students, consolidated from various and diverse university student databases. The data contained variables related to students’ academic, financial, and demographic characteristics. After merging and converting the multidimensional student data into a single flat file (a file with columns representing the variables and rows representing the student records), the resultant file was assessed and preprocessed to identify and remedy anomalies and unusable values. As an example, the study removed all international student records from the data set because they did not contain information about some of the most reputed predictors (e.g., high school GPA, SAT scores). In the data transformation phase, some of the variables were aggregated (e.g., “Major” and “Concentration” variables aggregated to binary variables *MajorDeclared* and *ConcentrationSpecified*) for better interpretation for the predictive modeling. In addition, some of the variables were used to derive new variables (e.g., *Earned/Registered* ratio and *YearsAfterHighSchool*).

$$\text{Earned/Registered} = \text{EarnedHours} / \text{RegisteredHours}$$

$$\text{YearsAfterHighSchool} = \text{FreshmenEnrollmentYear} - \text{HighSchoolGraduationYear}$$

The *Earned/Registered* ratio was created to have a better representation of the students’ resiliency and determination in their first semester of the first year. Intuitively, one would expect greater values for this variable to have a positive impact on retention/persistence. The *YearsAfterHighSchool* was created to measure the impact of the time taken between high school graduation and initial college enrollment. Intuitively, one would expect this variable to be a contributor to the prediction of attrition. These aggregations and derived variables are determined based on a number of experiments conducted for a number of logical hypotheses. The ones that made more common sense and the ones that led to better prediction accuracy were kept in the final variable set. Reflecting the true nature of the subpopulation (i.e., the freshmen students), the dependent variable (i.e., “Second Fall Registered”) contained many more *yes* records (~80%) than *no* records (~20%; see Figure 3.5).

Research shows that having such an imbalanced data has a negative impact on model performance. Therefore, the study experimented with the options of using and comparing the results of the same type of models built with the original imbalanced data (biased for the *yes* records) and the well-balanced data.

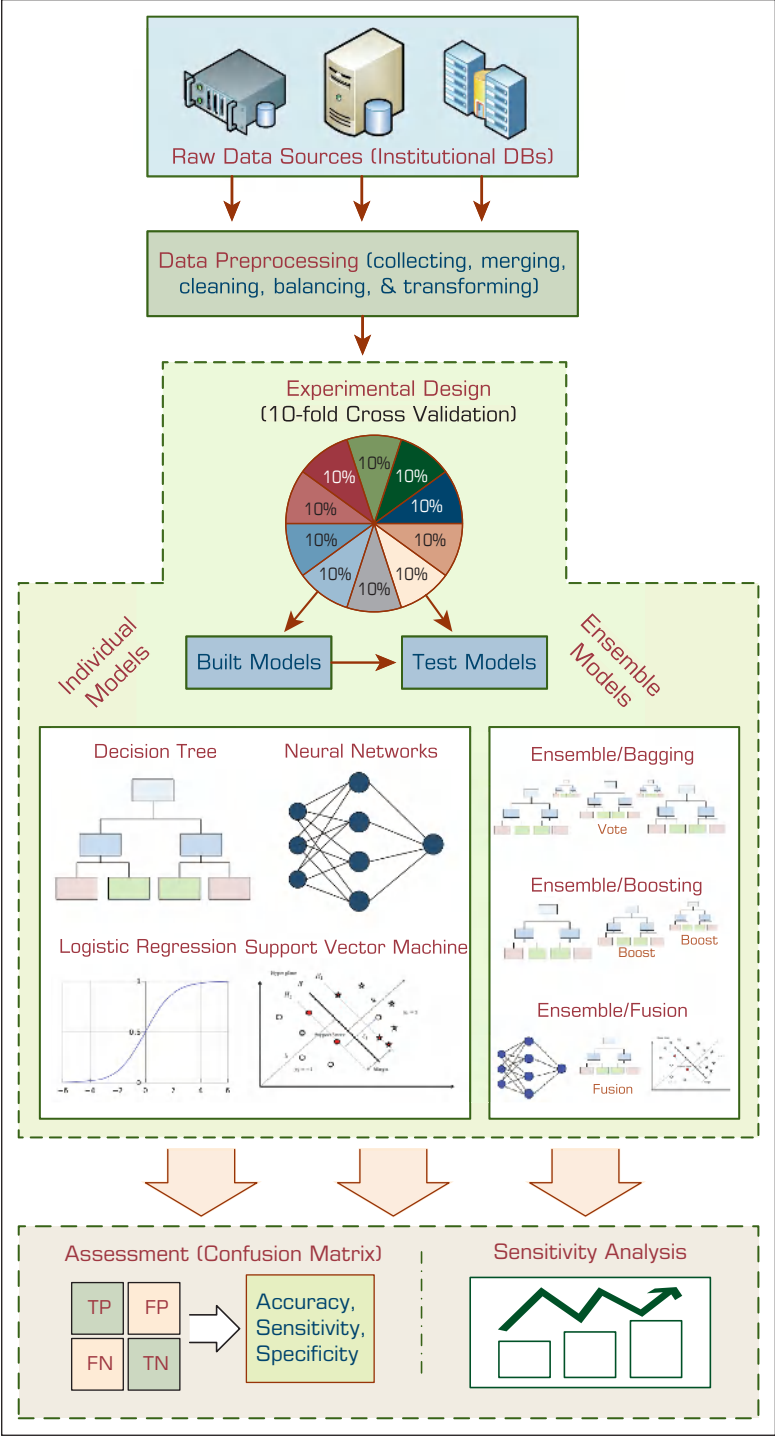


FIGURE 3.4 An Analytics Approach to Predicting Student Attrition.

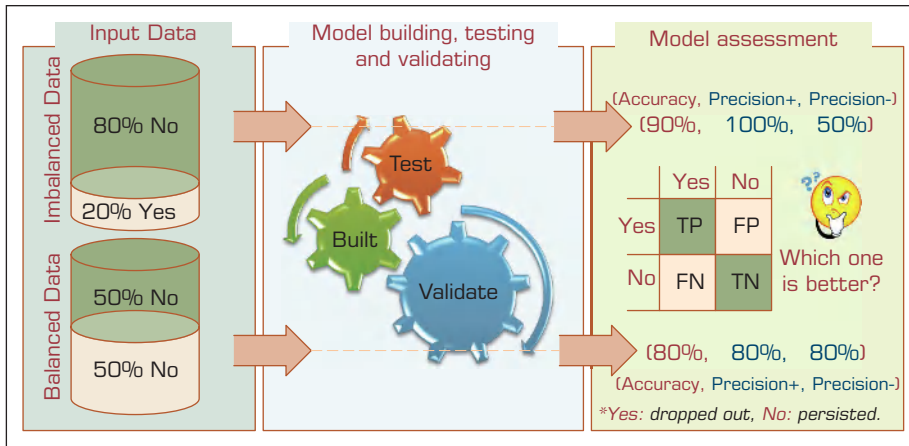


FIGURE 3.5 A Graphical Depiction of the Class Imbalance Problem.

Modeling and Assessment

The study employed four popular classification methods (i.e., artificial neural networks, decision trees, support vector machines, and logistic regression) along with three model ensemble techniques (i.e., bagging, busting, and information fusion). The results obtained from all model types were then compared to each other using regular classification model assessment methods (e.g., overall predictive accuracy, sensitivity, specificity) on the holdout samples.

In machine-learning algorithms (some of which will be covered in Chapter 5), sensitivity analysis is a method for identifying the “cause-and-effect” relationship between the inputs and outputs of a given prediction model. The fundamental idea behind sensitivity analysis is that it measures the importance of predictor variables based on the change in modeling performance that occurs if a predictor variable is not included in the model. This modeling and experimentation practice is also called a leave-one-out assessment. Hence, the measure of sensitivity of a specific predictor variable is the ratio of the error of the trained model without the predictor variable to the error of the model that includes this predictor variable. The more sensitive the network is to a particular variable, the greater the performance decrease would be in the absence of that variable, and therefore the greater the ratio of importance. In addition to the predictive power of the models, the study also conducted sensitivity analyses to determine the relative importance of the input variables.

Results

In the first set of experiments, the study used the original imbalanced data set. Based on the 10-fold cross-validation assessment results, the support vector machines produced the best accuracy with an overall prediction rate of 87.23%, the decision tree came out as the runner-up with an overall prediction rate of 87.16%, followed by artificial neural networks and logistic regression with overall prediction rates of 86.45% and 86.12%, respectively (see Table 3.2).

TABLE 3.2 Prediction Results for the Original/Unbalanced DataSet

	ANN(MLP)		DT(C5)		SVM		LR	
	No	Yes	No	Yes	No	Yes	No	Yes
No	1494	384	1518	304	1478	255	1438	376
Yes	1596	11142	1572	11222	1612	11271	1652	11150
SUM	3090	11526	3090	11526	3090	11526	3090	11526
Per-Class Accuracy	48.35%	96.67%	49.13%	97.36%	47.83%	97.79%	46.54%	96.74%
Overall Accuracy	86.45%		87.16%		87.23%		86.12%	

TABLE 3.3 Prediction Results for the Balanced Data Set

Confusion Matrix	ANN(MLP)		DT(C5)		SVM		LR	
	No	Yes	No	Yes	No	Yes	No	Yes
No	2309	464	2311	417	2313	386	2125	626
Yes	781	2626	779	2673	777	2704	965	2464
SUM	3090	3090	3090	3090	3090	3090	3090	3090
Per-class Accuracy	74.72%	84.98%	74.79%	86.50%	74.85%	87.51%	68.77%	79.74%
Overall Accuracy	79.85%		80.65%		81.18%		74.26%	

A careful examination of these results reveals that the predictions accuracy for the “Yes” class is significantly higher than the prediction accuracy of the “No” class. In fact, all four model types predicted the students who are likely to return for the second year with better than 90% accuracy, but they did poorly on predicting the students who are likely to drop out after the first year with less than 50% accuracy. Because the prediction of the “No” class is the main purpose of this study, less than 50% accuracy for this class was deemed not acceptable. Such a difference in prediction accuracy of the two classes can (and should) be attributed to the imbalanced nature of the training data set (i.e., ~80% “Yes” and ~20% “No” samples).

The next round of experiments used a well-balanced data set where the two classes are represented nearly equally in counts. In realizing this approach, the study took all the samples from the less-represented class (i.e., the “No” class herein) and randomly selected an equal number of samples from the majority class (i.e., the “Yes” class herein) and repeated this process for 10 times to reduce potential bias of random sampling. Each of these sampling processes resulted in a data set of 7,000+ records, of which both class labels (“Yes” and “No”) were equally represented. Again, using a 10-fold cross-validation methodology, the study developed and tested prediction models for all four model types. The results of these experiments are shown in Table 3.3. Based on the holdout sample results, support vector machines once again generated the best overall prediction accuracy with 81.18%, followed by decision trees, artificial neural networks, and logistic regression with an overall prediction accuracy of 80.65%, 79.85%, and 74.26%. As can be seen in the per-class accuracy figures, the prediction models did significantly better on predicting the “No” class with the well-balanced data than they did with the unbalanced data. Overall, the three machine-learning techniques performed significantly better than their statistical counterpart, logistic regression.

Next, another set of experiments were conducted to assess the predictive ability of the three ensemble models. Based on the 10-fold cross-validation methodology, the information fusion-type ensemble model produced the best results with an overall prediction rate of 82.10%, followed by the bagging-type ensembles and boosting-type ensembles with overall prediction rates of 81.80% and 80.21%, respectively (see Table 3.4). Even though the prediction results are slightly better than

TABLE 3.4 Prediction Results for the Three Ensemble Models

	Boosting (Boosted Trees)		Bagging (Random Forest)		Information Fusion (Weighted Average)	
	No	Yes	No	Yes	No	Yes
No	2242	375	2327	362	2335	351
Yes	848	2715	763	2728	755	2739
SUM	3090	3090	3090	3090	3090	3090
Per-Class Accuracy	72.56%	87.86%	75.31%	88.28%	75.57%	88.64%
Overall Accuracy	80.21%		81.80%		82.10%	

the individual models, ensembles are known to produce more robust prediction systems compared to a single-best prediction model (more on this can be found in Chapter 5).

In addition to assessing the prediction accuracy for each model type, a sensitivity analysis was also conducted using the developed prediction models to identify the relative importance of the independent variables (i.e., the predictors). In realizing the overall sensitivity analysis results, each of the four individual model types generated its own sensitivity measures ranking all the independent variables in a prioritized list. As expected, each model type generated slightly different sensitivity rankings of the independent variables. After collecting all four sets of sensitivity numbers, the sensitivity numbers are normalized and aggregated and plotted in a horizontal bar chart (see Figure 3.6).

Conclusions

The study showed that, given sufficient data with the proper variables, data mining methods are capable of predicting freshmen student attrition with approximately 80% accuracy. Results also showed that, regardless of the prediction model employed, the balanced data set (compared

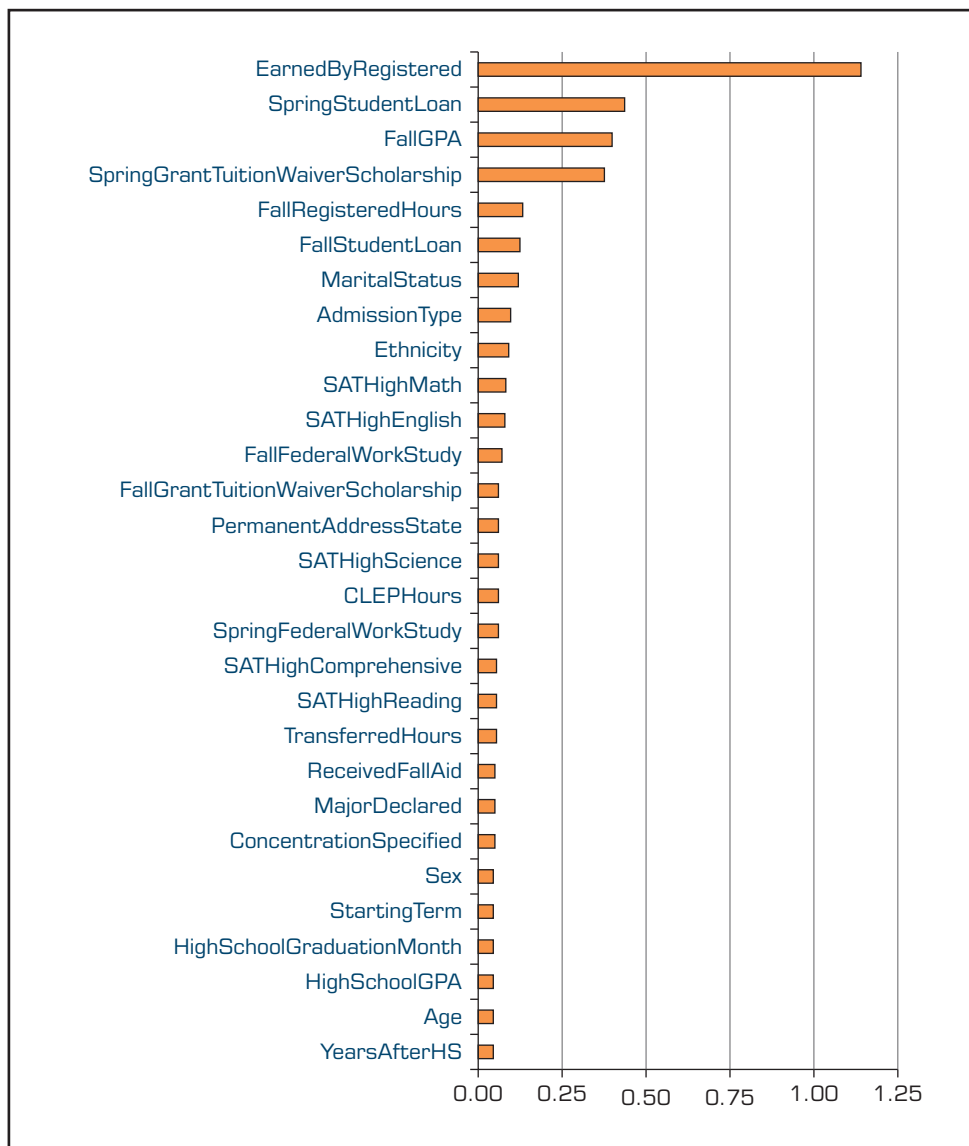


FIGURE 3.6 Sensitivity-Analysis-Based Variable Importance Results.

to unbalanced/original data set) produced better prediction models for identifying the students who are likely to drop out of the college prior to their sophomore year. Among the four individual prediction models used in this study, support vector machines performed the best, followed by decision trees, neural networks, and logistic regression. From the usability standpoint, despite the fact that support vector machines showed better prediction results, one might choose to use decision trees because compared to support vector machines and neural networks, they portray a more transparent model structure. Decision trees explicitly show the reasoning process of different predictions, providing a justification for a specific outcome, whereas support vector machines and artificial neural networks are mathematical models that do not provide such a transparent view of “how they do what they do.”

Sources: Delen, D., Topuz, K., & Eryarsoy, E. (2020). Development of a Bayesian Belief Network-based DSS for predicting and understanding freshmen student attrition. *European Journal of Operational Research*, 281(3), 575–587; Thammasiri, D., Delen, D., Meesad, P., & Kasap N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, 41(2), 321–330; Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention*, 13(1), 17–35; Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4), 498–506.

SECTION 3.4 REVIEW QUESTIONS

1. Why is the original/raw data not readily usable by analytics tasks?
2. What are the main data preprocessing steps?
3. What does it mean to clean/scrub the data? What activities are performed in this phase?
4. Why do we need data transformation? What are the commonly used data transformation tasks?
5. Data reduction can be applied to rows (sampling) and/or columns (variable selection). Which is more challenging?

3.5 DEFINITION OF BIG DATA

Using data to understand customers/clients and business operations to sustain (and foster) growth and profitability is an increasingly challenging task for today’s enterprises. As more and more data becomes available in various forms and fashions, timely processing of the data with traditional means becomes impractical. Nowadays, this phenomenon, usually called Big Data is receiving substantial press coverage and drawing increasing interest from both business users and IT professionals. The result is that Big Data is becoming an overhyped and overused marketing buzzword.

Big Data means different things to people with different backgrounds and interests. Traditionally, the term *Big Data* has been used to describe the massive volumes of data analyzed by huge organizations like Google or research science projects at NASA. But for most businesses, it’s a relative term: “Big” depends on an organization’s size. The point is more about finding new value within and outside conventional data sources. Pushing the boundaries of data analytics uncovers new insights and opportunities, and “big” depends on where you start and how you proceed. Consider the popular description of Big Data: Big Data exceeds the reach of commonly used hardware environments and/or capabilities of software tools to capture, manage, and process it within a tolerable time span for its user population. **Big Data** has become a popular term to describe the exponential growth, availability, and use of information, both structured and unstructured. Much has been written on the Big Data trend and how it can serve as the basis for innovation, differentiation, and growth. Because of the technology challenges in managing the large volume of data coming from multiple sources, sometimes at a rapid speed, additional

new technologies have been developed to overcome the technology challenges. Use of the term *Big Data* is usually associated with such technologies. Because a prime use of storing such data is generating insights through analytics, sometimes the term Big Data is expanded as Big Data analytics. But the term is becoming content free in that it can mean different things to different people. Because our goal is to introduce you to the large data sets and their potential in generating insights, we will use the original term in this chapter.

Where does Big Data come from? A simple answer is “everywhere.” The sources that were ignored because of the technical limitations are now treated as gold mines. Big Data may come from Web logs, radio-frequency identification (RFID), global positioning systems (GPS), sensor networks, social networks, Internet-based text documents, Internet search indexes, detail call records, astronomy, atmospheric science, biology, genomics, nuclear physics, biochemical experiments, medical records, scientific research, military surveillance, photography archives, video archives, and large-scale e-commerce practices.

Big Data is not new. What is new is that the definition and the structure of Big Data constantly change. Companies have been storing and analyzing large volumes of data since the advent of the data warehouses in the early 1990s. Whereas terabytes used to be synonymous with Big Data warehouses, now it’s exabytes, and the rate of growth in data volume continues to escalate as organizations seek to store and analyze greater levels of transaction details, as well as Web- and machine-generated data, to gain a better understanding of customer behavior and business drivers.

Many (academics and industry analysts/leaders alike) think that “Big Data” is a misnomer. What it says and what it means are not exactly the same. That is, Big Data is not just “big.” The sheer volume of the data is only one of many characteristics that are often associated with Big Data, including variety, velocity, veracity, variability, and value proposition, among others.

The “V”s That Define Big Data

Big Data is typically defined by three “V”s: volume, variety, velocity. In addition to these three, we see some of the leading Big Data solution providers adding other “V”s, such as veracity (IBM), variability (SAS), and value proposition.

Volume Volume is obviously the most common trait of Big Data. Many factors contributed to the exponential increase in data volume, such as transaction-based data stored through the years, text data constantly streaming in from social media, increasing amounts of sensor data being collected, automatically generated RFID and GPS data, and so on. In the past, excessive data volume created storage issues, both technical and financial. But with today’s advanced technologies coupled with decreasing storage costs, these issues are no longer significant; instead, other issues have emerged, including how to determine relevance amid the large volumes of data and how to create value from data that is deemed to be relevant.

As mentioned before, *big* is a relative term. It changes over time and is perceived differently by different organizations. With the staggering increase in data volume, even the naming of the next Big Data echelon has been a challenge. The highest mass of data that used to be called petabytes (PB) has left its place to zettabytes (ZB), which is a trillion gigabytes (GB) or a billion terabytes (TB). Technology Insights 3.1 provides an overview of the size and naming of Big Data volumes.

From a short historical perspective, in 2009 the world had about 0.8 ZB of data; in 2010, it exceeded the 1 ZB mark; at the end of 2011, the number was 1.8 ZB. It is expected to be 44 ZB in 2020 (Adshead, 2014). With the growth of sensors and the Internet of Things (IoT), these forecasts could all be wrong. Though these numbers are astonishing in size, so are the challenges and opportunities that come with them.

TECHNOLOGY INSIGHTS 3.1 The Data Size Is Getting Big, Bigger, and Bigger

The metrics of data size is having a hard time keeping up with new names. We all know kilobyte (KB, which is 1,000 bytes), megabyte (MB, which is 1,000,000 bytes), gigabyte (GB, which is 1,000,000,000 bytes), and terabyte (TB, which is 1,000,000,000,000 bytes). Beyond that, the names given to data sizes are relatively new to most of us. The following table shows what comes after terabyte and beyond.

Name	Symbol	Value
Kilobyte	kB	10 ³
Megabyte	MB	10 ⁶
Gigabyte	GB	10 ⁹
Terabyte	TB	10 ¹²
Petabyte	PB	10 ¹⁵
Exabyte	EB	10 ¹⁸
Zettabyte	ZB	10 ²¹
Yottabyte	YB	10 ²⁴
Brontobyte*	BB	10 ²⁷
Gegobyte*	GeB	10 ³⁰

*Not an official SI (International System of Units) name/symbol, yet.

Consider that an exabyte of data is created on the Internet each day, which equates to 250 million DVDs’ worth of information. And the idea of even larger amounts of data—a zettabyte—isn’t too far off when it comes to the amount of information traversing the Web in any one year. In fact, industry experts are already estimating that we will see 1.3 zettabytes of traffic annually over the Internet by 2016—and it could jump to 2.3 zettabytes by 2020. By 2020, Internet traffic is expected to reach 300 GB per capita per year. When referring to yottabytes, some of the Big Data scientists often wonder about how much data the NSA or FBI have on people altogether. Put in terms of DVDs, a yottabyte would require 250 trillion of them. A brontobyte, which is not an official SI prefix but is apparently recognized by some people in the measurement community, is a 1 followed by 27 zeros. The size of such a magnitude can be used to describe the amount of sensor data that we will get from the Internet in the next decade, if not sooner.

Sources: Backblaze. (2022). What is a yottabyte? at <https://www.backblaze.com/blog/what-is-a-yottabyte/> (accessed May 2022); Byte. (2022). Unit of information, at <https://en.wikipedia.org/wiki/Byte> (accessed April 2022).

Variety Data today come in all types of formats—ranging from traditional databases to hierarchical data stores created by the end users and OLAP systems to text documents, e-mail, XML, meter-collected and sensor-captured data, to video, audio, and stock ticker data. By some estimates, 80 to 85% of all organizations’ data are in some sort of unstructured or semistructured format (a format that is not suitable for traditional database schemas). But there is no denying its value, and hence, it must be included in the analyses to support decision making.

Velocity According to Gartner, velocity means both how fast data is being produced and how fast the data must be processed (i.e., captured, stored, and analyzed) to meet the need or demand. RFID tags, automated sensors, GPS devices, and smart meters are driving an increasing need to deal with torrents of data in near real time. Velocity is

perhaps the most overlooked characteristic of Big Data. Reacting quickly enough to deal with velocity is a challenge to most organizations. For time-sensitive environments, the opportunity cost clock of the data starts ticking the moment the data is created. As time passes, the value proposition of the data degrades and eventually becomes worthless. Whether the subject matter is the health of a patient, the well-being of a traffic system, or the health of an investment portfolio, accessing the data and reacting faster to the circumstances will always create more advantageous outcomes.

In the Big Data storm that we are currently witnessing, almost everyone is fixated on at-rest analytics, using optimized software and hardware systems to mine large quantities of variant data sources. Although this is critically important and highly valuable, there is another class of analytics, driven from the velocity of Big Data, called “data stream analytics” or “in-motion analytics,” which is evolving fast. If done correctly, data stream analytics can be as valuable, and in some business environments more valuable, than at-rest analytics. Later in this chapter, we will cover this topic in more detail.

Veracity *Veracity* is a term coined by IBM that is being used as the fourth “V” to describe Big Data. It refers to conformity to facts: accuracy, quality, truthfulness, or trustworthiness of the data. Tools and techniques are often used to handle Big Data’s veracity by transforming the data into quality and trustworthy insights.

Variability In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Is something big trending in the social media? Perhaps there is a high-profile IPO looming. Maybe swimming with manatees in the Bahamas is suddenly the must-do vacation activity. Daily, seasonal, and event-triggered peak data loads can be highly variable and thus challenging to manage—especially with social media involved.

Value Proposition The excitement around Big Data is its value proposition. A preconceived notion about “Big” data is that it contains (or has a greater potential to contain) more patterns and interesting anomalies than “small” data. Thus, by analyzing large and feature-rich data, organizations can gain greater business value that they may not have otherwise. Although users can detect the patterns in small data sets using simple statistical and machine-learning methods or ad hoc query and reporting tools, Big Data means “big” analytics. Big analytics means greater insight and better decisions, something that every organization needs.

Because the exact definition of Big Data (or its successor terms) is still a matter of ongoing discussion in academic and industrial circles, it is likely that more characteristics (perhaps more “V”s) are likely to be added to this list. Regardless of what happens, the importance and value proposition of Big Data is here to stay. Figure 3.7 shows a conceptual architecture where Big Data (at the left side of the figure) is converted to business insight through the use of a combination of advanced analytics and delivered to a variety of different users/roles for faster/better decision making.

► SECTION 3.5 REVIEW QUESTIONS

1. Why is Big Data important? What has changed to put it in the center of the analytics world?
2. How do you define Big Data? Why is it difficult to define?
3. Out of the “V”s that are used to define Big Data, in your opinion, which one is the most important? Why?
4. What do you think the future of Big Data will be like? Will it leave its popularity to something else? If so, what will it be?

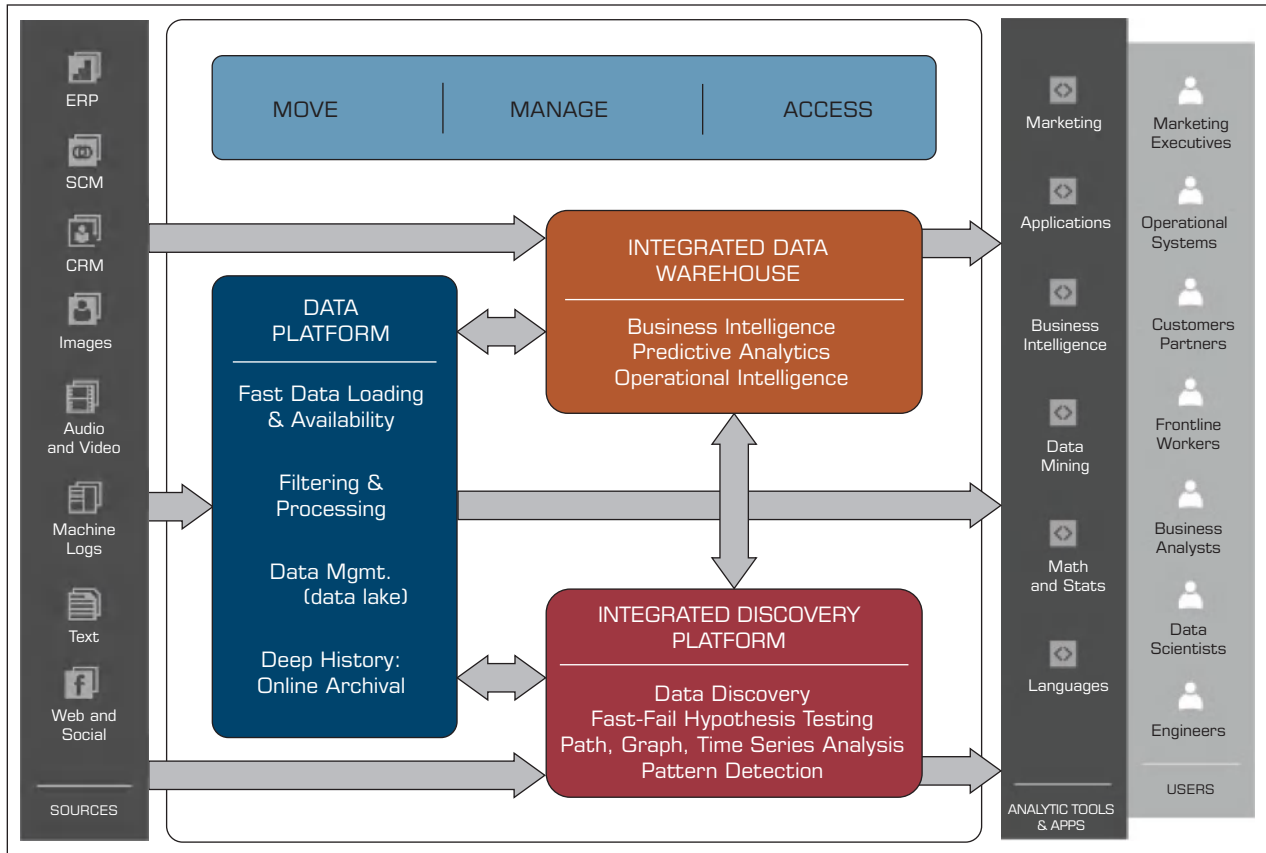


FIGURE 3.7 A High-Level Conceptual Architecture for Big Data Solutions. Source: AsterData—A Teradata Company.

3.6 FUNDAMENTALS OF BIG DATA ANALYTICS

Big Data by itself, regardless of the size, type, or speed, is worthless unless business users do something with it that delivers value to their organizations. That's where "big" analytics comes into the picture. Although organizations have always run reports and dashboards against data warehouses, most have not opened these repositories to in-depth on-demand exploration. This is partly because analysis tools are too complex for the average user but also because the repositories often do not contain all the data needed by the power user. But this is about to change (and has been changing, for some) in a dramatic fashion, thanks to the new Big Data analytics paradigm.

With the value proposition, Big Data also brought about big challenges for organizations. The traditional means for capturing, storing, and analyzing data are not capable of dealing with Big Data effectively and efficiently. Therefore, new breeds of technologies need to be developed (or purchased/hired/outsourced) to take on the Big Data challenge. Before making such an investment, organizations should justify the means. Here are some questions that may help shed light on this situation. If any of the following statements are true, then you need to seriously consider embarking on a Big Data journey.

- You can't process the amount of data that you want to because of the limitations posed by your current platform or environment.
- You want to involve new/contemporary data sources (e.g., social media, RFID, sensory, Web, GPS, textual data) into your analytics platform, but you can't because it

does not comply with the data storage schema-defined rows and columns without sacrificing fidelity or the richness of the new data.

- You need to (or want to) integrate data as quickly as possible to be current on your analysis.
- You want to work with a schema-on-demand (as opposed to predetermined schema used in relational database management systems [RDBMSs]) data storage paradigm because the nature of the new data may not be known, or there may not be enough time to determine it and develop schema for it.
- The data is arriving so fast at your organization's doorstep that your traditional analytics platform cannot handle it.

As is the case with any other large IT investment, the success in Big Data analytics depends on a number of factors. Figure 3.8 shows a graphical depiction of the most critical success factors.

The following are the most critical success factors for **Big Data analytics**:

1. **A clear business need (alignment with the vision and the strategy).** Business investments ought to be made for the good of the business, not for the sake of mere technology advancements. Therefore, the main driver for Big Data analytics should be the needs of the business, at any level—strategic, tactical, and operations.
2. **Strong, committed sponsorship (executive champion).** It is a well-known fact that if you don't have strong, committed executive sponsorship, it is difficult (if not impossible) to succeed. If the scope is a single or a few analytical applications, the sponsorship can be at the departmental level. However, if the target is enterprise-wide organizational transformation, which is often the case for Big Data initiatives, sponsorship needs to be at the highest levels and organization wide.



FIGURE 3.8 Critical Success Factors for Big Data Analytics.

3. **Alignment between the business and IT strategy.** It is essential to make sure that the analytics work is always supporting the business strategy, and not the other way around. Analytics should play the enabling role in successfully executing the business strategy.
4. **A fact-based decision-making culture.** In a fact-based decision-making culture, the numbers rather than intuition, gut feeling, or supposition drive decision making. There is also a culture of experimentation to see what works and what doesn't. To create a fact-based decision-making culture, senior management needs to:
 - Recognize that some people can't or won't adjust
 - Be a vocal supporter
 - Stress that outdated methods must be discontinued
 - Ask to see what analytics went into decisions
 - Link incentives and compensation to desired behaviors
5. **A strong data infrastructure.** Data warehouses have provided the data infrastructure for analytics. This infrastructure is changing and being enhanced in the Big Data era with new technologies. Success requires marrying the old with the new for a holistic infrastructure that works synergistically.

As the size and complexity increase, the need for more efficient analytical systems is also increasing. To keep up with the computational needs of Big Data, a number of new and innovative computational techniques and platforms have been developed. These techniques are collectively called **high-performance computing**, which includes the following:

- **In-memory analytics:** Solves complex problems in near real time with highly accurate insights by allowing analytical computations and Big Data to be processed in-memory and distributed across a dedicated set of nodes.
- **In-database analytics:** Speeds time to insights and enables better data governance by performing data integration and analytic functions inside the database so you won't have to move or convert data repeatedly.
- **Grid computing:** Promotes efficiency, lower cost, and better performance by processing jobs in a shared, centrally managed pool of IT resources.
- **Appliances:** Brings together hardware and software in a physical unit that is not only fast but also scalable on an as-needed basis.

Computational requirements are just a small part of the list of challenges that Big Data impose on today's enterprises. The following is a list of challenges that are found by business executives to have a significant impact on successful implementation of Big Data analytics. When considering Big Data projects and architecture, being mindful of these challenges will make the journey to analytics competency a less stressful one.

Data volume: The ability to capture, store, and process a huge volume of data at an acceptable speed so that the latest information is available to decision makers when they need it.

Data integration: The ability to combine data that is not similar in structure or source and to do so quickly and at a reasonable cost.

Processing capabilities: The ability to process data quickly, as it is captured. The traditional way of collecting and processing data may not work. In many situations, data needs to be analyzed as soon as it is captured to leverage the most value. (This is called *stream analytics*, which will be covered later in this chapter.)

Data governance: The ability to keep up with the security, privacy, ownership, and quality issues of Big Data. As the volume, variety (format and source), and velocity of data change, so should the capabilities of governance practices.

Skills availability: Big Data is being harnessed with new tools and is being looked at in different ways. There is a shortage of people (often called *data scientists*) with skills to do the job.

Solution cost: Because Big Data has opened up a world of possible business improvements, a great deal of experimentation and discovery is taking place to determine the patterns that matter and the insights that turn to value. To ensure a positive return on investment on a Big Data project, therefore, it is crucial to reduce the cost of the solutions used to find that value.

Though the challenges are real, so is the value proposition of Big Data analytics. Anything that you can do as a business analytics leader to help prove the value of new data sources to the business will move your organization beyond experimenting and exploring Big Data into adapting and embracing it as a differentiator. There is nothing wrong with exploration, but ultimately the value comes from putting those insights into action.

Business Problems Addressed by Big Data Analytics

The top business problems addressed by Big Data overall are process efficiency and cost reduction, as well as enhancing customer experience, but different priorities emerge when it is looked at by industry. Process efficiency and cost reduction are perhaps among the top-ranked problems that can be addressed with Big Data analytics for the manufacturing, government, energy and utilities, communications and media, transport, and healthcare sectors. Enhanced customer experience may be at the top of the list of problems addressed by insurance companies and retailers. Risk management usually is at the top of the list for companies in banking and education. Here is a partial list of problems that can be addressed using Big Data analytics:

- Process efficiency and cost reduction
- Brand management
- Revenue maximization, cross-selling, and up-selling
- Enhanced customer experience
- Churn identification, customer recruiting
- Improved customer service
- Identifying new products and market opportunities
- Risk management
- Regulatory compliance
- Enhanced security capabilities

SECTION 3.6 REVIEW QUESTIONS

1. What is Big Data analytics? How does it differ from regular analytics?
2. What are the critical success factors for Big Data analytics?
3. What are the big challenges that one should be mindful of when considering implementation of Big Data analytics?
4. What are the common business problems addressed by Big Data analytics?

3.7 BIG DATA TECHNOLOGIES

There are a number of technologies for processing and analyzing Big Data, but most have some common characteristics (Kelly, 2012). Namely, they take advantage of commodity hardware to enable scale-out and parallel-processing techniques; employ nonrelational data storage capabilities to process unstructured and semistructured data; and apply



Source: Hadoop.

advanced analytics and data visualization technology to Big Data to convey insights to end users. The three Big Data technologies that stand out that most believe will transform the business analytics and data management markets are Hadoop, MapReduce, and NoSQL.

Hadoop

Hadoop is an open-source framework for processing, storing, and analyzing massive amounts of distributed, unstructured data. Originally created by Doug Cutting at Yahoo!, Hadoop was inspired by MapReduce, a user-defined function developed by Google in the early 2000s for indexing the Web. It was designed to handle petabytes and exabytes of data distributed over multiple nodes in parallel. Hadoop clusters run on inexpensive commodity hardware so projects can scale-out without breaking the bank. Hadoop is now a project of the Apache Software Foundation, where hundreds of contributors continuously improve the core technology. Fundamental concept: rather than banging away at one huge block of data with a single machine, Hadoop breaks up Big Data into multiple parts so each part can be processed and analyzed at the same time.

How Does Hadoop Work?

A client accesses structured, unstructured and semistructured data from sources including log files, social media feeds, and internal data stores. Hadoop breaks the data up into smaller “parts,” which are then loaded into a file system made up of multiple nodes running on commodity hardware. The default file store in Hadoop is the **Hadoop Distributed File System**, or **HDFS**. File systems such as HDFS are adept at storing large volumes of unstructured and semistructured data as they do not require data to be organized into relational rows and columns. Each “part” is replicated multiple times and loaded into the file system so that if a node fails, another node has a copy of the data contained on the failed node. A Name Node acts as facilitator, communicating back to the client information such as which nodes are available, where in the cluster certain data resides, and which nodes have failed.

Once the data is loaded into the cluster, it is ready to be analyzed via the MapReduce framework. The client submits a “Map” job—usually a query written in Java—to one of the nodes in the cluster known as the Job Tracker. The Job Tracker refers to the Name Node to determine which data it needs to access to complete the job and where in the cluster that data is located. Once determined, the Job Tracker submits the query to the relevant nodes. Rather than bringing all the data back into a central location for processing, the processing occurs at each node simultaneously, or in parallel. This is an essential characteristic of Hadoop.

When each node has finished processing its given job, it stores the results. The client initiates a “Reduce” job through the Job Tracker in which results of the map phase stored locally on individual nodes are aggregated to determine the “answer” to the original query, and then are loaded onto another node in the cluster. The client accesses these results, which can then be loaded into one of a number of analytic environments for analysis. The MapReduce job has now been completed.

Once the MapReduce phase is complete, the processed data is ready for further analysis by data scientists and others with advanced data analytics skills. **Data scientists** can manipulate and analyze the data using any of a number of tools for any number of uses, including searching for hidden insights and patterns, or use as the foundation for building user-facing analytic applications. The data can also be modeled and transferred from Hadoop clusters into existing relational databases, data warehouses, and other traditional IT systems for further analysis and/or to support transactional processing.

MapReduce

MapReduce is a technique popularized by Google that distributes the processing of very large multistructured data files across a large cluster of machines. High performance is achieved by breaking the processing into small units of work that can be run in parallel across the hundreds, potentially thousands, of nodes in the cluster. To quote the seminal paper on MapReduce:

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system. (Maleki et al., 2019)

The key point to note from this quote is that MapReduce is a programming model, not a programming language, that is, it is designed to be used by programmers, rather than business users. The easiest way to describe how MapReduce works is through the use of an example—see the Colored Square Counter in Figure 3.9.

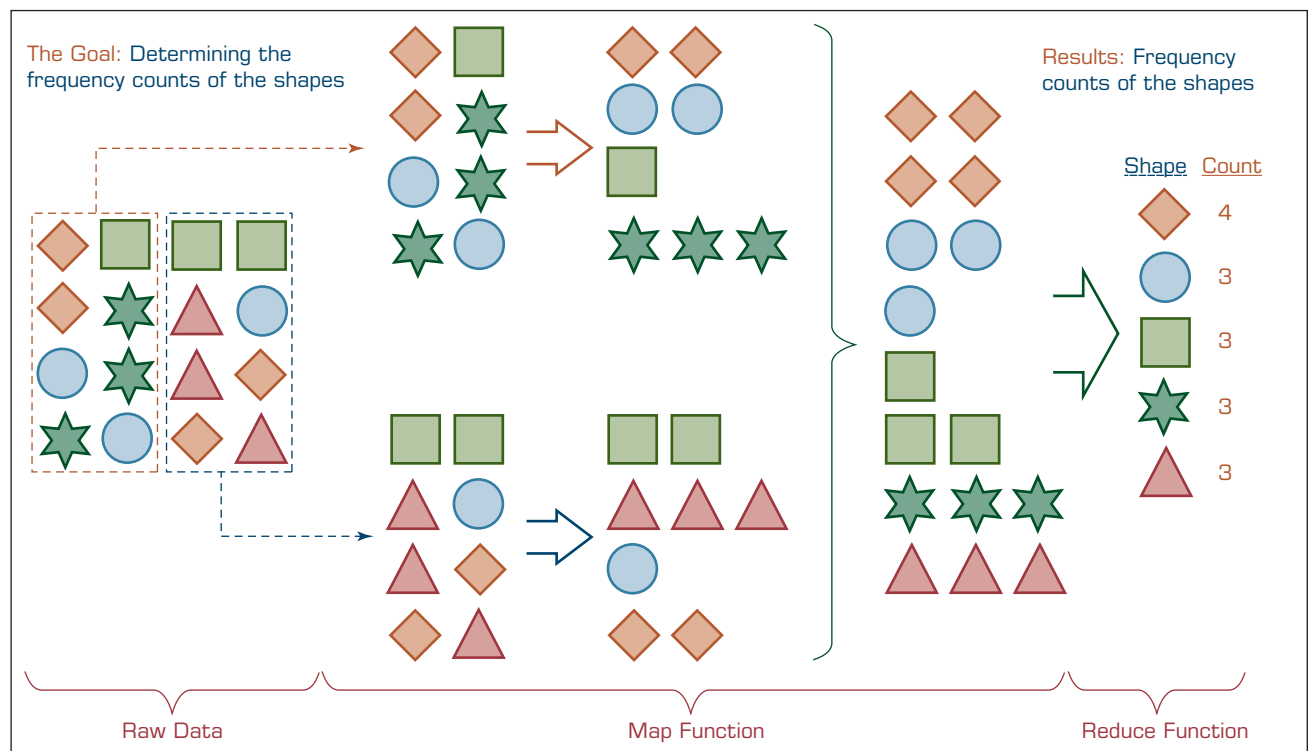


FIGURE 3.9 A Graphical Depiction of the MapReduce Process.

The input to the MapReduce process in Figure 3.9 is a set of colored squares. The objective is to count the number of squares of each color. The programmer in this example is responsible for coding the map and reducing programs; the remainder of the processing is handled by the software system implementing the MapReduce programming model.

The MapReduce system first reads the input file and splits it into multiple pieces. In this example, there are two splits, but in a real-life scenario, the number of splits would typically be much higher. These splits are then processed by multiple map programs running in parallel on the nodes of the cluster. The role of each map program in this case is to group the data in a split by color. The MapReduce system then takes the output from each map program and merges (shuffle/sort) the results for input to the reduce program, which calculates the sum of the number of squares of each color. In this example, only one copy of the reduce program is used, but there may be more in practice. To optimize performance, programmers can provide their own shuffle/sort program and can also deploy a combiner that combines local map output files to reduce the number of output files that have to be remotely accessed across the cluster by the shuffle/sort step.

Why Use MapReduce?

MapReduce aids organizations in processing and analyzing large volumes of multistructured data. Application examples include indexing and search, graph analysis, text analysis, machine learning, data transformation, and so forth. These types of applications are often difficult to implement using the standard SQL employed by relational DBMSs.

The procedural nature of MapReduce makes it easily understood by skilled programmers. It also has the advantage that developers do not have to be concerned with implementing parallel computing—this is handled transparently by the system. Although MapReduce is designed for programmers, nonprogrammers can exploit the value of pre-built MapReduce applications and function libraries. Both commercial and open-source MapReduce libraries are available that provide a wide range of analytic capabilities. Apache Mahout, for example, is an open-source machine-learning library of “algorithms for clustering, classification and batch-based collaborative filtering” that are implemented using MapReduce.

Hadoop Technical Components

In addition to MapReduce, a Hadoop “stack” is made up of a number of components, which include the following:

- **Hadoop Distributed File System (HDFS):** The default storage layer in any given Hadoop cluster.
- **Name Node:** The node in a Hadoop cluster that provides the client information on where in the cluster particular data is stored and if any nodes fail.
- **Secondary Node:** A backup to the Name Node, it periodically replicates and stores data from the Name Node should it fail.
- **Job Tracker:** The node in a Hadoop cluster that initiates and coordinates MapReduce jobs or the processing of the data.
- **Worker Nodes:** The grunts of any Hadoop cluster, worker nodes store data and take direction to process it from the Job Tracker.

In addition to these components, the Hadoop ecosystem is made up of a continuously expanding large number of complementary subprojects. NoSQL data stores like Cassandra and HBase are also used to store the results of MapReduce jobs in Hadoop. In addition to Java, some MapReduce jobs and other Hadoop functions are written in Pig, an open-source language designed specifically for Hadoop. **Hive** is an open-source data

warehouse originally developed by Facebook that allows for analytic modeling within Hadoop. Here are the most commonly referenced subprojects for Hadoop.

Hive Hive is a Hadoop-based data warehousing–like framework originally developed by Facebook. It allows users to write queries in an SQL-like language called HiveQL, which are then converted to MapReduce. This allows SQL programmers with no MapReduce experience to use the warehouse and makes it easier to integrate with business intelligence (BI) and visualization tools such as Microstrategy, Tableau, Revolutions Analytics, and so forth.

Pig **Pig** is a Hadoop-based query language developed by Yahoo! It is relatively easy to learn and is adept at very deep, very long data pipelines (a limitation of SQL).

HBase HBase is a nonrelational database that allows for low-latency, quick lookups in Hadoop. It adds transactional capabilities to Hadoop, allowing users to conduct updates, inserts, and deletes. eBay and Facebook use HBase heavily.

Flume Flume is a framework for populating Hadoop with data. Agents are populated throughout one's IT infrastructure—inside Web servers, application servers, and mobile devices, for example—to collect data and integrate it into Hadoop.

Oozie Oozie is a workflow processing system that lets users define a series of jobs written in multiple languages—such as MapReduce, Pig, and Hive—and then intelligently link them to one another. Oozie allows users to specify, for example, that a particular query is only to be initiated after specified previous jobs on which it relies for data are completed.

Ambari Ambari is a Web-based set of tools for deploying, administering, and monitoring Apache Hadoop clusters. Its development is being led by engineers from Hortonworks, which includes Ambari in its Hortonworks Data Platform.

Avro Avro is a data serialization system that allows for encoding the schema of Hadoop files. It is adept at parsing data and performing removed procedure calls.

Mahout Mahout is a data mining library. It takes the most popular data mining algorithms for performing clustering, regression testing, and statistical modeling and implements them using the MapReduce model.

Sqoop Sqoop is a connectivity tool for moving data from non-Hadoop data stores—such as relational databases and data warehouses—into Hadoop. It allows users to specify the target location inside of Hadoop and instructs Sqoop to move data from Oracle, Teradata, or other relational databases to the target.

HCatalog HCatalog is a centralized metadata management and sharing service for Apache Hadoop. It allows for a unified view of all data in Hadoop clusters and allows diverse tools, including Pig and Hive, to process any data elements without needing to know physically where in the cluster the data is stored.

Hadoop: The Pros and Cons

The main benefit of Hadoop is that it allows enterprises to process and analyze large volumes of unstructured and semistructured data, heretofore inaccessible to them, in a cost- and time-effective manner. Because Hadoop clusters can scale to petabytes and even

exabytes of data, enterprises no longer must rely on sample data sets but can process and analyze *all* relevant data. Data scientists can apply an iterative approach to analysis, continually refining and testing queries to uncover previously unknown insights. It is also inexpensive to get started with Hadoop. Developers can download the Apache Hadoop distribution for free and begin experimenting with Hadoop in less than a day.

The downside to Hadoop and its myriad components is that they are immature and still developing. As with any young, raw technology, implementing and managing Hadoop clusters and performing advanced analytics on large volumes of unstructured data require significant expertise, skill, and training. Unfortunately, there is currently a dearth of Hadoop developers and data scientists available, making it impractical for many enterprises to maintain and take advantage of complex Hadoop clusters. Further, as Hadoop's myriad components are improved on by the community and new components are created, there is, as with any immature open-source technology/approach, a risk of forking. Finally, Hadoop is a batch-oriented framework, meaning it does not support real-time data processing and analysis. In order to increase the efficiency and real-time data processing, another open-source big data technology under the name of Spark is developed by the Apache Foundation, which is explained in the next subsection.

The good news is that some of the brightest minds in IT are contributing to the open-source big data technology development projects, and a new generation of developers and data scientists is coming of age. As a result, the technology is advancing rapidly, becoming both more powerful and easier to implement and manage. An ecosystem of vendors, both recently founded start-ups like Cloudera and Hortonworks and well-known IT stalwarts like IBM, Microsoft, Teradata, and Oracle are working to offer commercial, enterprise-ready big data tools and services to make deploying and managing the underlying technology a practical reality for the traditional enterprise. Several other leading-edge start-ups are working to perfect NoSQL (Not Only SQL) data stores capable of delivering near-real-time insights in conjunction with Hadoop and Spark. Technology Insights 3.2 provides a few facts to clarify some misconceptions about Hadoop.

TECHNOLOGY INSIGHTS 3.2 A Few Demystifying Facts about Hadoop

Although Hadoop and related technologies have been around for more than five years now, most people still have several misconceptions about Hadoop and related technologies such as MapReduce and Hive. The following list of 10 facts intends to clarify what Hadoop is and does relative to BI, as well as in which business and technology situations Hadoop-based BI, data warehousing, and analytics can be useful (Russom, 2013).

Fact #1. Hadoop consists of multiple products. We talk about Hadoop as if it's one monolithic software, whereas it's actually a family of open-source products and technologies overseen by the Apache Software Foundation (ASF). (Some Hadoop products are also available via vendor distributions; more on that later.)

The Apache Hadoop library includes (in BI priority order) HDFS, MapReduce, Hive, Hbase, Pig, Zookeeper, Flume, Sqoop, Oozie, Hue, and so on. You can combine these in various ways, but HDFS and MapReduce (perhaps with Hbase and Hive) constitute a useful technology stack for applications in BI, data warehousing, and analytics.

Fact #2. Hadoop is open source but available from vendors, too. Apache Hadoop's open-source software library is available from ASF at apache.org. For users desiring a more enterprise-ready package, a few vendors now offer Hadoop distributions that include additional administrative tools and technical support.

Fact #3. Hadoop is an ecosystem, not a single product. In addition to products from Apache, the extended Hadoop ecosystem includes a growing list of vendor products that integrate

with or expand Hadoop technologies. One minute on your favorite search engine will reveal these.

Fact #4. HDFS is a file system, not a database management system (DBMS). Hadoop is primarily a distributed file system and lacks capabilities we would associate with a DBMS, such as indexing, random access to data, and support for SQL. That's okay, because HDFS does things DBMSs cannot do.

Fact #5. Hive resembles SQL but is not standard SQL. Many of us are handcuffed to SQL because we know it well and our tools demand it. People who know SQL can quickly learn to hand-code Hive, but that doesn't solve compatibility issues with SQL-based tools. TDWI feels that over time, Hadoop products will support standard SQL, so this issue will soon be moot.

Fact #6. Hadoop and MapReduce are related but don't require each other. Developers at Google developed MapReduce before HDFS existed, and some variations of MapReduce work with a variety of storage technologies, including HDFS, other file systems, and some DBMSs.

Fact #7. MapReduce provides control for analytics, not analytics per se. MapReduce is a general-purpose execution engine that handles the complexities of network communication, parallel programming, and fault tolerance for any kind of application that you can hand code—not just analytics.

Fact #8. Hadoop is about data diversity, not just data volume. Theoretically, HDFS can manage the storage and access of any data type as long as you can put the data in a file and copy that file into HDFS. As outrageously simplistic as that sounds, it's largely true, and it's exactly what brings many users to Apache HDFS.

Fact #9. Hadoop complements a DW; it's rarely a replacement. Most organizations have designed their DW for structured, relational data, which makes it difficult to wring BI value from unstructured and semistructured data. Hadoop promises to complement DWs by handling the multistructured data types most DWs can't.

Fact #10. Hadoop enables many types of analytics, not just Web analytics. Hadoop gets a lot of press about how Internet companies use it for analyzing Web logs and other Web data, but other use cases exist. For example, consider the Big Data coming from sensory devices, such as robotics in manufacturing, RFID in retail, or grid monitoring in utilities. Older analytic applications that need large data samples—such as customer-base segmentation, fraud detection, and risk analysis—can benefit from the additional Big Data managed by Hadoop. Likewise, Hadoop's additional data can expand 360-degree views to create a more complete and granular view.

Spark versus Hadoop

To leverage the value proposition of the Big Data phenomenon, over the last couple of decades, several novel data processing frameworks are proposed. Among the ones proposed to date, the two most commonly referenced and widely used ones have been Hadoop and Spark. Both of these open-source frameworks are developed by the Apache Software Foundation (Hadoop is released in 2004 and Spark is released in 2009) to manage and process modern-day Big Data sets.

Hadoop (or more appropriately Apache Hadoop), as described in more detail above, is a well-known, one of the first successfully developed and deployed framework to cope with Big Data. It can handle not only very large sizes of data (i.e., volume), but also a wide range of data types (i.e., variety) created at a unprecedented speed (i.e., velocity). Key benefits of Hadoop include (1) handling the Big Data with commodity hardware, (2) preventing loss of data and information due to hardware failures through replications, (3) scaling from a small cluster to a very large analytics system, and (4) enabling the discovery of knowledge from Big Data cost effectively and efficiently.

Spark (or more appropriately Apache Spark) is also an open-source Big Data processing system developed to address the challenges of Big Data, and to create value

from the new portrait of the digital world. Similar to Hadoop, Spark also decomposes large data management and analysis tasks into smaller pieces to be handled by large volumes of different nodes built on regular storage and computing hardware devices. The key difference is that Spark uses in-memory processing (i.e., random access memory, or RAM) to cache and analyze the data (instead of a file system, which has been the case in Hadoop) to create result significantly faster than Hadoop. Many believe that the key motivation behind developing Spark several years after Hadoop was the dire need for computational efficiency. Some of the most commonly cited benefits of Spark include (1) order of magnitude faster processing of big data, (2) a unified engine that supports highly efficient SQL queries, streaming data, machine learning and graph processing, and (3) a revamped APIs designed for ease of use, especially for processing of unstructured and semistructured data.

These two frameworks differ in their ecosystems. While Hadoop ecosystem consists of four primary modules (i.e., Hadoop Distributed File System (HDFS), Yet Another Resource Negotiator (YARN), Hadoop MapReduce, Hadoop Common (Hadoop Core)), Spark ecosystem consists of five modules (i.e., Spark Core, Spark SQL, Spark Streaming and Structured Streaming, Machine Learning Library (MLlib), GraphX). Some of the key comparison points between Hadoop and Spark are as follows:

- **Performance:** Spark is faster because it uses random access memory (RAM) instead of reading and writing intermediate data to disks. In contrast, Hadoop stores data on multiple sources and processes it in batches via MapReduce.
- **Cost:** Hadoop runs at a lower cost since it relies on any disk storage type for data processing. Spark runs at a higher cost because it relies on in-memory computations for real-time data processing, which requires it to use high quantities of RAM to spin up nodes.
- **Parallel processing:** Though both platforms process data in parallel in a distributed environment, Hadoop is ideal for batch processing and linear data processing whereas Spark is ideal for real-time processing and processing live unstructured data streams.
- **Scalability:** When data volume rapidly grows, Hadoop quickly scales to accommodate the demand via Hadoop Distributed File System (HDFS). In turn, Spark relies on the fault tolerant HDFS for large volumes of data.
- **Security:** Spark enhances security with authentication via shared secret or event logging, whereas Hadoop uses multiple authentication and access control methods. Though, overall, Hadoop is more secure, Spark can integrate with Hadoop to reach a higher security level.
- **Analytics:** Spark is the superior platform in this category because it includes MLlib, which performs iterative in-memory ML computations. It also includes tools that perform regression, classification, persistence, pipeline construction, evaluation, etc.

There are misconceptions when it comes to comparing Hadoop and Spark. Here are a few most common ones:

- Spark replaced Hadoop. Although there is a partial truth to it, it is not entirely correct. Both Hadoop and Spark have their respective use cases (as mentioned below).
- Hadoop is a database and Spark is an analytics engine. This is not true as both Hadoop and Spark can be used as both data management and data analysis tools.
- Hadoop is cheaper. Although both Hadoop and Spark are free open-source frameworks, professional installation and use of these tools are anything but free, requiring significant investment on consultancy, data scientists, and appropriate hardware infrastructure.

- Spark is always order of magnitude (up to 100 time) faster than Hadoop. That is not true. Although for small data processing tasks Spark can perform 100 times faster than Hadoop, its efficiency diminishes significantly for very large data processing tasks.

So, how do we decide when to use Hadoop and when to use Spark:

Use Hadoop for:

- Processing big data sets in environments where data size exceeds available memory
- Batch processing with tasks that exploit disk read and write operations
- Building data analysis infrastructure with a limited budget
- Completing jobs that are not time-sensitive
- Historical and archive data analysis

Use Spark for:

- Dealing with chains of parallel operations by using iterative algorithms
- Achieving quick results with in-memory computations
- Analyzing stream data analysis in real time
- Graph-parallel processing to model data
- All ML applications

NoSQL

A related new style of database called **NoSQL** (Not Only SQL) has emerged to, like Hadoop, process large volumes of multistructured data. However, whereas Hadoop is adept at supporting large-scale, batch-style historical analysis, NoSQL databases are aimed, for the most part (though there are some important exceptions), at serving up discrete data stored among large volumes of multistructured data to end-user and automated Big Data applications. This capability is sorely lacking from relational database technology, which simply can't maintain needed application performance levels at a Big Data scale.

In some cases, NoSQL and Hadoop work in conjunction. The aforementioned HBase, for example, is a popular NoSQL database modeled after Google BigTable that is often deployed on top of HDFS, the Hadoop Distributed File System, to provide low-latency, quick lookups in Hadoop. The downside of most NoSQL databases today is that they trade ACID (atomicity, consistency, isolation, durability) compliance for performance and scalability. Many also lack mature management and monitoring tools. Both of these shortcomings are in the process of being overcome by the open-source NoSQL communities and a handful of vendors that are attempting to commercialize the various NoSQL databases. NoSQL databases currently available include HBase, Cassandra, MongoDB, Accumulo, Riak, CouchDB, and DynamoDB, among others.

Data for Good

The recent years has witnessed a significant proliferation in volunteered and crowd-sourced community initiative. Although a significant amount of these initiatives has focused on algorithms (mathematical, statistical, and machine learning), programming languages (e.g., Python and R), and software tools (e.g., KNIME and Orange), some of them have been focusing on solving socially and environmentally meaningful problems and to discover new patterns and knowledge. The latter types of initiatives are often called “data for good” projects. The fundamental idea behind these projects is to do good with data by using mostly publicly available data sources and business analytics and data science tools and techniques to create new understanding, awareness, and solutions to problems that affect the environment, large populations, and often less fortunate groups or minorities. Table 3.5 lists and briefly defines some of the most prolific of these projects.

TABLE 3.5 Data for Good Project Examples

Name of Initiative	Sponsor	Description	URL
TIBCO4Good	TIBCO	An initiative is fueled by a passion for solving worldwide challenges through data. The goal is to make a bigger impact for humanitarian efforts around the world by mobilizing the ecosystem through strategic programs.	https://www.tibco.com/tibco4good .
Data for Good	SAS	SAS is proud to be part of the Data for Good movement, which encourages using data in meaningful ways to solve humanitarian issues around poverty, health, human rights, education, and the environment.	https://www.sas.com/en_us/data-for-good.html
Science for Social Good	IBM	Applying artificial intelligence, cloud, and deep science to scale social impact where applied science can help solve the world's toughest problems and inspire business innovations.	https://research.ibm.com/science-for-social-good/
Doing good with Data	Teradata	Teradata is dedicated to helping non-profit organizations around the globe use data to tackle the problems they face every day.	https://www.teradata.com/About-Us/Doing-Good-With-Data
DataKind	Tableau	An organization that aims to harness the power of data science in the service of humanity, and brings solutions to big social problems. DataKind is considered a leader in the Data for Good movement.	https://www.datakind.org/
Data for Good	A Canadian Foundation	A collective of do gooders, who want to use their powers for good, and not evil, to help make our communities better through data. This is a national not for profit organization, with chapters across the country (i.e., Canada), that help other not for profit, and non-governmental, organizations harness the power of their data to make more informed and better decisions in their quest to make their communities flourish.	https://dataforgood.ca/
Data for Good	Facebook	The goal is to empower partners with privacy-preserving data that strengthens communities and advances social issues. Data for Good is helping organizations respond to the COVID-19 pandemic.	https://dataforgood.facebook.com/
Data & Analytics for Good	Open Journal	Data & Analytics for Good is an open access, peer-reviewed data journal with an international focus that is dedicated to publishing novel research that furthers one or more of the United Nations' sustainable development goals.	https://data-for-good.pubpub.org/
Data for Good	Appsilon	The team behind this initiative has gathered significant expertise in interactive data visualization, machine learning, and developing artificial intelligence solutions and put these technologies to work in resolving the world's most pressing challenges.	https://appsilon.com/data-for-good/
Democratizing Data	Data.org	Data.org is a platform for partnerships to build the field of data science for social impact. We work with organizations from all over the world to increase the use of data science in order to improve the lives of millions of people.	https://data.org/

There are many local Data for Good initiatives. If you are interested in advancing your knowledge and experience in business analytics and data science, it would be a good idea to get involved in some of these community-based volunteered initiatives. That way, while learning analytics, you can also make an impact on environment preservation and improving the wellness of humanity.

SECTION 3.7 REVIEW QUESTIONS

1. What are the common characteristics of emerging Big Data technologies?
2. What is Hadoop? How does it work?
3. What are the main Hadoop components? What functions do they perform?
4. What is MapReduce? What does it do? How does it do it?
5. What is Spark and how does it compare to Hadoop? What are the similarities and differences?
6. What is NoSQL? How does it fit into the Big Data analytics picture?
7. What is Data for Good? Select, study, and report on two of the Data for Good initiatives. In your report, make sure to comment on the origins, motivations, and outcomes of the initiatives.

3.8 BIG DATA AND STREAM ANALYTICS

Along with volume and variety, as we have seen earlier in this chapter, one of the key characteristics that defines Big Data is velocity, which refers to the speed at which the data is created and streamed into the analytics environment. Organizations are looking for new means to process streaming data as it comes in to react quickly and accurately to problems and opportunities to please their customers and to gain a competitive advantage. In situations where data streams in rapidly and continuously, traditional analytics approaches that work with previously accumulated data (i.e., data at rest) often either arrive at the wrong decisions because of using too much out-of-context data, or they arrive at the correct decisions but too late to be of any use to the organization. Therefore, it is critical for a number of business situations to analyze the data soon after it is created and/or as soon as it is streamed into the analytics system.

The presumption that the vast majority of modern-day businesses are currently living by is that it is important and critical to record every piece of data because it might contain valuable information now or sometime in the near future. However, as long as the number of data sources increases, the “store-everything” approach becomes harder and harder and, in some cases, not even feasible. In fact, despite technological advances, current total storage capacity lags far behind the digital information being generated in the world. Moreover, in the constantly changing business environment, real-time detection of meaningful changes in data as well as of complex pattern variations within a given short time window is essential to come up with the actions that better fit with the new environment. These facts become the main triggers for a paradigm that we call *stream analytics*. The stream analytics paradigm was born as an answer to these challenges, namely, the unbounded flows of data that cannot be permanently stored to be subsequently analyzed, in a timely and efficient manner, and complex pattern variations that need to be detected and acted on as soon as they happen.

Stream analytics (also called *data-in-motion analytics* and *real-time data analytics*, among others) is a term commonly used for the analytic process of extracting actionable information from continuously flowing/streaming data. A stream is defined as a continuous sequence of data elements (Zikopoulos et al., 2013). The data elements in a stream are often called *tuples*. In a relational database sense, a tuple is similar to a row of data (a record, an object, an instance). However, in the context of semistructured or unstructured data, a tuple is an abstraction that represents a package of data, which can be characterized as a set of attributes for a given object. If a tuple by itself is not sufficiently informative for analysis or a correlation—or other collective relationships among tuples are needed—then a window of data that includes a set of tuples is used. A window of data is a finite number/sequence of tuples, where the windows are continuously updated

as new data become available. The size of the window is determined based on the system being analyzed. Stream analytics is becoming increasingly more popular because of two things. First, time-to-action has become an ever-decreasing value, and second, we have the technological means to capture and process the data while it is created.

Some of the most impactful applications of stream analytics were developed in the energy industry, specifically for smart grid (electric power supply chain) systems. The new smart grids are capable of not only real-time creation and processing of multiple streams of data to determine optimal power distribution to fulfill real customer needs, but also generating accurate short-term predictions aimed at covering unexpected demand and renewable energy generation peaks. Figure 3.10 shows a depiction of a generic use case for streaming analytics in the energy industry (a typical smart grid application). The goal is to accurately predict electricity demand and production in real time by using streaming data that is coming from smart meters, production system sensors, and meteorological models. The ability to predict near future consumption/production trends and detect anomalies in real time can be used to optimize supply decisions (how much to produce, what sources of production to use, and optimally adjust production capacities) as well as to adjust smart meters to regulate consumption and favorable energy pricing.

Applications of Stream Analytics

Because of its power to create insight instantly, helping decision makers to be on top of events as they unfold and allowing organizations to address issues before they become problems, the use of streaming analytics is on an exponentially increasing trend. The following are some of the application areas that have already benefited from stream analytics.

e-Commerce

Companies like Amazon and eBay (among many others) are trying to make the most out of the data that they collect while a customer is on their Web site. Every page visit, every product looked at, every search conducted, and every click made is recorded and

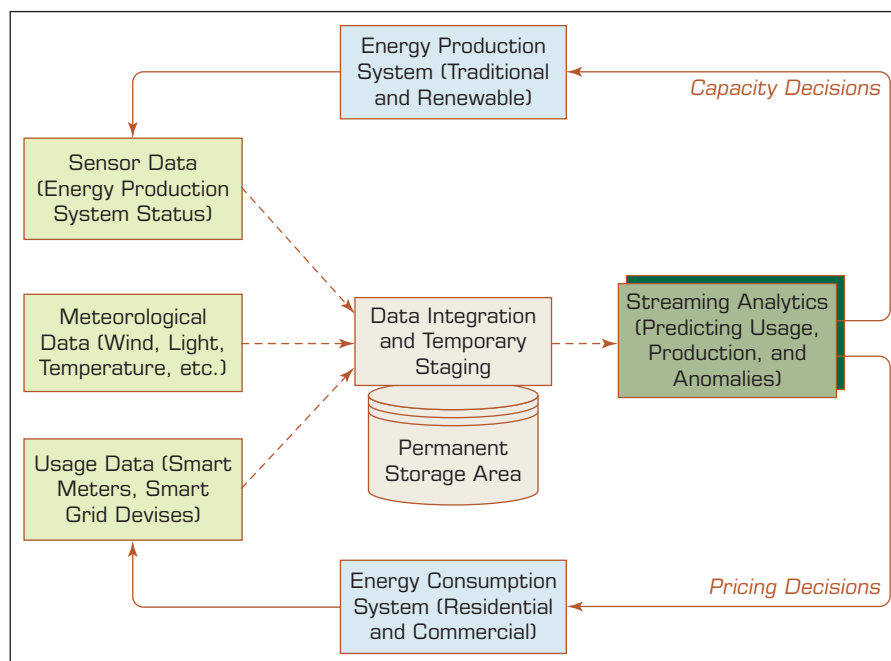


FIGURE 3.10 A Use Case of Streaming Analytics in the Energy Industry.

analyzed to maximize the value gained from a user's visit. If done quickly, analysis of such a stream of data can turn browsers into buyers and buyers into shopaholics. When we visit an e-commerce Web site, even the ones where we are not a member, after a few clicks here and there we start to get very interesting product and bundle price offers. Behind the scenes, advanced analytics are crunching the real-time data coming from our clicks, and the clicks of thousands of others, to "understand" what it is that we are interested in (in some cases, even we do not know that) and make the most of that information by making creative offerings.

Telecommunications

The volume of data that come from call detail records (CDR) for telecommunications companies is astounding. Although this information has been used for billing purposes for quite some time now, there is a wealth of knowledge buried deep inside this Big Data that the telecommunications companies are just now realizing to tap. For instance, CDR data can be analyzed to prevent churn by identifying networks of callers, influencers, leaders, and followers within those networks and proactively acting on this information. As we all know, influencers and leaders have the effect of changing the perception of the followers within their network toward the service provider, either positively or negatively. Using social network analysis techniques, telecommunication companies are identifying the leaders and influencers and their network participants to better manage their customer base. In addition to churn analysis, such information can also be used to recruit new members and maximize the value of the existing members.

Continuous streams of data that come from CDR can be combined with social media data (sentiment analysis) to assess the effectiveness of marketing campaigns. Insight gained from these data streams can be used to rapidly react to adverse effects (which may lead to loss of customers) or boost the impact of positive effects (which may lead to maximizing purchases of existing customers and recruitment of new customers) observed in these campaigns. Furthermore, the process of gaining insight from CDR can be replicated for data networks using Internet protocol detail records. Because most telecommunications companies provide both of these service types, a holistic optimization of all offerings and marketing campaigns could lead to extraordinary market gains.

Law Enforcement and Cybersecurity

Streams of Big Data provide excellent opportunities for improved crime prevention, law enforcement, and enhanced security. They offer unmatched potential when it comes to security applications that can be built in the space, such as real-time situational awareness, multimodal surveillance, cyber-security detection, legal wiretapping, video surveillance, and face recognition (Zikopoulos et al., 2013). As an application of information assurance, enterprises can use streaming analytics to detect and prevent network intrusions, cyberattacks, and malicious activities by streaming and analyzing network logs and other Internet activity monitoring resources.

Power Industry

Because of the increasing use of smart meters, the amount of real-time data collected by power utilities is increasing exponentially. Moving from once a month to every 15 minutes (or more frequently), meter reading accumulates large quantities of invaluable data for power utilities. These smart meters and other sensors placed all around the power grid are sending information back to the control centers to be analyzed in real time. Such analyses help utility companies to optimize their supply chain decisions (e.g., capacity adjustments, distribution network options, real-time buying or selling) based on the

up-to-the-minute consumer usage and demand patterns. In addition, utility companies can integrate weather and other natural conditions data into their analytics to optimize power generation from alternative sources (e.g., wind, solar) and to better forecast energy demand on different geographic granulations. Similar benefits also apply to other utilities such as water and natural gas.

Financial Services

Financial service companies are among the prime examples where analysis of Big Data streams can provide faster and better decisions, competitive advantage, and regulatory oversight. The ability to analyze fast-paced, high-volumes of trading data at very low latency across markets and countries offers a tremendous advantage to making the split-second buy/sell decisions that potentially translate into big financial gains. In addition to optimal buy/sell decisions, stream analytics can also help financial service companies in real-time trade monitoring to detect fraud and other illegal activities.

Health Sciences

Modern-era medical devices (e.g., electrocardiograms and equipment that measure blood pressure, blood oxygen level, blood sugar level, and body temperature) are capable of producing invaluable streaming diagnostic/sensory data at a very fast rate. Harnessing this data and analyzing it in real time offers benefits—the kind that we often call “life and death”—unlike any other field. In addition to helping healthcare companies become more effective and efficient (and hence more competitive and profitable), stream analytics is also improving patient conditions and saving lives.

Many hospital systems all around the world are developing care infrastructures and health systems that are futuristic. These systems aim to take full advantage of what the technology has to offer, and more. Using hardware devices that generate high-resolution data at a very rapid rate, coupled with super-fast computers that can synergistically analyze multiple streams of data, increases the chances of keeping patients safe by quickly detecting anomalies. These systems are meant to help human decision makers make faster and better decisions by being exposed to a multitude of information as soon as it becomes available.

Government

Governments around the world are trying to find ways to be more efficient (via optimal use of limited resources) and effective (providing the services that people need and want). As the practices for e-government become mainstream, coupled with widespread use and access to social media, very large quantities of data (both structured and unstructured) are at the disposal of government agencies. Proper and timely use of these Big Data streams differentiates proactive and highly efficient agencies from the ones who are still using traditional methods to react to situations as they unfold. Another way in which government agencies can leverage real-time analytics capabilities is to manage natural disasters such as snowstorms, hurricanes, tornadoes, and wildfires through a surveillance of streaming data coming from radar, sensors, and other smart detection devices. They can also use similar approaches to monitor water quality, air quality, and consumption patterns and detect anomalies before they become significant problems. Another area where government agencies use stream analytics is in traffic management in congested cities. By using the data coming from traffic flow cameras, GPS data coming from commercial vehicles, and traffic sensors embedded in roadways, agencies are able to change traffic light sequences and traffic flow lanes to ease the pain caused by traffic congestion problems.

SECTION 3.8 REVIEW QUESTIONS

1. What is stream analytics? How does it differ from regular analytics?
2. What are the motivations for stream analytics?
3. What are the most fruitful industries for stream analytics?
4. How can stream analytics be used in e-commerce?
5. In addition to what is listed in this section, can you think of other industries and/or application areas where stream analytics can be used?
6. Compared to regular analytics, do you think stream analytics will have more (or less) use cases in the era of Big Data analytics? Why?

3.9 STATISTICAL MODELING FOR BUSINESS ANALYTICS

Because of the increasing popularity of business analytics, the traditional statistical methods and underlying techniques are also regaining their attractiveness as enabling tools to support evidence-based managerial decision making. Not only are they regaining attention and admiration, but this time around, they are attracting business users in addition to statisticians and analytics professionals.

Statistics (statistical methods and underlying techniques) is usually considered as part of descriptive analytics (see Figure 3.11). Some of the statistical methods can also be considered as part of predictive analytics such as discriminant analysis, multiple regression, logistic regression, and k-means clustering. As shown in Figure 3.11, descriptive analytics has two main branches: statistics and **online analytics processing (OLAP)**. OLAP is the term used for analyzing, characterizing, and summarizing structured data stored in organizational databases (often stored in a data warehouse or in a data mart) using cubes (i.e., multidimensional data structures that are created to extract a subset of data values to answer a specific business question). The OLAP branch of descriptive analytics has also been called business intelligence. Statistics, on the other hand, helps to characterize the data either one variable at a time or multivariables all together, using either descriptive or inferential methods.

Statistics—a collection of mathematical techniques to characterize and interpret data—has been around for a very long time. Many methods and techniques have been developed to address the needs of the end users and the unique characteristics of the

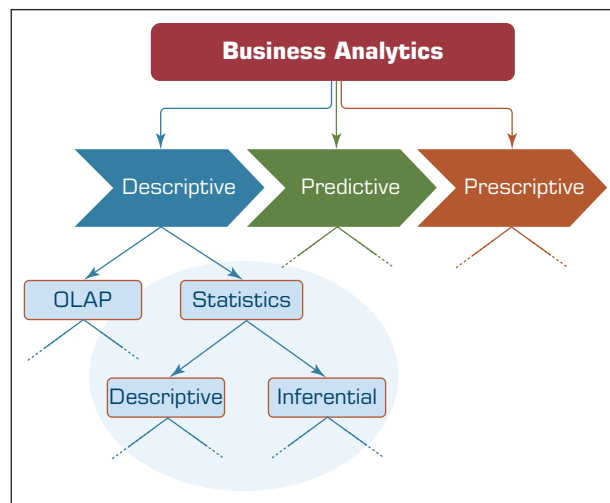


FIGURE 3.11 Relationship between Statistics and Descriptive Analytics.

data being analyzed. Generally speaking, at the highest level, statistical methods can be classified as either descriptive or inferential. The main difference between descriptive and inferential statistics is the data used in these methods—whereas **descriptive statistics** is all about describing the sample data on hand, and **inferential statistics** is about drawing inferences or conclusions about the characteristics of the population. In this section we will briefly describe descriptive statistics (because of the fact that it lays the foundation for, and is the integral part of, descriptive analytics), and in the following section we will cover regression (both linear and logistic regression) as part of inferential statistics.

Descriptive Statistics for Descriptive Analytics

Descriptive statistics, as the name implies, describes the basic characteristics of the data at hand, often one variable at a time. Using formulas and numerical aggregations, descriptive statistics summarizes the data in such a way that often meaningful and easily understandable patterns emerge from the study. Although it is very useful in data analytics and very popular among the statistical methods, descriptive statistics does not allow making conclusions (or inferences) beyond the sample of the data being analyzed. That is, it is simply a nice way to characterize and describe the data on hand, without making conclusions (inferences or extrapolations) regarding the population of related hypotheses we might have in mind.

In business analytics, descriptive statistics plays a critical role—it allows us to understand and explain/present our data in a meaningful manner using aggregated numbers, data tables, or charts/graphs. In essence, descriptive statistics helps us convert our numbers and symbols into meaningful representations for anyone to understand and use. Such an understanding not only helps business users in their decision-making processes, but also helps analytics professionals and data scientists to characterize and validate the data for other more sophisticated analytics tasks. Descriptive statistics allows analysts to identify data concentration, unusually large or small values (i.e., outliers), and unexpectedly distributed data values for numeric variables. Therefore, the methods in descriptive statistics can be classified as either measures for central tendency or measures of dispersion. In the following section, we will use a simple description and mathematical formulation/representation of these measures. In mathematical representation, we will use $x_1, x_2, \dots, x_n$ to represent individual values (observations) of the variable (measure) that we are interested in characterizing.

Measures of Centrality Tendency (May Also Be Called Measures of Location or Centrality)

Measures of centrality are the mathematical methods by which we estimate or describe central positioning of a given variable of interest. A measure of central tendency is a single numerical value that aims to describe a set of data by simply identifying or estimating the central position within the data. The mean (often called the arithmetic mean or the simple average) is the most commonly used measure of central tendency. In addition to mean, you could also see median or mode being used to describe the centrality of a given variable. Although, the mean, median, and mode are all valid measures of central tendency, under different circumstances, one of these measures of centrality becomes more appropriate than the others. What follows are short descriptions of these measures, including how to calculate them mathematically and pointers on the circumstances in which they are the most appropriate measure to use.

Arithmetic Mean

The **arithmetic mean** (or simply *mean* or *average*) is the sum of all the values/observations divided by the number of observations in the data set. It is by far the most popular and most commonly used measure of central tendency. It is used with continuous or

discrete numeric data. For a given variable x , if we happen to have n values/observations ($x_1, x_2, \dots, x_n$), we can write the arithmetic mean of the data sample ($\bar{x}$, pronounced as x-bar) as follows:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

or

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

The mean has several unique characteristics. For instance, the sum of the absolute deviations (differences between the mean and the observations) above the mean is the same as the sum of the deviations below the mean, balancing the values on either side of it. That said, it does not suggest, however, that half the observations are above and the other half are below the mean (a common misconception among those who do not know basic statistics). Also, the mean is unique for every data set and is meaningful and calculable for both interval- and ratio-type numeric data. One major downside is that the mean can be affected by outliers (observations that are considerably larger or smaller than the rest of the data points). Outliers can pull the mean toward their direction and, hence, bias the centrality representation. Therefore, if there are outliers or if the data is erratically dispersed and skewed, one should either avoid using mean as the measure of centrality or augment it with other central tendency measures, such as median and mode.

Median

The **median** is the measure of center value in a given data set. It is the number in the middle of a given set of data that has been arranged/sorted in order of magnitude (either ascending or descending). If the number of observation is an odd number, identifying the median is very easy—just sort the observations based on their values and pick the value right in the middle. If the number of observations is an even number, then identify the two middle values, and then take the simple average of these two values. The median is meaningful and calculable for ratio, interval, and ordinal data types. Once determined, half the data points in the data are above and the other half are below the median. In contrary to the mean, the median is not affected by outliers or skewed data.

Mode

The **mode** is the observation that occurs most frequently (the most frequent value in our data set). On a histogram it represents the highest bar in a bar chart, and hence it may be considered as being the most popular option/value. The mode is most useful for data sets that contain a relatively small number of unique values. That is, it may be useless if the data have too many unique values (as is the case in many engineering measurements that capture high precision with a large number of decimal places), rendering each value having either one or a very small number representing its frequency. Although it is a useful measure (especially for nominal data), mode is not a very good representation of centrality, and therefore it should not be used as the only measure of central tendency for a given data set.

In summary, which central tendency measure is the best? Although there is not a clear answer to this question, here are a few hints—use the mean when the data is not prone to outliers and there is no significant level of skewness; use the median when the data has outliers and/or it is ordinal in nature; use the mode when the data is nominal. Perhaps the best practice is to use all three together so that the central tendency of the data set can be captured and represented from three perspectives. Mostly because

“average” is a very familiar and highly used concept to everyone in regular daily activities, managers (as well as some scientists and journalists) often use the centrality measures (especially mean) inappropriately when other statistical information should be considered along with the centrality. It is a better practice to present descriptive statistics as a package—a combination of centrality and dispersion measures—as opposed to a single measure like mean.

Measures of Dispersion

Measures of **dispersion** (it may also be called measures of spread or decentrality) are the mathematical methods used to estimate or describe the degree of variation in a given variable of interest. They are a representation of the numerical spread (compactness or lack thereof) of a given data set. To describe this dispersion, a number of statistical measures are developed; the most notable ones are range, variance, and standard deviation (and also quartiles and absolute deviation). One of the main reasons why the measures of dispersion/spread of data values are important is the fact that it gives us a framework within which we can judge the central tendency—gives us the indication of how well the mean (or other centrality measures) represents the sample data. If the dispersion of values in the data set is large, the mean is not deemed to be a very good representation of the data. This is because a large dispersion measure indicates large differences between individual scores. Also, in research, it is often perceived as a positive sign to see a small variation within each data sample, as it may indicate homogeneity, similarity, and robustness within the collected data.

Range

The **range** is perhaps the simplest measure of dispersion. It is the difference between the largest and the smallest values in a given data set (i.e., variables). So, we calculate range by simply identifying the smallest value in the data set (minimum), identifying the largest value in the data set (maximum), and calculating the difference between them (range = maximum – minimum).

Variance

A more comprehensive and sophisticated measure of dispersion is the **variance**. It is a method used to calculate the deviation of all data points in a given data set from the mean. The larger the variance, the more the data are spread out from the mean and the more variability one can observe in the data sample. To prevent the offsetting of negative and positive differences, the variance takes into account the square of the distances from the mean. The formula for a data sample can be written as

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

where n is the number of samples, $\bar{x}$ is the mean of the sample, and x_i is the i^{th} value in the data set. The larger values of variance indicate more dispersion, whereas smaller values indicate compression in the overall data set. Because the differences are squared, larger deviations from the mean contribute significantly to the value of variance. Again, because the differences are squared, the numbers that represent deviation/variance become somewhat meaningless (as opposed to a dollar difference, herein you are given a squared dollar difference). Therefore, instead of variance, in many business applications we use a more meaningful dispersion measure, called standard deviation.

Standard Deviation

The **standard deviation** is also a measure of the spread of values within a set of data. The standard deviation is calculated by simply taking the square root of the variations. The following formula shows the calculation of standard deviation from a given sample of data points.

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

Mean Absolute Deviation

In addition to variance and standard deviation, sometimes we also use **mean absolute deviation** to measure dispersion in a data set. It is a simpler way to calculate the overall deviation from the mean. Specifically, it is calculated by measuring the absolute values of the differences between each data point and the mean and summing them. It provides a measure of spread without being specific about the data point being lower or higher than the mean. The following formula shows the calculation of the mean absolute deviation:

$$MAD = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

Quartiles and Interquartile Range

Quartiles help us identify spread within a subset of the data. A **quartile** is a quarter of the number of data points given in a data set. Quartiles are determined by first sorting the data and then splitting the sorted data into four disjoint smaller data sets. Quartiles are a useful measure of dispersion because they are much less affected by outliers or a skewness in the data set than the equivalent measures in the whole data set. Quartiles are often reported along with the median as the best choice of measure of dispersion and central tendency, respectively, when dealing with skewed and/or data with outliers. A common way of expressing quartiles is as an interquartile range, which describes the difference between the third quartile (Q3) and the first quartile (Q1), telling us about the range of the middle half of the scores in the distribution. The quartile-driven descriptive measures (both centrality and dispersion) are best explained with a popular plot called a box plot (or box-and-whiskers plot).

Box-and-Whiskers Plot

The **box-and-whiskers plot** (or simply a **box plot**) is a graphical illustration of several descriptive statistics about a given data set. They can be either horizontal or vertical, but vertical is the most common representation, especially in modern-day analytics software products. It is known to be first created and presented by John W. Tukey in 1969. Box plot is often used to illustrate both centrality and dispersion of a given data set (i.e., the distribution of the sample data) in an easy-to-understand graphical notation. Figure 3.12 shows a couple of box plots side by side, sharing the same y -axis. As shown therein, a single chart can have one or more box plots for visual comparison purposes. In such cases, the y -axis would be the common measure of magnitude (the numerical value of the variable), with the x -axis showing different classes/subsets such as different time dimensions (e.g., descriptive statistics for annual Medicare expenses in 2015 versus 2016) or different categories (e.g., descriptive statistics for marketing expenses versus total sales).

Although, historically speaking, the box plot was not used widely and often enough (especially in areas outside of statistics), with the emerging popularity of business

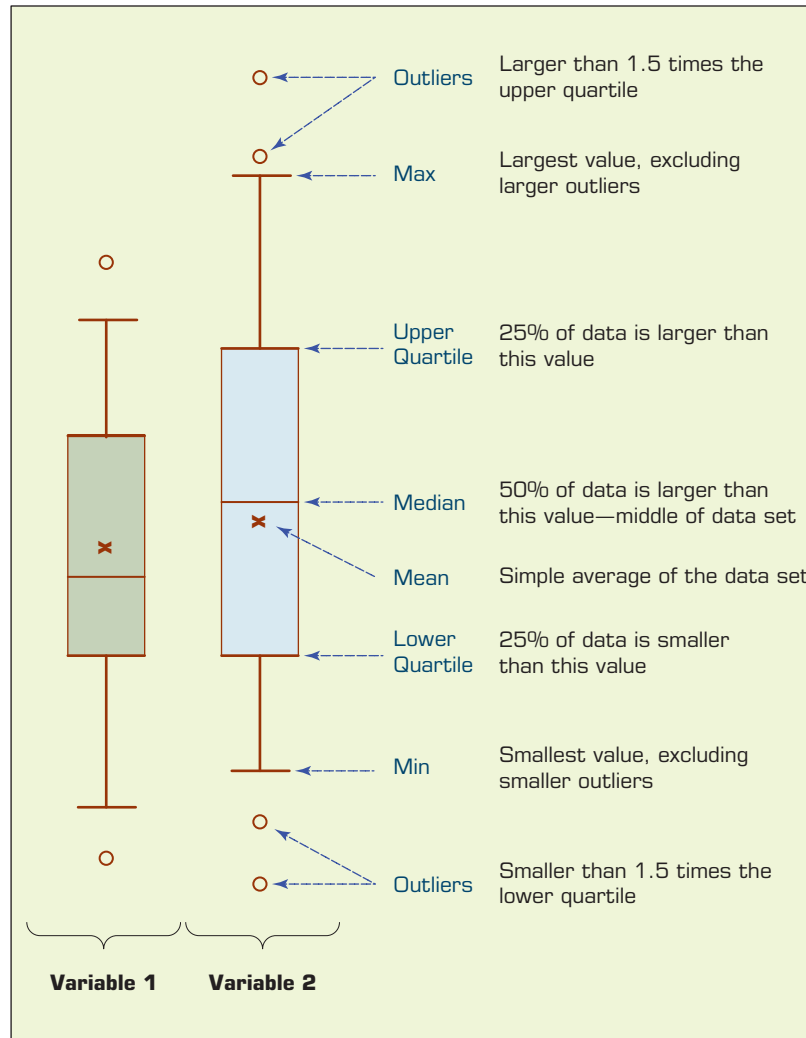


FIGURE 3.12 Understanding the Specifics about Box-and-Whiskers Plots.

analytics, it is gaining fame in less-technical areas of the business world. Its information richness and ease of understanding are largely to credit for its recent popularity.

The box plot shows the **centrality** (median, and sometimes also mean) as well as the dispersion (the density of the data within the middle half—drawn as a box between the first and third quartile), the minimum and maximum ranges (shown as extended lines from the box, looking like whiskers, that are calculated as 1.5 times the upper or lower end of the quartile box) along with the outliers that are larger than the limits of the whiskers. A box plot also shows whether the data is symmetrically distributed with respect to the mean, or it sways one way or another. The relative position of the median versus mean and the lengths of the whiskers on both side of the box give a good indication of the potential skewness in the data.

The Shape of a Distribution

Although not as common as the centrality and dispersion, the shape of the data distribution is also a useful measure for the descriptive statistics. Before delving into the shape of the distribution we first need to define the distribution itself. Simply put, distribution

is the frequency of data points counted and plotted over a small number of class labels or numerical ranges (i.e., bins). In a graphical illustration of distribution, the y -axis shows the frequency (count or %), and the x -axis shows the individual classes or bins in a rank-ordered fashion. A very well-known distribution is called normal distribution, which is perfectly symmetric on both sides of the mean and has several well-founded mathematical properties that make it a very useful statistical tool for research and practice. As the dispersion of a data set increases, so does the standard deviation, and the shape of the distribution looks wider. A graphic illustration of the relationship between dispersion and distribution shape (in the context of normal distribution) is shown in Figure 3.13.

There are two commonly used measures to calculate the shape characteristics of a distribution: skewness and kurtosis. A **histogram** (i.e., frequency plot) is often used to visually illustrate both skewness and kurtosis.

Skewness is a measure of asymmetry (sway) in a distribution of the data that portrays a unimodal (i.e., only one peak exists in the distribution) structure of the data. Because normal distribution is a perfectly symmetric unimodal distribution, it does not have skewness, that is, its skewness measure (i.e., the value of the coefficient of skewness) is equal to zero. The skewness measure/value can be either positive or negative. If the distribution sways left (i.e., the peak is on the left and the long tail is on the right side, and the mean is greater than median), then it produces a positive skewness measure, and if the distribution sways right (i.e., the peak is on the right and the long tail is

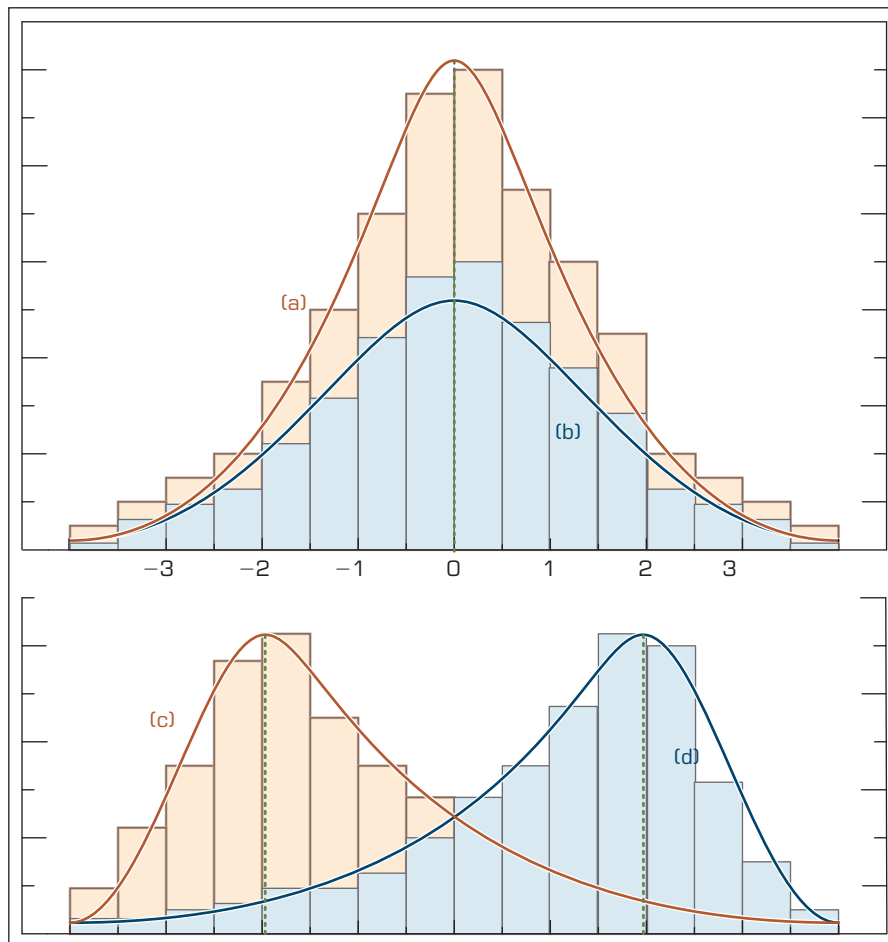


FIGURE 3.13 Relationship between Dispersion and Shape Properties.

on the left side, and the mean is smaller than median), then it produces a negative skewness measure. In Figure 3.13, (c) represents a positively skewed distribution, whereas (d) represents a negatively skewed distribution. In the same figure, both (a) and (b) represent perfect symmetry and hence zero measure for skewness.

$$\text{Skewness} = S = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)s^3}$$

where s is the standard deviation and n is the number of samples.

Kurtosis is another measure to use in characterizing the shape of a unimodal distribution. As opposed to the sway in shape, kurtosis focuses on characterizing the peakiness/tallness nature of the distribution. Specifically, kurtosis measures the degree to which a distribution is more or less peaked than a normal distribution. Whereas a positive kurtosis indicates a relatively peaked/tall distribution, a negative kurtosis indicates a relatively flat/short distribution. As a reference point, a normal distribution has a kurtosis of 3. The formula for kurtosis can be written as

$$\text{Kurtosis} = K = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{ns^4} - 3$$

Descriptive statistics (as well as inferential statistics) can easily be calculated using commercial statistical software packages (e.g., SAS, SPSS, Minitab, JMP, Statistica) or free/open-source tools (e.g., R, Python, KNIME). Perhaps the most convenient way to calculate descriptive and some of the inferential statistics is to use Excel. Technology Insight 3.3 describes in detail how to use Microsoft Excel to calculate descriptive statistics.

TECHNOLOGY INSIGHT 3.3 How to Calculate Descriptive Statistics in Microsoft Excel

Excel, arguably the most popular data analysis tool in the world, can easily be used for descriptive statistics. Although the base configuration of Excel does not seem to have the statistics function readily available for end users, those functions come with the installation and can be activated (turned on) with only a few mouse clicks. Figure 3.14 shows how these statistics functions (as part of the Analysis ToolPak) can be activated in Microsoft Excel.

Once activated, the *Analysis ToolPak* will appear in the *Data* menu option under the name of *Data Analysis*. When you click on Data Analysis in the Analysis group under the Data tab in the Excel menu bar, you will see Descriptive Statistics as one of the options within the list of data analysis tools (see Figure 3.15, steps [1, 2]); click on OK, and the Descriptive Statistics dialog box will appear (see middle of Figure 3.15). In this dialog box you need to enter the range of the data, which can be one or more numerical columns, along with the preference check boxes, and click OK (see Figure 3.15, steps [3, 4]). If the selection includes more than one numeric column, the tool treats each column as a separate data set and provides descriptive statistics for each column separately.

As a simple example, we selected two columns (labeled as Expense and Demand) and executed the Descriptive Statistics option. The bottom section of Figure 3.15 shows the output created by Excel. As can be seen, Excel produced all the descriptive statistics that are covered in the previous section and added a few more to the list. In Excel, it is also very easy (a few mouse clicks) to create a box-and-whiskers plot. Figure 3.16 shows the simple three-step process of creating a box-and-whiskers plot in Excel.

Although this is a very useful tool in Excel, one should be aware of an important point related to the results generated by Analysis ToolPak, which have a different behavior than other ordinary Excel functions: Although Excel functions dynamically change as the underlying

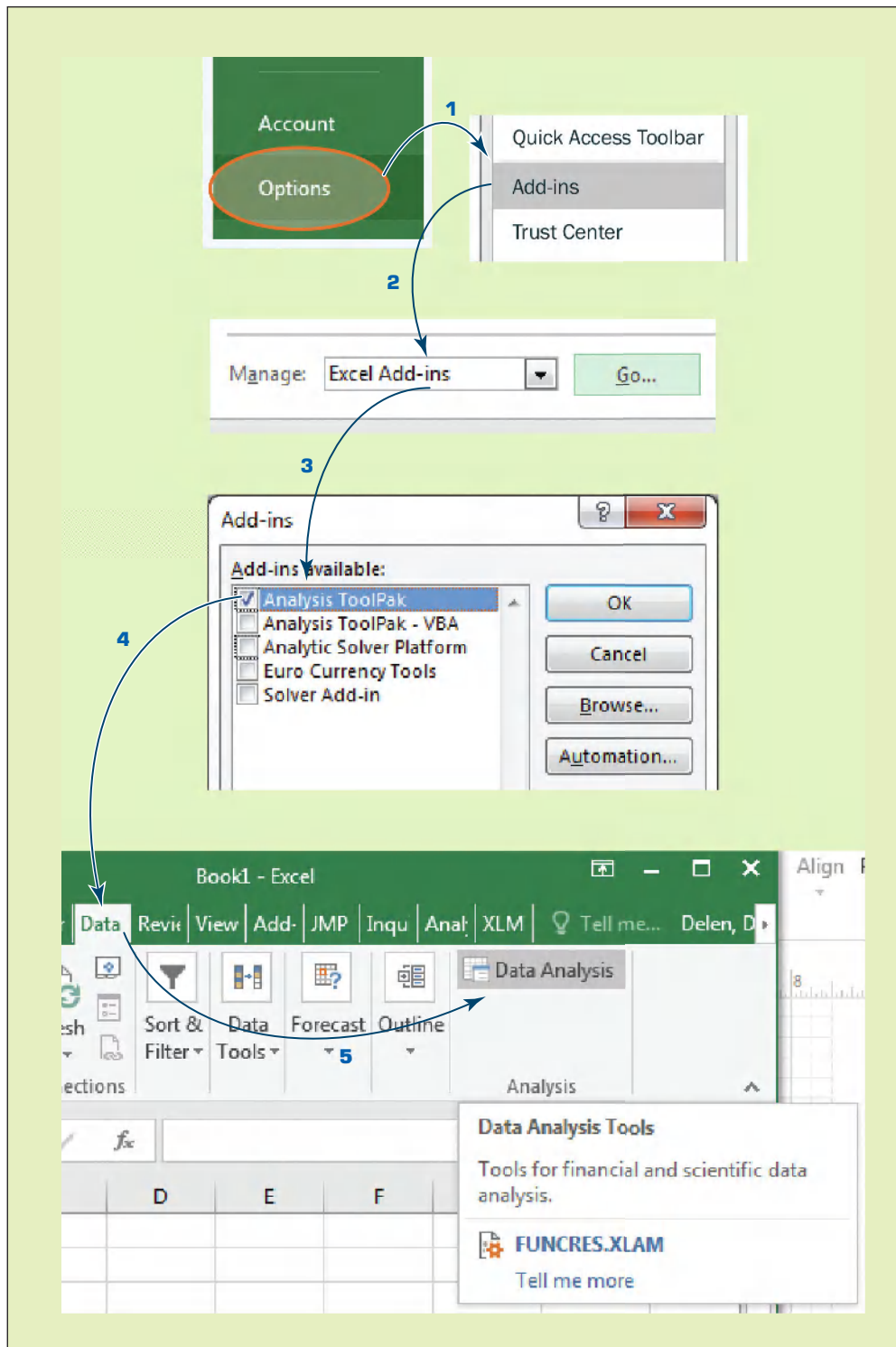
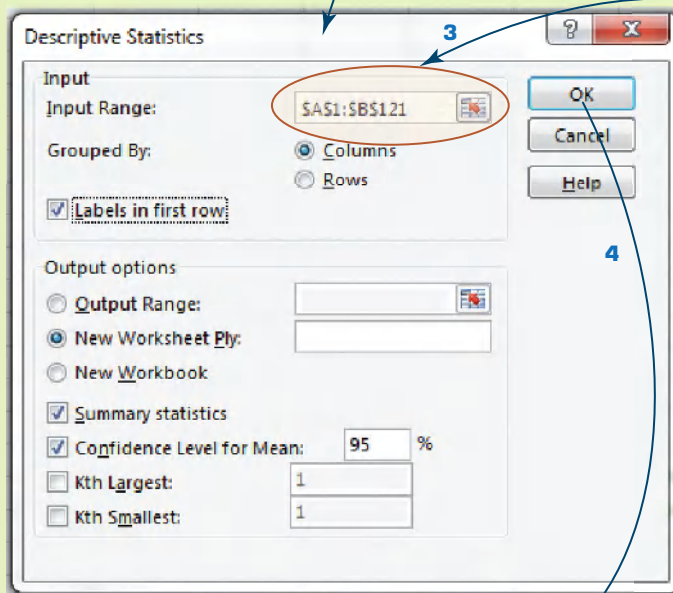
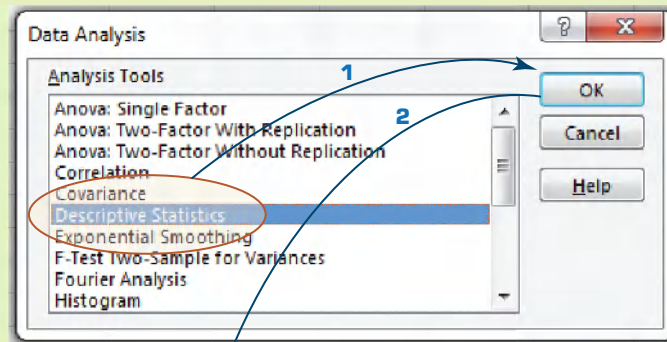


FIGURE 3.14 Activating Data Analysis Functions in Excel.



	A	B
1	Expense	Demand
2	8553	14824
3	9107	14567
4	8137	14509
5	6999	12395
6	6009	11031
7	8140	12945
8	5811	11558
9	7787	14018
10	6032	11828
11	7146	11359
12	7379	11663
13	7988	13167
14	6236	11943
15	6967	11524
16	8765	13722
17	6865	11095

	A	B	C	D
1	Expense		Demand	
2				
3	Mean	7595.808333	Mean	12976.53333
4	Standard Error	78.07347263	Standard Error	109.7357916
5	Median	7679	Median	12945
6	Mode	8140	Mode	14567
7	Standard Deviation	855.252042	Standard Deviation	1202.095369
8	Sample Variance	731456.0554	Sample Variance	1445033.276
9	Kurtosis	-0.82138214	Kurtosis	-1.290541404
10	Skewness	-0.235479407	Skewness	-0.002281156
11	Range	3526	Range	3944
12	Minimum	5811	Minimum	11019
13	Maximum	9337	Maximum	14963
14	Sum	911497	Sum	1557184
15	Count	120	Count	120
16	Confidence Level(95.0%)	154.5932735	Confidence Level(95.0%)	217.2878275

FIGURE 3.15 Obtaining Descriptive Statistics in Excel.

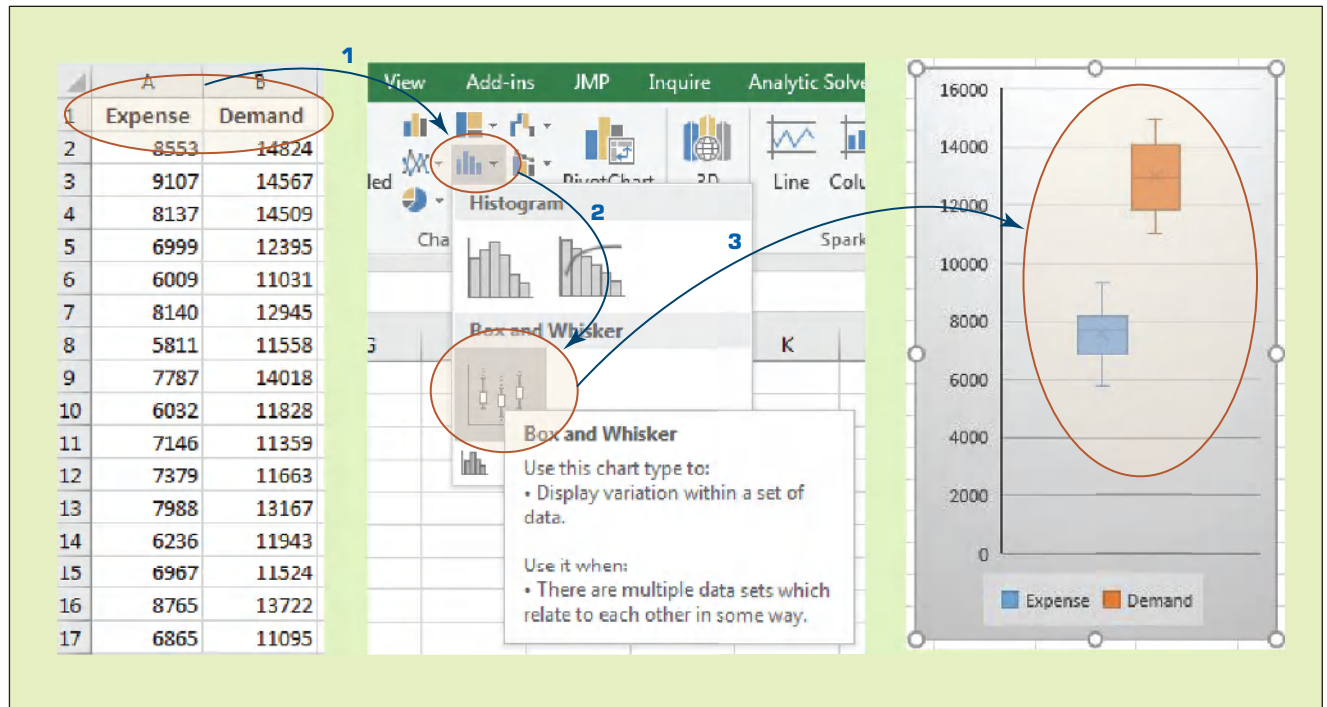


FIGURE 3.16 Creating a Box-and-Whiskers Plot in Excel.

data in the spreadsheet are changed, the results generated by the Analysis ToolPak do not. For example, if you change the values in either or both of these columns, the Descriptive Statistics results produced by the Analysis ToolPak will stay the same. However, the same is not true for ordinary Excel functions. If you were to calculate the mean value of a given column (using `=AVERAGE(A1:A121)`), and then change the values within the data range, the mean value would automatically change. In summary, the results produced by Analysis ToolPak do not have a dynamic link to the underlying data, and if the data changes, the analysis needs to be redone using the dialog box.

Successful applications of data analytics cover a wide range of business and organizational settings, addressing the problems once thought unsolvable. Analytics in Action 3.2 is an excellent illustration of those success stories where a small municipality administration adopts a data analytics approach to intelligently detect and solve problems by continuously analyzing demand and consumption patterns.

ANALYTICS IN ACTION 3.2 Town of Cary Uses Analytics to Analyze Data from Sensors, Assess Demand, and Detect Problems

A leaky faucet. A malfunctioning dishwasher. A cracked sprinkler head. These are more than just a headache for a home owner or business to fix. They can be costly, unpredictable, and, unfortunately, hard to pinpoint. Through a combination of wireless water meters and a data-analytics-driven, customer-accessible portal, the Town of Cary, North Carolina, is making it much easier to find and fix water loss issues. In the process, the town has gained a big-picture view of water usage critical to planning future water plant expansions and promoting targeted conservation efforts.

When the Town of Cary installed the wireless meters for 60,000 customers in 2010, it knew the new technology wouldn't just save money by eliminating manual monthly readings; the town also realized it would get more accurate and timely information about water consumption. The Aquastar wireless system reads meters once an hour—that's 8,760 data points per customer each year instead of 12 monthly readings. The data had tremendous potential, if it could be easily consumed.

"Monthly readings are like having a gallon of water's worth of data. Hourly meter readings are more like an Olympic-size pool of data," says Karen Mills, Finance Director for the Town of Cary. "SAS helps us manage the volume of that data nicely." In fact, the solution enables the town to analyze a half-billion data points on water usage and make them available, and easily consumable, to all customers.

The ability to visually look at data by household or commercial customer, by the hour, has led to some very practical applications:

- The town can notify customers of potential leaks within days.
- Customers can set alerts that notify them within hours if there is a spike in water usage.
- Customers can track their water usage online, helping them to be more proactive in conserving water.

Through the online portal, one business in the Town of Cary saw a spike in water consumption on weekends, when employees are away. This seemed odd, and the unusual reading helped the company learn that a commercial dishwasher was malfunctioning, running continuously over weekends. Without the wireless water-meter data and the customer-accessible portal, this problem could have gone unnoticed, continuing to waste water and money.

The town has a much more accurate picture of daily water usage per person, critical for planning future water plant expansions. Perhaps the most interesting perk is that the town was able to verify a hunch that has far-reaching cost ramifications: Cary residents are very economical in their use of water. "We calculate that with modern high-efficiency appliances, indoor water use could be as low as 35 gallons per person per day. Cary residents average 45 gallons, which is still phenomenally low," explains town Water Resource Manager Leila Goodwin. Why is this important? The town was spending money to encourage water efficiency—rebates on low-flow toilets or discounts on rain barrels. Now it can take a more targeted approach, helping specific consumers understand and manage both their indoor and outdoor water use.

SAS was critical not just for enabling residents to understand their water use, but also in working behind the scenes to link two disparate databases. "We have a billing database and the meter-reading database. We needed to bring that together and make it presentable," Mills says.

The town estimates that by just removing the need for manual readings, the Aquastar system will save more than \$10 million above the cost of the project. But the analytics component could provide even bigger savings. Already both the town and individual citizens have saved money by catching water leaks early. As the Town of Cary continues to plan its future infrastructure needs, having accurate information on water usage will help it invest in the right amount of infrastructure at the right time. In addition, understanding water usage will help the town if it experiences something detrimental like a drought.

"We went through a drought in 2007," says Goodwin. "If we go through another, we have a plan in place to use Aquastar data to see exactly how much water we are using on a day-by-day basis and communicate with customers. We can show 'here's what's happening, and here is how much you can use because our supply is low.' Hopefully, we'll never have to use it, but we're prepared."

Source: "Municipality puts wireless water meter-reading data to work (SAS® Analytics) – The Town of Cary, North Carolina uses SAS Analytics to analyze data from wireless water meters, assess demand, detect problems and engage customers" (accessed May 2022). Copyright © 2022 SAS Institute Inc., Cary, NC, USA. Reprinted with permission. All rights reserved.

SECTION 3.9 REVIEW QUESTIONS

1. What is the relationship between statistics and business analytics?
2. What are the main differences between descriptive and inferential statistics?
3. List and briefly define the central tendency measures of descriptive statistics.
4. List and briefly define the dispersion measures of descriptive statistics.
5. What is a box-and-whiskers plot? What types of statistical information does it represent?
6. What are the two most commonly used shape characteristics to describe a data distribution?

3.10 REGRESSION MODELING FOR INFERENTIAL STATISTICS

Regression, especially linear regression, is perhaps the most widely known and used analytics technique in statistics. Historically speaking, the roots of regression date back to the 1920s and 1930s, to the earlier work on inherited characteristics of sweet peas by Sir Francis Galton and subsequently by Karl Pearson. Since then regression has become the statistical technique for characterization of relationships between explanatory (input) variable(s) and response (output) variable(s).

As popular as it is, essentially, regression is a relatively simple statistical technique to model the dependence of a variable (response or output variable) on one (or more) explanatory (input) variables. Once identified, this relationship between the variables can be formally represented as a linear/additive function/equation. As is the case with many other modeling techniques, regression aims to capture the functional relationship between and among the characteristics of the real world and describe this relationship with a mathematical model, which may then be used to discover and understand the complexities of reality—explore and explain relationships or forecast future occurrences.

Regression can be used for one of two purposes: hypothesis testing—investigating potential relationships between different variables, and prediction/forecasting—estimating values of a response variables based on one or more explanatory variables. These two uses are not mutually exclusive. The explanatory power of regression is also the foundation of its prediction ability. In hypothesis testing (theory building), regression analysis can reveal the existence/strength and the directions of relationships between a number of explanatory variables (often represented with x_i) and the response variable (often represented with y). In prediction, regression identifies additive mathematical relationships (in the form of an equation) between one or more explanatory variables and a response variable. Once determined, this equation can be used to forecast the values of the response variable for a given set of values of the explanatory variables.

Correlation versus Regression Because regression analysis originated from correlation studies, and because both methods attempt to describe the association between two (or more) variables, these two terms are often confused by professionals and even by scientists. **Correlation** makes no a priori assumption of whether one variable is dependent on the other(s) and is not concerned with the relationship between variables; instead it gives an estimate on the degree of association between the variables. On the other hand, regression attempts to describe the dependence of a response variable on one (or more) explanatory variables where it implicitly assumes that there is a one-way causal effect from the explanatory variable(s) to the response variable, regardless of whether the path of effect is direct or indirect. Also, although correlation is interested in the low-level relationships between two variables, regression is concerned with the relationships between all explanatory variables and the response variable.

Simple versus Multiple Regression If the regression equation is built between one response variable and one explanatory variable, then it is called simple regression. For instance, the regression equation built to predict/explain the relationship between a height of a person (explanatory variable) and the weight of a person (response variable) is a good example of simple regression. Multiple regression is the extension of simple regression where the explanatory variables are more than one. For instance, in the previous example, if we were to include not only the height of the person but also other personal characteristics (e.g., BMI, gender, ethnicity) to predict the weight of a person, then we would be performing multiple regression analysis. In both cases, the relationship between the response variable and the explanatory variable(s) are linear and additive in nature. If the relationships are not linear, then we may want to use one of many other nonlinear regression methods to better capture the relationships between the input and output variables.

How Do We Develop the Linear Regression Model?

To understand the relationship between two variables, the simplest and most intuitive thing that one can do is to create a scatter plot, where the y -axis represents the values of the response variable, and the x -axis represents the values of the explanatory variable (see Figure 3.17). A **scatter plot** would show the changes in the response variable as a function of the changes in the explanatory variable. In the case shown in Figure 3.17, there seems to be a positive relationship between the two; as the explanatory variable values increase, so does the response variable.

Simple regression analysis aims to find a mathematical representation of the relationship between two variables. In reality, it tries to find the signature (i.e., algebraic representation) of a straight line passing through right in between the plotted dots (representing the observation/historical data) in such a way that the distance between the dots and the line is minimized (the predicted values on the theoretical regression line). Even

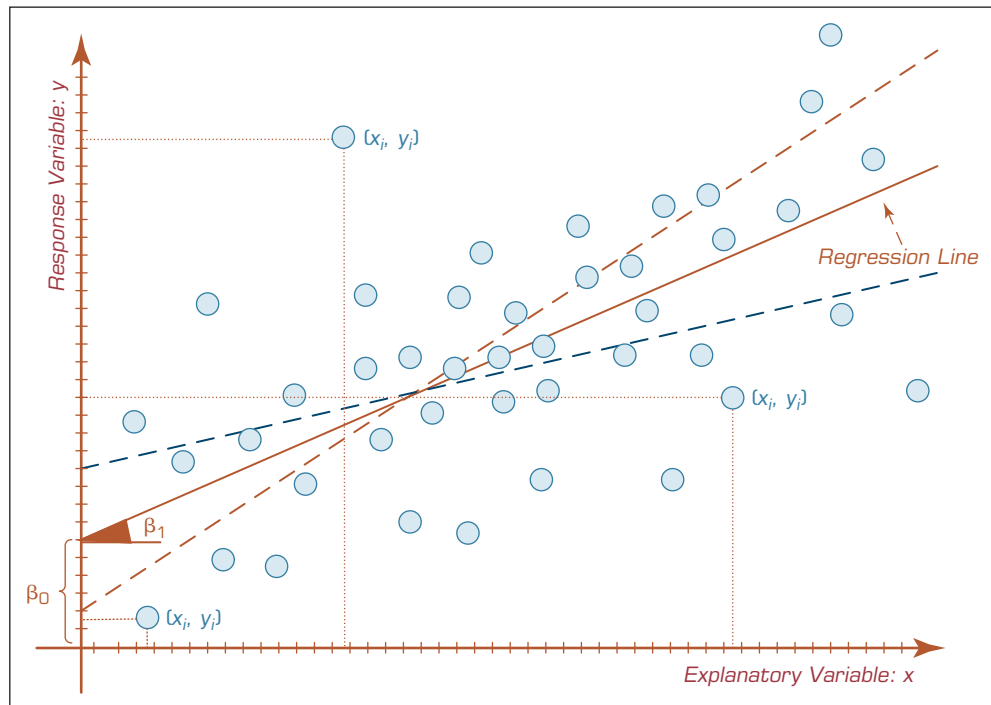


FIGURE 3.17 A Scatter Plot and a Linear Regression Line.

though there are several methods/algorithms proposed to identify the regression line, the one that is most commonly used is called the **ordinary least squares (OLS)** method. The OLS method aims to minimize the sum of squared residuals (squared vertical distances between the observation and the regression point) and leads to a mathematical expression for the estimated value of the regression line (which are known as β parameters). For simple **linear regression**, the aforementioned relationship between the response variable (y) and the explanatory variable(s) (x) can be shown as a simple equation as follows:

$$y = \beta_0 + \beta_1 x$$

In this equation, β_0 is called the intercept and β_1 is called the slope. Once OLS determines the values of these two coefficients, the simple equation can be used to forecast the values of y for given values of x . The sign and the value of β_1 also reveal the direction and the strengths of relationship between the two variables.

If the model is of a multiple linear regression type, then there would be more coefficients to be determined, one for each additional explanatory variable. As the following formula shows, the additional explanatory variable would be multiplied with the new β_i coefficients and summed together to establish a linear additive representation of the response variable.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_n x_n$$

How Do We Know If the Model Is Good Enough?

Because of a variety of reasons, sometimes models as representations of the reality do not prove to be good. Regardless of the number of explanatory variables included, there is always a possibility of not having a good model, and therefore the linear regression model needs to be assessed for its fit (the degree at which it represents the response variable). In the simplest sense, a well-fitting regression model results in predicted values close to the observed data values. For the numerical assessment, three statistical measures are often used in evaluating the fit of a regression model. R^2 (R -squared), the overall F-test, and the root mean square error (RMSE). All three of these measures are based on the sums of the square errors (how far the data are from the mean and how far the data are from the model's predicted values). Different combinations of these two values provide different information about how the regression model compares to the mean model.

Of the three, R^2 has the most useful and understandable meaning because of its intuitive scale. The value of R^2 ranges from zero to one (corresponding to the amount of variability explained in percentage) with zero indicating that the relationship and the prediction power of the proposed model is not good, and one indicating that the proposed model is a perfect fit that produces exact predictions (which is almost never the case). The good R^2 values would usually come close to one, and the closeness is a matter of the phenomenon being modeled—whereas an R^2 value of 0.3 for a linear regression model in social sciences can be considered good enough, an R^2 value of 0.7 in engineering may be considered as not a good-enough fit. The improvement in the regression model can be achieved by adding more explanatory variables, taking some of the variables out of the model, or using different data transformation techniques, which would result in comparative increases in an R^2 value. Figure 3.18 shows the process flow of developing regression models. As can be seen in the process flow, the model development task is followed by the model assessment task, where not only is the fit of the model assessed, but because of restrictive assumptions with which the linear models have to comply, also the validity of the model needs to be put under the microscope.

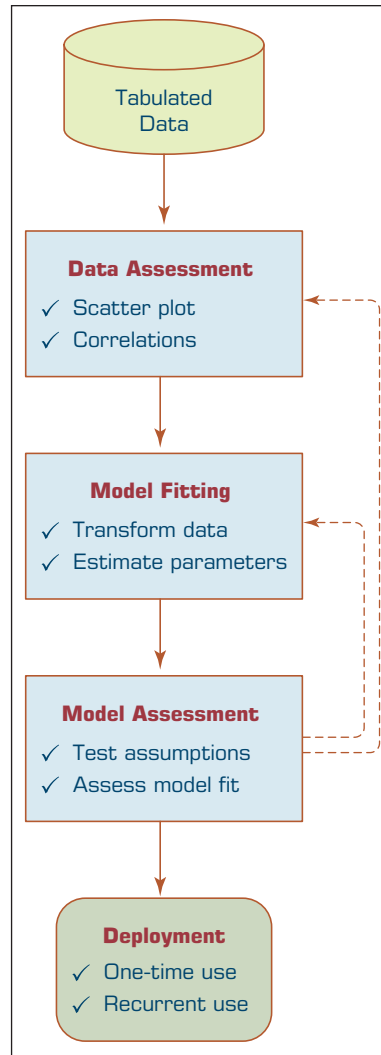


FIGURE 3.18 A Process Flow for Developing Regression Models.

What Are the Most Important Assumptions in Linear Regression?

Even though they are still the choice of many for data analyses (both for explanatory as well as for predictive modeling purposes), linear regression models suffer from several highly restrictive assumptions. The validity of the linear model built depends on its ability to comply with these assumptions. Here are the most commonly pronounced assumptions:

- **Linearity.** This assumption states that the relationship between the response variable and the explanatory variables are linear. That is, the expected value of the response variable is a straight-line function of each explanatory variable, while holding all other explanatory variables fixed. Also, the slope of the line does not depend on the values of the other variables. It also implies that the effects of different explanatory variables on the expected value of the response variable are additive in nature.
- **Independence** (of errors). This assumption states that the errors of the response variable are uncorrelated with each other. This independence of the errors is weaker than actual statistical independence, which is a stronger condition and is often not needed for linear regression analysis.

- **Normality** (of errors). This assumption states that the errors of the response variable are normally distributed. That is, they are supposed to be totally random and should not represent any nonrandom patterns.
- **Constant variance** (of errors). This assumption, also called homoscedasticity, states that the response variables have the same variance in their error, regardless of the values of the explanatory variables. In practice, this assumption is invalid if the response variable varies over a wide enough range/scale.
- **Multicollinearity**. This assumption states that the explanatory variables are not correlated (i.e., do not replicate the same but provide a different perspective of the information needed for the model). Multicollinearity can be triggered by having two or more perfectly correlated explanatory variables presented to the model (e.g., if the same explanatory variable is mistakenly included in the model twice, one with a slight transformation of the same variable). A correlation-based data assessment usually catches this error.

There are statistical techniques developed to identify the violation of these assumptions and techniques to mitigate them. The most important part for a modeler is to be aware of their existence and to put in place the means to assess the models to make sure that the models are compliant with the assumptions they are built on.

Logistic Regression

Logistic regression is a very popular, statistically sound, probability-based classification algorithm that employs supervised **learning**. It was developed in the 1940s as a complement to linear regression and linear discriminant analysis methods. It has been used extensively in numerous disciplines, including the medical and social sciences fields. Logistic regression is similar to linear regression in that it also aims to regress to a mathematical function that explains the relationship between the response variable and the explanatory variables using a sample of past observations (training data). It differs from linear regression with one major point: its output (response variable) is a class as opposed to a numerical variable. That is, whereas linear regression is used to estimate a continuous numerical variable, logistic regression is used to classify a categorical variable. Even though the original form of logistic regression was developed for a binary output variable (e.g., 1/0, yes/no, pass/fail, accept/reject), the present-day modified version is capable of predicting multiclass output variables (i.e., multinomial logistic regression). If there is only one predictor variable and one predicted variable, the method is called simple logistic regression (similar to calling linear regression models with only one independent variable as simple linear regression).

In predictive analytics, logistic regression models are used to develop probabilistic models between one or more explanatory/predictor variables (which may be a mix of both continuous and categorical in nature) and a class/response variable (which may be binomial/binary or multinomial/multiclass). Unlike ordinary linear regression, logistic regression is used for predicting categorical (often binary) outcomes of the response variable—treating the response variable as the outcome of a Bernoulli trial. Therefore, logistic regression takes the natural logarithm of the odds of the response variable to create a continuous criterion as a transformed version of the response variable. Thus the logit transformation is referred to as the link function in logistic regression—even though the response variable in logistic regression is categorical or binomial, the logit is the continuous criterion on which linear regression is conducted. Figure 3.19 shows a logistic regression function where the odds are represented in the *x*-axis (a linear function of the independent variables), whereas the probabilistic outcome is shown in the *y*-axis (i.e., response variable values change between 0 and 1).

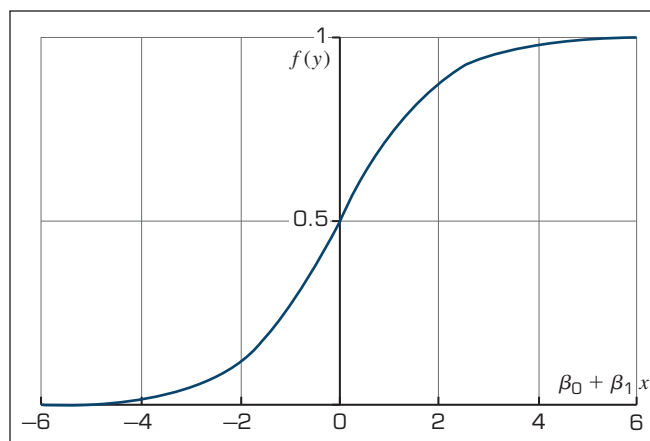


FIGURE 3.19 The Logistic Function.

The logistic function, $f(y)$ in Figure 3.19, is the core of logistic regression, which can only take values between 0 and 1. The following equation is a simple mathematical representation of this function:

$$f(y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

The logistic regression coefficients (the β s) are usually estimated using the maximum likelihood estimation method. Unlike linear regression with normally distributed residuals, it is not possible to find a closed-form expression for the coefficient values that maximizes the likelihood function, so an iterative process must be used instead. This process begins with a tentative starting solution, then revises the parameters slightly to see if the solution can be improved and repeats this iterative revision until no improvement can be achieved or are very minimal, at which point the process is said to have completed/converged.

Sports analytics—use of data and statistical/analytics techniques to better manage sports teams/organizations—has been gaining tremendous popularity. Use of data-driven analytics techniques has become mainstream for not only professional teams but also college and amateur sports. Analytics in Action 3.3 is an excellent example of how existing and readily available public data sources can be used to predict college football bowl game outcomes using both classification and regression-type prediction models.

ANALYTICS IN ACTION 3.3 Predicting NCAA Bowl Game Outcomes

Predicting the outcome of a college football game (or any sports game, for that matter) is an interesting and challenging problem. Therefore, challenge-seeking researchers from both academics and industry have spent a great deal of effort on forecasting the outcome of sporting events. Large quantities of historic data exist in different media outlets (often publicly available) regarding the structure and outcomes of sporting events in the form of a variety of numerically or symbolically represented factors that are assumed to contribute to those outcomes.

The end-of-season bowl games are very important to colleges both financially (bringing in millions of dollars of additional revenue) as well as reputational—for recruiting quality students and highly regarded high school athletes for their athletic programs (Freeman & Brewer, 2016). Teams that are selected to compete in a given bowl game split a purse, the size of which depends on the specific bowl (some bowls are more prestigious and have higher payouts for the two teams), and therefore securing an invitation to a bowl game is the main goal of any division I-A college football program. The decision makers of the bowl games are given the authority



Source: bestv/Shutterstock

to select and invite bowl-eligible (a team that has six wins against its Division I-A opponents in that season) successful teams (as per the ratings and rankings) that will play in an exciting and competitive game, attract fans of both schools, and keep the remaining fans tuned in via a variety of media outlets for advertising.

In an unusual data mining study, Delen, Cogdell, and Kasap (2012) used eight years of bowl game data along with three popular data mining techniques (decision trees, neural networks, and support vector machines) to predict both the classification-type outcome of a game (win versus loss) as well as the regression-type outcome (projected point difference between the scores of the two opponents). What follows is a shorthand description of their study.

METHODOLOGY

In this research, Delen and his colleagues followed a popular data mining methodology called CRISP-DM (Cross-Industry Standard Process for Data Mining), which is a six-step process. This popular methodology, which is covered in detail in Chapter 5, provided them with a systematic and structured way to conduct the underlying data mining study and hence improved the likelihood of obtaining accurate and reliable results. To objectively assess the prediction power of the different model types, they used a cross-validation methodology, called *k*-fold cross-validation. Details on *k*-fold cross-validation can be found in Chapter 5. Figure 3.20 graphically illustrates the methodology employed by the researchers.

DATA ACQUISITION AND DATA PREPROCESSING

The sample data for this study is collected from a variety of sports databases available on the Web, including jhowel.net, ESPN.com, Covers.com, ncaa.org, and raunuzusstreet.com. The data set included 244 bowl games, representing a complete set of eight seasons of college football bowl games played between 2002 and 2009. We also included an out-of-sample data set (2010–2011 bowl games) for additional validation purposes. Exercising one of the popular data mining rules-of-thumb, they included as much relevant information into the model as possible. Therefore, after an in-depth variable identification and collection process, they ended up with a data set that included 36 variables, of which the first 6 were the identifying variables (i.e., name and the year of the bowl game, home and away team names and their athletic conferences—see variables 1–6 in Table 3.6), followed by 28 input variables (which included variables delineating a team's seasonal statistics on offense and defense, game outcomes, team composition characteristics, athletic conference characteristics, and how they fared against the odds—see variables 7–34 in 6), and finally the last two were the output variables (i.e., ScoreDiff—the score difference between the home team and the away team represented with an integer number, and WinLoss—whether the home team won or lost the bowl game represented with a nominal label).

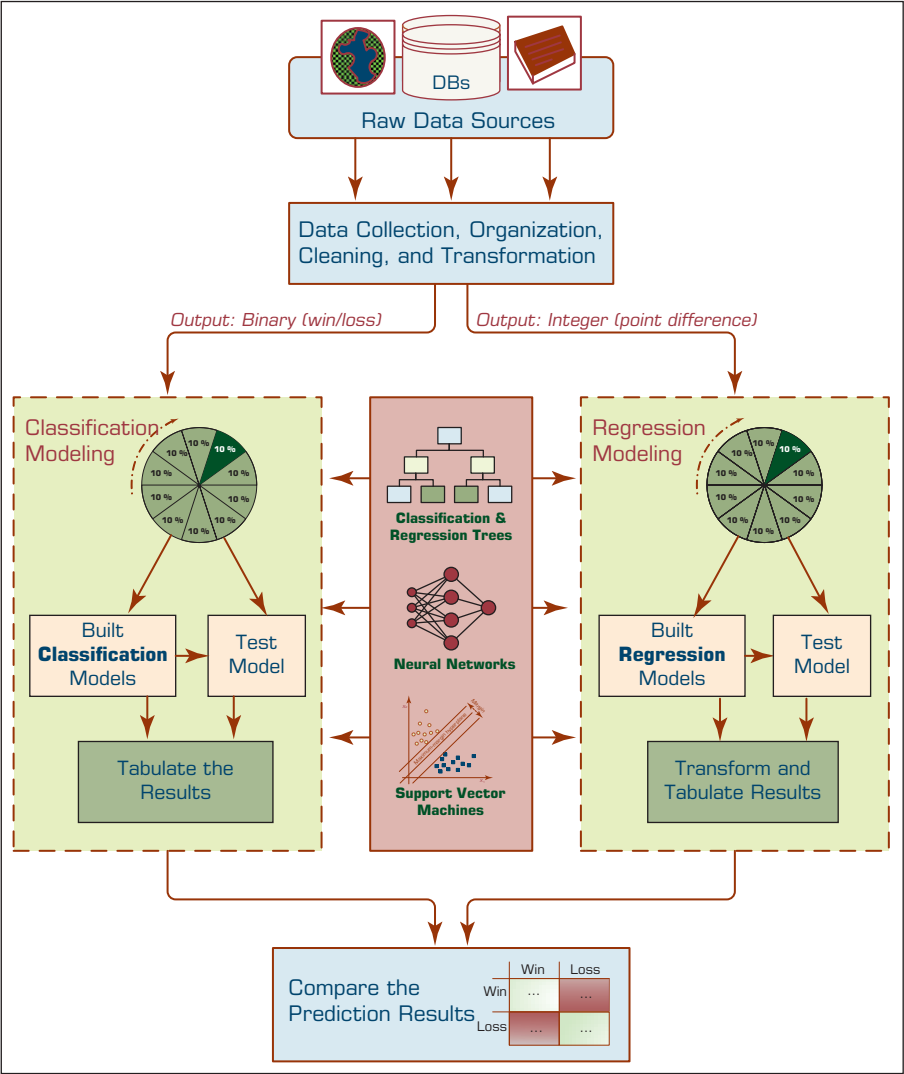


FIGURE 3.20 The Graphical Illustration of the Methodology Employed in the Study.

In the formulation of the data set, each row (a.k.a. tuple, case, sample, example, etc.) represented a bowl game, and each column stood for a variable (i.e., identifier/input or output type). To represent the game-related comparative characteristics of the two opponent teams, in the input variables, we calculated and used the differences between the measures of the home and away teams. All these variable values are calculated from the home team's perspective. For instance, the variable PPG (average number of points a team scored per game) represents the difference between the home team's PPG and away team's PPG. The output variables represent whether the home team wins or loses the bowl game. That is, if the ScoreDiff variable takes a positive integer number, then the home team is expected to win the game by that margin, otherwise (if the ScoreDiff variable takes a negative integer number) then the home team is expected to lose the game by that margin. In the case of WinLoss, the value of the output variable is a binary label, "Win" or "Loss" indicating the outcome of the game for the home team.

RESULTS AND EVALUATION

In this study, three popular prediction techniques are used to build models (and to compare them to each other): artificial neural networks, decision trees, and support vector machines. These prediction techniques are selected based on their capability of modeling both classification as

TABLE 3.6 Description of the Variables Used in the Study

No	Cat	Variable Name	Description
1	ID	YEAR	Year of the bowl game
2	ID	BOWLGAME	Name of the bowl game
3	ID	HOMETEAM	Home team (as listed by the bowl organizers)
4	ID	AWAYTEAM	Away team (as listed by the bowl organizers)
5	ID	HOMECONFERENCE	Conference of the home team
6	ID	AWAYCONFERENCE	Conference of the away team
7	I1	DEFPTPGM	Defensive points per game
8	I1	DEFYDPM	Defensive rush yards per game
9	I1	DEFYDPM	Defensive yards per game
10	I1	PPG	Average number of points a given team scored per game
11	I1	PYDPM	Average total pass yards per game
12	I1	RYDPM	Team's average total rush yards per game
13	I1	YRDPM	Average total offensive yards per game
14	I2	HMWIN%	Home winning percentage
15	I2	LAST7	How many games the team won out of their last 7 games
16	I2	MARGOVIC	Average margin of victory
17	I2	NCTW	Nonconference team winning percentage
18	I2	PREVAPP	Did the team appeared in a bowl game previous year
19	I2	RDWIN%	Road winning percentage
20	I2	SEASTW	Winning percentage for the year
21	I2	TOP25	Winning percentage against AP top 25 teams for the year
22	I3	TSOS	Strength of schedule for the year
23	I3	FR%	Percentage of games played by freshmen class players for the year
24	I3	SO%	Percentage of games played by sophomore class players for the year
25	I3	JR%	Percentage of games played by junior class players for the year
26	I3	SR%	Percentage of games played by senior class players for the year
27	I4	SEASOVUn%	Percentage of times a team went over the O/U* in the current season
28	I4	ATSCOV%	Against the spread cover percentage of the team in previous bowl games
39	I4	UNDER%	Percentage of times a team went under in previous bowl games
30	I4	OVER%	Percentage of times a team went over in previous bowl games

(Continued)

TABLE 3.6 (Continued)

No	Cat	Variable Name	Description
31	I4	SEASATS%	Percentage of covering against the spread for the current season
32	I5	CONCH	Did the team win their respective conference championship game
33	I5	CONFSOS	Conference strength of schedule
34	I5	CONFWIN%	Conference winning percentage
35	O1	ScoreDiff ^o	Score difference (HomeTeamScore – AwayTeamScore)
36	O2	WinLoss ^o	Whether the home team wins or loses the game

*Over/Under—Whether or not a team will go over or under of the expected score difference.

^o Output variables—ScoreDiff for regression models and WinLoss for binary classification models.

I1: Offense/defense; I2: game outcome; I3: team configuration; I4: against the odds; I5: conference stats.

ID: Identifier variables; O1: output variable for regression models; O2: output variable for classification models.

well as regression-type prediction problems and their popularity in recently published data mining literature. More details about these popular data mining methods can be found in Chapter 5.

To compare predictive accuracy of all models to one another, the researchers used a stratified k -fold cross-validation methodology. In a stratified version of k -fold cross-validation, the folds are created in a way that they contain approximately the same proportion of predictor labels (i.e., classes) as the original data set. In this study, the value of k is set to 10 (i.e., the complete set of 244 samples are split into 10 subsets, each having about 25 samples), which is a common practice in predictive data mining applications. A graphical depiction of the 10-fold cross-validations was shown earlier in this chapter. To compare the prediction models that were developed using the aforementioned three data mining techniques, the researchers chose to use three common performance criteria: accuracy, sensitivity, and specificity. The simple formulas for these metrics were also explained earlier in this chapter.

The prediction results of the three modeling techniques are presented in Tables 3.7 and 3.8. Table 3.7 presents the 10-fold cross-validation results of the classification methodology where the three data mining techniques are formulated to have a binary-nominal output variable (i.e., *WinLoss*). Table 3.8 presents the 10-fold cross-validation results of the regression-based classification methodology, where the three data mining techniques are formulated to have a numerical output variable (i.e., *ScoreDiff*). In the regression-based classification prediction, the numerical output of the models is converted to a classification type by labeling the positive *WinLoss* numbers with a “Win” and negative *WinLoss* numbers with a “Loss,” and then tabulating them in the confusion matrixes. Using the confusion matrices, the overall prediction accuracy, sensitivity, and specificity of each model type are calculated and presented in these two tables. As the results indicate, the classification-type prediction methods performed better than regression-based classification-type prediction methodology. Among the three data mining technologies, classification and regression trees produced better prediction accuracy in both prediction methodologies. Overall, classification and regression tree classification models produced a 10-fold cross-validation accuracy of 86.48%, followed by support vector machines (with a 10-fold cross-validation accuracy of 79.51%) and neural networks (with a 10-fold cross-validation accuracy of 75.00%). Using a t -test, researchers found that these accuracy values were significantly different at 0.05 alpha level, that is, the decision tree is a significantly better predictor of this domain than the neural network and support vector machine, and the support vector machine is a significantly better predictor than neural networks.

The results of the study showed that the classification-type models predict the game outcomes better than regression-based classification models. Even though these results are specific to the application domain and the data used in this study, and therefore should not be generalized beyond the scope of the study, they are exciting because decision trees are not only the best predictors but also the best in understanding and deployment, compared to the other two machine-learning techniques employed in this study.

TABLE 3.7 Prediction Results for the Direct Classification Methodology

Prediction Method (Classification*)		Confusion Matrix		Accuracy** (in %)	Sensitivity (in %)	Specificity (in %)
		Win	Loss			
ANN (MLP)	Win	92	42	75.00	68.66	82.73
	Loss	19	91			
SVM (RBF)	Win	105	29	79.51	78.36	80.91
	Loss	21	89			
DT (C&RT)	Win	113	21	86.48	84.33	89.09
	Loss	12	98			

*The output variable is a binary categorical variable (Win or Loss); differences were sig (** $p < 0.01$).

TABLE 3.8 Prediction Results for the Regression-Based Classification Methodology

Prediction Method (Regression-Based*)		Confusion Matrix		Accuracy**	Sensitivity	Specificity
		Win	Loss			
ANN (MLP)	Win	94	40	72.54	70.15	75.45
	Loss	27	83			
SVM (RBF)	Win	100	34	74.59	74.63	74.55
	Loss	28	82			
DT (C&RT)	Win	106	28	77.87	76.36	79.10
	Loss	26	84			

*The output variable is a numerical/integer variable (point-diff); differences were sig (** $p < 0.01$).

Sources: Delen, D., Cogdell, D., & Kasap, N. (2012). A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting*, 28, 543–552; Freeman, K. M., & Brewer, R. M. (2016). The politics of American college football. *Journal of Applied Business and Economics*, 18(2), 97–101.

Time Series Forecasting

Sometimes the variable that we are interested in (i.e., the response variable) may not have distinctly identifiable explanatory variables, or there may be too many of them in a highly complex relationship. In such cases, if the data is available in a desired format, a prediction model, the so-called time series, can be developed. A time series is a sequence of data points of the variable of interest, measured and represented at successive points in time spaced at uniform time intervals. Examples of time series include monthly rain volumes in a geographic area, the daily closing value of the stock market indexes, and daily sales totals for a grocery store. Often, time series are visualized using a line chart. Figure 3.21 shows an example time series of sales volumes for the years 2008 through 2012 in a quarterly basis.

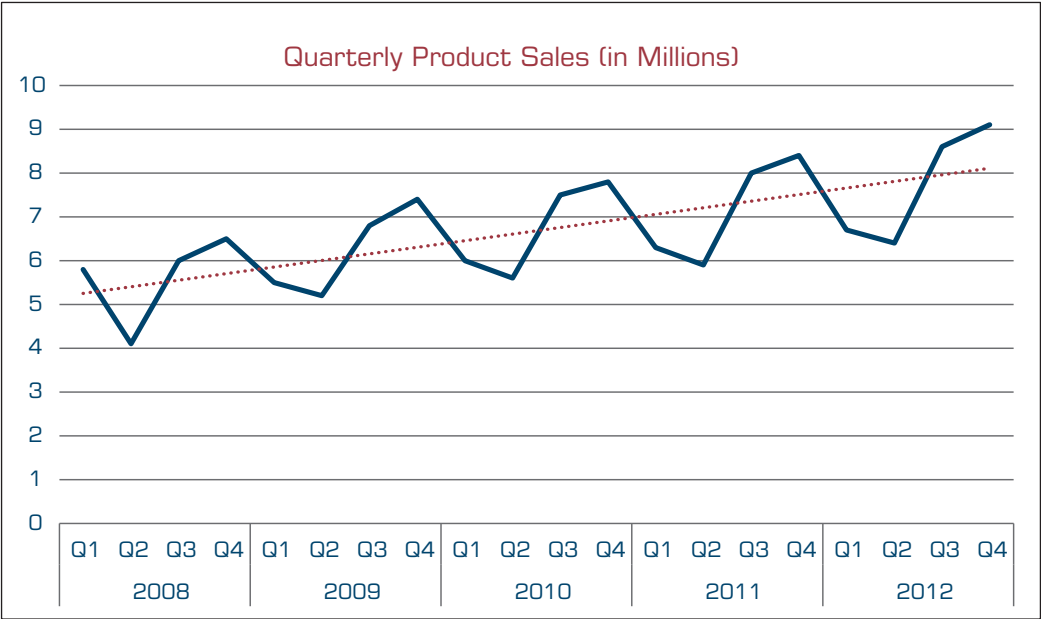


FIGURE 3.21 A Sample Time Series of Data on Quarterly Sales Volumes.

Time series forecasting is the use of mathematical modeling to predict future values of the variable of interest based on previously observed values. The time series plots/charts look and feel very similar to simple linear regression in that as was the case in simple linear regression, in time series there are two variables: the response variable and the time variable presented in a scatter plot. Beyond this look similarity, there is hardly any other commonality between the two. Although regression analysis is often employed in testing theories to see if current values of one or more explanatory variables explain (and hence predict) the response variable, the time series models are focused on extrapolating on their time-varying behavior to estimate the future values.

Time series forecasting assumes all the explanatory variables are aggregated and consumed in the response variable’s time-variant behavior. Therefore, capturing of the time-variant behavior is the way to predict the future values of the response variable. To do that the pattern is analyzed and decomposed into its main components: random variations, time trends, and seasonal cycles. The time series example shown in Figure 3.21 illustrates all these distinct patterns.

The techniques used to develop time series forecasts range from very simple (the naïve forecast that suggests today’s forecast is the same as yesterday’s actual) to very complex like ARIMA (a method that combines autoregressive and moving average patterns in data). Most popular techniques are perhaps the averaging methods that include simple average, moving average, weighted moving average, and exponential smoothing. Many of these techniques also have advanced versions where seasonality and trend can also be taken into account for better and more accurate forecasting. The accuracy of a method is usually assessed by computing its error (calculated deviation between actuals and forecasts for the past observations) via mean absolute error (MAE), mean squared error (MSE), or mean absolute percent error (MAPE). Even though they all use the same core error measure, these three assessment methods emphasize different aspects of the error, some penalizing larger errors more so than the others.

SECTION 3.10 REVIEW QUESTIONS

1. What is regression, and what statistical purpose does it serve?
2. What are the commonalities and differences between regression and correlation?
3. What is OLS? How does OLS determine the linear regression line?
4. List and describe the main steps to follow in developing a linear regression model.
5. What are the most commonly pronounced assumptions for linear regression?
6. What is logistics regression? How does it differ from linear regression?
7. What is time series? What are the main forecasting techniques for time series data?

Chapter Highlights

- Data has become one of the most valuable assets of today's organizations.
- Data is the main ingredient for any BI, data science, and business analytics initiative.
- Although its value proposition is undeniable, to live up its promise, the data has to comply with some basic usability and quality metrics.
- *Data* (datum in singular form) refers to a collection of facts usually obtained as the result of experiments, observations, transactions, or experiences.
- At the highest level of abstraction, data can be classified as structured and unstructured.
- Data in its original/raw form is not usually ready to be useful in analytics tasks.
- Data preprocessing is a tedious, time-demanding, yet crucial task in business analytics.
- Big Data means different things to people with different backgrounds and interests.
- Big Data exceeds the reach of commonly used hardware environments and/or capabilities of software tools to capture, manage, and process it within a tolerable time span.
- Big Data is typically defined by three "V"s: volume, variety, velocity.
- MapReduce is a technique to distribute the processing of very large multistructured data files across a large cluster of machines.
- Hadoop is an open-source framework for processing, storing, and analyzing massive amounts of distributed, unstructured data.
- Hive is a Hadoop-based data warehousing-like framework originally developed by Facebook.
- Pig is a Hadoop-based query language developed by Yahoo!
- NoSQL, which stands for Not Only SQL, is a new paradigm to store and process large volumes of unstructured, semistructured, and multistructured data.
- Data for Good is a popular concept where data and analytics is used for betterment of the humanity and environment.
- Data scientist is a new role or job commonly associated with Big Data or data science.
- Stream analytics is a term commonly used for extracting actionable information from continuously flowing/streaming data sources.
- Data stream mining, as an enabling technology for stream analytics, is the process of extracting novel patterns and knowledge structures from continuous, rapid data records.
- Statistics is a collection of mathematical techniques to characterize and interpret data.
- Statistical methods can be classified as either descriptive or inferential.
- Statistics in general, and descriptive statistics in particular, is a critical part of BI and business analytics.
- Descriptive statistics methods can be used to measure central tendency, dispersion, or the shape of a given data set.
- Regression, especially linear regression, is perhaps the most widely known and used analytics technique in statistics.
- Linear regression and logistic regression are the two major regression types in statistics.
- Logistics regression is a probability-based classification algorithm.
- Time series is a sequence of data points of a variable, measured and recorded at successive points in time spaced at uniform time intervals.

Key Terms

analytics ready	datum	learning	quartile
arithmetic mean	descriptive statistics	linear regression	range
Big Data	dimensional reduction	logistic regression	ratio data
Big Data analytics	dispersion	MapReduce	regression
box-and-whiskers plot	Hadoop	mean absolute deviation	scatter plot
box plot	Hadoop Distributed File System (HDFS)	median	skewness
categorical data	high-performance computing	mode	standard deviation
centrality	histogram	nominal data	statistics
correlation	Hive	NoSQL	stream analytics
data preprocessing	inferential statistics	online analytics processing (OLAP)	structured data
data quality	knowledge	ordinal data	time series forecasting
data scientists	kurtosis	ordinary least squares (OLS)	unstructured data
data security		Pig	variable selection
data taxonomy			variance

Questions for Discussion

- How do you describe the importance of data in analytics? Can we think of analytics without data? Explain.
- Considering the new and broad definition of business analytics, what are the main inputs and outputs to the analytics continuum?
- Where does the data for business analytics come from? What are the sources and the nature of that incoming data?
- What are the most common metrics that make for analytics-ready data?
- What are the main categories of data? What types of data can we use for BI and analytics?
- Can we use the same data representation for all analytics models (i.e., do different analytics models require different data representation schema)? Why, or why not?
- Why is the original/raw data not readily usable by analytics tasks?
- What are the main data preprocessing steps? List and explain their importance in analytics.
- What does it mean to clean/scrub the data? What activities are performed in this phase?
- Data reduction can be applied to rows (sampling) and/or columns (variable selection). Which is more challenging? Explain.
- What is Big Data? Why is it important? Where does Big Data come from?
- What do you think the future of Big Data will be? Will it lose its popularity to something else? If so, what will it be?
- What is Big Data analytics? How does it differ from regular analytics?
- What are the critical success factors for Big Data analytics?
- What are the big challenges that one should be mindful of when considering implementation of Big Data analytics?
- What are the common business problems addressed by Big Data analytics?
- In the era of Big Data, are we about to witness the end of data warehousing? Why?
- What is “Data for Good?” Select, study, and report on two of the Data for Good initiatives. In your report, make sure to comment on the origins, motivations, and outcomes of the initiatives.
- What is stream analytics? How does it differ from regular analytics?
- What are the most fruitful industries for stream analytics? What is common to those industries?
- Compared to regular analytics, do you think stream analytics will have more (or less) use cases in the era of Big Data analytics? Why?
- What is the relationship between statistics and business analytics (consider the placement of statistics in a business analytics taxonomy)?
- What are the main differences between descriptive and inferential statistics?
- What is a box-and-whiskers plot? What types of statistical information does it represent?
- What are the two most commonly used shape characteristics to describe a data distribution?
- List and briefly define the central tendency measures of descriptive statistics.
- What are the commonalities and differences between regression and correlation?
- List and describe the main steps to follow in developing a linear regression model.

29. What are the most commonly pronounced assumptions for linear regression? What is crucial to the regression models against these assumptions?
30. What are the commonalities and differences between linear regression and logistic regression?
31. What is time series? What are the main forecasting techniques for time series data?

Exercises

Hands-on Exercises

1. Download the “Voting Behavior” data and the brief data description from the book’s Web site. This is a data set manually compiled from counties all around the United States. The data is partially processed, that is, some derived variables are created. Your task is to thoroughly preprocess the data by identifying the error and anomalies and proposing remedies and solutions. At the end, you should have an analytics-ready version of this data. Once the preprocessing is completed, pull this data into Tableau (or into some other data visualization software tool) to extract useful visual information from it. To do so, conceptualize relevant questions and hypotheses (come up with at least three of them) and create proper visualizations that address those questions of “tests” of those hypotheses.
2. Download Visualization_MFG_Sample data set (available as an Excel file on this book’s Web site). Using Excel or any other software tools made available to you in your institution, answer the following questions:
 - a. What is the relationship between gross box office revenue and other movie-related parameters given in the data set?
 - b. How does this relationship vary across different years? Prepare a professional-looking written report that is enhanced with screenshots of your graphic findings.
3. Conduct a search on the Internet using your favorite Web browser to look for articles that deals with the nature of data, management of data, and/or governance of data as it relates to BI and analytics, and critically analyze the content of the article. Makes sure to include scholarly articles in your search by using your school’s library resources and scholar.google.com. Write a report about your findings.
4. Go to UCI data repository (archive.ics.uci.edu/ml/datasets.html), and identify a large data set that contains both numeric and nominal values. Using Microsoft Excel, or any other statistical software:
 - a. Calculate and interpret central tendency measures for each and every variable.
 - b. Calculate and interpret the dispersion/spread measures for each and every variable.
5. Go to UCI data repository (archive.ics.uci.edu/ml/datasets.html), and identify two data sets, one for estimation/regression and one for classification. Using Microsoft Excel, or any other statistical software:
 - a. Develop and interpret a linear regression model.
 - b. Develop and interpret a logistic regression model.
6. Go to kdnugget.com, and become familiar with the range of analytics resources available on this portal. Then, identify an article, a white paper, or an interview script that deals with the nature of data, management of data, and/or governance of data as it relates to BI and business analytics, and critically analyze the content of the article.
7. Go to teradata.com. Find at least three customer case studies on Big Data, and write a report where you discuss the commonalities and differences of these cases.
8. Go to IBM.com. Find at least three customer case studies on Big Data, and write a report where you discuss the commonalities and differences of these cases.
9. Go to SAS.com. Find at least three customer case studies on Big Data, and write a report where you discuss the commonalities and differences of these cases.
10. Go to claudera.com. Find at least three customer case studies on Hadoop implementation, and write a report where you discuss the commonalities and differences of these cases.
11. Go to mapr.com. Find at least three customer case studies on Hadoop implementation, and write a report where you discuss the commonalities and differences of these cases.
12. Go to hortonworks.com. Find at least three customer case studies on Hadoop implementation, and write a report in which you discuss the commonalities and differences of these cases.
13. Go to marklogic.com. Find at least three customer case studies on Hadoop implementation, and write a report where you discuss the commonalities and differences of these cases.
14. Go to youtube.com. Search for videos on Big Data computing. Watch at least two. Summarize your findings.
15. Go to google.com/scholar, and search for articles on stream analytics. Find at least three related articles. Read and summarize your findings.
16. Enter google.com/scholar, and search for articles on data stream mining. Find at least three related articles. Read and summarize your findings.
17. Enter google.com/scholar, and search for articles that talk about Big Data versus data warehousing. Find at least five articles. Read and summarize your findings.

Team Assignment and Role-Playing Project

1. Analytics starts with data. Identifying, accessing, obtaining, and processing of relevant data are the most essential tasks in any analytics study. As a team, you are

tasked to find a large enough real-world data (either from your own organization, which is the most preferred, or from the Internet that can start with a simple search, or from the data links posted on KDnuggets.com), one that has tens of thousands of rows and more than 20 variables to go through and document

a thorough data preprocessing project. In your processing of the data, identify anomalies and discrepancies using descriptive statistics methods and measures, and make the data analytics ready. List and justify your preprocessing steps and decisions in a comprehensive report.

References

- Adshead, A. (2014). Data set to grow 10-fold by 2020 as Internet of Things takes off. <http://www.computerweekly.com/news/2240217788/Data-set-to-grow-10-fold-by-2020-as-internet-of-things-takes-off> (accessed September 2016).
- Annas, G. J. (2003). HIPAA regulations—A new era of medical-record privacy? *New England Journal of Medicine* 348(15), 1486–1490.
- Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified data processing on large clusters. research.google.com/archive/mapreduce.html (accessed August 2021).
- Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4), 498–506.
- Delen, D. (2011). Predicting student attrition with data mining methods. *Journal of College Student Retention*, 13(1), 17–35.
- Delen, D. (2020). *Predictive Analytics: Data Mining, Machine Learning and Data Science for Practitioners*, 2nd Edition. Pearson Business Analytics Series. Upper Saddle River, NJ.
- Delen, D., Cogdell, D., & Kasap, N. (2012). A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting*, 28, 543–552.
- Delen, D., Topuz, K., & Eryarsoy, E. (2020). Development of a Bayesian Belief Network-based DSS for predicting and understanding freshmen student attrition. *European Journal of Operational Research*, 281(3), 575–587.
- Freeman, K. M., & Brewer, R. M. (2016). The politics of American college football. *Journal of Applied Business and Economics*, 18(2), 97–101.
- Hernández, M. A., & Stolfo, S. J. (1998, January). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, 2(1), 9–37.
- Kelly, L. (2012). Big data: Hadoop, business analytics, and beyond. wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_and_Beyond (accessed August 2021).
- Kock, N. F., McQueen, R. J., & Corner, J. L. (1997). The nature of data, information and knowledge exchanges in business processes: Implications for process improvement and organizational learning. *The Learning Organization*, 4(2), 70–80.
- Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised learning. *International Journal of Computer Science*, 1(2), 111–117.
- Maleki, N., Rahmani, A. M., & Conti, M. (2019). MapReduce: An infrastructure review and research insights. *The Journal of Supercomputing*, 75(10), 6934–7002.
- Quinn, C. (2016). Data-driven marketing at SiriusXM. Teradata Articles & News. at <http://bigdata.teradata.com/US/Articles-News/Data-Driven-Marketing-At-SiriusXM/> (accessed August 2016); Teradata customer success story. SiriusXM attracts and engages a new generation of radio consumers. <http://assets.teradata.com/resourceCenter/downloads/CaseStudies/EB8597.pdf?processed=1>.
- Russom, P. (2013). Busting 10 Myths about Hadoop: The Big Data Explosion. *TDWI's Best of Business Intelligence*, 10, 45–46.
- Silipo, R. & Rudnitchkaia, L. (2022). Will they Blend? (3rd Edition) at <https://www.knime.com/knimepress/will-they-blend> (accessed June 2022).
- Thammasiri, D., Delen, D., Meesad, P., & Kasap N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, 41(2), 321–330.
- Zikopoulos, P., DeRoos, D., Parasuraman, K., Deutsch, T., Corrigan, D., & Giles, J. (2013). *Harness the Power of Big Data*. New York: McGraw-Hill.

Descriptive Analytics II: Business Intelligence Data Warehousing, and Visualization

LEARNING OBJECTIVES

- Understand the basic definitions and concepts of data warehousing
- Understand data warehousing architectures
- Describe the processes used in developing and managing data warehouses
- Explain data warehousing operations
- Explain the role of data warehouses in decision support
- Explain data integration and the extraction, transformation, and load (ETL) processes
- Understand the importance of data warehouse administration, security issues, and future trends
- Define business reporting, and understand its historical evolution
- Understand the importance of data/information visualization
- Learn different types of visualization techniques
- Appreciate the value that visual analytics brings to business analytics
- Know the capabilities and limitations of dashboards

The concept of data warehousing has been around since the late 1980s. This chapter provides the foundation for an important type of database, called a *data warehouse*, which is primarily used for decision support and provides the informational foundation for improved analytical capabilities. We discuss data warehousing concepts and, relatedly, business intelligence and data visualization as critical components of the descriptive analytics continuum. Following the data warehousing sections are sections on business reporting and visualization. A report is a communication artifact prepared with the specific intention of converting data into information and knowledge and relaying that information in an easily understandable/digestible format. Nowadays,

these reports are more visually oriented, often using colors and graphical icons that collectively look like a dashboard to enhance the information content. Therefore, the latter part of the chapter is dedicated to subsections that present the design, implementation, and best practices for information visualization and storytelling.

- 4.1** Opening Vignette: Targeting Tax Fraud with Data Warehousing and Business Analytics 186
- 4.2** Business Intelligence and Data Warehousing 188
- 4.3** Data Warehousing Process 195
- 4.4** Data Warehousing Architectures 197
- 4.5** Data Management and Warehouse Development 203
- 4.6** Data Warehouse Administration, Security Issues, and Future Trends 217
- 4.7** Business Reporting 223
- 4.8** Data Visualization 225
- 4.9** Different Types of Charts and Graphs 228
- 4.10** The Emergence of Visual Analytics 232
- 4.11** Information Dashboards 238

4.1 OPENING VIGNETTE: Targeting Tax Fraud with Data Warehousing and Business Analytics

Governments have to work hard to keep tax fraud from taking a significant bite from their revenues. In 2013, the Internal Revenue Service (IRS) successfully foiled attempts, which were based on stolen identities, to cheat the federal government out of \$24.2 billion in tax refunds. However, that same year the IRS paid out \$5.8 billion on claims it only later identified as fraud. Identifying and preventing tax fraud has proven to be a challenging task.

States also lose money when fraudsters use stolen Social Security numbers, W-2 forms, and other personal information to file false refund claims. This kind of crime has increased in recent years at an alarming rate. “Virtually all Americans have heard of identity theft, but very few are aware of this explosive increase in tax return fraud,” says Maryland Comptroller Peter Franchot. “This is an alarming problem, affecting every state. It is, literally, systematic burglary of the taxpayer’s money.”

In Maryland, the people charged with rooting out false refund claims are members of the Questionable Return Detection Team (QRDT). Like their counterparts in many other states, these experts use software to identify suspicious returns. They then investigate the returns to pinpoint which ones are fraudulent.

CHALLENGE

In the past, Maryland used metrics that examined tax returns one by one. If a return displayed specific traits—for instance, a certain ratio of wages earned to wages withheld—the software suspended that return for further investigation. Members of the QRDT then researched each suspended return—for example, by comparing its wage and withholding information with figures from a W-2 form submitted by an employer. The process was labor-intensive and inefficient. Of the approximately 2.8 million tax returns Maryland received each year, the QRDT suspended about 110,000. But most of those turned out to be legitimate returns. “Only about 10% were found to be fraudulent,” says Andy Schaufele, director of the Bureau of Revenue Estimates for the Maryland Comptroller.

In a typical year, that process saved Maryland from mailing out \$5 million to \$10 million in fraudulent refunds. Although that’s a success, it’s only a modest one, considering the resources tied up in the process and the inconvenience to honest taxpayers whose returns

were flagged for investigation. “The thought that we were holding up 90,000 to 100,000 tax refunds was tough to stomach,” Schaufele says. “We wanted to get those refunds to the taxpayers faster, since many people count on that money as part of their income.”

SOLUTION

Maryland needed a more effective process. It also needed new strategies for staying ahead of fraudsters. “All the states, as well as the IRS, were using the same metrics we were using,” Schaufele says. “I don’t think it was hard for criminals to figure out what our defenses were.” Fortunately, Maryland had recently gained a powerful new weapon against tax fraud. In 2010, the Maryland Comptroller of the Treasury worked with Teradata of Dayton, Ohio, to implement a data warehouse designed to support a variety of compliance initiatives.

As officials discussed which initiatives to launch, one idea rose to the top. “We determined that we should prioritize our efforts to go after refund fraud,” says Sharonne Bonardi, Maryland’s deputy comptroller. So the state started working with Teradata and with ASR Analytics of Potomac, Maryland, to develop a better process for isolating fraudulent tax returns).

“The first step was to analyze our data and learn what we knew about fraud,” Schaufele says. Among other discoveries, the analysis showed that when multiple returns were suspended—even for completely different reasons—they often had traits in common. The state built a database of traits that characterize fraudulent returns and traits that characterize honest ones. “We worked with ASR to put that information together and develop linear regressions,” Schaufele says. “Instead of looking at one-off metrics, we began to bring many of those metrics together.” The result was a far more nuanced portrait of the typical fraudulent return.

Instead of flagging returns one by one, the new system identifies groups of returns that look suspicious for similar reasons. That strategy speeds up investigations. The analytics system also assigns a score to each return, based on how likely it is to be fraudulent. It then produces a prioritized list to direct the QRDT’s workflow. “We’re first working on the returns that are more likely not to be fraudulent, so we can get them out of the queue,” Schaufele says. The more suspicious-looking returns go back for further review.

RESULTS

“With these analytics models, we’re able to reduce false positives, so that we don’t overburden the taxpayers who have accurately reported their information to the state,” Bonardi says. Once investigators remove their returns from the queue, those taxpayers can get their refunds.

Thanks to the new technology, QRDT expects to suspend only 40,000 to 50,000 tax returns, compared with 110,000 in past years. “Of those we’ve worked so far, we’re getting an accuracy rate of about 65%,” says Schaufele. That’s a big improvement over the historical 10% success rate. “Once the returns are identified which may be fraudulent, the team of expert examiners can then carefully review them, one at a time, to eliminate returns that are found to be legitimate,” Maryland Comptroller Franchot says. “The entire operation is getting better and stronger all the time.”

As of late March, advanced analytics had helped the QRDT recover approximately \$10 million in the current filing season. Schaufele says, “Under the old system, that number would have been about \$3 million at this point.” Not only does the new technology help the QRDT work faster and more efficiently, but it also helps the team handle a heavier and more complex workload. As tax criminals have ramped up their efforts, the QRDT has had to deploy new strategies against them. For example, in 2015 the team received some 10,000 notifications from taxpayers whose identifications had been stolen. “So we have a new workflow: We look up their Social Security numbers

and try to find any incidences of fraud that might have been perpetrated with them,” says Schaufele. “That’s a new level of effort that this group is now completing without additional resources.”

To stay ahead of more sophisticated tax schemes, investigators now not only examine current W-2 forms, but also compare them with the same taxpayers’ forms from prior years, looking for inconsistencies. “The investigations are becoming more complex and taking longer,” Schaufele says. “If we hadn’t winnowed down the universe for review, we would have had some real problems pursuing them.”

► QUESTIONS FOR THE OPENING VIGNETTE

1. Why is it important for IRS and for U.S. state governments to use data warehousing and business intelligence (BI) tools in managing state revenues?
2. What were the challenges the state of Maryland was facing with regard to tax fraud?
3. What was the solution they adopted? Do you agree with their approach? Why?
4. What were the results that they obtained? Did the investment in BI and data warehousing pay off?
5. What other problems and challenges do you think federal and state governments are having that can benefit from BI and data warehousing?

WHAT WE CAN LEARN FROM THIS VIGNETTE

The opening vignette illustrates the value of BI, decision support systems, and data warehousing in management of government revenues. With their data warehouse implementation, the State of Maryland was able to leverage its data assets to make more accurate and timely decisions on identifying fraudulent tax returns. Consolidating and processing a wide variety of data sources within a unified data warehouse enabled Maryland to automate the identification of tax fraud signals/rules/traits from historic facts as opposed to merely relying on traditional ways where they have been implementing intuition-based filtering rules. By using data warehousing and BI, Maryland managed to significantly reduce the false positive rate (and by doing so ease the pain on the part of taxpayers) and improved the prediction accuracy rate from 10% to 65% (more than a sixfold improvement in accurate identification of fraudulent tax returns). The key lesson here is that a properly designed and implemented data warehouse combined with BI tools and techniques can and will result in significant improvement (both on accuracy and on timeliness) resulting in benefits (both financial and nonfinancial) for any organization, including state governments like Maryland.

Sources: Teradata case study (2022). Targeting Tax Fraud with Advanced Analytics, <https://www.teradata.com/Resources/Case-Studies/Targeting-Tax-Fraud-with-Advanced-Analytics> (accessed June 2022); Temple-West, P. (2013, November 7). Tax refund ID theft is growing “epidemic”: U.S. IRS watchdog. Reuters. <http://www.reuters.com/article/us-usa-tax-refund-idUSBRE9A61HB20131107> (accessed July 2016).

4.2 BUSINESS INTELLIGENCE AND DATA WAREHOUSING

Business intelligence (BI), as a term to describe evidence/fact-based managerial decision making, has been around for more than 20 years. With the emergence of business analytics as a new buzzword to describe pretty much the same managerial phenomenon, the popularity of BI as a term has gone down. As opposed to being an all-encompassing term, nowadays BI is used to describe the early stages of business analytics (i.e., descriptive analytics).

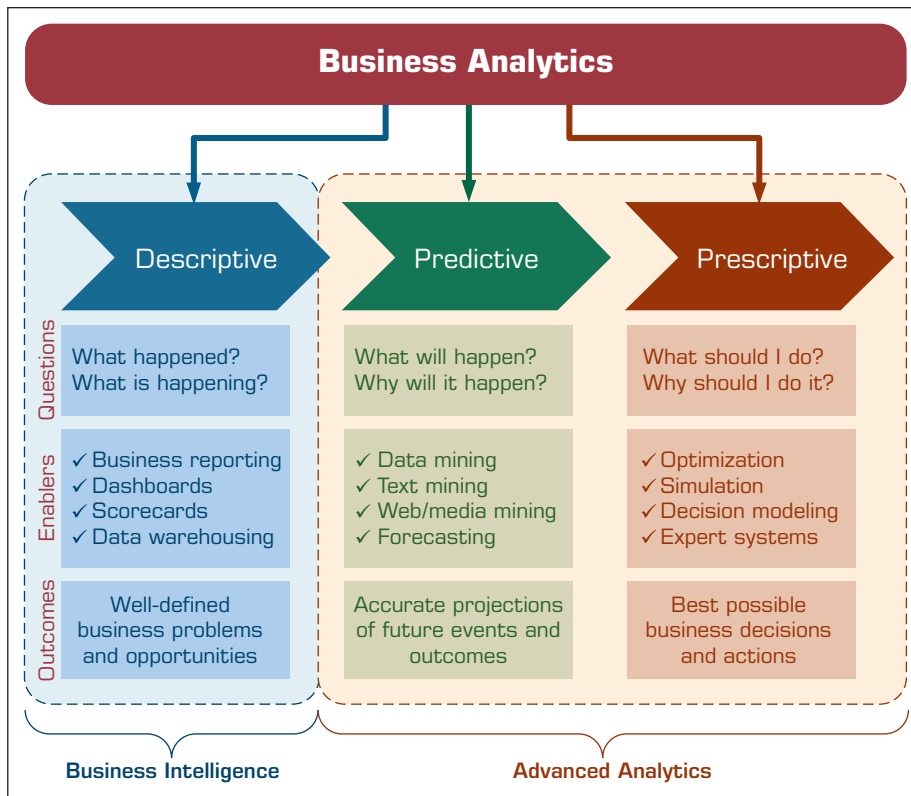


FIGURE 4.1 Relationship between Business Analytics and BI, and BI and Data Warehousing.

Figure 4.1 (a simplified version of which was shown and described in Chapter 1 to describe business analytics taxonomy) illustrates the relationship between BI and business analytics from a conceptual perspective. As shown therein, BI is the descriptive analytics portion of the business analytics continuum, the maturity of which leads to advanced analytics—a combination of predictive and prescriptive analytics.

Descriptive analytics (i.e., BI) is the entry level in the business analytics taxonomy. It is often called business reporting because of the fact that most of the analytics activities at this level deal with creating reports to summarize business activities to answer questions such as “What happened?” and “What is happening?” The spectrum of these reports includes static snapshots of business transactions delivered to knowledge workers (i.e., decision makers) on a fixed schedule (e.g., daily, weekly, quarterly); ad hoc reporting where the decision maker is given the capability of creating his or her own specific report (using an intuitive drag-and-drop graphical user interface) to address a specific or unique decision situation; and dynamic views of key business performance indicators (often captured and presented within a business performance management system) delivered to managers and executives in an easily digestible form (e.g., dashboard-looking graphical interfaces) on a continuous manner.

Generally speaking, and as depicted in Figure 4.1, BI systems rely on a data warehouse as the information source for creating insight and supporting managerial decisions. A multitude of organizational and external data is captured, transformed, and stored in a data warehouse to support timely and accurate decisions through enriched business insight. This chapter aims to cover the concepts, methods, and tools related to data warehousing and business performance management.

What Is a Data Warehouse?

In simple terms, a **data warehouse (DW)** is a pool of data produced to support decision making; it is also a repository of current and historical data of potential interest to managers throughout the organization. Data are usually structured to be available in a form ready for analytical processing activities (i.e., online analytical processing [OLAP], data mining, querying, reporting, and other decision support applications). A data warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management's decision-making process.

A Historical Perspective to Data Warehousing

Even though *data warehousing* is a relatively new term in information technology (IT), its roots can be traced back in time, even before computers were widely used. In the early 1900s, people were using data (though mostly via manual methods) to formulate trends to help business users make informed decisions, which is the most prevailing purpose of data warehousing.

The motivations that led to the development of data warehousing technologies go back to the 1970s, when the computing world was dominated by mainframes. Real business data-processing applications, the ones run on the corporate mainframes, had complicated file structures using early-generation databases (not the table-oriented relational databases most applications use today) in which they stored data. Although these applications did a decent job of performing routine transactional data-processing functions, the data created as a result of these functions (such as information about customers, the products they ordered, and how much money they spent) was locked away in the depths of the files and databases. When aggregated information such as sales trends by region and by product type was needed, one had to formally request it from the data-processing department, where it was put on a waiting list with a couple of hundred other report requests (Hammergren & Simon, 2009). Even though the need for information and the data used to generate it existed, the database technology was not there to satisfy it. Figure 4.2 shows a timeline where some of the significant events that led to the development of data warehousing are shown.

Later in the last century, commercial hardware and software companies began to emerge with solutions to this problem. Between 1976 and 1979, the concept for a new company, Teradata, grew out of research at the California Institute of Technology

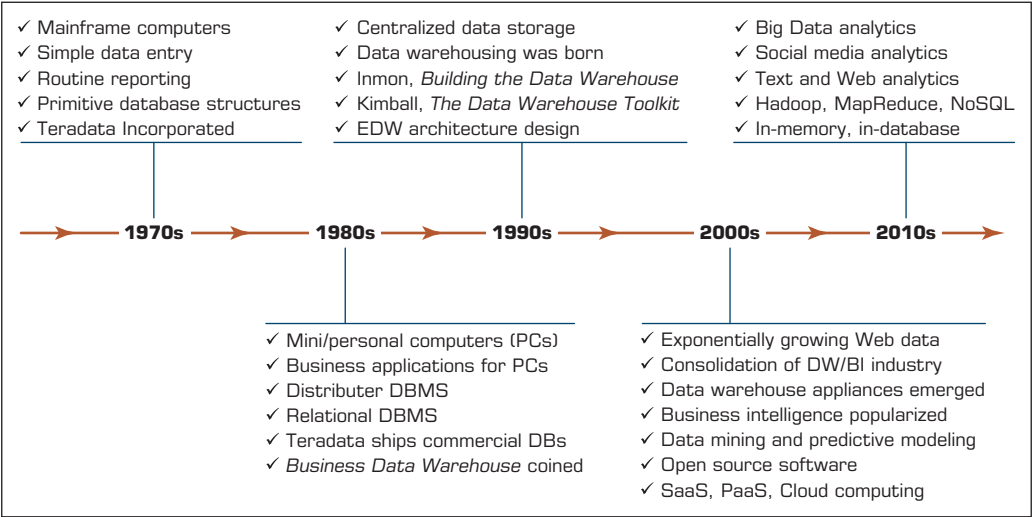


FIGURE 4.2 A List of Events That Led to Data Warehousing Development.

(Caltech), driven from discussions with Citibank's advanced technology group. Founders worked to design a database management system for parallel processing with multiple microprocessors, targeted specifically for decision support. Teradata was incorporated on July 13, 1979, and started in a garage in Brentwood, California. The name *Teradata* was chosen to symbolize the ability to manage terabytes (trillions of bytes) of data.

The 1980s were the decade of personal computers and minicomputers. Before anyone knew it, real computer applications were no longer only on mainframes; they were all over the place—everywhere you looked in an organization. That led to a portentous problem called *islands of data*. The solution to this problem led to a new type of software, called a *distributed database management system*, which would magically pull the requested data from databases across the organization, bring all the data back to the same place, and then consolidate it, sort it, and do whatever else was necessary to answer the user's question. Although the concept was a good one and early results from research were promising, the results were plain and simple: They just didn't work efficiently in the real world, and the islands-of-data problem still existed.

Meanwhile, Teradata began shipping commercial products to solve this problem. Wells Fargo Bank received the first Teradata test system in 1983, a parallel RDBMS (relational database management system) for decision support—the world's first. By 1984, Teradata released a production version of their product, and in 1986, *Fortune* magazine named Teradata "Product of the Year." Teradata, still in existence today, built the first data warehousing appliance—a combination of hardware and software to solve the data warehousing needs of many. Other companies began to formulate their strategies, as well.

During the 1980s, several other events occurred, collectively making it the decade of data warehousing innovation. For instance, Ralph Kimball founded Red Brick Systems in 1986. Red Brick began to emerge as a visionary software company by discussing how to improve data access; in 1988, Barry Devlin and Paul Murphy of IBM Ireland introduced the term *business data warehouse* as a key component of business information systems.

In the 1990s, a new approach to solving the islands-of-data problem surfaced. If the 1980s approach of reaching out and accessing data directly from the files and databases didn't work, the 1990s philosophy involved going back to the 1970s, in which data from those places was copied to another location—only doing it right this time; hence, data warehousing was born. In 1993, Bill Inmon wrote the seminal book *Building the Data Warehouse*. Many people recognize Inmon as the father of data warehousing. Additional publications emerged, including the 1996 book by Ralph Kimball, *The Data Warehouse Toolkit*, which discussed general-purpose dimensional design techniques to improve the data architecture for query-centered decision support systems.

In the 2000s, in the world of data warehousing, both popularity and the amount of data continued to grow. The vendor community and options began to consolidate. In 2006, Microsoft acquired ProClarity, jumping into the data warehousing market. In 2007, Oracle purchased Hyperion, SAP acquired Business Objects, and IBM merged with Cognos. The data warehousing leaders of the 1990s have been swallowed by some of the largest providers of information system solutions in the world. During this time, other innovations emerged, including data warehouse appliances from vendors such as Netezza (acquired by IBM), Greenplum (acquired by EMC), DATAlegro (acquired by Microsoft), and performance management appliances that enabled real-time performance monitoring. These innovative solutions provided cost savings because they were plug-compatible to legacy data warehouse solutions.

Since 2010, the big buzz has been *Big Data*. Many believe that Big Data is going to make an impact on data warehousing as we know it. Either they will find a way to coexist (which seems to be the most likely case, at least for several years) or Big Data (and the technologies that come with it) will make traditional data warehousing obsolete. The technologies that came with Big Data include Hadoop, MapReduce, NoSQL, and Hive.

Maybe we will see a new term coined in the world of data that combines the needs and capabilities of traditional data warehousing and the Big Data phenomenon.

Analytics in Action 4.1 is all about enhancing customer experience (CX) in financial sector through use of all kinds of data obtained from all channels along with business analytics.

ANALYTICS IN ACTION 4.1 Data-Driven Customer Experience in Financial Services

Introduction

The rapid migration to digital banking technologies driven by the pandemic will continue into the recovery. One-third of retail banking customers plan to increase their use of online and mobile banking services post COVID-19. Financial Services enterprises that make the right investments now could build an enduring advantage in serving customers.

Advanced customer analytics informed by new digital banking sources allows financial services firms to develop insights on customer cohorts and behaviors, and to more precisely target and tailor products and value propositions. Recent advances in analytics, such as non-linear machine-learning algorithms, combined with new and more granular data and are dramatically improving model efficacy and resulting service improvements, revenue increases, and lowering the cost to serve. Activating insights through triggers derived from real-time behaviors further contribute to these business outcomes.

Challenges

Optimizing the customer experience can be challenging. For one thing, most organizations have little visibility into the end-to-end customer journey that spans multiple customer touchpoints across multiple departments. Furthermore, enterprises suffer from an inability to sense and react to a multitude of customer pain points and opportunities. Finally, organizations struggle to meaningfully personalize customer interactions in a timely manner, particularly in real-time, at the moments of truth.

Solutions

Vantage Customer Experience (Vantage CX) uniquely overcomes these challenges, enabling financial services providers to execute a sustainable customer experience initiative, including the following:

- Enables the previously elusive 360° view of the customer by providing financial services industry data models that include 100% of online behaviors and transactions. Granular, individual level behavioral data is captured in real time to provide an enhance view of every customer interaction across all digital properties. Online behavioral data is unified with customer transactions and behaviors from branch operations, call center, all transactions covering deposits, trades, transfers, applications, and more. This unified view of the customer across multiple departments and systems provides visibility into the full customer experience.
- Turns data into customer insights by putting the marketer and CX professional in control with powerful, easy-to use machine learning analytics along with other advanced analytics that provide richer, faster insights into customer behavior. No coding is required to leverage intuitive advanced analytics in Vantage, including Path for understanding the customer journey, Text for sentiment analysis, Cluster for hyper-segmentation, and Machine Learning to optimize the next best offer.
- Turns customer insights into action by delivering personalized customer interactions in real time, seamlessly across channels. Vantage CX enables real-time dynamic personalization by determining the optimal message to deliver based on both what the customer is doing right now combined with historical, integrated information on that customer's experiences and behaviors. Machine learning optimizes the goal selected by the business user,



such as maximizing revenue or minimizing cost to serve. These highly personalized offers are delivered at scale to reach millions of customers informed by billions of transactions and interactions. Real-time personalization makes it easy to deliver these dynamic messages through a company's Web site, SMS, POS, and other digital screens.

Highly detailed and structured Celebris data captured from all digital channels is fed into Vantage CX to enable highly effective customer journey analytics and real-time personalization at scale. Previously undiscovered issues with customer facing channels and pages can be quickly identified and resolved. Millisecond omnichannel data feeds combined with historical integrated data and machine learning insights from Vantage CX enable the creation of definitive and highly accurate customer profiles and the ability to orchestrate personalized customer experiences in-the-moment and across all channels.

Results

Working with Teradata and Celebris, the bank:

- Utilized identity resolution to connect web, mobile app, and offline customer data.
- Derived new digital variables such as page scores, recurring visits, duration, and others.
- Uncovered optimization opportunities across multiple touch points using Path analytics.
- Triggered hyper-personalized communications to recover abandoned applications.

In addition, the data and resultant insight enabled them to not only develop new products specifically designed for their digitally engaged customers based on enhanced customer segmentation, but also maximize acquisition rates and reduce service costs for those new offerings. Customer acquisition is now 3–5 times more effective overall than before for the same level of investment. Credit card conversion significantly improved for both the online and mobile channels: These successes fundamentally shifted the bank's digital banking department from product-centric to customer centric and accelerated the speed of the transformation program across the entire global organization.

Source: Teradata case study. Data-Driven Customer Experience in Financial Services. <https://www.teradata.com/Resources/Case-Studies/Data-Driven-Customer-Experience-in-Financial-Services> (accessed June 2022).

Characteristics of Data Warehousing

A common way to introduce data warehousing is to refer to its fundamental characteristics (see Inmon, 2005):

- **Subject oriented.** Data are organized by detailed subject, such as sales, products, or customers, containing only information relevant for decision support. Subject orientation enables users to determine not only how their business is performing, but also why. A data warehouse differs from an operational database in that most operational databases have a product orientation and are tuned to handle transactions that update the database. Subject orientation provides a more comprehensive view of the organization.

- **Integrated.** Integration is closely related to subject orientation. Data warehouses must place data from different sources into a consistent format. To do so, they must deal with naming conflicts and discrepancies among units of measure. A data warehouse is presumed to be totally integrated.
- **Time variant (time series).** A warehouse maintains historical data. The data do not necessarily provide current status (except in real-time systems). They detect trends, deviations, and long-term relationships for forecasting and comparisons, leading to decision making. Every data warehouse has a temporal quality. Time is the one important dimension that all data warehouses must support. Data for analysis from multiple sources contain multiple time points (e.g., daily, weekly, monthly views).
- **Nonvolatile.** After data are entered into a data warehouse, users cannot change or update the data. Obsolete data are discarded, and changes are recorded as new data.

These characteristics enable data warehouses to be tuned almost exclusively for data access. Some additional characteristics may include the following:

- **Web based.** Data warehouses are typically designed to provide an efficient computing environment for Web-based applications.
- **Relational/multidimensional.** A data warehouse uses either a relational structure or a multidimensional structure. A survey on multidimensional structures can be found in Romero and Abelló (2009).
- **Client/server.** A data warehouse uses the client/server architecture to provide easy access for end users.
- **Real time.** Newer data warehouses provide real-time, or active, data-access and analysis capabilities (see Basu, 2003; Bonde & Kuckuk, 2004).
- **Include metadata.** A data warehouse contains metadata (data about data) about how the data are organized and how to effectively use them.

Whereas a data warehouse is a repository of data, data warehousing is literally the entire process (see Watson, 2002). Data warehousing is a discipline that results in applications that provide decision support capability, allows ready access to business information, and creates business insight. The three main types of data warehouses are data marts (DMs), operational data stores (ODS), and enterprise data warehouses (EDW). In addition to discussing these three types of warehouses next, we also discuss metadata.

Data Marts

Whereas a data warehouse combines databases across an entire enterprise, a **data mart (DM)** is usually smaller and focuses on a particular subject or department. A DM is a subset of a data warehouse, typically consisting of a single subject area (e.g., marketing, operations). A DM can be either dependent or independent. A **dependent data mart** is a subset that is created directly from the data warehouse. It has the advantages of using a consistent data model and providing quality data. Dependent DMs support the concept of a single enterprise-wide data model, but the data warehouse must be constructed first. A dependent DM ensures that the end user is viewing the same version of the data that is accessed by all other data warehouse users. The high cost of data warehouses limits their use to large companies. As an alternative, many firms use a lower-cost, scaled-down version of a data warehouse referred to as an *independent DM*. An **independent data mart** is a small warehouse designed for a strategic business unit or a department, but its source is not an EDW.

Operational Data Stores

An **operational data store (ODS)** provides a fairly recent form of customer information file. This type of database is often used as an interim staging area for a data warehouse. Unlike the static contents of a data warehouse, the contents of an ODS are updated throughout the course of business operations. An ODS is used for short-term decisions

involving mission-critical applications rather than for the medium- and long-term decisions associated with an EDW. An ODS is similar to short-term memory in that it stores only very recent information. In comparison, a data warehouse is like long-term memory because it stores permanent information. An ODS consolidates data from multiple source systems and provides a near-real-time, integrated view of volatile, current data. The exchange, transfer, and load (ETL) processes (discussed later in this chapter) for an ODS are identical to those for a data warehouse. Finally, **oper marts** (see Imhoff, 2001) are created when operational data needs to be analyzed multidimensionally. The data for an oper mart come from an ODS.

Enterprise Data Warehouses (EDW)

An **enterprise data warehouse (EDW)** is a large-scale data warehouse that is used across the enterprise for decision support. The large-scale nature of an EDW provides integration of data from many sources into a standard format for effective BI and decision support applications. EDWs are used to provide data for many types of decision support systems (DSS), including customer relationship management (CRM), supply chain management (SCM), business performance management (BPM), business activity monitoring, product life cycle management, revenue management, and sometimes even knowledge management systems.

Metadata

Metadata are data about data (e.g., see Sen, 2004; Zhao, 2005). Metadata describe the structure of and some meaning about data, thereby contributing to their effective or ineffective use. Mehra (2005) indicated that few organizations really understand metadata, and fewer understand how to design and implement a metadata strategy. Metadata are generally defined in terms of usage as technical or business metadata. Patterns are another way to view metadata. According to the pattern view, we can differentiate between syntactic metadata (i.e., data describing the syntax of data), structural metadata (i.e., data describing the structure of the data), and semantic metadata (i.e., data describing the meaning of the data in a specific domain).

SECTION 4.2 REVIEW QUESTIONS

1. What is a data warehouse?
2. How does a data warehouse differ from a transactional database?
3. What is an ODS?
4. Differentiate among a DM, an ODS, and an EDW.
5. What is metadata? Explain the importance of metadata.

4.3 DATA WAREHOUSING PROCESS

Organizations, private and public, continuously collect data, information, and knowledge at an increasingly accelerated rate and store them in computerized systems. Maintaining and using these data and information becomes extremely complex, especially as scalability issues arise. In addition, the number of users needing to access the information continues to increase as a result of improved reliability and availability of network access, especially the Internet. Working with multiple databases, either integrated in a data warehouse or not, has become an extremely difficult task requiring considerable expertise, but it can provide immense benefits far exceeding its cost. As an illustrative example, Figure 4.3 shows business benefits of the EDW built by Teradata for a major automobile manufacturer.

Many organizations need to create data warehouses—massive data stores of time series data for decision support. Data are imported from various external and internal

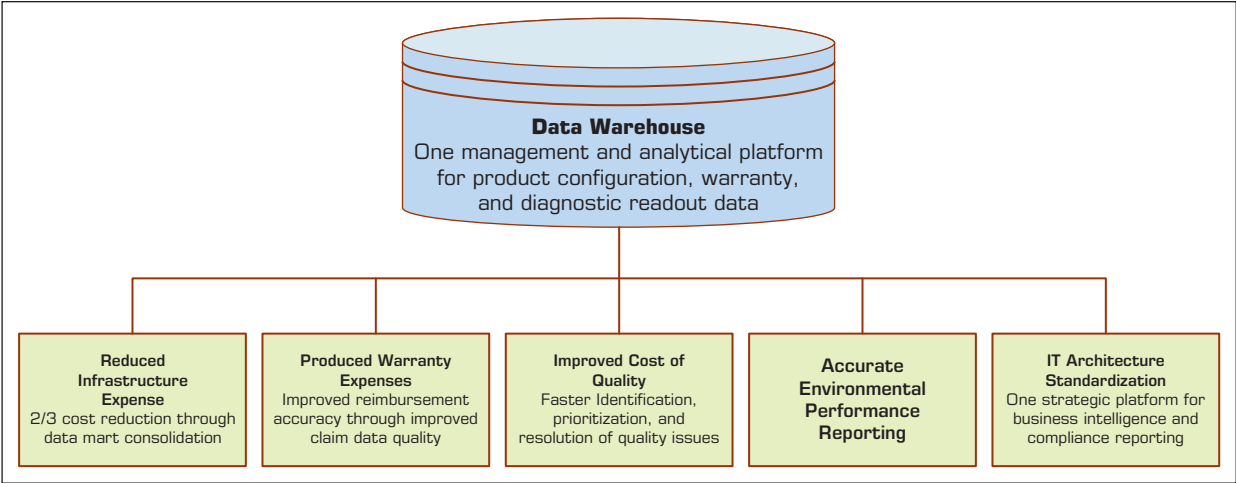


FIGURE 4.3 Data-Driven Decision Making—Business Benefits of the Data Warehouse. *Source: Teradata Corp.*

resources and are cleansed and organized in a manner consistent with the organization’s needs. After the data are populated in the data warehouse, DMs can be loaded for a specific area or department. Alternatively, DMs can be created first, as needed, and then integrated into an EDW. Often, though, DMs are not developed, but data are simply loaded onto PCs or left in their original state for direct manipulation using BI tools.

In Figure 4.4, we show a conceptual depiction of a typical data framework.

The following are the major components of the data warehousing process:

- **Data sources.** Data are sourced from multiple independent operational “legacy” systems and possibly from external data providers (such as the U.S. Census). Data may also come from an OLTP or enterprise resource planning (ERP) system. Web data in the form of Web logs may also feed to a data warehouse.
- **Data extraction and transformation.** Data are extracted and properly transformed using custom-written or commercial software called ETL.

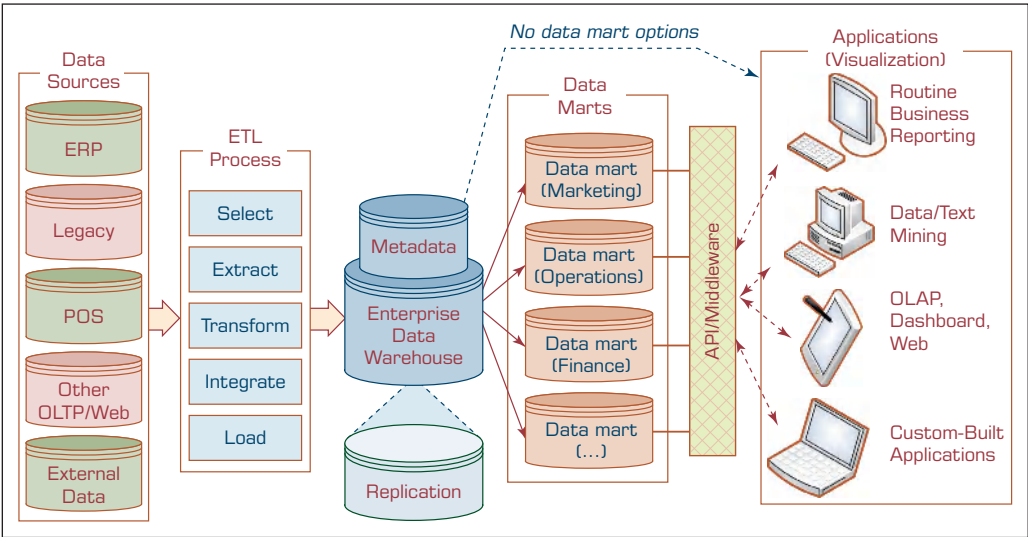


FIGURE 4.4 High-Level Depiction of a Data Warehouse Framework.

- **Data loading.** Data are loaded into a staging area, where they are transformed and cleansed. The data are then ready to load into the data warehouse and/or DMs.
- **Comprehensive database.** Essentially, this is the EDW to support all decision analysis by providing relevant summarized and detailed information originating from many different sources.
- **Metadata.** Metadata are maintained so that they can be assessed by IT personnel and users. Metadata include software programs about data and rules for organizing data summaries that are easy to index and search, especially with Web tools.
- **Middleware tools.** Middleware tools enable access to the data warehouse. Power users such as analysts may write their own SQL queries. Others may employ a managed query environment, such as Business Objects, to access data. There are many front-end applications that business users can use to interact with data stored in the data repositories, including data mining, OLAP, reporting tools, and data visualization tools.

SECTION 4.3 REVIEW QUESTIONS

1. Describe the data warehousing process.
2. Describe the major components of a data warehouse.
3. Identify and discuss the role of middleware tools.

4.4 DATA WAREHOUSING ARCHITECTURES

Several basic information system architectures can be used for data warehousing. Generally speaking, these architectures are commonly called client/server or n-tier architectures, of which two-tier and three-tier architectures are the most common (see Figures 4.5 and 4.6), but sometimes there is simply one tier. These types of multitiered architectures are known to be capable of serving the needs of large-scale, performance-demanding information systems such as data warehouses. Referring to the traditional use of n-tiered architectures for IT systems, we can distinguish among these architectures by dividing the data warehouse into three parts:

1. The data warehouse itself, which contains the data and associated software
2. Data acquisition (back-end) software, which extracts data from legacy systems and external sources, consolidates and summarizes them, and loads them into the data warehouse
3. Client (front-end) software, which allows users to access and analyze data from the warehouse (a DSS/BI/business analytics [BA] engine)

In a three-tier architecture, operational systems contain the data and the software for data acquisition in one tier (i.e., the server), the data warehouse is another tier, and the third tier includes the DSS/BI/BA engine (i.e., the application server) and the client (see Figure 4.5). Data from the warehouse are processed twice and deposited in an

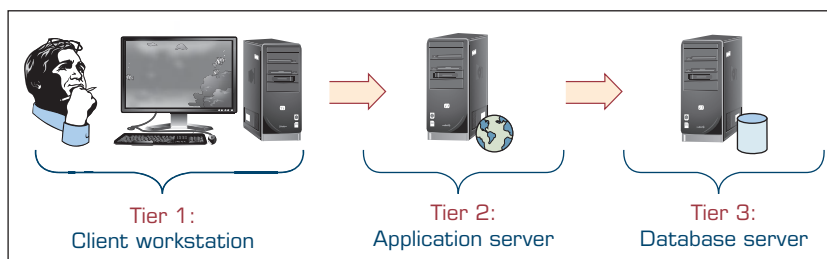


FIGURE 4.5 Architecture of a Three-Tier Data Warehouse.

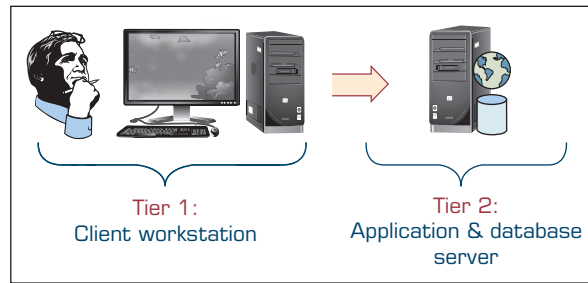


FIGURE 4.6 Architecture of a Two-Tier Data Warehouse.

additional multidimensional database, organized for easy multidimensional analysis and presentation, or replicated in DMs. The advantage of the three-tier architecture is its separation of the functions of the data warehouse, which eliminates resource constraints and makes it possible to easily create DMs.

In a two-tier architecture, the DSS engine physically runs on the same hardware platform as the data warehouse (see Figure 4.6). Therefore, it is more economical than the three-tier structure. The two-tier architecture can have performance problems for large data warehouses that work with data-intensive applications for decision support.

Much of the common wisdom assumes an absolutist approach, maintaining that one solution is better than the other, despite the organization's circumstances and unique needs. To further complicate these architectural decisions, many consultants and software vendors focus on one portion of the architecture, therefore limiting their capacity and motivation to assist an organization through the options based on its needs. But these aspects are being questioned and analyzed.

Data warehousing and the Internet are two key technologies that offer important solutions for managing corporate data. The integration of these two technologies produces Web-based data warehousing. In Figure 4.7, we show the architecture of Web-based data warehousing. The architecture is three-tiered and includes the PC client, Web server, and application server. On the client side, the user needs an Internet connection and a Web browser (preferably Java-enabled) through the familiar graphical user interface (GUI). The Internet/intranet/extranet is the communication medium between client

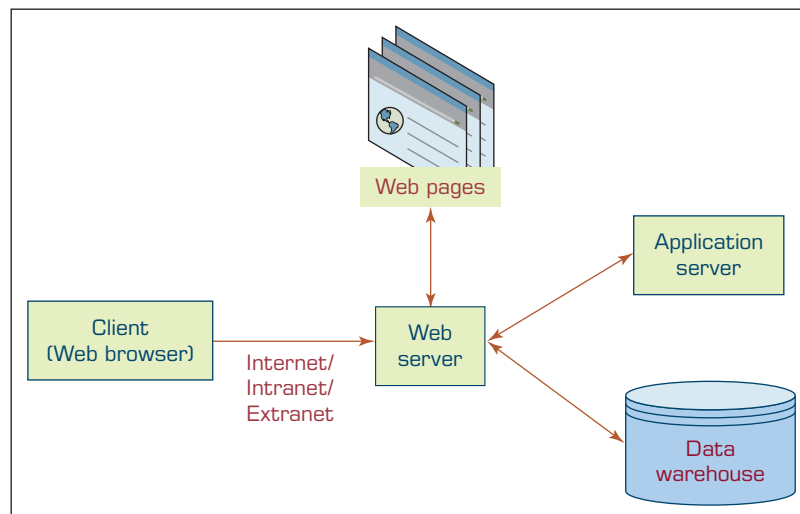


FIGURE 4.7 Architecture of Web-Based Data Warehousing.

and servers. On the server side, a Web server is used to manage the inflow and outflow of information between client and server. It is backed by both a data warehouse and an application server. Web-based data warehousing offers several compelling advantages, including ease of access, platform independence, and lower cost.

Web architectures for data warehousing are similar in structure to other data warehousing architectures, requiring a design choice for housing the Web data warehouse with the transaction server or as a separate server(s). Page-loading speed is an important consideration in designing Web-based applications; therefore, server capacity must be planned carefully.

Several issues must be considered when deciding which architecture to use. Among them are the following:

- ***Which database management system (DBMS) should be used?*** Most data warehouses are built using RDBMS. Oracle (Oracle Corporation, oracle.com), SQL Server (Microsoft Corporation, <http://microsoft.com/sql>), and DB2 (IBM Corporation, <http://www-01.ibm.com/software/data/db2>) are the ones most commonly used. Each of these products supports both client/server and Web-based architectures.
- ***Will parallel processing and/or partitioning be used?*** Parallel processing enables multiple central processing units (CPUs) to process data warehouse query requests simultaneously and provides scalability. Data warehouse designers need to decide whether the database tables will be partitioned (i.e., split into smaller tables) for access efficiency and what the criteria will be. This is an important consideration that is necessitated by the large amounts of data contained in a typical data warehouse. A recent survey on parallel and distributed data warehouses can be found in Furtado (2009). Teradata (teradata.com) has successfully adopted and is often commended on its novel implementation of this approach.
- ***Will data migration tools be used to load the data warehouse?*** Moving data from an existing system into a data warehouse is a tedious and laborious task. Depending on the diversity and the location of the data assets, migration may be a relatively simple procedure or (on the contrary) a months-long project. The results of a thorough assessment of the existing data assets should be used to determine whether to use migration tools, and if so, what capabilities to seek in those commercial tools.
- ***What tools will be used to support data retrieval and analysis?*** Often it is necessary to use specialized tools to periodically locate, access, analyze, extract, transform, and load necessary data into a data warehouse. A decision has to be made on (1) developing the migration tools in-house, (2) purchasing them from a third-party provider, or (3) using the ones provided with the data warehouse system. Overly complex, real-time migrations warrant specialized third-party ETL tools.

Alternative Data Warehousing Architectures

At the highest level, data warehouse architecture design viewpoints can be categorized into enterprise-wide data warehouse (EDW) design and DM design (Golfarelli & Rizzi, 2009). In Figure 4.8a–e, we show some alternatives to the basic architectural design types that are neither pure EDW nor pure DM, but in between or beyond the traditional architectural structures. Notable new ones include hub-and-spoke and federated architectures. The five architectures shown in Figure 4.8a–e, are proposed by Ariyachandra and Watson (2005, 2006a,b). Previously, in an extensive study, Sen and Sinha (2005) identified 15 different data warehousing methodologies. The sources of these methodologies are classified into three broad categories: core-technology vendors, infrastructure vendors, and information-modeling companies.

- ***Independent data marts.*** This is arguably the simplest and the least costly architecture alternative. The DMs are developed to operate independent of each other to serve the needs of individual organizational units. Because of their independence,

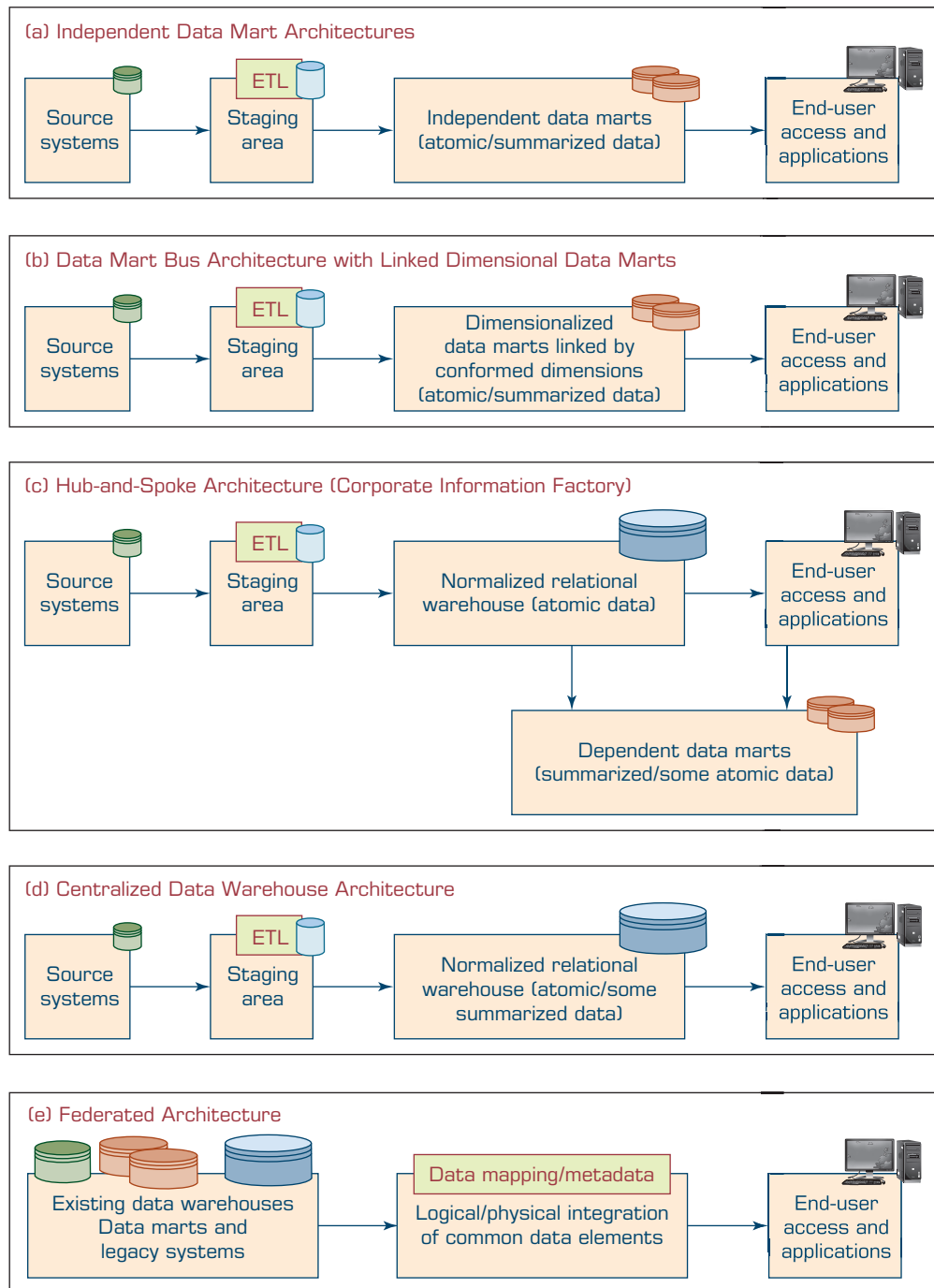


FIGURE 4.8 Alternative Data Warehouse Architectures. Source: Adapted from Ariyachandra, T., & Watson, H. (2006b) Which data warehouse architecture is most successful? *Business Intelligence Journal*, 11(1), 4–6.

they may have inconsistent data definitions and different dimensions and measures, making it difficult to analyze data across the DMs (i.e., it is difficult, if not impossible, to get to the “one version of the truth”).

- **Data mart bus architecture.** This architecture is a viable alternative to the independent DMs where the individual marts are linked to each other via some

kind of middleware. Because the data are linked among the individual marts, there is a better chance of maintaining data consistency across the enterprise (at least at the metadata level). Even though it allows for complex data queries across DMs, the performance of these types of analysis may not be at a satisfactory level.

- **Hub-and-spoke architecture.** This is perhaps the most famous data warehousing architecture today. Here the attention is focused on building a scalable and maintainable infrastructure (often developed in an iterative way, subject area by subject area) that includes a centralized data warehouse and several dependent DMs (each for an organizational unit). This architecture allows for easy customization of user interfaces and reports. On the negative side, this architecture lacks the holistic enterprise view and may lead to data redundancy and data latency.
- **Centralized data warehouse.** The centralized data warehouse architecture is similar to the hub-and-spoke architecture except that there are no dependent DMs; instead, there is a gigantic EDW that serves the needs of all organizational units. This centralized approach provides users with access to all data in the data warehouse instead of limiting them to DMs. In addition, it reduces the amount of data the technical team has to transfer or change, therefore simplifying data management and administration. If designed and implemented properly, this architecture provides a timely and holistic view of the enterprise to whoever, whenever, and wherever they may be within the organization.
- **Federated data warehouse.** The federated approach is a concession to the natural forces that undermine the best plans for developing a perfect system. It uses all possible means to integrate analytical resources from multiple sources to meet changing needs or business conditions. Essentially, the federated approach involves integrating disparate systems. In a federated architecture, existing decision support structures are left in place, and data are accessed from those sources as needed. The federated approach is supported by middleware vendors that propose distributed query and join capabilities. These eXtensible Markup Language (XML)-based tools offer users a global view of distributed data sources, including data warehouses, DMs, Web sites, documents, and operational systems. When users choose query objects from this view and press the submit button, the tool automatically queries the distributed sources, joins the results, and presents them to the user. Because of performance and data quality issues, most experts agree that federated approaches work well to supplement data warehouses, not replace them (see Eckerson, 2005).

Following are the top 10 factors that potentially affect the architecture selection decision:

1. Information interdependence between organizational units
2. Upper management's information needs
3. Urgency of need for a data warehouse
4. Nature of end-user tasks
5. Constraints on resources
6. Strategic view of the data warehouse prior to implementation
7. Compatibility with existing systems
8. Perceived ability of the in-house IT staff
9. Technical issues
10. Social/political factors

These factors are similar to many success factors described in the literature for information system projects and DSS and BI projects. Technical issues, beyond providing technology that is feasibly ready for use, are important, but often not as important as behavioral issues, such as meeting upper management's information needs and user

involvement in the development process (a social/political factor). Each data warehousing architecture has specific applications for which it is most (and least) effective and thus provides maximal benefits to the organization. However, overall, the DM structure seems to be the least effective in practice. See Ariyachandra and Watson (2006a) for some additional details.

Which Architecture Is the Best?

Ever since data warehousing became a critical part of modern enterprises, the question of which data warehouse architecture is the best has been a topic of regular discussion. The two gurus of the data warehousing field, Bill Inmon and Ralph Kimball, are at the heart of this discussion. Inmon advocates the hub-and-spoke architecture (e.g., the Corporate Information Factory), whereas Kimball promotes the DM bus architecture with conformed dimensions. Other architectures are possible, but these two options are fundamentally different approaches, and each has strong advocates. To shed light on this controversial question, Ariyachandra and Watson (2006b) conducted an empirical study. To collect the data, they used a Web-based survey targeted at individuals involved in data warehouse implementations. Their survey included questions about the respondent, the respondent's company, the company's data warehouse, and the success of the data warehouse architecture.

In total, 454 respondents provided usable information. Surveyed companies ranged from small (less than \$10 million in revenue) to large (in excess of \$10 billion). Most of the companies were located in the United States (60%) and represented a variety of industries, with the financial services industry (15%) providing the most responses. The predominant architecture was the hub-and-spoke architecture (39%), followed by the bus architecture (26%), the centralized architecture (17%), independent DMs (12%), and the federated architecture (4%). The most common platform for hosting the data warehouses was Oracle (41%), followed by Microsoft (19%) and IBM (18%). The average (mean) gross revenue varied from \$3.7 billion for independent DMs to \$6 billion for the federated architecture.

They used four measures to assess the success of the architectures: (1) information quality, (2) system quality, (3) individual impacts, and (4) organizational impacts. The questions used a 7-point scale, with the higher score indicating a more successful architecture. Table 4.1 shows the average scores for the measures across the architectures.

TABLE 4.1 Average Assessment Scores for the Success of the Architectures					
	Independent DMs	Bus Architecture	Hub-and-Spoke Architecture	Centralized Architecture (No Dependent DMs)	Federated Architecture
Information Quality	4.42	5.16	5.35	5.23	4.73
System Quality	4.59	5.60	5.56	5.41	4.69
Individual Impacts	5.08	5.80	5.62	5.64	5.15
Organizational Impacts	4.66	5.34	5.24	5.30	4.77

As the results of the study indicate, independent DMs scored the lowest on all measures. This finding confirms the conventional wisdom that independent DMs are a poor architectural solution. Next lowest on all measures was the federated architecture. Firms sometimes have disparate decision-support platforms resulting from mergers and acquisitions, and they may choose a federated approach, at least in the short term. The findings suggest that the federated architecture is not an optimal long-term solution. What is interesting, however, is the similarity of the averages for the bus, hub-and-spoke, and centralized architectures. The differences are sufficiently small that no claims can be made for a particular architecture's superiority over the others, at least based on a simple comparison of these success measures.

They also collected data on the domain (e.g., varying from a subunit to company-wide) and the size (i.e., amount of data stored) of the warehouses. They found that the hub-and-spoke architecture is typically used with more enterprise-wide implementations and larger warehouses. They also investigated the cost and time required to implement the different architectures. Overall, the hub-and-spoke architecture was the most expensive and time-consuming to implement.

► SECTION 4.4 REVIEW QUESTIONS

1. What are the key similarities and differences between a two-tiered architecture and a three-tiered architecture?
2. How has the Web influenced data warehouse design?
3. List the alternative data warehousing architectures discussed in this section.
4. What issues should be considered when deciding which architecture to use in developing a data warehouse? List the 10 most important factors.
5. Which data warehousing architecture is the best? Why?

4.5 DATA MANAGEMENT AND WAREHOUSE DEVELOPMENT

A data warehousing project is a major undertaking for any organization and is more complicated than a simple mainframe selection and implementation project because it comprises and influences many departments and many input and output interfaces and it can be part of a CRM business strategy. A data warehouse provides several benefits that can be classified as direct and indirect. Direct benefits include the following:

- End users can perform extensive analysis in numerous ways.
- A consolidated view of corporate data (i.e., a single version of the truth) is possible.
- Better and more timely information is possible. A data warehouse permits information processing to be relieved from costly operational systems onto low-cost servers; therefore, many more end-user information requests can be processed more quickly.
- Enhanced system performance can result. A data warehouse frees production processing because some operational system reporting requirements are moved to DSS.
- Data access is simplified.

Indirect benefits result from end users using these direct benefits. On the whole, these benefits enhance business knowledge, present a competitive advantage, improve customer service and satisfaction, facilitate decision making, and help in reforming business processes; therefore, they are the strongest contributions to competitive advantage (Parzinger & Frolick, 2001). For a detailed discussion of how organizations can obtain exceptional levels of payoffs, see Watson, Goodhue, and Wixom (2002). Given the potential benefits that a data warehouse can provide and the substantial investments

in time and money that such a project requires, it is critical that an organization structure its data warehouse project to maximize the chances of success. In addition, the organization must, obviously, take costs into consideration. Kelly (2001) described an ROI approach that considers benefits in the categories of keepers (i.e., money saved by improving traditional decision support functions), gatherers (i.e., money saved due to automated collection and dissemination of information), and users (i.e., money saved or gained from decisions made using the data warehouse). Costs include those related to hardware, software, network bandwidth, internal development, internal support, training, and external consulting. The net present value is calculated over the expected life of the data warehouse. Because the benefits are broken down approximately as 20% for keepers, 30% for gatherers, and 50% for users, Kelly indicated that users should be involved in the development process, a success factor typically mentioned as critical for systems that imply change in an organization.

Clearly defining the business objective, gathering project support from management end users, setting reasonable time frames and budgets, and managing expectations are critical to a successful data warehousing project. A data warehousing strategy is a blueprint for the successful introduction of the data warehouse. The strategy should describe where the company wants to go, why it wants to go there, and what it will do when it gets there. It needs to take into consideration the organization's vision, structure, and culture. See Matney (2003) for the steps that can help in developing a flexible and efficient support strategy. When the plan and support for a data warehouse are established, the organization needs to examine data warehouse vendors. (See Table 4.2 for a sample list of vendors; also see The Data Warehousing Institute [twidi.org] and Information Builders [informationbuilders.com].) Many vendors provide software demos of their data warehousing and BI products.

Data Warehouse Development Approaches

Many organizations need to create the data warehouses used for decision support. Two competing approaches are employed. The first approach is that of Bill Inmon, who is often called “the father of data warehousing.” Inmon supports a top-down development approach that adapts traditional relational database tools to the development needs of an enterprise-wide data warehouse, also known as the EDW approach. The second approach is that of Ralph Kimball, who proposed a bottom-up approach that employs dimensional modeling, also known as the DM approach. Knowing how these two models are alike and how they differ helps us understand the basic data warehouse concepts. Table 4.3 compares the two approaches. We describe these approaches in detail next.

THE INMON MODEL: THE EDW APPROACH Inmon's approach emphasizes top-down development, employing established database development methodologies and tools, such as entity-relationship diagrams (ERD) and an adjustment of the spiral development approach. The EDW approach does not preclude the creation of DMs. The EDW is the ideal in this approach because it provides a consistent and comprehensive view of the enterprise. Murtaza (1998) presented a framework for developing EDW.

THE KIMBALL MODEL: THE DATA MART APPROACH Kimball's DM strategy is a “plan big, build small” approach. A DM is a subject-oriented or department-oriented data warehouse. It is a scaled-down version of a data warehouse that focuses on the requests of a specific department, such as marketing or sales. This model applies dimensional data modeling, which starts with tables. Kimball advocated a development methodology that entails a bottom-up approach, which in the case of data warehouses means building one DM at a time.

TABLE 4.2 Sample List of Data Warehousing Vendors

Vendor	Product Offerings
Business Objects (businessobjects.com)	A Comprehensive set of BI and data visualization software (now owned by SAP)
Computer Associates (cai.com)	Comprehensive set of data warehouse (DW) tools and products
DataMirror (datamirror.com)	DW administration, management, and performance products
Data Advantage Group (dataadvantagegroup.com)	Metadata software
Dell (dell.com)	DW servers
Embarcadero Technologies (embarcadero.com)	DW administration, management, and performance products
Greenplum (greenplum.com)	Data warehousing and data appliance solution provider (now owned by EMC)
Harte-Hanks (harte-hanks.com)	Customer relationship management (CRM) products and services
HP (hp.com)	DW servers
Hummingbird Ltd. (hummingbird.com)	DW engines and exploration warehouses
Hyperion Solutions (hyperion.com)	Comprehensive set of DW tools, products, and applications
IBM InfoSphere (www-01.ibm.com/software/data/infosphere)	Data integration, DW, master data management, Big Data products
Informatica (informatica.com)	DW administration, management, and performance products
Microsoft (microsoft.com)	DW tools and products
Netezza	DW software and hardware (DW appliance) provider (now owned by IBM)
Oracle (including PeopleSoft and Siebel; oracle.com)	DW, ERP, and CRM tools, products, and applications
SAS Institute (sas.com)	DW tools, products, and applications
Siemens (siemens.com)	DW servers
Sybase (sybase.com)	Comprehensive set of DW tools and applications
Teradata (teradata.com)	DW tools, DW appliances, DW consultancy, and applications

WHICH MODEL IS BEST? There is no one-size-fits-all strategy to data warehousing. An enterprise's data warehousing strategy can evolve from a simple DM to a complex data warehouse in response to user demands, the enterprise's business requirements, and the enterprise's maturity in managing its data resources. For many enterprises, a DM is frequently a convenient first step to acquiring experience in constructing and managing a data warehouse while presenting business users with the benefits of better access to their data; in addition, a DM commonly indicates the business value of data warehousing. Ultimately, engineering an EDW that consolidates old DMs and data warehouses is the ideal solution.

TABLE 4.3 Contrasts between the DM and EDW Development Approaches

Effort	DM Approach	EDW Approach
Scope	One subject area	Several subject areas
Development time	Months	Years
Development cost	\$10,000 to \$100,000+	\$1,000,000+
Development difficulty	Low to medium	High
Data prerequisite for sharing	Common (within business area)	Common (across enterprise)
Sources	Only some operational and external systems	Many operational and external systems
Size	Megabytes to several gigabytes	Gigabytes to petabytes
Time horizon	Near-current and historical data	Historical data
Data transformations	Low to medium	High
Update frequency	Hourly, daily, weekly	Weekly, monthly
Technology		
Hardware	Workstations and departmental servers	Enterprise servers and mainframe computers
Operating system	Windows and Linux	Unix, Z/OS, OS/390
Databases	Workgroup or standard database servers	Enterprise database servers
Usage		
Number of simultaneous users	10s	100s to 1,000s
User types	Business area analysts and managers	Enterprise analysts and senior executives
Business spotlight	Optimizing activities within the business area	Crossfunctional optimization and decision making

Sources: Adapted from Van den Hoven, J. (2003). Data marts: Plan big, build small. In *IS Management Handbook*, 8th ed., Boca Raton, FL: CRC Press; Ariyachandra, T., & Watson, H. (2006b). Which data warehouse architecture is most successful? *Business Intelligence Journal*, 11(1), 4–6.

However, the development of individual DMs can often provide many benefits along the way toward developing an EDW, especially if the organization is unable or unwilling to invest in a large-scale project. DMs can also demonstrate feasibility and success in providing benefits. This could potentially lead to an investment in an EDW. Table 4.4 summarizes the most essential characteristic differences between the two models.

Additional Data Warehouse Development Considerations

Some organizations want to completely outsource their data warehousing efforts. They simply do not want to deal with software and hardware acquisitions, and they do not want to manage their information systems. One alternative is to use hosted data warehouses. In this scenario, another firm—ideally, one that has a lot of experience and expertise—develops and maintains the data warehouse. However, there are security and privacy concerns with this approach. See Technology Insights 4.1 for some details.

TABLE 4.4 Essential Differences between Inmon's and Kimball's Approaches

Characteristic	Inmon	Kimball
<i>Methodology and Architecture</i>		
Overall approach	Top-down	Bottom-up
Architecture structure	Enterprise-wide (atomic) data warehouse "feeds" departmental database	DMs model a single business process, and enterprise consistency is achieved through a data bus and conformed dimensions
Complexity of the method	Quite complex	Fairly simple
Comparison with established development methodologies	Derived from the spiral methodology	Four-step process; a departure from RDBMS methods
Discussion of physical design	Fairly thorough	Fairly light
<i>Data Modeling</i>		
Data orientation	Subject- or data-driven	Process-oriented
Tools	Traditional (entity-relationship diagrams [ERD], data flow diagrams [DFD])	Dimensional modeling; a departure from relational modeling
End-user accessibility	Low	High
<i>Philosophy</i>		
Primary audience	IT professionals	End users
Place in the organization	Integral part of the corporate information factory	Transformer and retainer of operational data
Objective	Deliver a sound technical solution based on proven database methods and technology	Deliver a solution that makes it easy for end users to directly query the data and still get reasonable response times

Sources: Adapted from Breslin, M. (2004, Winter). Data warehousing battle of the giants: Comparing the basics of Kimball and Inmon models. *Business Intelligence Journal*, 9(1), 6–20; Ariyachandra, T., & Watson, H. (2006b). Which data warehouse architecture is most successful? *Business Intelligence Journal*, 11(1).

TECHNOLOGY INSIGHTS 4.1 Hosted Data Warehouses

A hosted data warehouse has nearly the same, if not more, functionality as an on-site data warehouse, but it does not consume computer resources on client premises. A hosted data warehouse offers the benefits of BI minus the cost of computer upgrades, network upgrades, software licenses, in-house development, and in-house support and maintenance.

A hosted data warehouse offers the following benefits:

- Requires minimal investment in infrastructure
- Frees up capacity on in-house systems
- Frees up cash flow
- Makes powerful solutions affordable
- Enables powerful solutions that provide for growth
- Offers better-quality equipment and software

- Provides faster connections
- Enables users to access data from remote locations
- Allows a company to focus on core business
- Meets storage needs for large volumes of data

Despite its benefits, a hosted data warehouse is not necessarily a good fit for every organization. Large companies with revenue upwards of \$500 million could lose money if they already have underused internal infrastructure and IT staff. Furthermore, companies that see the paradigm shift of outsourcing applications as loss of control of their data are not likely to use a BI service provider. Finally, the most significant and common argument against implementing a hosted data warehouse is that it may be unwise to outsource sensitive applications for reasons of security and privacy.

Sources: Compiled from Thornton, M., & Lampa, M. (2002). Hosted data warehouse. *Journal of Data Warehousing*, 7(2), 27–34; Thornton, M. (2002, March 18). What about security? The most common, but unwarranted, objection to hosted data warehouses. *DM Review*, 12(3), 30–43.

Representation of Data in Data Warehouse

A typical data warehouse structure is shown in Figure 4.4. Many variations of data warehouse architecture are possible (see Figure 4.8). No matter what the architecture was, the design of data representation in the data warehouse has always been based on the concept of dimensional modeling. **Dimensional modeling** is a retrieval-based system that supports high-volume query access. Representation and storage of data in a data warehouse should be designed in such a way that not only accommodates but also boosts the processing of complex multidimensional queries. Often, the star schema and the snowflake schema are the means by which dimensional modeling is implemented in data warehouses.

The **star schema** (sometimes referenced as star join schema) is the most commonly used and the simplest style of dimensional modeling. A star schema contains a central fact table surrounded by and connected to several **dimension tables** (Adamson, 2009). The fact table contains a large number of rows that correspond to observed facts and external links (i.e., foreign keys). A fact table contains the descriptive attributes needed to perform decision analysis and query reporting, and foreign keys are used to link to dimension tables. The decision analysis attributes consist of performance measures, operational metrics, aggregated measures (e.g., sales volumes, customer retention rates, profit margins, production costs, scrap rate), and all the other metrics needed to analyze the organization's performance. In other words, the fact table primarily addresses what the data warehouse supports for decision analysis.

Surrounding the central fact tables (and linked via foreign keys) are dimension tables. The dimension tables contain classification and aggregation information about the central fact rows. Dimension tables contain attributes that describe the data contained within the fact table; they address how data will be analyzed and summarized. Dimension tables have a one-to-many relationship with rows in the central fact table. In querying, the dimensions are used to slice and dice the numerical values in the fact table to address the requirements of an ad hoc information need. The star schema is designed to provide fast query-response time, simplicity, and ease of maintenance for read-only database structures. A simple star schema is shown in Figure 4.9a. The star schema is considered a special case of the snowflake schema.

The **snowflake schema** is a logical arrangement of tables in a multidimensional database in such a way that the entity-relationship diagram resembles a snowflake in shape. Closely related to the star schema, the snowflake schema is represented by centralized fact tables (usually only one), which are connected to multiple dimensions. In the snowflake schema, however, dimensions are normalized into multiple

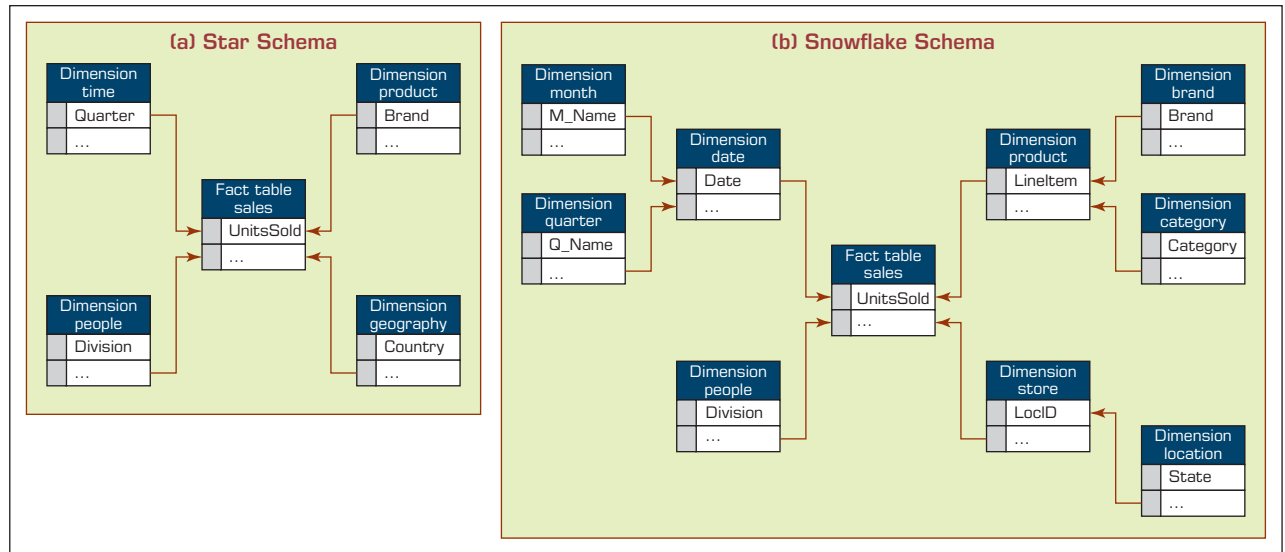


FIGURE 4.9 (a) The Star Schema, and (b) the Snowflake Schema.

related tables, whereas the star schema's dimensions are denormalized, with each dimension being represented by a single table. A simple snowflake schema is shown in Figure 4.9b.

Analysis of Data in Data Warehouse

Once the data is properly stored in a data warehouse, it can be used in various ways to support organizational decision making. OLAP is arguably the most commonly used data analysis technique in data warehouses, and it has been growing in popularity due to the exponential increase in data volumes and the recognition of the business value of data-driven analytics. Simply, OLAP is an approach to quickly answer ad hoc questions by executing multidimensional analytical queries against organizational data repositories (i.e., data warehouses, DMs).

OLAP versus OLTP

OLTP (online transaction processing system) is a term used for a transaction system that is primarily responsible for capturing and storing data related to day-to-day business functions such as ERP, CRM, SCM, POS, and so forth. An OLTP system addresses a critical business need, automating daily business transactions, and running real-time reports and routine analysis. But these systems are not designed for ad hoc analysis and complex queries that deal with a number of data items. OLAP, on the other hand, is designed to address this need by providing ad hoc analysis of organizational data much more effectively and efficiently. OLAP and OLTP rely heavily on each other: OLAP uses the data captured by OLTP, and OLTP automates the business processes that are managed by decisions supported by OLAP. Table 4.5 provides a multicriteria comparison between OLTP and OLAP.

OLAP Operations

The main operational structure in OLAP is based on a concept called cube. A **cube** in OLAP is a multidimensional data structure (actual or virtual) that allows fast analysis of data. It can also be defined as the capability of efficiently manipulating and analyzing data from multiple perspectives. The arrangement of data into cubes aims to overcome a limitation of relational databases: Relational databases are not well suited for

TABLE 4.5 A Comparison between OLTP and OLAP

Criteria	OLTP	OLAP
Purpose	To carry out day-to-day business functions	To support decision making and provide answers to business and management queries
Data source	Transaction database (a normalized data repository primarily focused on efficiency and consistency)	Data warehouse or DM (a nonnormalized data repository primarily focused on accuracy and completeness)
Reporting	Routine, periodic, narrowly focused reports	Ad hoc, multidimensional, broadly focused reports and queries
Resource requirements	Ordinary relational database	Multiprocessor, large-capacity specialized database
Execution speed	Fast (recording of business transaction and routine reports)	Slow (resource intensive, complex, large-scale queries)

near instantaneous analysis of large amounts of data. Instead, they are better suited for manipulating records (adding, deleting, and updating data) that represent a series of transactions. Although many report-writing tools exist for relational databases, these tools are slow when a multidimensional query that encompasses many database tables needs to be executed.

Using OLAP, an analyst can navigate through the database and screen for a particular subset of the data (and its progression over time) by changing the data's orientations and defining analytical calculations. These types of user-initiated navigation of data through the specification of slices (via rotations) and **drill down/up** (via aggregation and disaggregation) is sometimes called "slice and dice." Commonly used OLAP operations include slice and dice, drill down, roll-up, and pivot.

- **Slice.** A slice is a subset of a multidimensional array (usually a two-dimensional representation) corresponding to a single value set for one (or more) of the dimensions not in the subset. A simple slicing operation on a three-dimensional cube is shown in Figure 4.10.
- **Dice.** The dice operation is a slice on more than two dimensions of a data cube.
- **Drill Down/Up.** Drilling down or up is a specific OLAP technique whereby the user navigates among levels of data ranging from the most summarized (up) to the most detailed (down).
- **Roll-up.** A roll-up involves computing all the data relationships for one or more dimensions. To do this, a computational relationship or formula might be defined.
- **Pivot.** This is used to change the dimensional orientation of a report or ad hoc query-page display.

Data Integration and the Extraction, Transformation, and Load (ETL) Processes

Global competitive pressures, demand for return on investment (ROI), management and investor inquiry, and government regulations are forcing business managers to rethink how they integrate and manage their businesses. A decision maker typically needs access to multiple sources of data that must be integrated. Before data warehouses, DMs, and BI software, providing access to data sources was a major, laborious process. Even with modern Web-based data management tools, recognizing what data to access and providing

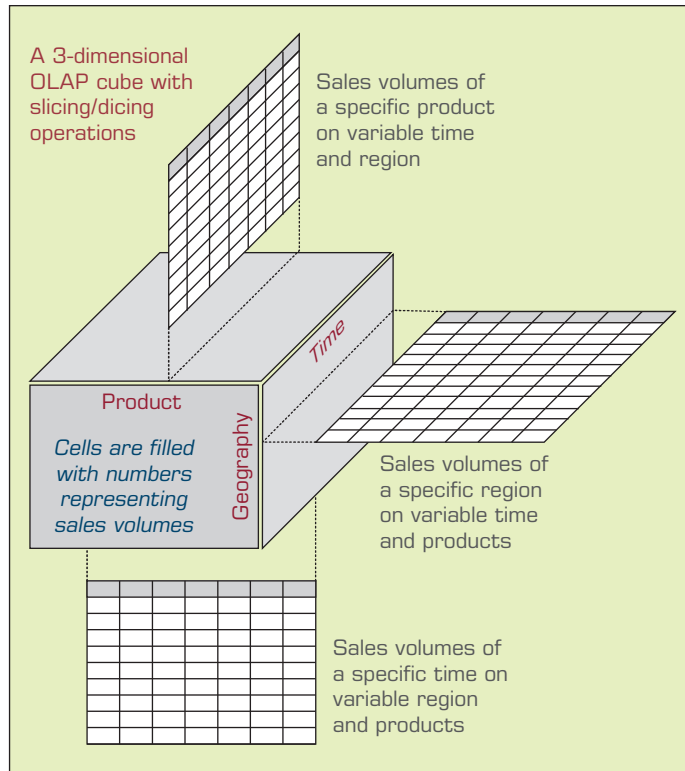


FIGURE 4.10 Slicing Operations on a Simple Three-Dimensional Data Cube.

them to the decision maker is a nontrivial task that requires database specialists. As data warehouses grow in size, the issues of integrating data grow as well.

The business analysis needs continue to evolve. Mergers and acquisitions, regulatory requirements, and the introduction of new channels can drive changes in BI requirements. In addition to historical, cleansed, consolidated, and point-in-time data, business users increasingly demand access to real-time, unstructured, and/or remote data. And everything must be integrated with the contents of an existing data warehouse. Moreover, access via PDAs and through speech recognition and synthesis is becoming more commonplace, further complicating integration issues (Edwards, 2003). Many integration projects involve enterprise-wide systems. Orovic (2003) provided a checklist of what works and what does not work when attempting such a project. Properly integrating data from various databases and other disparate sources is difficult. When it is not done properly, though, it can lead to disaster in enterprise-wide systems such as CRM, ERP, and supply-chain projects (Nash, 2002).

Data Integration

Data integration comprises three major processes that, when correctly implemented, permit data to be accessed and made accessible to an array of ETL and analysis tools and the data warehousing environment: data access (i.e., the ability to access and extract data from any data source), data federation (i.e., the integration of business views across multiple data stores), and change capture (based on the identification, capture, and delivery of the changes made to enterprise data sources). Analytics in Action 4.2 shows how a large nonprofit organization achieved extraordinary results by transforming and modernizing their BI infrastructure with modern-day data warehousing appliances.

ANALYTICS IN ACTION 4.2 AARP Transforms Its BI Infrastructure and Achieves a 347% ROI in Three Years

AARP, Inc., formerly the American Association of Retired Persons, is a U.S.-based membership and interest group, founded in 1958 by Ethel Percy Andrus, PhD, a retired educator from California, and Leonard Davis, founder of Colonial Penn Group of insurance companies. As described in their Web site (aarp.org), AARP is a nonprofit, nonpartisan, social welfare organization with a membership of nearly 38 million that helps people turn their goals and dreams into real possibilities, strengthens communities, and fights for the issues that matter most to families—such as healthcare, employment and income security, and protection from financial abuse.

A Growing Demand for BI

In 2002, the organization first launched a BI initiative that would centralize information (AARP has offices in all 50 states as well as the District of Columbia) and empower its staff with current, relevant, accurate, and flexible analytics to:

- Match services and product offerings to membership base and expectations.
- Improve member profitability, retention, and acquisition.
- Protect AARP brand image by managing relationships with third-party service providers.

This insight helped fuel AARP's success, and with this success came larger data volumes and an increased demand for new analytics.

By 2009, the BI team faced a new challenge. Its data warehouse—based on an SQL relational database from Oracle—could no longer keep up with the demand. The team experienced more than 30 system failures that year. This was both unacceptable and costly.

System performance was a key concern as well. As the data volumes grew, daily loads into the warehouse couldn't be completed until 3:00 P.M.—which affected how long staff had to wait for reports. “Our analysts would run a report, then go for coffee or for lunch, and, maybe if they were lucky, by 5:00 P.M. they would get the response,” says Bruni, Practice Director, Business Intelligence, AARP. “It was unacceptable. The system was so busy writing the new daily data that it didn't give any importance to the read operations performed by users.”

Analysts also couldn't create ad hoc queries without IT intervention. When IT received a request for a new type of report, the BI team would have to optimize the queries and send a report sample back to the requestors for review. The process, from start to finish, could take weeks to months. Finally, with more than 36 terabytes of data in the data warehouse, staff found it impossible to back up the system each night. Backups were limited to a few critical tables, making it difficult for staff to create an effective disaster recovery plan.

According to Bruni, if left unsolved, these challenges could have affected AARP's work. “Analytics provide key metrics that are critical to evaluate how well our membership and social goals are being attained,” says Bruni. “It is essential to enabling continuous improvement and decision making to support member needs.”

Creating an Agile BI Environment

As Bruni's team looked to modernize the BI environment, they evaluated two options—upgrading the existing environment or moving to a single data warehouse appliance. “We found the cost of each option comparable, but only the appliance provided us a paradigm shift in terms of the performance we needed,” says Bruni. “Among the different partners we looked at, the IBM Netezza data warehouse appliance provided the safest bet because it didn't require the data model fine-tuning that other data warehouses do. We were also able to try the solution before we bought it to see whether it really could do everything we needed. Most vendors do not provide this type of ‘try-before-you-buy’ option.”

In building the new environment, the organization adopted a “Scrum” development model, usually used by software developers, to provide a framework that shortens development cycles and speeds time to market for BI requests. “Using Scrum in data warehousing is kind of unheard of,” says Bruni. “But the basic premise it provides is an agile, iterative process that enables us to rapidly transform our users’ analytic needs into operating reports that show meaningful data.”

Within nine months from the acquisition of its new platform, the team had converted all the scripts and procedures from Oracle Database into the IBM® Netezza® data warehouse appliance. Core accounts and membership data (which resides on an IBM DB2® for z/OS® database running on an IBM System z® server), financial and human resource data from other smaller databases, and campaign analysis and segmentation data from third-party data sources are now loaded in the IBM Netezza data warehouse appliance nightly and accessible via the organization’s BI tools without interruption.

Running Complex Queries at Lightning Speed

In terms of performance (which was the BI team’s most pressing concern), daily data loads are now completed before 8:00 A.M.—a 1400% improvement—and reports that previously took minutes to run are completed in seconds—a 1700% improvement. The solution also helped compress the data size from 36 terabytes to just 1.5 terabytes, enabling staff to easily back up the data warehouse in only 30 minutes.

Equally important, the nearly 220 human resources, finance, marketing, and campaign staff members that use the system can now conduct what Bruni refers to as “train-of-thought analysis”—creating ad hoc reports to test theories regarding membership needs. “The IBM Netezza data warehouse appliance is like driving a Ferrari,” says Bruni. “We have opened a whole new realm of possibilities to our internal customers, who are actually able to create reports on-the-fly and get the results back in a matter of seconds. In the first few months of operation, we saw a huge spike in the number of reports being created—nearly three times the number that we had previously supported. With the deep dive they can conduct now, we’ve seen a steady growth in member renewals, acquisitions and engagement.”

Achieving Rapid ROI

The new platform has also enabled the organization to redeploy IT support staff from the BI group to other areas. Previously, the team needed one full-time database administrator (DBA) along with part-time support from the organization’s storage area network (SAN) and midrange service teams. “It’s amazing,” says Bruni. “We no longer need IT support. The IBM Netezza data warehouse appliance is shipped already optimized. Give it power, give it network, and you’re done. It doesn’t need anything else.”

These improvements have enabled the organization to realize a 9% return on investment in the first year, with an anticipated 274% ROI by the second year, and a 347% investment by the third year. “Our initial analysis projected a positive ROI already in the first year—which is very unusual for infrastructure upgrades given all costs are incurrent in the first year,” says Bruni. “Our actual ROI post-implementation was even higher as we completed the swap three months ahead of schedule.”

Expanding the Influence of BI

By modernizing its infrastructure, Bruni’s team has elevated the value and perception of BI in the organization. “After we moved to IBM Netezza, the word spread that we were doing things right and that leveraging us as an internal service was really smart,” says Bruni. “We’ve gained new mission-critical areas, such as the social-impact area which supports our Drive to End Hunger and Create the Good campaigns, based on the fact that we have such a robust infrastructure and that we changed our approach to business. We can develop in a more agile way from a development standpoint. From a program management standpoint, it shrinks our release cycles from months, which is typical with traditional data warehouse infrastructures, to just weeks.”

QUESTIONS FOR DISCUSSION

1. What were the challenges AARP was facing?
2. What was the approach for a potential solution?
3. What were the results obtained in the short term, and what were the future plans?

Source: IBM customer success story. (2021). “AARP transforms its business intelligence infrastructure—Achieving a 347% ROI in three years from BI modernization effort,” <http://www-03.ibm.com/software/businesscasestudies/us/en/corp?synkey=A735189Y23828M82> (accessed June 2021) and <https://www.merkle.com/thought-leadership/case-studies/aarpr-launches-digital-transformation-drive-personalization> (accessed May 2023).

A major purpose of a data warehouse is to integrate data from multiple systems. Various integration technologies enable data and metadata integration:

- Enterprise application integration (EAI)
- Service-oriented architecture (SOA)
- Enterprise information integration (EII)
- Extraction, transformation, and load (ETL)

Enterprise application integration (EAI) provides a vehicle for pushing data from source systems into the data warehouse. It involves integrating application functionality and is focused on sharing functionality (rather than data) across systems, thereby enabling flexibility and reuse. Traditionally, EAI solutions have focused on enabling application reuse at the application programming interface level. Recently, EAI is accomplished by using SOA coarse-grained services (a collection of business processes or functions) that are well defined and documented. Using Web services is a specialized way of implementing an SOA. EAI can be used to facilitate data acquisition directly into a near-real-time data warehouse or to deliver decisions to the OLTP systems. There are many different approaches to and tools for EAI implementation.

Enterprise information integration (EII) is an evolving tool space that promises real-time data integration from a variety of sources, such as relational databases, Web services, and multidimensional databases. It is a mechanism for pulling data from source systems to satisfy a request for information. EII tools use predefined metadata to populate views that make integrated data appear relational to end users. XML may be the most important aspect of EII because XML allows data to be tagged either at creation time or later. These tags can be extended and modified to accommodate almost any area of knowledge (see Kay, 2005).

Physical data integration has conventionally been the main mechanism for creating an integrated view with data warehouses and DMs. With the advent of EII tools (see Kay, 2005), new virtual data integration patterns are feasible. Manglik and Mehra (2005) discussed the benefits and constraints of new data integration patterns that can expand traditional physical methodologies to present a comprehensive view for the enterprise.

We next turn to the approach for loading data into the warehouse: ETL.

Extraction, Transformation, and Load

At the heart of the technical side of the data warehousing process is **extraction, transformation, and load (ETL)**. ETL technologies, which have existed for some time, are instrumental in the process and use of data warehouses. The ETL process is an integral component in any data-centric project. IT managers are often faced with challenges because the ETL process typically consumes 70% of the time in a data-centric project.

The ETL process consists of extraction (i.e., reading data from one or more databases), transformation (i.e., converting the extracted data from its previous form into the form in which it needs to be so that it can be placed into a data warehouse or simply another database), and load (i.e., putting the data into the data warehouse). Transformation occurs by using rules or lookup tables or by combining the data with other data. The three database functions are integrated into one tool to pull data out of one or more databases and place them into another, consolidated database or a data warehouse.

ETL tools also transport data between sources and targets, document how data elements (e.g., metadata) change as they move between source and target, exchange metadata with other applications as needed, and administer all runtime processes and operations (e.g., scheduling, error management, audit logs, statistics). ETL is extremely important for data integration as well as for data warehousing. The purpose of the ETL process is to load the warehouse with integrated and cleansed data. The data used in ETL processes can come from any source: a mainframe application, an ERP application, a CRM tool, a flat file, an Excel spreadsheet, or even a message queue. In Figure 4.11, we outline the ETL process.

The process of migrating data to a data warehouse involves the extraction of data from all relevant sources. Data sources may consist of files extracted from OLTP databases, spreadsheets, personal databases (e.g., Microsoft Access), or external files. Typically, all the input files are written to a set of staging tables, which are designed to facilitate the load process. A data warehouse contains numerous business rules that define such things as how the data will be used, summarization rules, standardization of encoded attributes, and calculation rules. Any data quality issues pertaining to the source files need to be corrected before the data are loaded into the data warehouse. One of the benefits of a well-designed data warehouse is that these rules can be stored in a metadata repository and applied to the data warehouse centrally. This differs from an OLTP approach, which typically has data and business rules scattered throughout the system. The process of loading data into a data warehouse can be performed either through data transformation tools that provide a GUI to aid in the development and maintenance of business rules or through more traditional methods, such as developing programs or utilities to load the data warehouse, using programming languages such as PL/SQL, C++, Java, or .NET Framework languages. This decision is not easy for organizations. Several issues affect

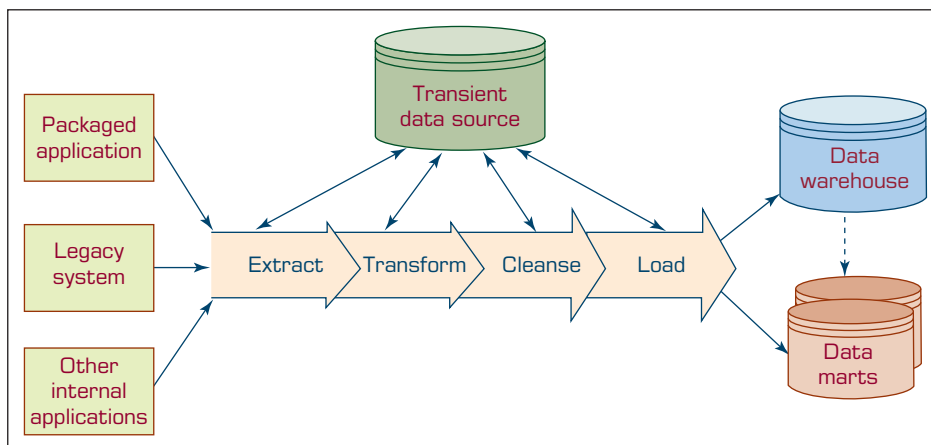


FIGURE 4.11 The ETL Process.

whether an organization will purchase data transformation tools or build the transformation process itself:

- Data transformation tools are expensive.
- Data transformation tools may have a long learning curve.
- It is difficult to measure how the IT organization is doing until it has learned to use the data transformation tools.

In the long run, a transformation-tool approach should simplify the maintenance of an organization's data warehouse. Transformation tools can also be effective in detecting and scrubbing (i.e., removing any anomalies in the data). OLAP and data mining tools rely on how well the data are transformed.

As an example of effective ETL, Motorola, Inc., uses ETL to feed its data warehouses. Motorola collects information from 30 different procurement systems and sends them to its global SCM data warehouse for analysis of aggregate company spending (see Songini, 2004).

Solomon (2005) classified ETL technologies into four categories: sophisticated, enabler, simple, and rudimentary. It is generally acknowledged that tools in the sophisticated category will result in the ETL process being better documented and more accurately managed as the data warehouse project evolves.

Even though it is possible for programmers to develop software for ETL, it is simpler to use an existing ETL tool. The following are some of the important criteria in selecting an ETL tool (see Brown, 2004):

- Ability to read from and write to an unlimited number of data source architectures
- Automatic capturing and delivery of metadata
- A history of conforming to open standards
- An easy-to-use interface for the developer and the functional user

Performing extensive ETL may be a sign of poorly managed data and a fundamental lack of a coherent data management strategy. There is a direct correlation between the extent of redundant and inconsistent data and the number of ETL tasks performed in the data pre-processes. When data are managed correctly as an enterprise asset, ETL efforts are significantly reduced, and redundant data are eliminated. This leads to huge savings in data maintenance and greater efficiency in new developments while also improving data quality. Poorly designed ETL processes are costly to maintain, change, and update. Consequently, it is crucial to make the proper choices in terms of the technology and tools to use for developing and maintaining the ETL process.

Numerous packaged ETL tools are available in the market. Database and data warehouse vendors currently offer ETL capabilities that both enhance and compete with independent ETL tools. SAS acknowledges the importance of data quality and offers the industry's first fully integrated solution that merges ETL and data quality to transform data into strategic valuable assets. Other integrated ETL software providers include Microsoft, Oracle, IBM, Informatica, and Salesforce.

► SECTION 4.5 REVIEW QUESTIONS

1. List the benefits of data warehouses.
2. List several criteria for selecting a data warehouse vendor, and describe why they are important.
3. What is OLAP, and how does it differ from OLTP?
4. What is a cube? What do drill down, roll-up, slice, and dice mean?
5. Describe data integration.
6. Describe the three steps of the ETL process.
7. Why is the ETL process so important for data warehousing efforts?

4.6 DATA WAREHOUSE ADMINISTRATION, SECURITY ISSUES, AND FUTURE TRENDS

Data warehouses provide a distinct competitive edge to enterprises that effectively create and use them. Due to its huge size and its intrinsic nature, a data warehouse requires especially strong monitoring to sustain satisfactory efficiency and productivity. The successful administration and management of a data warehouse entails skills and proficiency that go past what is required of a traditional database administrator (DBA). A **data warehouse administrator (DWA)** should be familiar with high-performance software, hardware, and networking technologies. He or she should also possess solid business insight. Because data warehouses feed BI systems and DSS that help managers with their decision-making activities, the DWA should be familiar with the decision-making processes to suitably design and maintain the data warehouse structure. It is particularly significant for a DWA to keep the existing requirements and capabilities of the data warehouse stable while simultaneously providing flexibility for rapid improvements. Finally, a DWA must possess excellent communication skills. See Benander, Benander, Fadlalla, and James (2000) for a description of the key differences between a DBA and a DWA.

Security and privacy of information are main and significant concerns for a data warehouse professional. The U.S. government has passed regulations (e.g., the Gramm-Leach-Bliley privacy and safeguards rules, the Health Insurance Portability and Accountability Act of 1996 [HIPAA]), instituting obligatory requirements in the management of customer information. Hence, companies must create security procedures that are effective yet flexible to conform to numerous privacy regulations. According to Elson and LeClerc (2005), effective security in a data warehouse should focus on four main areas:

- Establishing effective corporate and security policies and procedures. An effective security policy should start at the top, with executive management, and should be communicated to all individuals within the organization.
- Implementing logical security procedures and techniques to restrict access. This includes user authentication, access controls, and encryption technology.
- Limiting physical access to the data center environment.
- Establishing an effective internal control review process with an emphasis on security and privacy.

In the near term, data warehousing developments will be determined by noticeable factors (e.g., data volumes, increased intolerance for latency, the diversity and complexity of data types) and less noticeable factors (e.g., unmet end-user requirements for dashboards, balanced scorecards, master data management, information quality). Given these drivers, Moseley (2009) and Agosta (2006) suggested that data warehousing trends will lean toward simplicity, value, and performance.

The Future of Data Warehousing

The field of data warehousing has been a vibrant area in IT in the last couple of decades, and the evidence in the BI/BA and Big Data world shows that the importance of the field will only get even more interesting. Following are some of the recently popularized concepts and technologies that will play a significant role in defining the future of data warehousing.

Sourcing (mechanisms for acquisition of data from diverse and dispersed sources):

- **Web, social media, and Big Data.** The recent upsurge in the use of the Web for personal as well as business purposes coupled with the tremendous interest in social media creates opportunities for analysts to tap into very rich data sources. Because of the sheer volume, velocity, and variety of the data, a new term, “Big

Data,” has been coined to name the phenomenon. Taking advantage of Big Data requires development of new and dramatically improved BI/BA technologies, which will result in a revolutionized data warehousing world.

- **Open source software.** Use of open source software tools is increasing at an unprecedented level in warehousing, BI, and data integration. There are good reasons for the upswing of open source software used in data warehousing (Russom, 2009): (1) the recession has driven up interest in low-cost open source software, (2) open source tools are coming into a new level of maturity, and (3) open source software augments traditional enterprise software without replacing it.
- **SaaS (software as a service), “The Extended ASP Model.”** SaaS is a creative way of deploying information systems applications where the provider licenses its applications to customers for use as a service on demand (usually over the Internet). SaaS software vendors may host the application on their own servers or upload the application to the consumer site. In essence, SaaS is the new and improved version of the ASP model. For data warehouse customers, finding SaaS-based software applications and resources that meet specific needs and requirements can be challenging. As these software offerings become more agile, the appeal and the actual use of SaaS as the choice of data warehousing platform will also increase.
- **Cloud computing.** Cloud computing is perhaps the newest and the most innovative platform choice to come along in years. Numerous hardware and software resources are pooled and virtualized, so that they can be freely allocated to applications and software platforms as resources are needed. This enables information systems applications to dynamically scale up as workloads increase. Although cloud computing and similar virtualization techniques are fairly well established for operational applications today, they are just now starting to be used as data warehouse platforms of choice. The dynamic allocation of a cloud is particularly useful when the data volume of the warehouse varies unpredictably, making capacity planning difficult.
- **Data lakes.** With the emergence of Big Data, there came a new data platform: data lake, which is a large storage location that can hold vast quantities of data (mostly unstructured) in its native/raw format for future/potential analytics consumption. Traditionally speaking, whereas a data warehouse stores structured data, a data lake stores all kinds of data. While they are both data storage mechanisms, a data warehouse is all about structured/tabular data and a data lake is about all types of data. Although much has been said and written about the relationship between the two (some of which suggests that data lake is the future name of data warehouses), as it stands, a data lake is not a replacement for a data warehouse; rather, they are complementary to one another. Inmon et al. (2021) provide some invaluable insight in developing successful data lake houses. Technology Insights 4.2 digs deeper into explaining data lakes and their role in the worlds of data warehousing and big data analytics.

TECHNOLOGY INSIGHTS 4.2 Data Lakes

With the emergence of Big Data phenomenon, a new term, “data lake,” has been coined. Many believe that a data lake is just the reincarnation of the good old data warehouse. The underlying assumption suggests that in the age of Big Data, the old way of data storage is not suitable (or sufficient) and therefore a new way of data storage/management is needed, which is paving the way for data lakes. Although most believe that a data lake is the way to go and are ready to jump in, others are standing back and being more cautious (and perhaps skeptical of its viability), calling it a swamp. So, what really is a data lake? Simply put, a data lake is a large storage location that can hold huge quantities of data (structured, unstructured, or semistructured) in its

native/raw format for a potential future use. Whereas a data warehouse stores structured data in related tables, files, or folders, a data lake uses a loosely defined (i.e., unstructured) architecture to store all kinds of data. The main commonality between a data lake and data warehouse is that they are both data storage mechanisms, and conversely, the main difference is that one is all about structured/tabular data and the other is about all kinds of data (i.e., Big Data).

Although the definition of data lake changes among data scientists, the most commonly used definition of it comes from James Dixon, the founder and CTO of Pentaho, who has also been credited with coming up with the term itself. This is how he describes a data lake (Dixon, 2010):

If you think of a data mart as a store of bottled water—cleansed and packaged and structured for easy consumption—the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples.

Perhaps the best way to characterize a data lake is to compare it to a data warehouse in a multidimensional table. Table 4.6 is a summary table (followed by brief descriptions) of the most commonly used dimensions to compare a data lake to a data warehouse (Dull, 2021; Campbell, 2015).

Data. A data warehouse only stores data that has been modeled/aggregated/structured, whereas a data lake stores all kinds of data—structured, semistructured, and unstructured—in its native/raw format.

Processing. Before loading data into a data warehouse, we first need to give it some shape and structure—that is, we need to model it into a star or snowflake schema, which is called schema-on-write. With a data lake, we just load in the raw data, as-is, and then when we are ready to use the data, we give it a shape or structure, which is called schema-on-read. These are two very different processing approaches.

Retrieval speed. For more than two decades, many algorithms have been developed to improve the speed at which the data is retrieved from large and feature-rich data warehouses. Such techniques included triggers, columnar data representation, in-database processing. As of now, the retrieval of data (which can be in any form or fashion—including unstructured text) is a time-demanding activity.

Storage. One of the primary features of Big Data technologies like Hadoop is that the cost of storing data is relatively low as compared to the data warehouse. There are two key reasons for this: First, Hadoop is open source software, so the licensing and community support is free. And second, Hadoop is designed to be installed on low-cost commodity hardware.

Agility. A data warehouse is a highly structured repository, by definition. It's not technically hard to change the structure, but it can be very time-consuming given all the business

TABLE 4.6 A Simple Comparison between a Data Warehouse and a Data Lake

Dimension	Data Warehouse	Data Lake
The nature of data	Structured, processed	Any data in raw/native format
Processing	Schema-on-write (SQL)	Schema-on-read (NoSQL)
Retrieval speed	Very fast	Slow
Cost	Expensive for large data volumes	Designed for low-cost storage
Agility	Less agile, fixed configuration	Highly agile, flexible configuration
Novelty/newness	Not new/matured	Very new/maturing
Security	Well-secured	Not yet well-secured
Users	Business professionals	Data scientists

processes that are tied to it. A data lake, on the other hand, lacks the structure of a data warehouse—which gives developers and data scientists the ability to easily configure and reconfigure their models, queries, and apps on-the-fly.

Novelty/newness. The technologies underlying data warehousing have been around for a long time. Most of the innovations have been accomplished in the last 20–30 years. Therefore, there is very little, if any, newness coming out of data warehousing (with the exclusion of the technologies to leverage and use Big Data within a data warehouse). On the other hand, data lakes are new and are going through a novelty/innovation phase to become a mainstream data storage technology.

Security. Because data warehouse technologies have been around for decades the ability to secure data in a data warehouse is much more mature than securing data in a data lake. It should be noted, however, that there is a significant effort being placed on security right now in the Big Data industry. It is not a question of if, but when, the security of the data lakes will meet the needs and wants of the analytics professionals and other end users.

Users. For a long time, the motto in the analytic world has been “Business intelligence and analytics for everyone!” We have built the data warehouse and invited “everyone” to come, but have they come? On average, only 20–25% of them have. Is it the same cry for the data lake? Will we build the data lake and invite everyone to come? Maybe in the future. For now, a data lake, at this point in its maturity, is best suited for the data scientists.

In summary, a data lake and a data warehouse are not the same. Also, the data lake is not Data Warehouse 2.0 (as suggested by some) or a replacement for the data warehouse. They both are needed and therefore optimized for different data mediums and different tasks/purposes. That is, they need to coexist in the analytics world (at least for a while—until data lakes mature to a level to accomplish what nowadays data warehouses are good for). The goal is to design and properly use each for what they were intended to do—use the best option for the job, which may turn out to be a hybrid of data warehouse and data lake storage mediums.

Sources: Inmon, B. (2021). Five Steps to a Successful Data Lakehouse, retrieved at databricks.com/wp-content/uploads/2021/10/The-Data-Lakehouse.pdf (accessed June 2022); Campbell, C. (2015). Top five differences between data lakes and data warehouses. www.blue-granite.com/blog/bid/402596/Top-Five-Differences-between-Data-Lakes-and-Data-Warehouses (accessed July 2022); Dull, T. (2021). Data lake vs data warehouse: Key differences. KDnuggets.com. <http://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html> (accessed August 2016).

Infrastructure (architectural—hardware and software—enhancements):

- **Columnar (a new way to store and access data in the database).** A column-oriented database management system (also commonly called a *columnar database*) is a system that stores data tables as sections of columns of data rather than as rows of data (which is the way most RDBMS do it). That is, these columnar databases store data by columns instead of rows (all values of a single column are stored consecutively on disk memory). Such a structure gives a much finer grain of control to the RDBMS. It can access only the columns required for the query as opposed to being forced to access all columns of the row. It performs significantly better for queries that need a small percentage of the columns in the tables they are in, but performs significantly worse when you need most of the columns due to the overhead in attaching all of the columns together to form the result sets. Comparisons between row-oriented and column-oriented data layouts are typically concerned with the efficiency of hard disk access for a given workload (which happens to be one of the most time-consuming operations in a computer). Based on the task at hand, one may have significant advantages over the other. Column-oriented organizations are more efficient when (1) an aggregate needs to be computed over many rows but only for a notably smaller subset of all columns of data, because reading that smaller subset of data can be faster

than reading all data; and (2) new values of a column are supplied for all rows at once because that column data can be written efficiently and replace old column data without touching any other columns for the rows. Row-oriented organizations are more efficient when (1) many columns of a single row are required at the same time, and when row size is relatively small, as the entire row can be retrieved with a single-disk seek; and (2) writing a new row if all of the column data is supplied at the same time, as the entire row can be written with a single disk seek. In addition, because the data stored in a column is of uniform type, it lends itself better for compression. That is, significant storage size optimization is available in column-oriented data that is not available in row-oriented data. Such optimal compression of data reduces storage size, making it more economically justifiable to pursue in-memory or solid-state storage alternatives.

- **Real-time data warehousing.** **Real-time data warehousing (RDW)** implies that the refresh cycle of an existing data warehouse to update the data is more frequent (almost at the same time as the data becomes available at operational databases). These RDW systems can achieve near real-time updates of data, where the data latency typically is in the range from minutes to hours. As the latency gets smaller, the cost of data updates seems to be increasing exponentially. Future advancements in many technological fronts (ranging from automatic data acquisition to intelligent software agents) are needed to make RDW a reality with an affordable price tag.
- **Data warehouse appliances (all-in-one solutions to DW).** A data warehouse appliance consists of an integrated set of servers, storage, operating system(s), database management systems, and software specifically preinstalled and preoptimized for data warehousing. In practice, data warehouse appliances provide solutions for the mid-warehouse to Big Data warehouse market, offering low-cost performance on data volumes in the terabyte to petabyte range. To improve performance, most data warehouse appliance vendors use massively parallel processing architectures. Even though most database and data warehouse vendors provide appliances nowadays, many believe that Teradata was the first to provide a commercial data warehouse appliance product. What is most currently observed is the emergence of data warehouse bundles where vendors combine their hardware and database software as a data warehouse platform. From a benefits standpoint, data warehouse appliances have significantly low total cost of ownership, which includes initial purchase costs, ongoing maintenance costs, and the cost of changing capacity as the data grows. The resource cost for monitoring and tuning the data warehouse makes up a large part of the total cost of ownership, often as much as 80%. DW appliances reduce administration for day-to-day operations, setup, and integration. Because data warehouse appliances provide a single-vendor solution, they tend to better optimize the hardware and software within the appliance. Such a unified integration maximizes the chances of successful integration and testing of the DBMS, storage, and operating system by avoiding some of the compatibility issues that arise from multivendor solutions. A data warehouse appliance also provides a single point of contact for problem resolution and a much simpler upgrade path for both software and hardware.
- **Data management technologies and practices.** Some of the most pressing needs for a next-generation data warehouse platform involve technologies and practices that we generally don't think of as part of the platform. In particular, many users need to update the data management tools that process data for use through the data warehousing. The future holds strong growth for master data management (MDM). This relatively new, but extremely important, concept is gaining popularity for many reasons, including the following: (1) tighter integration with

operational systems demands MDM; (2) most data warehouses still lack MDM and data quality functions; and (3) regulatory and financial reports must be perfectly clean and accurate.

- ***In-database processing technology (putting the algorithms where the data is).*** In-database processing (also called *in-database analytics*) refers to the integration of the algorithmic extent of data analytics into data warehousing. By doing so, the data and the analytics that work off the data live within the same environment. Having the two in close proximity increases the efficiency of the computationally intensive analytics procedures. Today, many large database-driven decision support systems, such as those used for credit card fraud detection and investment bank risk management, use this technology because it provides significant performance improvements over traditional methods in a decision environment where time is of the essence. In-database processing is a complex endeavor compared to the traditional way of conducting analytics, where the data is moved out of the database (often in a flat-file format that consists of rows and columns) into a separate analytics environment (such as SAS Enterprise Modeler, Statistica Data Miner, or IBM SPSS Modeler) for processing. In-database processing makes more sense for high-throughput, real-time application environments, including fraud detection, credit scoring, risk management, transaction processing, pricing and margin analysis, usage-based micro-segmenting, behavioral ad targeting, and recommendation engines, such as those used by customer service organizations to determine next-best actions. In-database processing is performed and promoted as a feature by many of the major data warehousing vendors, including Teradata (integrating SAS analytics capabilities into the data warehouse appliances), IBM Netezza, EMC Greenplum, and Sybase, among others.
- ***In-memory storage technology (moving the data in the memory for faster processing).*** Conventional database systems, such as RDBMS, typically use physical hard drives to store data for an extended period of time. When a data-related process is requested by an application, the database management system loads the data (or parts of the data) into the main memory, processes it, and responds back to the application. Although data (or parts of the data) are cached temporarily in the main memory in a database management system, the primary storage location remains a magnetic hard disk. In contrast, an in-memory database system keeps the data permanently in the main memory. When a data-related process is requested by an application, the database management system directly accesses the data, which is already in the main memory, processes it, and responds back to the requesting application. This direct access to data in main memory makes the processing of data orders of magnitude faster than the traditional method. So the main benefit of in-memory technology (may be the only benefit of it) is the incredible speed at which it accesses the data. The disadvantages include the cost of paying for a very large main memory (even though it is getting cheaper, it still costs a great deal to have a large enough main memory to hold all of a company's data) and the need for sophisticated data recovery strategies (because main memory is volatile and can be wiped out accidentally).
- ***New database management systems.*** A data warehouse platform consists of several basic components, of which the most critical is the database management system (DBMS). This is only natural, given the fact that the DBMS is the component of the platform where the most work must be done to implement a data model and optimize it for query performance. Therefore, the DBMS is where many next-generation innovations are expected to happen.
- ***Advanced analytics.*** There are different analytic methods users can choose as they move beyond basic OLAP-based methods and into advanced analytics. Some

users choose advanced analytic methods based on data mining, predictive analytics, statistics, artificial intelligence, and so on. Still, the majority of users seem to be choosing SQL-based methods. Whether SQL-based or not, advanced analytics seem to be among the most important promises of next-generation data warehousing.

In this evolving big data era, the future of data warehousing seems to be full of promises, but at the same time, many uncertainties and significant challenges. Whatever the name may be, let it be data warehousing, data lakes, data lake houses, data reservoir, or any creative combination of these popular terms, the importance of collecting and analyzing data for timely and accurate decision making will not go away. In fact, as the world of business becomes more global and complex, the need for analytics and data warehousing tools will become more prominent. The fast-improving IT tools and techniques seem to be moving in the right direction to address the needs of the future decision support systems.

► SECTION 4.6 REVIEW QUESTIONS

1. What steps can an organization take to ensure the security and confidentiality of customer data in its data warehouse?
2. What skills should a DWA possess? Why?
3. What recent technologies may shape the future of data warehousing? Why?

4.7 BUSINESS REPORTING

Decision makers are in need of information to make accurate and timely decisions. Information is essentially the contextualization of data. In addition to statistical means that were explained in the previous section, information (descriptive analytics) can also be obtained using online analytics processing [OLTP] systems (see the simple taxonomy of descriptive analytics in Chapter 3, Figure 3.11). The information is usually provided to the decision makers in the form of a written report (digital or on paper), although it can also be provided orally. Simply put, a **report** is any communication artifact prepared with the specific intention of conveying information in a digestible form to whoever needs it, whenever and wherever they may need it. It is usually a document that contains information (usually driven from data) organized in a narrative, graphic, and/or tabular form, prepared periodically (recurring) or on an as-needed (ad hoc) basis, referring to specific time periods, events, occurrences, or subjects. Business reports can fulfill many different (but often related) functions. Here are a few of the most prevailing ones:

- To ensure that all departments are functioning properly
- To provide information
- To provide the results of an analysis
- To persuade others to act
- To create an organizational memory (as part of a knowledge management system)

Business reporting (also called OLAP or BI) is an essential part of the larger drive toward improved, evidence-based, optimal managerial decision making. The foundation of these **business reports** is various sources of data coming from both inside and outside the organization (online transaction processing [OLTP] systems). Creation of these reports involves ETL (extract, transform, and load) procedures in coordination with a data warehouse and then using one or more reporting tools (see Chapter 3 for a detailed description of these concepts).

Due to the rapid expansion of information technology coupled with the need for improved competitiveness in business, there has been an increase in the use of computing power to produce unified reports that join different views of the enterprise in one

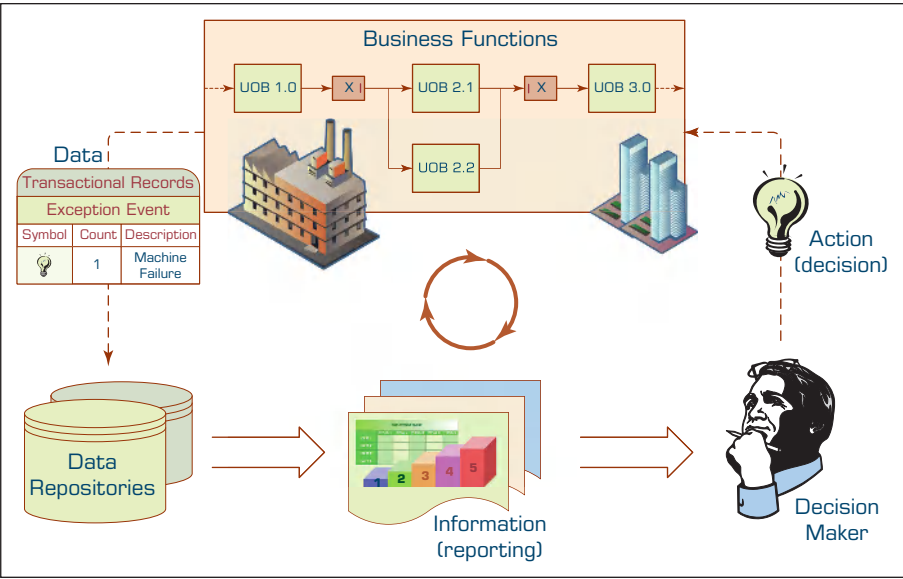


FIGURE 4.12 The Role of Information Reporting in Managerial Decision Making.

place. Usually, this reporting process involves querying structured data sources, most of which were created using different logical data models and data dictionaries, to produce a human-readable, easily digestible report. These types of business reports allow managers and coworkers to stay informed and involved, review options and alternatives, and make informed decisions. Figure 4.12 shows the continuous cycle of data acquisition → information generation → decision making → business process management. Perhaps the most critical task in this cyclical process is the reporting (i.e., information generation)—converting data from different sources into actionable information.

Key to any successful report are clarity, brevity, completeness, and correctness. The nature of the report and the level of importance of these success factors change significantly based on for whom the report is created. Most of the research in effective reporting is dedicated to internal reports that inform stakeholders and decision makers within the organization. There are also external reports between businesses and the government (e.g., for tax purposes or for regular filings to the Securities and Exchange Commission). Even though there are a wide variety of business reports, the ones that are often used for managerial purposes can be grouped into three major categories (Hill, 2016).

METRIC MANAGEMENT REPORTS In many organizations, business performance is managed through outcome-oriented metrics. For external groups, these are service-level agreements. For internal management, they are **key performance indicators (KPIs)**. Typically, there are enterprise-wide agreed targets to be tracked against over a period of time. They may be used as part of other management strategies such as Six Sigma or Total Quality Management.

DASHBOARD-TYPE REPORTS A popular idea in business reporting in recent years has been to present a range of different performance indicators on one page, like a dashboard in a car. Typically, dashboard vendors would provide a set of predefined reports with static elements and fixed structure, but also allow for customization of the dashboard widgets, views, and set targets for various metrics. It's common to have color-coded traffic lights defined for performance (red, orange, green) to draw management's attention to particular areas. A more detailed description of dashboards can be found in later part of this chapter.

BALANCED SCORECARD-TYPE REPORTS This is a method developed by Kaplan and Norton that attempts to present an integrated view of success in an organization. In addition to financial performance, balanced scorecard-type reports also include customer, business process, and learning and growth perspectives. More details on balanced scorecards are provided later in this chapter.

► SECTION 4.7 REVIEW QUESTIONS

1. What is a report? What are reports used for?
2. What is a business report? What are the main characteristics of a good business report?
3. Describe the cyclic process of management, and comment on the role of business reports.
4. List and describe the three major categories of business reports.
5. What are the main components of a business reporting system?

4.8 DATA VISUALIZATION

Data visualization (or more appropriately, information visualization) has been defined as “the use of visual representations to explore, make sense of, and communicate data” (Few, 2007). Although the name that is commonly used is *data visualization*, usually what is meant by this is information visualization. Because information is the aggregation, summarization, and contextualization of data (raw facts), what is portrayed in visualizations is the information and not the data. However, because the two terms *data visualization* and *information visualization* are used interchangeably and synonymously, in this chapter we will follow suit.

Data visualization is closely related to the fields of information graphics, information visualization, scientific visualization, and statistical graphics. Until recently, the major forms of data visualization available in both BI applications have included charts and graphs, as well as the other types of visual elements used to create scorecards and dashboards.

To better understand the current and future trends in the field of data visualization, it helps to begin with some historical context.

A Brief History of Data Visualization

Despite the fact that predecessors to data visualization date back to the second century AD, most developments have occurred in the last two and a half centuries, predominantly during the last 30 years (Few, 2007). Although visualization has not been widely recognized as a discipline until fairly recently, today’s most popular visual forms date back a few centuries. Geographical exploration, mathematics, and popularized history spurred the creation of early maps, graphs, and timelines as far back as the 1600s, but William Playfair is widely credited as the inventor of the modern chart, having created the first widely distributed line and bar charts in his *Commercial and Political Atlas of 1786* and what is generally considered to be one of the first time series portraying line charts in his *Statistical Breviary*, published in 1801 (see Figure 4.13).

Perhaps the most notable innovator of information graphics during this period was Charles Joseph Minard, who graphically portrayed the losses suffered by Napoleon’s army in the Russian campaign of 1812 (see Figure 4.14). Beginning at the Polish–Russian border, the thick band shows the size of the army at each position. The path of Napoleon’s retreat from Moscow in the bitterly cold winter is depicted by the dark lower band, which is tied to temperature and time scales. Popular visualization expert, author, and critic Edward Tufte

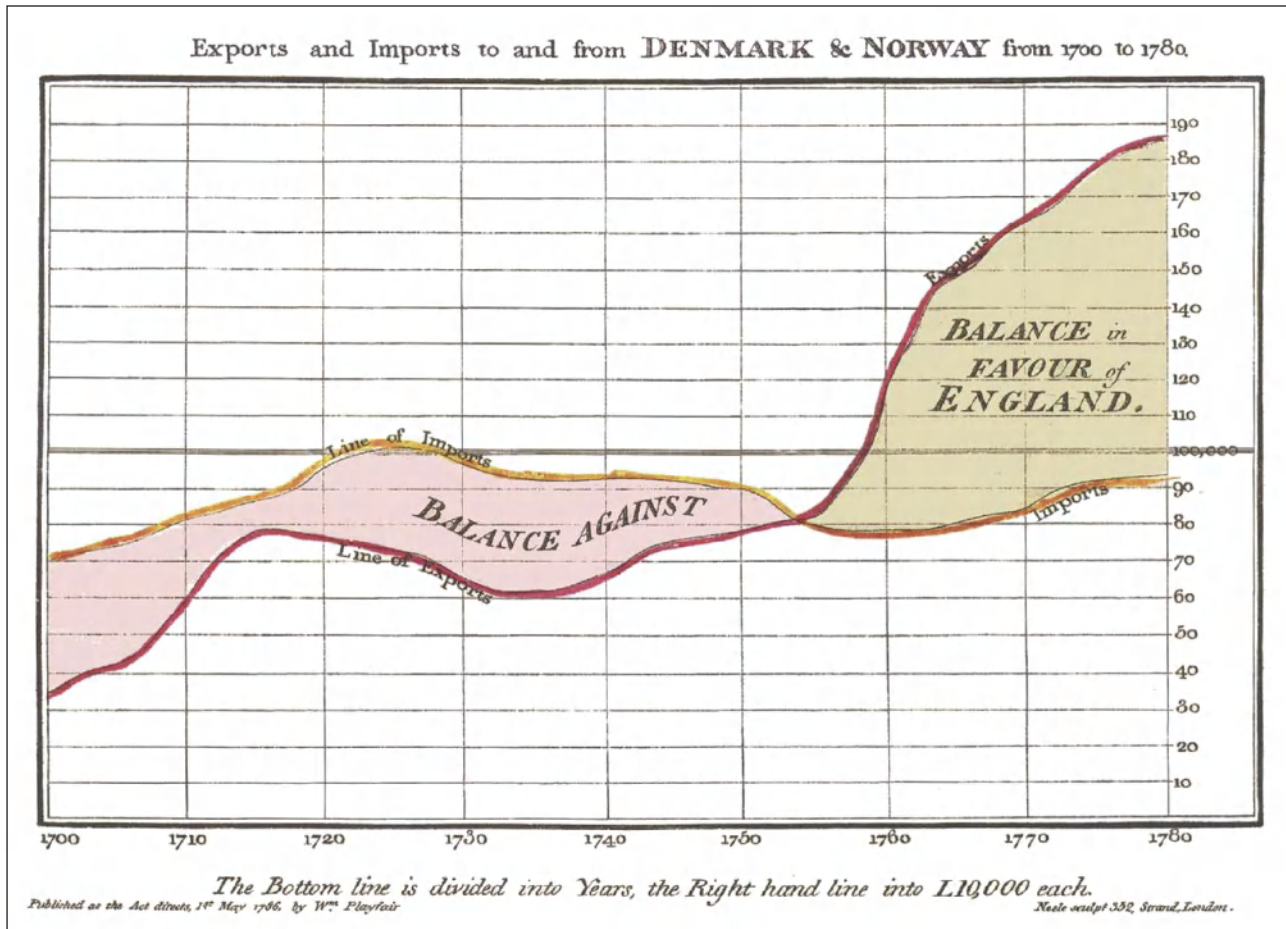


FIGURE 4.13 One of The First Time Series Line Chart (Created by William Playfair in 1801).

says that this “may well be the best statistical graphic ever drawn.” In this graphic Minard managed to simultaneously represent several data dimensions (the size of the army, direction of movement, geographic locations, outside temperature, etc.) in an artistic and informative manner. Many more excellent visualizations were created in the 1800s, and most of them are chronicled on Tufte’s Web site (edwardtufte.com) and his visualization books.

The 1900s saw the rise of a more formal, empirical attitude toward visualization, which tended to focus on aspects such as color, value scales, and labeling. In the mid-1900s, cartographer and theorist Jacques Bertin published his *Semiologie Graphique*, which some say serves as the theoretical foundation of modern information visualization. Although most of his patterns are either outdated by more recent research or completely inapplicable to digital media, many are still very relevant.

In the 2000s, the Internet emerged as a new medium for visualization and brought with it a whole lot of new tricks and capabilities. Not only has the worldwide, digital distribution of both data and visualization made them more accessible to a broader audience (raising visual literacy along the way), but it has also spurred the design of new forms that incorporate interaction, animation, and graphics-rendering technology unique to screen media, and real-time data feeds to create immersive environments for communicating and consuming data.

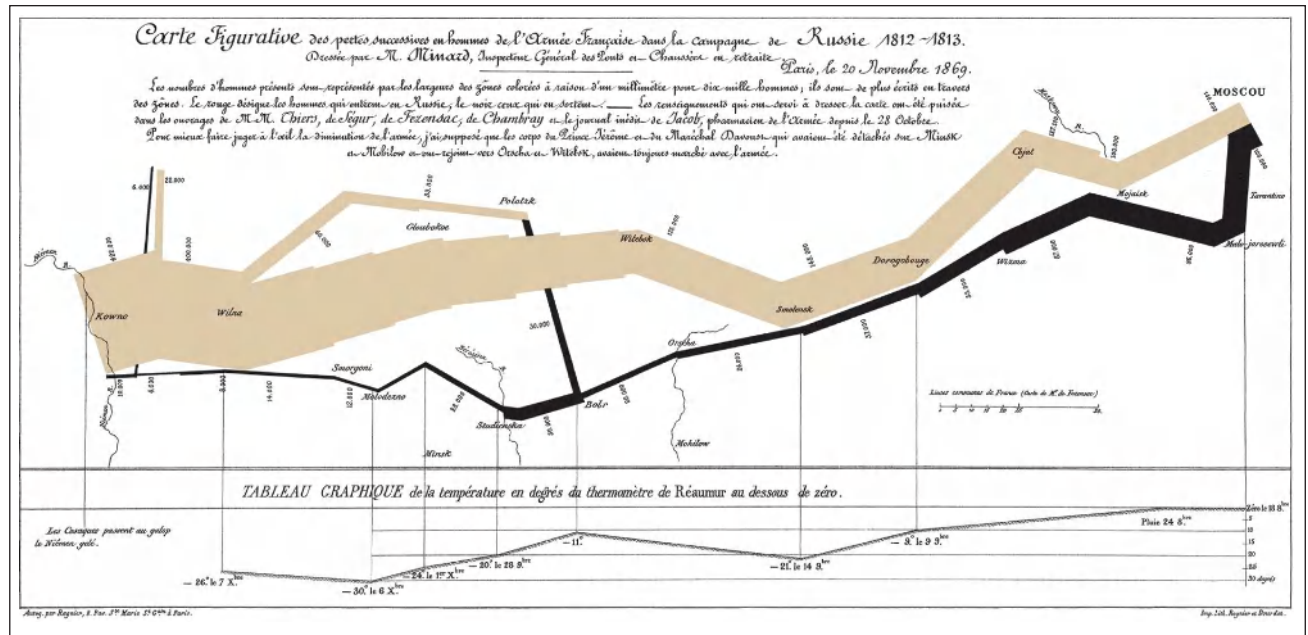


FIGURE 4.14 Decimation of Napoleon's Army during the 1812 Russian Campaign.

Companies and individuals are, seemingly all of a sudden, interested in data; that interest has in turn sparked a need for visual tools that help them understand it. Cheap hardware sensors and do-it-yourself frameworks for building your own system are driving down the costs of collecting and processing data. Countless other applications, software tools, and low-level code libraries are springing up to help people collect, organize, manipulate, visualize, and understand data from practically any source. The Internet has also served as a fantastic distribution channel for visualizations; a diverse community of designers, programmers, cartographers, tinkerers, and data wonks has assembled to disseminate all sorts of new ideas and tools for working with data in both visual and nonvisual forms.

Google Maps has also single-handedly democratized both the interface conventions (click to pan, double-click to zoom) and the technology (256-pixel square map tiles with predictable file names) for displaying interactive geography online, to the extent that most people just know what to do when they're presented with a map online. Flash has served well as a cross-browser platform on which to design and develop rich, beautiful Internet applications incorporating interactive data visualization and maps; now, new browser-native technologies such as canvas and SVG (sometimes collectively included under the umbrella of HTML5) are emerging to challenge Flash's supremacy and extend the reach of dynamic visualization interfaces to mobile devices.

The future of data/information visualization is very hard to predict. We can only extrapolate from what has already been invented: more three-dimensional visualization, more immersive experience with multidimensional data in a virtual reality environment, and holographic visualization of information. There is a pretty good chance that we will see something that we have never seen in the information visualization realm invented before the end of this decade.

► SECTION 4.8 REVIEW QUESTIONS

1. What is data visualization? Why is it needed?
2. What are the historical roots of data visualization?
3. Carefully analyze Charles Joseph Minard's graphical portrayal of Napoleon's march. Identify and comment on all the information dimensions captured in this ancient diagram.
4. Who is Edward Tufte? Why do you think we should know about his work?
5. What do you think is the "next big thing" in data visualization?

4.9 DIFFERENT TYPES OF CHARTS AND GRAPHS

Often end users of business analytics systems are not sure what type of chart or graph to use for a specific purpose. Some charts or graphs are better at answering certain types of questions. Some look better than others. Some are simple; some are rather complex and crowded. What follows is a short description of the types of charts and/or graphs commonly found in most business analytics tools and what types of questions they are better at answering/analyzing. This material is compiled from several published articles and other literature (Abela, 2008; Hardin et al., 2012).

Basic Charts and Graphs

What follows are the basic charts and graphs that are commonly used for information visualization.

LINE CHART Line charts are perhaps the most frequently used graphical visuals for time series data. Line charts (or a line graphs) show the relationship between two variables; they are most often used to track changes or trends over time (having one of the variables set to time on the x -axis). Line charts sequentially connect individual data points to help infer changing trends over a period of time. Line charts are often used to show time-dependent changes in the values of some measure, such as changes on a specific stock price over a 5-year period or changes on the number of daily customer service calls over a month.

BAR CHART Bar charts are among the most basic visuals used for data representation. Bar charts are effective when you have nominal data or numerical data that splits nicely into different categories so you can quickly see comparative results and trends within your data. Bar charts are often used to compare data across multiple categories such as percent of advertising spending by departments or by product categories. Bar charts can be vertically or horizontally oriented. They can also be stacked on top of each other to show multiple dimensions in a single chart.

PIE CHART **Pie charts** are visually appealing, as the name implies, pie-looking charts. Because they are so visually attractive, they are often incorrectly used. Pie charts should only be used to illustrate relative proportions of a specific measure. For instance, they can be used to show the relative percentage of an advertising budget spent on different product lines, or they can show relative proportions of majors declared by college students in their sophomore year. If the number of categories to show is more than just a few (say more than four), one should seriously consider using a bar chart instead of a pie chart.

SCATTER PLOT **Scatter plots** are often used to explore the relationship between two or three variables (in 2-D or 2-D visuals). Because they are visual exploration tools, having more than three variables, translating them into more than three dimensions is not easily

achievable. Scatter plots are an effective way to explore the existence of trends, concentrations, and outliers. For instance, in a two-variable (two-axis) graph, a scatter plot can be used to illustrate the corelationship between age and weight of heart disease patients or it can illustrate the relationship between the number of customer care representatives and the number of open customer service claims. Often, a trend line is superimposed on a two-dimensional scatter plot to illustrate the nature of the relationship.

BUBBLE CHART **Bubble charts** are often enhanced versions of scatter plots. Bubble charts, though, are not a new visualization type; instead, they should be viewed as a technique to enrich data illustrated in scatter plots (or even geographic maps). By varying the size and/or color of the circles, one can add additional data dimensions, offering more enriched meaning about the data. For instance, a bubble chart can be used to show a competitive view of college-level class attendance by major and by time of the day, or it can be used to show profit margin by product type and by geographic region.

Specialized Charts and Graphs

The graphs and charts that we review in this section are either derived from the basic charts as special cases or they are relatively new and are specific to a problem type and/or an application area.

HISTOGRAM Graphically speaking, a **histogram** looks just like a bar chart. The difference between histograms and generic bar charts is the information that is portrayed. Histograms are used to show the frequency distribution of a variable or several variables. In a histogram, the *x*-axis is often used to show the categories or ranges, and the *y*-axis is used to show the measures/values/frequencies. Histograms show the distributional shape of the data. That way, one can visually examine if the data is normally or exponentially distributed. For instance, one can use a histogram to illustrate the exam performance of a class, where distribution of the grades as well as comparative analysis of individual results can be shown, or one can use a histogram to show age distribution of the customer base.

GANTT CHART Gantt charts are a special case of horizontal bar charts that are used to portray project timelines, project tasks/activity durations, and overlap among the tasks/activities. By showing start and end dates/times of tasks/activities and the overlapping relationships, Gantt charts provide an invaluable aid for management and control of projects. For instance, Gantt charts are often used to show project timelines, task overlaps, relative task completions (a partial bar illustrating the completion percentage inside a bar that shows the actual task duration), resources assigned to each task, milestones, and deliverables.

PERT CHART PERT charts (also called network diagrams) are developed primarily to simplify the planning and scheduling of large and complex projects. They show precedence relationships among the project activities/tasks. A PERT chart is composed of nodes (represented as circles or rectangles) and edges (represented with directed arrows). Based on the selected PERT chart convention, either nodes or the edges may be used to represent the project activities/tasks (activity-on-node versus activity-on-arrow representation schema).

GEOGRAPHIC MAP When the data set includes any kind of location data (e.g., physical addresses, postal codes, state names or abbreviations, country names, latitude/longitude, or some type of custom geographic encoding), it is better and more informative to see the data on a map. Maps usually are used in conjunction with other charts and graphs, as opposed to by themselves. For instance, one can use maps to show distribution of

customer service requests by product type (depicted in pie charts) by geographic locations. Often a large variety of information (e.g., age distribution, income distribution, education, economic growth, or population changes) can be portrayed in a geographic map to help decide where to open a new restaurant or a new service station. These types of systems are often called geographic information systems (GIS).

BULLET Bullet graphs are often used to show progress toward a goal. A bullet graph is essentially a variation of a bar chart. Often they are used in place of gauges, meters, and thermometers in a dashboard to more intuitively convey the meaning within a much smaller space. Bullet graphs compare a primary measure (e.g., year-to-date revenue) to one or more other measures (e.g., annual revenue target) and present this in the context of defined performance metrics (e.g., sales quotas). A bullet graph can intuitively illustrate how the primary measure is performing against overall goals (e.g., how close a sales representative is to achieving his/her annual quota).

HEAT MAP Heat maps are great visuals to illustrate the comparison of continuous values across two categories using color. The goal is to help the user quickly see where the intersection of the categories is strongest and weakest in terms of numerical values of the measure being analyzed. For instance, one can use heat maps to show segmentation analysis of target markets where the measure (color gradient would be the purchase amount) and the dimensions would be age and income distribution.

HIGHLIGHT TABLE Highlight tables are intended to take heat maps one step further. In addition to showing how data intersects by using color, highlight tables add a number on top to provide additional detail. That is, they are two-dimensional tables with cells populated with numerical values and gradients of colors. For instance, one can show sales representatives' performance by product type and by sales volume.

TREE MAP Tree maps display hierarchical (tree-structured) data as a set of nested rectangles. Each branch of the tree is given a rectangle, which is then tiled with smaller rectangles representing subbranches. A leaf node's rectangle has an area proportional to a specified dimension on the data. Often the leaf nodes are colored to show a separate dimension of the data. When the color and size dimensions are correlated in some way with the tree structure, one can often easily see patterns that would be difficult to spot in other ways, such as if a certain color is particularly relevant. A second advantage of tree maps is that, by construction, they make efficient use of space. As a result, they can legibly display thousands of items on the screen simultaneously.

Which Chart or Graph Should You Use?

Which chart or graph that we explained in the previous section is the best? The answer is rather easy: there is not one best chart or graph, because if there was we would not have these many chart and graph types. They all have somewhat different data representation "skills." Therefore, the right question should be, "Which chart or graph is the best for a given task?" The capabilities of the charts given in the previous section can help in selecting and using the right chart/graph for a specific task, but it still is not easy to sort out. Several different chart/graph types can be used for the same visualization task. One rule of thumb is to select and use the simplest one from the alternatives to make it easy for the intended audience to understand and digest.

Although there is not a widely accepted, all-encompassing chart selection algorithm or chart/graph taxonomy, Figure 4.15 presents a rather comprehensive and highly logical organization of chart/graph types in a taxonomy-like structure (the original version was

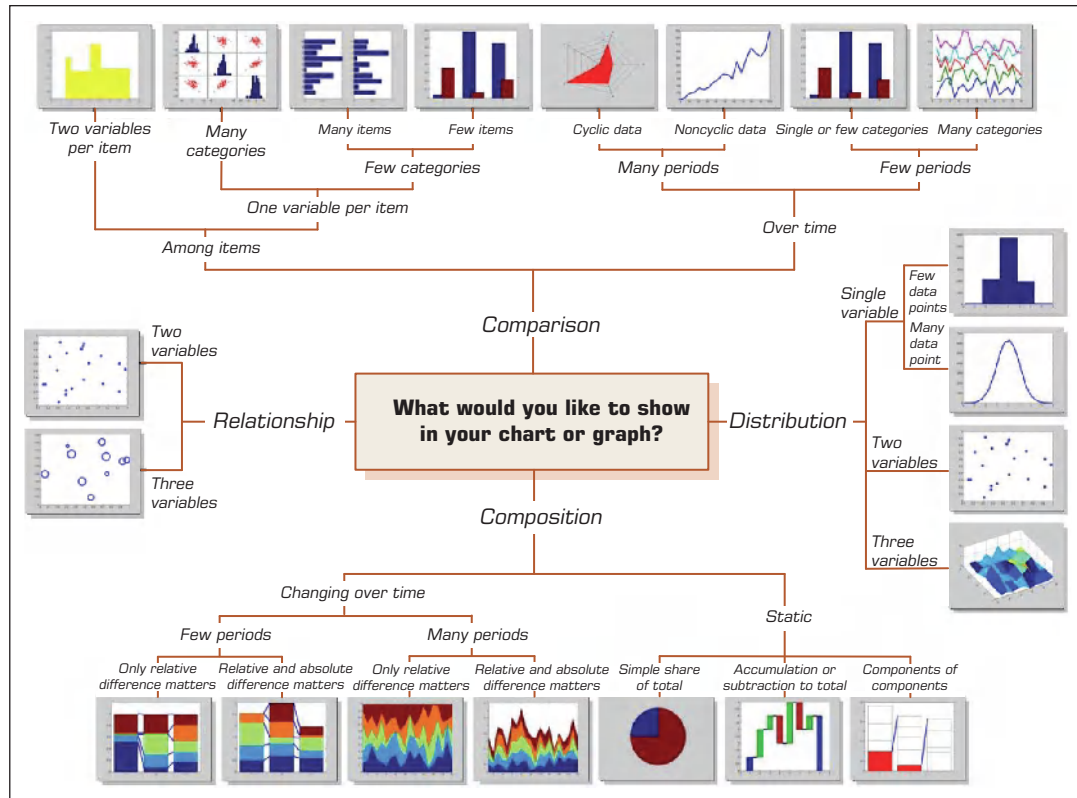


FIGURE 4.15 A Taxonomy of Charts and Graphs. Source: Adapted from Abela, A. (2008). *Advanced Presentations by Design: Creating Communication that Drives Action*. New York: Wiley.

published in Abela 2008). The taxonomic structure is organized around the questions of “What would you like to show in your chart or graph?” That is, what the purpose of the chart or graph will be. At that level, the taxonomy divides the purpose into four different types—relationship, comparison, distribution, and composition—and further divides the branches into subcategories based on the number of variables involved and time dependency of the visualization.

Even though these charts and graphs cover a major part of what is commonly used in information visualization, they by no means cover it all. Nowadays, one can find many other specialized graphs and charts that serve a specific purpose. Furthermore, the current trend is to combine/hybridize and animate these charts for better-looking and more intuitive visualization of today’s complex and volatile data sources. For instance, the interactive, animated, bubble charts available at the Gapminder Web site (gapminder.org) provide an intriguing way of exploring world health, wealth, and population data from a multidimensional perspective. Figure 4.16 depicts the sorts of displays available at the site. In this graph, population size, life expectancy, and per capita income at the continent level are shown; also given is a time-varying animation that shows how these variables change over time.

SECTION 4.9 REVIEW QUESTIONS

1. Why do you think there are many different types of charts and graphs?
2. What are the main differences among line, bar, and pie charts? When should you use one over the others?

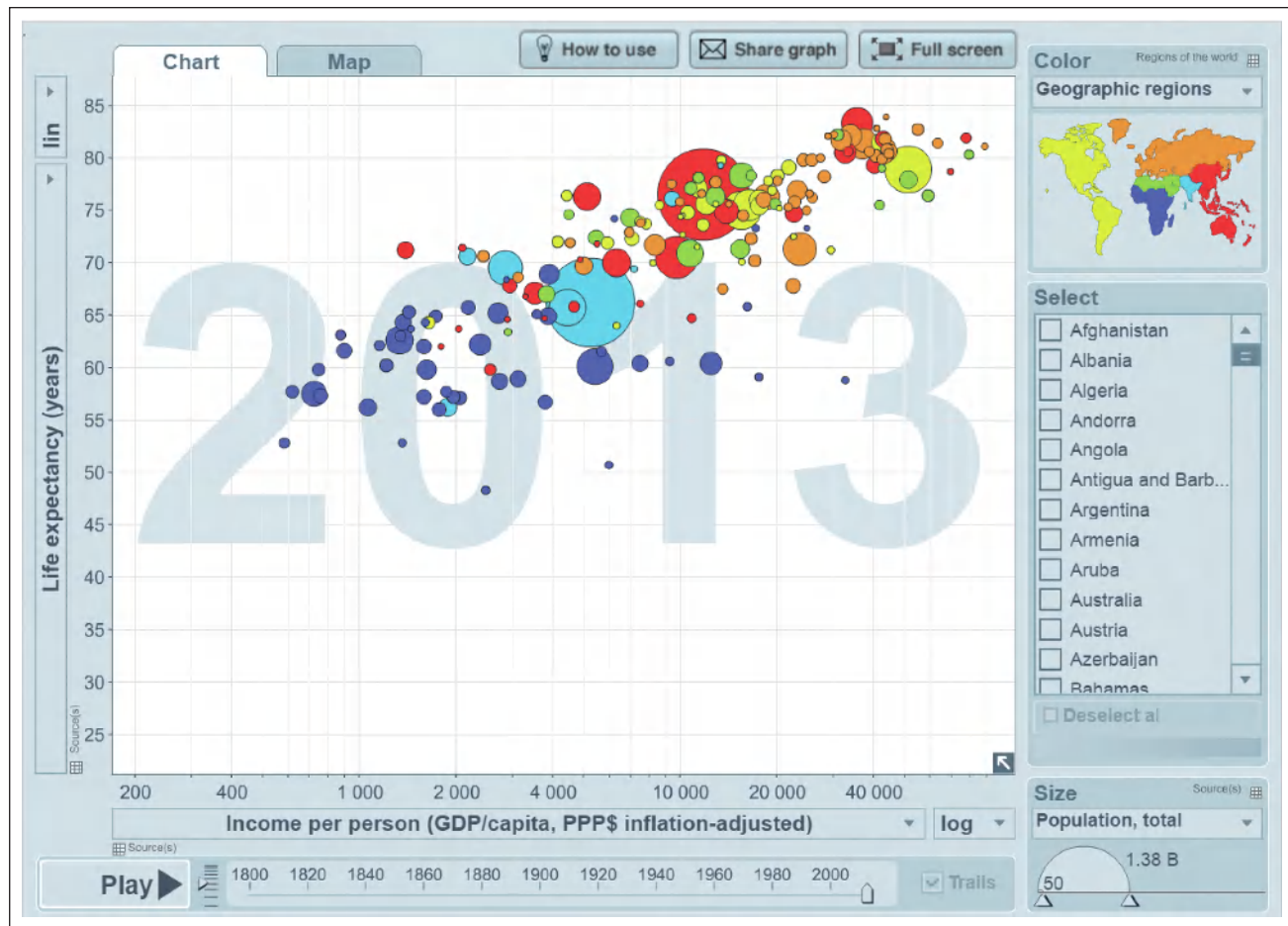


FIGURE 4.16 A Gapminder Chart that Shows the Wealth and Health of Nations. Source: gapminder.org.

3. Why would you use a geographic map? What other types of charts can be combined with a geographic map?
4. Find and explain the role of two types of charts that are not covered in this section.

4.10 THE EMERGENCE OF VISUAL ANALYTICS

As Seth Grimes (2009a,b) has noted, there is a “growing palate” of data visualization techniques and tools that enable the users of business analytics and BI systems to better “communicate relationships, add historical context, uncover hidden correlations, and tell persuasive stories that clarify and call to action.” The latest Magic Quadrant on Business Intelligence and Analytics Platforms released by Gartner in February 2016 further emphasizes the importance of data visualization in BI and analytics. As the chart shows, all the solution providers in the *Leaders* and *Visionary* quadrants are either relatively recently founded information visualization companies (e.g., Tableau Software, QlikTech) or well-established large analytics companies (e.g., Microsoft, SAS, IBM, SAP, MicroStrategy, Alteryx) that are increasingly focusing their efforts on information visualization and visual analytics. More details on Gartner’s latest Magic Quadrant are given in Technology Insights 4.3.

TECHNOLOGY INSIGHTS 4.3 Gartner Magic Quadrant for Business Intelligence and Analytics Platforms

Gartner, Inc., the creator and publisher of Magic Quadrants, is the leading information technology research and advisory company publically traded in the United States with over \$4.7 billion annual revenues in 2021. Founded in 1979, Gartner has 16,600 associates, a significant portion of which are research analysts and consultants, and numerous clients in more than 100 countries.

Magic Quadrant is a research method designed and implemented by Gartner to monitor and evaluate the progress and positions of companies in a specific, technology-based market. By applying a graphical treatment and a uniform set of evaluation criteria, Magic Quadrant helps users to understand how technology providers are positioned within a market.

Gartner changed the name of this Magic Quadrant from “Business Intelligence Platforms” to “Business Intelligence and Analytics Platforms” to emphasize the growing importance of analytics capabilities to the information systems that organizations are now building. Gartner defines the BI and analytics platform market as a software platform that delivers 15 capabilities across three categories: integration, information delivery, and analysis. These capabilities enable organizations to build precise systems of classification and measurement to support decision making and improve performance.

Figure 4.17 illustrates the latest Magic Quadrant for Business Intelligence and Analytics Platforms. Magic Quadrant places providers in four groups (niche players, challengers,



FIGURE 4.17 Magic Quadrant for Business Intelligence and Analytics Platforms. Source: gartner.com.

visionaries, and leaders) along two dimensions: completeness of vision (x -axis) and ability to execute (y -axis). As the quadrant clearly shows, most of the well-known data visualization, visual analytics, BI/BA providers are positioned in the “leaders” category while many of the lesser known, relatively new, emerging providers are positioned in the “niche players” category.

The BI and analytics platform market’s multiyear shift from IT-led enterprise reporting to business-led self-service analytics seems to have passed the tipping point. Most new buying is of modern, business-user-centric visual analytics platforms forcing a new market perspective, significantly reordering the vendor landscape. Most of the activity in the BI and analytics platform market is from organizations that are trying to mature their visualization capabilities and to move from descriptive to predictive and prescriptive analytics echelons. The vendors in the market have overwhelmingly concentrated on meeting this user demand. If there were a single market theme in 2022, it would be that data discovery/visualization became a mainstream architecture. While data discovery/visualization vendors such as Microsoft (with PowerBI), Salesforce (with Tableau), and Qlik are solidifying their position in the *Leaders* quadrant, others (both emerging and large, well-established tool/solution providers) are trying to move out of *Visionaries* into the *Leaders* quadrant.

This emphasis on data discovery/visualization from most of the leaders and visionaries in the market—which are now promoting tools with business-user-friendly data integration, coupled with embedded storage and computing layers and unfettered drilling—continues to accelerate the trend toward decentralization and user empowerment of BI and analytics and greatly enables organizations’ ability to perform diagnostic analytics.

Source: Gartner Magic Quadrant, released in February 2022, gartner.com (accessed June 2022).

In BI and analytics, the key challenges for visualization have revolved around the intuitive representation of large, complex data sets with multiple dimensions and measures. For the most part, the typical charts, graphs, and other visual elements used in these applications usually involve two dimensions, sometimes three, and fairly small subsets of data sets. In contrast, the data in these systems reside in a data warehouse. At a minimum, these warehouses involve a range of dimensions (e.g., product, location, organizational structure, time), a range of measures, and millions of cells of data. In an effort to address these challenges, a number of researchers have developed a variety of new visualization techniques.

Visual Analytics

Visual analytics is a recently coined term that is often used loosely to mean nothing more than information visualization. What is meant by **visual analytics** is the combination of visualization and predictive analytics. Whereas information visualization is aimed at answering, “What happened?” and “What is happening?” and is closely associated with BI (routine reports, scorecards, and dashboards), visual analytics is aimed at answering, “Why is it happening?” “What is more likely to happen?” and is usually associated with business analytics (forecasting, segmentation, correlation analysis). Many of the information visualization vendors are adding the capabilities to call themselves visual analytics solution providers. One of the top, long time analytics solution providers, SAS Institute, is approaching it from another direction. They are embedding their analytics capabilities into a high-performance data visualization environment that they call visual analytics.

Visual or not visual, automated or manual, online or paper based, business reporting is not much different than telling a story. Technology Insights 4.4 provides a different, unorthodox viewpoint to better business reporting.

TECHNOLOGY INSIGHTS 4.4 Telling Great Stories with Data and Visualization

Everyone who has data to analyze has stories to tell, whether it's diagnosing the reasons for manufacturing defects, selling a new idea in a way that captures the imagination of your target audience, or informing colleagues about a particular customer service improvement program. And when it's telling the story behind a big strategic choice so that you and your senior management team can make a solid decision, providing a fact-based story can be especially challenging. In all cases, it's a big job. You want to be interesting and memorable; you know you need to keep it simple for your busy executives and colleagues. Yet you also know you have to be factual, detail oriented, and data driven, especially in today's metric-centric world.

It's tempting to present just the data and facts, but when colleagues and senior management are overwhelmed by data and facts without context, you lose. We have all experienced presentations with large slide decks, only to find that the audience is so overwhelmed with data that they don't know what to think, or they are so completely tuned out that they take away only a fraction of the key points.

Start engaging your executive team and explaining your strategies and results more powerfully by approaching your assignment as a story. You will need the "what" of your story (the facts and data) but you also need the "Who?" "How?" "Why?" and the often-missed "So what?" It's these story elements that will make your data relevant and tangible for your audience. Creating a good story can aid you and senior management in focusing on what is important.

Why Story?

Stories bring life to data and facts. They can help you make sense and order out of a disparate collection of facts. They make it easier to remember key points and can paint a vivid picture of what the future can look like. Stories also create interactivity—people put themselves into stories and can relate to the situation.

Cultures have long used **storytelling** to pass on knowledge and content. In some cultures, storytelling is critical to their identity. Among the many examples from disparate parts of the world, this one is from New Zealand, where some of the Maori people tattoo their faces with *mokus*. A *moku* is a facial tattoo containing a story about ancestors—the family tribe. A man may have a tattoo design on his face that shows features of a hammerhead to highlight unique qualities about his lineage. The design he chooses signifies what is part of his "true self" and his ancestral home.

Likewise, when we are trying to understand a story, the storyteller navigates to finding the "true north." If senior management is looking to discuss how they will respond to a competitive change, a good story can make sense and order out of a lot of noise. For example, you may have facts and data from two studies, one including results from an advertising study and one from a product satisfaction study. Developing a story for what you measured across both studies can help people see the whole where there were disparate parts. For rallying your distributors around a new product, you can employ a story to give vision to what the future can look like. Most important, storytelling is interactive—typically the presenter uses words and pictures that audience members can put themselves into. As a result, they become more engaged and better understand the information.

So What Is a Good Story?

Most people can easily rattle off their favorite film or book. Or they remember a funny story that a colleague recently shared. Why do people remember these stories? Because they contain certain characteristics. First, a good story has great characters. In some cases, the reader or viewer has a vicarious experience where they become involved with the character. The character then has to be faced with a challenge that is difficult but believable. There must be hurdles that the character overcomes. And finally, the outcome or prognosis is clear by the end of the story. The situation may not be resolved—but the story has a clear endpoint.

Think of Your Analysis as a Story—Use a Story Structure

When crafting a data-rich story, the first objective is to find the story. Who are the characters? What is the drama or challenge? What hurdles have to be overcome? And at the end of your story, what do you want your audience to do as a result?

Once you know the core story, craft your other story elements: define your characters, understand the challenge, identify the hurdles, and crystallize the outcome or decision question. Make sure you are clear with what you want people to do as a result. This will shape how your audience will recall your story. With the story elements in place, write out the storyboard, which represents the structure and form of your story. Although it's tempting to skip this step, it is better first to understand the story you are telling and then to focus on the presentation structure and form. Once the storyboard is in place, the other elements will fall into place. The storyboard will help you to think about the best analogies or metaphors, to clearly set up challenge or opportunity, and to finally see the flow and transitions needed. The storyboard also helps you focus on key visuals (graphs, charts, and graphics) that you need your executives to recall. Figure 2.24 shows a storyline for the impact of small loans in a worldwide view within the Tableau visual analytics environment.

In summary, don't be afraid to use data to tell great stories. Being factual, detail oriented, and data driven is critical in today's metric-centric world, but it does not have to mean being boring and lengthy. In fact, by finding the real stories in your data and following the best practices, you can get people to focus on your message—and thus on what's important. Here are those best practices:

- Think of your analysis as a story—use a story structure.
- Be authentic—your story will flow.
- Be visual—think of yourself as a film editor.
- Make it easy for your audience and you.
- Invite and direct discussion.

Sources: Fink, E., & Moore, S. J. (2022). 5 Best Practices for Telling Great Data Stories—And Why It Will Make You a Better Analyst. White paper by Tableau Software, Inc., <https://www.tableau.com/whitepapers/telling-data-stories> (accessed June 2022); Knaflitz, C. N. (2015). *Storytelling with Data: A Data Visualization Guide for Business Professionals*. John Wiley & Sons.

High-Powered Visual Analytics Environments

Due to the increasing demand for visual analytics coupled with fast-growing data volumes, there is an exponential movement toward investing in highly efficient visualization systems. With their latest move into visual analytics, the statistical software giant SAS Institute is now among those who are leading this wave. Their new product, SAS Visual Analytics, is a very **high-performance computing**, in-memory solution for exploring massive amounts of data in a very short time (almost instantaneously). It empowers users to spot patterns, identify opportunities for further analysis, and convey visual results via Web reports or a mobile platform such as tablets and smartphones. Figure 4.18 shows the high-level architecture of the SAS Visual Analytics platform. On one end of the architecture, there is a universal data builder and administrator capabilities, leading into explorer, report designer, and mobile BI modules, collectively providing an end-to-end visual analytics solution.

Some of the key benefits proposed by SAS analytics are the following:

- Empowers all users with data exploration techniques and approachable analytics to drive improved decision making. SAS Visual Analytics enables different types of users to conduct fast, thorough explorations on all available data. Sampling to reduce the data is not required and not preferred.
- Easy-to-use, interactive Web interfaces broaden the audience for analytics, enabling everyone to glean new insights. Users can look at more options, make more precise decisions, and drive success even faster than before.
- Answer complex questions faster, enhancing the contributions from your analytic talent. SAS Visual Analytics augments the data discovery and exploration process by

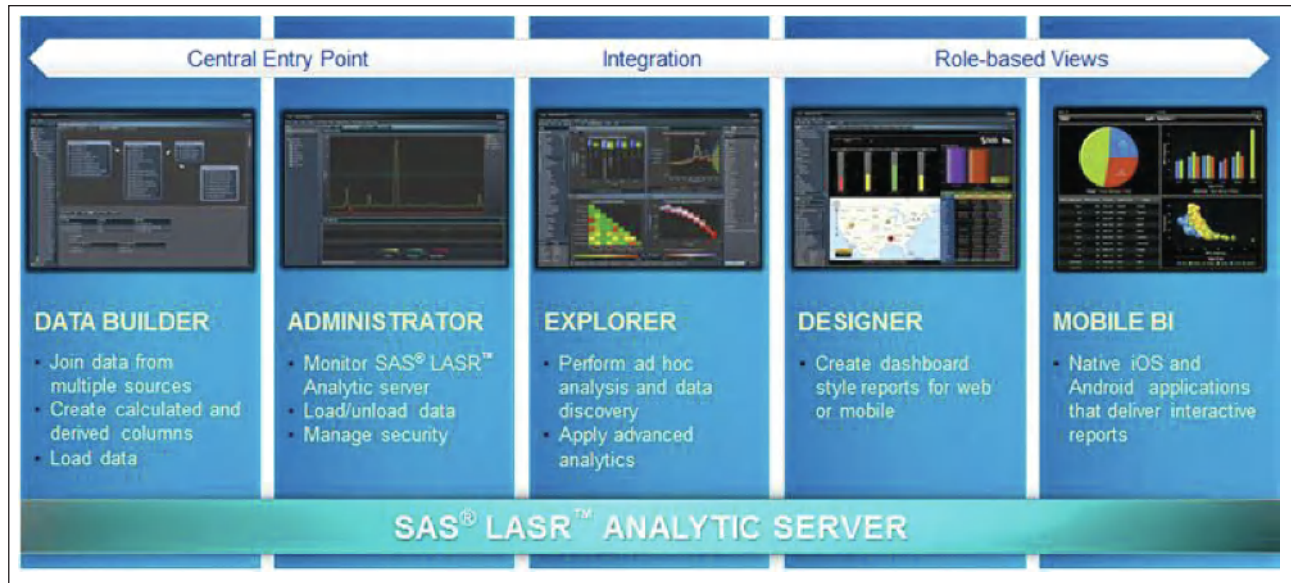


FIGURE 4.18 An Overview of SAS Visual Analytics Architecture. Source: SAS.com.

providing extremely fast results to enable better, more focused analysis. Analytically savvy users can identify areas of opportunity or concern from vast amounts of data so further investigation can take place quickly.

- Improves information sharing and collaboration. Large numbers of users, including those with limited analytical skills, can quickly view and interact with reports and charts via the Web, Adobe PDF files, and iPad mobile devices, while IT maintains control of the underlying data and security. SAS Visual Analytics provides the right information to the right person at the right time to improve productivity and organizational knowledge.
- Liberates IT by giving users a new way to access the information they need. Frees IT from the constant barrage of demands from users who need access to different amounts of data, different data views, ad hoc reports, and one-off requests for information. SAS Visual Analytics enables IT to easily load and prepare data for multiple users. Once data is loaded and available, users can dynamically explore data, create reports, and share information on their own.
- Provides room to grow at a self-determined pace. SAS Visual Analytics provides the option of using commodity hardware or database appliances from EMC Greenplum and Teradata. It is designed from the ground up for performance optimization and scalability to meet the needs of any size organization.

Figure 4.19 shows a screenshot of an SAS Analytics platform where time series forecasting and confidence interval around the forecast are depicted.

► SECTION 4.10 REVIEW QUESTIONS

1. What are the main reasons for the recent emergence of visual analytics?
2. Look at Gartner's Magic Quadrant for Business Intelligence and Analytics Platforms. What do you see? Discuss and justify your observations.
3. What is the difference between information visualization and visual analytics?
4. Why should storytelling be a part of your reporting and data visualization?
5. What is a high-powered visual analytics environment? Why do we need it?

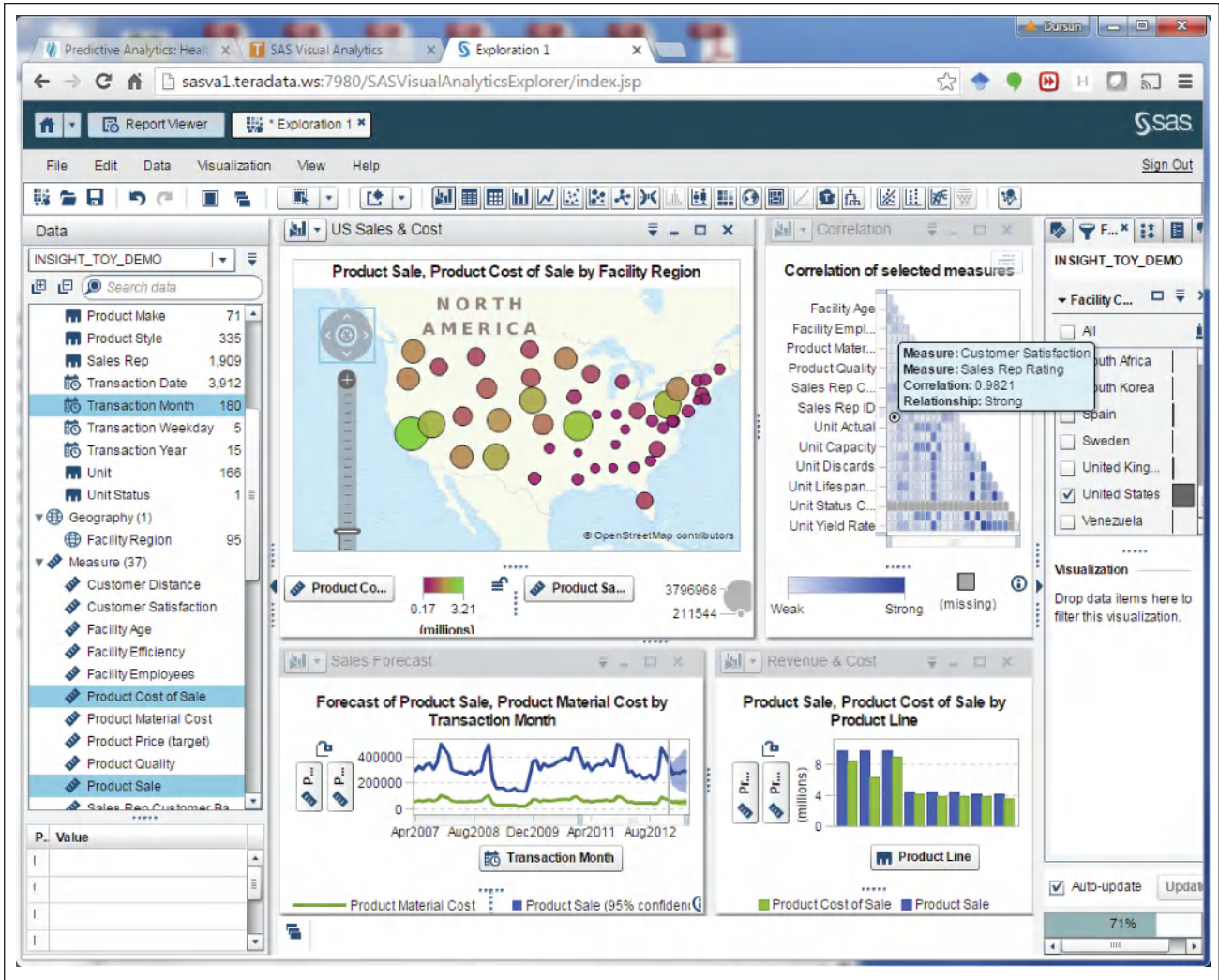


FIGURE 4.19 A Screenshot from SAS Visual Analytics. Source: SAS.com.

4.11 INFORMATION DASHBOARDS

Information dashboards are common components of most, if not all, BI or business analytics platforms, business performance management systems, and performance measurement software suites. **Dashboards** provide visual displays of important information that is consolidated and arranged on a single screen so that information can be digested at a single glance and easily drilled in and further explored. A typical dashboard is shown in Figure 4.20. This particular executive dashboard displays a variety of KPIs for a hypothetical software company called Sonatica (selling audio tools). This executive dashboard shows a high-level view of the different functional groups surrounding the products, starting from a general overview to the marketing efforts, sales, finance, and support departments. All of this is intended to give executive decision makers a quick and accurate idea of what is going on within the organization. On the left side of the dashboard, we can see (in a time series fashion) the quarterly changes in revenues, expenses, and margins, as well as the comparison of those figures to previous years'



FIGURE 4.20 A Sample Executive Dashboard. Source: dundas.com.

monthly numbers. On the upper-right side we see two dials with color-coded regions showing the amount of monthly expenses for support services (dial on the left) and the amount of other expenses (dial on the right). As the color coding indicates, although the monthly support expenses are well within the normal ranges, the other expenses are in the red region, indicating excessive values. The geographic map on the bottom right shows the distribution of sales at the country level throughout the world. Behind these graphical icons, there are a variety of mathematical functions aggregating numerous data points to their highest level of meaningful figures. By clicking on these graphical icons, the consumer of this information can drill down to more granular levels of information and data.

Nowadays, a significant portion of the digitized data for business reports and dashboards come from the Internet, especially from social media. In Analytics in Action 4.3, you will find the summary of a successful implementation of data blending, efficient reports, and better informed decisioning.

ANALYTICS IN ACTION 4.3 Increasing the Efficiency of Social Media Campaign Reporting to Get to Insights Quicker

Motivation

Sport England is an arms-length body of government responsible for growing and developing grassroots sport and getting more people active across England. They're working to keep movement, sport, and physical activity central to the lives of everyone.

Sport England run cross-platform media campaigns to encourage participation in sport and physical activity. Reporting, however, is always a manual task. Sport England wanted to increase the efficiency of this process to not only generate insights more quickly, but also reduce unnecessary manual work.

Previously, the process required logging in to each of the social media platforms, collecting statistics from accounts, transforming the data so it was comparable, before visualizing it in Excel and producing a report in PowerPoint. This was a labor-intensive process, which could only be done by one person and created a significant lag in reporting. The delay in reporting meant that while a view of overall success was available to senior management, information was not available for operational decisions where a nimble response to events might be required. Off the shelf solutions were considered but could not provide the targeted approach or custom visuals required by Sport England for their "This Girl Can" campaign.

Implementation (Blending Technologies for an Automated, Bespoke, and Branded Reporting Solution)

Atos (an international IT services company and a trusted KNIME partner) created workflows in KNIME Analytics Platform to collect, transform, join, and write data from social media platforms to an SQL database on Azure, where it was visualized in Power BI. KNIME Server was used to schedule data collection and an analytical application was made available via the KNIME WebPortal to allow a supplementary data file to be uploaded in a user-friendly way by their admin team.

Atos used functionality already built into KNIME to connect to Twitter, Python nodes to connect to, navigate, and retrieve data from Facebook and Instagram graph databases, as well as Web-scraping for data from Medium. Python nodes were also used to query Google and YouTube APIs. Data was transformed and joined in KNIME to provide a comparable snapshot of activity over the previous 24 hours when the collected data ranged from cumulative counts to six-week windows.

The workflows were scheduled to run shortly after midnight on KNIME Server, which was running in a virtual machine on Sport England's Azure platform. Data was saved to an SQL database where it was available to Power BI (using the KNIME PowerBI Integration) for visualization.

Initially Azure Cognitive Services were to perform sentiment analysis on tweets in Power BI, but this functionality was folded back into KNIME. Therefore, a tweet was only ever passed to Cognitive Services once, and the score was stored in the database for cost and efficiency savings.

Atos created custom metrics for rating features such as a user's influence, which could be explained, understood, and adapted as necessary. These were then visualized to Sport England's specification in adherence with their internal style guides and branding to make it their tool. This highlighted that the data that was important to them and the success of their campaign.

Results (Automatic Data Processing and Increased Reporting Frequency)

The frequency of reporting was increased from monthly to daily. Instead of relying on a single person to generate the report, which would take half a day, the latest data was collected and processed automatically and made available via a Power BI report at the beginning of the day. The Power BI report was more useful than the PowerPoint report because it allowed users to apply multiple filters to find the data they needed—instead of relying on the analyst to anticipate requirements and generate insight.

Why KNIME?

KNIME Analytics Platform makes it easy to seamlessly blend core functionality for data manipulation with highly complex and bespoke operations for data collection and cleaning using R and Python. The visual workflow builder enables workflow creators to explain the process to the media team and others outside data specialist roles who don't necessarily have the technical data science knowledge, drawing parallels to how data was treated in the manual process. This means the solution is less of a "black box," but rather something that those who use the data can have confidence in because they understand how the parts fit together.

KNIME Server makes scheduling data collection a simple task and provides feedback that the workflows are executed without errors. Additional benefits include the ability to work remotely in the cloud, and the ability for Sport England to easily manage the data governance to sandbox this project from their other data.

Source: KNIME (2023). "Combining the Power of KNIME and PowerBI for Automated Sentiment Analysis" can be found at <https://www.knime.com/solutions/success-story/automated-sentiment-analysis> (accessed May 2023).

Dashboard Design

Dashboards are not a new concept. Their roots can be traced at least to the executive information system of the 1980s. Today, dashboards are ubiquitous. For example, a few years back, Forrester Research estimated that over 40% of the largest 2,000 companies in the world used the technology (Ante & McGregor, 2006). Since then, one can safely assume that this number has gone up quite significantly. In fact, nowadays it would be rather unusual to see a large company using a BI system that does not employ some sort of performance dashboards. The Dashboard Spy Web site (dashboardspy.com/about) provides further evidence of their ubiquity. The site contains descriptions and screenshots of thousands of BI dashboards, scorecards, and BI interfaces used by businesses of all sizes and industries, nonprofits, and government agencies.

According to Eckerson (2006), a well-known expert on BI in general and dashboards in particular, the most distinctive feature of a dashboard is its three layers of information:

- **Monitoring:** Graphical, abstracted data to monitor key performance metrics.
- **Analysis:** Summarized dimensional data to analyze the root cause of problems.
- **Management:** Detailed operational data that identify what actions to take to resolve a problem.

Because of these layers, dashboards pack a lot of information into a single screen. According to Few (2005), "The fundamental challenge of dashboard design is to display all the required information on a single screen, clearly and without distraction, in a manner that can be assimilated quickly." To speed assimilation of the numbers, the numbers need to be placed in context. This can be done by comparing the numbers of interest to other baseline or target numbers, by indicating whether the numbers are good or bad, by denoting whether a trend is better or worse, and by using specialized display widgets or components to set the comparative and evaluative context. Some of the common comparisons that are typically made in BI systems include comparisons against past values, forecasted values, targeted values, benchmark or average values, multiple instances of the same measure, and the values of other measures (e.g., revenues versus costs).

Even with comparative measures, it is important to specifically point out whether a particular number is good or bad and whether it is trending in the right direction. Without

these types of evaluative designations, it can be time consuming to determine the status of a particular number or result. Typically, either specialized visual objects (e.g., traffic lights, dials, and gauges) or visual attributes (e.g., color coding) are used to set the evaluative context.

What to Look for in a Dashboard

Although performance dashboards and other information visualization frameworks differ, they all do share some common design characteristics. First, they all fit within the larger BI and/or performance measurement system. This means that their underlying architecture is the BI or performance management architecture of the larger system. Second, all well-designed dashboard and other information visualizations possess the following characteristics (Novell, 2009):

- They use visual components (e.g., charts, performance bars, sparklines, gauges, meters, stoplights) to highlight, at a glance, the data and exceptions that require action.
- They are transparent to the user, meaning that they require minimal training and are extremely easy to use.
- They combine data from a variety of systems into a single, summarized, unified view of the business.
- They enable drill-down or drill-through to underlying data sources or reports, providing more detail about the underlying comparative and evaluative context.
- They present a dynamic, real-world view with timely data refreshes, enabling the end user to stay up to date with any recent changes in the business.
- They require little, if any, customized coding to implement, deploy, and maintain.

Best Practices in Dashboard Design

The real estate saying “location, location, location” makes it obvious that the most important attribute for a piece of real estate property is where it is located. For dashboards, it is “data, data, data.” An often overlooked aspect, data is one of the most important things to consider in designing dashboards. Even if a dashboard’s appearance looks professional, is aesthetically pleasing, and includes graphs and tables created according to accepted visual design standards, it is also important to ask about the data: Is it reliable? Is it timely? Is any data missing? Is it consistent across all dashboards? Here are some of the experience-driven best practices in dashboard design (Radha, 2008).

Benchmark Key Performance Indicators with Industry Standards

Many customers, at some point in time, want to know if the metrics they are measuring are the right metrics to monitor. Sometimes customers have found that the metrics they are tracking are not the right ones to track. Doing a gap assessment with industry benchmarks aligns you with industry best practices.

Wrap the Dashboard Metrics with Contextual Metadata

Often when a report or a visual dashboard/scorecard is presented to business users, questions remain unanswered. The following are some examples:

- Where did you source this data from?
- While loading the data warehouse, what percentage of the data got rejected/encountered data quality problems?

- Is the dashboard presenting “fresh” information or “stale” information?
- When was the data warehouse last refreshed?
- When is it going to be refreshed next?
- Were any high-value transactions that would skew the overall trends rejected as a part of the loading process?

Validate the Dashboard Design by a Usability Specialist

In most dashboard environments, the dashboard is designed by a tool specialist without giving consideration to usability principles. Even though it's a well-engineered data warehouse that can perform well, many business users do not use the dashboard, as it is perceived as not being user friendly, leading to poor adoption of the infrastructure and change management issues. Up-front validation of the dashboard design by a usability specialist can mitigate this risk.

Prioritize and Rank Alerts/Exceptions Streamed to the Dashboard

Because there are tons of raw data, it is important to have a mechanism by which important exceptions/behaviors are proactively pushed to the information consumers. A business rule can be codified, which detects the alert pattern of interest. It can be coded into a program, using database-stored procedures, which can crawl through the fact tables and detect patterns that need immediate attention. This way, information finds the business user as opposed to the business user polling the fact tables for the occurrence of critical patterns.

Enrich the Dashboard with Business-User Comments

When the same dashboard information is presented to multiple business users, a small text box can be provided that can capture the comments from an end-user's perspective. This can often be tagged to the dashboard to put the information in context, adding perspective to the structured KPIs being rendered.

Present Information in Three Different Levels

Information can be presented in three layers depending on the granularity of the information: the visual dashboard level, the static report level, and the self-service cube level. When a user navigates the dashboard, a simple set of 8 to 12 KPIs can be presented, which would give a sense of what is going well and what is not.

Pick the Right Visual Construct Using Dashboard Design Principles

In presenting information in a dashboard, some information is presented best with bar charts, some with time series line graphs, and when presenting correlations, a scatter plot is useful. Sometimes merely rendering it as simple tables is effective. Once the dashboard design principles are explicitly documented, all the developers working on the front end can adhere to the same principles while rendering the reports and dashboard.

Provide for Guided Analytics

In a typical organization, business users can be at various levels of analytical maturity. The capability of the dashboard can be used to guide the “average” business user to access the same navigational path as that of an analytically savvy business user.

SECTION 4.11 REVIEW QUESTIONS

- 1. What is an information dashboard? Why are they so popular?
- 2. What are the graphical widgets commonly used in dashboards? Why?
- 3. List and describe the three layers of information portrayed on dashboards.
- 4. What are the common characteristics of dashboards and other information visuals?

Chapter Highlights

- A data warehouse is a specially constructed data repository where data are organized so that they can be easily accessed by end users for several applications.
- DMs contain data on one topic (e.g., marketing). A DM can be a replication of a subset of data in the data warehouse. DMs are a less-expensive solution that can be replaced by or can supplement a data warehouse. DMs can be independent of or dependent on a data warehouse.
- An ODS is a type of customer-information-file database that is often used as a staging area for a data warehouse.
- Data integration comprises three major processes: data access, data federation, and change capture. When these three processes are correctly implemented, data can be accessed and made accessible to an array of ETL and analysis tools and data warehousing environments.
- ETL technologies pull data from many sources, cleanse them, and load them into a data warehouse. ETL is an integral process in any data-centric project.
- Real-time or active data warehousing supplements and expands traditional data warehousing, moving into the realm of operational and tactical decision making by loading data in real time and providing data to users for active decision making.
- The security and privacy of data and information are critical issues for a data warehouse professional.
- A report is any communication artifact prepared with the specific intention of conveying information in a presentable form.
- A business report is a written document that contains information regarding business matters.
- The key to any successful business report is clarity, brevity, completeness, and correctness.
- Data visualization is the use of visual representations to explore, make sense of, and communicate data.
- Perhaps the most notable information graphic of the past was developed by Charles J. Minard, who graphically portrayed the losses suffered by Napoleon’s army in the Russian campaign of 1812.
- Basic chart types include line, bar, and pie chart.
- Specialized charts are often derived from the basic charts as exceptional cases.
- Data visualization techniques and tools make the users of business analytics and BI systems better information consumers.
- Visual analytics is the combination of visualization and predictive analytics.
- Increasing demand for visual analytics coupled with fast-growing data volumes led to exponential growth in highly efficient visualization systems investment.
- Dashboards provide visual displays of important information that is consolidated and arranged on a single screen so that information can be digested at a single glance and easily drilled in and further explored.

Key Terms

bubble chart	data integration	data warehouse	dimensional modeling
business report	data mart (DM)	administrator (DWA)	drill down
cube	data visualization	dependent data mart	enterprise application
dashboards	data warehouse (DW)	dimension table	integration (EAI)

enterprise data warehouse (EDW)	high-performance computing histogram	OLTP	report
enterprise information integration (EII)	independent data mart	oper mart	scatter plot
extraction, transformation, and load (ETL)	key performance indicator (KPI) metadata	operational data store (ODS)	snowflake schema
		pie chart	star schema
		real-time data warehousing (RDW)	storytelling
			visual analytics

Questions for Discussion

1. Compare data integration and ETL. How are they related?
2. What is a data warehouse, and what are its benefits? Why is Web accessibility important with a data warehouse?
3. A data mart can replace a data warehouse or complement it. Compare and discuss these options.
4. Discuss the major drivers and benefits of data warehousing to end users.
5. List the differences and/or similarities between the roles of a database administrator and a data warehouse administrator.
6. Describe how data integration can lead to higher levels of data quality.
7. Compare the Kimball and Inmon approaches toward data warehouse development. Identify when each one is most effective.
8. Discuss security concerns involved in building a data warehouse.
9. Investigate current data warehouse development implementation through offshoring. Write a report about it. In class, debate the issue in terms of the benefits and costs, as well as social factors.
10. What is a business report? Why is it needed?
11. What are the best practices in business reporting? How can we make our reports stand out?
12. Describe the cyclic process of management, and comment on the role of business reports.
13. List and describe the three major categories of business reports.
14. Why has information visualization become a centerpiece in BI and business analytics? Is there a difference between information visualization and visual analytics?
15. What are the main types of charts/graphs? Why are there so many of them?
16. How do you determine the right chart for the job? Explain and defend your reasoning.
17. What is the difference between information visualization and visual analytics?
18. Why should storytelling be a part of your reporting and data visualization?
19. What is an information dashboard? What do they present?
20. What are the best practices in designing highly informative dashboards?
21. Do you think information/performance dashboards are here to stay? Or are they about to be outdated? What do you think will be the next big wave in BI and business analytics in terms of data/information visualization?

Exercises

Hands-On Exercises

1. Consider the case of developing a data warehouse for Coca-Cola Japan available at the DSS Resources Web site, <http://dssresources.com/cases/coca-colajapan>. Read the case, and answer the nine questions for further analysis and discussion.
2. Explain when you should implement a two- or three-tiered architecture when considering developing a data warehouse.
3. Read the full Continental Airlines case (a hugely popular data warehousing success story) at the book Web site and answer the case questions.
4. Download the seminal case study “Harrah’s High Payoff from Customer Information” from the book Web site, read and answer the questions at the end of the case. Relate Harrah’s results to how other service-oriented enterprises use their customer data for better decision making.
5. Consider the future of data warehousing in the age of big data. Perform a Web search on this topic. In addition to magazine articles and blog posts you will find, make sure to search on scholar.google.com for academic papers on the topic. Use at least three academic papers and three non-academic articles for your report.

Summarize, compare and contrast your understanding and findings.

6. Go to YouTube.com, and search for “Teradata BSI Cases,” where BSI stands for “Business Solutions Inc.” Select three interesting data warehousing cases, watch them carefully, and write a report to discuss your findings about the business problems and proposed investigative solutions.
7. Download a data/information visualization tool, such as Tableau, PowerBI, QlikView, or Spotfire. If your school does not have an educational agreement with these companies, then a trial version would be sufficient for this exercise. Use your own data (if you have any) or use one of the data sets that comes with the tool (they usually have one or more data sets for demonstration purposes). Study the data, come up with a couple of business problems, and use data visualization to analyze, visualize, and potentially solve those problems.
8. Find at least three articles (one journal article and two white papers) that talk about storytelling, especially within the context of analytics (i.e., data-driven storytelling). Read and critically analyze the article and paper, and write a report to reflect your understanding and opinions about the importance of storytelling in BI and business analytics.
9. Go to Data.gov—a U.S. government-sponsored data portal that has a very large number of data sets on a wide variety of topics ranging from healthcare to education, climate to public safety. Pick a topic that you are most passionate about. Go through the topic-specific information and explanation provided on the site. Explore the possibilities of downloading the data, and use your favorite data visualization tool to create your own meaningful information and visualizations.
10. Go To SAS Viya for Learners site (https://www.sas.com/en_us/software/viya-for-learners.html). Create a SAS profile for yourself. Follow the “Access for Students” link to gain entry to the cloud-based analytics platform. Using the existing data sets in this educational platform (a good data set is called INSIGHT-TOY_DEMO), create a visual analytics report using the “Explore and Visualize” functionality. Once you are done, export/print your report into a pdf file (e.g., YourFirstName_LastName_HW1.pdf) from within the platform and submit it to your instructor for evaluation.
11. Go to Stephen Few’s blog, “The Perceptual Edge” (perceptualedge.com). Go to the section of “Examples.” In this section, he provides critiques of various dashboard examples. Read a handful of these examples. Now go to dundas.com. Select the “Gallery” section of the site. Once there, click the “Digital Dashboard” selection. You will be shown a variety of different dashboard demos. Run a couple of the demos. (1) What sorts of information and metrics are shown on the demos?

What sorts of actions can you take? (2) Using some of the basic concepts from Few’s critiques, describe some of the good design points and bad design points of the demos.

Team Assignments and Role-Playing Projects

1. Kathryn Avery has been a DBA with a nationwide retail chain (Big Chain) for the past six years. She has recently been asked to lead the development of Big Chain’s first data warehouse. The project has the sponsorship of senior management and the CIO. The rationale for developing the data warehouse is to advance the reporting systems, particularly in sales and marketing and, in the longer term, to improve Big Chain’s CRM. Kathryn has been to a Data Warehousing Institute conference and has been doing some reading, but she is still mystified about development methodologies. She knows there are two groups—EDW (Inmon) and architected DMs (Kimball)—that have equally robust features.

Initially, she believed that the two methodologies were extremely dissimilar, but as she has examined them more carefully, she isn’t so certain. Kathryn has a number of questions that she would like answered:

- What are the real differences between the methodologies?
- What factors are important in selecting a particular methodology?
- What should be her next steps in thinking about a methodology?

Help Kathryn answer these questions. (This exercise was adapted from Duncan, K., Reeves, L., & Griffin, J. (2003, Fall). BI experts’ perspective. *Business Intelligence Journal*, 8(4), 14–19.)

2. Jeet Kumar is the administrator of data warehousing at a big regional bank. He was appointed five years ago to implement a data warehouse to support the bank’s CRM business strategy. Using the data warehouse, the bank has been successful in integrating customer information, understanding customer profitability, attracting customers, enhancing customer relationships, and retaining customers.

Over the years, the bank’s data warehouse has moved closer to real time by moving to more frequent refreshes of the data warehouse. Now, the bank wants to implement customer self-service and call center applications that require even fresher data than is currently available in the warehouse.

Jeet wants some support in considering the possibilities for presenting fresher data. One alternative is to entirely commit to implementing RDW. His ETL vendor is prepared to assist him make this change. Nevertheless, Jeet has been informed about EAI and EII technologies and wonders how they might fit into his plans.

In particular, he has the following questions:

- What exactly are EAI and EII technologies?
- How are EAI and EII related to ETL?
- How are EAI and EII related to RDW?
- Are EAI and EII required, complementary, or alternatives to RDW?

Help Jeet answer these questions. (This exercise was adapted from Brobst, S., Levy, E., & Muzilla, C. (2005, Spring). Enterprise application integration and enterprise information integration. *Business Intelligence Journal*, 10(2), 27–33.)

3. Interview administrators in your college or executives in your organization to determine how data warehousing could assist them in their work. Write a proposal describing your findings. Include cost estimates and benefits in your report.
4. Do a search and identify a list of data warehousing potential risks and challenges. Rank order them based on their likelihood of happening and probable negative impact.
5. Go to a well-known information dashboard provider Web site (dundas.com, idashboards.com, enterprise-dashboard.com). These sites provide a number of examples of executive dashboards. As a team, select a particular industry (e.g., healthcare, banking, airline). Locate a handful of example dashboards for that industry. Describe the types of metrics found on the dashboards. What types of displays are used to provide the information? Using what you know about dashboard design, provide a paper prototype of a dashboard for this information.
6. Go to University of Arkansas data sources. Choose one of the large data sets, and download a large number of records (this may require you to write an SQL statement that creates the variables that you want to include in the data set). Come up with at least 10 questions that can be addressed with information visualization. Using your

favorite data visualization tool (e.g., PowerBI, Tableau), analyze the data, and prepare a detailed report that includes screenshots and other visuals.

Internet Exercises

1. Search the Internet to find information about data warehousing. Identify some newsgroups that have an interest in this concept. Explore digital databases in your library, e-library, and Google for recent articles on the topic. Begin with tdwi.org, technologyevaluation.com, and the major vendors: teradata.com, sas.com, oracle.com, and ncr.com. Also check cio.com, dmreview.com, dssresources.com, and db2mag.com.
2. Survey some ETL tools and vendors. Start with fairisaac.com and egain.com. Also consult dmreview.com (nowadays called informationbuilders.com).
3. Contact some data warehouse vendors and obtain information about their products. Give special attention to vendors that provide tools for multiple purposes, such as IBM Cognos, Software A&G, SAS Institute, Teradata, and Oracle. Free online demos are available from some of these vendors. Download a demo or two and try them. Write a report describing your experience.
4. Explore teradata.com for developments and success stories about data warehousing in the age of big data and cloud computing. Write a report about what you have discovered.
5. Explore teradata.com for white papers and Web-based courses on data warehousing. Read the former and watch the latter. (Divide the class so that all the sources are covered.) Write a report about what you have discovered.
6. Find recent cases of successful data warehousing applications. Go to data warehouse vendors' sites and look for cases or success stories. Select two, and write a brief summary to present to your class.

References

- Abela, A. (2008). *Advanced presentations by design: Creating communication that drives action*. New York: Wiley.
- Adamson, C. (2009). *The star schema handbook: The complete reference to dimensional data warehouse design*. Hoboken, NJ: Wiley.
- Agosta, L. (2006). The data strategy adviser: The year ahead—Data warehousing trends 2006. *DM Review*, 16(1).
- Ante, S. E., & McGregor, J. (2006). Giving the boss the big picture: A dashboard pulls up everything the CEO needs to run the show. *Business Week*, 43–51.
- Ariyachandra, T., & Watson, H. (2005). Key factors in selecting a data warehouse architecture. *Business Intelligence Journal*, 10(3).
- Ariyachandra, T., & Watson, H. (2006a, January). Benchmarks for BI and data warehousing success. *DM Review*, 16(1).
- Ariyachandra, T., & Watson, H. (2006b). Which data warehouse architecture is most successful? *Business Intelligence Journal*, 11(1).
- Basu, R. (2003, November). Challenges of real-time data warehousing. *DM Review*. <http://www.information-management.com/specialreports/20031111/7684-1.html> (accessed September 2020).
- Benander, A., Benander, B., Fadlalla, A., & James, G. (2000, Winter). Data warehouse administration and management. *Information Systems Management*, 17(1).

- Bonde, A., & Kuckuk, M. (2004, April). Real world business intelligence: The implementation perspective. *DM Review*, 14(4).
- Brown, M. (2004, May 9–12). 8 characteristics of a successful data warehouse. Proceedings of the 29th Annual SAS Users Group International Conference (SUGI 29). Montreal, Canada.
- Campbell, C. (2015). Top five differences between data lakes and data warehouses. www.blue-granite.com/blog/bid/402596/Top-Five-Differences-between-Data-Lakes-and-Data-Warehouses (accessed December 2022).
- Dixon, J. (2010). James Dixon's Blog on "Pentaho, Hadoop, and Data Lakes." <https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/> (accessed January 2023).
- Dull, T. (2021). Data lake vs data warehouse: Key differences. KDnuggets.com. <http://www.kdnuggets.com/2015/09/datalake-vs-data-warehouse-key-differences.html> (accessed June 2022).
- Eckerson, W. (2005, April 1). Data warehouse builders advocate for different architectures. Application Development Trends. <https://adtmag.com/articles/2005/04/01/data-warehouse-builders-advocate-for-different-architectures.aspx> (accessed September 2021).
- Eckerson, W. (2006). *Performance dashboards*. New York: Wiley.
- Edwards, M. (2003, Fall). 2003 Best Practices Awards winners: Innovators in business intelligence and data warehousing. *Business Intelligence Journal*, 8(4).
- Elson, R., & LeClerc, R. (2005). Security and privacy concerns in the data warehouse environment. *Business Intelligence Journal*, 10(3).
- Few, S. (2005, Winter). Dashboard design: Beyond meters, gauges, and traffic lights. *Business Intelligence Journal*, 10(1).
- Few, S. (2007). Data visualization: Past, present, and future. perceptualedge.com/articles/Whitepapers/Data_Visualization.pdf (accessed July 2016).
- Furtado, P. (2009). A survey of parallel and distributed data warehouses. *International Journal of Data Warehousing and Mining*, 5(2), 57–78.
- Golfarelli, M., & Rizzi, S. (2009). *Data warehouse design: Modern principles and methodologies*. San Francisco: McGraw-Hill Osborne Media.
- Grimes, S. (2009a, May 2). Seeing connections: Visualizations makes sense of data. Intelligent Enterprise. i.cmpnet.com/intelligententerprise/next-era-business-intelligence/Intelligent_Enterprise_Next_Era_BI_Visualization.pdf (accessed January 2010).
- Hammergren, T. C., & Simon, A. R. (2009). An intelligent look at business intelligence. *Data warehousing for dummies*.
- Hardin, M., Hom, D., Perez, R., & Williams, L. (2012). Which chart or graph is right for you? Tableau Software: Tell Impassioned Stories with DataCE. Tableau Software. http://www.tableau.com/sites/default/files/media/which_chart_v6_final_0.pdf (accessed August 2016).
- Hill, G. (2016). A Guide to enterprise reporting. ghill.customer.netspace.net.au/reporting/definition.html (accessed July 2016).
- Imhoff, C. (2001, May). Power up your enterprise portal. E-Business Advise.
- Inmon, B. Levins, M. and Srivastava, R. (2021). Five Steps to a Successful Data Lakehouse, retrieved from <https://databricks.com/wp-content/uploads/2021/10/The-Data-Lakehouse.pdf> (accessed June 2022).
- Inmon, W. H. (2005). *Building the data warehouse*, 4th ed. New York: Wiley.
- Kay, R. (2005, September 19). EII. *Computerworld*, 39(38).
- Kelly, C. (2001, June 14). Calculating data warehousing ROI. SearchSQLServer.com.
- Manglik, A., & Mehra, V. (2005, Winter). Extending enterprise BI capabilities: New patterns for data integration. *Business Intelligence Journal*, 10(1).
- Matney, D. (2003, Spring). End-user support strategy. *Business Intelligence Journal*, 8(3).
- Mehra, V. (2005, Summer). Building a metadata-driven enterprise: A holistic approach. *Business Intelligence Journal*, 10(3).
- Moseley, M. (2009). Eliminating data warehouse pressures with master data services and SOA. *Business Intelligence Journal*, 14(2), 33–43.
- Murtaza, A. (1998, Fall). A framework for developing enterprise data warehouses. *Information Systems Management*, 15(4).
- Nash, K. S. (2002). Chemical reaction. *Baseline*, (8), 27–36.
- Novell. (2009, April). Executive dashboards elements of success. Novell white paper. www.novell.com/docrep/documents/3rkw3etfc3/Executive%20Dashboards_Elements_of_Success_White_Paper_en.pdf (accessed June 2016).
- Orovic, V. (2003). To do & not to do. *eAI Journal*, 37–43.
- Parzinger, M. J., & Frolick, M. N. (2001, July). Creating competitive advantage through data warehousing. *Information Strategy*, 17(4).
- Radha, R. (2008). Eight best practices in dashboard design. Information Management. www.information-management.com/news/columns/10001129-1.html (accessed July 2016).
- Romero, O., & Abelló, A. (2009). A survey of multidimensional modeling methodologies. *International Journal of Data Warehousing and Mining*, 5(2), 1–24.
- Russom, P. (2009). Next generation data warehouse platforms. TDWI best practices report. tdwi.org/research/reportseries/reports.aspx?pid=842 (accessed January 2016).
- Sen, A. (2004, April). Metadata management: Past, present and future. *Decision Support Systems*, 37(1).

- Sen, A., & Sinha, P. (2005). A comparison of data warehousing methodologies. *Communications of the ACM*, 48(3).
- Solomon, M. (2005, Winter). Ensuring a successful data warehouse initiative. *Information Systems Management Journal*, 22(1), 26–36.
- Songini, M. L. (2004, February 2). ETL quickstudy. *Computer-world*, 38(5).
- Watson, H. J. (2002). Recent developments in data warehousing. *Communications of the ACM*, 8(1).
- Watson, H. J., Goodhue, D. L., & Wixom, B. H. (2002). The benefits of data warehousing: Why some organizations realize exceptional payoffs. *Information & Management*, 39.
- Zhao, X. (2005, October 7). Meta data management maturity model. *DM Direct Newsletter*.

Predictive Analytics I: Data Mining Process, Methods, and Algorithms

LEARNING OBJECTIVES

- Define data mining as an enabling technology for business analytics
- Understand the objectives and benefits of data mining
- Become familiar with the wide range of applications of data mining
- Learn the standardized data mining processes
- Learn different methods and algorithms of data mining
- Build awareness of the existing data mining software tools
- Understand the privacy issues, pitfalls, and myths of data mining

Generally speaking, data mining is a way to develop intelligence (i.e., actionable information or knowledge) from data that an organization collects, organizes, and stores. A wide range of data mining techniques are being used by organizations to gain a better understanding of their customers and their operations and to solve complex organizational problems. In this chapter, we study data mining as an enabling technology for business analytics and predictive analytics, learn about the standard processes of conducting data mining projects, understand and build expertise in the use of major data mining techniques, develop awareness of the existing software tools, and explore privacy issues, common myths, and pitfalls that are often associated with data mining.

- 5.1** Opening Vignette: Police Departments Are Using Predictive Analytics to Foresee and Fight Crime 251
- 5.2** Data Mining Concepts and Applications 254
- 5.3** Data Mining Applications 261
- 5.4** Data Mining Process 264
- 5.5** Data Mining Methods 272
- 5.6** Data Mining Software Tools 286
- 5.7** Data Mining Privacy Issues, Myths, and Blunders 292

5.1 OPENING VIGNETTE: Police Departments Are Using Predictive Analytics to Foresee and Fight Crime

Big data, predictive analytics, and data mining have become an integral part of many law enforcement agencies. Police departments in some of the largest U.S. cities, Los Angeles, New York, and Chicago police departments, have been using predictive policing as a way to foresee criminal activity. Essentially, predictive policing uses computers and data mining to analyze large variety of historical data sets, including prior crime data, to predict where the crime will take place so that they can pre-deploy police to prevent or mitigate the criminal activities, or to identify and deal with individuals who are more likely to commit or be a victim of a crime.

Predictive policing, a term commonly used for smart policing nowadays, involves using data science to analyze massive amounts of information in order to predict and help prevent potential future crimes. Place-based predictive policing is the most widely practiced method where preexisting crime data are used to identify places and times that have a high risk of crime. Person-based predictive policing, on the other hand, attempts to identify individuals or groups who are likely to commit a crime—or to be victim of one—by analyzing for risk factors such as past arrests or victimization patterns.

One of the exemplary police departments successfully using predictive analytics is Miami-Dade County and its Police Department (whose mission is not only to protect the safety of Florida’s largest county with 2.5 million citizens (making it the seventh largest in the United States), but also to provide a safe and inviting climate for the millions of tourists that come from around the world to enjoy the county’s natural beauty, warm climate, and stunning beaches). With tourists spending over \$20 billion every year and generating nearly a third of Florida’s sales taxes, it is hard to overstate the importance of tourism to the region’s economy. So, although few of the county’s police officers would likely list economic development in their job description, nearly all grasp the vital link between safe streets and the region’s tourist-driven prosperity.

That connection is paramount for Lieutenant Arnold Palmer, currently supervising the Robbery Investigations Section, and a former supervisor of the department’s Robbery Intervention Detail. This specialized team of detectives is focused on intensely policing the county’s robbery hot spots and worst repeat offenders. He and the team occupy modest offices on the second floor of a modern-looking concrete building, set back from a palm-lined street on the western edge of Miami. In his 10 years in the unit, out of 23 in total on the force, Palmer has seen a lot of changes. It’s not just in policing practices, like the way his team used to mark street crime hot spots with colored pushpins on a map.

POLICING WITH LESS

Palmer and the team have also seen the impact of a growing population, shifting demographics, and a changing economy on the streets they patrol. Like any good police force, they’ve continually adapted their methods and practices to meet a policing challenge that has grown in scope and complexity. But like nearly all branches of the county’s government, intensifying budget pressures have placed the department in a squeeze between rising demands and shrinking resources.

Palmer, who sees detectives as front-line fighters against a rising tide of street crime and the looming prospect of ever-tightening resources, put it this way: “Our basic challenge was how to cut street crime even as tighter resources have reduced the number of cops on the street.” Over the years, the team had been open to trying new tools, the most notable of which was a program called “analysis-driven enforcement” that used crime history data as the basis for positioning teams of detectives. “We’ve evolved a lot since

then in our ability to predict where robberies are likely to occur, both through the use of analysis and our own collective experience.”

NEW THINKING ON COLD CASES

The more confounding challenge for Palmer and his team of investigators, one shared with the police of all major urban areas, is in closing the hardest cases, where leads, witnesses, video—any facts or evidence that can help solve a case—are lacking. It’s not surprising, explains Palmer, because “the standard practices we used to generate leads, like talking to informants or to the community or to patrol officers, haven’t changed much, if at all,” says Palmer. “That kind of an approach works okay, but it relies a lot on the experience our detectives carry in their head. When the detectives retire or move on, that experience goes with them.”

Palmer’s conundrum was that turnover, due to the retirement of many of his most experienced detectives, was on an upward trend. True, he saw the infusion of young blood as an inherently good thing, especially given their greater comfort with the new types of information—from e-mails, social media, and traffic cameras, to name a few—that his team had access to. But as Palmer recounts, the problem came when the handful of new detectives coming into the unit turned to look for guidance from the senior officers “and it’s just not there. We knew at that point we needed a different way to fill the experience gap going forward.”

His ad hoc efforts to come up with a solution led to blue-sky speculation. What if new detectives on the squad could pose the same questions to a computer database as they would to a veteran detective? That speculation planted a seed in Palmer’s mind that wouldn’t go away.

THE BIG PICTURE STARTS SMALL

What was taking shape within the robbery unit demonstrated how big ideas can come from small places. But more important, it showed that for these ideas to reach fruition, the “right” conditions need to be in alignment at the right time. On a leadership level, that means a driving figure in the organization who knows what it takes to nurture top-down support as well as crucial bottom-up buy-in from the ranks, while at the same time keeping the department’s information technology (IT) personnel on the same page. That person was Palmer. At the organizational level, the robbery unit served as a particularly good launching point for lead modeling because of the prevalence of repeat offenders among perpetrators. Ultimately, the department’s ability to unleash the broader transformative potential of lead modeling would hinge in large part on the team’s ability to deliver results on a smaller scale.

When early tests and demos proved encouraging—with the model yielding accurate results when the details of solved cases were fed into it—the team started gaining attention. The initiative received a critical boost when the robbery bureau’s unit major and captain voiced their support for the direction of the project, telling Palmer that “if you can make this work, run with it.” But more important than the encouragement, Palmer explains, was their willingness to advocate for the project among the department’s higher-ups. “I can’t get it off the ground if the brass doesn’t buy in,” says Palmer. “So, their support was crucial.”

SUCCESS BRINGS CREDIBILITY

Having been appointed the official liaison between IT and the robbery unit, Palmer set out to strengthen the case for the lead-modeling tool—now officially called Blue PALMS, for Predictive Analytics Lead Modeling Software—by building up a series of successes. His

constituency was not only the department brass, but also the detectives whose support would be critical to its successful adoption as a robbery-solving tool. In his attempts to introduce Blue PALMS, resistance was predictably stronger among veteran detectives, who saw no reason to give up their long-standing practices. Palmer knew that dictates or coercion wouldn't win their hearts and minds. He would need to build a beachhead of credibility.

Palmer found that opportunity in one of his best and most experienced detectives. Early in a robbery investigation, the detective indicated to Palmer that he had a strong hunch who the perpetrator was and wanted, in essence, to test the Blue PALMS system. So, at the detective's request, the department analyst fed key details of the crime into the system, including the *modus operandi*, or MO. The system's statistical models compared these details to a database of historical data, looking for important correlations and similarities in the crime's signature. The report that came out of the process included a list of 20 suspects ranked in order of match strength, or likelihood. When the analyst handed the detective the report, his "hunch" suspect was listed in the top five. Soon after his arrest, he confessed, and Palmer had gained a solid convert.

Though it was a useful exercise, Palmer realized that the true test wasn't in confirming hunches but in breaking cases that had come to a dead end. Such was the situation in a carjacking that had, in Palmer's words, "no witnesses, no video and no crime scene—nothing to go on." When the senior detective on the stalled case went on leave after three months, the junior detective to whom it was assigned requested a Blue PALMS report. Shown photographs of the top people on the suspect list, the victim made a positive identification of the suspect leading to the successful conclusion of the case. That suspect was number one on the list.

JUST THE FACTS

The success that Blue PALMS continues to build has been a major factor in Palmer's success in getting his detectives on board. But if there's a part of his message that resonates even more with his detectives, it's the fact that Blue PALMS is designed not to change the basics of policing practices, but to enhance them by giving them a second chance of cracking the case. "Police work is at the core about human relations—about talking to witnesses, to victims, to the community—and we're not out to change that," says Palmer. "Our aim is to give investigators factual insights from information we already have that might make a difference, so even if we're successful 5% of the time, we're going to take a lot of offenders off the street."

The growing list of cold cases solved has helped Palmer in his efforts to reinforce the merits of Blue PALMS. But, in showing where his loyalty lies, he sees the detectives who've closed these cold cases—not the program—as most deserving of the spotlight, and that approach has gone over well. At his chief's request, Palmer is beginning to use his liaison role as a platform for reaching out to other areas in the Miami-Dade Police Department.

SAFER STREETS FOR A SMARTER CITY

When he speaks of the impact of tourism, a thread that runs through Miami-Dade's Smarter Cities vision, Palmer sees Blue PALMS as an important tool to protect one of the county's greatest assets. "The threat to tourism posed by rising street crime was a big reason the unit was established," says Palmer. "The fact that we're able to use analytics and intelligence to help us close more cases and keep more criminals off the street is good news for our citizens and our tourist industry."

► QUESTIONS FOR THE OPENING VIGNETTE

1. Why do law enforcement agencies and departments like LAPD, NYPD, CPD, and Miami-Dade Police Department embrace advanced analytics and data mining?

2. What are the top challenges for law enforcement agencies and departments? Can you think of other challenges (not mentioned in this case) that can benefit from data mining?
3. What are the sources of data that law enforcement agencies and departments like Miami-Dade Police Department use for their predictive modeling and data mining projects?
4. What type of analytics do law enforcement agencies and departments use to fight crime?
5. What does “the big picture starts small” mean in this case? Explain.

WHAT WE CAN LEARN FROM THIS VIGNETTE

The law enforcement agencies and departments are under tremendous pressure to carry out their mission of safeguarding people with limited resources. The environment within which they perform their duties is becoming increasingly more challenging so that they have to constantly adopt and perhaps stay a few steps ahead to prevent the likelihood of catastrophes. Understanding the changing nature of crime and criminals is an ongoing challenge. In the midst of these challenges, what works in favor of these agencies is the availability of the data and analytics technologies to better analyze past occurrences and to foresee future events. Data has become available more than it has in the past. Applying advanced analytics and data mining tools (i.e., knowledge discovery techniques) to these large and rich data sources provides them with the insight that they need to better prepare and act on their duties. Therefore, law enforcement agencies are becoming one of the leading users of the new face of analytics. Data mining is a prime candidate for better understanding and management of these mission critical tasks with a high level of accuracy and timeliness. The study described in the opening vignette clearly illustrates the power of analytics and data mining to create a holistic view of the world of crime and criminals for better and faster reaction and management. In this chapter, you will see a wide variety of data mining applications solving complex problems in a variety of industries and organizational settings where the data are used to discover actionable insight to improve mission readiness, operational efficiency, and competitive advantage.

Sources: Building a smarter, safer county with intelligent, analytics-driven operations management, IBM Customer Case Studies, retrieved from <https://www.ibm.com/case-studies/miami-dade-county> (accessed June 2022); Predictive Policing: The Future of Law Enforcement? by US Department of Justice, retrieved from <https://www.ojp.gov/pdffiles1/nij/230414.pdf> (accessed June 2022). Predictive Policing Explained, Brennan Center for Justice, retrieved from <https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained> (accessed June 2022).

5.2 DATA MINING CONCEPTS AND APPLICATIONS

Data mining, a new and exciting technology of only a few years ago, has become a common practice for a vast majority of organizations. In an interview with *Computerworld* magazine in January 1999, Dr. Arno Penzias (Nobel laureate and former chief scientist of Bell Labs) identified data mining from organizational databases as a key application for corporations of the near future. In response to *Computerworld's* age-old question of “What will be the killer applications in the corporation?” Dr. Penzias replied: “Data mining.” He then added, “Data mining will become much more important and companies will throw away nothing about their customers because it will be so valuable. If you’re not doing this, you’re out of

business.” Similarly, in an article in *Harvard Business Review*, Thomas Davenport (2006) argued that the latest strategic weapon for companies is analytical decision-making, providing examples of companies such as Amazon.com, Capital One, Marriott International, and others that have used analytics to better understand their customers and optimize their extended supply chains to maximize their returns on investment while providing the best customer service. This level of success is highly dependent on a company understanding its customers, vendors, business processes, and the extended supply chain very well.

A large portion of “understanding the customer” can come from analyzing the vast amount of data that a company collects. The cost of storing and processing data has decreased dramatically in the recent past, and, as a result, the amount of data stored in electronic form has grown at an explosive rate. With the creation of large databases, the possibility of analyzing the data stored in them has emerged. The term *data mining* was originally used to describe the process through which previously unknown patterns in data were discovered. This definition has since been stretched beyond those limits by some software vendors to include most forms of data analysis in order to increase sales with the popularity of the data mining label. In this chapter, we accept the original definition of data mining.

Although the term *data mining* is relatively new, the ideas behind it are not. Many of the techniques used in data mining have their roots in traditional statistical analysis and artificial intelligence work done since the early part of the 1980s. Why, then, has it suddenly gained the attention of the business world? Following are some of most pronounced reasons:

- More intense competition at the global scale driven by customers’ ever-changing needs and wants in an increasingly saturated marketplace.
- General recognition of the untapped value hidden in large data sources.
- Consolidation and integration of database records, which enables a single view of customers, vendors, transactions, and so on.
- Consolidation of databases and other data repositories into a single location in the form of a data warehouse.
- The exponential increase in data processing and storage technologies.
- Significant reduction in the cost of hardware and software for data storage and processing.
- Movement toward the demassification (conversion of information resources into nonphysical form) of business practices.

Data generated by the Internet is increasing rapidly in both volume and complexity. Large amounts of genomic data are being generated and accumulated all over the world. Disciplines such as astronomy and nuclear physics create huge quantities of data on a regular basis. Medical and pharmaceutical researchers constantly generate and store data that can then be used in data mining applications to identify better ways to accurately diagnose and treat illnesses and to discover new and improved drugs.

On the commercial side, perhaps the most common use of data mining has been in the finance, retail, and healthcare sectors. Data mining is used to detect and reduce fraudulent activities, especially in insurance claims and credit card use (Chan et al., 1999); to identify customer buying patterns (Hoffman, 1999); to reclaim profitable customers (Hoffman, 1998); to identify trading rules from historical data; and to aid in increased profitability using market-basket analysis. Data mining is already widely used to better target clients, and with the widespread development of e-commerce, this can only become more imperative with time.

Definitions, Characteristics, and Benefits

Simply defined, **data mining** is a term used to describe discovering or “mining” knowledge from large amounts of data. When considered by analogy, one can easily realize that the term *data mining* is a misnomer; that is, mining of gold from within rocks or dirt is referred to as “gold” mining rather than “rock” or “dirt” mining. Therefore, data mining perhaps should have been named “knowledge mining” or “knowledge discovery.” Despite the mismatch between the term and its meaning, *data mining* has become the choice of the community. Many other names that are associated with data mining include *knowledge extraction*, *pattern analysis*, *data archaeology*, *information harvesting*, *pattern searching*, and *data dredging*.

Technically speaking, data mining is a process that uses statistical, mathematical, and artificial intelligence techniques to extract and identify useful information and subsequent knowledge (or patterns) from large sets of data. These patterns can be in the form of business rules, affinities, correlations, trends, or prediction models (see Nemati and Barko, 2001). Most literature defines data mining as “the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases,” where the data are organized in records structured by categorical, ordinal, and continuous variables (Fayyad et al., 1996, pp. 40–41). In this definition, the meanings of the key term are as follows:

- *Process* implies that data mining comprises many iterative steps.
- *Nontrivial* means that some experimentation-type search or inference is involved; that is, it is not as straightforward as a computation of predefined quantities.
- *Valid* means that the discovered patterns should hold true on new data with a sufficient degree of certainty.
- *Novel* means that the patterns are not previously known to the user within the context of the system being analyzed.
- *Potentially useful* means that the discovered patterns should lead to some benefit to the user or task.
- *Ultimately understandable* means that the pattern should make business sense that leads to the user saying, “Mmm! It makes sense; why didn’t I think of that,” if not immediately, at least after some postprocessing.

Data mining is not a new discipline, but rather a new definition for the use of many disciplines. Data mining is tightly positioned at the intersection of many disciplines, including statistics, artificial intelligence, machine learning, management science, information systems (IS), and databases (see Figure 5.1). Using advances in all of these disciplines, data mining strives to make progress in extracting useful information and knowledge from large databases. It is an emerging field that has attracted much attention in a very short time.

The following are the major characteristics and objectives of data mining:

- Data are often buried deep within very large databases, which sometimes contain data from several years. In many cases, the data are cleansed and consolidated into a data warehouse. Data may be presented in a variety of formats (see Chapter 2 for a brief taxonomy of data).
- The data mining environment is usually a client/server architecture or a Web-based IS architecture.
- Sophisticated new tools, including advanced visualization tools, help to remove the information ore buried in corporate files or archival public records. Finding it involves massaging and synchronizing the data to get the right results. Cutting-edge data miners are also exploring the usefulness of soft data (i.e., unstructured text stored in such places as Lotus Notes databases, text files on the Internet, or enterprise-wide intranets).

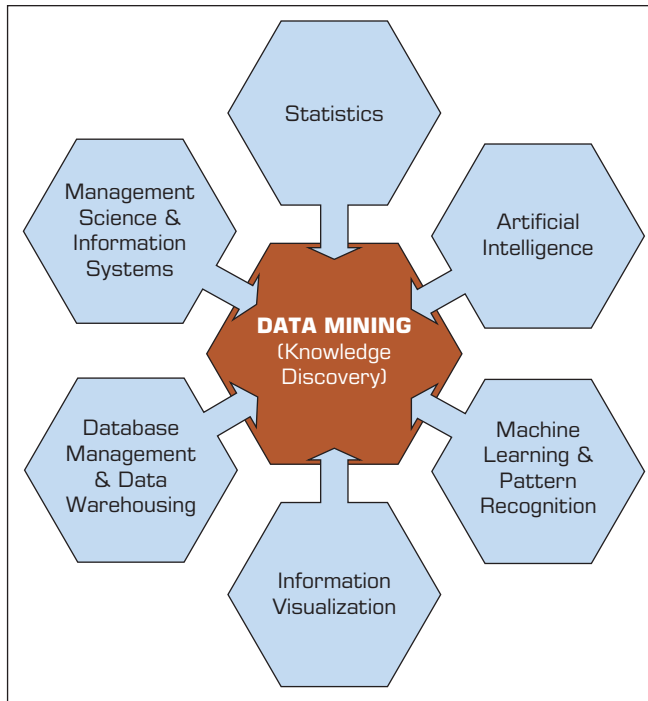


FIGURE 5.1 Data Mining Is a Blend of Multiple Disciplines.

- The miner is often an end user, empowered by data drills and other powerful query tools to ask ad hoc questions and obtain answers quickly, with little or no programming skill.
- Striking it rich often involves finding an unexpected result and requires end users to think creatively throughout the process, including the interpretation of the findings.
- Data mining tools are readily combined with spreadsheets and other software development tools. Thus, the mined data can be analyzed and deployed quickly and easily.
- Because of the large amounts of data and massive search efforts, it is sometimes necessary to use parallel processing for data mining.

A company that effectively leverages data mining tools and technologies can acquire and maintain a strategic competitive advantage. Data mining offers organizations an indispensable decision-enhancing environment to exploit new opportunities by transforming data into a strategic weapon. See Nemati and Barko (2001) for a more detailed discussion on the strategic benefits of data mining.

How Data Mining Works

Using existing and relevant data obtained from within and outside the organization, data mining builds models to discover patterns among the attributes presented in the data set. Models are the mathematical representations (simple linear relationships/affinities and/or complex and highly nonlinear relationships) that identify the patterns among the attributes of the things (e.g., customers, events) described within the data set. Some of these patterns are explanatory (explaining the interrelationships and affinities among the attributes), whereas others are predictive (foretelling future values of certain attributes). In general, data mining seeks to identify four major types of patterns:

- *Associations* find the commonly co-occurring groupings of things, such as beer and diapers going together in market-basket analysis.

- *Predictions* tell the nature of future occurrences of certain events based on what has happened in the past, such as predicting the winner of the Super Bowl or forecasting the absolute temperature of a particular day.
- *Clusters* identify natural groupings of things based on their known characteristics, such as assigning customers in different segments based on their demographics and past purchase behaviors.
- *Sequential relationships* discover time-ordered events, such as predicting that an existing banking customer who already has a checking account will open a savings account followed by an investment account within a year.

These types of patterns have been *manually* extracted from data by humans for centuries, but the increasing volume of data in modern times has created a need for more automatic approaches. As data sets have grown in size and complexity, direct manual data analysis has increasingly been augmented with indirect, automatic data processing tools that use sophisticated methodologies, methods, and algorithms. The manifestation of such evolution of automated and semiautomated means of processing large data sets is now commonly referred to as *data mining*.

Generally speaking, data mining tasks can be classified into three main categories: prediction, association, and clustering. Based on the way in which the patterns are extracted from the historical data, the learning algorithms of data mining methods can be classified as either supervised or unsupervised. With supervised learning algorithms, the training data includes both the descriptive attributes (i.e., independent variables or decision variables) as well as the class attribute (i.e., output variable or result variable). In contrast, with unsupervised learning the training data includes only the descriptive attributes. Figure 5.2 shows a simple taxonomy for data mining tasks, along with the learning methods and popular algorithms for each of the data mining tasks.

PREDICTION **Prediction** is commonly referred to as the act of telling about the future. It differs from simple guessing by taking into account the experiences, opinions, and other relevant information in conducting the task of foretelling. A term that is commonly associated with prediction is *forecasting*. Even though many believe that these two terms are synonymous, there is a subtle but critical difference between the two. Whereas prediction is largely experience and opinion based, forecasting is data and model based. That is, in order of increasing reliability, one might list the relevant terms as *guessing*, *predicting*, and *forecasting*, respectively. In data mining terminology, *prediction* and *forecasting* are used synonymously, and the term *prediction* is used as the common representation of the act. Depending on the nature of what is being predicted, prediction can be named more specifically as classification (where the predicted thing, such as tomorrow's forecast, is a class label such as "rainy" or "sunny") or regression (where the predicted thing, such as tomorrow's temperature, is a real number, such as "65°F").

CLASSIFICATION **Classification**, or supervised induction, is perhaps the most common of all data mining tasks. The objective of classification is to analyze the historical data stored in a database and automatically generate a model that can predict future behavior. This induced model consists of generalizations over the records of a training data set, which help distinguish predefined classes. The hope is that the model can then be used to predict the classes of other unclassified records and, more important, to accurately predict actual future events.

Common classification tools include neural networks and decision trees (from machine learning), logistic regression and discriminant analysis (from traditional statistics), and emerging tools such as rough sets, support vector machines (SVMs), and genetic algorithms. Statistics-based classification techniques (e.g., logistic regression and

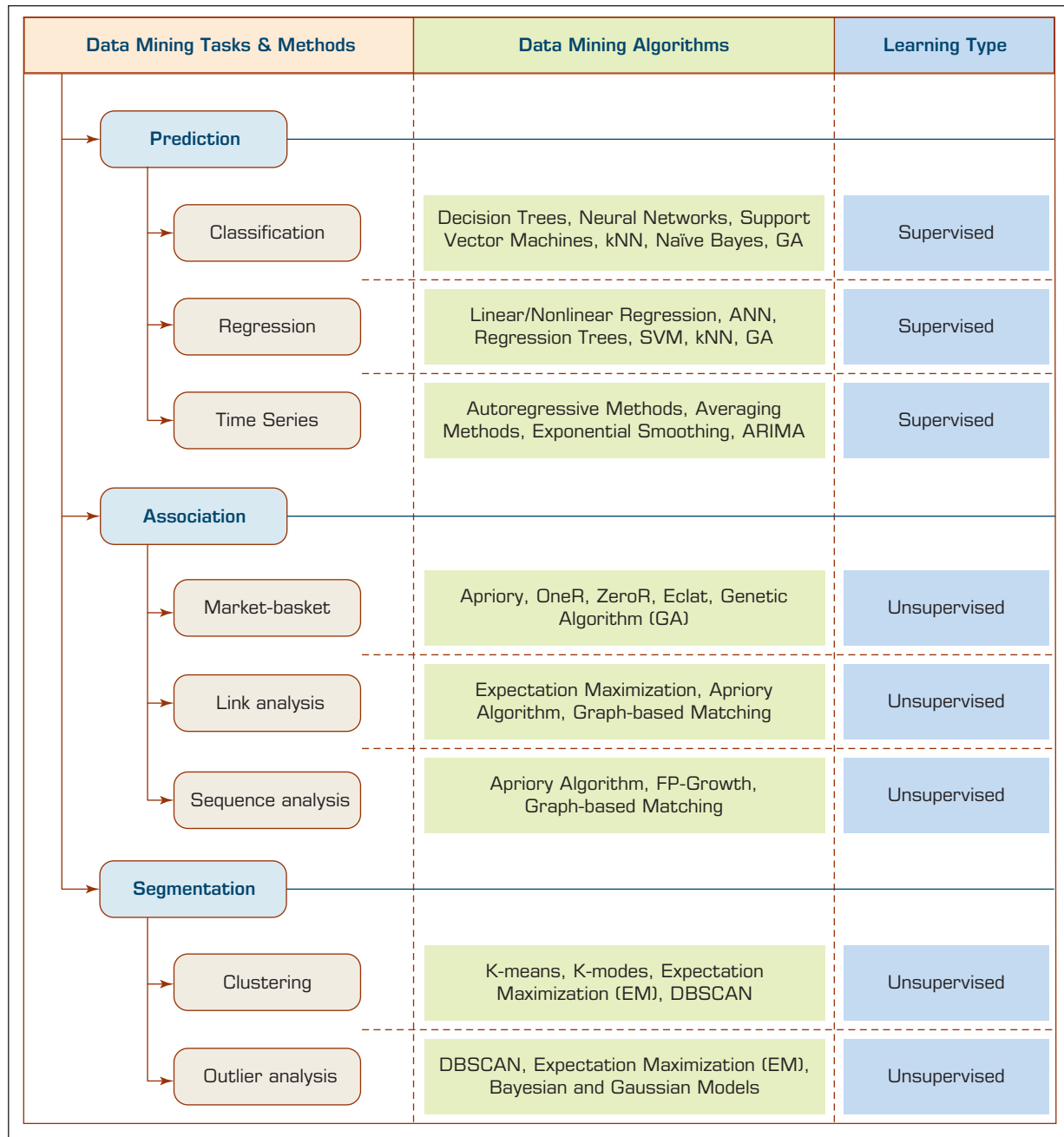


FIGURE 5.2 A Simple Taxonomy for Data Mining Tasks, Methods, and Algorithms.

discriminant analysis) have received their share of criticism—that they make unrealistic assumptions about the data, such as independence and normality—which limit their use in classification-type data mining projects.

Neural networks involve the development of mathematical structures (somewhat resembling the biological neural networks in the human brain) that have the capability to learn from past experiences presented in the form of well-structured data sets. They tend to be more effective when the number of variables involved is rather large and the relationships among them are complex and imprecise. Neural networks have disadvantages

as well as advantages. For example, it is usually very difficult to provide a good rationale for the predictions made by a neural network. Also, neural networks tend to need considerable training. Unfortunately, the time needed for training tends to increase exponentially as the volume of data increases, and in general, neural networks cannot be trained on very large databases. These and other factors have limited the applicability of neural networks in data-rich domains.

Decision trees classify data into a finite number of classes based on the values of the input variables. Decision trees are essentially a hierarchy of if-then statements and are thus significantly faster than neural networks. They are most appropriate for **categorical** and **interval data**. Therefore, incorporating continuous variables into a decision tree framework requires *discretization*; that is, converting continuous valued numerical variables to ranges and categories.

A related category of classification tools is rule induction. Unlike with a decision tree, with rule induction the if-then statements are induced from the training data directly, and they need not be hierarchical in nature. Other, more recent techniques such as SVM, rough sets, and genetic algorithms are gradually finding their way into the arsenal of classification algorithms.

CLUSTERING **Clustering** partitions a collection of things (e.g., objects, events, presented in a structured data set) into segments (or natural groupings) whose members share similar characteristics. Unlike in classification, in clustering the class labels are unknown. As the selected algorithm goes through the data set, identifying the commonalities of things based on their characteristics, the clusters are established. Because the clusters are determined using a heuristic-type algorithm, and because different algorithms may end up with different sets of clusters for the same data set, before the results of clustering techniques are put to actual use it may be necessary for an expert to interpret, and potentially modify, the suggested clusters. After reasonable clusters have been identified, they can be used to classify and interpret new data.

Not surprisingly, clustering techniques include optimization. The goal of clustering is to create groups so that the members within each group have maximum similarity and the members across groups have minimum similarity. The most commonly used clustering techniques include *k*-means (from statistics) and self-organizing maps (from machine learning), which is a unique neural network architecture developed by Kohonen (1982).

Firms often effectively use their data mining systems to perform market segmentation with cluster analysis. Cluster analysis is a means of identifying classes of items so that items in a cluster have more in common with each other than with items in other clusters. It can be used in segmenting customers and directing appropriate marketing products to the segments at the right time in the right format at the right price. Cluster analysis is also used to identify natural groupings of events or objects so that a common set of characteristics of these groups can be identified to describe them.

ASSOCIATIONS **Associations**, or *association rule learning in data mining*, is a popular and well-researched technique for discovering interesting relationships among variables in large databases. Thanks to automated data-gathering technologies such as bar code scanners, the use of association rules for discovering regularities among products in large-scale transactions recorded by point-of-sale systems in supermarkets has become a common knowledge discovery task in the retail industry. In the context of the retail industry, association rule mining is often called *market-basket analysis*.

Two commonly used derivatives of association rule mining are **link analysis** and **sequence mining**. With link analysis, the linkage among many objects of interest is discovered automatically, such as the link between Web pages and referential relationships among groups of academic publication authors. With sequence mining, relationships

are examined in terms of their order of occurrence to identify associations over time. Algorithms used in association rule mining include the popular Apriori (where frequent itemsets are identified) and FP-Growth, OneR, ZeroR, and Eclat.

VISUALIZATION AND TIME-SERIES FORECASTING Two techniques often associated with data mining are *visualization* and *time-series forecasting*. Visualization can be used in conjunction with other data mining techniques to gain a clearer understanding of underlying relationships. As the importance of visualization has increased in recent years, a new term, *visual analytics*, has emerged. The idea is to combine analytics and visualization in a single environment for easier and faster knowledge creation. Visual analytics is covered in detail in Chapter 3. In time-series forecasting, the data consists of values of the same variable that is captured and stored over time in regular intervals. These data are then used to develop forecasting models to extrapolate the future values of the same variable.

Data Mining versus Statistics

Data mining and statistics have a lot in common. They both look for relationships within data. Most people call statistics the “foundation of data mining.” The main difference between the two is that statistics starts with a well-defined proposition and hypothesis, whereas data mining starts with a loosely defined discovery statement. Statistics collects sample data (i.e., primary data) to test the hypothesis, whereas data mining and analytics use all the existing data (i.e., often observational, secondary data) to discover novel patterns and relationships. Another difference comes from the size of data that they use. Data mining looks for data sets that are as “big” as possible, whereas statistics looks for the right size of data (if the data are larger than what is needed/required for the statistical analysis, a sample of the data is used). The meaning of “large data” is rather different between statistics and data mining. A few hundred to a thousand data points are large enough to a statistician, but several million to a few billion data points are considered large for data mining studies.

SECTION 5.2 REVIEW QUESTIONS

1. Define *data mining*. Why are there many different names and definitions for data mining?
2. What recent factors have increased the popularity of data mining?
3. Is data mining a new discipline? Explain.
4. What are some major data mining methods and algorithms?
5. What are the key differences between the major data mining tasks?

5.3 DATA MINING APPLICATIONS

Data mining has become a popular tool in addressing many complex businesses problems and opportunities. It has been proven to be very successful and helpful in many areas, some of which are shown by the following representative examples. The goal of many of these business data mining applications is to solve a pressing problem or to explore an emerging business opportunity to create a sustainable competitive advantage.

- **Customer relationship management.** Customer relationship management (CRM) is the extension of traditional marketing. The goal of CRM is to create one-on-one relationships with customers by developing an intimate understanding of their needs and wants. As businesses build relationships with their customers over time through a variety of interactions (e.g., product inquiries, sales, service requests,

warranty calls, product reviews, social media connections), they accumulate tremendous amounts of data. When combined with demographic and socioeconomic attributes, this information-rich data can be used to (1) identify most likely responders/buyers of new products/services (i.e., customer profiling), (2) understand the root causes of customer attrition to improve customer retention (i.e., churn analysis), (3) discover time-variant associations between products and services to maximize sales and customer value, and (4) identify the most profitable customers and their preferential needs to strengthen relationships and to maximize sales.

- **Banking.** Data mining can help banks with the following: (1) automating the loan application process by accurately predicting the most probable defaulters, (2) detecting fraudulent credit card and online banking transactions, (3) identifying ways to maximize customer value by selling them products and services that they are most likely to buy, and (4) optimizing the cash return by accurately forecasting the cash flow on banking entities (e.g., ATM machines, banking branches).
- **Retailing and logistics.** In the retailing industry, data mining can be used to (1) predict accurate sales volumes at specific retail locations to determine correct inventory levels, (2) identify sales relationships between different products (with market-basket analysis) to improve the store layout and optimize sales promotions, (3) forecast consumption levels of different product types (based on seasonal and environmental conditions) to optimize logistics and, hence, maximize sales, and (4) discover interesting patterns in the movement of products (especially for the products that have a limited shelf life because they are prone to expiration, perishability, and contamination) in a supply chain by analyzing sensory and radio-frequency identification (RFID) data.
- **Manufacturing and production.** Manufacturers can use data mining to (1) predict machinery failures before they occur through the use of sensory data (enabling what is called *condition-based maintenance*), (2) identify anomalies and commonalities in production systems to optimize manufacturing capacity, and (3) discover novel patterns to identify and improve product quality.
- **Brokerage and securities trading.** Brokers and traders use data mining to (1) predict when and how much certain bond prices will change, (2) forecast the range and direction of stock fluctuations, (3) assess the effect of particular issues and events on overall market movements, and (4) identify and prevent fraudulent activities in securities trading.
- **Insurance.** The insurance industry uses data mining techniques to (1) forecast claim amounts for property and medical coverage costs for better business planning, (2) determine optimal rate plans based on the analysis of claims and customer data, (3) predict which customers are more likely to buy new policies with special features, and (4) identify and prevent incorrect claim payments and fraudulent activities.
- **Computer hardware and software.** Data mining can be used to (1) predict disk drive failures well before they actually occur, (2) identify and filter unwanted Web content and e-mail messages, (3) detect and prevent computer network security breaches and (4) identify potentially insecure software products.
- **Government and defense.** Data mining also has a number of military applications. It can be used to (1) forecast the cost of moving military personnel and equipment; (2) predict an adversary's moves and, hence, develop more successful strategies for military engagements; (3) predict resource consumption for better planning and budgeting; and (4) identify classes of unique experiences, strategies, and lessons learned from military operations for better knowledge sharing throughout the organization.
- **Travel industry (airlines, hotels/resorts, rental car companies).** Data mining has a variety of uses in the travel industry. It is successfully used to (1)

predict sales of different services (seat types in airplanes, room types in hotels/resorts, car types in rental car companies) in order to optimally price services to maximize revenues as a function of time-varying transactions (commonly referred to as *yield management*), (2) forecast demand at different locations to better allocate limited organizational resources, (3) identify the most profitable customers and provide them with personalized services to maintain their repeat business, and (4) retain valuable employees by identifying and acting on the root causes for attrition.

- **Healthcare.** Data mining has a number of healthcare applications. It can be used to (1) identify people without health insurance and the factors underlying this undesired phenomenon, (2) identify novel cost–benefit relationships between different treatments to develop more effective strategies, (3) forecast the level and the time of demand at different service locations to optimally allocate organizational resources, and (4) understand the underlying reasons for customer and employee attrition.
- **Medicine.** Use of data mining in medicine should be viewed as an invaluable complement to traditional medical research, which is mainly clinical and biological in nature. Data mining analyses can (1) identify novel patterns to improve survivability of patients with cancer, (2) predict success rates of organ transplantation patients to develop better organ donor matching policies, (3) identify the functions of different genes in the human chromosome (known as genomics), and (4) discover the relationships between symptoms and illnesses (as well as illnesses and successful treatments) to help medical professionals make informed and correct decisions in a timely manner.
- **Entertainment industry.** Data mining is successfully used by the entertainment industry to (1) analyze viewer data to decide what programs to show during prime time and how to maximize returns by knowing where to insert advertisements, (2) predict the financial success of movies before they are produced to make investment decisions and to optimize the returns (Delen et al., 2007), (3) forecast the demand at different locations and different times to better schedule entertainment events and to optimally allocate resources, and (4) develop optimal pricing policies to maximize revenues.
- **Homeland security and law enforcement.** Data mining has a number of homeland security and law enforcement applications. Data mining is often used to (1) identify patterns of terrorist behaviors, (2) discover crime patterns (e.g., locations, timings, criminal behaviors, and other related attributes) to help solve criminal cases in a timely manner, (3) predict and eliminate potential biological and chemical attacks to the nation's critical infrastructure by analyzing special-purpose sensory data, and (4) identify and stop malicious attacks on critical information infrastructures (often called *information warfare*).
- **Sports.** Data mining was used to improve the performance of National Basketball Association (NBA) teams in the United States. Major League Baseball teams are into predictive analytics and data mining to optimally utilize their limited resources for a winning season (see Moneyball article in Chapter 1). In fact, most, if not all, professional sports nowadays employ data crunchers and use data mining to increase their chances of winning. Data mining applications are not limited to professional sports. In a 2012 article, Delen, Cogdell, and Kasap (2012) developed data mining models to predict National Collegiate Athletic Association (NCAA) Bowl Game outcomes using a wide range of variables about the two opposing teams' previous games statistics (more details about this case study are provided in Chapter 2). Wright (2012) used a variety of predictors for examination of the NCAA men's basketball championship bracket (a.k.a. March Madness).

SECTION 5.3 REVIEW QUESTIONS

1. What are the major application areas for data mining?
2. Identify at least five specific applications of data mining and list five common characteristics of these applications.
3. What do you think is the most prominent application area for data mining? Why?
4. Can you think of other application areas for data mining not discussed in this section? Explain.

5.4 DATA MINING PROCESS

To systematically carry out data mining projects, a general process is usually followed. Based on best practices, data mining researchers and practitioners have proposed several processes (workflows or simple step-by-step approaches) to maximize the chances of success in conducting data mining projects. These efforts have led to several standardized processes, some of which (a few of the most popular ones) are described in this section.

One such standardized process, arguably the most popular one, Cross-Industry Standard Process for Data Mining—**CRISP-DM**—was proposed in the mid-1990s by a European consortium of companies to serve as a nonproprietary standard methodology for data mining (CRISP-DM, 2013 and Delen, 2014). Figure 5.3 illustrates this proposed process, which is a sequence of six steps that starts with a good understanding of the business and the need for the data mining project (i.e., the application domain) and ends with the deployment of the solution that satisfies the specific business need. Even though these steps are sequential in nature, there is usually a great deal of backtracking. Because the data mining is driven by experience and experimentation, depending on the problem

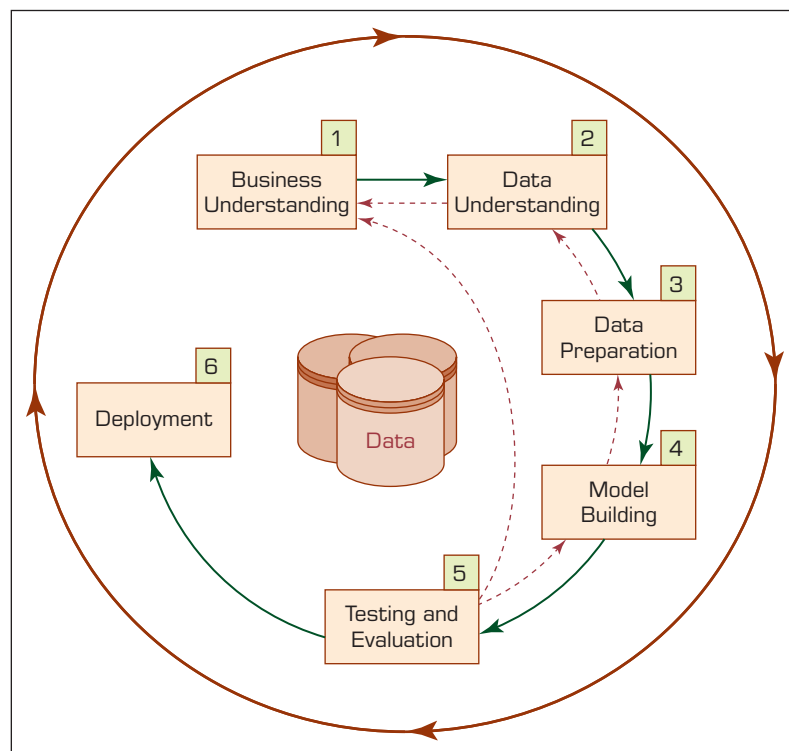


FIGURE 5.3 The Six-Step CRISP-DM Data Mining Process.

situation and the knowledge/experience of the analyst, the whole process can be very iterative (i.e., one should expect to go back and forth through the steps quite a few times) and time-consuming. Because later steps are built on the outcomes of the former ones, one should pay extra attention to the earlier steps in order not to put the whole study on an incorrect path from the onset.

Step 1: Business Understanding

The key element of any data mining study is to know what the study is for. Answering such a question begins with a thorough understanding of the managerial need for new knowledge and an explicit specification of the business objective regarding the study to be conducted. Specific goals such as “What are the common characteristics of the customers we have lost to our competitors recently?” or “What are typical profiles of our customers, and how much value does each of them provide to us?” are needed. Then a project plan for finding such knowledge is developed that specifies the people responsible for collecting the data, analyzing the data, and reporting the findings. At this early stage, a budget to support the study should also be established, at least at a high level with rough numbers.

Step 2: Data Understanding

A data mining study is specific to addressing a well-defined business task, and different business tasks require different sets of data. Following the business understanding, the main activity of the data mining process is to identify the relevant data from many available databases. Some key points must be considered in the data identification and selection phase. First and foremost, the analyst should be clear and concise about the description of the data mining task so that the most relevant data can be identified. For example, a retail data mining project may seek to identify spending behaviors of shoppers who purchase seasonal clothes based on their demographics, credit card transactions, and socioeconomic attributes. Furthermore, the analyst should build an intimate understanding of the data sources (e.g., where the relevant data are stored and in what form; what the process of collecting the data is—automated versus manual; who the collectors of the data are and how often the data are updated) and the variables (e.g., What are the most relevant variables? Are there any synonymous and/or homonymous variables? Are the variables independent of each other—do they stand as a complete information source without overlapping or conflicting information?).

To better understand the data, the analyst often uses a variety of statistical and graphical techniques, such as simple statistical summaries of each variable (e.g., for numeric variables the average, minimum/maximum, median, and standard deviation are among the calculated measures, whereas for categorical variables the mode and frequency tables are calculated), correlation analysis, scatterplots, histograms, and box plots. A careful identification and selection of data sources and the most relevant variables can make it easier for data mining algorithms to quickly discover useful knowledge patterns.

Data sources for data selection can vary. Traditionally, data sources for business applications include demographic data (such as income, education, number of households, and age), sociographic data (such as hobby, club membership, and entertainment), transactional data (sales record, credit card spending, issued checks), and so on. Nowadays, data sources also use external (open or commercial) data repositories, social media, and machine-generated data.

Data can be categorized as quantitative and qualitative. Quantitative data are measured using numeric values, or **numeric data**. It can be discrete (such as integers) or continuous (such as real numbers). Qualitative data, also known as categorical data, contains both nominal and ordinal data. **Nominal data** has finite nonordered values (e.g., gender data, which may have several values). **Ordinal data** has finite ordered values. For

example, customer credit ratings are considered ordinal data because the ratings can be excellent, fair, and bad. A simple taxonomy of data (i.e., the nature of data) is provided in Chapter 2.

Quantitative data can be readily represented by some sort of probability distribution. A probability distribution describes how the data are dispersed and shaped. For instance, normally distributed data are symmetric and is commonly referred to as being a bell-shaped curve. Qualitative data may be coded to numbers and then described by frequency distributions. Once the relevant data are selected according to the data mining business objective, data preprocessing should be pursued.

Step 3: Data Preparation

The purpose of data preparation (more commonly called *data preprocessing*) is to take the data identified in the previous step and prepare it for analysis by data mining methods. Compared to the other steps in CRISP-DM, data preprocessing consumes the most time and effort; most believe that this step accounts for roughly 80% of the total time spent on a data mining project. The reason for such an enormous effort spent on this step is the fact that real-world data are generally incomplete (lacking attribute values, lacking certain attributes of interest, or containing only aggregate data), noisy (containing errors or outliers), and inconsistent (containing discrepancies in codes or names). The nature of the data and the issues related to preprocessing of data for analytics are explained in detail in Chapter 2.

Step 4: Model Building

In this step, various modeling techniques are selected and applied to an already prepared data set to address the specific business need. The model-building step also encompasses the assessment and comparative analysis of the various models built. Because there is not a universally known *best* method or algorithm for a data mining task, one should use a variety of viable model types along with a well-defined experimentation and assessment strategy to identify the “best” method for a given purpose. Even for a single method or algorithm, a number of parameters need to be calibrated to obtain optimal results. Some methods may have specific requirements in the way that the data are to be formatted; thus, stepping back to the data preparation step is often necessary. Analytics in Action 5.1 presents a research study where a number of model types are developed and compared to each other.

Depending on the business need, the data mining task can be of a prediction (either classification or regression), an association, or a clustering type. Each of these data mining tasks can use a variety of data mining methods and algorithms. Some of these data mining methods were explained earlier in this chapter, and some of the most popular algorithms, including decision trees for classification, *k*-means for clustering, and the Apriori algorithm for association rule mining, are described later in this chapter.

ANALYTICS IN ACTION 5.1 Data Mining Helps in Cancer Research

According to the American Cancer Society, half of all men and one-third of all women in the United States will develop cancer during their lifetimes; approximately 1.5 million new cancer cases were expected to be diagnosed in 2013. Cancer is the second-most-common cause of death in the United States and in the world, exceeded only by cardiovascular disease. This year, over 500,000 Americans are expected to die of cancer—more than 1,300 people a day—accounting for nearly one of every four deaths.

Cancer is a group of diseases generally characterized by uncontrolled growth and spread of abnormal cells. If the growth and/or spread are not controlled, it can result in death. Even though the exact reasons are not known, cancer is believed to be caused by both external factors (e.g., tobacco, infectious organisms, chemicals, and radiation) and internal factors (e.g.,

inherited mutations, hormones, immune conditions, and mutations that occur from metabolism). These causal factors may act together or in sequence to initiate or promote carcinogenesis. Cancer is treated with surgery, radiation, chemotherapy, hormone therapy, biological therapy, and targeted therapy. Survival statistics vary greatly by cancer type and stage at diagnosis.

The five-year relative survival rate for all cancers is improving, and decline in cancer mortality had reached 20%, translating into the avoidance of about 1.2 million deaths from cancer since 1991. That's more than 400 lives saved per day! The improvement in survival reflects progress in diagnosing certain cancers at an earlier stage and improvements in treatment. Further improvements are needed to prevent and treat cancer.

Even though cancer research has traditionally been clinical and biological in nature, in recent years data-driven analytic studies have become a common complement. In medical domains where data- and analytics-driven research have been applied successfully, novel research directions have been identified to further advance the clinical and biological studies. Using various types of data, including molecular, clinical, literature-based, and clinical trial data, along with suitable data mining tools and techniques, researchers have been able to identify novel patterns, paving the road toward a cancer-free society.

In one study, Delen (2009) used three popular data mining techniques (decision trees, artificial neural networks, and SVMs) in conjunction with logistic regression to develop prediction models for prostate cancer survivability. The data set contained around 120,000 records and 77 variables. A k -fold cross-validation methodology was used in model building, evaluation, and comparison. The results showed that support vector models are the most accurate predictor (with a test set accuracy of 92.85%) for this domain, followed by artificial neural networks and decision trees. Furthermore, using a sensitivity analysis-based evaluation method, the study also revealed novel patterns related to prognostic factors of prostate cancer.

In a related study, Delen, Walker, and Kadam (2005) used two data mining algorithms (artificial neural networks and decision trees) and logistic regression to develop prediction models for breast cancer survival using a large data set (more than 200,000 cases). Using a 10-fold cross-validation method to measure the unbiased estimate of the prediction models for performance comparison purposes, the results indicated that the decision tree (C5 algorithm) was the best predictor, with 93.6% accuracy on the holdout sample (which was the best prediction accuracy reported in the literature), followed by artificial neural networks, with 91.2% accuracy, and logistic regression, with 89.2% accuracy. Further analysis of prediction models revealed prioritized importance of the prognostic factors, which can then be used as a basis for further clinical and biological research studies.

In the most recent study, Zolbanin, Delen, and Zadeh (2015) studied the impact of comorbidity in cancer survivability. Although prior research has shown that diagnostic and treatment recommendations might be altered based on the severity of comorbidities, chronic diseases are still being investigated in isolation from one another in most cases. To illustrate the significance of concurrent chronic diseases in the course of treatment, their study used the Surveillance, Epidemiology, and End Results (SEER) Program's cancer data to create two comorbid data sets: one for breast and female genital cancers and another for prostate and urinal cancers. Several popular machine-learning techniques are then applied to the resultant data sets to build predictive models (see Figure 5.4). Comparison of the results has shown that having more information about comorbid conditions of patients can improve models' predictive power, which in turn can help practitioners make better diagnostic and treatment decisions. Therefore, the study suggested that proper identification, recording, and use of patients' comorbidity status can potentially lower treatment costs and ease the healthcare-related economic challenges.

These examples (among many others in the medical literature) show that advanced data mining techniques can be used to develop models that possess a high degree of predictive as well as explanatory power. Although data mining methods are capable of extracting patterns and relationships hidden deep in large and complex medical databases, without the cooperation and feedback from the medical experts, their results are not of much use. The patterns found via data mining methods should be evaluated by medical professionals who have years of experience in the problem domain to decide whether they are logical, actionable, and novel enough to warrant new research directions. In short, data mining is not meant to replace medical professionals and researchers, but to complement their invaluable efforts to provide data-driven new research directions and to ultimately save more human lives.

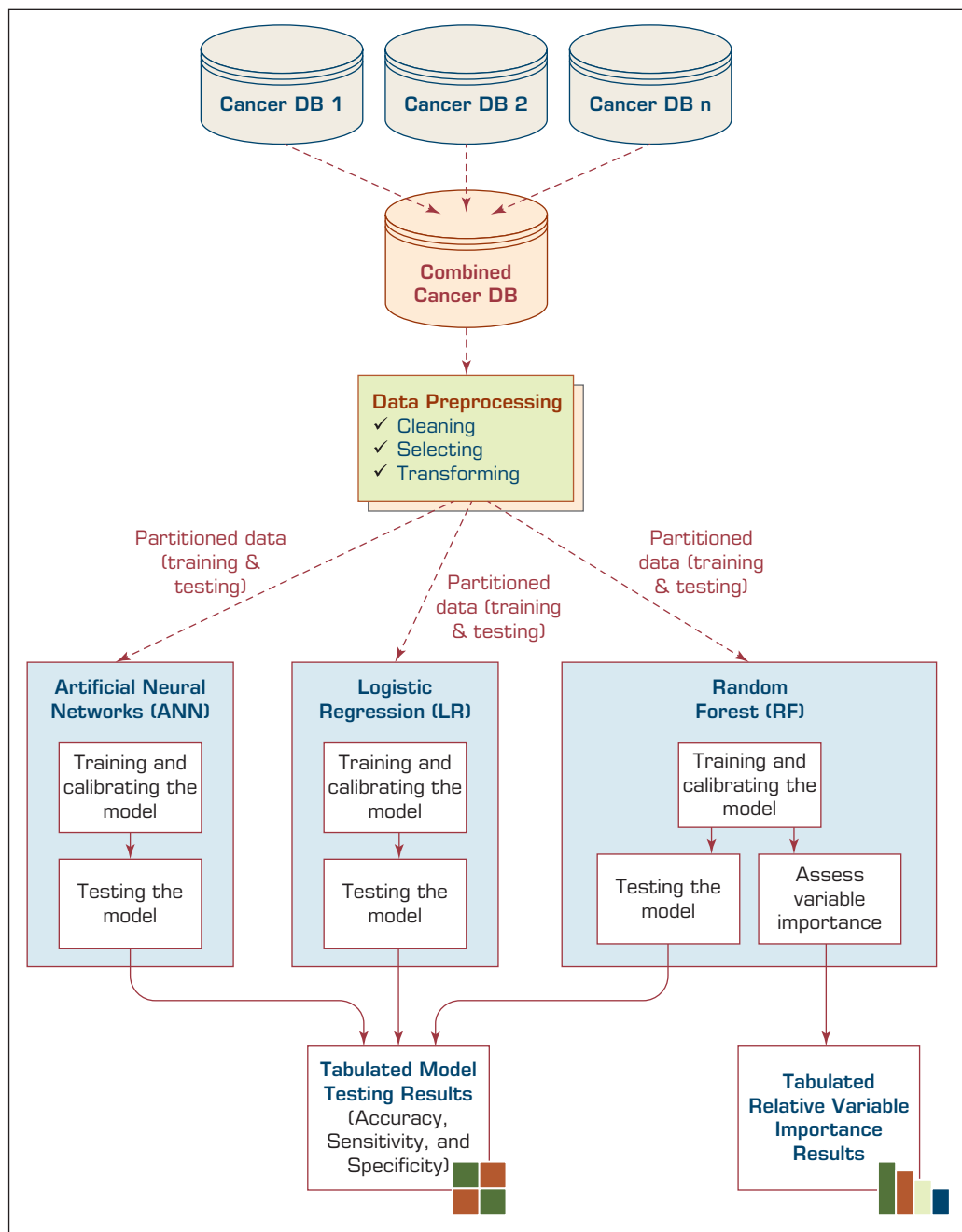


FIGURE 5.4 A Data Mining Methodology for Investigation of Comorbidity in Cancer Survivability.

Sources: Zolbanin, H. M., Delen, D., & Zadeh, A. H. (2015). Predicting overall survivability in comorbidity of cancers: A data mining approach. *Decision Support Systems*, 74, 150–161; Delen, D. (2009). Analysis of cancer data: A data mining approach. *Expert Systems*, 26(1), 100–112; Thongkam, J., Xu, G., Zhang, Y., & Huang, F. (2009). Toward breast cancer survivability prediction models through improving training space. *Expert Systems with Applications*, 36(10), 12200–12209; Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine*, 34(2), 113–127.

Step 5: Testing and Evaluation

In step 5, the developed models are assessed and evaluated for their accuracy and generality. This step assesses the degree to which the selected model (or models) meets the business objectives and, if so, to what extent (i.e., Do more models need to be developed and assessed?). Another option is to test the developed model(s) in a real-world scenario if time and budget constraints permit. Even though the outcome of the developed models is expected to relate to the original business objectives, other findings that are not necessarily related to the original business objectives but that might also unveil additional information or hints for future directions often are discovered.

The testing and evaluation step is a critical and challenging task. No value is added by the data mining task until the business value obtained from discovered knowledge patterns is identified and recognized. Determining the business value from discovered knowledge patterns is somewhat similar to playing with puzzles. The extracted knowledge patterns are pieces of the puzzle that need to be put together in the context of the specific business purpose. The success of this identification operation depends on the interaction among data analysts, business analysts, and decision makers (such as business managers). Because data analysts may not have the full understanding of the data mining objectives and what they mean to the business and the business analysts, and decision makers may not have the technical knowledge to interpret the results of sophisticated mathematical solutions, interaction among them is necessary. To properly interpret knowledge patterns, it is often necessary to use a variety of tabulation and visualization techniques (e.g., pivot tables, cross-tabulation of findings, pie charts, histograms, box plots, scatterplots).

Step 6: Deployment

Development and assessment of the models is not the end of the data mining project. Even if the purpose of the model is to have a simple exploration of the data, the knowledge gained from such exploration will need to be organized and presented in a way that the end user can understand and benefit from. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases, it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will not carry out the deployment effort, it is important for the customer to understand up front what actions need to be carried out to actually make use of the created models.

The deployment step may also include maintenance activities for the deployed models. Because everything about the business is constantly changing, the data that reflect the business activities also are changing. Over time, the models (and the patterns embedded within them) built on the old data may become obsolete, irrelevant, or misleading. Therefore, monitoring and maintenance of the models are important if the data mining results are to become a part of the day-to-day business and its environment. A careful preparation of a maintenance strategy helps to avoid unnecessarily long periods of incorrect usage of data mining results. To monitor the deployment of the data mining result(s), the project needs a detailed plan on the monitoring process, which may not be a trivial task for complex data mining models.

Other Data Mining Standardized Processes and Methodologies

To be applied successfully, a data mining study must be viewed as a process that follows a standardized methodology rather than as a set of automated software tools and techniques. In addition to CRISP-DM, there is another well-known methodology developed by the SAS Institute, called SEMMA (2009). The acronym **SEMMA** stands for “sample, explore, modify, model, and assess.”

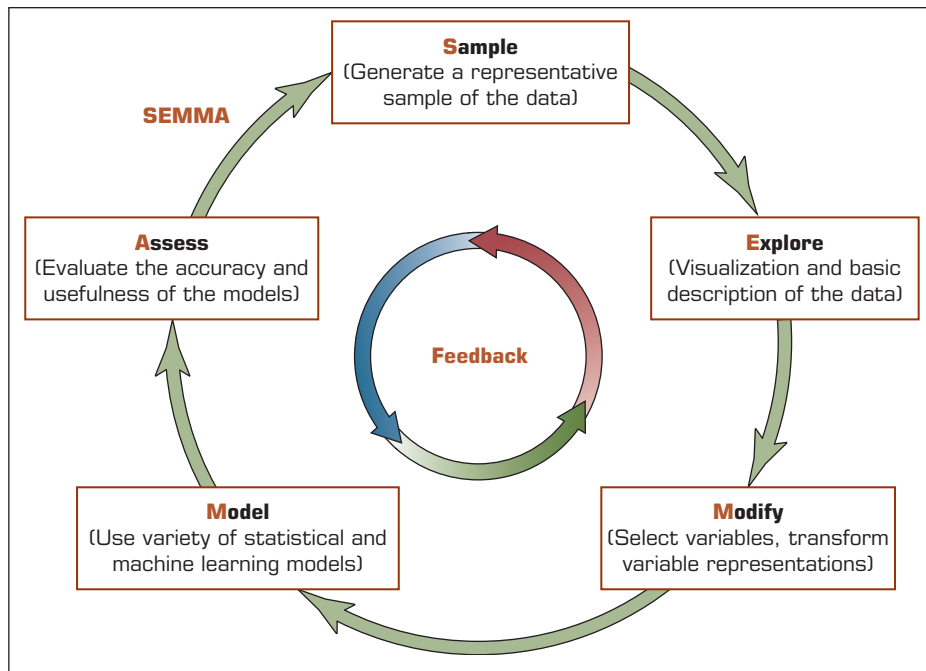


FIGURE 5.5 SEMMA Data Mining Process.

Beginning with a statistically representative sample of the data, SEMMA makes it easy to apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and confirm a model's accuracy. A pictorial representation of SEMMA is given in Figure 5.5.

By assessing the outcome of each stage in the SEMMA process, the model developer can determine how to model new questions raised by the previous results and thus proceed back to the exploration phase for additional refinement of the data; that is, as with CRISP-DM, SEMMA is driven by a highly iterative experimentation cycle. The main difference between CRISP-DM and SEMMA is that CRISP-DM takes a more comprehensive approach—including understanding of the business and the relevant data—to data mining projects, whereas SEMMA implicitly assumes that the data mining project's goals and objectives along with the appropriate data sources have been identified and understood.

Some practitioners commonly use the term **knowledge discovery in databases (KDD)** as a synonym for data mining. Fayyad et al. (1996) defined *knowledge discovery in databases* as a process of using data mining methods to find useful information and patterns in the data, as opposed to data mining, which involves using algorithms to identify patterns in data derived through the KDD process (see Figure 5.6). KDD is a comprehensive process that encompasses data mining. The input to the KDD process consists of organizational data. The enterprise data warehouse enables KDD to be implemented efficiently because it provides a single source for data to be mined. Dunham (2003) summarized the KDD process as consisting of the following steps: data selection, data preprocessing, data transformation, data mining, and interpretation/evaluation. KDD process is often criticized by analytics professionals because it portrays data mining as a single-step in the overall process, as opposed to the complete process itself.

Figure 5.7 shows the polling results for the question, “What main methodology are you using for data mining?” (conducted by KDnuggets.com in August 2007).

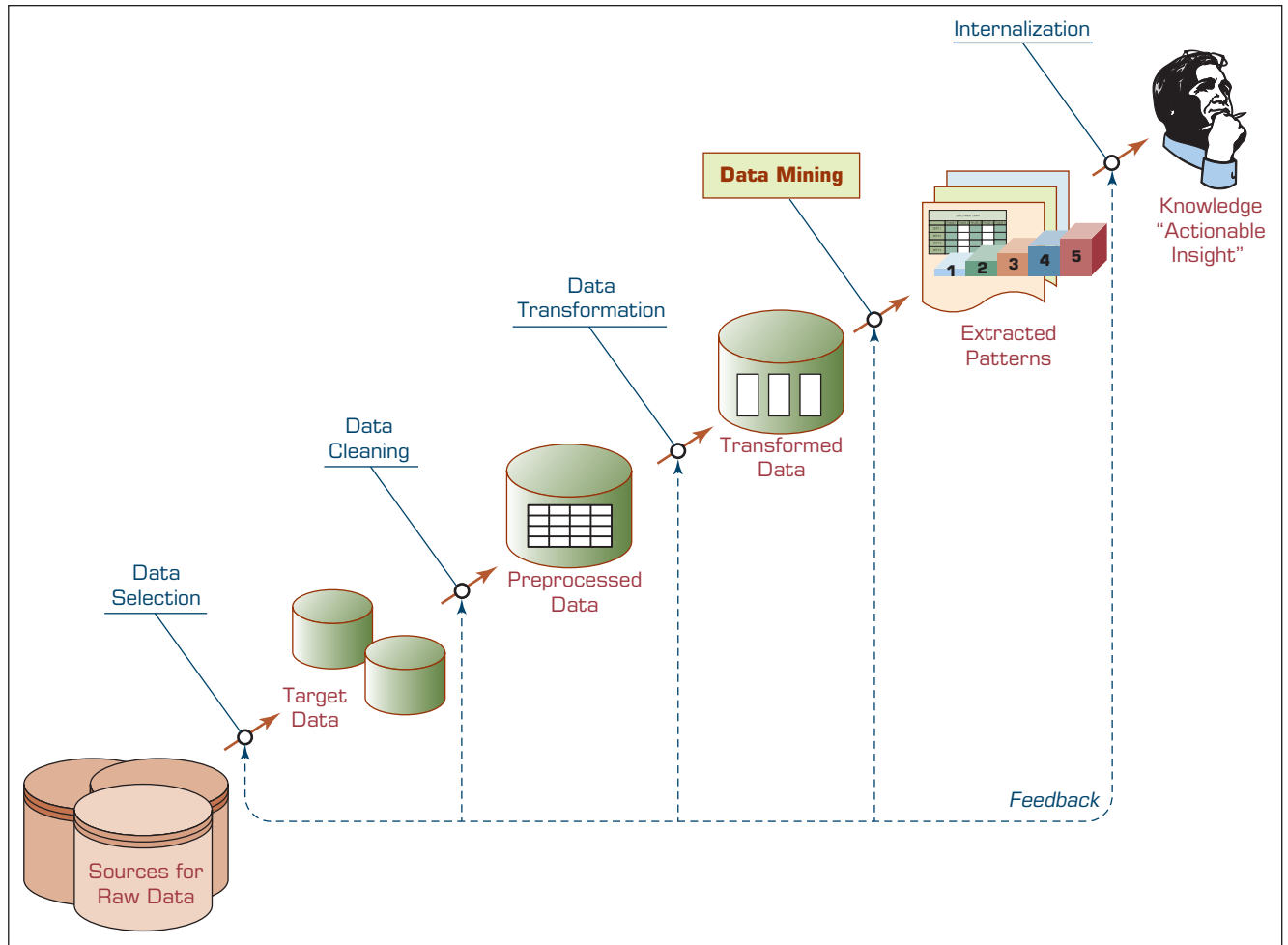


FIGURE 5.6 KDD (Knowledge Discovery in Databases) Process.

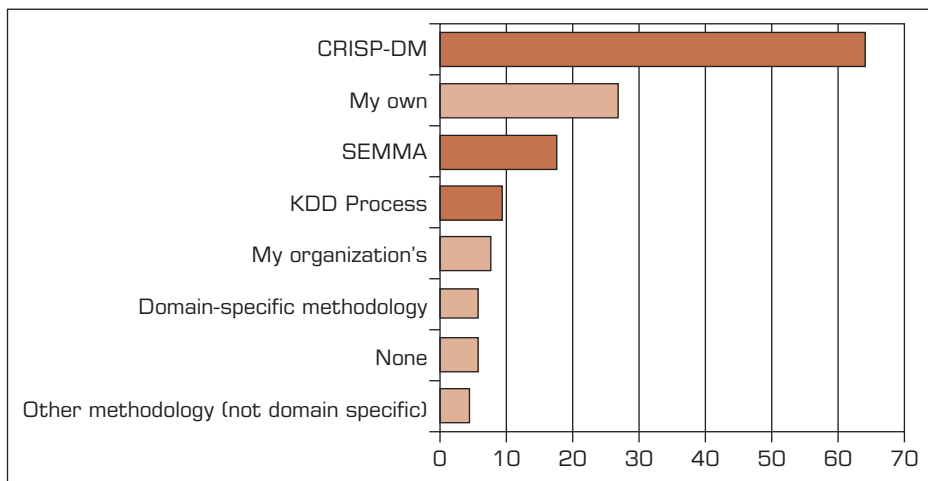


FIGURE 5.7 Ranking of Data Mining Methodologies/Processes. *Source:* Used with permission from KDnuggets.com.

► SECTION 5.4 REVIEW QUESTIONS

1. What are the major data mining processes?
2. Why do you think the early phases (understanding of the business and understanding of the data) take the longest in data mining projects?
3. List and briefly define the phases in the CRISP-DM process.
4. What are the main data preprocessing steps? Briefly describe each step, and provide relevant examples.
5. How does CRISP-DM differ from SEMMA?

5.5 DATA MINING METHODS

A variety of methods are available for performing data mining studies, including classification, regression, clustering, and association. Most data mining software tools employ more than one technique (or algorithm) for each of these methods. This section describes the most popular data mining methods and explains their representative techniques.

Classification

Classification is perhaps the most frequently used data mining method for real-world problems. As a popular member of the machine-learning family of techniques, classification learns patterns from past data (a set of information—traits, variables, features—on characteristics of the previously labeled items, objects, or events) to place new instances (with unknown labels) into their respective groups or classes. For example, one could use classification to predict whether the weather on a particular day will be “sunny,” “rainy,” or “cloudy.” Popular classification tasks include credit approval (i.e., good or bad credit risk), store location (e.g., good, moderate, bad), target marketing (e.g., likely customer, no hope), fraud detection (i.e., yes/no), and telecommunication (e.g., likely to turn to another phone company, yes/no). If what is being predicted is a class label (e.g., “sunny,” “rainy,” or “cloudy”), the prediction problem is called a classification, whereas if it is a numeric value (e.g., temperature, such as 68°F), the prediction problem is called a **regression**.

Even though clustering (another popular data mining method) can also be used to determine groups (or class memberships) of things, there is a significant difference between the two. Classification learns the function between the characteristics of things (i.e., independent variables) and their membership (i.e., output variable) through a supervised learning process where both types (input and output) of variables are presented to the algorithm; in clustering, the membership of the objects is learned through an unsupervised learning process where only the input variables are presented to the algorithm. Unlike classification, clustering does not have a supervising (or controlling) mechanism that enforces the learning process; instead, clustering algorithms use one or more heuristics (e.g., multidimensional distance measure) to discover natural groupings of objects.

The most common two-step methodology of classification-type prediction involves model development/training and model testing/deployment. In the model development phase, a collection of input data, including the actual class labels, is used. After a model has been trained, the model is tested against the holdout sample for accuracy assessment and eventually deployed for actual use where it is to predict classes of new data instances (where the class label is unknown). Several factors are considered in assessing the model, including the following.

- **Predictive accuracy.** The model’s ability to correctly predict the class label of new or previously unseen data. Prediction accuracy is the most commonly used assessment factor for classification models. To compute this measure, actual class labels of a test data set are matched against the class labels predicted by the model. The accuracy can then be computed as the *accuracy rate*, which is the percentage

of test data set samples correctly classified by the model (more on this topic is provided later in the chapter).

- **Speed.** The computational costs involved in generating and using the model, where faster is deemed to be better.
- **Robustness.** The model's ability to make reasonably accurate predictions, given noisy data or data with missing and erroneous values.
- **Scalability.** The ability to construct a prediction model efficiently given a rather large amount of data.
- **Interpretability.** The level of understanding and insight provided by the model (e.g., how and/or what the model concludes on certain predictions).

Estimating the True Accuracy of Classification Models

In classification problems, the primary source for accuracy estimation is the *confusion matrix* (also called a *classification matrix* or a *contingency table*). Figure 5.8 shows a confusion matrix for a two-class classification problem. The numbers along the diagonal from the upper left to the lower right represent correct decisions (i.e., classifications), and the numbers outside this diagonal represent the errors (i.e., misclassifications).

Using the information tabulated in the confusion matrix, Table 5.1 provides equations used as accuracy metrics for binary classification models. Among the ones listed, the first three, namely Accuracy, True Positive Rate, and True Negative Rate are the most commonly used ones.

When the classification problem is not binary, the confusion matrix gets bigger (a square matrix with the size of the unique number of class labels), and accuracy metrics become limited to *per class accuracy rates* (e.g., True Classification Accuracy for Class A) and the *overall classifier accuracy* (e.g., Overall Classification Accuracy for All Classes), as shown in the following formulations.

$$\text{Overall Classification Accuracy for All Classes} = \frac{\text{Correct Prediction of All Classes}}{\text{Total Number of Cases for All Classes}}$$

Estimating the accuracy of a classification model (or classifier) induced by a supervised learning algorithm is important for the following two reasons: First, it can be used to estimate its future prediction accuracy, which could imply the level of confidence one should have in the classifier's output in the prediction system. Second, it can be used for choosing a classifier from a given set (identifying the “best” classification model among

		Predicted Class	
		Positive	Negative
True/Observed Class	Positive	True Positive Count (TP)	False Negative Count (FN)
	Negative	False Positive Count (FP)	True Negative Count (TN)

FIGURE 5.8 A Simple Confusion Matrix for Tabulation of Two-Class Classification Results.

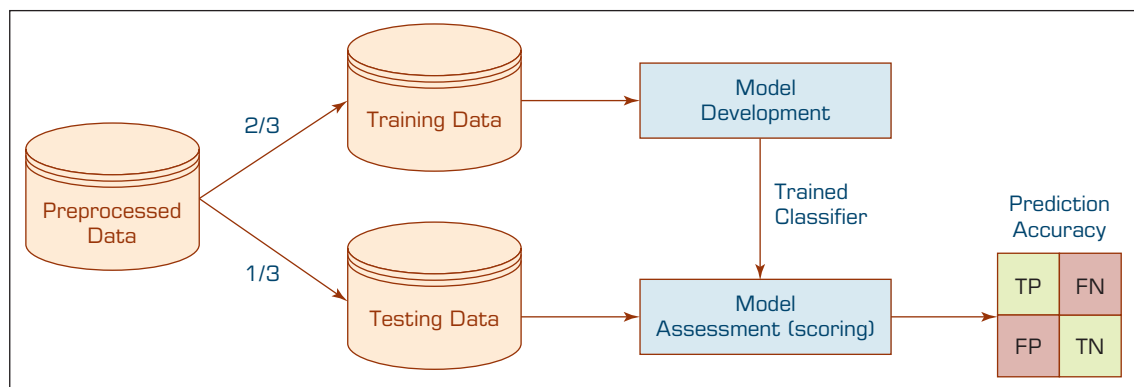
TABLE 5.1 Common Accuracy Metrics for Classification Models

Metric	Description
Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$	The ratio of correctly classified instances (positives and negative) divided by the total numbers of instances
True Positive Rate = $\frac{TP}{TP + FN}$	(a.k.a Sensitivity) The ratio of correctly classified positives divided by the total positive count (i.e., hit rate or recall)
True Negative Rate = $\frac{TN}{TN + FP}$	(a.k.a Specificity) The ratio of correctly classified negatives divided by the total negative count (i.e., false alarm rate)
Precision = $\frac{TP}{TP + FP}$	The ratio of correctly classified positives divided by the sum of correctly classified positives and incorrectly classified positives
Recall = $\frac{TP}{TP + FN}$	Ratio of correctly classified positives divided by the sum of correctly classified positives and incorrectly classified negatives

the many trained). The following are among the most popular estimation methodologies used for classification-type data mining models.

SIMPLE SPLIT The **simple split** (or holdout or test sample estimation) partitions the data into two mutually exclusive subsets called a *training set* and a *test set* (or *holdout set*). It is common to designate two-thirds of the data as the training set and the remaining one-third as the test set. The training set is used by the inducer (model builder), and the built classifier is then tested on the test set. An exception to this rule occurs when the classifier is an artificial neural network. In this case, the data are partitioned into three mutually exclusive subsets: training, validation, and testing. The validation set is used during model building to prevent overfitting. Figure 5.9 shows the simple split methodology.

The main criticism of this method is that it makes the assumption that the data in the two subsets are of the same kind (i.e., have the exact same properties). Because this is a simple random partitioning, in most realistic data sets where the data are skewed on the classification variable, such an assumption may not hold true. To improve this situation, stratified sampling is suggested, where the strata become the output variable. Even

**FIGURE 5.9** Simple Random Data Splitting.

though this is an improvement over the simple split, it still has a bias associated from the single random partitioning.

K-FOLD CROSS-VALIDATION To minimize the bias associated with the random sampling of the training and holdout data samples in comparing the predictive accuracy of two or more methods, one can use a methodology called **k-fold cross-validation**. In *k*-fold cross-validation, also called *rotation estimation*, the complete data set is randomly split into *k* mutually exclusive subsets of approximately equal size, often using the stratified sampling technique to produce unbiased sample for each fold. The classification model is trained and tested *k* times. Each time it is trained on all but one fold and then tested on the remaining single fold. The cross-validation estimate of the overall accuracy of a model is calculated by simply averaging the *k* individual accuracy measures, as shown in the following equation:

$$CVA = \frac{1}{k} \sum_{i=1}^k A_i$$

where *CVA* stands for cross-validation accuracy, *k* is the number of folds used, and *A* is the accuracy measure (e.g., hit rate, sensitivity, specificity) of each fold. Figure 5.10 shows a graphical illustration of *k*-fold cross-validation where *k* is set to 10.

ADDITIONAL CLASSIFICATION ASSESSMENT METHODOLOGIES Other popular assessment methodologies include the following:

- **Leave-one-out.** The leave-one-out method is similar to the *k*-fold cross-validation where the *k* takes the value of 1; that is, every data point is used for testing once on as many models developed as there are number of data points. This is a time-consuming methodology, but sometimes for small data sets, it is a viable option.
- **Bootstrapping.** With **bootstrapping**, a fixed number of instances from the original data are sampled (with replacement) for training, and the rest of the data set is used for testing. This process is repeated as many times as desired.
- **Jackknifing.** Though similar to the leave-one-out methodology, with jackknifing the accuracy is calculated by leaving one sample out at each iteration of the estimation process.
- **Area under the ROC curve.** The **area under the ROC curve** is a graphical assessment technique where the true positive rate is plotted on the *y*-axis and the false positive rate is plotted on the *x*-axis. The area under the ROC curve (of ten denoted

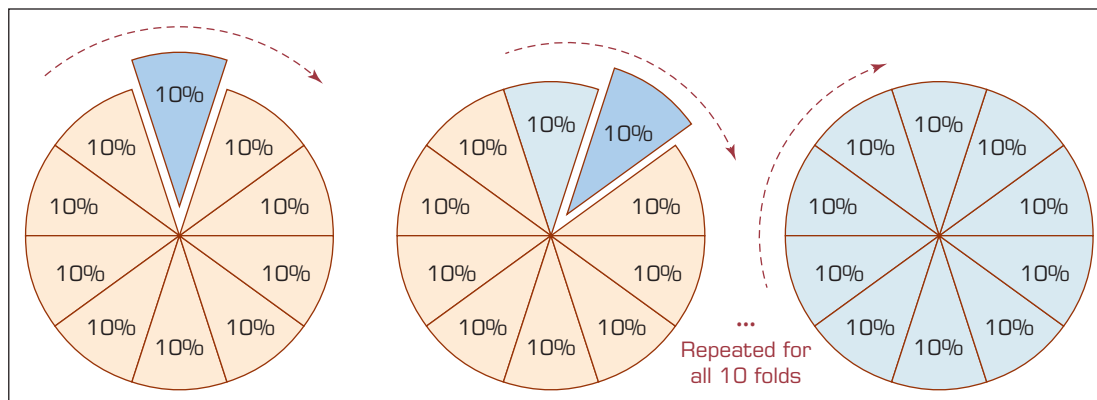


FIGURE 5.10 A Graphical Depiction of *k*-Fold Cross-Validation.

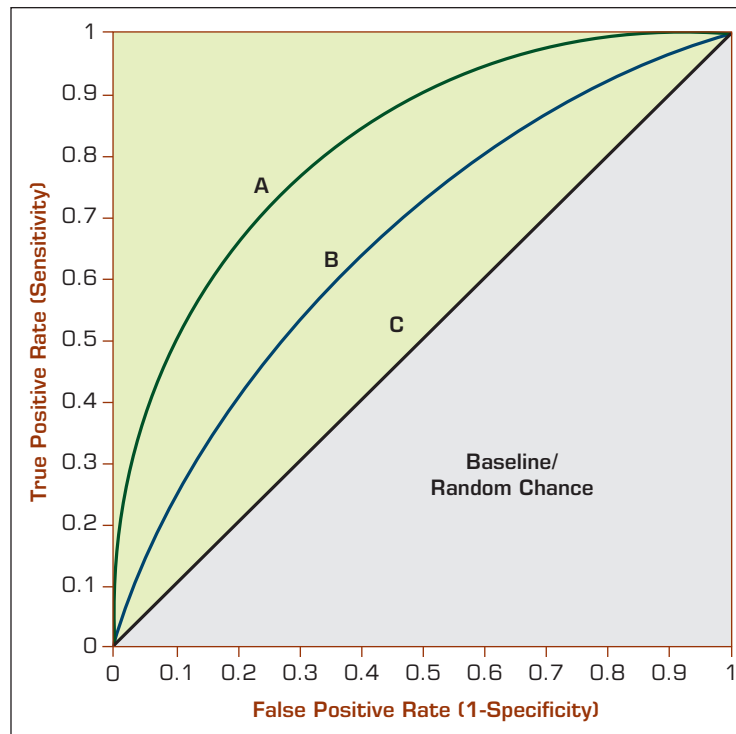


FIGURE 5.11 A Sample ROC Curve.

as AUC) determines the predictive power of a classifier: A value of 1 indicates a perfect classifier, whereas 0.5 indicates no better than random chance; in reality, the values would range between the two extreme cases. For example in Figure 5.11, *A* has a better overall classification performance than *B*, whereas *C* is not any better than the baseline, which is the same as random chance (a value of 0.5).

CLASSIFICATION TECHNIQUES A number of techniques (or algorithms) are used for classification modeling, including the following:

- **Decision tree analysis.** Decision tree analysis (a machine-learning technique) is arguably the most popular classification technique in the data mining arena. A detailed description of this technique is given in the following section.
- **Statistical analysis.** Statistical techniques were the primary classification algorithm for many years until the emergence of machine-learning techniques. Statistical classification techniques include logistic regression and discriminant analysis, both of which make the assumptions that the relationships between the input and output variables are linear in nature, the data are normally distributed, and the variables are not correlated and are independent of each other. The questionable nature of these assumptions has led to the shift toward machine-learning techniques.
- **Neural networks.** These are among the most popular machine-learning techniques that can be used for classification-type problems.
- **Case-based reasoning.** This approach uses historical cases to recognize commonalities to assign a new case into the most probable category.
- **Bayesian classifiers.** This approach uses probability theory to build classification models based on the past occurrences that are capable of placing a new instance into a most probable class (or category).

- **Genetic algorithms.** The use of the analogy of natural evolution to build directed-search-based mechanisms to classify data samples.
- **Rough sets.** This method takes into account the partial membership of class labels to predefined categories in building models (collection of rules) for classification problems.

A complete description of all of these classification techniques is beyond the scope of this book; thus, only several of the most popular ones are presented here.

ENSEMBLE MODELS FOR BETTER PREDICTIVE ANALYTICS Creating ensembles is essentially the process of intelligently combining the information (forecasts or predictions) created and provided by two or more information sources (i.e., prediction models). While there is an ongoing debate about the sophistication level of the ensemble methods to be employed, there is a general consensus that ensemble models produce more robust and reliable information for business decisions (Seni & Elder, 2010). That is, combining forecasts can (and often will) improve accuracy and robustness of information outcomes, while reducing uncertainty and bias associated with individual models.

As we all know, in data mining and prediction modeling there is not a universally accepted “best model” that works for any problem. The best model depends on the scenario being analyzed and the data set being used and can only be obtained through an extensive trial-and-error experimentation (and only happens if time and resources permit). Just as there is not a single best model, there is also not a single best implementation of different model types—for instance decision trees, neural networks, and support vector machines have different architectures and parameter sets that need to be “optimized” for the best possible results. Data scientists are developing new ways to improve the accuracy and efficiency of today’s prediction models. One proven way to do so has been combining the outputs of prediction models into a single composite score, which is a model ensemble. Ensembles have been the winners of many data mining and prediction modeling competitions in recent years (see kaggle.org for a list of recent predictive analytics competitions and winners).

Ensemble models can be categorized as either homogeneous or heterogamous (Abbott 2014, p. 307). As the name implies, homogeneous model ensembles combine the outcomes of two or more of the same type of models such as decision trees. In fact, a vast majority of homogeneous model ensembles are developed using a combination of decision tree structures. The two common categories of decision tree ensembles are bagging and boosting. A well-recognized and hugely successful example of bagging-type decision tree ensembles is called Random Forest—as opposed to building a large tree, Random Forest develops a forest of many small trees. And a good example of boosting-type decision tree ensembles is called AdaBoosting (a short name for “Adaptive Boosting”)—an algorithm that changes the weight (i.e., the importance or contribution) assigned to each data sample in each iteration in the learning process based on the misclassification outcomes, so that the accuracy of the classifier/predictor will be optimized for all class labels.

Heterogeneous model ensembles, again as the name implies, combines the outcomes of two or more different types of models such as decision trees, artificial neural networks, logistic regression, and support vector machines. One of the key success factors in ensemble modeling is to use models that are fundamentally different than one another, ones that look at the data from a different perspective. Because of the way it combines the outcomes of different models of different model types, heterogeneous model ensembles are also called information fusion models (Delen & Sharda, 2010). In the process of combining the outcomes of multiple models, either a simple voting (each model contributes equally, one vote) or a weighted combination of voting (each model is contributing based on its prediction accuracy—more accurate models are having higher weight value) can be used. Regardless of the combination method, ensembles have been

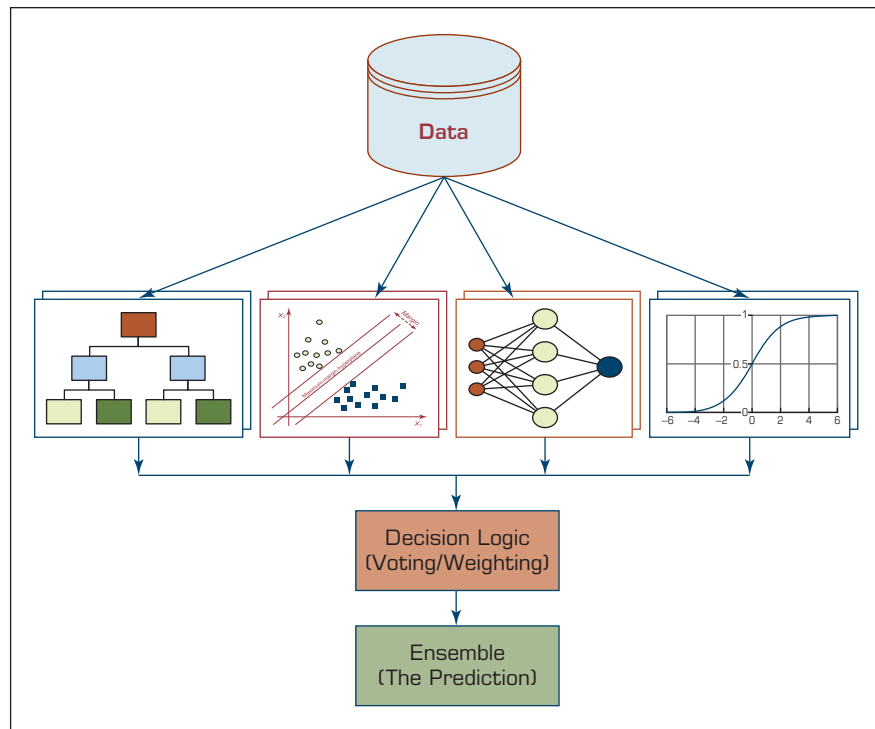


FIGURE 5.12 Graphical Illustration of a Heterogeneous Ensemble.

shown to be an invaluable addition to any data mining and predictive modeling project. Although ensemble modeling improves accuracy and robustness, on the negative side, it also increases the model complexity, and hence the lack of interpretability (i.e., transparency). Figure 5.12 graphically illustrates a heterogeneous model ensemble process.

DECISION TREES Before describing the details of **decision trees**, we need to discuss some simple terminology. First, decision trees include many input variables that may have an impact on the classification of different patterns. These input variables are usually called *attributes*. For example, if we were to build a model to classify loan risks on the basis of just two characteristics—income and a credit rating—these two characteristics would be the attributes, and the resulting output would be the *class label* (e.g., low, medium, or high risk). Second, a tree consists of branches and nodes. A *branch* represents the outcome of a test to classify a pattern (on the basis of a test) using one of the attributes. A *leaf node* at the end represents the final class choice for a pattern (a chain of branches from the root node to the leaf node, which can be represented as a complex if-then statement).

The basic idea behind a decision tree is that it recursively divides a training set until each division consists entirely or primarily of examples from one class. Each nonleaf node of the tree contains a *split point*, which is a test on one or more attributes and determines how the data are to be divided further. Decision tree algorithms, in general, build an initial tree from the training data such that each leaf node is pure, and they then prune the tree to increase its generalization, and hence, the prediction accuracy on test data.

In the growth phase, the tree is built by recursively dividing the data until each division is either pure (i.e., contains members of the same class) or relatively small. The basic idea is to ask questions whose answers would provide the most information, similar to what we may do when playing the game “Twenty Questions.”

The split used to partition the data depends on the type of the attribute used in the split. For a continuous attribute A , splits are of the form $\text{value}(A) < x$, where x is some “optimal” split value of A . For example, the split based on income could be “Income < 50,000.” For the categorical attribute A , splits are of the form $\text{value}(A)$ belongs to x , where x is a subset of A . As an example, the split could be on the basis of gender: “Male versus Female.”

A general algorithm for building a decision tree is as follows:

- Create a root node and assign all of the training data to it.
- Select the *best* splitting attribute.
- Add a branch to the root node for each value of the split. Split the data into mutually exclusive (nonoverlapping) subsets along the lines of the specific split and move to the branches.
- Repeat steps 2 and 3 for each and every leaf node until the stopping criteria is reached (e.g., the node is dominated by a single class label).

Many different algorithms have been proposed for creating decision trees. These algorithms differ primarily in terms of the way in which they determine the splitting attribute (and its split values), the order of splitting the attributes (splitting the same attribute only once or many times), the number of splits at each node (binary versus ternary), the stopping criteria, and the pruning of the tree (pre- versus postpruning). Some of the most well-known algorithms are ID3 (followed by C4.5 and C5 as the improved versions of ID3) from machine learning, classification and regression trees (CART) from statistics, and the chi-squared automatic interaction detector (CHAID) from pattern recognition.

When building a decision tree, the goal at each node is to determine the attribute and the split point of that attribute that best divides the training records to purify the class representation at that node. To evaluate the goodness of the split, some splitting indices have been proposed. Two of the most common ones are the Gini index and information gain. The Gini index is used in CART and SPRINT (Scalable PaRallelizable INduction of Decision Trees) algorithms. Versions of information gain are used in ID3 (and its newer versions, C4.5 and C5).

The **Gini index** has been used in economics to measure the diversity of a population. The same concept can be used to determine the purity of a specific class as a result of a decision to branch along a particular attribute or variable. The best split is the one that increases the purity of the sets resulting from a proposed split. Let us briefly look into a simple calculation of the Gini index.

If a data set S contains examples from n classes, the Gini index is defined as

$$gini(S) = 1 - \sum_{j=1}^n p_j^2$$

where p_j is a relative frequency of class j in S . If a data set S is split into two subsets, S_1 and S_2 , with sizes N_1 and N_2 , respectively, the Gini index of the split data contains examples from n classes, and the Gini index is defined as

$$gini_{split}(S) = \frac{N_1}{N} gini(S_1) + \frac{N_2}{N} gini(S_2)$$

The attribute/split combination that provides the smallest $gini_{split}(S)$ is chosen to split the node. In such a determination, one should enumerate all possible splitting points for each attribute.

Information gain is the splitting mechanism used in ID3, which is perhaps the most widely known decision tree algorithm. It was developed by Ross Quinlan in 1986, and since then he has evolved this algorithm into the C4.5 and C5 algorithms. The basic

idea behind ID3 (and its variants) is to use a concept called *entropy* in place of the Gini index. **Entropy** measures the extent of uncertainty or randomness in a data set. If all the data in a subset belong to just one class, there is no uncertainty or randomness in that data set, so the entropy is zero. The objective of this approach is to build subtrees so that the entropy of each final subset is zero (or close to zero). Let us also look at the calculation of the information gain.

Assume that there are two classes, P (positive) and N (negative). Let the set of examples S contain p counts of class P and n counts of class N . The amount of information needed to decide if an arbitrary example in S belongs to P or N is defined as

$$I(p, n) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}$$

Assume that using attribute A , the set S will be partitioned into sets $\{S_1, S_2, \dots, S_v\}$. If S_i contains p_i examples of P and n_i examples of N , the entropy, or the expected information needed to classify objects in all subtrees, S_i , is

$$E(A) = \sum_{i=1}^v \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

Then, the information that would be gained by branching on attribute A would be

$$Gain(A) = I(p, n) - E(A)$$

These calculations are repeated for each and every attribute, and the one with the highest information gain is selected as the splitting attribute. The basic ideas behind these splitting indices are rather similar to each other, but the specific algorithmic details vary. A detailed definition of the ID3 algorithm and its splitting mechanism can be found in Quinlan (1986).

Cluster Analysis for Data Mining

Cluster analysis is an essential data mining method for classifying items, events, or concepts into common groupings called *clusters*. The method is commonly used in biology, medicine, genetics, social network analysis, anthropology, archaeology, astronomy, character recognition, and even in management information systems (MIS) development. As data mining has increased in popularity, the underlying techniques have been applied to business, especially to marketing. Cluster analysis has been used extensively for fraud detection (both credit card and e-commerce fraud) and market segmentation of customers in contemporary CRM systems. More applications in business continue to be developed as the strength of cluster analysis is recognized and used.

Cluster analysis is an exploratory data analysis tool for solving classification problems. The objective is to sort cases (e.g., people, things, events) into groups, or clusters, so that the degree of association is strong among members of the same cluster and weak among members of different clusters. Each cluster describes the class to which its members belong. An obvious one-dimensional example of cluster analysis is to establish score ranges into which to assign class grades for a college class. This is similar to the cluster analysis problem that the U.S. Treasury faced when establishing new tax brackets in the 1980s. A fictional example of clustering occurs in J. K. Rowling's *Harry Potter* books. The Sorting Hat determines to which House (e.g., dormitory) to assign first-year students at the Hogwarts School. Another example involves determining how to seat guests at a wedding. As far as data mining goes, the importance of cluster analysis is that it may reveal associations and structures in data that were not previously apparent but are sensible and useful once found.

Cluster analysis results may be used to:

- Identify a classification scheme (e.g., types of customers)
- Suggest statistical models to describe populations
- Indicate rules for assigning new cases to classes for identification, targeting, and diagnostic purposes
- Provide measures of definition, size, and change in what were previously broad concepts
- Find typical cases to label and represent classes
- Decrease the size and complexity of the problem space for other data mining methods
- Identify outliers in a specific domain (e.g., rare-event detection)

DETERMINING THE OPTIMAL NUMBER OF CLUSTERS Clustering algorithms usually require one to specify the number of clusters to find. If this number is not known from prior knowledge, it should be chosen in some way. Unfortunately, there is not an optimal way of calculating what this number is supposed to be. Therefore, several different heuristic methods have been proposed. The following are among the most commonly referenced ones:

- Look at the percentage of variance explained as a function of the number of clusters; that is, choose a number of clusters so that adding another cluster would not give much better modeling of the data. Specifically, if one graphs the percentage of variance explained by the clusters, there is a point at which the marginal gain will drop (giving an angle in the graph), indicating the number of clusters to be chosen.
- Set the number of clusters to $\sqrt{n}$, where n is the number of data points.
- Use the Akaike information criterion (AIC), which is a measure of the goodness of fit (based on the concept of entropy) to determine the number of clusters.
- Use Bayesian information criterion, which is a model-selection criterion (based on maximum likelihood estimation) to determine the number of clusters.

ANALYSIS METHODS Cluster analysis may be based on one or more of the following general methods:

- Statistical methods (including both hierarchical and nonhierarchical), such as k -means or k -modes
- Neural networks (with the architecture called self-organizing map)
- Fuzzy logic (e.g., fuzzy c -means algorithm)
- Genetic algorithms

Each of these methods generally works with one of two general method classes:

- **Divisive.** With divisive classes, all items start in one cluster and are broken apart.
- **Agglomerative.** With agglomerative classes, all items start in individual clusters, and the clusters are joined together.

Most cluster analysis methods involve the use of a **distance measure** to calculate the closeness between pairs of items. Popular distance measures include Euclidian distance (the ordinary distance between two points that one would measure with a ruler) and Manhattan distance (also called the rectilinear distance, or taxicab distance, between two points). Often, they are based on true distances that are measured, but this need not be so, as is typically the case in IS development. Weighted averages may be used to establish these distances. For example, in an IS development project, individual modules of the system may be related by the similarity between their inputs, outputs, processes, and

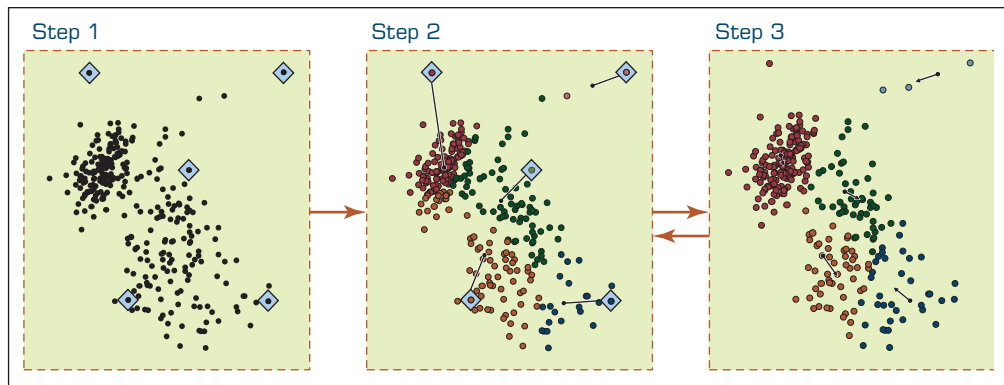


FIGURE 5.13 A Graphical Illustration of the Steps in the k -Means Algorithm.

the specific data used. These factors are then aggregated, pairwise by item, into a single distance measure.

K-MEANS CLUSTERING ALGORITHM The k -means algorithm (where k stands for the pre-determined number of clusters) is arguably the most referenced clustering algorithm. It has its roots in traditional statistical analysis. As the name implies, the algorithm assigns each data point (customer, event, object, etc.) to the cluster whose center (also called the *centroid*) is the nearest. The center is calculated as the average of all the points in the cluster; that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. The algorithm steps are listed below and shown graphically in Figure 5.13:

Initialization step: Choose the number of clusters (i.e., the value of k).

Step 1: Randomly generate k random points as initial cluster centers.

Step 2: Assign each point to the nearest cluster center.

Step 3: Recompute the new cluster centers.

Repetition step: Repeat steps 2 and 3 until some convergence criterion is met (usually that the assignment of points to clusters becomes stable).

Association Rule Mining

Association rule mining (also known as affinity analysis or market-basket analysis) is a popular data mining method that is commonly used as an example to explain what data mining is and what it can do to a technologically less-savvy audience. Most of you might have heard the famous (or infamous, depending on how you look at it) relationship discovered between the sales of beer and diapers at grocery stores. As the story goes, a large supermarket chain (maybe Walmart, maybe not; there is no consensus on which supermarket chain it was) did an analysis of customers' buying habits and found a statistically significant correlation between purchases of beer and purchases of diapers. It was theorized that the reason for this was that fathers (presumably young men) were stopping off at the supermarket to buy diapers for their babies (especially on Thursdays), and because they could no longer go down to the sports bar as often would buy beer as well. As a result of this finding, the supermarket chain is alleged to have placed the diapers next to the beer, resulting in increased sales of both.

In essence, association rule mining aims to find interesting relationships (affinities) between variables (items) in large databases. Because of its successful application to retail

business problems, it is commonly called *market-basket analysis*. The main idea in market-basket analysis is to identify strong relationships among different products (or services) that are usually purchased together (show up in the same basket together, either a physical basket at a grocery store or a virtual basket at an e-commerce Web site). For example, 65% of those who buy comprehensive automobile insurance also buy health insurance; 80% of those who buy books online also buy music online; 60% of those who have high blood pressure and are overweight have high cholesterol; 70% of the customers who buy a laptop computer and virus protection software also buy extended service plans.

The input to market-basket analysis is the simple point-of-sale transaction data, where a number of products and/or services purchased together (just like the content of a purchase receipt) are tabulated under a single transaction instance. The outcome of the analysis is invaluable information that can be used to better understand customer-purchase behavior to maximize the profit from business transactions. A business can take advantage of such knowledge by (1) putting the items next to each other to make it more convenient for the customers to pick them up together and not forget to buy one when buying the others (increasing sales volume); (2) promoting the items as a package (do not put one on sale if the other(s) are on sale); and (3) placing them apart from each other so that the customer has to walk the aisles to search for it, and by doing so potentially seeing and buying other items.

Applications of market-basket analysis include cross-marketing, cross-selling, store design, catalog design, e-commerce site design, optimization of online advertising, product pricing, and sales/promotion configuration. In essence, market-basket analysis helps businesses infer customer needs and preferences from their purchase patterns. Outside the business realm, association rules are successfully used to discover relationships between symptoms and illnesses, diagnosis and patient characteristics and treatments (which can be used in a medical decision support system), and genes and their functions (which can be used in genomics projects), among others. Here are a few common areas and uses for association rule mining:

- **Sales transactions:** Combinations of retail products purchased together can be used to improve product placement on the sales floor (placing products that go together in close proximity) and promotional pricing of products (not having promotions on both products that are often purchased together).
- **Credit card transactions:** Items purchased with a credit card provide insight into other products the customer is likely to purchase or fraudulent use of credit card numbers.
- **Banking services:** The sequential patterns of services used by customers (checking account followed by savings account) can be used to identify other services they may be interested in (investment account).
- **Insurance service products:** Bundles of insurance products bought by customers (car insurance followed by home insurance) can be used to propose additional insurance products (life insurance), or unusual combinations of insurance claims can be a sign of fraud.
- **Telecommunication services:** Commonly purchased groups of options (e.g., call waiting, caller ID, three-way calling) help better structure product bundles to maximize revenue; the same is also applicable to multichannel telecom providers with phone, television, and Internet service offerings.
- **Medical records:** Certain combinations of conditions can indicate increased risk of various complications; or, certain treatment procedures at certain medical facilities can be tied to certain types of infections.

A good question to ask with respect to the patterns/relationships that association rule mining can discover is “Are all association rules interesting and useful?” To answer such a question, association rule mining uses three common metrics: **support**,

confidence, and **lift**. Before defining these terms, let's get a little technical by showing what an association rule looks like:

$$X \Rightarrow Y [Supp(\%), Conf(\%)]$$

$$\{\text{Laptop Computer, Antivirus software}\} \Rightarrow \{\text{Extended Service Plan}\} [30\%, 70\%]$$

Here, X (products and/or service; called the *left-hand side*, *LHS*, or the antecedent) is associated with Y (products and/or service; called the *right-hand side*, *RHS*, or *consequent*). S is the support, and C is the confidence for this particular rule. Here are the simple formulas for *Supp*, *Conf*, and *Lift*.

$$\text{Support} = Supp(X \Rightarrow Y) = \frac{\text{number of baskets that contains both } X \text{ and } Y}{\text{total number of baskets}}$$

$$\text{Confidence} = Conf(X \Rightarrow Y) = \frac{Supp(X \Rightarrow Y)}{Supp(X)}$$

$$\text{Lift}(X \Rightarrow Y) = \frac{Conf(X \Rightarrow Y)}{\text{Expected } Conf(X \Rightarrow Y)} = \frac{\frac{Supp(X \Rightarrow Y)}{Supp(X)}}{\frac{Supp(X) * Supp(Y)}{Supp(X)}} = \frac{Supp(X \Rightarrow Y)}{Supp(X) * Supp(Y)}$$

The support (S) of a collection of products is the measure of how often these products and/or services (i.e., LHS + RHS = Laptop Computer, Antivirus Software, and Extended Service Plan) appear together in the same transaction; that is, the proportion of transactions in the data set that contain all of the products and/or services mentioned in a specific rule. In this example, 30% of all transactions in the hypothetical store database had all three products present in a single sales ticket. The confidence of a rule is the measure of how often the products and/or services on the RHS (consequent) go together with the products and/or services on the LHS (antecedent), that is, the proportion of transactions that include LHS while also including the RHS. In other words, it is the conditional probability of finding the RHS of the rule present in transactions where the LHS of the rule already exists. The lift value of an association rule is the ratio of the confidence of the rule and the expected confidence of the rule. The expected confidence of a rule is defined as the product of the support values of the LHS and the RHS divided by the support of the LHS.

Several algorithms are available for discovering association rules. Some well-known algorithms include Apriori, Eclat, and FP-Growth. These algorithms only do half the job, which is to identify the frequent itemsets in the database. Once the frequent itemsets are identified, they need to be converted into rules with antecedent and consequent parts. Determination of the rules from frequent itemsets is a straightforward matching process, but the process may be time-consuming with large transaction databases. Even though there can be many items on each section of the rule, in practice the consequent part usually contains a single item. In the following section, one of the most popular algorithms for identification of frequent itemsets is explained.

APRIORI ALGORITHM The **Apriori algorithm** is the most commonly used algorithm to discover association rules. Given a set of itemsets (e.g., sets of retail transactions, each listing individual items purchased), the algorithm attempts to find subsets that are common to at least a minimum number of the itemsets (i.e., complies with a minimum support). Apriori uses a bottom-up approach, where frequent subsets are extended one item at a time (a method known as *candidate generation*, whereby the size of frequent subsets

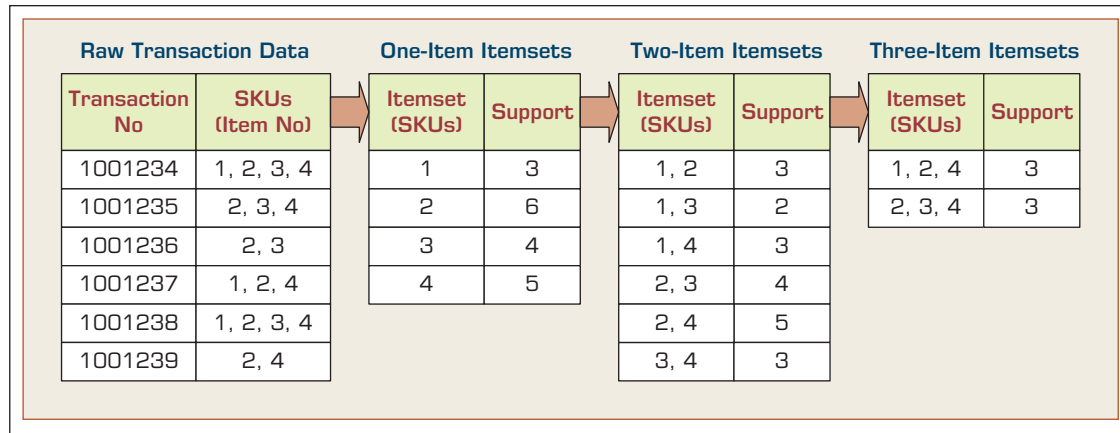


FIGURE 5.14 Identification of Frequent Itemsets in the Apriori Algorithm.

increases from one-item subsets to two-item subsets, then three-item subsets, etc.), and groups of candidates at each level are tested against the data for minimum support. The algorithm terminates when no further successful extensions are found.

As an illustrative example, consider the following. A grocery store tracks sales transactions by SKU (stock keeping unit) and thus knows which items are typically purchased together. The database of transactions, along with the subsequent steps in identifying the frequent itemsets, is shown in Figure 5.14. Each SKU in the transaction database corresponds to a product, such as “1 = butter,” “2 = bread,” “3 = water,” and so on. The first step in Apriori is to count up the frequencies (i.e., the supports) of each item (one-item itemsets). For this overly simplified example, let us set the minimum support to 3 (or 50%, meaning an itemset is considered to be a frequent itemset if it shows up in at least 3 out of 6 transactions in the database). Because all the one-item itemsets have at least 3 in the support column, they are all considered frequent itemsets. However, had any of the one-item itemsets not been frequent, they would not have been included as a possible member of possible two-item pairs. In this way, Apriori *prunes* the tree of all possible itemsets. As Figure 5.14 shows, using one-item itemsets, all possible two-item itemsets are generated and the transaction database is used to calculate their support values. Because the two-item itemset {1, 3} has a support less than 3, it should not be included in the frequent itemsets that will be used to generate the next-level itemsets (three-item itemsets). The algorithm seems deceptively simple, but only for small data sets. In much larger data sets, especially those with huge amounts of items present in low quantities and small amounts of items present in big quantities, the search and calculation become a computationally intensive process.

SECTION 5.5 REVIEW QUESTIONS

1. Identify at least three of the main data mining methods.
2. Give examples of situations in which classification would be an appropriate data mining technique. Give examples of situations in which regression would be an appropriate data mining technique.
3. List and briefly define at least two classification techniques.
4. What are some of the criteria for comparing and selecting the best classification technique?
5. Briefly describe the general algorithm used in decision trees.

6. Define *Gini index*. What does it measure?
7. What is an ensemble model in data mining? What are the pros and cons of ensemble models?
8. Give examples of situations in which cluster analysis would be an appropriate data mining technique.
9. What is the major difference between cluster analysis and classification?
10. What are some of the methods for cluster analysis?
11. Give examples of situations in which association would be an appropriate data mining technique.

5.6 DATA MINING SOFTWARE TOOLS

Many software vendors provide powerful data mining tools. Examples of these vendors include IBM (IBM SPSS Modeler, formerly known as SPSS PASW Modeler and Clementine), SAS (Enterprise Miner), Dell (Statistica, formerly known as StatSoft Statistica Data Miner), SAP (Infinite Insight, formerly known as KXEN Infinite Insight), Salford Systems (CART, MARS, TreeNet, RandomForest), Angoss (KnowledgeSTUDIO, KnowledgeSEEKER), and Megaputer (PolyAnalyst). Noticeably but not surprisingly, the most popular data mining tools are developed by the well-established statistical software companies (SAS, SPSS, and StatSoft)—largely because statistics is the foundation of data mining, and these companies have the means to cost-effectively develop them into full-scale data mining systems. Most of the business intelligence tool vendors (e.g., IBM Cognos, Oracle Hyperion, SAP Business Objects, Tableau, Tibco, Qlik, MicroStrategy, Teradata, and Microsoft) also have some level of data mining capabilities integrated into their software offerings. These BI tools are still primarily focused on multidimensional modeling and data visualization and are not considered to be direct competitors of the data mining tool vendors. As described in Analytics in Action 5.2, IBM SPSS Modeler is used as the main data mining tool in developing prediction models for Hollywood movies' box-office receipts.

In addition to these commercial tools, several open-source and/or free data mining software tools are available online. Traditionally, especially in educational circles, the most popular free and open-source data mining tool is **Weka**, which was developed by a number of researchers from the University of Waikato in New Zealand (the tool can be downloaded from cs.waikato.ac.nz/ml/weka). Weka includes a large number of algorithms for different data mining tasks and has an intuitive user interface. Recently, a number of free open-source, highly capable data mining tools emerged: leading the pack are **KNIME** (knime.org) and **RapidMiner** (rapidminer.com). Their graphically enhanced user interfaces, employment of a rather large number of algorithms, and incorporation of a variety of data visualization features set them apart from the rest of the free tools. These two free software tools are also platform agnostic (i.e., can natively run on both Windows and Mac operating systems). With a recent change in their offerings, RapidMiner has created a scaled-down version of their analytics tool for free (i.e., community edition) while making the full commercial product. Therefore, once listed under the free/open-source tools category, RapidMiner nowadays is often listed under commercial tools. The main difference between commercial tools, such as SAS Enterprise Miner, IBM SPSS Modeler, and Statistica, and free tools, such as Weka, RapidMiner (community edition), and KNIME, is the computational efficiency. The same data mining task involving a rather large and feature-rich data set may take a whole lot longer to complete with the free software tools, and for some algorithms the job may not even complete (i.e., crashing due to the inefficient use of computer memory). Table 5.2 lists a few of the major products and their Web sites.

TABLE 5.2 Selected Data Mining Software

Product Name	Web Site (URL)
IBM SPSS Modeler	www-01.ibm.com/software/analytics/spss/products/modeler/
IBM Watson Analytics	ibm.com/analytics/watson-analytics/
SAS Enterprise Miner	sas.com/en_id/software/analytics/enterprise-miner.html
Dell Statistica	statsoft.com/products/statistica/product-index
PolyAnalyst	megaputer.com/site/polyanalyst.php
CART, RandomForest	salford-systems.com
Insightful Miner	solutionmetrics.com.au/products/iminer/default.html
XLMiner	solver.com/xlminer-data-mining
SAP InfitelInsight (KXEN)	help.sap.com/ii
GhostMiner	fqs.pl/ghostminer
SQL Server Data Mining	msdn.microsoft.com/en-us/library/bb510516.aspx
Knowledge Miner	knowledgeminer.com
Teradata Warehouse Miner	teradata.com/products-and-services/teradata-warehouse-miner/
Oracle Data Mining (ODM)	oracle.com/technetwork/database/options/odm/
FICO Decision Management	fico.com/en/analytics/decision-management-suite/
Orange Data Mining Tool	orange.biolab.si/
Zementis Predictive Analytics	zementis.com

ANALYTICS IN ACTION 5.2 Data Mining Goes to Hollywood: Predicting Financial Success of Movies

Analytics in Action 5.2 is about a research study where a number of software tools and data mining techniques are used to build data mining models to predict financial success (box-office receipts) of Hollywood movies while they are nothing more than ideas.

Predicting box-office receipts (i.e., financial success) of a particular motion picture is an interesting and challenging problem. According to some domain experts, the movie industry is the “land of hunches and wild guesses” due to the difficulty associated with forecasting product demand, making the movie business in Hollywood a risky endeavor. In support of such observations, Jack Valenti (the longtime president and CEO of the Motion Picture Association of America) once mentioned that “no one can tell you how a movie is going to do in the marketplace ... not until the film opens in darkened theatre and sparks fly up between the screen and the audience.” Entertainment industry trade journals and magazines have been full of examples, statements, and experiences that support such a claim.

Like many other researchers who have attempted to shed light on this challenging real-world problem, Ramesh Sharda and Dursun Delen have been exploring the use of data mining to predict the financial performance of a motion picture at the box office before it even enters production (while the movie is nothing more than a conceptual idea). In their highly publicized prediction models, they convert the forecasting (or regression) problem into a classification problem; that is, rather than forecasting the point estimate of box-office receipts, they classify a movie based on its box-office receipts in one of nine categories, ranging from “flop” to “block-buster,” making the problem a multinomial classification problem. Table 5.3 illustrates the definition of the nine classes in terms of the range of box-office receipts.

TABLE 5.3 Movie Classification Based on Receipts

Class No.	1	2	3	4	5	6	7	8	9
Range (in millions of dollars)	> 1 (Flop)	> 1 < 610	> 10 < 20	> 20 < 640	> 40 < 665	> 65 < 6100	> 100 < 6150	> 150 < 6200	> 200 (Blockbuster)

TABLE 5.4 Summary of Independent Variables

Independent Variable	Number of Values	Possible Values
MPAA Rating	5	G, PG, PG-13, R, NR
Competition	3	High, Medium, Low
Star value	3	High, Medium, Low
Genre	10	Sci-Fi, Historic Epic Drama, Modern Drama, Politically Related, Thriller, Horror, Comedy, Cartoon, Action, Documentary
Special effects	3	High, Medium, Low
Sequel	2	Yes, No
Number of screens	1	A positive integer between 1 and 3876

Data

Data were collected from a variety of movie-related databases (e.g., ShowBiz, IMDb, IMSDb, AllMovie, BoxofficeMojo) and consolidated into a single data set. The data set for the most recently developed models contained 2,632 movies released between 1998 and 2006. A summary of the independent variables along with their specifications is provided in Table 5.4. For more descriptive details and justification for inclusion of these independent variables, the reader is referred to Sharda and Delen (2006).

Methodology

Using a variety of data mining methods, including neural networks, decision trees, SVMs, and three types of ensembles, Sharda and Delen developed the prediction models. The data from 1998 to 2005 were used as training data to build the prediction models, and the data from 2006 was used as the test data to assess and compare the models' prediction accuracy. Figure 5.15 shows a screenshot of IBM SPSS Modeler (formerly Clementine data mining tool) depicting the process map employed for the prediction problem. The upper-left side of the process map shows the model development process, and the lower-right corner of the process map shows the model assessment (i.e., testing or scoring) process (more details on IBM SPSS Modeler tool and its usage can be found on the book's Web site).

Results

Table 5.5 provides the prediction results of all three data mining methods as well as the results of the three different ensembles. The first performance measure is the percent correct classification rate, which is called *Bingo*. Also reported in the table is the *1-Away* correct classification rate (i.e., within one category). The results indicate that SVM performed the best among the individual prediction models, followed by ANN; the worst of the three was the CART decision tree algorithm. In general, the ensemble models performed better than the individual prediction models, of which the fusion algorithm performed the best. What is probably more important to decision makers, and standing out in the results table, is the significantly low standard deviation obtained from the ensembles compared to the individual models.

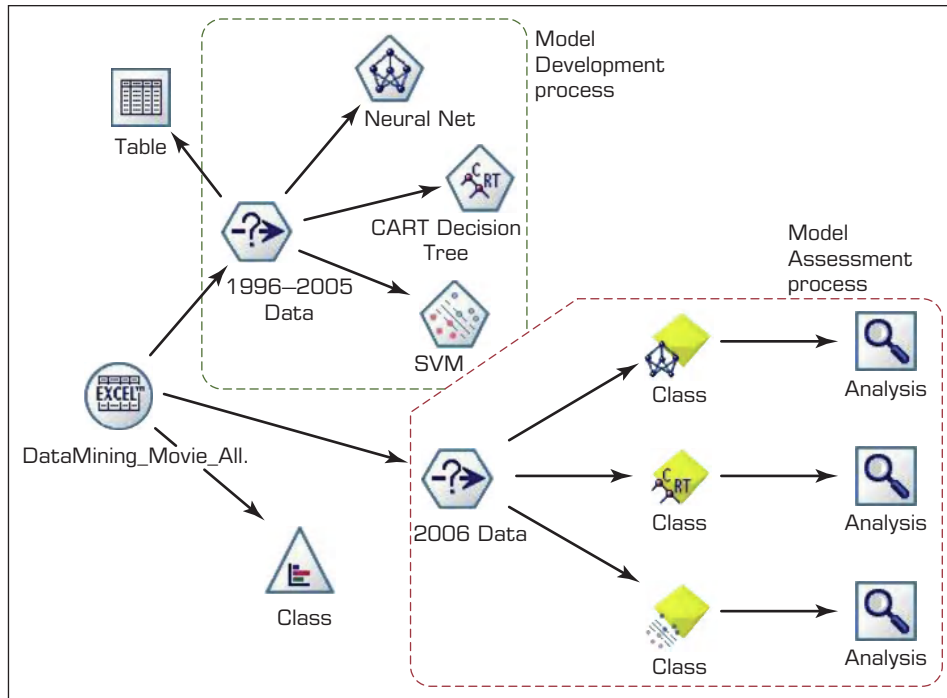


FIGURE 5.15 Process Flow Screenshot for the Box-Office Prediction System. *Source:* Used with permission from IBM SPSS.

TABLE 5.5 Tabulated Prediction Results for Individual and Ensemble Models

Performance Measure	Prediction Models					
	Individual Models			Ensemble Models		
	SVM	ANN	CART	Random Forest	Boosted Tree	Fusion (Average)
Count (Bingo)	192	182	140	189	187	194
Count (1-Away)	104	120	126	121	104	120
Accuracy (% Bingo)	55.49%	52.60%	40.46%	54.62%	54.05%	56.07%
Accuracy (% 1-Away)	85.55%	87.28%	76.88%	89.60%	84.10%	90.75%
Standard deviation	0.93	0.87	1.05	0.76	0.84	0.63

Conclusion

The researchers claim that these prediction results are better than any reported in the published literature for this problem domain. Beyond the attractive accuracy of their prediction results of the box-office receipts, these models could also be used to further analyze (and potentially optimize) the decision variables to maximize the financial return. Specifically, the parameters used for modeling could be altered using the already trained prediction models to better understand the impact of different parameters on the end results. During this process, which is commonly referred to as *sensitivity analysis*, the decision maker of a given entertainment firm could find out, with a fairly high accuracy level, how much value a specific actor (or a specific release date, or the addition of more technical effects, etc.) brings to the financial success of a film, making the underlying system an invaluable decision aid.

Sources: Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30, 243–254; Delen, D., Sharda, R., & Kumar, P. (2007). Movie forecast Guru: A Web-based DSS for Hollywood managers. *Decision Support Systems*, 43(4), 1151–1170.

A suite of business intelligence and analytics capabilities that has become increasingly more popular for data mining studies is **Microsoft's SQL Server** (it has been including increasingly more analytics capabilities, such as BI and predictive modeling modules, starting with the SQL Server 2012 version), where data and the models are stored in the same relational database environment, making model management a considerably easier task. The **Microsoft Enterprise Consortium** serves as the worldwide source for access to Microsoft's SQL Server software suite for academic purposes—teaching and research. The consortium has been established to enable universities around the world to access enterprise technology without having to maintain the necessary hardware and software on their own campus. The consortium provides a wide range of business intelligence development tools (e.g., data mining, cube building, business reporting) as well as a number of large, realistic data sets from Sam's Club, Dillard's, and Tyson Foods. The Microsoft Enterprise Consortium is free of charge and can only be used for academic purposes. The Sam M. Walton College of Business at the University of Arkansas hosts the enterprise system and allows consortium members and their students to access these resources using a simple remote desktop connection. The details about becoming a part of the consortium along with easy-to-follow tutorials and examples can be found at walton.uark.edu/enterprise/.

In May 2016, KDnuggets.com conducted the 13th Annual Software Poll on the following question: “What software you used for Analytics, Data Mining, Data Science, Machine Learning projects in the past 12 months?” The poll has received remarkable participation from analytics and data science community and vendors, attracting 2,895 voters, who chose from a record number of 102 different tools. Here are some of the interesting findings that came out of the poll:

- R remains the leading tool, with 49% shares (up from 46.9% in 2015), but Python usage grew faster and it almost caught up to R with 45.8% shares (up from 30.3%). Later polls conducted by KDnuggets.com show that Python has already overtook the first spot in the analytics tools popularity rankings.
- RapidMiner remains the most popular general platform for data mining/data science, with about 33% shares. Notable tools with the most growth in popularity include Dato, Dataiku, MLlib, H2O, Amazon Machine Learning, scikit-learn, and IBM Watson.
- The increased choice of tools is reflected in wider usage. The average number of tools used was 6.0 (versus 4.8 in May 2015).
- The usage of Hadoop/Big Data tools grew to 39%, up from 29% in 2015 (and 17% in 2014), driven by Apache Spark, MLlib (Spark Machine Learning Library), and H2O.
- The participation by region was: US/Canada (40%), Europe (39%), Asia (9.4%), Latin America (5.8%), Africa/MidEast (2.9%), and Australia/NZ (2.2%).
- This year, 86% of voters used commercial software, and 75% used free software. About 25% used only commercial software, and 13% used only open-source/free software. A majority of 61% used both free and commercial software, similar to 64% in 2015.
- The usage of Hadoop/Big Data tools grew to 39%, up from 29% in 2015 and 17% in 2014, driven mainly by big growth in Apache Spark, MLlib (Spark Machine Learning Library), and H2O, which we included among Big Data tools.
- For the second year KDnuggets.com's poll included Deep Learning tools. This year, 18% of voters used Deep Learning tools, doubling the 9% in 2015—Google Tensorflow jumped to first place, displacing last year's leader, Theano/Pylearn2 ecosystem.
- In the programming languages category, Python, Java, Unix tools, and Scala grew in popularity, while C/C++, Perl, Julia, F#, Clojure, and Lisp declined.

To reduce bias through multiple voting, in this poll KDnuggets.com used e-mail verification and, by doing so, aimed to make results more representative of the reality in the analytics world. The results for the top 40 software tools (as per total number of votes received) are shown in Figure 5.16. The horizontal bar chart also makes a differentiation among free/open-source, commercial, and Big Data/Hadoop tools using a color-coding schema. Recent trends are mostly inline with these outcomes with the exception that Python has surpassed R and become the most popular analytics tool in both industry and academia.

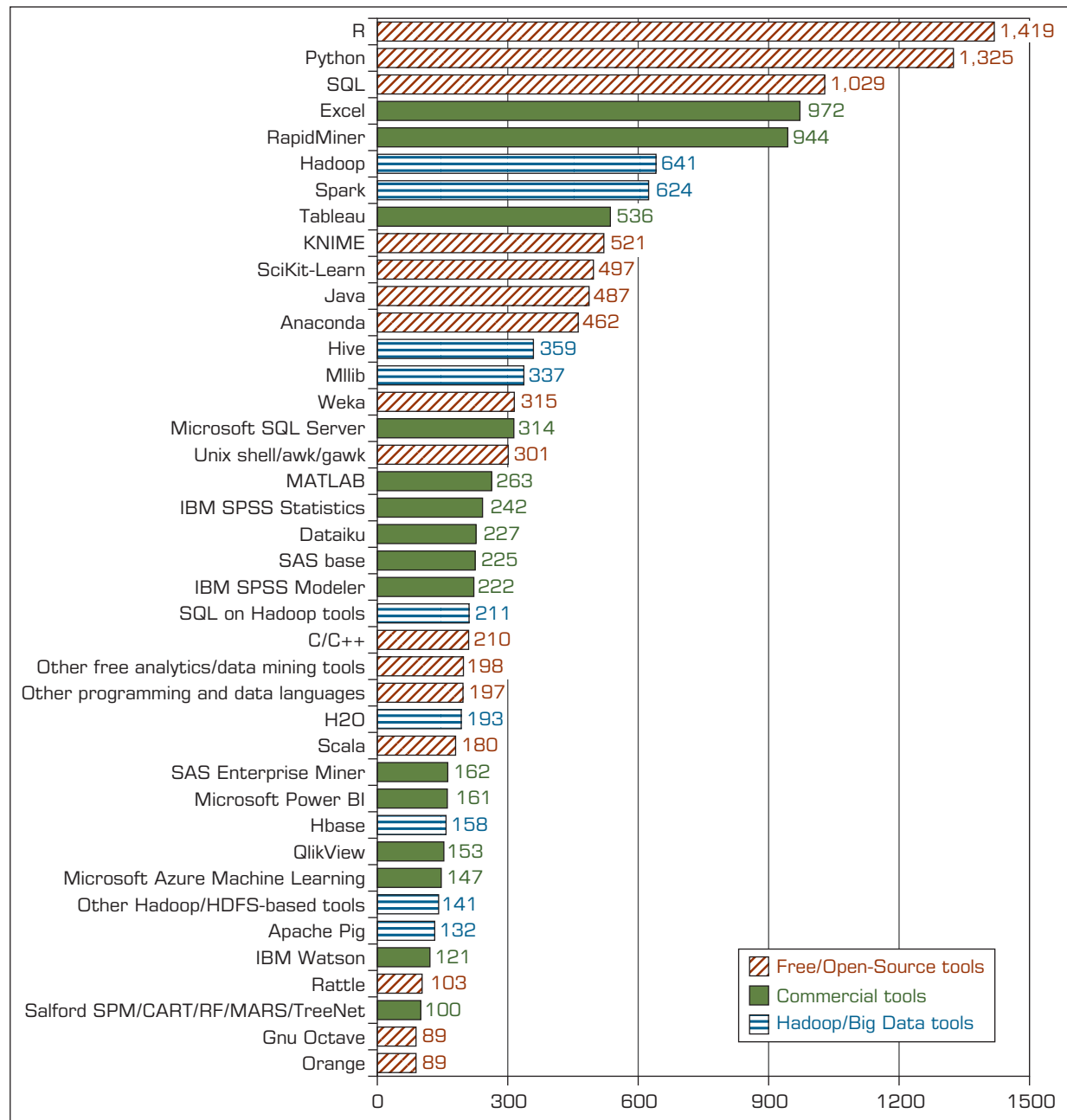


FIGURE 5.16 Popular Data Analytics Software Tools (Poll Results). Source: Used with permission from KDnuggets.com.

► SECTION 5.6 REVIEW QUESTIONS

1. What are the most popular commercial data mining tools?
2. Why do you think the most popular tools are developed by statistics-based companies?
3. What are the most popular free data mining tools? Why are they gaining overwhelming popularity (especially R)?
4. What are the main differences between commercial and free data mining software tools?
5. What would be your top five selection criteria for a data mining tool? Explain.

5.7 DATA MINING PRIVACY ISSUES, MYTHS, AND BLUNDERS

Data that is collected, stored, and analyzed in data mining often contains information about real people. Such information may include identification data (name, address, Social Security number, driver's license number, employee number, etc.), demographic data (e.g., age, sex, ethnicity, marital status, number of children), financial data (e.g., salary, gross family income, checking or savings account balance, home ownership, mortgage or loan account specifics, credit card limits and balances, investment account specifics), purchase history (i.e., what is bought from where and when—either from vendor's transaction records or from credit card transaction specifics), and other personal data (e.g., anniversary, pregnancy, illness, loss in the family, bankruptcy filings). Most of these data can be accessed through some third-party data providers. The main question here is the privacy of the person to whom the data belongs. To maintain the privacy and protection of individuals' rights, data mining professionals have ethical (and often legal) obligations. One way to accomplish this is the process of de-identification of the customer records prior to applying data mining applications, so that the records cannot be traced to an individual. Many publicly available data sources (e.g., CDC data, SEER data, UNOS data) are already de-identified. Prior to accessing these data sources, users are often asked to consent that, under no circumstances, will they try to identify the individuals behind those figures.

There have been a number of instances in the recent past where the companies shared their customer data with others without seeking the explicit consent of their customers. For instance, as most of you might recall, in 2003, JetBlue Airlines provided more than a million passenger records of their customers to Torch Concepts, a U.S. government contractor. Torch then subsequently augmented the passenger data with additional information such as family sizes and Social Security numbers—information purchased from a data broker called Acxiom. The consolidated personal database was intended to be used for a data mining project to develop potential terrorist profiles. All of this was done without notification or consent of passengers. When news of the activities got out, however, dozens of privacy lawsuits were filed against JetBlue, Torch, and Acxiom, and several U.S. senators called for an investigation into the incident (Wald, 2004). Similar, but not as dramatic, privacy-related news came out in the recent past about the popular social network companies, which allegedly were selling customer-specific data to other companies for personalized target marketing.

There was another peculiar story about privacy concerns that made it to the headlines in 2012. In this instance, the company did not even use any private and/or personal data. Legally speaking, there was no violation of any laws. It was about Target and is summarized in Analytics in Action 5.3.

ANALYTICS IN ACTION 5.3 Predicting Customer Buying Patterns—the Target Story

In early 2012, an infamous story appeared concerning Target's practice of predictive analytics. The story was about a teenage girl who was being sent advertising flyers and coupons by Target for the kinds of things that a new mother-to-be would buy from a store like Target. The story goes like this: An angry person went into a Target outside of Minneapolis, demanding to talk to a manager: "My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?" The manager didn't have any idea what the person was talking about. He looked at the mailer. Sure enough, it was addressed to the man's daughter and contained advertisements for maternity clothing, nursery furniture, and pictures of smiling infants. The manager apologized and then called a few days later to apologize again. On the phone, though, the father was somewhat abashed. "I had a talk with my daughter," he said. "It turns out there's been some activities in my house I haven't been completely aware of. She's due in August. I owe you an apology."

As it turns out, Target figured out a teen girl was pregnant before her father did! Here is how they did it. Target assigns every customer a Guest ID number (tied to their credit card, name, or e-mail address) that becomes a placeholder that keeps a history of everything they have bought. Target augments this data with any demographic information that they had collected from them or bought from other information sources. Using this information, Target looked at historical buying data for all the females who had signed up for Target baby registries in the past. They analyzed the data from all directions, and soon enough some useful patterns emerged. For example, lotions and special vitamins were among the products with interesting purchase patterns. Lots of people buy lotion, but what they noticed was that women on the baby registry were buying larger quantities of unscented lotion around the beginning of their second trimester. Another analyst noted that sometime in the first 20 weeks, pregnant women loaded up on supplements like calcium, magnesium, and zinc. Many shoppers purchase soap and cotton balls, but when someone suddenly starts buying lots of scent-free soap and extra-large bags of cotton balls, in addition to hand sanitizers and washcloths, it signals that they could be getting close to their delivery date. In the end, they were able to identify about 25 products that, when analyzed together, allowed them to assign each shopper a "pregnancy prediction" score. More important, they could also estimate a woman's due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

If you look at this practice from a legal perspective, you would conclude that Target did not use any information that violates customer privacy; rather, they used transactional data that almost every other retail chain is collecting and storing (and perhaps analyzing) about their customers. What was disturbing in this scenario was perhaps the targeted concept: pregnancy. There are certain events or concepts that should be off limits or treated extremely cautiously, such as terminal disease, divorce, and bankruptcy.

Sources: Hill, K. (2012, February 16). How Target figured out a teen girl was pregnant before her father did. *Forbes*; Nolan, R. (2012, February 21). Behind the cover story: How much does Target know? *NYTimes.com*.

Data Mining Myths and Blunders

Data mining is a powerful analytical tool that enables business executives to advance from describing the nature of the past (looking at a rearview mirror) to predicting the future (looking ahead with the help of headlights) to better manage their business operations (making accurate and timely decisions). Data mining helps marketers find patterns that unlock the mysteries of customer behavior. The results of data mining can be used to increase revenue and reduce cost by identifying fraud and discovering business opportunities, offering a whole new range of tools for competitive advantage. As an evolving and maturing field, data mining is often associated with a number of myths, and corresponding realities, including those of the ones listed in Table 5.6 (Delen, 2021; Zaima, 2003).

TABLE 5.6 Data Mining Myths

Myth	Reality
Data mining provides instant, crystal-ball-like predictions.	Data mining is a multistep process that requires deliberate, proactive design and use.
Data mining is not yet viable for mainstream business applications.	They current state of the art is ready to go for almost any business type and/or size.
Data mining requires a separate, dedicated database.	Because of the advances in database technology, a dedicated database is not required.
Only those with advanced degrees can do data mining.	Never Web-based tools enable managers of all educational levels to do data mining.
Data mining is only for large firms that have lots of customer data.	If the data accurately reflect the business or its customers, any company can use data mining.

Data mining visionaries have gained enormous competitive advantage by understanding that these myths are just that: myths.

Although the value proposition and therefore the necessity of it is obvious to anyone, those who carry out data mining projects (from novice to seasoned data scientist) sometimes make mistakes that result in projects with less-than-desirable outcomes. The following 16 data mining mistakes (also called blunders, pitfalls, or bloopers) are often made in practice (Delen 2021, Nesbit et al., 2009 Shultz, 2004; Skalak, 2001), and data scientists should be aware of them, and to the extent that is possible, do their best to avoid them:

- Selecting the wrong problem for data mining. Not every business problem can be solved with data mining (i.e., the magic bullet syndrome). When there is no representative data (large and feature rich), there cannot be a practicable data mining project.
- Ignoring what your sponsor thinks data mining is and what it really can and cannot do. Expectation management is the key for successful data mining projects.
- Beginning without the end in mind. Although data mining is a process of knowledge discovery, one should have a goal/objective (a stated business problem) in mind to succeed. Because, as the saying goes, “if you don’t know where you are going, you will never get there.”
- Define the project around a foundation that your data can’t support. Data mining is all about data; that is, the biggest constraint that you have in a data mining project is the richness of the data. Knowing what the limitations of data are help you craft feasible projects that deliver results and meet expectations.
- Leaving insufficient time for data preparation. It takes more effort than is generally understood. The common knowledge suggests that up to a third of the total project time is spent on data acquisition, understanding, and preparation tasks. To succeed, avoid proceeding into modeling until after your data are properly processed (aggregated, cleaned, and transformed).
- Looking only at aggregated results and not at individual records. Data mining is at its best when the data are at a granular representation. Try to avoid unnecessarily aggregating and overly simplifying data to help data mining algorithms—they don’t really need your help; they are more than capable of figuring it out themselves.
- Being sloppy about keeping track of the data mining procedure and results. Because it is a discovery process that involves many iterations and experimentations, it is

highly likely to lose track of the findings. Success requires a systematic and orderly planning, execution, and tracking/recording of all data mining tasks.

- Using data from the future to predict the future. Because of the lack of description and understanding of the data, oftentimes analysts include variables that are unknown at the time when the prediction is supposed to be made. By doing so, their prediction models produce unbelievable accurate results (a phenomenon that is often called “fool’s gold”). If your prediction results are too good to be true, they usually are; in that case, the first thing that you need to look for is the incorrect use of a variable from the future.
- Ignoring suspicious findings and quickly moving on. The unexpected findings are often the indicators of real novelties in data mining projects. Proper investigation of such oddities can lead to surprisingly pleasing discoveries.
- Starting with a high-profile complex project that will make you a superstar. Data mining projects often fail if they are not thought out carefully from start to end. Success often comes with a systematic and orderly progression of projects from smaller/simpler to larger/complex ones. The goal should be to show incremental and continuous value added, as opposed to taking on a large project that will consume resources without producing any valuable outcomes.
- Running data mining algorithms repeatedly and blindly. Although today’s data mining tools are capable of consuming data and setting algorithmic parameters to produce results, one should know how to transform the data and set the proper parameter values to obtain the best possible results. Each algorithm has its own unique way of processing data, and knowing that is necessary to get the most out of each model type.
- Ignore the subject matter experts. Understanding the problem domain and the related data requires a highly involved collaboration between the data mining and the domain experts. Working together helps the data mining expert to go beyond the syntactic representation and also obtain semantic nature (i.e., the true meaning of the variables) of the data.
- Believing everything you are told about the data. Although it is necessary to talk to domain experts to better understand the data and the business problem, the data scientist should not take anything for granted. Validation and verification through a critical analysis is the key to intimate understanding and processing of the data.
- Assuming that the keepers of the data will be fully on board with cooperation. Many data mining projects fail because the data mining expert did not know/understand the organizational politics. One of the biggest obstacles in data mining projects can be the people who own and control the data. Understanding and managing the politics is a key to identify, access, and properly understand the data to produce a successful data mining project.
- Measuring your results differently from the way your sponsor measures them. The results should talk/appeal to the end user (manager/decision maker) who will be using them. Therefore, producing the results in a measure and format that appeals to the end user tremendously increases the likelihood of true understanding and proper use of the data mining outcomes.
- If you build it, they will come: don’t worry about how to serve it up. Usually, data mining experts think they are done once they build models that meet and hopefully exceed the needs/wants/expectations of the end user (i.e., the customer). Without a proper deployment, the value deliverance of data mining outcomes is rather limited. Therefore, deployment is a necessary last step in the data mining process where models are integrated into the organizational decision support infrastructure for enablement of better and faster decision-making.

SECTION 5.7 REVIEW QUESTIONS

1. What are the privacy issues in data mining?
2. How do you think the discussion between privacy and data mining will progress? Why?
3. What are the most common myths about data mining?
4. What do you think are the reasons for these myths about data mining?
5. What are the most common data mining mistakes/blunders? How can they be alleviated or completely eliminated?

Chapter Highlights

- Data mining is the process of discovering new knowledge from structured databases.
- Data mining can use simple flat files as data sources, or it can be performed on data in data warehouses.
- There are many alternative names and definitions for data mining.
- Data mining is at the intersection of many disciplines, including statistics, artificial intelligence, and mathematical modeling.
- Companies use data mining to better understand their customers and optimize their operations.
- Data mining applications can be found in virtually every area of business and government, including healthcare, finance, marketing, and homeland security.
- Three broad categories of data mining tasks are prediction (classification or regression), clustering, and association.
- Similar to other IS initiatives, a data mining project must follow a systematic project management process to be successful.
- Several data mining processes have been proposed: CRISP-DM, SEMMA, KDD, and so on.
- CRISP-DM provides a systematic and orderly way to conduct data mining projects.
- The earlier steps in data mining projects (i.e., understanding the domain and the relevant data) consume most of the total project time (often more than 80% of the total time).
- Data preprocessing is essential to any successful data mining study. Good data leads to good information; good information leads to good decisions.
- Data preprocessing includes four main steps: data consolidation, data cleaning, data transformation, and data reduction.
- Classification methods learn from previous examples containing inputs and the resulting class labels, and once properly trained they are able to classify future cases.
- Clustering partitions pattern records into natural segments or clusters. Each segment's members share similar characteristics.
- A number of different algorithms are commonly used for classification. Commercial implementations include ID3, C4.5, C5, CART, CHAID, and SPRINT.
- Decision trees partition data by branching along different attributes so that each leaf node has all the patterns of one class.
- The Gini index and information gain (entropy) are two popular ways to determine branching choices in a decision tree.
- The Gini index measures the purity of a sample. If everything in a sample belongs to one class, the Gini index value is zero.
- Several assessment techniques can measure the prediction accuracy of classification models, including simple split, k -fold cross-validation, bootstrapping, and area under the ROC curve.
- Cluster algorithms are used when the data records do not have predefined class identifiers (i.e., it is not known to what class a particular record belongs).
- Cluster algorithms compute measures of similarity in order to group similar cases into clusters.
- The most commonly used similarity measure in cluster analysis is a distance measure.
- The most commonly used clustering algorithms are k -means and self-organizing maps.
- Association rule mining is used to discover two or more items (or events or concepts) that go together.
- Association rule mining is commonly referred to as market-basket analysis.
- The most commonly used association algorithm is Apriori, whereby frequent itemsets are identified through a bottom-up approach.

- Association rules are assessed based on their support and confidence measures.
- Many commercial and free data mining tools are available.
- The most popular commercial data mining tools are SPSS PASW and SAS Enterprise Miner.
- The most popular free data mining tools are Weka and RapidMiner.

Key Terms

Apriori algorithm	data mining	knowledge discovery in	ordinal data
area under the ROC curve	decision tree	databases (KDD)	prediction
association	distance measure	lift	RapidMiner
bootstrapping	entropy	link analysis	regression
categorical data	Gini index	Microsoft Enterprise Con-	SEMMA
classification	information gain	sortium	sequence mining
clustering	interval data	Microsoft SQL Server	simple split
confidence	<i>k</i> -fold cross-validation	nominal data	support
CRISP-DM	KNIME	numeric data	Weka

Questions for Discussion

1. Define *data mining*. Why are there many names and definitions for data mining?
2. What are the main reasons for the recent popularity of data mining?
3. Discuss what an organization should consider before making a decision to purchase data mining software.
4. Distinguish data mining from other analytical tools and techniques.
5. Discuss the main data mining methods. What are the fundamental differences among them?
6. What are the main data mining application areas? Discuss the commonalities of these areas that make them a prospect for data mining studies.
7. Why do we need a standardized data mining process? What are the most commonly used data mining processes?
8. Discuss the differences between the two most commonly used data mining processes.
9. Are data mining processes a mere sequential set of activities? Explain.
10. Why do we need data preprocessing? What are the main tasks and relevant techniques used in data preprocessing?
11. Discuss the reasoning behind the assessment of classification models.
12. What is the main difference between classification and clustering? Explain using concrete examples.
13. Moving beyond the chapter discussion, where else can association be used?
14. What are the privacy issues with data mining? Do you think they are substantiated?
15. What are the most common myths and mistakes about data mining?

Exercises

Teradata University Network (TUN) and Other Hands-On Exercises

1. Conduct a Web search to identify case studies and white papers about data mining. Based on your findings, describe the recent developments in the field of data mining and predictive modeling.
2. Do a Web search to locate Web-based seminars (i.e., webinars) related to data mining. Watch at least two webinars and write a short report to answer the following questions:
 - a. What are some of the interesting applications of data mining?
 - b. What types of payoffs and costs can organizations expect from data mining initiatives?
3. For this exercise, your goal is to build a model to identify inputs or predictors that differentiate risky customers from others (based on patterns pertaining to previous customers) and then use those inputs to predict new risky customers. This sample case is typical for this domain.

The sample data to be used in this exercise are in Online File W5.1 in the file CreditRisk.xlsx. The data set has 425 cases and 15 variables pertaining to past and current customers who have borrowed from a bank for

various reasons. The data set contains customer-related information such as financial standing, reason for the loan, employment, demographic information, and the outcome or dependent variable for credit standing, classifying each case as good or bad, based on the institution's past experience.

Take 400 of the cases as training cases and set aside the other 25 for testing. Build a decision tree model on 400 cases to learn the characteristics of the problem. Test its performance on the other 25 cases. Report on your model's learning and testing performance. Prepare a report that identifies the decision tree model and training parameters, as well as the resulting performance on the test set. You can use any data mining software that has decision tree induction module (the easiest would be to use one of the free/open-source tools like R, Python, or KNIME).

4. For this exercise, you will replicate (on a smaller scale) the box-office prediction modeling explained in Analytics in Action 5.2. Download the training data set from Online File W5.2, *MovieTrain.xlsx*, which is in Microsoft Excel format. Use the data description given in Analytics in Action 5.2 to understand the domain and the problem you are trying to solve. Pick and choose your independent variables. Develop at least three classification models (e.g., decision tree, logistic regression, neural networks). Compare the accuracy results using 10-fold cross-validation and percentage split techniques, use confusion matrices, and comment on the outcome. Test the models you have developed on the test set (see Online File W5.3, *MovieTest.xlsx*). Analyze the results with different models, and come up with the best classification model, supporting it with your results.
5. This exercise is aimed at introducing you to association rule mining. The Excel data set *basketsIntrans.xlsx* has around 2,800 observations/records of supermarket trans products data. Each record contains the customer's ID and Products that they have purchased. Use this data set to understand the relationships among products (i.e., which products are purchased together). Look for interesting relationships and add screenshots of any subtle association patterns that you might find. More specifically, answer the following questions.
 - Which association rules do you think are most important?
 - Based on some of the association rules you found, make at least three business recommendations that might be beneficial to the company. These recommendations up-selling may include ideas about shelf organization, upselling, or cross-selling products. (Bonus points will be given to new/innovative ideas.)
 - What are the Support, Confidence, and Lift values for the following rule?

wine, canned Veg → Frozen Meal

In this assignment you will use a free/open-source data mining tool, KNIME (knime.org) to build predictive models for a relatively small Customer Churn Analysis data set. You are to analyze the given data set (about the customer retention/attrition behavior for 1,000 customers) to develop and compare at least three prediction (i.e., classification) models. For example, you can include decision trees, neural networks, SVM, k nearest neighbor, and/or logistic regression models in your comparison. Here are the specifics for this assignment:

- Install and use the KNIME software tool from (knime.org).
- You can also use MS Excel to preprocess the data (if you need to/want to).
- Download *CustomerChurnData.csv* data file from the book's Web site.
 - The data are given in CSV (Comma Separated Value) format. This format is the most common flat-file format that many software tools can easily open/handle (including KNIME and MS Excel).
- Present your results in a well-organized professional document.
 - Include a cover page (with proper information about you and the assignment).
 - Make sure to nicely integrate figures (graphs, charts, tables, screenshots) within your textual description in a professional manner. The report should have six main sections (resembling CRISP-DM phases).
 - Try not to exceed 15 pages in total, including the cover (use 12-point Times New Roman fonts, and 1.5-line spacing).
- 6. In this assignment, you will use a free/open-source data mining tool, KNIME (knime.org) to build predictive models for a relatively small Customer Churn Analysis data set. You are to analyze the given data set (about the customer retention/attrition behavior for 1,000 customers) to develop and compare at least three prediction (i.e., classification) models. For example, you can include decision trees, neural networks, SVM, k nearest neighbor, and/or logistic regression models in your comparison. Here are the specifics for this assignment:
 - Install and use the KNIME software tool from (knime.org).
 - You can also use MS Excel to preprocess the data (if you need to/want to).
 - Download *CustomerChurnData.csv* data file from the book's Web site.
 - The data are given in CSV (Comma Separated Value) format. This format is the most common flat-file format that many software tools can easily open/handle (including KNIME and MS Excel).
 - Present your results in a well-organized professional document.
 - Include a cover page (with proper information about you and the assignment).

- Make sure to nicely integrate figures (graphs, charts, tables, screenshots) within your textual description in a professional manner. The report should have six main sections (resembling CRISP-DM phases).
- Try not to exceed 15 pages in total, including the cover (use 12-point Times New Roman fonts, and 1.5-line spacing).

Team Assignments and Role-Playing Projects

1. Examine how new data capture devices such as RFID tags help organizations accurately identify and segment their customers for activities such as targeted marketing. Many of these applications involve data mining. Scan the literature and the Web and then propose five potential new data mining applications that can use the data created with RFID technology. What issues could arise if a country's laws required such devices to be embedded in everyone's body for a national identification system?
2. Interview administrators in your college or executives in your organization to determine how data mining, data warehousing, OLAP, and visualization tools could assist them in their work. Write a proposal describing your findings. Include cost estimates and benefits in your report.
3. A very good repository of data that has been used to test the performance of many data mining algorithms is available at ics.uci.edu/~mlearn/MLRepository.html. Some of the data sets are meant to test the limits of current machine-learning algorithms and to compare their performance with new approaches to learning. However, some of the smaller data sets can be useful for exploring the functionality of any data mining software, such as RapidMiner or KNIME. Download at least one data set from this repository (e.g., Credit Screening Databases, Housing Database) and apply decision tree or clustering methods, as appropriate. Prepare a report based on your results. (Some of these exercises, especially the ones that involve large/challenging data/problem may be used as semester-long term projects.)
4. Large and feature-rich data sets are made available by the U.S. government or its subsidiaries on the Internet. For instance, see a large collection of government data sets (data.gov), the Centers for Disease Control and Prevention data sets (www.cdc.gov/DataStatistics), Surveillance, Cancer.org's Epidemiology and End Results data sets (<http://seer.cancer.gov/data>), and the Department of Transportation's Fatality Analysis Reporting System crash data sets (www.nhtsa.gov/FARS). These data sets are not preprocessed for data mining, which makes them a great resource to experience the complete data mining process. Another rich source for a collection of analytics data sets is listed on KDnuggets. com (kdnuggets.com/datasets/index.html).
5. Consider the following data set, which includes three attributes and a classification for admission decisions into an MBA program:

GMAT	GPA	Quantitative GMAT Score (percentile)	Decision
650	2.75	35	No
580	3.50	70	No
600	3.50	75	Yes
450	2.95	80	No
700	3.25	90	Yes
590	3.50	80	Yes
400	3.85	45	No
640	3.50	75	Yes
540	3.00	60	?
690	2.85	80	?
490	4.00	65	?

- Using the data shown, develop your own manual expert rules for decision-making.
- Use the Gini index to build a decision tree. You can use manual calculations or a spreadsheet to perform the basic calculations.
- Use an automated decision tree software program to build a tree for the same data.

Internet Exercises

1. Visit the AI Exploratorium at cs.ualberta.ca/~aixplore. Click the Decision Tree link. Read the narrative on basketball game statistics. Examine the data, and then build a decision tree. Report your impressions of the accuracy of this decision tree. Also, explore the effects of different algorithms.
2. Survey some data mining tools and vendors. Start with fico.com and egain.com. Consult dmreview.com, and identify some data mining products and service providers that are not mentioned in this chapter.
3. Find recent cases of successful data mining applications. Visit the Web sites of some data mining vendors, and look for cases or success stories. Prepare a report summarizing five new case studies.
4. Go to vendor Web sites (especially those of SAS, SPSS, Cognos, Teradata, KNIME, and Fair Isaac) and look at success stories for BI (OLAP and data mining) tools. What do the various success stories have in common? How do they differ?
5. Go to knime.com. Download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?
6. Go to tibco.com. Download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?
7. Go to sas.com. Download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?

8. Go to spss.com (an IBM company). Download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?
9. Go to teradata.com. Download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?
10. Go to fico.com. Download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?
11. Go to salfordsystems.com. Download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?
12. Go to rulequest.com. Download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?
13. Go to kdnuggets.com. Explore the sections on applications as well as software. Find names of at least three additional packages for data mining and text mining.

References

- Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst*. John Wiley & Sons.
- Chan, P. K., Phan, W., Prodromidis, A., & Stolfo, S. (1999). Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems*, 14(6), 67–74.
- CRISP-DM. (2013). Cross-Industry Standard Process for Data Mining (CRISP-DM). <http://crisp-dm.org/www.the-modeling-agency.com/crisp-dm.pdf> (accessed February 2, 2013).
- Davenport, T. H. (2006, January). Competing on analytics. *Harvard Business Review*, 99–107.
- Delen, D. (2009). Analysis of cancer data: A data mining approach. *Expert Systems*, 26(1), 100–112.
- Delen, D. (2014). *Real-world data mining: Applied business analytics and decision making*. Upper Saddle River, NJ: Pearson.
- Delen, D. (2021). *Predictive Analytics: Data Mining, Machine Learning and Data Science for Practitioners., 2nd Edition*. Upper Saddle River, NJ: Pearson.
- Delen, D., Cogdell, D., & Kasap, N. (2012). A comparative analysis of data mining methods in predicting NCAA Bowl outcomes. *International Journal of Forecasting*, 28, 543–552.
- Delen, D., & Sharda, R. (2010). Predicting the financial success of Hollywood movies using an information fusion approach. *Industrial Engineering Journal*, 21(1), 30–37.
- Delen, D., Sharda, R., & Kumar, P. (2007). Movie forecast Guru: A Web-based DSS for Hollywood managers. *Decision Support Systems*, 43(4), 1151–1170.
- Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine*, 34(2), 113–127.
- Dunham, M. (2003). *Data mining: Introductory and advanced topics*. Upper Saddle River, NJ: Prentice Hall.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From knowledge discovery in databases. *AI Magazine*, 17(3), 37–54.
- Hoffman, T. (1998, December 7). Banks turn to IT to reclaim most profitable customers. *Computerworld*.
- Hoffman, T. (1999, April 19). Insurers mine for age-appropriate offering. *Computerworld*.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1), 59–69.
- Nemati, H. R., & Barko, C. D. (2001). Issues in organizational data mining: A survey of current practices. *Journal of Data Warehousing*, 6(1), 25–36.
- Nisbet, R., Miner, G., & Elder IV, J. (2009). “Top 10 Data Mining Mistakes” in the Handbook of statistical analysis and data mining applications. pp. 733–754. Academic Press.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1, 81–106.
- SEMMA. (2009). SAS’s data mining process: Sample, explore, modify, model, assess. sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html (accessed August 2009).
- Seni, G., & Elder, J. F. (2010). Ensemble methods in data mining: Improving accuracy through combining predictions. Synthesis Lectures on Data Mining and Knowledge Discovery, 2(1), 1–126.
- Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30, 243–254.
- Shultz, R. (2004, December 7). Live from NCDM: Tales of database buffoonery. directmag.com/news/ncdm-12-07-04/index.html (accessed April 2009).
- Skalak, D. (2001). Data mining blunders exposed! *DB2 Magazine*, 6(2), 10–13.
- Thongkam, J., Xu, G., Zhang, Y., & Huang, F. (2009). Toward breast cancer survivability prediction models through improving training space. *Expert Systems with Applications*, 36(10), 12200–12209.
- Wald, M. L. (2004, February 21). U.S. calls release of JetBlue data improper. *The New York Times*.
- Wright, C. (2012). *Statistical predictors of March Madness: An examination of the NCAA Men’s Basketball Championship*. <http://economics-files.pomona.edu/GarySmith/Econ190/Wright%20March%20Madness%20Final%20Paper.pdf> (accessed February 2, 2013).
- Zaima, A. (2003). The five myths of data mining. *What Works: Best practices in business intelligence and data warehousing*, Vol. 15. Chatsworth, CA: The Data Warehousing Institute, pp. 42–43.
- Zolbanin, H. M., Delen, D., & Zadeh, A. H. (2015). Predicting overall survivability in comorbidity of cancers: A data mining approach. *Decision Support Systems*, 74, 150–161.

Predictive Analytics II: Text, Web, and Social Media Analytics

LEARNING OBJECTIVES

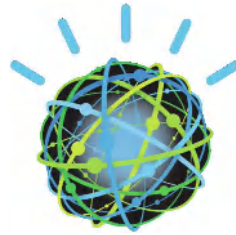
- Describe text analytics and understand the need for text mining
- Differentiate among text analytics, text mining, and data mining
- Understand the different application areas for text mining
- Know the process of carrying out a text mining project
- Appreciate the different methods to introduce structure to text-based data
- Describe sentiment analysis
- Develop familiarity with popular applications of sentiment analysis
- Learn the common methods for sentiment analysis
- Understand the foundations of topic modeling and its analytics techniques
- Become familiar with social media analytics

This chapter provides a comprehensive overview of text analytics (also known as text mining or text processing) and Web analytics/mining along with their popular application areas such as search engines, sentiment analysis, and social network/media analytics. As we have been witnessing in the recent years, the unstructured data generated over the Internet (through user-generated content on the Web pages, social networks links, text and images/videos, internet of things [IoT] platforms, sensor networks, radio-frequency identification [RFID]-enabled supply chain systems, surveillance networks, etc.) is increasing at an exponential pace, and there is no indication of it slowing down. This changing nature of data is forcing organizations to make text and Web analytics a critical part of their business intelligence/analytics infrastructure.

- 6.1** Opening Vignette: Machine versus Human on *Jeopardy!*: The Story of Watson 302
- 6.2** Text Analytics and Text Mining Overview 305
- 6.3** Natural Language Processing (NLP) 308
- 6.4** Text Mining Applications 313

- 6.5 Text Mining Process 318
- 6.6 Sentiment Analysis and Topic Modeling 326
- 6.7 Web Mining Overview 339
- 6.8 Search Engines 344
- 6.9 Web Usage Mining (Web Analytics) 349
- 6.10 Social Analytics 355

6.1 OPENING VIGNETTE: Machine Versus Human on *Jeopardy!*: The Story of Watson



Can a machine beat the best of human in what human is supposed to be the best at? Evidently, yes, and the machine's name is Watson. Watson is an extraordinary computer system (a novel combination of advanced hardware and software) designed to answer questions posed in natural human language. It was developed in 2010 by an IBM Research team as part of a DeepQA project and was named after IBM's first president, Thomas J. Watson.

BACKGROUND

Roughly three years ago, IBM Research was looking for a major research challenge to rival the scientific and popular interest of Deep Blue, the computer chess-playing champion, which would also have clear relevance to IBM business interests. The goal was to advance computer science by exploring new ways for computer technology to affect science, business, and society. Accordingly, IBM Research undertook a challenge to build a computer system that could compete at the human champion level in real time on the American TV quiz show *Jeopardy!* The extent of the challenge included fielding a real-time automatic contestant on the show, capable of listening, understanding, and responding—not merely a laboratory exercise.

COMPETING AGAINST THE BEST

In 2011, as a test of its abilities, Watson competed on the quiz show *Jeopardy!*, which was the first ever human-versus-machine matchup for the show. In a two-game, combined-point match (broadcast in three *Jeopardy!* episodes during February 14–16), Watson beat Brad Rutter, the biggest all-time money winner on *Jeopardy!*, and Ken Jennings, the record holder for the longest championship streak (75 days). In these episodes, Watson consistently outperformed its human opponents on the game's signaling device, but had trouble responding to a few categories, notably those having short clues containing only a few words. Watson had access to 200 million pages of structured and unstructured content consuming 4 terabytes of disk storage. During the game Watson was not connected to the Internet.

Meeting the *Jeopardy!* Challenge required advancing and incorporating a variety of QA technologies (text mining and natural language processing), including parsing,

question classification, question decomposition, automatic source acquisition and evaluation, entity and relation detection, logical form generation, and knowledge representation and reasoning. Winning at *Jeopardy!* required accurately computing confidence in your answers. The questions and content are ambiguous and noisy, and none of the individual algorithms are perfect. Therefore, each component must produce a confidence in its output, and individual component confidences must be combined to compute the overall confidence of the final answer. The final confidence is used to determine whether the computer system should risk choosing to answer at all. In *Jeopardy!* parlance, this confidence is used to determine whether the computer will “ring in” or “buzz in” for a question. The confidence must be computed during the time the question is read and before the opportunity to buzz in. This is roughly between 1 and 6 seconds with an average around 3 seconds.

HOW DOES WATSON DO IT?

The system behind Watson, which is called DeepQA, is a massively parallel, text mining-focused, probabilistic evidence-based computational architecture. For the *Jeopardy!* Challenge, Watson used more than 100 different techniques for analyzing natural language, identifying sources, finding and generating hypotheses, finding and scoring evidence, and merging and ranking hypotheses. What is far more important than any particular technique that they used was how they combine them in DeepQA such that overlapping approaches can bring their strengths to bear and contribute to improvements in accuracy, confidence, and speed.

DeepQA is an architecture with an accompanying methodology, which is not specific to the *Jeopardy!* Challenge. The overarching principles in DeepQA are massive parallelism, many experts, pervasive confidence estimation, and integration of the latest and greatest in text analytics.

- **Massive parallelism:** Exploits massive parallelism in the consideration of multiple interpretations and hypotheses.
- **Many experts:** Facilitate the integration, application, and contextual evaluation of a wide range of loosely coupled probabilistic question and content analytics.
- **Pervasive confidence estimation:** No component commits to an answer; all components produce features and associated confidences, scoring different question and content interpretations. An underlying confidence-processing substrate learns how to stack and combine the scores.
- **Integrate shallow and deep knowledge:** Balance the use of strict semantics and shallow semantics, leveraging many loosely formed ontologies.

Figure 6.1 illustrates a high-level depiction of the DeepQA architecture. More technical details about the various architectural components and their specific roles and capabilities can be found in Ferrucci et al. (2010).

CONCLUSION

The *Jeopardy!* challenge helped IBM address requirements that led to the design of the DeepQA architecture and the implementation of Watson. After three years of intense research and development by a core team of about 20 researchers, Watson is performing at human expert levels in terms of precision, confidence, and speed at the *Jeopardy!* quiz show.

IBM claims to have developed many computational and linguistic algorithms to address different kinds of issues and requirements in QA. Even though the internals of these algorithms are not known, it is imperative that they made the most out of text analytics and text mining. In the last decade, IBM has been working on and deploying various implementation of Watson to take on surmountable problems in healthcare and medicine.

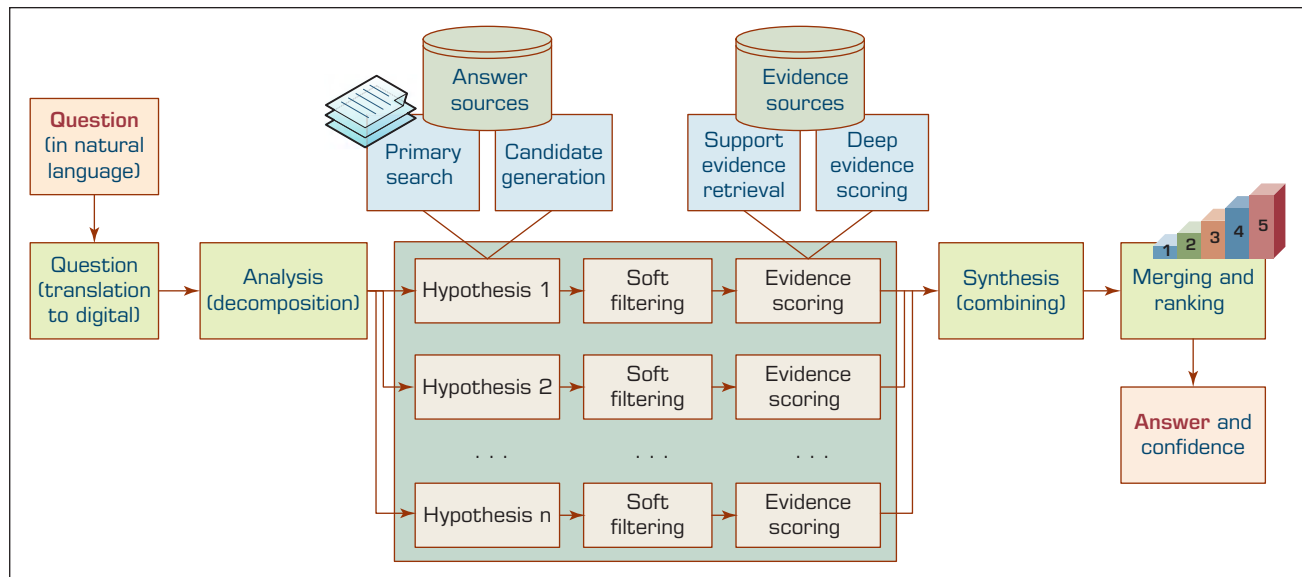


FIGURE 6.1 A High-Level Depiction of DeepQA Architecture.

► QUESTIONS FOR THE OPENING VIGNETTE

1. What is Watson? What is special about it?
2. What technologies were used in building Watson (both hardware and software)?
3. What are the innovative characteristics of DeepQA architecture that made Watson superior?
4. Why did IBM spend all that time and money to build Watson? Where is the return on investment (ROI)?

WHAT WE CAN LEARN FROM THIS VIGNETTE

It is safe to say that computer technology, both on the hardware and software fronts, is advancing faster than anything else in the last 50-plus years. Things that were too big, too complex, impossible to solve are now well within the reach of information technology. One of the enabling technologies is perhaps text analytics/text mining. We created databases to structure the data so that it can be processed by computers. Text, on the other hand, has always been meant for humans to process. Can machines do the things that require human creativity and intelligence, and were not originally designed for machines? Evidently, yes! Watson is a great example of the distance that we have traveled in addressing the impossible. Computers are now intelligent enough to take on men at what we think men are the best at. Understanding the question that was posed in spoken human language, processing and digesting it, searching for an answer, and replying within a few seconds was something that we could not have imagined possible before Watson actually did it. In this chapter, you will learn the tools and techniques embedded in Watson and many other smart machines to create miracles in tackling problems that were once believed impossible to solve.

Sources: Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., ... Welty, C. (2010). Building Watson: An overview of the DeepQA Project. *AI Magazine*, 31(3); The DeepQA Research Team (2021). researcher.watson.ibm.com/researcher/view_group_subpage.php?id=2159 (accessed January 2022); Feldman, S., Hanover, J., Burghard, C., & Schubmehl, D. (2012). Recommend with confidence—Unlocking the power of unstructured data. IBM Watson (2022). www.ibm.com/watson/advantages/recommend?mbsrc=ibmsearch_a&mbq=Unlocking%20Unstructured%20Data (accessed June 2022).

6.2 TEXT ANALYTICS AND TEXT MINING OVERVIEW

The information age that we are living in is characterized by the rapid growth in the amount of data and information collected, stored, and made available in electronic format. A vast majority of business data are stored in text documents that are virtually unstructured. According to a study by Merrill Lynch and Gartner, 85% of all corporate data are captured and stored in some sort of unstructured form (McKnight, 2005). The same study also stated that these unstructured data are doubling in size every 18 months. Because knowledge is power in today's business world, and knowledge is derived from data and information, businesses that effectively and efficiently tap into their text data sources will have the necessary knowledge to make better decisions, leading to a competitive advantage over those businesses that lag behind. This is where the need for text analytics and text mining fits into the big picture of today's businesses.

Even though the overarching goal for both text analytics and text mining is to turn unstructured textual data into actionable information through the application of natural language processing (NLP) and analytics, their definitions are somewhat different, at least to some experts in the field. According to them, text analytics is a broader concept that includes information retrieval (e.g., searching and identifying relevant documents for a given set of key terms), as well as information extraction, data mining, and Web mining, whereas text mining is primarily focused on discovering new and useful knowledge from the textual data sources. Figure 6.2 illustrates the relationships between text analytics and text mining along with other related application areas (Miner et al., 2012).

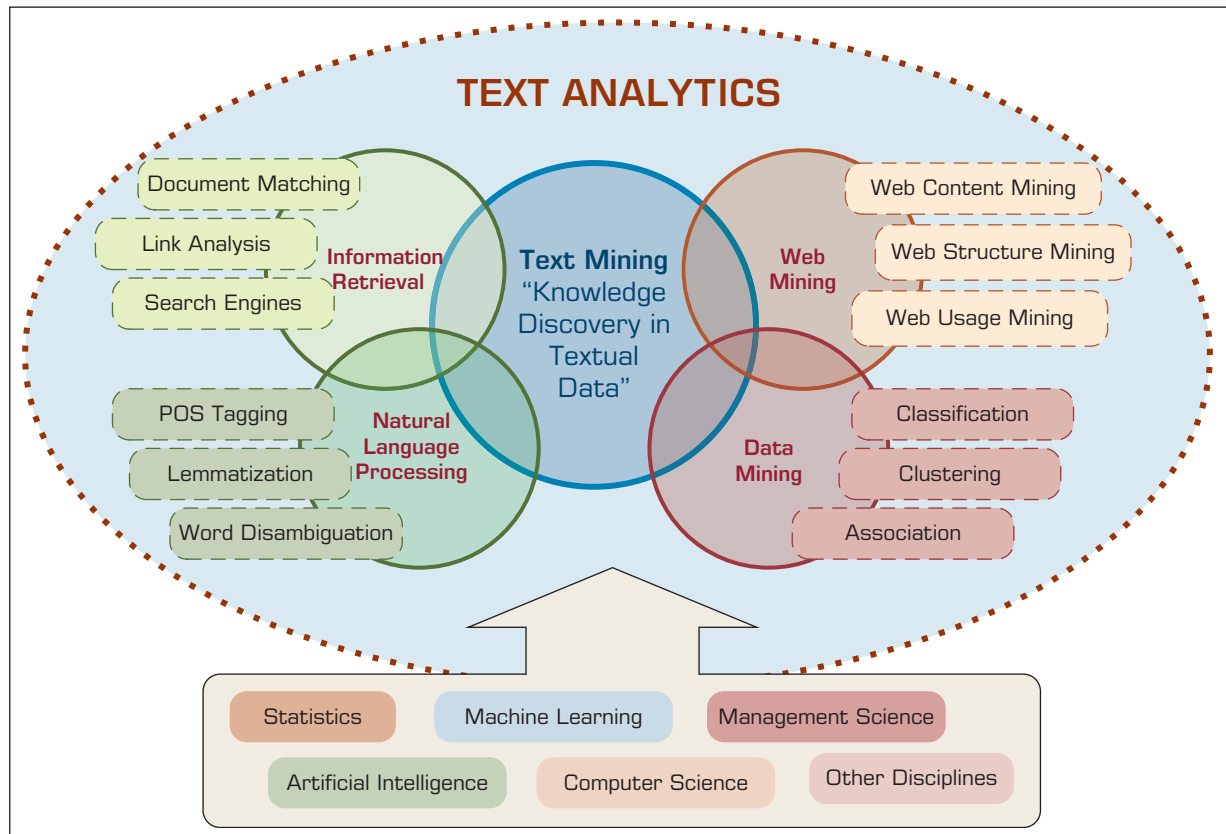


FIGURE 6.2 Text Analytics, Related Application Areas, and Enabling Disciplines.

The bottom of Figure 6.2 lists the main disciplines (the foundation of the house) that play a critical role in the development of these increasingly more popular application areas. Based on this definition of text analytics and text mining, one could simply formulate the difference between the two as follows:

$$\text{Text Analytics} = \text{Information Retrieval} + \text{Information Extraction} + \text{Data Mining} + \text{Web Mining},$$

or simply

$$\text{Text Analytics} = \text{Information Retrieval} + \text{Text Mining}$$

Compared to text mining, *text analytics* is a relatively new term. With the recent emphasis on *analytics*, as has been the case in many other related technical application areas (e.g., consumer analytics, competitive analytics, visual analytics, social analytics), the field of text has also wanted to get on the analytics bandwagon. Although the term *text analytics* is more commonly used in a business application context, text mining is frequently used in academic research circles. Even though they may be defined somewhat differently at times, text analytics and text mining are usually used synonymously, and we (authors of this book) concur with this.

Text mining (also known as *text data mining* or *knowledge discovery in textual databases*) is the semi-automated process of extracting patterns (useful information and knowledge) from large amounts of unstructured data sources. Remember that data mining is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases, where the data are organized in records structured by categorical, ordinal, or continuous variables. Text mining is the same as data mining in that it has the same purpose and uses the same processes, but with text mining the input to the process is a collection of unstructured (or less structured) data files such as Word documents, PDF files, text excerpts, and XML files. In essence, text mining can be thought of as a process (with two main steps) that starts with imposing structure on the text-based data sources followed by extracting relevant information and knowledge from this structured text-based data using data mining techniques and tools.

The benefits of text mining are obvious in the areas where very large amounts of textual data are being generated, such as law (court orders), academic research (research articles), finance (quarterly reports), medicine (discharge summaries), biology (molecular interactions), technology (patent files), and marketing (customer comments). For example, the free-form text-based interactions with customers in the form of complaints (or praises) and warranty claims can be used to objectively identify product and service characteristics that are deemed to be less than perfect and can be used as input to better product development and service allocations. Likewise, market outreach programs and focus groups generate large amounts of data. By not restricting product or service feedback to a codified form, customers can present, in their own words, what they think about a company's products and services. Another area where the automated processing of unstructured text has had a lot of impact is in electronic communications and e-mail. Text mining not only can be used to classify and filter junk e-mail, but it can also be used to automatically prioritize e-mail based on importance level as well as generate automatic responses (Weng & Liu, 2004). Following are among the most popular application areas of text mining:

- **Information extraction.** Identification of key phrases and relationships within text by looking for predefined objects and sequences in text by way of pattern matching.
- **Topic tracking.** Based on a user profile and documents that a user views, text mining can predict other documents of interest to the user.

- **Summarization.** Summarizing a document to save time on the part of the reader.
- **Categorization.** Identifying the main themes of a document and then placing the document into a predefined set of categories based on those themes.
- **Clustering.** Grouping similar documents without having a predefined set of categories.
- **Concept linking.** Connects related documents by identifying their shared concepts and, by doing so, helps users find information that they perhaps would not have found using traditional search methods.
- **Question answering.** Finding the best answer to a given question through knowledge-driven pattern matching.

See Technology Insights 6.1 for explanations of some of the terms and concepts used in text mining.

TECHNOLOGY INSIGHTS 6.1 Text Mining Terminology

The following list describes some commonly used text mining terms:

- **Unstructured data (versus structured data).** Structured data has a predetermined format. It is usually organized into records with simple data values (categorical, ordinal, and continuous variables) and stored in databases. In contrast, **unstructured data** does not have a predetermined format and is stored in the form of textual documents. In essence, the structured data are for the computers to process while the unstructured data are for humans to process and understand.
- **Corpus.** In linguistics, a **corpus** (plural *corpora*) is a large and structured set of texts (now usually stored and processed electronically) prepared for the purpose of conducting knowledge discovery.
- **Terms.** A *term* is a single word or multiword phrase extracted directly from the corpus of a specific domain by means of NLP methods.
- **Concepts.** *Concepts* are features generated from a collection of documents by means of manual, statistical, rule-based, or hybrid categorization methodology. Compared to terms, concepts are the result of higher level abstraction.
- **Stemming.** **Stemming** is the process of reducing inflected words to their stem (or base or root) form. For instance, *stemmer*, *stemming*, *stemmed* are all based on the root *stem*.
- **Stop words.** **Stop words** (or *noise words*) are words that are filtered out prior to or after processing of natural language data (i.e., text). Even though there is no universally accepted list of stop words, most NLP tools use a list that includes articles (*a*, *am*, *the*, *of*, etc.), auxiliary verbs (*is*, *are*, *was*, *were*, etc.), and context-specific words that are deemed not to have differentiating value.
- **Synonyms and polysemes.** Synonyms are syntactically different words (i.e., spelled differently) with identical or at least similar meanings (e.g., *movie*, *film*, and *motion picture*). In contrast, **polysemes**, which are also called *homonyms*, are syntactically identical words (i.e., spelled exactly the same) with different meanings (e.g., *bow* can mean “to bend forward,” “the front of the ship,” “the weapon that shoots arrows,” or “a kind of tied ribbon”).
- **Tokenizing.** A *token* is a categorized block of text in a sentence. The block of text corresponding to the token is categorized according to the function it performs. This assignment of meaning to blocks of text is known as **tokenizing**. A token can look like anything; it just needs to be a useful part of the structured text.
- **Term dictionary.** A collection of terms specific to a narrow field that can be used to restrict the extracted terms within a corpus.
- **Word frequency.** The number of times a word is found in a specific document.
- **Part-of-speech tagging.** The process of marking up the words in a text as corresponding to a particular part of speech (nouns, verbs, adjectives, adverbs, etc.) based on a word's definition and the context in which it is used.

- **Morphology.** A branch of the field of linguistics and a part of NLP that studies the internal structure of words (patterns of word formation within a language or across languages).
 - **Term-by-document matrix (occurrence matrix).** A common representation schema of the frequency-based relationship between the terms and documents in tabular format where terms are listed in columns, documents are listed in rows, and the frequency between the terms and documents is listed in cells as integer values.
 - **Singular value decomposition (latent semantic indexing).** A dimensionality reduction method used to transform the term-by-document matrix to a manageable size by generating an intermediate representation of the frequencies using a matrix manipulation method similar to principal component analysis.
-

► SECTION 6.2 REVIEW QUESTIONS

1. What is text analytics? How does it differ from text mining?
2. What is text mining? How does it differ from data mining?
3. Why is the popularity of text mining as an analytics tool increasing?
4. What are some of the most popular application areas of text mining?

6.3 NATURAL LANGUAGE PROCESSING (NLP)

Some of the early text mining applications used a simplified representation called *bag-of-words* when introducing structure to a collection of text-based documents to classify them into two or more predetermined classes or to cluster them into natural groupings. In the bag-of-words model, text, such as a sentence, paragraph, or complete document, is represented as a collection of words, disregarding the grammar or the order in which the words appear. The bag-of-words model is still used in some simple document classification tools. For instance, in spam filtering an e-mail message can be modeled as an unordered collection of words (a bag-of-words) that is compared against two different predetermined bags. One bag is filled with words found in spam messages and the other is filled with words found in legitimate e-mails. Although some of the words are likely to be found in both bags, the “spam” bag will contain spam-related words such as *stock*, *Viagra*, and *buy* much more frequently than the legitimate bag, which will contain more words related to the user’s friends or workplace. The level of match between a specific e-mail’s bag-of-words and the two bags containing the descriptors determines the membership of the e-mail as either spam or legitimate.

Naturally, we (humans) do not use words without some order or structure. We use words in sentences, which have semantic as well as syntactic structure. Thus, automated techniques (such as text mining) need to look for ways to go beyond the bag-of-words interpretation and incorporate more and more semantic structure into their operations. The current trend in text mining is toward including many of the advanced features that can be obtained using NLP.

It has been shown that the bag-of-words method may not produce good enough information content for text mining tasks (e.g., classification, clustering, association). A good example of this can be found in evidence-based medicine. A critical component of evidence-based medicine is incorporating the best available research findings into the clinical decision-making process, which involves appraisal of the information collected from the printed media for validity and relevance. Several researchers from the University of Maryland developed evidence assessment models using a bag-of-words method (Lin & Demner-Fushman, 2005). They employed popular machine-learning methods along with more than half a million research articles collected from MEDLINE (Medical Literature Analysis and Retrieval System Online). In their models, they represented each abstract as

a bag-of-words, where each stemmed term represented a feature. Despite using popular classification methods with proven experimental design methodologies, their prediction results were not much better than simple guessing, which may indicate that the bag-of-words is not generating a good enough representation of the research articles in this domain; hence, more advanced techniques such as NLP are needed.

Natural language processing (NLP) is an important component of text mining and is a subfield of artificial intelligence and computational linguistics. It studies the problem of “understanding” the natural human language, with the view of converting depictions of human language (such as textual documents) into more formal representations (in the form of numeric and symbolic data) that are easier for computer programs to manipulate. The goal of NLP is to move beyond syntax-driven text manipulation (which is often called “word counting”) to a true understanding and processing of natural language that considers grammatical and semantic constraints as well as the context.

The definition and scope of the word *understanding* is one of the major discussion topics in NLP. Considering that the natural human language is vague and that a true understanding of meaning requires extensive knowledge of a topic (beyond what is in the words, sentences, and paragraphs), will computers ever be able to understand natural language the same way and with the same accuracy that humans do? Probably not! NLP has come a long way from the days of simple word counting, but it has an even longer way to go to really understanding natural human language. The following are just a few of the challenges commonly associated with the implementation of NLP:

- **Part-of-speech tagging.** It is difficult to mark up terms in a text as corresponding to a particular part of speech (such as nouns, verbs, adjectives, or adverbs) because the part of speech depends not only on the definition of the term but also on the context within which it is used.
- **Text segmentation.** Some written languages, such as Chinese, Japanese, and Thai, do not have single-word boundaries. In these instances, the text-parsing task requires the identification of word boundaries, which is often a difficult task. Similar challenges in speech segmentation emerge when analyzing spoken language because sounds representing successive letters and words blend into each other.
- **Word sense disambiguation.** Many words have more than one meaning. Selecting the meaning that makes the most sense can only be accomplished by taking into account the context within which the word is used.
- **Syntactic ambiguity.** The grammar for natural languages is ambiguous; that is, multiple possible sentence structures often need to be considered. Choosing the most appropriate structure usually requires a fusion of semantic and contextual information.
- **Imperfect or irregular input.** Foreign or regional accents and vocal impediments in speech and typographical or grammatical errors in texts make the processing of the language an even more difficult task.
- **Speech acts.** A sentence can often be considered an action by the speaker. The sentence structure alone may not contain enough information to define this action. For example, “Can you pass the class?” requests a simple yes/no answer, whereas “Can you pass the salt?” is a request for a physical action to be performed.

It is a long-standing dream of the artificial intelligence community to have algorithms that are capable of automatically reading and obtaining knowledge from text. By applying a learning algorithm to parsed text, researchers from Stanford University’s NLP lab have developed methods that can automatically identify the concepts and relationships between those concepts in the text. By applying a unique procedure to large amounts of text, their algorithms automatically acquire hundreds of thousands of items of world knowledge and use them to produce significantly enhanced repositories for

WordNet. **WordNet** is a laboriously hand-coded database of English words, their definitions, sets of synonyms, and various semantic relations between synonym sets. It is a major resource for NLP applications, but it has proven to be very expensive to build and maintain manually. By automatically inducing knowledge into WordNet, the potential exists to make WordNet an even greater and more comprehensive resource for NLP at a fraction of the cost. One prominent area where the benefits of NLP and WordNet are already being harvested is in customer relationship management (CRM). Broadly speaking, the goal of CRM is to maximize customer value by better understanding and effectively responding to their actual and perceived needs. An important area of CRM, where NLP is making a significant impact, is sentiment analysis. **Sentiment analysis** is a technique used to detect favorable and unfavorable opinions toward specific products and services using a large number of textual data sources (customer feedback in the form of Web postings). A detailed coverage of sentiment analysis and WordNet is given in Section 6.6.

Analytics in general and text analytics and text mining in specific can be used to understand customer sentiments and provide means to deliver timely and innovative products and services. **Analytics In Action 6.1** provides an example where customer generated textual contents (obtained from the internet and social media) is used as the voice of the customer to enhance loyalty and profitability.

ANALYTICS IN ACTION 6.1 Deliver Innovation by Understanding Customer Sentiments

Analyzing online reviews and social media provides valuable insights into what consumers want, how they interact with products, and where they encounter usability issues. These insights can lead to new feature designs and development, or even new products. Understanding customer sentiment and knowing what consumers truly think about products or a brand are traditional pain points. Customer journey analytics provide insights into these areas, yet these solutions are not all designed to integrate vital sources of unstructured data such as call center notes or social media feedback.

In today's world, unstructured notes are part of core communications in virtually every industry. For example:

- Medical professionals record patient observations.
- Auto technicians write down safety information.
- Retailers track social media for consumer comments.
- Call centers monitor customer feedback and take notes.

Bringing together notes, which are usually available as free-form text, with other data for analysis has been difficult. That's because each industry has its own unique terms, slang, shorthand, and acronyms embedded in the data. Finding meaning and business insights first requires the text to be changed into a structured form. This manual process is expensive, time consuming, and prone to errors, especially as data scales to ever-increasing volumes. One way companies can leverage notes without codifying the text is to use text clustering. This analytic technique quickly identifies common words or phrases for rapid insights.

GROUPING WORDS ISOLATES ISSUES

Clustering words or phrases allows companies to identify trends or popular themes. Analytics can be performed on a variety of text, including:

- Employee e-mails
- Customer complaints
- Call center notes
- Business forms
- Social media posts

Think Big Analytics, a Teradata company, recently performed a Teradata® Rapid Analytic Consulting Engagement™ (RACE™) for an automobile manufacturer. Data scientists and business consultants leveraged RACE because it allows high-value outcomes in just six weeks or less. Data scientists used Term Frequency-Inverse Document Frequency (TF-IDF) and Cosine Similarity to analyze text information related to automobile safety issues. TF-IDF determines how important an individual word is compared to other words in the document. Cosine Similarity brings together and compares the document against another document for similarities. This helps identify common phrases. Together, TF-IDF and Cosine Similarity are a powerful combination that identify important phrases or themes that are commonly seen in a large corpus of customer notes.

Both TF-IDF and Cosine Similarity are part of the larger effort to include natural language processing (NLP) capabilities to cluster phrases and then identify the underlying sentiments. Clustering helps isolate important areas of discussion. NLP provides an understanding into the nature of these discussions (e.g., positive, negative, neutral). With the auto manufacturer, the analytics isolated unique topics taken from call center and safety notes. The company could see specific issues and customer concerns. For example, the clusters showed customers with brake problems, power steering leaks, and other safety issues. Buckets could also refer to a safety problem at a specific factory. The manufacturer could then determine the origin of the problem and take corrective action.

TEXT AND NOTES CAN LEAD TO NEW AND IMPROVED PRODUCTS

Leveraging the insights and customer sentiment uncovered during a text and sentiment analysis can spark innovation. Companies such as vehicle manufacturers can use the intelligence to improve customer service and deliver an elevated customer experience. By learning what customers like and dislike about current products, companies can improve their design, such as adding new features to a car to enhance the driving experience.

Forming word clusters also allows companies to identify safety issues. If an auto manufacturer sees that numerous customers are expressing negative sentiment about black smoke coming from their cars, the company can respond. Likewise, manufacturers can address safety issues that are a concern to customers. With comments grouped into buckets, companies have the ability to focus on specific customers who experienced a similar problem. This allows a company to, for instance, offer a rebate or special promotion to those who experienced black smoke.

Understanding sentiments can better inform a car manufacturer's policies. For example, customers have different lifetime values. A customer who complains just once but has a very large lifetime value can be a more urgent candidate for complaint redressal than a customer with a lower lifetime value with multiple issues. One may have spent \$5,000 buying the vehicle from a used car lot. Another may have a history of buying new cars from the manufacturer and spent \$30,000 to buy the car off the showroom floor.

Using a Teradata analytical platform, companies can create a campaign that segments customers based on their value. This could allow repeat customers to get a \$500 rebate, with low value customers receiving a discount on their next service. Companies can also segment customers by location. A manufacturer may want to take action with customers complaining of black smoke who live in California immediately because of the state's stringent laws, whereas the company can wait until the next scheduled service for customers with the same issue in Nebraska.

EVERY INDUSTRY CAN BENEFIT FROM TEXT AND SENTIMENT ANALYSIS

All industries can use text clustering and analytics to innovate, address quality or safety issues, or improve customer service. Physician offices and hospitals can analyze medical transcriptions to uncover medical issues with a particular group of patients. Companies with call centers can identify common complaints by monitoring customer calls. Financial advisors can review emails to ensure compliance. Utility companies can identify safety issues with a utility grid.

Text clustering also allows retailers to quickly understand what product lines their customers are excited about on social media, such as a new line of luxury good products. A retailer can drill down into a cluster to see specific issues, which can reveal that customers are raving about a new men's cologne that was just released.

Understanding customer sentiments, patient issues, or product safety considerations empowers companies to make more informed decisions. This can result in creating or refining products, enacting new policies or procedures, or developing solutions that meet customers' evolving needs.

Source: Teradata Case Study. Deliver Innovation by Understanding Customer Sentiments, <https://assets.teradata.com/resourceCenter/downloads/CaseStudies/EB9859.pdf> (accessed June 2022).

NLP has successfully been applied to a variety of domains for a wide range of tasks via computer programs to automatically process natural human language that previously could only be done by humans. Following are among the most popular of these tasks:

- **Question answering.** The task of automatically answering a question posed in natural language; that is, producing a human language answer when given a human language question. To find the answer to a question, the computer program may use either a prestructured database or a collection of natural language documents (a text corpus such as the World Wide Web).
- **Automatic summarization.** The creation of a shortened version of a textual document by a computer program that contains the most important points of the original document.
- **Natural language generation.** Systems convert information from computer databases into readable human language.
- **Natural language understanding.** Systems convert samples of human language into more formal representations that are easier for computer programs to manipulate.
- **Machine translation.** The automatic translation of one human language to another.
- **Foreign language reading.** A computer program that assists a nonnative language speaker to read a foreign language with correct pronunciation and accents on different parts of the words.
- **Foreign language writing.** A computer program that assists a nonnative language user in writing in a foreign language.
- **Speech recognition.** Converts spoken words to machine-readable input. Given a sound clip of a person speaking, the system produces a text dictation.
- **Text-to-speech.** Also called *speech synthesis*, a computer program automatically converts normal language text into human speech.
- **Text proofing.** A computer program reads a proof copy of a text to detect and correct any errors.
- **Optical character recognition.** The automatic translation of images of handwritten, typewritten, or printed text (usually captured by a scanner) into machine-editable textual documents.

The success and popularity of text mining depends greatly on advancements in NLP in both generation as well as understanding of human languages. NLP enables the extraction of features from unstructured text so that a wide variety of data mining techniques can be used to extract knowledge (novel and useful patterns and relationships) from it. In that sense, simply put, text mining is a combination of NLP and data mining.

► SECTION 6.3 REVIEW QUESTIONS

1. What is NLP?
2. How does NLP relate to text mining?
3. What are some of the benefits and challenges of NLP?
4. What are the most common tasks addressed by NLP?

6.4 TEXT MINING APPLICATIONS

As the amount of unstructured data collected by organizations increases, so does the value proposition and popularity of text mining tools. Many organizations are now realizing the importance of extracting knowledge from their document-based data repositories through the use of text mining tools. Following is only a small subset of the exemplary application categories of text mining.

Marketing Applications

Text mining can be used to increase cross-selling and up-selling by analyzing the unstructured data generated by call centers. Text generated by call center notes as well as transcriptions of voice conversations with customers can be analyzed by text mining algorithms to extract novel, actionable information about customers' perceptions toward a company's products and services. In addition, blogs, user reviews of products at independent Web sites, and discussion board postings are a gold mine of customer sentiments. This rich collection of information, once properly analyzed, can be used to increase satisfaction and the overall lifetime value of the customer (Coussement & Van den Poel, 2008).

Text mining has become invaluable for CRM. Companies can use text mining to analyze rich sets of unstructured text data, combined with the relevant structured data extracted from organizational databases, to predict customer perceptions and subsequent purchasing behavior. Coussement and Van den Poel (2009) successfully applied text mining to significantly improve the ability of a model to predict customer churn (i.e., customer attrition) so that those customers identified as most likely to leave a company are accurately identified for retention tactics.

Ghani et al. (2006) used text mining to develop a system capable of inferring implicit and explicit attributes of products to enhance retailers' ability to analyze product databases. Treating products as sets of attribute–value pairs rather than as atomic entities can potentially boost the effectiveness of many business applications, including demand forecasting, assortment optimization, product recommendations, assortment comparison across retailers and manufacturers, and product supplier selection. The proposed system allows a business to represent its products in terms of attributes and attribute values without much manual effort. The system learns these attributes by applying supervised and semi-supervised learning techniques to product descriptions found on retailers' Web sites.

Security Applications

One of the largest and most prominent text mining applications in the security domain is probably the highly classified ECHELON surveillance system. As rumor has it, ECHELON is assumed to be capable of identifying the content of telephone calls, faxes, e-mails, and other types of data, intercepting information sent via satellites, public-switched telephone networks, and microwave links.

In 2007, EUROPOL developed an integrated system capable of accessing, storing, and analyzing vast amounts of structured and unstructured data sources to track transnational organized crime. Called the Overall Analysis System for Intelligence Support (OASIS), this system aims to integrate the most advanced data and text mining technologies available in today's market. The system has enabled EUROPOL to make significant progress in supporting its law enforcement objectives at the international level (EUROPOL, 2007).

The U.S. Federal Bureau of Investigation (FBI) and the Central Intelligence Agency (CIA), under the direction of the Department for Homeland Security, are jointly developing a supercomputer data and text mining system. The system is expected to create a gigantic data warehouse along with a variety of data and text mining modules to meet the knowledge-discovery needs of federal, state, and local law enforcement agencies.

Prior to this project, the FBI and CIA each had its own separate database, with little or no interconnection.

Another security-related application of text mining is in the area of **deception detection**. Applying text mining to a large set of real-world criminal (person-of-interest) statements, Fuller, Biros, and Delen (2008) developed prediction models to differentiate deceptive statements from truthful ones. Using a rich set of cues extracted from the textual statements, the model predicted the holdout samples with 70% accuracy, which is believed to be a significant success considering that the cues are extracted only from textual statements (no verbal or visual cues are present). Furthermore, compared to other deception-detection techniques, such as polygraph, this method is nonintrusive and widely applicable to not only textual data, but also (potentially) to transcriptions of voice recordings. A more detailed description of text mining-based deception detection is provided in Analytics in Action 6.2.

ANALYTICS IN ACTION 6.2 Mining for Lies

Driven by advancements in Web-based information technologies and increasing globalization, computer-mediated communication continues to filter into everyday life, bringing with it new venues for deception. The volume of text-based chat, instant messaging, text messaging, and text generated by online communities of practice is increasing rapidly. Even e-mail continues to grow in use. With the massive growth of text-based communication, the potential for people to deceive others through computer-mediated communication has also grown, and such deception can have disastrous results.

Unfortunately, in general, humans tend to perform poorly at deception-detection tasks. This phenomenon is exacerbated in text-based communications. A large part of the research on deception detection (also known as credibility assessment) has involved face-to-face meetings and interviews. Yet, with the growth of text-based communication, text-based deception-detection techniques are essential.

Techniques for successfully detecting deception—that is, lies—have wide applicability. Law enforcement can use decision support tools and techniques to investigate crimes, conduct security screening in airports, and monitor communications of suspected terrorists. Human resources professionals might use deception-detection tools to screen applicants. These tools and techniques also have the potential to screen e-mails to uncover fraud or other wrongdoings committed by corporate officers. Although some people believe that they can readily identify those who are not being truthful, a summary of deception research showed that, on average, people are only 54% accurate in making veracity determinations (Bond & DePaulo, 2006). This figure may actually be worse when humans try to detect deception in text.

Using a combination of text mining and data mining techniques, Fuller et al. (2008) analyzed person-of-interest statements completed by people involved in crimes on military bases. In these statements, suspects and witnesses are required to write their recollection of the event in their own words. Military law enforcement personnel searched archival data for statements that they could conclusively identify as being truthful or deceptive. These decisions were made on the basis of corroborating evidence and case resolution. Once labeled as truthful or deceptive, the law enforcement personnel removed identifying information and gave the statements to the research team. In total, 371 usable statements were received for analysis. The text-based deception-detection method used by Fuller et al. (2008) was based on a process known as message feature mining, which relies on elements of data and text mining techniques. A simplified depiction of the process is provided in Figure 6.3 (Fuller et al., 2011).

First, the researchers prepared the data for processing. The original handwritten statements had to be transcribed into a word processing file. Second, features (i.e., cues) were

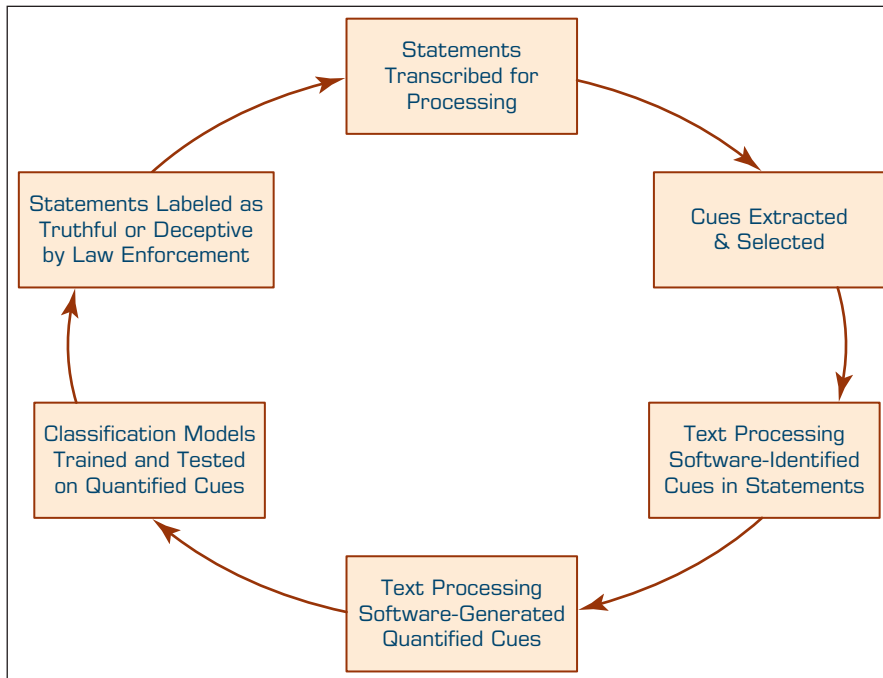


FIGURE 6.3 Text-Based Deception-Detection Process.

identified. The researchers identified 31 features representing categories or types of language that are relatively independent of the text content and that can be readily analyzed by automated means. For example, first-person pronouns such as I or me can be identified without analysis of the surrounding text. Table 6.1 lists the categories and an example list of features used in this study.

The features were extracted from the textual statements and input into a flat file for further processing. Using several feature-selection methods along with 10-fold cross-validation, the researchers compared the prediction accuracy of three popular data mining methods. Their results indicated that neural network models performed the best, with 73.46% prediction accuracy

TABLE 6.1 Categories and Examples of Linguistic Features Used in Deception Detection

Number	Construct (Category)	Example Cues
1	Quantity	Verb count, noun-phrase count, etc.
2	Complexity	Average number of clauses, average sentence length, etc.
3	Uncertainty	Modifiers, modal verbs, etc.
4	Nonimmediacy	Passive voice, objectification, etc.
5	Expressivity	Emotiveness
6	Diversity	Lexical diversity, redundancy, etc.
7	Informality	Typographical error ratio
8	Specificity	Spatiotemporal information, perceptual information, etc.
9	Affect	Positive affect, negative affect, etc.

on test data samples; decision trees performed second best, with 71.60% accuracy; and logistic regression was last, with 65.28% accuracy.

The results indicate that automated text-based deception detection has the potential to aid those who must try to detect lies in text and can be successfully applied to real-world data. The accuracy of these techniques exceeded the accuracy of most other deception-detection techniques, even though it was limited to textual cues.

Sources: Fuller, C. M., Biros, D. P., & Delen, D. (2011). An investigation of data and text mining methods for real world deception detection. *Expert Systems with Applications*, 38(7), 8392-8398; Fuller, C. M., Biros, D., & Delen, D. (2008). Exploration of feature selection and advanced classification models for high-stakes deception detection. *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS)*, Big Island, HI: IEEE Press, 80-99; Bond C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Reports*, 10(3), 214-234.

Biomedical Applications

Text mining holds great potential for the medical field in general and biomedicine in particular for several reasons. First, the published literature and publication outlets (especially with the advent of the open-source journals) in the field are expanding at an exponential rate. Second, compared to most other fields, the medical literature is more standardized and orderly, making it a more “minable” information source. Finally, the terminology used in this literature is relatively constant, having a fairly standardized ontology. What follows are a few exemplary studies where text mining techniques were successfully used in extracting novel patterns from biomedical literature.

Experimental techniques such as DNA microarray analysis, serial analysis of gene expression (SAGE), and mass spectrometry proteomics, among others, are generating large amounts of data related to genes and proteins. As in any other experimental approach, it is necessary to analyze this vast amount of data in the context of previously known information about the biological entities under study. The literature is a particularly valuable source of information for experiment validation and interpretation. Therefore, the development of automated text mining tools to assist in such interpretation is one of the main challenges in current bioinformatics research.

Knowing the location of a protein within a cell can help to elucidate its role in biological processes and to determine its potential as a drug target. Numerous location-prediction systems are described in the literature; some focus on specific organisms, whereas others attempt to analyze a wide range of organisms. Shatkay et al. (2007) proposed a comprehensive system that uses several types of sequence- and text-based features to predict the location of proteins. The main novelty of their system lies in the way in which it selects its text sources and features and integrates them with sequence-based features. They tested the system on previously used data sets and on new data sets devised specifically to test its predictive power. The results showed that their system consistently beat previously reported results.

Chun et al. (2006) described a system that extracts disease-gene relationships from literature accessed via MEDLINE. They constructed a dictionary for disease and gene names from six public databases and extracted relation candidates by dictionary matching. Because dictionary matching produces a large number of false positives, they developed a method of machine-learning-based named entity recognition (NER) to filter out false recognitions of disease/gene names. They found that the success of relation extraction is heavily dependent on the performance of NER filtering and that the filtering improved the precision of relation extraction by 26.7%, at the cost of a small reduction in recall.

Figure 6.4 shows a simplified depiction of a multilevel text analysis process for discovering gene-protein relationships (or protein-protein interactions) in the biomedical

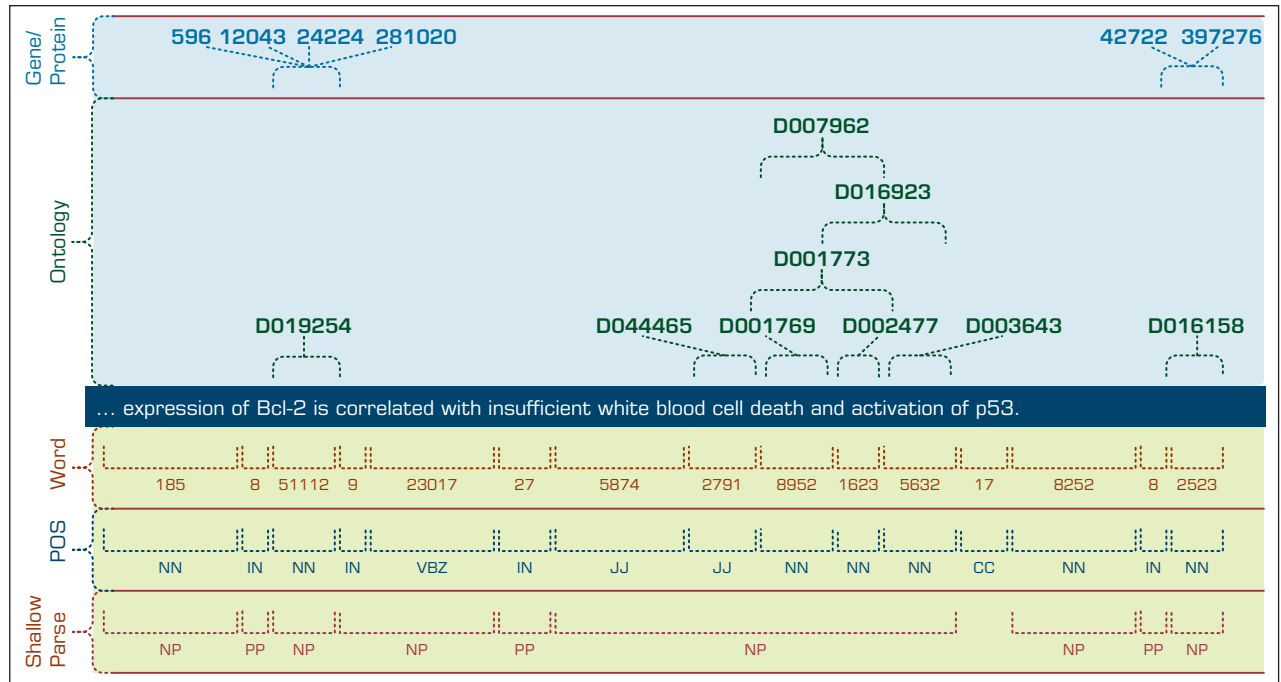


FIGURE 6.4 Multilevel Analysis of Text for Gene/Protein Interaction Identification. *Source:* Based on Nakov, P., Schwartz, A., Wolf, B., & Hearst, M. A. (2005). Supporting annotation layers for natural language processing. *Proceedings of the Association for Computational Linguistics (ACL), Interactive Poster and Demonstration Sessions, Ann Arbor, MI. Association for Computational Linguistics*, 65–68.

literature (Nakov et al., 2005). As can be seen in this simplified example that uses a simple sentence from biomedical text, first (at the bottom three levels) the text is tokenized using **part-of-speech tagging** and shallow-parsing. The tokenized terms (words) are then matched (and interpreted) against the hierarchical representation of the domain ontology to derive the gene–protein relationship. Application of this method (and/or some variation of it) to the biomedical literature offers great potential to decode the complexities in the Human Genome Project.

Academic Applications

The issue of text mining is of great importance to publishers who hold large databases of information requiring indexing for better retrieval. This is particularly true in scientific disciplines, in which highly specific information is often contained within written text. Initiatives have been launched, such as *Nature's* proposal for an Open Text Mining Interface and the National Institutes of Health's common Journal Publishing Document Type Definition, which would provide semantic cues to machines to answer specific queries contained within text without removing publisher barriers to public access.

Academic institutions have also launched text mining initiatives. For example, the National Centre for Text Mining, a collaborative effort between the Universities of Manchester and Liverpool, provides customized tools, research facilities, and advice on text mining to the academic community. With an initial focus on text mining in the biological and biomedical sciences, research has since expanded into the social sciences. In the United States, the School of Information at the University of California, Berkeley, is developing a program called BioText to assist bioscience researchers in text mining and analysis.

SECTION 6.4 REVIEW QUESTIONS

1. List and briefly discuss some of the text mining applications in marketing.
2. How can text mining be used in security and counterterrorism?
3. What are some promising text mining applications in biomedicine?

6.5 TEXT MINING PROCESS

To be successful, text mining studies should follow a sound methodology based on best practices. A standardized process model is needed similar to Cross-Industry Standard Process for Data Mining (CRISP-DM), which is the industry standard for data mining projects (see Chapter 4). Even though most parts of CRISP-DM are also applicable to text mining projects, a specific process model for text mining would include much more elaborate data preprocessing activities. Figure 6.5 depicts a high-level context diagram of a typical text mining process (Delen & Crossland, 2008). This context diagram presents the scope of the process, emphasizing its interfaces with the larger environment. In essence, it draws boundaries around the specific process to explicitly identify what is included in (and excluded from) the text mining process.

As the context diagram indicates, the input (inward connection to the left edge of the box) into the text-based knowledge-discovery process is the unstructured as well as structured data collected, stored, and made available to the process. The output (outward extension from the right edge of the box) of the process is the context-specific knowledge that can be used for decision-making. The controls, also called the *constraints* (inward connection to the top edge of the box), of the process include software and hardware limitations, privacy issues, and the difficulties related to processing of the text that is presented in the form of natural language. The mechanisms (inward connection to the bottom edge of the box) of the process include proper techniques, software tools, and domain expertise. The primary purpose of text mining (within the context of knowledge discovery) is to process unstructured (textual) data (along with structured data, if relevant to the problem being addressed and available) to extract meaningful and actionable patterns for better decision-making.

At a very high level, the text mining process can be broken down into three consecutive tasks, each of which has specific inputs to generate certain outputs (see Figure 6.6).

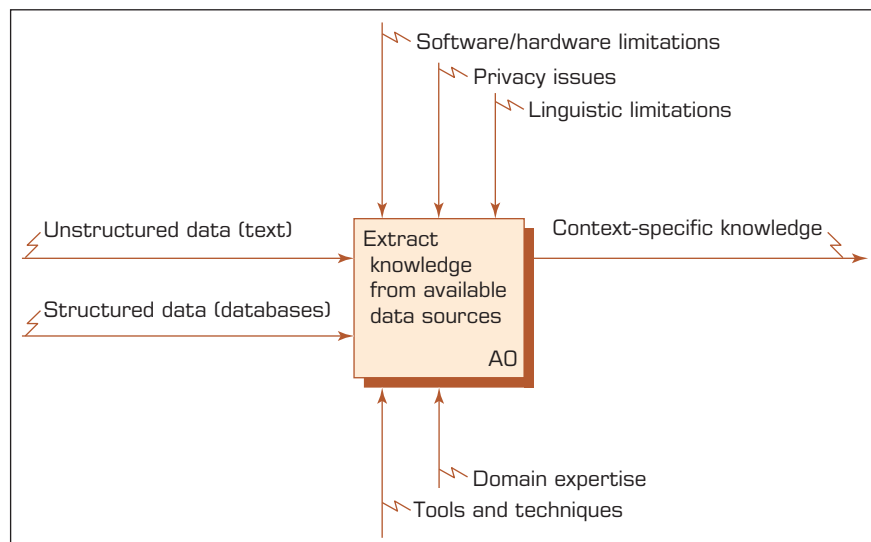


FIGURE 6.5 Context Diagram for the Text Mining Process.

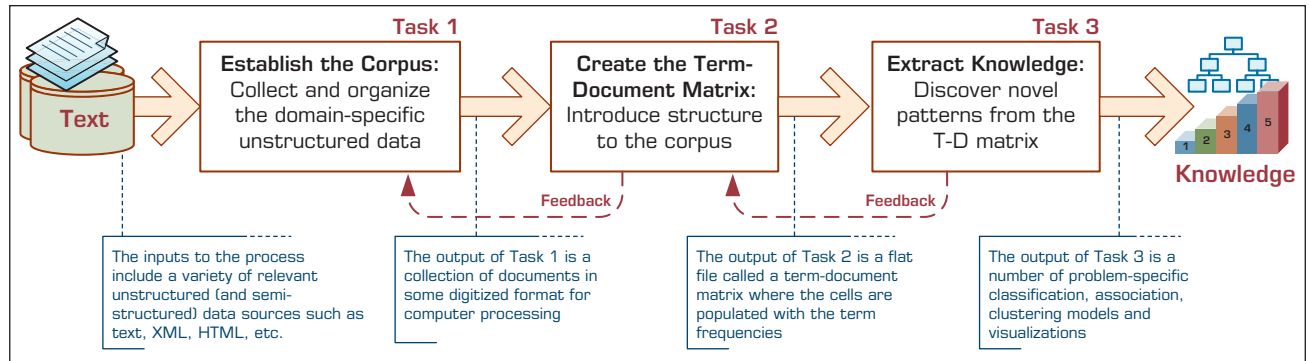


FIGURE 6.6 The Three-Step/Task Text Mining Process.

If, for some reason, the output of a task is not that which is expected, a backward redirection to the previous task execution is necessary.

Task 1: Establish the Corpus

The main purpose of the first task activity is to collect all the documents related to the context (domain of interest) being studied. This collection may include textual documents, XML files, e-mails, Web pages, and short notes. In addition to the readily available textual data, voice recordings may also be transcribed using speech-recognition algorithms and made a part of the text collection.

Once collected, the text documents are transformed and organized in a manner such that they are all in the same representational form (e.g., ASCII text files) for computer processing. The organization of the documents can be as simple as a collection of digitized text excerpts stored in a file folder or it can be a list of links to a collection of Web pages in a specific domain. Many commercially available text mining software tools could accept these as input and convert them into a flat file for processing. Alternatively, the flat file can be prepared outside the text mining software and then presented as the input to the text mining application.

Task 2: Create the Term–Document Matrix

In this task, the digitized and organized documents (the corpus) are used to create the **term–document matrix (TDM)**. In the TDM, rows represent the documents and columns represent the terms. The relationships between the terms and documents are characterized by indices (i.e., a relational measure that can be as simple as the number of occurrences of the term in respective documents). Figure 6.7 is a typical example of a TDM.

The goal is to convert the list of organized documents (the corpus) into a TDM where the cells are filled with the most appropriate indices. The assumption is that the essence of a document can be represented with a list and frequency of the terms used in that document. However, are all terms important when characterizing documents? Obviously, the answer is “no.” Some terms, such as articles, auxiliary verbs, and terms used in almost all the documents in the corpus, have no differentiating power and, therefore, should be excluded from the indexing process. This list of terms, commonly called *stop terms* or *stop words*, is specific to the domain of study and should be identified by the domain experts. On the other hand, one might choose a set of predetermined terms under which the documents are to be indexed (this list of terms is conveniently called

Terms Documents							
	Investment Risk	Project Management	Software Engineering	Development	SAP	...	
Document 1	1			1			
Document 2		1					
Document 3			3		1		
Document 4		1					
Document 5			2	1			
Document 6	1			1			
...							

FIGURE 6.7 A Simple Term–Document Matrix.

include terms or *dictionary*). In addition, synonyms (pairs of terms that are to be treated the same) and specific phrases (e.g., “Eiffel Tower”) can also be provided so that the index entries are more accurate.

Another filtration that should take place to accurately create the indices is *stemming*, which refers to the reduction of words to their roots so that, for example, different grammatical forms or declinations of a verb are identified and indexed as the same word. For example, stemming will ensure that *modeling* and *modeled* will be recognized as the word *model*.

The first generation of the TDM includes all the unique terms identified in the corpus (as its columns), excluding the ones in the stop term list; all the documents (as its rows); and the occurrence count of each term for each document (as its cell values). If, as is commonly the case, the corpus includes a rather large number of documents, then there is a very good chance that the TDM will have a very large number of terms. Processing such a large matrix might be time-consuming and, more important, might lead to extraction of inaccurate patterns. At this point, one has to decide the following: (1) What is the best representation of the indices? and (2) How can we reduce the dimensionality of this matrix to a manageable size?

REPRESENTING THE INDICES Once the input documents are indexed and the initial word frequencies (by document) computed, a number of additional transformations can be performed to summarize and aggregate the extracted information. The raw term frequencies generally reflect on how salient or important a word is in each document. Specifically, words that occur with greater frequency in a document are better descriptors of the contents of that document. However, it is not reasonable to assume that the word counts themselves are proportional to their importance as descriptors of the documents. For example, if a word occurs one time in document *A*, but three times in document *B*, then it is not necessarily reasonable to conclude that this word is three times as important a descriptor of document *B* as compared to document *A*. To have a more consistent TDM for further analysis, these raw indices need to be normalized. As opposed to showing the actual frequency counts, the numerical representation between terms and documents can be normalized using a number of alternative methods, such as log frequencies, binary frequencies, and inverse document frequencies, among others.

REDUCING THE DIMENSIONALITY OF THE MATRIX Because the TDM is often very large and rather sparse (most of the cells filled with zeros), another important question is, “How do we reduce the dimensionality of this matrix to a manageable size?” Several options are available for managing the matrix size.

- A domain expert goes through the list of terms and eliminates those that do not make much sense for the context of the study (this is a manual, labor-intensive process).
- Eliminate terms with very few occurrences in very few documents.
- Transform the matrix using SVD.

Singular value decomposition (SVD), which is closely related to principal components analysis, reduces the overall dimensionality of the input matrix (number of input documents by number of extracted terms) to a lower-dimensional space, where each consecutive dimension represents the largest degree of variability (between words and documents) possible (Manning & Schütze, 1999). Ideally, the analyst might identify the two or three most salient dimensions that account for most of the variability (differences) between the words and documents, thus identifying the latent semantic space that organizes the words and documents in the analysis. Once such dimensions are identified, the underlying “meaning” of what is contained (discussed or described) in the documents has been extracted.

Task 3: Extract the Knowledge

Using the well-structured TDM, and potentially augmented with other structured data elements, novel patterns are extracted in the context of the specific problem being addressed. The main categories of knowledge extraction methods are classification, clustering, association, and trend analysis. A short description of these methods follows.

CLASSIFICATION Arguably the most common knowledge-discovery topic in analyzing complex data sources is the **classification** (or categorization) of certain objects. The task is to classify a given data instance into a predetermined set of categories (or classes). As it applies to the domain of text mining, the task is known as *text categorization*, where for a given set of categories (subjects, topics, or concepts) and a collection of text documents the goal is to find the correct topic (subject or concept) for each document using models developed with a training data set that includes both the documents and actual document categories. Today, automated text classification is applied in a variety of contexts, including automatic or semiautomatic (interactive) indexing of text, spam filtering, Web page categorization under hierarchical catalogs, automatic generation of metadata, detection of genre, and many others.

The two main approaches to text classification are knowledge engineering and machine learning (Feldman & Sanger, 2007). With the knowledge-engineering approach, an expert’s knowledge about the categories is encoded into the system either declaratively or in the form of procedural classification rules. With the machine-learning approach, a general inductive process builds a classifier by learning from a set of reclassified examples. As the number of documents increases at an exponential rate and as knowledge experts become harder to come by, the popularity trend between the two is shifting toward the machine-learning approach.

CLUSTERING **Clustering** is an unsupervised process whereby objects are classified into “natural” groups called *clusters*. Compared to categorization, where a collection of pre-classified training examples is used to develop a model based on the descriptive features of the classes to classify a new unlabeled example, in clustering the problem is to group

an unlabeled collection of objects (e.g., documents, customer comments, Web pages) into meaningful clusters without any prior knowledge.

Clustering is useful in a wide range of applications, from document retrieval to enabling better Web content searches. In fact, one of the prominent applications of clustering is the analysis and navigation of very large text collections, such as Web pages. The basic underlying assumption is that relevant documents tend to be more similar to each other than to irrelevant ones. If this assumption holds, the clustering of documents based on the similarity of their content improves search effectiveness (Feldman & Sanger, 2007):

- **Improved search recall.** Clustering, because it is based on overall similarity as opposed to the presence of a single term, can improve the recall of a query-based search in such a way that when a query matches a document its whole cluster is returned.
- **Improved search precision.** Clustering can also improve search precision. As the number of documents in a collection grows, it becomes difficult to browse through the list of matched documents. Clustering can help by grouping the documents into a number of much smaller groups of related documents, ordering them by relevance and returning only the documents from the most relevant group (or groups).

The two most popular clustering methods are scatter/gather clustering and query-specific clustering:

- **Scatter/gather.** This document browsing method uses clustering to enhance the efficiency of human browsing of documents when a specific search query cannot be formulated. In a sense, the method dynamically generates a table of contents for the collection and adapts and modifies it in response to the user selection.
- **Query-specific clustering.** This method employs a hierarchical clustering approach where the most relevant documents to the posed query appear in small tight clusters that are nested in larger clusters containing less-similar documents, creating a spectrum of relevance levels among the documents. This method performs consistently well for document collections of realistically large sizes.

ASSOCIATION A formal definition and detailed description of **association** was provided in the chapter on data mining (Chapter 4). Associations or association rule learning in data mining is a popular and well-researched technique for discovering interesting relationships among variables in large databases. The main idea in generating association rules (or solving market-basket problems) is to identify the frequent sets that go together.

In text mining, associations specifically refer to the direct relationships between concepts (terms) or sets of concepts. The concept set association rule $A + C$ relating two frequent concept sets A and C can be quantified by the two basic measures of support and confidence. In this case, confidence is the percentage of documents that include all the concepts in C within the same subset of those documents that include all the concepts in A . Support is the percentage (or number) of documents that include all the concepts in A and C . For instance, in a document collection the concept “Software Implementation Failure” may appear most often in association with “Enterprise Resource Planning” and “Customer Relationship Management” with significant support (4%) and confidence (55%), meaning that 4% of the documents had all three concepts represented together in the same document, and of the documents that included “Software Implementation Failure,” 55% of them also included “Enterprise Resource Planning” and “Customer Relationship Management.”

Text mining with association rules was used to analyze published literature (news and academic articles posted on the Web) to chart the outbreak and progress of the bird flu (Mahgoub et al., 2008). The idea was to automatically identify the association among the geographic areas, spreading across species, and countermeasures (treatments).

TREND ANALYSIS Recent methods of trend analysis in text mining have been based on the notion that the various types of concept distributions are functions of document collections; that is, different collections lead to different concept distributions for the same set of concepts. It is, therefore, possible to compare two distributions that are otherwise identical except that they are from different subcollections. One notable direction of this type of analysis is having two collections from the same source (such as from the same set of academic journals) but from different points in time. Delen and Crossland (2008) applied **trend analysis** to a large number of academic articles (published in the three highest-rated academic journals) to identify the evolution of key concepts in the field of information systems.

As described in this section, a number of methods are available for text mining. Analytics in Action 6.3 describes the use of clustering and topic modeling techniques in analyzing a large collection of academic literature.

ANALYTICS IN ACTION 6.3 Research Literature Survey with Text Mining

Researchers conducting searches and reviews of relevant literature face an increasingly complex and voluminous task. In extending the body of relevant knowledge, it has always been important to work hard to gather, organize, analyze, and assimilate existing information from the literature, particularly from one's home discipline. With the increasing abundance of potentially significant research being reported in related fields, and even in what are traditionally deemed to be nonrelated fields of study, the researcher's task is ever more daunting, if a thorough job is desired. For instance, in the last couple of years a large volume of papers is published on COVID-19 pandemic. A recent study by Zengul et al. (2022) proposes a semi-automated text-mining process to discover the major research topics and related trends in this field using a NIH curated collection of 85,268 COVID-19-related articles (published within the first year of the pandemic, until November 2021).

In new streams of research, such as the ones dealing with information systems and technologies, the researcher's task is often overly time-demanding and complex. Trying to ferret out relevant work that others have reported may be difficult, at best, and perhaps even near impossible if traditional, largely manual reviews of published literature are required. Even with a legion of dedicated graduate students or helpful colleagues, trying to cover all potentially relevant published work is problematic.

Many scholarly conferences take place every year. In addition to extending the body of knowledge of the current focus of a conference, organizers often desire to offer additional minitracks and workshops. In many cases, these additional events are intended to introduce the attendees to significant streams of research in related fields of study and to try to identify the "next big thing" in terms of research interests and focus. Identifying reasonable candidate topics for such minitracks and workshops is often subjective rather than derived objectively from the existing and emerging research.

In the study referenced herein, Delen and Crossland (2008) proposed a method to assist and greatly enhance the efforts of the researchers by enabling a semiautomated analysis of large volumes of published literature through the application of text mining. Using standard digital libraries and online publication search engines, the authors downloaded and collected all the available articles for the three major journals in the field of management information systems: MIS Quarterly (MISQ), Information Systems Research (ISR), and the Journal of Management Information Systems (JMIS). To maintain the same time interval for all three journals (for potential comparative longitudinal studies), the journal with the most recent starting date for its digital publication availability was used as the start time for this study (i.e., JMIS articles have been digitally available since 1994). For each article, they extracted the title, abstract, author list, published keywords, volume, issue number, and year of publication. They then loaded all the article data into a simple database file. Also included in the combined data set was a field that designated the journal type of each article for likely discriminatory analysis. Editorial notes, research notes, and executive overviews were omitted from the collection. Table 6.2 shows how the data were presented in a tabular format.

TABLE 6.2 Tabular Representation of the Fields Included in the Combined Data Set

A1		fx		ID						
	A	B	C	D	E	F	G	H	I	J
1	ID	YEAR	JOURNAL	ABSTRACT						
2	PID001	2005	MISQ	The need for continual value innovation is driving supply chains to evolve from						
3	PID002	1999	ISR	Although much contemporary thought considers advanced information technoc						
4	PID003	2001	JMIS	When producers of goods (or services) are confronted by a situation in which						
5	PID004	1995	ISR	Preservation of organizational memory becomes increasingly important to org						
6	PID005	1994	ISR	The research reported here is an adaptation of a model developed to measure						
7	PID006	1995	MISQ	This study evaluates the extent to which the added value to customers from a						
8	PID007	2003	MISQ	This paper reports the results(-) of a field-study of six medical project teams t						
9	PID008	1999	JMIS	Researchers and managers are beginning to realize that the full advantages c						
10	PID009	2000	JMIS	The Internet commerce technologies have significantly reduced sellers' costs						
11	PID010	1997	ISR	Adaptive Structuration Theory (AST) is rapidly becoming an influential theoret						
12	PID011	1995	JMIS	Research shows that group support systems (GSS) have dramatically increa						
13	PID012	2000	MISQ	Increasingly, business leaders are demanding that IT play the role of a busine						
14	PID013	2001	ISR	Alignment between business strategy and IS strategy is widely believed to in						
15	PID014	1999	JMIS	A framework is outlined that includes the planning of and setting goals for IT,						
16	PID015	1999	JMIS	The continuously growing importance of information technology (IT) requires c						
17	PID016	1994	MISQ	Identifying the best way to organize the IS functions within an enterprise has b						
18	PID017	1996	ISR	Reasons for the mixed reactions to todays electronic off-exchange trading sy						
19	PID018	1996	JMIS	The performance impacts of information technology investments in organizati						
20	PID019	1997	JMIS	Anonymity is a fundamental concept in group support systems (GSS) resear						
21	PID020	2002	ISR	Although electronic commerce (EC) has created new opportunities for busine						
22	PID021	2005	JMIS	Understanding the successful adoption of information technology is largely ba						
23	PID022	2005	MISQ	Enterprise resource planning (ERP) systems and other complex information s						
24	PID023	1994	JMIS	Model management systems support modelers in various phases of the mode						
25	PID024	1995	ISR	While computer training is widely recognized as an essential contributor to th						

In the analysis phase, they chose to use only the abstract of an article as the source of information extraction. They chose not to include the keywords listed with the publications for two main reasons: (1) under normal circumstances, the abstract would already include the listed keywords, and therefore inclusion of the listed keywords for the analysis would mean repeating the same information and potentially giving them unmerited weight; and (2) the listed keywords may be terms that authors would like their article to be associated with (as opposed to what is really contained in the article), therefore potentially introducing unquantifiable bias to the analysis of the content.

The first exploratory study was to look at the longitudinal perspective of the three journals (i.e., evolution of research topics over time). In order to conduct a longitudinal study, they divided the 12-year period (from 1994 to 2005) into four 3-year periods for each of the three journals. This framework led to 12 text mining experiments with 12 mutually exclusive data sets. At this point, for each of the 12 data sets they used text mining to extract the most descriptive terms from these collections of articles represented by their abstracts. The results were tabulated and examined for time-varying changes in the terms published in these three journals.

As a second exploration, using the complete data set (including all three journals and all four periods), they conducted a clustering analysis. Clustering is arguably the most commonly used text mining technique. Clustering was used in this study to identify the natural groupings of the articles (by putting them into separate clusters) and then to list the most descriptive terms that characterized those clusters. They used SVD to reduce the dimensionality of the term-by-document matrix and then an expectation-maximization algorithm to create the clusters. They conducted several experiments to identify the optimal number of clusters, which turned out to be nine. After the construction of the nine clusters, they analyzed the content of those clusters from two perspectives: (1) representation of the journal type (see Figure 6.8a) and (2) representation of time (Figure 6.8b). The idea was to explore the potential differences and/or commonalities among the three journals and potential changes in the emphasis on those clusters; that is, to answer questions such as “Are there clusters that represent different research themes specific to a single journal?” and “Is there a time-varying characterization of those clusters?” They discovered and discussed several interesting patterns using tabular and graphical representation of their findings (for further information see Delen & Crossland, 2008).

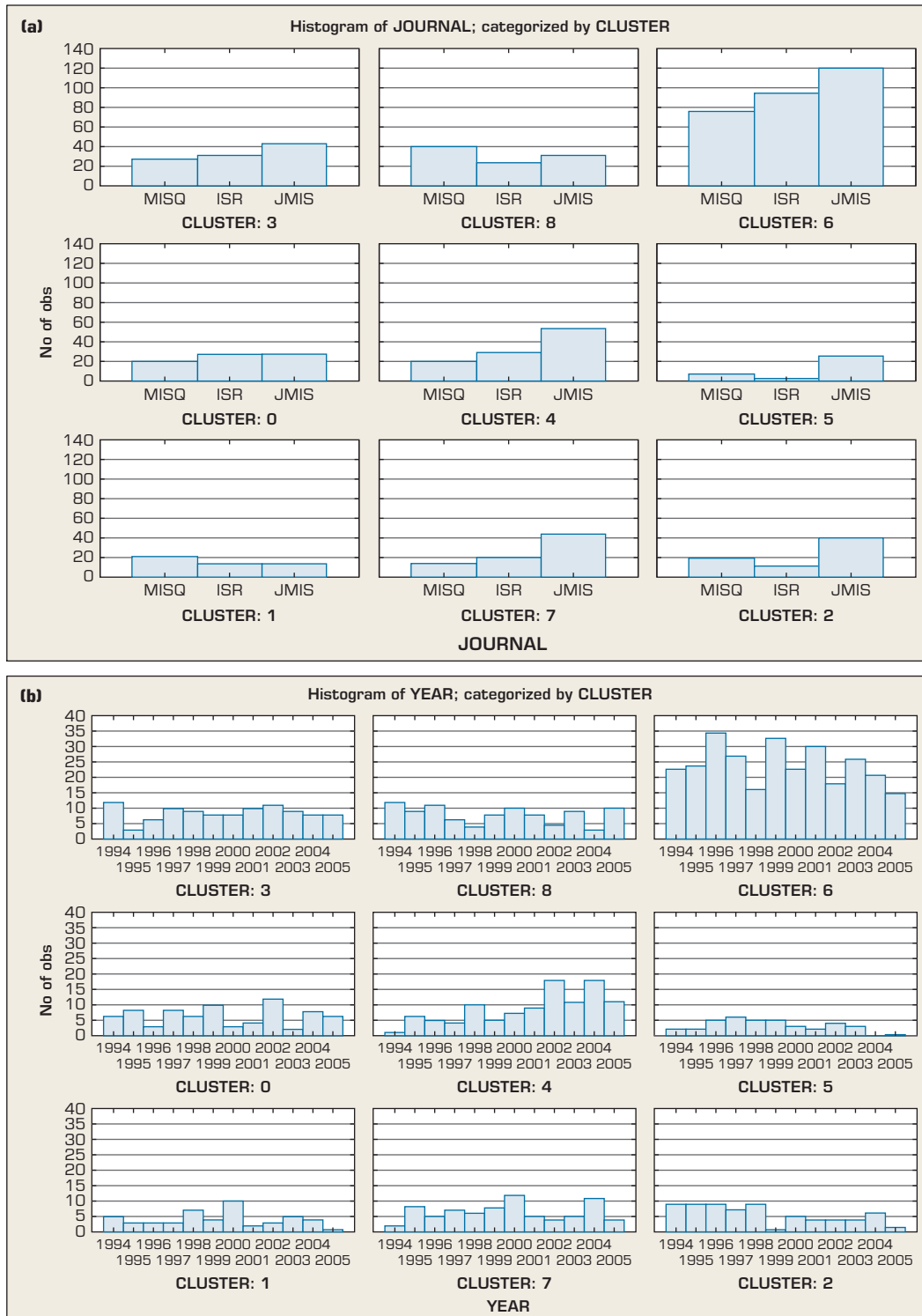


FIGURE 6.8 (a) Distribution of the Number of Articles for the Three Journals over the Nine Clusters; (b) Development of the Nine Clusters over the Years. Source: Delen, D., & Crossland, M. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, 34(3), 1707–1720. Source: Zengul, F. D., Zengul, A. G., Mugavero, M. J., Oner, N., Ozaydin, B., Delen, D., ... & Cimino, J. (2022). A critical analysis of COVID-19 research literature: Text mining approach. *Intelligence-based Medicine*, 5, 100036; Delen, D., & Crossland, M. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, 34(3), 1707–1720.

► SECTION 6.5 REVIEW QUESTIONS

1. What are the main steps in the text mining process?
2. What is the reason for normalizing word frequencies? What are the common methods for normalizing word frequencies?
3. What is SVD? How is it used in text mining?
4. What are the main knowledge extraction methods from corpus?

6.6 SENTIMENT ANALYSIS AND TOPIC MODELING

Text analytics has numerous application areas. Wherever there is enough volume of textual content, there usually is a plenty of opportunities, and hence, plenty of already completed and published/publicized text mining applications and case studies exist. Among them there are two application areas that seem to be leading the pack: sentiment analytics and topic modeling. The following sections provide an overview on the conceptual and technical details of these two text mining application areas.

Sentiment Analysis

We humans are social beings. We are adept at utilizing a variety of means to communicate. We often consult financial discussion forums before making an investment decision; ask our friends for their opinions on a newly opened restaurant or a newly released movie; and conduct Internet searches and read consumer reviews and expert reports before making a big purchase like a house, a car, or an appliance. We rely on others' opinions to make better decisions, especially in an area where we don't have a lot of knowledge or experience. Thanks to the growing availability and popularity of opinion-rich Internet resources such as social media outlets (e.g., Twitter, Facebook), online review sites, and personal blogs, it is now easier than ever to find opinions of others (thousands of them, as a matter of fact) on everything from the latest gadgets to political and public figures. Even though not everybody expresses opinions over the Internet—due mostly to the fast-growing number and capabilities of social communication channels—the numbers are increasing exponentially.

Sentiment is a difficult word to define. It is often linked to or confused with other terms like *belief*, *view*, *opinion*, and *conviction*. Sentiment suggests a settled opinion reflective of one's feelings (Mejova, 2009). Sentiment has some unique properties that set it apart from other concepts that we may want to identify in text. Often we want to categorize text by topic, which may involve dealing with whole taxonomies of topics. Sentiment classification, on the other hand, usually deals with two classes (positive versus negative), a range of polarity (e.g., star ratings for movies), or even a range in strength of opinion (Pang & Lee, 2008). These classes span many topics, users, and documents. Although dealing with only a few classes may seem like an easier task than standard text analysis, this is far from the truth.

As a field of research, sentiment analysis is closely related to computational linguistics, NLP, and text mining. Sentiment analysis has many names. It's often referred to as opinion mining, subjectivity analysis, and appraisal extraction, with some connections to affective computing (computer recognition and expression of emotion). The sudden upsurge of interest and activity in the area of sentiment analysis (i.e., opinion mining), which deals with the automatic extraction of opinions, feelings, and subjectivity in text, is creating opportunities and threats for businesses and individuals alike. The ones who embrace and take advantage of it will greatly benefit from it. Every opinion put on the Internet by an individual or a company will be accredited to the originator (good or bad) and will be retrieved and mined by others (often automatically by computer programs).

Sentiment analysis is trying to answer the question “What do people feel about a certain topic?” by digging into opinions of many using a variety of automated tools. Bringing together researchers and practitioners in business, computer science, computational linguistics, data mining, text mining, psychology, and even sociology, sentiment analysis aims to expand the traditional fact-based text analysis to new frontiers, to realize opinion-oriented information systems. In a business setting, especially in marketing and CRM, sentiment analysis seeks to detect favorable and unfavorable opinions toward specific products and/or services using large numbers of textual data sources (customer feedback in the form of Web postings, tweets, blogs, etc.).

Sentiment that appears in text comes in two flavors: explicit, where the subjective sentence directly expresses an opinion (“It’s a wonderful day”), and implicit, where the text implies an opinion (“The handle breaks too easily”). Most of the earlier work done in sentiment analysis focused on the first kind of sentiment because it is easier to analyze. Current trends are to implement analytical methods to consider both implicit and explicit sentiments. Sentiment polarity is a particular feature of text that sentiment analysis primarily focuses on. It is usually dichotomized into two—positive and negative—but polarity can also be thought of as a range. A document containing several opinionated statements will have a mixed polarity overall, which is different from not having a polarity at all (being objective or factual. Timely collection and analysis of textual data, which may be coming from a variety of sources—ranging from customer call center transcripts to social media postings—is a crucial part of the capabilities of proactive and customer-focused companies, nowadays. These real-time analyses of textual data are often visualized in easy-to-understand dashboards. Analytics in Action 6.4 provides a customer success story, where a collection of analytics solutions is used to enhance viewers’ experience at the Wimbledon tennis tournament.

ANALYTICS IN ACTION 6.4 Creating a Unique Digital Experience to Capture the Moments That Matter at Wimbledon

Known to millions of fans simply as “Wimbledon,” The Championships is the oldest of tennis’ four Grand Slams, and one of the world’s highest-profile sporting events. Organized by the All England Lawn Tennis Club (AELTC) it has been a global sporting and cultural institution since 1877.

The organizers of The Championships, Wimbledon, the AELTC, have a simple objective: every year, they want to host the best tennis championships in the world—in every way, and by every metric.

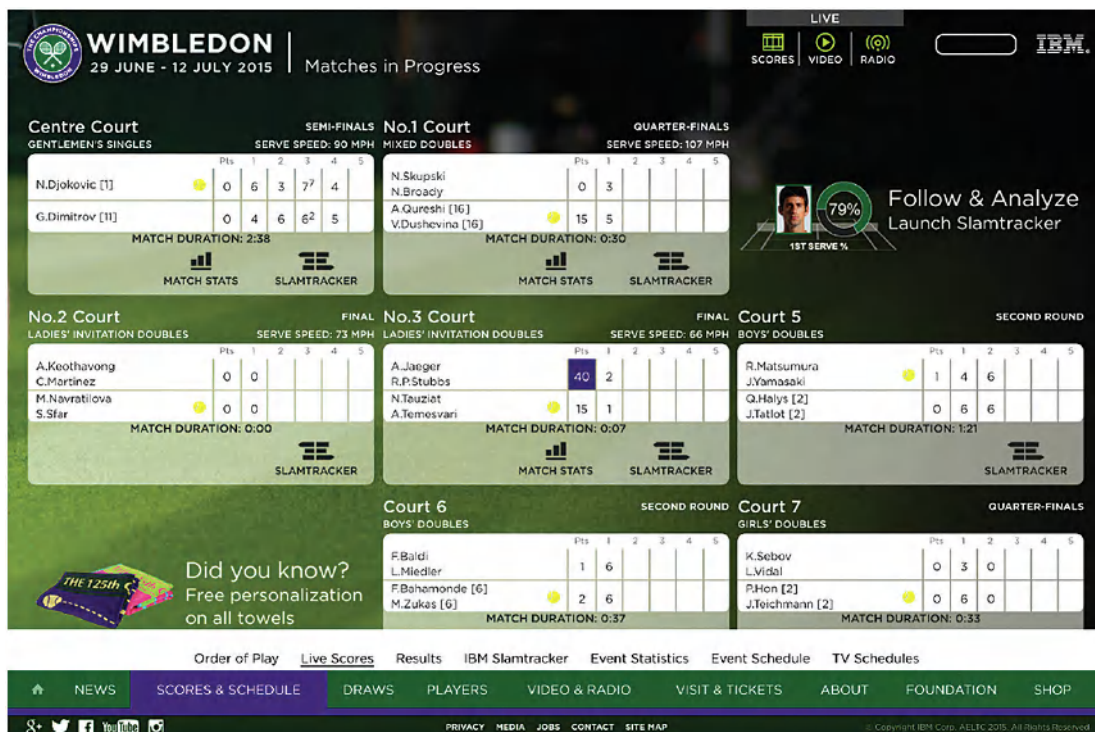
The motivation behind this commitment is not simply pride; it also has a commercial basis. Wimbledon’s brand is built on its premier status: this is what attracts both fans and partners. The world’s best media organizations and greatest corporations—IBM included—want to be associated with Wimbledon precisely because of its reputation for excellence.

For this reason, maintaining the prestige of The Championships is one of the AELTC’s top priorities, but there are only two ways that the organization can directly control how The Championships are perceived by the rest of the world.

The first, and most important, is to provide an outstanding experience for the players, journalists, and spectators who are lucky enough to visit and watch the tennis courtside. The AELTC has vast experience in this area. Since 1877 it has delivered two weeks of memorable, exciting competition in an idyllic setting: tennis in an English country garden.

The second is The Championships’ online presence, which is delivered via the wimbledon.com Web site, mobile apps, and social media channels. The constant evolution of these digital platforms is the result of a 26-year partnership between the AELTC and IBM.

Mick Desmond, Commercial and Media Director at the AELTC, explains: “When you watch Wimbledon on TV, you are seeing it through the broadcaster’s lens. We do everything we can to help our media partners put on the best possible show, but at the end of the day, their broadcast is their presentation of The Championships.



"Digital is different: it's our platform, where we can speak directly to our fans—so it's vital that we give them the best possible experience. No sporting event or media channel has the right to demand a viewer's attention, so if we want to strengthen our brand, we need people to see our digital experience as the number-one place to follow The Championships online."

To that end, the AELTC set a target of attracting 70 million visits, 20 million unique devices, and 8 million social followers during the two weeks of The Championships 2015. It was up to IBM and AELTC to find a way to deliver.

DELIVERING A UNIQUE DIGITAL EXPERIENCE

IBM and the AELTC embarked on a complete redesign of the digital platform, using their intimate knowledge of The Championships' audience to develop an experience tailor-made to attract and retain tennis fans from across the globe.

"We recognized that while mobile is increasingly important, 80% of our visitors are using desktop computers to access our website," says Alexandra Willis, Head of Digital and Content at the AELTC. "Our challenge for 2015 was how to update our digital properties to adapt to a mobile-first world, while still offering the best possible desktop experience. We wanted our new site to take maximum advantage of that large screen size and give desktop users the richest possible experience in terms of high-definition visuals and video content—while also reacting and adapting seamlessly to smaller tablet or mobile formats.

"Second, we placed a major emphasis on putting content in context—integrating articles with relevant photos, videos, stats and snippets of information, and simplifying the navigation so that users could move seamlessly to the content that interests them most."

On the mobile side, the team recognized that the wider availability of high-bandwidth 4G connections meant that the mobile Web site would become more popular than ever—and ensured that it would offer easy access to all rich media content. At the same time, The Championships' mobile apps were enhanced with real-time notifications of match scores and events—and could even greet visitors as they passed through stations on the way to the grounds.

The team also built a special set of Web sites for the most important tennis fans of all: the players themselves. Using IBM® Bluemix® technology, it built a secure Web application

that provided players with a personalized view of their court bookings, transport, and on-court times, as well as helping them review their performance with access to stats on every match they played.

TURNING DATA INTO INSIGHT—AND INSIGHT INTO NARRATIVE

To supply its digital platforms with the most compelling possible content, the team took advantage of a unique advantage: its access to real-time, shot-by-shot data on every match played during The Championships. Over the course of the Wimbledon fortnight, 48 courtside experts capture approximately 3.4 million data points, tracking the type of shot, the strategies, and the outcome of each and every point.

These data are collected and analyzed in real time to produce statistics for TV commentators and journalists—and also for the digital platform's own editorial team.

"This year IBM gave us an advantage that we had never had before—using data streaming technology to provide our editorial team with real-time insight into significant milestones and breaking news," says Alexandra Willis.

"The system automatically watched the streams of data coming in from all 19 courts, and whenever something significant happened—such as Sam Groth hitting the second-fastest serve in Championships, history—it let us know instantly. Within seconds, we were able to bring that news to our digital audience and share it on social media to drive even more traffic to our site.

"The ability to capture the moments that matter and uncover the compelling narratives within the data, faster than anyone else, was key. If you wanted to experience the emotions of The Championships live, the next best thing to being there in person was to follow the action on wimbledon.com."

HARNESSING THE POWER OF NATURAL LANGUAGE

Another new capability trialed this year was the use of IBM's NLP technologies to help mine the AELTC's huge library of tennis history for interesting contextual information. The team trained IBM Watson™ Engagement Advisor to digest this rich unstructured data set and use it to answer queries from the press desk.

The same NLP front-end was also connected to a comprehensive structured database of match statistics, dating back to the first Championships in 1877—providing a one-stop shop for both basic questions and more complex inquiries.

"The Watson trial showed a huge amount of potential. Next year, as part of our annual innovation planning process, we will look at how we can use it more widely—ultimately in pursuit of giving fans more access to this incredibly rich source of tennis knowledge," says Mick Desmond.

The whole digital environment was hosted by IBM in its Hybrid Cloud. IBM used sophisticated modeling techniques to predict peaks in demand based on the schedule, the popularity of each player, the time of day, and many other factors—enabling it to dynamically allocate cloud resources appropriately to each piece of digital content and ensure a seamless experience for millions of visitors around the world.

In addition to the powerful private cloud platform that has supported The Championships for several years, IBM also used a separate SoftLayer® cloud to host the Wimbledon Social Command Centre and also provide additional incremental capacity to supplement the main cloud environment during times of peak demand.

The elasticity of the cloud environment is key, as The Championships' digital platforms need to be able to scale efficiently by a factor of more than 100 within a matter of days as the interest builds ahead of the first match on Centre Court.

KEEPING WIMBLEDON SAFE AND SECURE

Online security is a key concern nowadays for all organizations. For major sporting events in particular, brand reputation is everything—and while the world is watching, it is particularly important to avoid becoming a high-profile victim of cyber-crime. For these reasons, security has a vital role to play in IBM's partnership with the AELTC.

Over the first five months of 2015, IBM security systems detected a 94% increase in security events on the wimbledon.com infrastructure, compared to the same period in 2014.

As security threats—and in particular distributed denial of service (DDoS) attacks—become ever more prevalent, IBM continually increases its focus on providing industry-leading levels of security for the AELTC's whole digital platform.

A full suite of IBM security products, including IBM QRadar® SIEM and IBM Preventia Intrusion Prevention, enabled this year's Championships to run smoothly and securely and the digital platform to deliver a high-quality user experience at all times.

CAPTURING HEARTS AND MINDS

The success of the new digital platform for 2015—supported by IBM cloud, analytics, mobile, social, and security technologies—was immediate and complete. Targets for total visits and unique visitors were not only met, but exceeded. Achieving 71 million visits and 542 million page views from 21.1 million unique devices demonstrates the platform's success in attracting a larger audience than ever before and keeping those viewers engaged throughout The Championships.

“Overall, we had 13% more visits from 23% more devices than in 2014, and the growth in the use of wimbledon.com on mobile was even more impressive,” says Alexandra Willis. “We saw 125% growth in unique devices on mobile, 98% growth in total visits, and 79% growth in total page views.”

Mick Desmond concludes: “The results show that in 2015, we won the battle for fans' hearts and minds. People may have favorite newspapers and sports websites that they visit for 50 weeks of the year—but for two weeks, they came to us instead.

“That's a testament to the sheer quality of the experience we can provide—harnessing our unique advantages to bring them closer to the action than any other media channel. The ability to capture and communicate relevant content in real time helped our fans experience The Championships more vividly than ever before.”

Source: IBM Case Study. Wimbledon. (1) Protecting the oldest brand in tennis with the latest in cognitive security, www.ibm.com/case-studies/wimbledon-2017; (2) Digital experience for the global audience. www.ibm.com/services/ibmix/case-studies/wimbledon.html (accessed July 2022).

Sentiment Analysis Applications

Compared to traditional sentiment analysis methods, which were survey based or focus group centered, costly, and time consuming (and therefore driven from a small sample of participants), the new face of text analytics-based sentiment analysis is a limit breaker. Current solutions automate very large-scale data collection, filtering, classification, and clustering methods via NLP and data mining technologies that handle both factual and subjective information. Sentiment analysis is perhaps the most popular application of text analytics, tapping into data sources like tweets, Facebook posts, online communities, discussion boards, Web logs, product reviews, call center logs and recordings, product rating sites, chat rooms, price comparison portals, search engine logs, and newsgroups. The following applications of sentiment analysis are meant to illustrate the power and the widespread coverage of this technology.

VOICE OF THE CUSTOMER (VOC) **Voice of the customer** (VOC) is an integral part of analytic CRM and customer experience management systems. As the enabler of VOC, sentiment analysis can access a company's product and service reviews (either continuously or periodically) to better understand and better manage customer complaints and praises. For instance, a motion picture advertising/marketing company may detect negative sentiments toward a movie that is about to open in theatres (based on its trailers) and quickly change the composition of trailers and advertising strategy (on all media outlets) to mitigate the negative impact. Similarly, a software company may detect the negative

buzz regarding the bugs found in their newly released product early enough to release patches and quick fixes to alleviate the situation.

Often, the focus of VOC is individual customers, their service- and support-related needs, wants, and issues. VOC draws data from the full set of customer touch points, including e-mails, surveys, call center notes/recordings, and social media postings, and matches customer voices to transactions (inquiries, purchases, returns) and individual customer profiles captured in enterprise operational systems. VOC, mostly driven by sentiment analysis, is a key element of customer experience management initiatives, where the goal is to create an intimate relationship with the customer.

VOICE OF THE MARKET (VOM) VOM is about understanding aggregate opinions and trends. It's about knowing what stakeholders—customers, potential customers, influencers, whoever—are saying about your (and your competitors') products and services. A well-done VOM analysis helps companies with competitive intelligence and product development and positioning.

VOICE OF THE EMPLOYEE (VOE) Traditionally, VOE has been limited to employee satisfaction surveys. Text analytics in general (and sentiment analysis in particular) is a huge enabler of assessing the VOE. Using rich, opinionated textual data is an effective and efficient way to listen to what employees are saying. As we all know, happy employees empower customer experience efforts and improve customer satisfaction.

BRAND MANAGEMENT Brand management focuses on listening to social media where anyone (past/current/prospective customers, industry experts, other authorities) can post opinions that can damage or boost your reputation. A number of relatively newly launched start-up companies offer analytics-driven brand management services for others. Brand management is product and company (rather than customer) focused. It attempts to shape perceptions rather than to manage experiences using sentiment analysis techniques.

FINANCIAL MARKETS Predicting the future values of individual (or a group of) stocks has been an interesting and seemingly unsolvable problem. What makes a stock (or a group of stocks) move up or down is anything but an exact science. Many believe that the stock market is mostly sentiment driven, making it anything but rational (especially for short-term stock movements). Therefore, the use of sentiment analysis in financial markets has gained significant popularity. Automated analysis of market sentiment using social media, news, blogs, and discussion groups seems to be a proper way to compute the market movements. If done correctly, sentiment analysis can identify short-term stock movements based on the buzz in the market, potentially impacting liquidity and trading.

POLITICS As we all know, opinions matter a great deal in politics. Because political discussions are dominated by quotes, sarcasm, and complex references to persons, organizations, and ideas, politics is one of the most difficult, and potentially fruitful, areas for sentiment analysis. By analyzing the sentiment on election forums, one may predict who is more likely to win or lose. Sentiment analysis can help understand what voters are thinking and can clarify a candidate's position on issues. Sentiment analysis can help political organizations, campaigns, and news analysts to better understand which issues and positions matter the most to voters. The technology was successfully applied by both parties to the 2008 and 2012 American presidential election campaigns.

GOVERNMENT INTELLIGENCE Government intelligence is another application that has been used by intelligence agencies. For example, it has been suggested that one could monitor sources for increases in hostile or negative communications. Sentiment analysis

can allow the automatic analysis of the opinions that people submit about pending policy or government-regulation proposals. Furthermore, monitoring communications for spikes in negative sentiment may be of use to agencies like Homeland Security.

OTHER INTERESTING AREAS Sentiments of customers can be used to better design e-commerce sites (product suggestions, up-sell/cross-sell advertising), better place advertisements (e.g., placing dynamic advertisements of products and services that consider the sentiment on the page the user is browsing), and manage opinion- or review-oriented search engines (i.e., an opinion-aggregation Web site, an alternative to sites like Epinions, summarizing user reviews). Sentiment analysis can help with e-mail filtration by categorizing and prioritizing incoming e-mails (e.g., it can detect strongly negative or flaming e-mails and forward them to a proper folder), as well as citation analysis, where it can determine whether an author is citing a piece of work as supporting evidence or as research that he or she dismisses.

Sentiment Analysis Process

Because of the complexity of the problem (underlying concepts, expressions in text, context in which the text is expressed, etc.), there is not a readily available and universally accepted standardized process to conduct sentiment analysis. Because sentiment analysis can be conducted using one of two learning methods (i.e., supervised or unsupervised), each method may require a different process to follow. While supervised learning relies on manually labeled data set (text samples are labeled as positive or negative by human, and the objective samples are excluded) and machine learning algorithms, unsupervised learning utilizes language dictionary (i.e., a lexicon) as a look-up table to identify the sentiment of each and every term in the textual data. While supervised sentiment analysis is rather straightforward, resembling closely to the ordinary classification-type prediction modeling process, unsupervised sentiment analysis is rather unique and more interesting because of the fact that it does not rely on a time demanding and tedious process of manual labeling of the data. Based on the published work in the field of sensitivity analysis so far (both on research methods and wide range of applications), a multistep logical process, as given in Figure 6.9, seems to be the dominating methodology to follow for unsupervised, lexicon-based sentiment analysis. These logical steps are iterative (i.e., feedback, corrections, and iterations are part of the discovery process) and experimental in nature, and once completed and combined, capable of automatically producing desired insight about the opinions in the text collection.

STEP 1: SENTIMENT DETECTION After the retrieval and preparation of the text documents, the first main task in sensitivity analysis is the detection of objectivity. Here the goal is to differentiate between a fact and an opinion, which may be viewed as classification of text as objective or subjective. This may also be characterized as calculation of O–S Polarity (Objectivity–Subjectivity Polarity, which may be represented with a numerical value ranging from 0 to 1). If the objectivity value is close to 1, then there is no opinion to mine (i.e., it is a fact); therefore, the process goes back and grabs the next text data to analyze. Usually opinion detection is based on the examination of adjectives in text. For example, the polarity of “what a wonderful work” can be determined relatively easily by looking at the adjective.

STEP 2: N–P POLARITY CLASSIFICATION The second main task is that of polarity classification. Given an opinionated piece of text, the goal is to classify the opinion as falling under one of two opposing sentiment polarities, or locate its position on the continuum between these two polarities (Pang & Lee, 2008). When viewed as a binary feature,

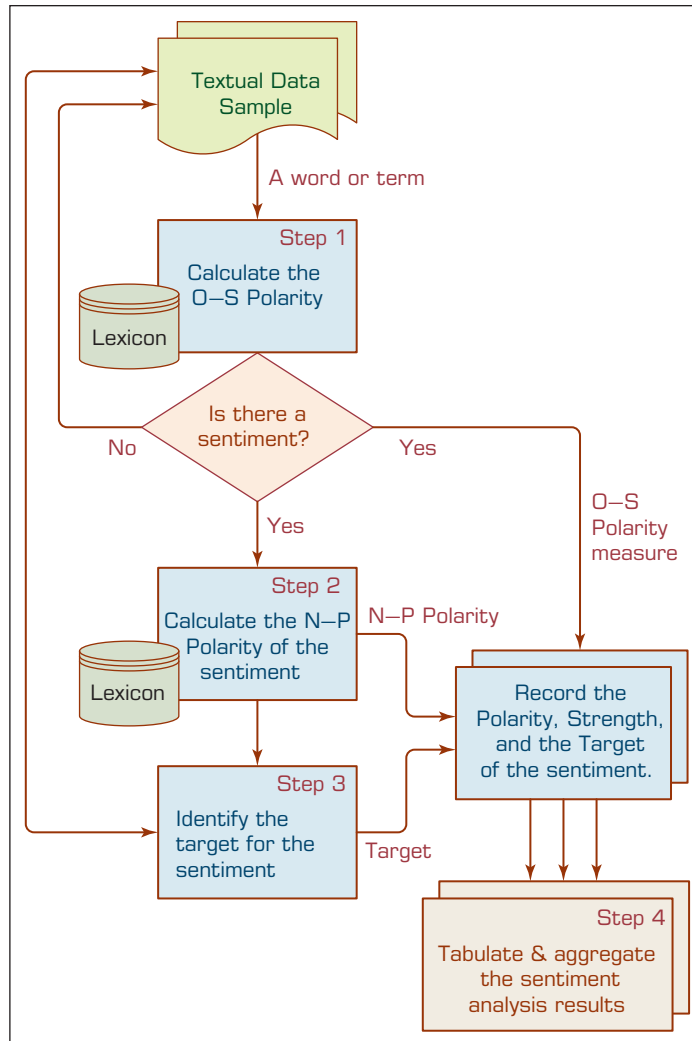


FIGURE 6.9 A Multistep Process to Lexicon-based Sentiment Analysis.

polarity classification is the binary classification task of labeling an opinionated document as expressing either an overall positive or an overall negative opinion (e.g., thumbs up or thumbs down). In addition to the identification of N-P polarity, one should also be interested in identifying the strength of the sentiment (as opposed to just positive, it may be expressed as mildly, moderately, strongly, or very strongly positive). Most of this research was done on product or movie reviews where the definitions of “positive” and “negative” are quite clear. Other tasks, such as classifying news as “good” or “bad,” present some difficulty. For instance, an article may contain negative news without explicitly using any subjective words or terms. Furthermore, these classes usually appear intermixed when a document expresses both positive and negative sentiments. Then the task can be to identify the main (or dominating) sentiment of the document. Still, for lengthy texts, the tasks of classification may need to be done at several levels: term, phrase, sentence, and perhaps document level. For those, it is common to use the outputs of one level as the inputs for the next higher layer. Several methods used to identify the polarity and strengths of the polarity are explained in the next section.

STEP 3: TARGET IDENTIFICATION The goal of this step is to accurately identify the target of the expressed sentiment (e.g., a person, a product, an event). The difficulty of this task depends largely on the domain of the analysis. Even though it is usually easy to accurately identify the target for product or movie reviews because the review is directly connected to the target, it may be quite challenging in other domains. For instance, lengthy, general-purpose text such as Web pages, news articles, and blogs do not always have a predefined topic that they are assigned to, and often mention many objects, any of which may be deduced as the target. Sometimes there is more than one target in a sentiment sentence, which is the case in comparative texts. A subjective comparative sentence orders objects in order of preferences—for example, “This laptop computer is better than my desktop PC.” These sentences can be identified using comparative adjectives and adverbs (more, less, better, longer), superlative adjectives (most, least, best), and other words (such as *same*, *differ*, *win*, *prefer*). Once the sentences have been retrieved, the objects can be put in an order that is most representative of their merits, as described in the text.

STEP 4: COLLECTION AND AGGREGATION Once the sentiments of all text data points in the document are identified and calculated, in this step they are aggregated and converted to a single sentiment measure for the whole document. This aggregation may be as simple as summing up the polarities and strengths of all texts, or as complex as using semantic aggregation techniques from NLP to come up with the ultimate sentiment.

Methods for Polarity Identification

As mentioned in the previous section, **polarity identification** can be made at the word, term, sentence, or document level. The most granular level for polarity identification is at the word level. Once the polarity identification is made at the word level, then it can be aggregated to the next higher level, and then the next until the level of aggregation desired from the sentiment analysis is reached. There seem to be two dominant techniques used for identification of polarity at the word/term level, each having its advantages and disadvantages:

- Using a lexicon as a reference library (developed either manually or automatically, by an individual for a specific task or developed by an institution for general use)
- Using a collection of training documents as the source of knowledge about the polarity of terms within a specific domain (i.e., inducing predictive models from opinionated textual documents)

Using a Lexicon

A lexicon is essentially the catalog of words, their synonyms, and their meanings for a given language. In addition to lexicons for many other languages, there are several general-purpose lexicons created for English. Often general-purpose lexicons are used to create a variety of special-purpose lexicons for use in sentiment analysis projects. Perhaps the most popular general-purpose lexicon is WordNet, created at Princeton University, which has been extended and used by many researchers and practitioners for sentiment analysis purposes. As described on the WordNet Web site (wordnet.princeton.edu), it is a large lexical database of English, including nouns, verbs, adjectives, and adverbs grouped into sets of cognitive synonyms (i.e., synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual–semantic and lexical relations.

An interesting extension of WordNet was created by Esuli and Sebastiani (2006) where they added polarity (Positive–Negative; P–N) and objectivity (Subjective–Objective; S–O) labels for each term in the lexicon. To label each term, they classify the synset

(a group of synonyms) to which this term belongs using a set of ternary classifiers (a measure that attaches to each object exactly one out of three labels), each of them capable of deciding whether a synset is Positive, or Negative, or Objective. The resulting scores range from 0.0 to 1.0, giving a graded evaluation of opinion-related properties of the terms. These can be summed up visually as in Figure 6.10. The edges of the triangle represent one of the three classifications (positive, negative, and objective). A term can be located in this space as a point, representing the extent to which it belongs to each of the classifications.

A similar extension methodology is used to create SentiWordNet, a publicly available lexicon specifically developed for opinion mining (sentiment analysis) purposes. **SentiWordNet** assigns to each synset of WordNet three sentiment scores: positivity, negativity, objectivity. More about SentiWordNet can be found at sentiwordnet.isti.cnr.it.

Another extension to WordNet is WordNet-Affect, developed by Strapparava and Valitutti (2004). They label WordNet synsets using affective labels representing different affective categories like emotion, cognitive state, attitude, and feeling. WordNet has been also directly used in sentiment analysis. For example, Kim and Hovy (2004) and Liu, Hu, and Cheng (2005) generate lexicons of positive and negative terms by starting with a small list of “seed” terms of known polarities (e.g., love, like, nice) and then using the antonymy and synonymy properties of terms to group them into either of the polarity categories.

Using a Collection of Training Documents

It is possible to perform sentiment classification using statistical analysis and machine-learning tools that take advantage of the vast resources of labeled (manually by annotators or using a star/point system) documents available. Product review Web sites like Amazon, C-NET, eBay, RottenTomatoes, and the Internet Movie Database have all been extensively used as sources of annotated data. The star (or tomato, as it were) system provides an explicit label of the overall polarity of the review, and it is often taken as a gold standard in algorithm evaluation. This type of analytics studies fall under the supervised learning-based sentiment analysis while the lexicon-based sentiment analysis (explained above) fall under unsupervised learning/classification type analytics.

A variety of manually labeled textual data are available through evaluation efforts such as the Text REtrieval Conference, NII Test Collection for IR Systems, and Cross Language Evaluation Forum. The data sets these efforts produce often serve as a standard in the text mining community, including for sentiment analysis researchers. Individual

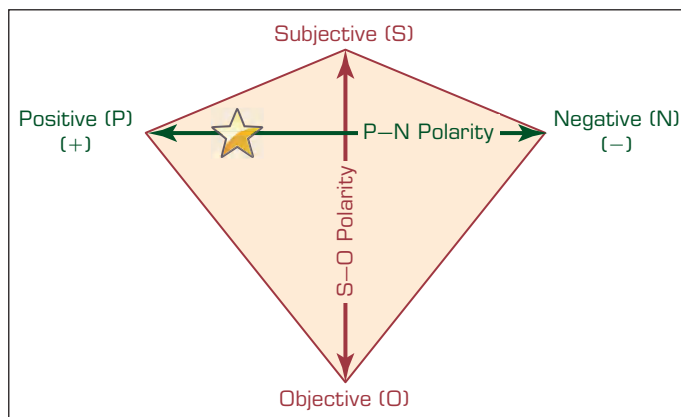


FIGURE 6.10 A Graphical Representation of the P–N Polarity and S–O Polarity Relationship.

TECHNOLOGY INSIGHTS 6.2 Large Textual Data Sets for Predictive Text Mining and Sentiment Analysis

Congressional Floor-Debate Transcripts: Published by Thomas, Pang, and Lee (2006); contains political speeches that are labeled to indicate whether the speaker supported or opposed the legislation discussed.

Econominig: Published by Stern School at New York University; consists of feedback postings for merchants at Amazon.com.

Cornell Movie-Review Data Sets: Introduced by Pang and Lee (2008); contains 1,000 positive and 1,000 negative automatically derived document-level labels and 5,331 positive and 5,331 negative sentences/snippets.

Stanford—Large Movie Review Data Set: A set of 25,000 highly polar movie reviews for training and 25,000 for testing. There is additional unlabeled data for use as well. Raw text and already processed bag-of-words formats are provided. (See <http://ai.stanford.edu/~amaas/data/sentiment>.)

MPQA Corpus: Corpus and Opinion Recognition System corpus; contains 535 manually annotated news articles from a variety of news sources containing labels for opinions and private states (beliefs, emotions, speculations, etc.).

Multiple-Aspect Restaurant Reviews: Introduced by Snyder and Barzilay (2007); contains 4,488 reviews with an explicit 1-to-5 rating for five different aspects: food, ambiance, service, value, and overall experience.

researchers and research groups have also produced many interesting data sets. Technology Insights 6.2 lists some of the most popular ones. Once an already labeled textual data set is obtained, a variety of predictive modeling and other machine-learning algorithms can be used to train sentiment classifiers. Some of the most popular algorithms used for this task include Artificial Neural Networks, Support Vector Machines, k -Nearest Neighbor, Naive Bayes, Decision Trees, and expectation maximization-based Clustering.

Identifying Semantic Orientation of Sentences and Phrases

Once the semantic orientation of individual words has been determined, it is often desirable to extend this to the phrase or sentence the word appears in. The simplest way to accomplish such aggregation is to use some type of averaging for the polarities of words in the phrases or sentences. Though rarely applied, such aggregation can be as complex as using one or more machine-learning techniques to create a predictive relationship between the words (and their polarity values) and phrases or sentences.

Identifying Semantic Orientation of Documents

Even though the vast majority of the work in this area is done in determining semantic orientation of words and phrases/sentences, some tasks like summarization and information retrieval may require semantic labeling of the whole document (Ramage et al., 2009). Similar to the case in aggregating sentiment polarity from word level to phrase or sentence level, aggregation to document level is also accomplished by some type of averaging. Sentiment orientation of the document may not make sense for very large documents; therefore, it is often used on small to medium-sized documents posted on the Internet.

Topic Modeling

Topic modeling, also known as topic detection, is a suite of probabilistic machine learning algorithms that aim to discover and annotate large archives of documents with thematic information. Topic modeling algorithms are statistical methods that analyze the

words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change overtime. The common assumptions behind all topic modeling methods are (1) each document consists of a mix of topics, and (2) each topic consists of a collection of words/terms; the topics are “hidden” or “latent” constructs in between documents and words. The goal of the topic modeling is to discover these latent variables (i.e., topics) that shape the meaning/semantics in the document collection.

There have been numerous attempts to accurately identify inherent topics in a collection of documents. Perhaps the earliest and most rudimentary attempt toward topics detection was to use clustering to group document based on word frequencies of the documents (in the form of a term by document matrix) using one of the traditional clustering techniques—hierarchical or k-means clustering algorithms. This method provided mathematical means to group documents (put them into pre-determined number of clusters) based on how frequently certain words/terms appear in them, and then, one can label these clusters with arbitrary topic names based on the most frequently used terms.

The next popular topic modeling approach was called latent semantic analysis (LSA), or latent semantic indexing (LSI). LSA and LSI are synonymously used terms, while machine leaning/text mining community refers to it as LSA, the information retrieval community refers to it as LSI. The core idea behind LSA is to take the numericized representation of a given corpus (i.e., a term–document matrix) and disperse/factor it into its component matrixes. Specifically, in the process of conducting LSA, the first step is to generate the document-term matrix. Given m documents and n words in the vocabulary, we construct an $m \times n$ matrix A in which each row represents a document and each column represents a word. In the simplest version of LSA, each entry can simply be a raw count of the number of times the j th word appeared in the i th document. In practice, however, raw counts do not work particularly well because they overstate the importance of counts and do not consider the significance of unique words in the document. Instead, the common practice nowadays is to use tf-idf (term frequency $\times$ inverse document frequency) for a better representation of the indices. The simple formula for tf-idf is given below.

$$\underbrace{[tf - idf]_{i,j}}_{\text{Score for document } i \text{ and term } j} = \underbrace{[tf]_{i,j}}_{\substack{\text{Number of} \\ \text{occurrences} \\ \text{of term } j \text{ in} \\ \text{document } i}} \times \log \frac{\underbrace{N}_{\substack{\text{Total number} \\ \text{of documents}}}}{\underbrace{[df]_j}_{\substack{\text{Number of} \\ \text{document} \\ \text{containing} \\ \text{the term } j}}}$$

Once we have the document-term matrix (A) containing the tf-idf generated indices, we can start thinking about identifying the latent topics. Document-term matrix (A) usually is a very sparse, noisy, and redundant matrix across its dimensions. Therefore, to find the latent topics that capture the relationships among the terms and documents, we need to perform some type of dimensionality reduction on A . In LSA, this dimensionality reduction is performed using singular value decomposition (SVD). SVD is a linear algebra technique used to factorize the matrix (A) into the product of three separate matrices: a term-concept matrix (U), a singular value matrix (S), and a concept-document matrix (V) in the form of $A = U \cdot S \cdot V$, where S is a diagonal matrix of the singular values of A . Then, as a final step, documents are grouped into latent topics by taking the cosine of the angle between two vectors, where the resultant values close to 1 representing very similar documents while values close to 0 representing very dissimilar documents. One of the shortcomings of the earlier topic modeling methods is that they assumed the existence of an explicit relationships between the topics and documents (e.g., a document belongs

to only one topic). However, in reality, a document can belong to multiple topics with variant degree and each word/term can belong to each topic with a variant degree. This shortcoming is addressed in the latent Dirichlet allocation, which is the topic of the next sub-section.

Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is arguably the most popular and perhaps most effective topic detection technique of this decade. Although, there are more contemporary attempts to better detect/identify latent topics in textual data, some leveraging deep learning (RNN/LSTM type neural network architectures, which are covered in the Deep Learning chapter later in this book) while other using Word Embedding/Word2Vec type of approaches, vast majority of current practices are still employing on LDA. LDA uses Dirichlet priors/distributions for the document-to-topic and topic-to-term associations/allocation. Dirichlet distribution, the core function of LDA, is a family of continuous multivariate probability distributions parameterized by a vector α of positive reals. It is, in fact, a multivariate generalization of the beta distribution, and hence, it is also called multivariate beta distribution.

LDA is a generative probabilistic model that employs an unsupervised learning process—given a set of training data, it aims to identify the underlying distribution by generating samples from the same distribution. At the highest level, LDA portrays a three-level hierarchical probabilistic allocation model, where each document is modeled as a weighted mixture of the underlying set of topics, and each topic is, in turn, modeled as a weighted mixture of the underlying set of terms. Figure 6.11 illustrates this hierarchical structure of LDA.

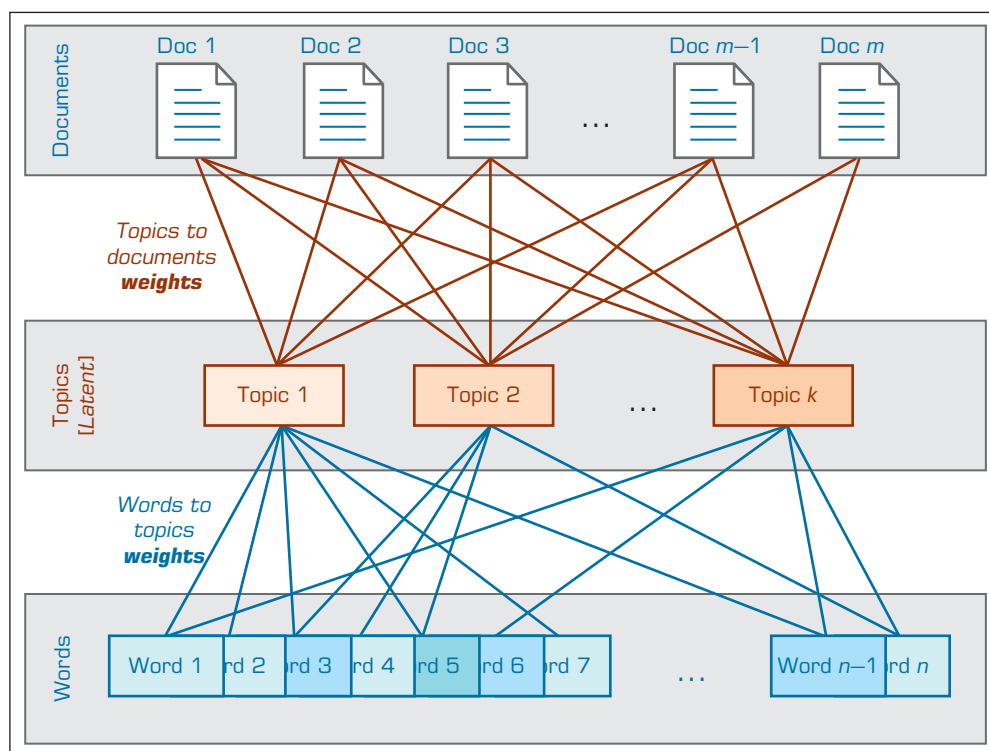


FIGURE 6.11 A Graphical Depiction of the LDA Components—Documents, Words, and Latent Topics.

The algorithmic details of LDA can be found in Blei et al. (2003), considered as the seminal work for LDA algorithm. In order to further explain the internal structure and superior applicability of LDA, the lead author on the seminal work, David M. Blei, a professor at UC Berkeley, presented several illustrative examples in his CACM review article (Blei, 2012). In one of those examples, he fit a 100-topic LDA model to 17,000 articles from the journal *Science*; and the results of this interesting topic modeling exercise along with intuitive graphs and tables can be found in (Blei, 2012).

► SECTION 6.6 REVIEW QUESTIONS

1. What is sentiment analysis? How does it relate to text mining?
2. What are the most popular application areas for sentiment analysis? Why?
3. What would be the expected benefits and beneficiaries of sentiment analysis in politics?
4. What are the main steps in carrying out sentiment analysis projects?
5. What are the two common methods for polarity identification? Explain.
6. What is topic modeling and where can we use it?
7. What differentiated LDA from the previous topic modeling methods? What is its major advantages?
8. Think of an example application area where you could use topic modeling to discover new insights?

6.7 WEB MINING OVERVIEW

The Internet has changed the landscape for conducting business forever. Because of the highly connected, flattened world, and broadened competition field, today's companies are increasingly facing greater opportunities (being able to reach customers and markets that they may have never thought possible) and bigger challenges (a globalized and ever-changing competitive marketplace). Ones with the vision and capabilities to deal with such a volatile environment are greatly benefiting from it, whereas others who resist adapting are having a hard time surviving. Having an engaged presence on the Internet is not a choice anymore; it is a business requirement. Customers are expecting companies to offer their products and/or services over the Internet. They are not only buying products and services but also talking about companies and sharing their transactional and usage experiences with others over the Internet.

The growth of the Internet and its enabling technologies has made data creation, data collection, and data/information/opinion exchange easier. Delays in service, manufacturing, shipping, delivery, and customer inquiries are no longer private incidents and are accepted as necessary evils. Now, thanks to social media tools and technologies on the Internet, everybody knows everything. Successful companies are the ones who embrace these Internet technologies and use them for the betterment of their business processes so that they can better communicate with their customers, understand their needs and wants, and serve them thoroughly and expeditiously. Being customer focused and keeping customers happy has never been as important a concept for businesses as they are now, in this age of Internet and social media.

The World Wide Web (or, for short, Web) serves as an enormous repository of data and information on virtually everything one can conceive; business, personal, you name it, an abundant amount of it is there. The Web is perhaps the world's largest data and text repository, and the amount of information on the Web is growing rapidly. A lot of interesting information can be found online: whose home page is linked to which other pages, how many people have links to a specific Web page, and how a particular site

is organized. In addition, each visitor to a Web site, each search on a search engine, each click on a link, and each transaction on an e-commerce site creates additional data. Although unstructured textual data in the form of Web pages coded in HTML or XML is the dominant content of the Web, the Web infrastructure also contains hyperlink information (connections to other Web pages) and usage information (logs of visitors' interactions with Web sites), all of which provide rich data for knowledge discovery. Analysis of this information can help us make better use of Web sites and also aid us in enhancing relationships and value for the visitors to our own Web sites.

Because of its sheer size and complexity, mining the Web is not an easy undertaking by any means. The Web also poses great challenges for effective and efficient knowledge discovery (Han & Kamber, 2006):

- ***The Web is too big for effective data mining.*** The Web is so large and growing so rapidly that it is difficult to even quantify its size. Because of the sheer size of the Web, it is not feasible to set up a data warehouse to replicate, store, and integrate all of the data on the Web, making data collection and integration a challenge.
- ***The Web is too complex.*** The complexity of a Web page is far greater than that of a page in a traditional text document collection. Web pages lack a unified structure. They contain far more authoring style and content variation than any set of books, articles, or other traditional text-based document.
- ***The Web is too dynamic.*** The Web is a highly dynamic information source. Not only does the Web grow rapidly, but also its content is constantly being updated. Blogs, news stories, stock market results, weather reports, sports scores, prices, company advertisements, and numerous other types of information are updated regularly on the Web.
- ***The Web is not specific to a domain.*** The Web serves a broad diversity of communities and connects billions of workstations. Web users have very different backgrounds, interests, and usage purposes. Most users may not have good knowledge of the structure of the information network and may not be aware of the heavy cost of a particular search that they perform.
- ***The Web has everything.*** Only a small portion of the information on the Web is truly relevant or useful to someone (or some task). It is said that 99% of the information on the Web is useless to 99% of Web users. Although this may not seem obvious, it is true that a particular person is generally interested in only a tiny portion of the Web, whereas the rest of the Web contains information that is uninteresting to the user and may swamp desired results. Finding the portion of the Web that is truly relevant to a person and the task being performed is a prominent issue in Web-related research.

These challenges have prompted many research efforts to enhance the effectiveness and efficiency of discovering and using data assets on the Web. A number of index-based Web search engines constantly search the Web and index Web pages under certain keywords. Using these search engines, an experienced user may be able to locate documents by providing a set of tightly constrained keywords or phrases. However, a simple keyword-based search engine suffers from several deficiencies. First, a topic of any breadth can easily contain hundreds or thousands of documents. This can lead to a large number of document entries returned by the search engine, many of which are marginally relevant to the topic. Second, many documents that are highly relevant to a topic may not contain the exact keywords defining them. As we will cover in more detail later in this chapter, compared to keyword-based Web search, Web mining is a prominent (and more challenging) approach that can be used to substantially enhance the power of Web search engines because Web mining can identify

authoritative Web pages, classify Web documents, and resolve many ambiguities and subtleties raised in keyword-based Web search engines.

Web mining (or Web data mining) is the process of discovering intrinsic relationships (i.e., interesting and useful information) from Web data, which are expressed in the form of textual, linkage, or usage information. The term *Web mining* was first used by Etzioni (1996); today, many conferences, journals, and books focus on Web data mining. It is a continually evolving area of technology and business practice. Web mining is essentially the same as data mining that uses data generated over the Web. The goal is to turn vast repositories of business transactions, customer interactions, and Web site usage data into actionable information (i.e., knowledge) to promote better decision-making throughout the enterprise. Because of the increased popularity of the term *analytics*, nowadays many have started to refer to Web mining as *Web analytics*. However, these two terms are not the same. Whereas Web analytics is primarily Web site usage data focused, Web mining is inclusive of all data generated via the Internet including transaction, social, and usage data. Where Web analytics aims to describe what has happened on the Web site (employing a predefined, metrics-driven descriptive analytics methodology), Web mining aims to discover previously unknown patterns and relationships (employing a novel predictive or prescriptive analytics methodology). From a big-picture perspective, Web analytics can be considered to be a part of Web mining. Figure 6.12 presents a simple taxonomy of Web mining, where it is divided into three main areas: Web content mining, Web structure mining and Web usage mining. In the figure, the data sources used in these three main areas are also specified. Although these three areas are shown separately, as you will see in the following section, they are often used collectively and synergistically to address business problems and opportunities.

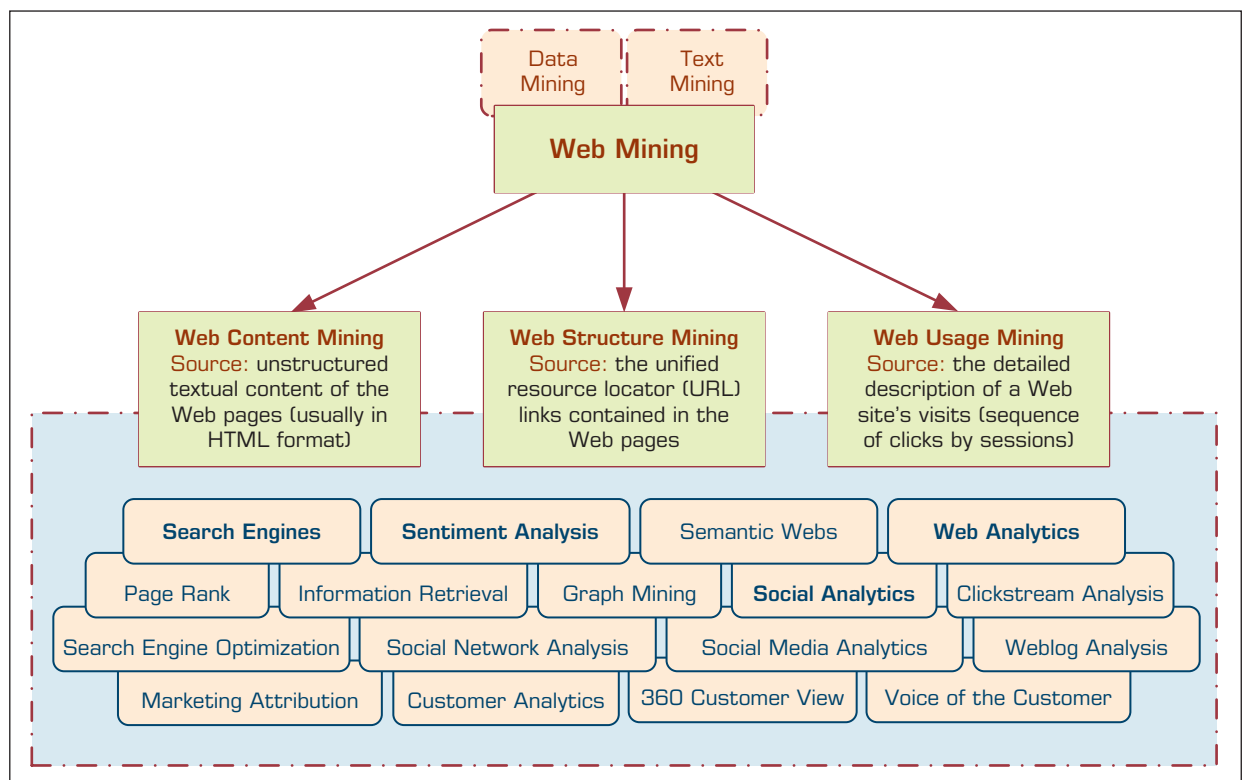


FIGURE 6.12 A Simple Taxonomy of Web Mining.

As Figure 6.12 indicates, Web mining relies heavily on data mining and text mining and their enabling tools and techniques, which we have covered in detail early in this chapter and in the previous chapter (Chapter 4). The figure also indicates that these three generic areas are further extended into several very well-known application areas. Some of these areas were explained in the previous chapters, and some of the others will be covered in detail in this chapter.

Web Content and Web Structure Mining

Web content mining refers to the extraction of useful information from Web pages. The documents may be extracted in some machine-readable format so that automated techniques can extract some information from these Web pages. **Web crawlers** (also called **spiders**) are used to read through the content of a Web site automatically. The information gathered may include document characteristics similar to what is used in text mining, but it may also include additional concepts, such as the document hierarchy. Such an automated (or semiautomated) process of collecting and mining of Web content can be used for competitive intelligence (collecting intelligence about competitors' products, services, and customers). It can also be used for information/news/opinion collection and summarization, sentiment analysis, and automated data collection and structuring for predictive modeling. As an illustrative example to using Web content mining as an automated data collection tool, consider the following. For more than 10 years now, two of the three authors of this book (Drs. Sharda and Delen) have been developing models to predict the financial success of Hollywood movies before their theatrical release. The data that they use for training of the models come from several Web sites, each having a different hierarchical page structure. Collecting a large set of variables on thousands of movies (from the past several years) from these Web sites is a time-demanding, error-prone process. Therefore, they use Web content mining and spiders as an enabling technology to automatically collect, verify, validate (if the specific data item is available on more than one Web site, then the values are validated against each other and anomalies are captured and recorded), and store these values in a relational database. That way, they ensure the quality of the data while saving valuable time (days or weeks) in the process.

In addition to text, Web pages also contain hyperlinks pointing one page to another. Hyperlinks contain a significant amount of hidden human annotation that can potentially help to automatically infer the notion of centrality or *authority*. When a Web page developer includes a link pointing to another Web page, this may be regarded as the developer's endorsement of the other page. The collective endorsement of a given page by different developers on the Web may indicate the importance of the page and may naturally lead to the discovery of authoritative Web pages (Miller, 2005). Therefore, the vast amount of Web linkage information provides a rich collection of information about the relevance, quality, and structure of the Web's contents, and thus is a rich source for Web mining.

Web content mining can also be used to enhance the results produced by search engines. In fact, search is perhaps the most prevailing application of Web content mining and Web structure mining. A search on the Web to obtain information on a specific topic (presented as a collection of keywords or a sentence) usually returns a few relevant, high-quality Web pages and a larger number of unusable Web pages. Use of a relevance index based on keywords and authoritative pages (or some measure of it) improves the search results and ranking of relevant pages. The idea of authority (or **authoritative pages**) stems from earlier information retrieval work using citations among journal articles to evaluate the impact of research papers (Miller, 2005). Though that was the origination of the idea, there are significant differences between the citations in research articles

and hyperlinks on Web pages. First, not every hyperlink represents an endorsement (some links are created for navigation purposes and some are for paid advertisements). Although this is true, if the majority of the hyperlinks are of the endorsement type, then the collective opinion will still prevail. Second, for commercial and competitive interests, one authority will rarely have its Web page point to rival authorities in the same domain. For example, Microsoft may prefer not to include links on its Web pages to Apple's Web sites because this may be regarded as an endorsement of its competitor's authority. Third, authoritative pages are seldom particularly descriptive. For example, the main Web page of Yahoo! may not contain the explicit self-description that it is in fact a Web search engine.

The structure of Web hyperlinks has led to another important category of Web pages called a **hub**. A hub is one or more Web pages that provide a collection of links to authoritative pages. Hub pages may not be prominent, and only a few links may point to them; however, they provide links to a collection of prominent sites on a specific topic of interest. A hub could be a list of recommended links on an individual's home page, recommended reference sites on a course Web page, or a professionally assembled resource list on a specific topic. Hub pages play the role of implicitly conferring the authorities on a narrow field. In essence, a close symbiotic relationship exists between good hubs and authoritative pages; a good hub is good because it points to many good authorities, and a good authority is good because it is being pointed to by many good hubs. Such relationships between hubs and authorities make it possible to automatically retrieve high-quality content from the Web.

The most popular publicly known and referenced algorithm used to calculate hubs and authorities is **hyperlink-induced topic search (HITS)**. It was originally developed by Kleinberg (1999) and has since been improved on by many researchers. HITS is a link-analysis algorithm that rates Web pages using the hyperlink information contained within them. In the context of Web search, the HITS algorithm collects a base document set for a specific query. It then recursively calculates the hub and authority values for each document. To gather the base document set, a root set that matches the query is fetched from a search engine. For each document retrieved, a set of documents that points to the original document and another set of documents that is pointed to by the original document are added to the set as the original document's neighborhood. A recursive process of document identification and link analysis continues until the hub and authority values converge. These values are then used to index and prioritize the document collection generated for a specific query.

Web structure mining is the process of extracting useful information from the links embedded in Web documents. It is used to identify authoritative pages and hubs, which are the cornerstones of the contemporary page-rank algorithms that are central to popular search engines such as Google and Yahoo! Just as links going to a Web page may indicate a site's popularity (or authority), links within the Web page (or the complete Web site) may indicate the depth of coverage of a specific topic. Analysis of links is very important in understanding the interrelationships among large numbers of Web pages, leading to a better understanding of a specific Web community, clan, or clique.

► SECTION 6.7 REVIEW QUESTIONS

1. What are some of the main challenges the Web poses for knowledge discovery?
2. What is Web mining? How does it differ from regular data mining or text mining?
3. What are the three main areas of Web mining?
4. What is Web content mining? How can it be used for competitive advantage?
5. What is Web structure mining? How does it differ from Web content mining?

6.8 SEARCH ENGINES

In this day and age, there is no denying the importance of Internet search engines. As the size and complexity of the World Wide Web increase, finding what you want is becoming a complex and laborious process. People use search engines for a variety of reasons. We use them to learn about a product or service before committing to buy it (including who else is selling it, what the prices are at different locations/sellers, the common issues people are discussing about it, how satisfied previous buyers are, what other products or services might be better, etc.), and to search for places to go, people to meet, things to do. In a sense, search engines have become the centerpiece of most Internet-based transactions and other activities. The incredible success and popularity of Google, the most popular search engine company, is a good testament to this claim. What is somewhat of a mystery to many is how a search engine actually does what it is meant to do. In simplest terms, a **search engine** is a software program that searches for documents (Internet sites or files), based on the keywords (individual words, multiword terms, or a complete sentence) users have provided, that have to do with the subject of their inquiry. Search engines are the workhorses of the Internet, responding to billions of queries in hundreds of different languages every day.

Technically speaking, “search engine” is the popular term for information retrieval systems. Although Web search engines are the most popular, search engines are often used in contexts other than the Web, such as desktop search engines and document search engines. As you will see in this section, many of the concepts and techniques that we have covered in text analytics and text mining early in this chapter also apply here. The overall goal of a search engine is to return one or more documents/pages (if more than one document/page applies, then a ranked-order list is often provided) that best match the user’s query. The two metrics that are often used to evaluate search engines are *effectiveness* (or quality—finding the right documents/pages) and *efficiency* (or speed—returning a response quickly). These two metrics tend to work in reverse directions; improving one tends to worsen the other. Often, based on the user expectation, search engines focus on one at the expense of the other. Better search engines are the ones that excel in both at the same time. Because search engines not only search but in fact find and return documents/pages, perhaps a more appropriate name for them would have been *finding engines*.

Anatomy of a Search Engine

Now let us dissect a search engine and look inside it. At the highest level, a search engine system is composed of two main cycles: a development cycle and a responding cycle (see the structure of a typical Internet search engine in Figure 6.13). While one is interfacing with the World Wide Web, the other is interfacing with the user. One can think of the development cycle as a production process (manufacturing and inventorying documents/pages) and the responding cycle as a retailing process (providing customers/users what they want). In the following section, these two cycles are explained in more detail.

1. Development Cycle

The two main components of the development cycle are the Web crawler and document indexer. The purpose of this cycle is to create a huge database of documents/pages organized and indexed based on their content and information value. The reason for developing such a repository of documents/pages is quite obvious: Due to its sheer size and complexity, searching the Web to find pages in response to a user query is not practical (or feasible within a reasonable time frame); therefore, search engines “cache the Web” into their database and use the cached version of the Web for searching and finding. Once created, this database allows search engines to rapidly and accurately respond to user queries.

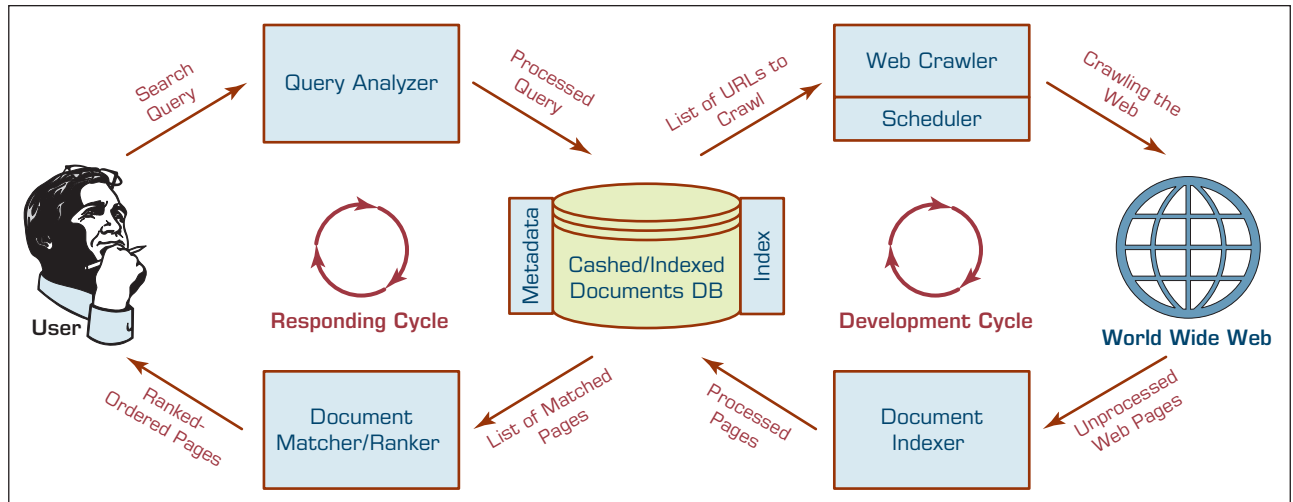


FIGURE 6.13 Structure of a Typical Internet Search Engine.

WEB CRAWLER A Web crawler (also called a spider or a Web spider) is a piece of software that systematically browses (crawls through) the World Wide Web for the purpose of finding and fetching Web pages. Often Web crawlers copy all the pages they visit for later processing by other functions of a search engine.

A Web crawler starts with a list of URLs to visit, which are listed in the scheduler and often are called the *seeds*. These URLs may come from submissions made by Webmasters or, more often, they come from the internal hyperlinks of previously crawled documents/pages. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit (i.e., the scheduler). URLs in the scheduler are recursively visited according to a set of policies determined by the specific search engine. Because there are large volumes of Web pages, the crawler can only download a limited number of them within a given time; therefore, it may need to prioritize its downloads.

DOCUMENT INDEXER As the documents are found and fetched by the crawler, they are stored in a temporary staging area for the document indexer to grab and process. The document indexer is responsible for processing the documents (Web pages or document files) and placing them into the document database. To convert the documents/pages into the desired, easily searchable format, the document indexer performs the following tasks.

STEP 1: PREPROCESSING THE DOCUMENTS Because the documents fetched by the crawler may all be in different formats, for the ease of processing them further, in this step they all are converted to some type of standard representation. For instance, different content types (text, hyperlink, image, etc.) may be separated from each other, formatted (if necessary), and stored in a place for further processing.

STEP 2: PARSING THE DOCUMENTS This step is essentially the application of text mining (i.e., computational linguistic, NLP) tools and techniques to a collection of documents/pages. In this step, first the standardized documents are parsed into components to identify index-worthy words/terms. Then, using a set of rules, the words/terms are indexed. More specifically, using tokenization rules, the words/terms/entities are extracted from the sentences in these documents. Using proper lexicons, the spelling errors and other anomalies in

these words/terms are corrected. Not all the terms are discriminators. The nondiscriminating words/terms (also known as stop words) are eliminated from the list of index-worthy words/terms. Because the same word/term can be in many different forms, stemming is applied to reduce the words/terms to their root forms. Again, using lexicons and other language-specific resources (e.g., WordNet), synonyms and homonyms are identified, and the word/term collection is processed before moving into the indexing phase.

STEP 3: CREATING THE TERM-BY-DOCUMENT MATRIX In this step, the relationships between the words/terms and documents/pages are identified. The weight can be as simple as assigning 1 for presence or 0 for absence of the word/term in the document/page. Usually more sophisticated weight schemas are used. For instance, as opposed to binary, one may choose to assign frequency of occurrence (number of times the same word/term is found in a document) as a weight. As we have seen early in this chapter, text mining research and practice have clearly indicated that the best weighting may come from the use of *term frequency* divided by inverse document frequency (TF/IDF). This algorithm measures the frequency of occurrence of each word/term within a document and then compares that frequency against the frequency of occurrence in the document collection. As we all know, not all high-frequency words/term are good document discriminators, and a good document discriminator in a domain may not be one in another domain. Once the weighing schema is determined, the weights are calculated and the term-by-document index file is created.

2. Response Cycle

The two main components of the responding cycle are the query analyzer and document matcher/ranker.

QUERY ANALYZER The query analyzer is responsible for receiving a search request from the user (via the search engine's Web server interface) and converting it into a standardized data structure, so that it can be easily queried/matched against the entries in the document database. How the query analyzer does what it is supposed to do is quite similar to what the document indexer does (as we have just explained). The query analyzer parses the search string into individual words/terms using a series of tasks that include tokenization, removal of stop words, stemming, and word/term disambiguation (identification of spelling errors, synonyms, and homonyms). The close similarity between the query analyzer and document indexer is not coincidental. In fact, it is quite logical because both are working off the document database; one is putting in documents/pages using a specific index structure, and the other is converting a query string into the same structure so that it can be used to quickly locate most relevant documents/pages.

DOCUMENT MATCHER/RANKER This is where the structured query data are matched against the document database to find the most relevant documents/pages and also rank them in the order of relevance/importance. The proficiency of this step is perhaps the most important component when different search engines are compared to one another. Every search engine has its own (often proprietary) algorithm that it uses to carry out this important step.

The early search engines used a simple keyword match against the document database and returned a list of ordered documents/pages, where the determinant of the order was a function that used the number of words/terms matched between the query and the document along with the weights of those words/terms. The quality and the usefulness of the search results were not all that good. Then, in 1997, the creators of Google came up with a new algorithm, called PageRank. As the name implies, PageRank is an algorithmic

way to rank-order documents/pages based on their relevance and value/importance. Even though PageRank is an innovative way to rank documents/pages, it is an augmentation to the process of retrieving relevant documents from the database and ranking them based on the weights of the words/terms. Google does all these collectively and more to come up with the most relevant list of documents/pages for a given search request. Once an ordered list of documents/pages is created, it is pushed back to the user in an easily digestible format. At this point, users may choose to click on any of the documents in the list, and it may not be the one at the top. If they click on a document/page link that is not at the top of the list, then can we assume that the search engine did not do a good job ranking them? Perhaps, yes. Leading search engines like Google monitor the performance of their search results by capturing, recording, and analyzing postdelivery user actions and experiences. These analyses often lead to more and more rules to further refine the ranking of the documents/pages so that the links at the top are more preferable to the end users.

Search Engine Optimization

Search engine optimization (SEO) is the intentional activity of affecting the visibility of an e-commerce site or a Web site in a search engine's natural (unpaid or organic) search results. In general, the higher ranked on the search results page, and the more frequently a site appears in the search results list, the more visitors it will receive from the search engine's users. As an Internet marketing strategy, SEO considers how search engines work, what people search for, the actual search terms or keywords typed into search engines, and which search engines are preferred by their targeted audience. Optimizing a Web site may involve editing its content, HTML, and associated coding to both increase its relevance to specific keywords and to remove barriers to the indexing activities of search engines. Promoting a site to increase the number of backlinks, or inbound links, is another SEO tactic.

In the early days, in order to be indexed, all Webmasters needed to do was to submit the address of a page, or URL, to the various engines, which would then send a “spider” to “crawl” that page, extract links to other pages from it, and return information found on the page to the server for indexing. The process, as explained before, involves a search engine spider downloading a page and storing it on the search engine's own server, where a second program, known as an indexer, extracts various information about the page, such as the words it contains and where these are located, as well as any weight for specific words, and all links the page contains, which are then placed into a scheduler for crawling at a later date. Nowadays, search engines are no longer relying on Webmasters submitting URLs (even though they still can); instead, they are proactively and continuously crawling the Web and finding, fetching, and indexing everything about it.

Being indexed by search engines like Google, Bing, and Yahoo! is not good enough for businesses. Getting ranked on the most widely used search engines (see Technology Insights 6.3 for a list of most widely used search engines) and getting ranked higher than your competitors are what make the difference. A variety of methods can increase the ranking of a Web page within the search results. Cross-linking between pages of the same Web site to provide more links to the most important pages may improve its visibility. Writing content that includes frequently searched keyword phrases, so as to be relevant to a wide variety of search queries, will tend to increase traffic. Updating content to keep search engines crawling back frequently can give additional weight to a site. Adding relevant keywords to a Web page's metadata, including the title tag and metadescription, will tend to improve the relevancy of a site's search listings, thus increasing traffic. URL normalization of Web pages so that they are accessible via multiple URLs and using canonical link elements, and redirects can help make sure links to different versions of the URL all count toward the page's link popularity score.

TECHNOLOGY INSIGHTS 6.3 Top 15 Most Popular Search Engines

Following are the 15 most popular search engines in 2016 (as derived from eBizMBA Rank).

Rank	Name	Estimated Unique Monthly Visitors
1	Google	1,600,000,000
2	Bing	400,000,000
3	Yahoo! Search	300,000,000
4	Ask	245,000,000
5	AOL Search	125,000,000
6	Wow	100,000,000
7	WebCrawler	65,000,000
8	MyWebSearch	60,000,000
9	Infospace	24,000,000
10	Info	13,500,000
11	DuckDuckGo	11,000,000
12	Contentko	10,500,000
13	Dogpile	7,500,000
14	Alhea	4,000,000
15	ixQuick	1,000,000

A more recent ranking (as of June 2022) of the top-10 search-engines by ReliableSoft (<https://www.reliablesoft.net/top-10-search-engines-in-the-world/>) reveals that, between 2016 and 2022, the top players, led by the undisputed leader, Google, remain the same, while some newcomers make it to the list whereas some lost their spots in the popular search engines list.

- 1. Google
- 2. Microsoft Bing
- 3. Yahoo
- 4. Baidu (only in China)
- 5. Yandex
- 6. DuckDuckGo
- 7. Ask.com
- 8. Ecosia
- 9. Aol.com
- 10. Internet Archive

Although there is no one list that everybody agrees on, the top players seem to be consistently making it to the top of any list created by any organization or technology expert.

Methods for Search Engine Optimization

In general, SEO techniques can be classified into two broad categories: techniques that search engines recommend as part of good site design, and those techniques of which search engines do not approve. The search engines attempt to minimize the effect of the latter, which is often called *spamdexing* (also known as *search spam*, *search engine spam*, or *search engine poisoning*). Industry commentators have classified these methods, and the practitioners who employ them, as either white-hat SEO or black-hat SEO (Goodman, 2005). White hats tend to produce results that last a long time, whereas black hats anticipate that their sites may eventually be banned either temporarily or permanently once the search engines discover what they are doing.

An SEO technique is considered white hat if it conforms to the search engine's guidelines and involves no deception. Because search engine guidelines are not written as a series of rules or commandments, this is an important distinction to note. White-hat SEO is not just about following guidelines, but about ensuring that the content a search engine indexes and subsequently ranks is the same content a user will see. White-hat advice is generally summed up as creating content for users, not for search engines, and then making that content easily accessible to the spiders, rather than attempting to trick the algorithm from its intended purpose. White-hat SEO is in many ways similar to Web development that promotes accessibility, although the two are not identical.

Black-hat SEO attempts to improve rankings in ways that are not approved by the search engines, or involve deception. One black-hat technique uses text that is hidden, either as text colored similar to the background, in an invisible div tag, or positioned off-screen. Another method gives a different page depending on whether the page is being requested by a human visitor or a search engine, a technique known as *cloaking*. Search engines may penalize sites they discover using black-hat methods, either by reducing their rankings or eliminating their listings from their databases altogether. Such penalties can be applied either automatically by the search engines' algorithms or by a manual site review. One example was the February 2006 Google removal of both BMW Germany and Ricoh Germany for use of unapproved practices (Cutts, 2006). Both companies, however, quickly apologized, fixed their practices, and were restored to Google's list.

For some businesses, SEO may generate a significant return on investment. However, one should keep in mind that search engines are not paid for organic search traffic, their algorithms change constantly, and there are no guarantees of continued referrals. Due to this lack of certainty and stability, a business that relies heavily on search engine traffic can suffer major losses if the search engine decides to change its algorithms and stop sending visitors. According to Google's CEO, Eric Schmidt, in 2010, Google made over 500 algorithm changes—almost 1.5 per day. Because of the difficulty in keeping up with changing search engine rules, companies that rely on search traffic practice one or more of the following: (1) hire a company that specializes in SEO (there seem to be an abundant number of those nowadays) to continuously improve your site's appeal to changing practices of the search engines; (2) pay the search engine providers to be listed on the paid sponsors' sections; and (3) consider liberating yourself from dependence on search engine traffic.

Either originating from a search engine (organically or otherwise) or coming from other sites and places, what is most important for an e-commerce site is to maximize the likelihood of customer transactions. Having a lot of visitors without sales is not what a typical e-commerce site is built for.

► SECTION 6.8 REVIEW QUESTIONS

1. What is a search engine? Why are they important for today's businesses?
2. What is a Web crawler? What is it used for? How does it work?
3. What is "search engine optimization?" Who benefits from it?
4. What things can help Web pages rank higher in the search engine results?

6.9 WEB USAGE MINING (WEB ANALYTICS)

Web usage mining (also called **Web analytics**) is the extraction of useful information from data generated through Web page visits and transactions. Analysis of the information collected by Web servers can help us better understand user behavior. Analysis of this data is often called **clickstream analysis**. By using the data and text mining techniques, a company might be able to discern interesting patterns from the clickstreams.

For example, it might learn that 60% of visitors who searched for “hotels in Maui” had searched earlier for “airfares to Maui.” Such information could be useful in determining where to place online advertisements. Clickstream analysis might also be useful for knowing *when* visitors access a site. For example, if a company knew that 70% of software downloads from its Web site occurred between 7 and 11 p.m., it could plan for better customer support and network bandwidth during those hours. Figure 6.14 shows the process of extracting knowledge from clickstream data and how the generated knowledge is used to improve the process, improve the Web site, and, most important, increase the customer value.

Web Analytics Technologies

There are numerous tools and technologies for Web analytics in the marketplace. Because of their power to measure, collect, and analyze Internet data to better understand and optimize Web usage, the popularity of Web analytics tools is increasing. Web analytics holds the promise of revolutionizing how business is done on the Web. Web analytics is not just a tool for measuring Web traffic; it can also be used as a tool for e-business and market research and to assess and improve the effectiveness of e-commerce Web sites. Web analytics applications can also help companies measure the results of traditional print or broadcast advertising campaigns. It can help estimate how traffic to a Web site changes after the launch of a new advertising campaign. Web analytics provides information about the number of visitors to a Web site and the number of page views. It helps gauge traffic and popularity trends, which can be used for market research.

There are two main categories of Web analytics: off-site and on-site. Off-site Web analytics refers to Web measurement and analysis about you and your products that takes place outside your Web site. It includes the measurement of a Web site’s potential audience (prospect or opportunity), share of voice (visibility or word-of-mouth), and buzz (comments or opinions) that is happening on the Internet.

What is more mainstream has been on-site Web analytics. Historically, Web analytics has referred to on-site visitor measurement. However, in recent years this has blurred, mainly because vendors are producing tools that span both categories. On-site Web analytics measure visitors’ behavior once they are on your Web site. This includes its drivers and conversions—for example, the degree to which different landing pages are associated with online purchases. On-site Web analytics measure the performance of your Web site in a commercial context. The data collected on the Web site is then compared against key performance indicators for performance and used to improve a Web site’s or marketing

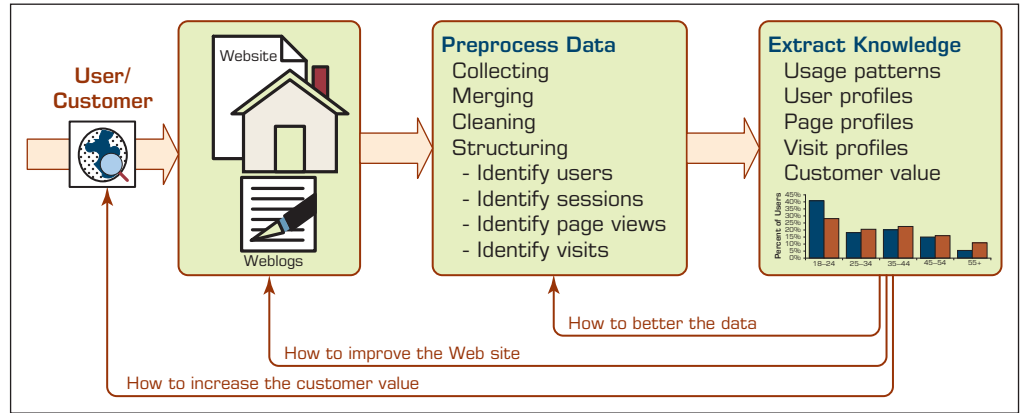


FIGURE 6.14 Extraction of Knowledge from Web Usage Data.

campaign's audience response. Even though Google Analytics is the most widely used on-site Web analytics service, others are provided by Yahoo! and Microsoft, and newer and better tools are emerging constantly that provide additional layers of information.

For on-site Web analytics, there are two technical ways of collecting the data. The first and more traditional method is the server log file analysis, where the Web server records file requests made by browsers. The second method is page tagging, which uses JavaScript embedded in the site page code to make image requests to a third-party analytics-dedicated server whenever a page is rendered by a Web browser (or when a mouse click occurs). Both collect data that can be processed to produce Web traffic reports. In addition to these two main streams, other data sources may also be added to augment Web site behavior data. These other sources may include e-mail, direct mail campaign data, sales and lead history, or social media-originated data.

Web Analytics Metrics

Using a variety of data sources, Web analytics programs provide access to a lot of valuable marketing data, which can be leveraged for better insights to grow your business and better document your return on investment (ROI). The insight and intelligence gained from Web analytics can be used to effectively manage the marketing efforts of an organization and its various products or services. Web analytics programs provide nearly real-time data, which can document your marketing campaign successes or empower you to make timely adjustments to your current marketing strategies.

Whereas Web analytics provides a broad range of metrics, there are four categories of metrics that are generally actionable and can directly impact your business objectives (The Westover Group, 2013). These categories include:

- Web site usability: How were they using my Web site?
- Traffic sources: Where did they come from?
- Visitor profiles: What do my visitors look like?
- Conversion statistics: What does it all mean for the business?

Web Site Usability

Beginning with your Web site, let's take a look at how well it works for your visitors. This is where you can learn how "user friendly" it really is or whether or not you are providing the right content.

- **Page views.** The most basic of measurements, this metric is usually presented as the "average page views per visitor." If people come to your Web site and don't view many pages, then your Web site may have issues with its design or structure. Another explanation for low page views is a disconnect in the marketing messages that brought them to the site and the content that is actually available.
- **Time on site.** Similar to page views, it's a fundamental measurement of a visitor's interaction with your Web site. Generally, the longer a person spends on your Web site, the better it is. That could mean they're carefully reviewing your content, utilizing interactive components you have available, and building toward an informed decision to buy, respond, or take the next step you've provided. On the contrary, the time on site also needs to be examined against the number of pages viewed to make sure the visitors aren't spending their time trying to locate content that should be more readily accessible.
- **Downloads.** This includes PDFs, videos, and other resources you make available to your visitors. Consider how accessible these items are as well as how well they're promoted. If your Web statistics, for example, reveal that 60% of the individuals

who watch a demo video also make a purchase, then you'll want to strategize to increase viewership of that video.

- **Click map.** Most analytics programs can show you the percentage of clicks each item on your Web page received. This includes clickable photos, text links in your copy, downloads, and, of course, any navigation you may have on the page. Are they clicking the most important items?
- **Click paths.** Although an assessment of click paths is more involved, it can quickly reveal where you might be losing visitors in a specific process. A well-designed Web site uses a combination of graphics and information architecture to encourage visitors to follow “predefined” paths through your Web site. These are not rigid pathways but rather intuitive steps that align with the various processes you've built into the Web site. One process might be that of “educating” a visitor who has minimum understanding of your product or service. Another might be a process of “motivating” a returning visitor to consider an upgrade or repurchase. A third process might be structured around items you market online. You'll have as many process pathways in your Web site as you have target audiences, products, and services. Each can be measured through Web analytics to determine how effective it is.

Traffic Sources

Your Web analytics program is an incredible tool for identifying where your Web traffic originates. Basic categories such as search engines, referral Web sites, and visits from bookmarked pages (i.e., direct) are compiled with little involvement by the marketer. With a little effort, however, you can also identify Web traffic that was generated by your various offline or online advertising campaigns.

- **Referral Web sites.** Other Web sites that contain links that send visitors directly to your Web site are considered referral Web sites. Your analytics program will identify each referral site your traffic comes from, and a deeper analysis will help you determine which referrals produce the greatest volume, the highest conversions, the most new visitors, and so on.
- **Search engines.** Data in the search engine category is divided between paid search and organic (or natural) search. You can review the top keywords that generated Web traffic to your site and see if they are representative of your products and services. Depending upon your business, you might want to have hundreds (or thousands) of keywords that draw potential customers. Even the simplest product search can have multiple variations based on how the individual phrases the search query.
- **Direct.** Direct searches are attributed to two sources. An individual who bookmarks one of your Web pages in their favorites and clicks that link will be recorded as a direct search. Another source occurs when someone types your URL directly into their browser. This happens when someone retrieves your URL from a business card, brochure, print ad, radio commercial, and so on. That's why it's a good strategy to use coded URLs.
- **Offline campaigns.** If you utilize advertising options other than Web-based campaigns, your Web analytics program can capture performance data if you include a mechanism for sending them to your Web site. Typically, this is a dedicated URL that you include in your advertisement (i.e., “www.mycompany.com/offer50”) that delivers those visitors to a specific landing page. You now have data on how many responded to that ad by visiting your Web site.
- **Online campaigns.** If you are running a banner ad campaign, search engine advertising campaign, or even e-mail campaign, you can measure individual campaign effectiveness by simply using a dedicated URL similar to the offline campaign strategy.

Visitor Profiles

One of the ways you can leverage your Web analytics into a really powerful marketing tool is through segmentation. By blending data from different analytics reports, you'll begin to see a variety of user profiles emerge.

- **Keywords.** Within your analytics report, you can see what keywords visitors used in search engines to locate your Web site. If you aggregate your keywords by similar attributes, you'll begin to see distinct visitor groups that are using your Web site. For example, the particular search phrase that was used can indicate how well they understand your product or its benefits. If they use words that mirror your own product or service descriptions, then they probably are already aware of your offerings from effective advertisements, brochures, and so on. If the terms are more general in nature, then your visitor is seeking a solution for a problem and has happened upon your Web site. If this second group of searchers is sizable, then you'll want to ensure that your site has a strong education component to convince them they've found their answer and then move them into your sales channel.
- **Content groupings.** Depending on how you group your content, you may be able to analyze sections of your Web site that correspond with specific products, services, campaigns, and other marketing tactics. If you conduct a lot of trade shows and drive traffic to your Web site for specific product literature, then your Web analytics will highlight the activity in that section.
- **Geography.** Analytics permits you to see where your traffic geographically originates, including country, state, and city locations. This can be especially useful if you use geo-targeted campaigns or want to measure your visibility across a region.
- **Time of day.** Web traffic generally has peaks at the beginning of the workday, during lunch, and toward the end of the workday. It's not unusual, however, to find strong Web traffic entering your Web site up until the late evening. You can analyze this data to determine when people browse versus buy and also make decisions on what hours you should offer customer service.
- **Landing page profiles.** If you structure your various advertising campaigns properly, you can drive each of your targeted groups to a different landing page, which your Web analytics will capture and measure. By combining these numbers with the demographics of your campaign media, you can know what percentage of your visitors fit each demographic.

Conversion Statistics

Each organization will define a “conversion” according to its specific marketing objectives. Some Web analytics programs use the term *goal* to benchmark certain Web site objectives, whether that be a certain number of visitors to a page, a completed registration form, or an online purchase.

- **New visitors.** If you're working to increase visibility, you'll want to study the trends in your new visitors data. Analytics identifies all visitors as either new or returning.
- **Returning visitors.** If you're involved in loyalty programs or offer a product that has a long purchase cycle, then your returning visitors data will help you measure progress in this area.
- **Leads.** Once a form is submitted and a thank-you page is generated, you have created a lead. Web analytics will permit you to calculate a completion rate (or abandonment rate) by dividing the number of completed forms by the number of Web visitors that came to your page. A low completion percentage would indicate a page that needs attention.

- **Sales/conversions.** Depending on the intent of your Web site, you can define a “sale” by an online purchase, a completed registration, an online submission, or any number of other Web activities. Monitoring these figures will alert you to any changes (or successes!) that occur further upstream.
- **Abandonment/exit rates.** Just as important as those moving through your Web site are those who began a process and quit or came to your Web site and left after a page or two. In the first case, you’ll want to analyze where the visitor terminated the process and whether there are a number of visitors quitting at the same place. Then, investigate the situation for resolution. In the latter case, a high exit rate on a Web site or a specific page generally indicates an issue with expectations. Visitors click to your Web site based on some message contained in an advertisement, a presentation, and so on, and expect some continuity in that message. Make sure you’re advertising a message that your Web site can reinforce and deliver.

Within each of these items are metrics that can be established for your specific organization. You can create a weekly dashboard that includes specific numbers or percentages that will indicate where you’re succeeding—or highlight a marketing challenge that should be addressed. When these metrics are evaluated consistently and used in conjunction with other available marketing data, they can lead you to a highly quantified marketing program. Figure 6.15 shows a Web analytics dashboard created with freely available Google Analytics tools.

SECTION 6.9 REVIEW QUESTIONS

1. What are the three types of data generated through Web page visits?
2. What is clickstream analysis? What is it used for?
3. What are the main applications of Web mining?
4. What are commonly used Web analytics metrics? What is the importance of metrics?

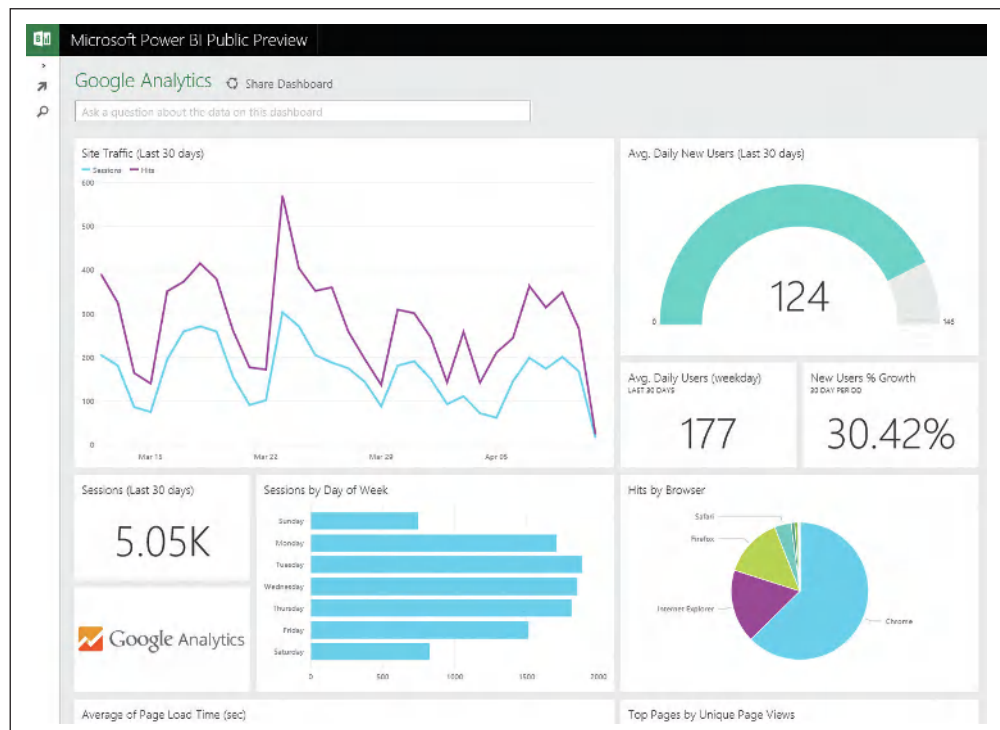


FIGURE 6.15 A Sample Web Analytics Dashboard.

6.10 SOCIAL ANALYTICS

Social analytics may mean different things to different people, based on their worldview and field of study. For instance, the dictionary definition of social analytics refers to a philosophical perspective developed by the Danish historian and philosopher Lars-Henrik Schmidt in the 1980s. The theoretical object of the perspective is *socius*, a kind of “commonness” that is neither a universal account nor a communality shared by every member of a body (Schmidt, 1996). Thus, social analytics differs from traditional philosophy as well as sociology. It might be viewed as a perspective that attempts to articulate the contentions between philosophy and sociology.

Our definition of social analytics is somewhat different; as opposed to focusing on the “social” part (as is done in its philosophical definition), we are more interested in the “analytics” part of the term. Gartner (a very well-known global IT consultancy company) defined social analytics as “monitoring, analyzing, measuring and interpreting digital interactions and relationships of people, topics, ideas and content” (gartner.com/it-glossary/social-analytics/). Social analytics include mining the textual content created in social media (e.g., sentiment analysis, NLP) and analyzing socially established networks (e.g., influencer identification, profiling, prediction) for the purpose of gaining insight about existing and potential customers’ current and future behaviors, and about the likes and dislikes toward a firm’s products and services. Based on this definition and the current practices, social analytics can be classified into two different, but not necessarily mutually exclusive, branches: social network analysis (SNA) and social media analytics.

Social Network Analysis

A **social network** is a social structure composed of individuals/people (or groups of individuals or organizations) linked to one another with some type of connections/relationships. The social network perspective provides a holistic approach to analyzing the structure and dynamics of social entities. The study of these structures uses SNA to identify local and global patterns, locate influential entities, and examine network dynamics. Social networks and the analysis of them is essentially an interdisciplinary field that emerged from social psychology, sociology, statistics, and graph theory. Development and formalization of the mathematical extent of SNA dates back to the 1950s; the development of foundational theories and methods of social networks dates back to the 1980s (Scott & Davis, 2003). SNA is now one of the major paradigms in business analytics, consumer intelligence, and contemporary sociology and is also employed in a number of other social and formal sciences.

A social network is a theoretical construct useful in the social sciences to study relationships between individuals, groups, organizations, or even entire societies (social units). The term is used to describe a social structure determined by such interactions. The ties through which any given social unit connects represent the convergence of the various social contacts of that unit. In general, social networks are self-organizing, emergent, and complex, such that a globally coherent pattern appears from the local interaction of the elements (individuals and groups of individuals) that make up the system.

Following are a few typical social network types that are relevant to business activities.

COMMUNICATION NETWORKS Communication studies are often considered a part of both the social sciences and the humanities, drawing heavily on fields such as sociology, psychology, anthropology, information science, biology, political science, and economics. Many communications concepts describe the transfer of information from one source to another and thus can be represented as a social network. Telecommunication companies are tapping into this rich information source to optimize their business practices and to improve customer relationships.

COMMUNITY NETWORKS Traditionally, community referred to a specific geographic location, and studies of community ties had to do with who talked, associated, traded, and attended social activities with whom. Today, however, there are extended “online” communities developed through social networking tools and telecommunications devices. Such tools and devices continuously generate large amounts of data, which can be used by companies to discover invaluable, actionable information.

CRIMINAL NETWORKS In criminology and urban sociology, much attention has been paid to the social networks among criminal actors. For example, studying gang murders and other illegal activities as a series of exchanges between gangs can lead to better understanding and prevention of such criminal activities. Now that we live in a highly connected world (thanks to the Internet), much of the criminal networks’ formations and their activities are being watched/pursued by security agencies using state-of-the-art Internet tools and tactics. Even though the Internet has changed the landscape for criminal networks and law enforcement agencies, the traditional social and philosophical theories still apply to a large extent.

INNOVATION NETWORKS Business studies on diffusion of ideas and innovations in a network environment focus on the spread and use of ideas among the members of the social network. The idea is to understand why some networks are more innovative, and why some communities are early adopters of ideas and innovations (i.e., examining the impact of social network structure on influencing the spread of an innovation and innovative behavior).

Social Network Analysis Metrics

SNA is the systematic examination of social networks. SNA views social relationships in terms of network theory, consisting of nodes (representing individuals or organizations within the network) and ties/connections (which represent relationships between the individuals or organizations, such as friendship, kinship, or organizational position). These networks are often represented using social network diagrams, where nodes are represented as points and ties are represented as lines.

Over the years, various metrics (or measurements) have been developed to analyze social network structures from different perspectives. These metrics are often grouped into three categories: connections, distributions, and segmentation.

Connections

Homophily: The extent to which actors form ties with similar versus dissimilar others. Similarity can be defined by gender, race, age, occupation, educational achievement, status, values, or any other salient characteristic.

Multiplexity: The number of content forms contained in a tie. For example, two people who are friends and also work together would have a multiplexity of two. Multiplexity has been associated with relationship strength.

Mutuality/reciprocity: The extent to which two actors reciprocate each other’s friendship or other interaction.

Network closure: A measure of the completeness of relational triads. An individual’s assumption of network closure (i.e., that their friends are also friends) is called *transitivity*. Transitivity is an outcome of the individual or situational trait of need for cognitive closure.

Proximity: The tendency for actors to have more ties with geographically close others.

Distributions

Bridge: An individual whose weak ties fill a structural hole, providing the only link between two individuals or clusters. It also includes the shortest route when a longer one is unfeasible due to a high risk of message distortion or delivery failure.

Centrality: Refers to a group of metrics that aim to quantify the importance or influence (in a variety of senses) of a particular node (or group) within a network. Examples of common methods of measuring centrality include betweenness centrality, closeness centrality, eigenvector centrality, alpha centrality, and degree centrality.

Density: The proportion of direct ties in a network relative to the total number possible.

Distance: The minimum number of ties required to connect two particular actors.

Structural holes: The absence of ties between two parts of a network. Finding and exploiting a structural hole can give an entrepreneur a competitive advantage. This concept was developed by sociologist Ronald Burt and is sometimes referred to as an alternate conception of social capital.

Tie strength: Defined by the linear combination of time, emotional intensity, intimacy, and reciprocity (i.e., mutuality). Strong ties are associated with homophily, propinquity, and transitivity, whereas weak ties are associated with bridges.

Segmentation

Cliques and social circles: Groups are identified as *cliques* if every individual is directly tied to every other individual or *social circles* if there is less stringency of direct contact, which is imprecise, or as structurally cohesive blocks if precision is wanted.

Clustering coefficient: A measure of the likelihood that two members of a node are associates. A higher clustering coefficient indicates a greater *cliquishness*.

Cohesion: The degree to which actors are connected directly to each other by cohesive bonds. Structural cohesion refers to the minimum number of members who, if removed from a group, would disconnect the group.

Social Media Analytics

Social media refers to the enabling technologies of social interactions among people in which they create, share, and exchange information, ideas, and opinions in virtual communities and networks. It is a group of Internet-based software applications that build on the ideological and technological foundations of Web 2.0 and that allow the creation and exchange of user-generated content (Kaplan & Haenlein, 2010). Social media depends on mobile and other Web-based technologies to create highly interactive platforms for individuals and communities to share, co-create, discuss, and modify user-generated content. It introduces substantial changes to communication among organizations, communities, and individuals.

Since their emergence in the early 1990s, Web-based social media technologies have seen a significant improvement in both quality and quantity. These technologies take on many different forms, including online magazines, Internet forums, Web logs, social blogs, microblogging, wikis, social networks, podcasts, pictures, video, and product/service evaluations/ratings. By applying a set of theories in the field of media research (social presence, media richness) and social processes (self-presentation, self-disclosure), Kaplan and Haenlein (2010) created a classification scheme with six different types of social media: collaborative projects (e.g., Wikipedia), blogs and microblogs (e.g., Twitter),

content communities (e.g., YouTube), social networking sites (e.g., Facebook), virtual game worlds (e.g., World of Warcraft), and virtual social worlds (e.g. Second Life).

Web-based social media are different from traditional/industrial media, such as newspapers, television, and film, as they are comparatively inexpensive and accessible to enable anyone (even private individuals) to publish or access/consume information. Industrial media generally require significant resources to publish information, as in most cases the articles (or books) go through many revisions before being published (as was the case in the publication of this very book). Here are some of the most prevailing characteristics that help differentiate between social and industrial media (Morgan, Jones, & Hodges, 2010):

Quality: In industrial publishing—mediated by a publisher—the typical range of quality is substantially narrower than in niche, unmediated markets. The main challenge posed by content in social media sites is the fact that the distribution of quality has high variance: from very high-quality items to low-quality, sometimes abusive, content.

Reach: Both industrial and social media technologies provide scale and are capable of reaching a global audience. Industrial media, however, typically use a centralized framework for organization, production, and dissemination, whereas social media are by their very nature more decentralized, less hierarchical, and distinguished by multiple points of production and utility.

Frequency: Compared to industrial media, updating and reposting on social media platforms is easier, faster, and cheaper, and therefore practiced more frequently, resulting in fresher content.

Accessibility: The means of production for industrial media are typically government and/or corporate (privately owned) and are costly, whereas social media tools are generally available to the public at little or no cost.

Usability: Industrial media production typically requires specialized skills and training. Conversely, most social media production requires only modest reinterpretation of existing skills; in theory, anyone with access can operate the means of social media production.

Immediacy: The time lag between communications produced by industrial media can be long (weeks, months, or even years) compared to social media (which can be capable of virtually instantaneous responses).

Updatability: Industrial media, once created, cannot be altered (once a magazine article is printed and distributed, changes cannot be made to that same article), whereas social media can be altered almost instantaneously by comments or editing.

How Do People Use Social Media?

Not only are the numbers on social networking sites growing, but so is the degree to which they are engaged with the channel. Brogan and Bastone (2011) presented research results that stratify users according to how actively they use social media and tracked evolution of these user segments over time. They listed six different engagement levels (Figure 6.16).

According to the research results, the online user community has been steadily migrating upward on this engagement hierarchy. The most notable change is among Inactives. Forty-four percent of the online population fell into this category. Two years later, more than half of those Inactives had jumped into social media in some form or another. “Now roughly 82% of the adult population online is in one of the upper categories,” said Bastone. “Social media has truly reached a state of mass adoption.”

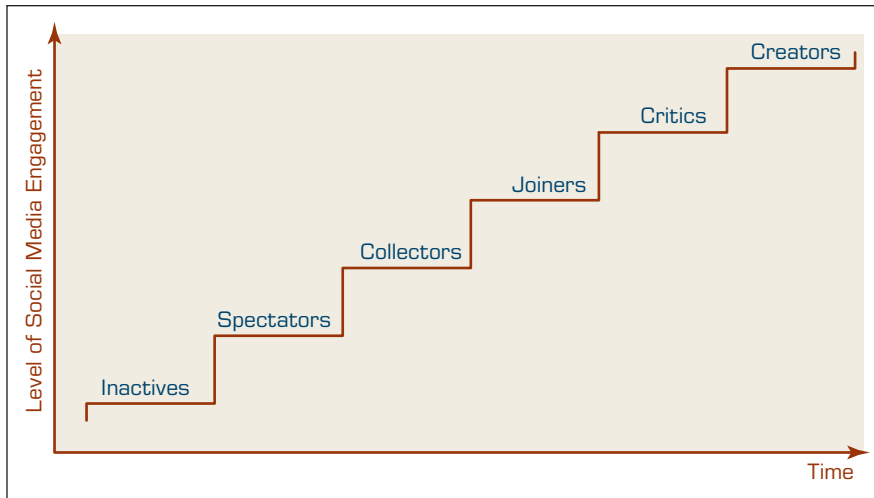


FIGURE 6.16 Evolution of Social Media User Engagement.

Social media analytics refers to the systematic and scientific ways to consume the vast amount of content created by Web-based social media outlets, tools, and techniques for the betterment of an organization's competitiveness. Social media analytics is rapidly becoming a new force in organizations around the world, allowing them to reach out to and understand consumers as never before. In many companies, it is becoming the tool for integrated marketing and communications strategies. Analytics in Action 6.5 summarizes how reporting of social media campaigns outcomes can be automated using KNIME analytics solutions.

ANALYTICS IN ACTION 6.5 Increasing the Efficiency of Social Media Campaigns

Motivation

Sport England is an arms-length body of government responsible for growing and developing grassroots sport and getting more people active across England. They're working to keep movement, sport, and physical activity central to the lives of everyone.

Sport England run cross-platform media campaigns to encourage participation in sport and physical activity. Reporting, however, is always a manual task. Sport England wanted to increase the efficiency of this process to not only generate insights more quickly, but also reduce unnecessary manual work.

Previously, the process required logging in to each of the social media platforms, collecting statistics from accounts, transforming the data so it was comparable, before visualizing it in Excel and producing a report in PowerPoint. This was a labor-intensive process, which could only be done by one person and created a significant lag in reporting. The delay in reporting meant that while a view of overall success was available to senior management, information was not available for operational decisions where a nimble response to events might be required. Off the shelf solutions were considered but could not provide the targeted approach or custom visuals required by Sport England for their "This Girl Can" campaign.

Implementation – *Blending Technologies for an Automated, Bespoke, and Branded Reporting Solution*

Atos (an international IT services company and a trusted KNIME partner) created workflows in KNIME Analytics Platform to collect, transform, join, and write data from social media platforms to an SQL database on Azure, where it was visualized in Power BI. KNIME Server was used to schedule data collection and an analytical application was made available via the KNIME WebPortal to allow a supplementary data file to be uploaded in a user-friendly way by their admin team.

Atos used functionality already built into KNIME to connect to Twitter, Python nodes to connect to, navigate, and retrieve data from Facebook and Instagram graph databases, as well as web-scraping for data from Medium. Python nodes were also used to query Google and YouTube APIs. Data were transformed and joined in KNIME to provide a comparable snapshot of activity over the previous 24 hours when the collected data ranged from cumulative counts to six-week windows.

The workflows were scheduled to run shortly after midnight on KNIME Server, which was running in a virtual machine on Sport England's Azure platform. Data were saved to an SQL database where it was available to Power BI (using the KNIME PowerBI Integration) for visualization.

Initially Azure Cognitive Services were to perform sentiment analysis on tweets in Power BI, but this functionality was folded back into KNIME. Therefore, a tweet was only ever passed to Cognitive Services once, and the score was stored in the database for cost and efficiency savings.

Atos created custom metrics for rating features such as a user's influence, which could be explained, understood, and adapted as necessary. These were then visualized to Sport England's specification in adherence with their internal style guides and branding to make it their tool. This highlighted that the data that was important to them and the success of their campaign.

Results – Automatic Data Processing and Increased Reporting Frequency

The frequency of reporting was increased from monthly to daily. Instead of relying on a single person to generate the report, which would take half a day, the latest data were collected and processed automatically and made available via a Power BI report at the beginning of the day. The Power BI report was more useful than the PowerPoint report because it allowed users to apply multiple filters to find the data they needed—instead of relying on the analyst to anticipate requirements and generate insight.

Why KNIME?

KNIME Analytics Platform makes it easy to seamlessly blend core functionality for data manipulation with highly complex and bespoke operations for data collection and cleaning using R and Python. The visual workflow builder enables workflow creators to explain the process to the media team and others outside data specialist roles who don't necessarily have the technical data science knowledge, drawing parallels to how data were treated in the manual process. This means the solution is less of a "black box," but rather something that those who use the data can have confidence in because they understand how the parts fit together.

KNIME Server makes scheduling data collection a simple task and provides feedback that the workflows are executed without errors. Additional benefits include the ability to work remotely in the cloud, and the ability for Sport England to easily manage the data governance to sandbox this project from their other data.

Sources: KNIME (2022). "Combining the Power of KNIME and PowerBI for Automated Sentiment Analysis" can be found at <https://www.knime.com/solutions/success-story/automated-sentiment-analysis> (accessed May 2022).

The exponential growth of social media outlets, from blogs, Facebook, and Twitter to LinkedIn and YouTube, and analytics tools that tap into these rich data sources offer organizations the chance to join a conversation with millions of customers around the globe every day. This aptitude is why nearly two-thirds of the 2,100 companies who participated in a recent survey by Harvard Business Review Analytic Services said they are either currently using social media channels or have social media plans in the works (Harvard Business Review, 2010). But many still say social media is an experiment, as they try to understand how to best use the different channels, gauge their effectiveness, and integrate social media into their strategy.

Measuring the Social Media Impact

For organizations, small or large, there is valuable insight hidden in all the user-generated content on social media sites. But how do you dig it out of dozens of review sites, thousands of blogs, millions of Facebook posts, and billions of tweets? Once you do that, how do you measure the impact of your efforts? These questions can be addressed by the analytics extension of the social media technologies. Once you decide on your goal for social media (what it is that you want to accomplish), there is a multitude of tools to help you get there. These analysis tools usually fall into three broad categories:

- **Descriptive analytics:** Uses simple statistics to identify activity characteristics and trends, such as how many followers you have, how many reviews were generated on Facebook, and which channels are being used most often.
- **Social network analysis:** Follows the links between friends, fans, and followers to identify connections of influence as well as the biggest sources of influence.
- **Advanced analytics:** Includes predictive analytics and text analytics that examine the *content* in online conversations to identify themes, sentiments, and connections that would not be revealed by casual surveillance.

Sophisticated tools and solutions to social media analytics use all three categories of analytics (i.e., descriptive, predictive, and prescriptive) in a somewhat progressive fashion.

Best Practices in Social Media Analytics

As an emerging tool, social media analytics is practiced by companies in a somewhat haphazard fashion. Because there are not well-established methodologies, everybody is trying to create their own by trial and error. What follows are some of the field-tested best practices for social media analytics proposed by Paine and Chaves (2012).

THINK OF MEASUREMENT AS A GUIDANCE SYSTEM, NOT A RATING SYSTEM Measurements are often used for punishment or rewards; they should not be. They should be about figuring out what the most effective tools and practices are, what needs to be discontinued because it doesn't work, and what needs to be done more because it does work very well. A good analytics system should tell you where you need to focus. Maybe all that emphasis on Facebook doesn't really matter because that is not where your audience is. Maybe they are all on Twitter, or vice versa. According to Paine and Chaves, channel preference won't necessarily be intuitive, "We just worked with a hotel that had virtually no activity on Twitter for one brand but lots of Twitter activity for one of their higher brands." Without an accurate measurement tool, you would not know.

TRACK THE ELUSIVE SENTIMENT Customers want to take what they are hearing and learning from online conversations and act on it. The key is to be precise in extracting and tagging their intentions by measuring their sentiments. As we have seen earlier in this chapter, text analytic tools can categorize online content, uncover linked concepts, and reveal the sentiment in a conversation as "positive," "negative," or "neutral," based on the words people use. Ideally, you would like to be able to attribute sentiment to a specific product, service, and business unit. The more precise you can get in understanding the tone and perception that people express, the more actionable the information becomes, because you are mitigating concerns about mixed polarity. A mixed-polarity phrase, such as "hotel in great location but bathroom was smelly," should not be tagged as "neutral" because you have positives and negatives offsetting each other. To be actionable, these types of phrases are to be treated separately; "bathroom was smelly" is something

someone can own and improve on. One can classify and categorize these sentiments, look at trends over time, and see significant differences in the way people speak either positively or negatively about you. Furthermore, you can compare sentiment about your brand to your competitors.

CONTINUOUSLY IMPROVE THE ACCURACY OF TEXT ANALYSIS An industry-specific text analytics package will already know the vocabulary of your business. The system will have linguistic rules built into it, but it learns over time and gets better and better. Much as you would tune a statistical model as you get more data, better parameters, or new techniques to deliver better results, you would do the same thing with the NLP that goes into sentiment analysis. You set up rules, taxonomies, categorization, and meaning of words; watch what the results look like; and then go back and do it again.

LOOK AT THE RIPPLE EFFECT It is one thing to get a great hit on a high-profile site, but that's only the start. There's a difference between a great hit that just sits there and goes away versus a great hit that is tweeted, retweeted, and picked up by influential bloggers. Analysis should show you which social media activities go "viral" and which quickly go dormant—and why.

LOOK BEYOND THE BRAND One of the biggest mistakes people make is to be concerned only about their brand. To successfully analyze and act on social media, you need to understand not just what is being said about your brand, but the broader conversation about the spectrum of issues surrounding your product or service, as well. Customers don't usually care about a firm's message or its brand; they care about themselves. Therefore, you should pay attention to what they are talking about, where they are talking, and where their interests are.

IDENTIFY YOUR MOST POWERFUL INFLUENCERS Organizations struggle to identify who has the most power in shaping public opinion. It turns out, your most important influencers are not necessarily the ones who advocate specifically for your brand; they are the ones who influence the whole realm of conversation about your topic. You need to understand whether they are saying nice things, expressing support, or simply making observations or critiquing. What is the nature of their conversations? How is my brand being positioned relative to the competition in that space?

LOOK CLOSELY AT THE ACCURACY OF YOUR ANALYTIC TOOL Until recently, computer-based automated tools were not as accurate as humans for sifting through online content. Even now, accuracy varies depending on the media. For product review sites, hotel review sites, and Twitter, it can reach anywhere between 80 and 90% accuracy because the context is more boxed in. When you start looking at blogs and discussion forums, where the conversation is more wide-ranging, the software can deliver 60 to 70% accuracy (Paine & Chaves, 2012). These figures will increase over time because the analytics tools are continually upgraded with new rules and improved algorithms to reflect field experience, new products, changing market conditions, and emerging patterns of speech.

INCORPORATE SOCIAL MEDIA INTELLIGENCE INTO PLANNING Once you have a big-picture perspective and detailed insight, you can begin to incorporate this information into your planning cycle. But that is easier said than done. A quick audience poll revealed that very few people currently incorporate learning from online conversations into their planning cycles (Paine & Chaves, 2012). One way to achieve this is to find time-linked associations between social media metrics and other business activities or market events. Social media is typically either organically invoked or invoked by something your organization does; therefore, if you see a spike in activity at some point in time, you want to know what was behind that.

SECTION 6.10 REVIEW QUESTIONS

1. What is meant by social analytics? Why is it an important business topic?
2. What is a social network? What is the need for SNA?
3. What is social media? How does it relate to Web 2.0?
4. What is social media analytics? What are the reasons behind its increasing popularity?
5. How can you measure the impact of social media analytics?

Chapter Highlights

- Text mining is the discovery of knowledge from unstructured (mostly text-based) data sources. Given that a great deal of information is in text form, text mining is one of the fastest-growing branches of the business intelligence field.
- Text mining applications are in virtually every area of business and government, including marketing, finance, healthcare, medicine, and homeland security.
- Text mining uses NLP to induce structure into the text collection and then uses data mining algorithms such as classification, clustering, association, and sequence discovery to extract knowledge from it.
- Sentiment can be defined as a settled opinion reflective of one's feelings.
- Sentiment analysis deals with differentiating between two classes, positive and negative.
- As a field of research, sentiment analysis is closely related to computational linguistics, NLP, and text mining.
- Sentiment analysis is trying to answer the question, "What do people feel about a certain topic?" by digging into opinions of many using a variety of automated tools.
- The VOC is an integral part of an analytic CRM and customer experience management systems and is often powered by sentiment analysis.
- The VOM is about understanding aggregate opinions and trends at the market level.
- Polarity identification in sentiment analysis is accomplished either by using a lexicon as a reference library or by using a collection of training documents.
- WordNet is a popular general-purpose lexicon created at Princeton University.
- SentiWordNet is an extension of WordNet to be used for sentiment identification.
- Speech analytics is a growing field of science that allows users to analyze and extract information from both live and recorded conversations.
- Web mining can be defined as the discovery and analysis of interesting and useful information from the Web, about the Web, and usually using Web-based tools.
- Web mining can be viewed as consisting of three areas: Web content mining, Web structure mining, and Web usage mining.
- Web content mining refers to the automatic extraction of useful information from Web pages. It may be used to enhance search results produced by search engines.
- Web structure mining refers to generating interesting information from the links included in Web pages.
- Web structure mining can also be used to identify the members of a specific community and perhaps even the roles of the members in the community.
- Web usage mining refers to developing useful information through analysis of Web server logs, user profiles, and transaction information.
- Text and Web mining are emerging as critical components of the next generation of business intelligence tools to enable organizations to compete successfully.
- A search engine is a software program that searches for documents (Internet sites or files), based on the keywords (individual words, multiword terms, or a complete sentence) users have provided, that have to do with the subject of their inquiry.
- SEO is the intentional activity of affecting the visibility of an e-commerce site or a Web site in a search engine's natural (unpaid or organic) search results.
- VOC is a term usually used to describe the analytic process of capturing a customer's expectations, preferences, and aversions.
- Social analytics is the monitoring, analyzing, measuring, and interpreting of digital interactions and relationships of people, topics, ideas, and content.
- A social network is a social structure composed of individuals/people (or groups of individuals or organizations) linked to one another with some type of connections/relationships.
- Social media analytics refers to the systematic and scientific ways to consume the vast amount of content created by Web-based social media outlets, tools, and techniques for the betterment of an organization's competitiveness.

Key Terms

association	natural language	social media analytics	voice of the customer
authoritative pages	processing (NLP)	social network	(VOC)
classification	part-of-speech (POS)	spider	Web analytics
clickstream analysis	tagging	stemming	Web content mining
clustering	polarity identification	stop words	Web crawler
corpus	polyseme	term–document matrix	Web mining
deception detection	search engine	(TDM)	Web structure mining
hubs	sentiment analysis	text mining	Web usage mining
hyperlink-induced topic	SentiWordNet	tokenizing	WordNet
search (HITS)	singular value	topic modeling	
Latent Dirichlet Allocation	decomposition	trend analysis	
(LDA)	(SVD)	unstructured data	

Questions for Discussion

1. Explain the relationship among data mining, text mining, and sentiment analysis.
2. In your own words, define text mining, and discuss its most popular applications.
3. What does it mean to induce structure into the text-based data? Discuss the alternative ways of inducing structure into text-based data.
4. What is the role of NLP in text mining? Discuss the capabilities and limitations of NLP in the context of text mining.
5. List and discuss three prominent application areas for text mining. What is the common theme among the three application areas you chose?
6. What is sentiment analysis? How does it relate to text mining?
7. What are the common challenges that sentiment analysis has to deal with?
8. What are the most popular application areas for sentiment analysis? Why?
9. What are the main steps in carrying out sentiment analysis projects?
10. What are the two common methods for polarity identification? Explain.
11. What Is topic modeling and where can it be used in business analytics?
12. What are the popular techniques used in topic modeling? Why LDA is the most popular topic modeling technique?
13. Discuss the differences and commonalities between text mining and Web mining.
14. In your own words, define Web mining, and discuss its importance.
15. What are the three main areas of Web mining? Discuss the differences and commonalities among these three areas.
16. What is a search engine? Why are they important for businesses?
17. What is SEO? Who benefits from it? How?
18. What is Web analytics? What are the metrics used in Web analytics?
19. Define *social analytics*, *social network*, and *social network analysis*. What are the relationships among them?
20. What is social media analytics? How is it done? Who does it? What comes out of it?

Exercises

Hands-On Exercises

1. Conduct a Web search to identify case studies about text mining. Based on your findings, describe the recent developments in the field in a report.
2. Search the Web to locate white papers, Web-based seminars (i.e., webinars), and other learning materials related to text mining. Synthesize your findings into a short report.
3. Within the book's supplemental material, locate the case study titled "eBay Analytics." Read the case carefully, extend your understanding of the case by searching the Internet for additional information, and answer the case questions.
4. Within the book's supplemental material, locate the case study related to sentiment analysis named "How Do We Fix an App Like That?" Read the description and follow the directions to download the data and the tool to carry out the exercise.
5. Search the Web to find cases about Web mining. Describe recent developments in the field. Write a short report to summarize your findings.

6. Browse the Web and your library's digital databases to identify articles that make the linkage between text/Web mining and contemporary business intelligence systems.

Team Assignments and Role-Playing Projects

1. Examine how textual data can be captured automatically using Web-based technologies. Once captured, what are the potential patterns that you can extract from these unstructured data sources?
2. Interview administrators in your college or executives in your organization to determine how text mining and Web mining could assist them in their work. Write a proposal describing your findings. Include a preliminary cost-benefit analysis in your report.
3. Go to your library's online resources. Learn how to download attributes of a collection of literature (journal articles) in a specific topic. Download and process the data using a methodology similar to the one explained in Analytics in Action 6.3.
4. Find a readily available sentiment text data set (see Technology Insights 6.2 for a list of popular data sets) and download it into your computer. If you have an analytics tool that is capable of text mining, use that. If not, download KNIME Analytics Platform (<http://knime.com>) and install it. Also install the Text Processing extension for KNIME Analytics Platform (see Chapter 9 for more details on KNIME and its Extension installations). Process the downloaded data using your text mining tool (i.e., convert the data into a structured form). Build models and assess the sentiment detection accuracy of several classification models (e.g., support vector machines, decision trees, neural networks, logistic regression). Write a detailed report in which you explain your findings and your experiences.
5. Go to <https://ai.stanford.edu/~amaas/data/sentiment/> and download the movie reviews dataset. The data are already manually labelled for positive and negative sentiments. The complete data set includes 50,000 reviews. Use half of the data for training and the remaining second half for testing. Using a text mining tool of your choice (whichever you have access to), develop binary classification models using five different machine learning methods. Create a written report organized around the six CRISP-DM steps, to report on your findings and experiences.
6. Examine how Web-based data can be captured automatically using the latest technologies. Once captured, what

are the potential patterns that you can extract from these content-rich, mostly unstructured data sources?

Internet Exercises

1. Find recent cases of successful text mining and Web mining applications. Try text and Web mining software vendors and consultancy firms and look for cases or success stories. Prepare a report summarizing five new case studies.
2. Go to statsoft.com. Select Downloads, and download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?
3. Go to sas.com. Download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?
4. Go to ibm.com. Download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?
5. Go to teradata.com. Download at least three white papers on applications. Which of these applications may have used the data/text/Web mining techniques discussed in this chapter?
6. Go to clarabridge.com. Download at least three white papers on applications. Which of these applications may have used text mining in a creative way?
7. Go to knime.com/hub and search for text mining-related workflows developed and shared by other data scientists. Identify one that you like, download it and open it in KNIME Analytics Platform. First, investigate how the text mining process is created. Then, provide your own ideas as to what other alternative text mining functions can be used (using term frequency as opposed to binary value to represent indices; using lemmatization as supposed to stemming; etc.) for at least five separate steps within the text mining workflow.
8. Go to kdnuggets.com. Explore the sections on applications as well as software. Find names of at least three additional packages for data mining and text mining.
9. Survey some Web mining tools and vendors. Identify some Web mining products and service providers that are not mentioned in this chapter.
10. Go to attensity.com. Download at least three white papers on Web analytics applications. Which of these applications may have used a combination of data/text/Web mining techniques?

References

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993–1022.
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.
- Bond C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Reports*, 10(3), 214–234.
- Brogan, C., & Bastone, J. (2011). Acting on customer intelligence from social media: The new edge for building customer loyalty and your brand. SAS white paper.
- Chun, H. W., Tsuruoka, Y., Kim, J. D., Shiba, R., Nagata, N., & Hishiki, T. (2006). Extraction of gene-disease relations from MEDLINE using domain dictionaries and machine learning. *Proceedings of the 11th Pacific Symposium on Biocomputing*, 4–15.

- Coussement, K., & Van Den Poel, D. (2008). Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decision Support Systems*, 44(4), 870–882.
- Coussement, K., & Van Den Poel, D. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications*, 36(3), 6127–6134.
- Cutts, M. (2006, February 4). Ramping Up on International Webspam. mattcutts.com/blog/mattcutts.com/blog/ramping-up-on-international-webspam (accessed March 2013).
- Delen, D., & Crossland, M. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*, 34(3), 1707–1720.
- Esuli, A., & Sebastiani, F. (2006, May). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, 6, 417–422.
- Etzioni, O. (1996). The World Wide Web: Quagmire or gold mine? *Communications of the ACM*, 39(11), 65–68.
- EUROPOL. (2007). EUROPOL Work Program 2005. statewatch.org/news/2006/apr/europol-work-programme-2005.pdf (accessed October 2008).
- Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Boston: ABS Ventures.
- Fuller, C. M., Biros, D. P., & Delen, D. (2011). An investigation of data and text mining methods for real world deception detection. *Expert Systems with Applications*, 38(7), 8392–8398.
- Fuller, C. M., Biros, D., & Delen, D. (2008). Exploration of feature selection and advanced classification models for high-stakes deception detection. *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS)*, Big Island, HI: IEEE Press, 80–99.
- Ghani, R., Probst, K., Liu, Y., Krema, M., & Fano, A. (2006). Text mining for product attribute extraction. *SIGKDD Explorations*, 8(1), 41–48.
- Goodman, A. (2005). Search engine showdown: Black hats versus white hats at SES. SearchEngineWatch.com/article/2066090/Search-Engine-Showdown-Black-Hats-vs.-White-Hats-at-SES (accessed February 2013).
- Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., ... Welty, C. (2010). Building Watson: An overview of the DeepQA Project. *AI Magazine*, 31(3); The DeepQA Research Team (2021). researcher.watson.ibm.com/researcher/view_group_subpage.php?id=2159 (accessed January 2022); Feldman, S., Hanover, J., Burghard, C., & Schubmehl, D. (2012). Recommend with confidence - Unlocking the power of unstructured data. IBM Watson (2022). www.ibm.com/watson/advantages/recommend?mbsrc=ibmsearch_a&mbq=Unlocking%20Unstructured%20Data (accessed June 2022).
- Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques*, 2nd ed. San Francisco: Morgan Kaufmann.
- Harvard Business Review*. (2010). The new conversation: Taking social media from talk to action. A SAS-Sponsored Research Report by Harvard Business Review Analytic Services. sas.com/resources/whitepaper/wp_23348.pdf (accessed March 2013).
- Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53(1), 59–68.
- Kim, S. M., & Hovy, E. (2004, August). Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics* (p. 1367). Association for Computational Linguistics.
- Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5), 604–632.
- Lin, J., & Demner-Fushman, D. (2005). “Bag of words” is not enough for strength of evidence classification. *AMIA Annual Symposium Proceedings*, 1031–1032. pubmedcentral.nih.gov/articlerender.fcgi?artid=1560897.
- Liu, B., Hu, M., & Cheng, J. (2005, May). Opinion observer: Analyzing and comparing opinions on the Web. In *Proceedings of the 14th International Conference on World Wide Web* (pp. 342–351). ACM.
- Mahgoub, H., Rösner, D., Ismail, N., & Torkey, F. (2008). A text mining technique using association rules extraction. *International Journal of Computational Intelligence*, 4(1), 21–28.
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
- McKnight, W. (2005, January 1). Text data mining in business intelligence. *Information Management Magazine*. information-management.com/issues/20050101/1016487-1.html (accessed May 22, 2009).
- Mejova, Y. (2009). Sentiment analysis: An overview. Comprehensive exam paper. <http://www.cs.uiowa.edu/~ymejova/publications/CompsYelenaMejova.pdf> (accessed February 2013).
- Miller, T. W. (2005). *Data and text mining: A business applications approach*. Upper Saddle River, NJ: Prentice Hall.
- Miner, G., Elder IV, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press (Elsevier Publishing, Amsterdam, Netherlands).
- Morgan, N., Jones, G., & Hodges, A. (2010). The complete guide to social media from the social media guys. thesocialmediaguys.co.uk/wp-content/uploads/downloads/2011/03/CompleteGuidetoSocialMedia.pdf (accessed February 2013).
- Nakov, P., Schwartz, A., Wolf, B., & Hearst, M. A. (2005). Supporting annotation layers for natural language processing. *Proceedings of the ACL*, Interactive Poster and Demonstration Sessions, Ann Arbor, MI. Association for Computational Linguistics, 65–68.
- Paine, K. D., & Chaves, M. (2012). Social media metrics. SAS white paper. sas.com/resources/whitepaper/wp_19861.pdf (accessed February 2013).
- Pang, B., & Lee, L. (2008). *Opinion mining and sentiment analysis*. Hanover, MA: Now Publishers, available at <http://books.google.com>.

- Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009, August). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1* (pp. 248–256). Association for Computational Linguistics.
- Schmidt, L.-H. (1996). Commonness across cultures. In A. N. Balslev (Ed.), *Cross-cultural conversation: Initiation* (pp. 119–132). New York: Oxford University Press.
- Scott, W. R., & Davis, G. F. (2003). Networks in and around organizations. In *Organizations and Organizing*. Upper Saddle River, NJ: Pearson Prentice Hall.
- Shatkay, H., Höglund, A., Brady, S., Blum, T., Dönnies, P., & Kohlbacher, O. (2007). SherLoc: High-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics*, 23(11), 1410–1415.
- Snyder, B., & Barzilay, R. (2007, April). Multiple aspect ranking using the good grief algorithm. In *HLT-NAACL* (pp. 300–307).
- Strapparava, C., & Valitutti, A. (2004, May). WordNet affect: An affective extension of WordNet. In *LREC* (Vol. 4, pp. 1083–1086).
- The Westover Group. (2013). 20 key Web analytics metrics and how to use them. <http://www.thewestovergroup.com> (accessed February 2013).
- Thomas, M., Pang, B., & Lee, L. (2006, July). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 327–335). Association for Computational Linguistics.
- Weng, S. S., & Liu, C. K. (2004). Using text classification and multiple concepts to answer e-mails. *Expert Systems with Applications*, 26(4), 529–543.
- Zengul, F. D., Zengul, A. G., Mugavero, M. J., Oner, N., Ozaydin, B., Delen, D., ... & Cimino, J. (2022). A critical analysis of COVID-19 research literature: Text mining approach. *Intelligence-based Medicine*, 5, 100036.

Deep Learning and Cognitive Computing

LEARNING OBJECTIVES

- Learn what deep learning is and how it is changing the world of computing
- Know the placement of deep learning within the broad family of artificial intelligence (AI) learning methods
- Understand how traditional “shallow” artificial neural networks (ANN) work
- Become familiar with the development and learning processes of ANN
- Develop an understanding of the methods to shed light into the ANN black box
- Know the underlying concept and methods for deep neural networks
- Become familiar with different types of deep learning methods
- Understand how convolutional neural networks (CNN) work
- Learn how recurrent neural networks (RNN) and long short-memory networks (LSTM) work
- Become familiar with the computer frameworks for implementing deep learning
- Know the foundational details about cognitive computing
- Learn how IBM Watson works and what types of application it can be used for

Artificial intelligence (AI) is making a re-entrance into the world of computing and in our lives, this time far stronger and much more promising than before. This unprecedented re-emergence and the new level of expectations can largely be attributed to deep learning and cognitive computing. These two latest buzzwords define the leading edge of AI and machine learning today. Evolving out of the traditional **artificial neural networks (ANN)**, deep learning is changing the very foundation of how machine learning works. Thanks to large collections of data and improved computational resources, deep learning is making a profound impact on how computers can discover complex patterns using the self-extracted features from the data (as opposed to a data scientist providing the feature vector to the learning algorithm). Cognitive

computing—first popularized by **IBM Watson** and its success against the best human players in the game show *Jeopardy!*—makes it possible to deal with a new class of problems, the type of problems that are thought to be solvable only by human ingenuity and creativity, ones that are characterized by ambiguity and uncertainty. This chapter covers the concepts, methods, and application of these two cutting-edge AI technology trends.

- 7.1** Opening Vignette: Fighting Fraud with Deep Learning and Artificial Intelligence 369
- 7.2** Introduction to Deep Learning 373
- 7.3** Basics of “Shallow” Neural Networks 376
- 7.4** Process of Developing Neural Network–Based Systems 383
- 7.5** Illuminating the Black Box of ANN 389
- 7.6** Deep Neural Networks 395
- 7.7** Convolutional Neural Networks 400
- 7.8** Recurrent Networks and Long Short-Term Memory Networks 411
- 7.9** Computer Frameworks for Implementation of Deep Learning 420
- 7.10** Cognitive Computing 425

7.1 OPENING VIGNETTE: Fighting Fraud with Deep Learning and Artificial Intelligence

THE BUSINESS PROBLEM

Danske Bank is a Nordic universal bank with strong local roots and bridges to the rest of the world. Founded in October 1871, Danske Bank has helped people and businesses in the Nordics realize their ambitions for over 145 years. Its headquarters is in Denmark, with core markets in Denmark, Finland, Norway, and Sweden.

Mitigating fraud is a top priority for banks. According to the Association of Certified Fraud Examiners, businesses lose more than \$3.5 trillion each year to fraud. The problem is pervasive across the financial industry and is becoming more prevalent and sophisticated each month. As customers conduct more banking online across a wider variety of channels and devices, there are more opportunities for fraud to occur. Adding to the problem, fraudsters are becoming more creative and technologically savvy—they are also using advanced technologies such as machine learning—and new schemes to defraud banks are evolving rapidly.

Old methods for identifying fraud, such as using human-written rules engines, catch only a small percentage of fraud cases and produce a significantly high number of false positives. While false negatives end up costing money to the bank, chasing after a large number of false positives not only costs time and money but also blemishes customer trust and satisfaction. To improve probability predictions and identify a much higher percentage of actual cases of fraud while reducing false alarms, banks need new forms of analytics. This includes using artificial intelligence.

Danske Bank, like other global banks, is seeing a seismic shift in customer interactions. In the past, most customers handled their transactions in a bank branch. Today, almost all interactions take place digitally through a mobile phone, tablet, ATM, or call center. This provides more “surface area” for fraud to occur. The bank needed to modernize its fraud detection defenses. It struggled with a low 40 percent fraud detection rate and was managing up to 1200 false positives per day—and 99.5 percent of all cases the bank was investigating were not fraud related. That large number of false alarms required a substantial investment of people, time, and money to investigate what turned out to be

dead ends. Working with Think Big Analytics, a Teradata company, Danske Bank made a strategic decision to apply innovative analytic techniques, including AI, to better identify instances of fraud while reducing false positives.

THE SOLUTION: DEEP LEARNING ENHANCES FRAUD DETECTION

Danske Bank integrated deep learning with **graphics processing unit (GPU)** appliances that were also optimized for deep learning. The new software system helps the analytics team to identify potential cases of fraud while intelligently avoiding false positives. Operational decisions are shifted from users to AI systems. However, human intervention is still necessary in some cases. For example, the model can identify anomalies, such as debit card purchases taking place around the world, but analysts are needed to determine whether that is fraud or a bank customer simply made an online purchase that sent a payment to China and then bought an item the next day from a retailer based in London.

Danske Bank's analytic approach employs a "champion/challenger" methodology. With this approach, deep learning systems compare models in real time to determine which one is most effective. Each challenger processes data in real time, learning as it goes which traits are more likely to indicate fraud. If a process dips below a certain threshold, the model is fed more data, such as the geolocation of customers or recent ATM transactions. When a challenger outperforms other challengers, it transforms into a champion, giving the other models a roadmap to successful fraud detection.

THE RESULTS

Danske Bank implemented a modern enterprise analytic solution leveraging AI and deep learning, and it has paid big dividends. The bank was able to:

- Realize a 60 percent reduction in false positives with an expectation to reach as high as 80 percent.
- Increase true positives by 50 percent.
- Focus resources on actual cases of fraud.

The following graph (see Figure 7.1) shows how true and false positive rates improved with advanced analytics (including deep learning). The red dot represents the old rules engine, which caught only about 40 percent of all fraud. Deep learning improved significantly upon machine learning, allowing Danske Bank to better detect fraud with much lower false positives.

Enterprise analytics is rapidly evolving and moving into new learning systems enabled by AI. At the same time, hardware and processors are becoming more powerful and specialized, and algorithms more accessible, including those available through open source. This gives banks the powerful solutions needed to identify and mitigate fraud. As Danske Bank learned, building and deploying an enterprise-grade analytics solution that meets its specific needs and leverages its data sources deliver more value than traditional off-the-shelf tools could have provided. With AI and deep learning, Danske Bank now has the ability to better uncover fraud without being burdened by an unacceptable number of false positives. The solution also allows the bank's engineers, data scientists, lines of business, and investigative officers from Interpol, local police, and other agencies to collaborate to uncover fraud, including sophisticated fraud rings. With its enhanced capabilities, the enterprise analytic solution is now being used across other business areas of the bank to deliver additional value.

Because these technologies are still evolving, implementing deep learning and AI solutions can be difficult for companies to achieve on their own. They can

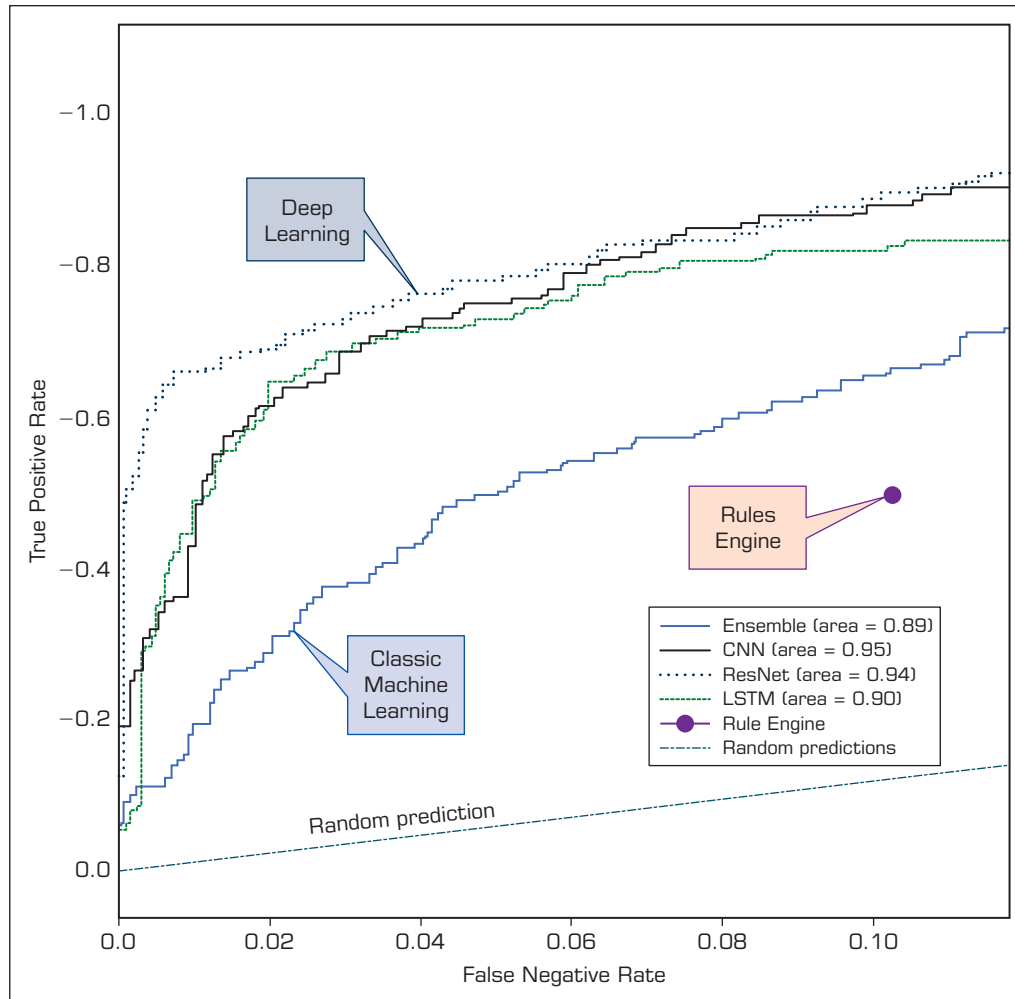


FIGURE 7.1 Deep Learning Improves Both True Positives and True Negatives.

benefit by partnering with a company that has the proven capabilities to implement technology-enabled solutions that deliver high-value outcomes. As shown in this case, Think Big Analytics, a Teradata company, has the expertise to configure specialized hardware and software frameworks to enable new operational processes. The project entailed integrating open-source solutions, deploying production models, and then applying deep learning analytics to extend and improve the models. A framework was created to manage and track the models in the production system and to make sure the models could be trusted. These models enabled the underlying system to make autonomous decisions in real time that aligned with the bank's procedural, security, and high-availability guidelines. The solution provided new levels of detail, such as time series and sequences of events, to better assist the bank with its fraud investigations. The entire solution was implemented very quickly—from kickoff to live in only five months. Figure 7.2 shows a generalized framework for AI and deep learning-based enterprise-level analytics solutions.

In summary, Danske Bank undertook a multi-step project to productionize machine-learning techniques while developing deep learning models to test those techniques. The integrated models helped identify the growing problem of fraud. For a visual summary, watch the video (<https://www.teradata.com/Resources/>

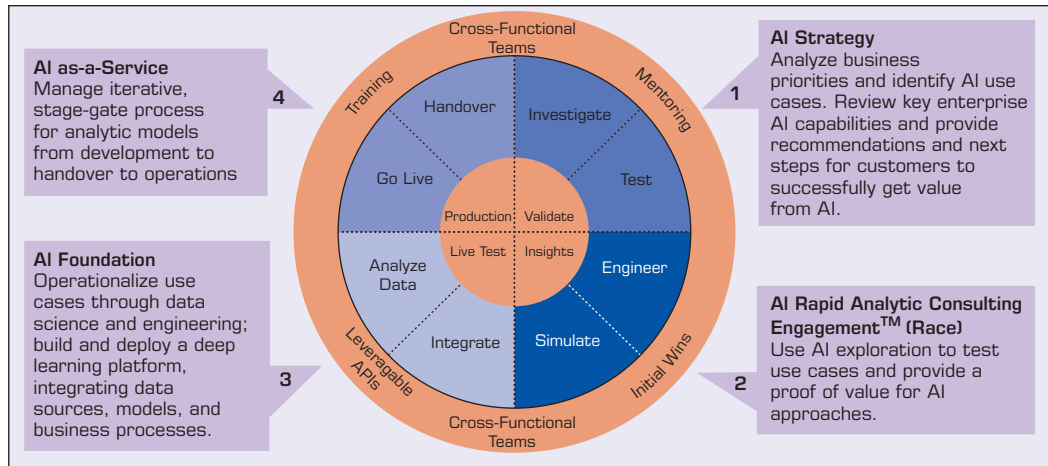


FIGURE 7.2 A Generalized Framework for AI and Deep Learning–Based Analytics Solutions.

Customer-Videos/Danske-Bank-Innovating-in-Artificial-Intelligence-and-Deep-Learning—accessed July 2022) and/or read the blog (<https://www.teradata.com/Blogs/Danske-Bank-Innovating-in-Artificial-Intelligence—accessed July 2022>).

► QUESTIONS FOR THE OPENING VIGNETTE

1. What is fraud in banking?
2. What are the types of fraud that banking firms are facing today?
3. What do you think are the implications of fraud on banks and on their customers?
4. Compare the old and new methods for identifying and mitigating fraud.
5. Why do you think deep learning methods provided better prediction accuracy?
6. Discuss the trade-off between false positive and false negative (type 1 and type 2 errors) within the context of predicting fraudulent activities.

WHAT WE CAN LEARN FROM THIS VIGNETTE

As you will see in this chapter, AI in general and the methods of machine learning in specific are evolving and advancing rapidly. The use of large digitized data sources, both from inside and outside the organization, both structured and unstructured, along with advanced computing systems (software and hardware combinations), has paved the way toward dealing with problems that were thought to be unsolvable just a few years ago. Deep learning and cognitive computing (as the ramifications of the cutting edge in AI systems) are helping enterprises to make accurate and timely decisions by harnessing the rapidly expanding Big Data resources. As shown in this opening vignette, this new generation of AI systems is capable of solving problems much better than their older counterparts. In the domain of fraud detection, traditional methods have always been marginally useful, having higher than desired false positive rates and causing unnecessary investigations and thereby dissatisfaction for their customers. As difficult problems such as fraud detection are, new AI technologies like deep learning are making them solvable with a high level of accuracy and applicability.

Source: Teradata Case Study. “Danske Bank Fights Fraud with Deep Learning and AI.” <https://www.teradata.com/Resources/Case-Studies/Danske-Bank-Fight-Fraud-With-Deep-Learning-and-AI> (accessed August 2018). Used with permission.

7.2 INTRODUCTION TO DEEP LEARNING

About a decade ago, conversing with an electronic device (in human language, intelligently) would have been unconceivable, something that could only be seen in sci-fi movies. Today, however, thanks to the advances in AI methods and technologies, almost everyone has experienced this unthinkable phenomenon. You probably have already asked Siri or Google Assistant several times to dial a number from your phone address book or to find an address and give you the specific directions while you were driving. Sometimes when you were bored in the afternoon, you may have asked the Google Home or Amazon's Alexa to play some music in your favorite genre on the device or your TV. You might have been surprised at times when you uploaded a group photo of your friends on Facebook and observed its tagging suggestions where the name tags often exactly match your friends' faces in the picture. Translating a manuscript from a foreign language does not require hours of struggling with a dictionary; it is as easy as taking a picture of that manuscript in the Google Translate mobile app and giving it a fraction of a second. These are only a few of the many, ever-increasing applications of deep learning that have promised to make life easier for people.

Deep learning, as the newest and perhaps at this moment the most popular member of the AI and machine-learning family, has a goal similar to those of the other machine-learning methods that came before it: mimic the thought process of humans—using mathematical algorithms to learn from data pretty much the same way that humans learn. So, what is really different (and advanced) in deep learning? Here is the most commonly pronounced differentiating characteristic of deep learning over traditional machine learning. The performance of traditional machine-learning algorithms such as decision trees, support vector machines, logistic regression, and neural networks relies heavily on the representation of the data. That is, only if we (analytics professionals or data scientists) provide those traditional machine-learning algorithms with relevant and sufficient pieces of information (a.k.a. features) in proper format are they able to “learn” the patterns and thereby perform their prediction (classification or estimation), clustering, or association tasks with an acceptable level of accuracy. In other words, these algorithms need humans to manually identify and derive features that are theoretically and/or logically relevant to the objectives of the problem on hand and feed these features into the algorithm in a proper format. For example, in order to use a decision tree to predict whether a given customer will return (or churn), the marketing manager needs to provide the algorithm with information such as the customer's socioeconomic characteristics—income, occupation, educational level, and so on (along with demographic and historical interactions/transactions with the company). But the algorithm itself is not able to define such socioeconomic characteristics and extract such features, for instance, from survey forms completed by the customer or obtained from social media.

While such a structured, human-mediated machine-learning approach has been working fine for rather abstract and formal tasks, it is extremely challenging to have the approach work for some informal, yet seemingly easy (to humans), tasks such as face identification or speech recognition since such tasks require a great deal of knowledge about the world (Goodfellow et al., 2016). It is not straightforward, for instance, to train a machine-learning algorithm to accurately recognize the real meaning of a sentence spoken by a person just by manually providing it with a number of grammatical or semantic features. Accomplishing such a task requires a “deep” knowledge about the world that is not easy to formalize and explicitly present. What deep learning has added to the classic machine-learning methods is in fact the ability to automatically acquire the knowledge required to accomplish such informal tasks and consequently extract some advanced features that contribute to the superior system performance.

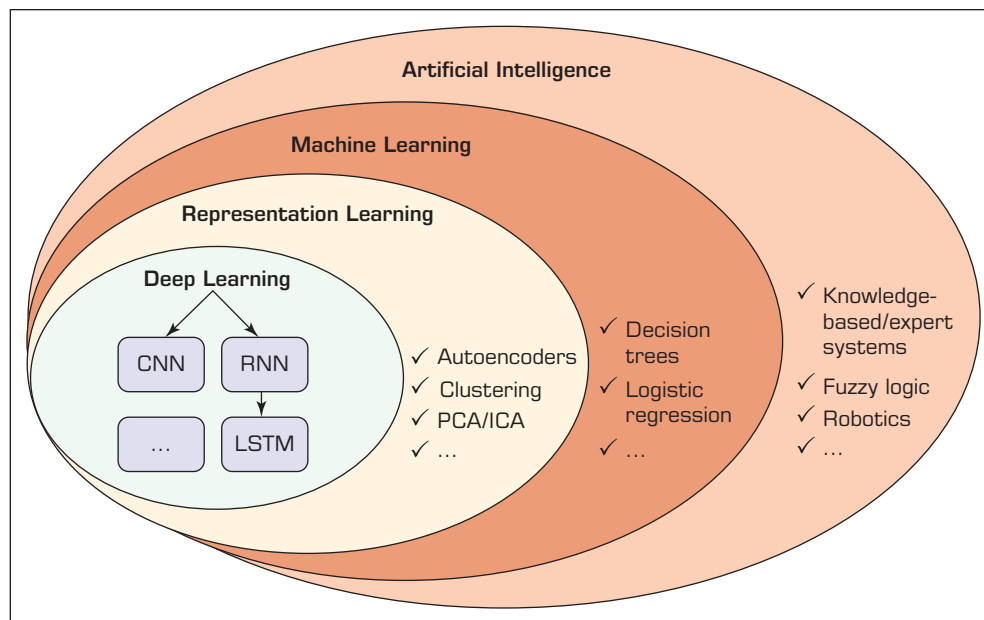


FIGURE 7.3 A Venn Diagram Showing the Placement of Deep Learning within the Overarching AI-Based Learning Methods.

To develop an intimate understanding of deep learning, one should learn where it fits in the big picture of all other AI family of methods. A simple hierarchical relationship diagram, or a taxonomy-like representation, may in fact provide such a holistic understanding. In an attempt to do this, Goodfellow and his colleagues (2016) categorized deep learning as part of the **representation learning** family of methods. Representation learning techniques entail one type of **machine learning** (which is also a part of AI) in which the emphasis is on learning and discovering features by the system in addition to discovering the mapping from those features to the output/target. Figure 7.3 uses a Venn diagram to illustrate the placement of deep learning within the overarching family of AI-based learning methods.

Figure 7.4 highlights the differences in the steps/tasks that need to be performed when building a typical deep learning model versus the steps/tasks performed when building models with classic machine-learning algorithms. As shown in the top two workflows, knowledge-based systems and classic machine-learning methods require data scientists to manually create the features (i.e., the representation) to achieve the desired output. The bottommost workflows show that deep learning enables the computer to derive some complex features from simple concepts that would be very effort intensive (or perhaps impossible in some problem situations) to be discovered by humans manually, and then it maps those advanced features to the desired output.

From a methodological viewpoint, although deep learning is generally believed to be a new area in machine learning, its initial idea goes back to the late 1980s, just a few decades after the emergence of artificial neural networks when LeCun and colleagues (1989) published an article about applying backpropagation networks for recognizing handwritten ZIP codes. In fact, as it is being practiced today, deep learning seems to be nothing but an extension of neural networks with the idea that deep learning is able to deal with more complicated tasks with a higher level of sophistication by employing many layers of connected **neurons** along with much larger data sets to automatically characterized variables and solve the problems but only at the expense of a great deal of computational effort.

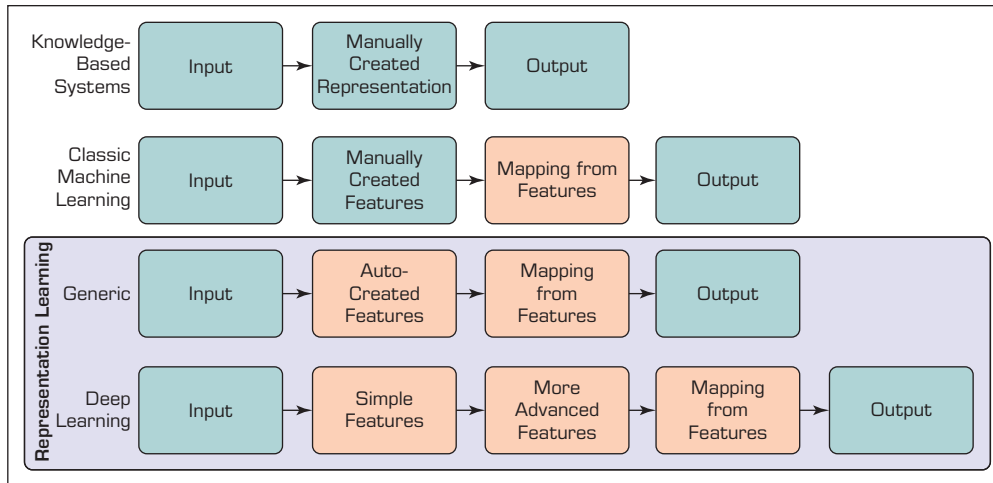


FIGURE 7.4 Illustration of the Key Differences between Classic Machine-Learning Methods and Representation Learning/Deep Learning (shaded boxes indicate components that are able to learn directly from data).

This very high computational requirement and the need for very large data sets were the two main reasons why the initial idea had to wait more than two decades until some advanced computational and technological infrastructure emerged for deep learning's practical realization. Although the scale of neural networks has dramatically increased in the past decade by the advancement of related technologies, it is still estimated that having artificial **deep neural networks** with the comparable number of neurons and level of complexity existing in the human brain will take several more decades.

In addition to the computer infrastructures, as mentioned, the availability of large and feature-rich digitized data sets was another key reason for the development of successful deep learning applications in recent years. Obtaining good performance from a deep learning algorithm used to be a very difficult task that required extensive skills and experience/understanding to design task-specific networks, and therefore, not many were able to develop deep learning for practical and/or research purposes. Large training data sets, however, have greatly compensated for the lack of intimate knowledge and reduced the level of skill needed for implementing deep neural networks. Nevertheless, although the size of available data sets has exponentially increased in recent years, a great challenge, especially for supervised learning of deep networks, is now the labeling of the cases in these huge data sets. As a result, a great deal of research is ongoing, focusing on how we can take advantage of large quantities of unlabeled data for semisupervised or unsupervised learning or how we can develop methods to label examples in bulk in a reasonable time.

The following section of this chapter provides a general introduction to **neural networks** from where deep learning has originated. Following the overview of these “shallow” neural networks, the chapter introduces different types of deep learning architectures and how they work, some common applications of these deep learning architectures, and some popular computer frameworks to use in implementing deep learning in practice. Since, as mentioned, the basics of deep learning are the same as those of artificial neural networks, in the following section, we provide a brief coverage of the neural network architecture (namely, multilayered **perceptron** [MLP]-type neural networks, which was omitted in the neural network section in Chapter 5 because it was to be covered here) to focus on their mathematical principles and then explain how the various types of deep learning architectures/approaches were derived from these foundations.

SECTION 7.2 REVIEW QUESTIONS

1. What is deep learning? What can deep learning do?
2. Compared to traditional machine learning, what is the most prominent difference of deep learning?
3. List and briefly explain different learning methods in AI.
4. What is representation learning, and how does it relate to deep learning?

7.3 BASICS OF “SHALLOW” NEURAL NETWORKS

Artificial neural networks are essentially simplified abstractions of the human brain and its complex biological networks of neurons. The human brain has a set of billions of interconnected neurons that facilitate our thinking, learning, and understanding of the world around us. Theoretically speaking, learning is nothing but the establishment and adaptation of new or existing interneuron connections. In the artificial neural networks, however, neurons are processing units (also called **processing elements [PEs]**) that perform a set of predefined mathematical operations on the numerical values coming from the input variables or from the other neuron outputs to create and push out its own outputs. Figure 7.5 shows a schematic representation of a single-input and single-output neuron (more accurately, the processing element in artificial neural networks).

In this figure, p represents a numerical input. Each input goes into the neuron with an *adjustable* weight w and a bias term b . A multiplication *weight function* applies the weight to the input, and a *net input function* shown by Σ adds the bias term to the weighted input z . The output of the net input function (n , known as the *net input*) then goes through another function called the *transfer* (a.k.a. activation) function (shown by f) for conversion and the production of the actual output a . In other words,

$$a = f(wp + b)$$

A numerical example: if $w = 2$, $p = 3$, and $b = -1$, then $a = f(2 * 3 - 1) = f(5)$

Various types of transfer functions are commonly used in the design of neural networks. Table 7.1 shows some of the most common transfer functions and their corresponding operations. Note that in practice, selection of proper transfer functions for a network requires a broad knowledge of neural networks—characteristics of the data as well as the specific purpose for which the network is created.

Just to provide an illustration, if in the previous example we had a hard limit transfer function, the actual output a would be $a = \text{hardlim}(5) = 1$. There are some guidelines for choosing the appropriate **transfer function** for each set of neurons in a network. These guidelines are especially robust for the neurons located at the output layer of the network. For example, if the nature of the output for a model is binary, we are advised to use *Sigmoid* transfer functions at the output layer so that it produces an output between 0 and 1, which represents the conditional probability of $y = 1$ given x or $P(y = 1|x)$. Many

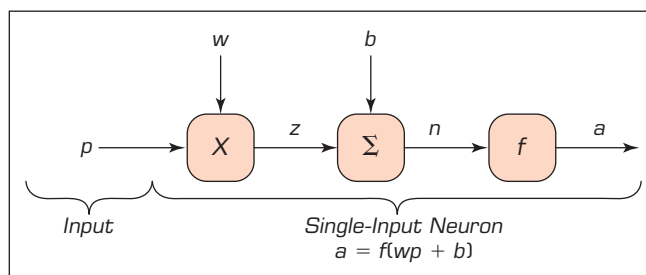
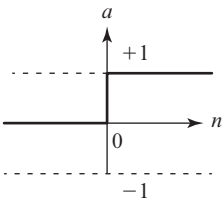
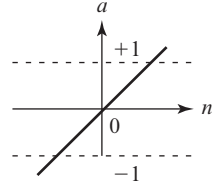
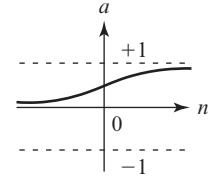
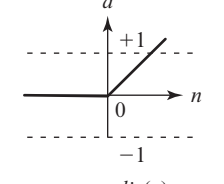


FIGURE 7.5 General Single-Input Artificial Neuron Representation.

TABLE 7.1 Common Transfer (Activation) Functions in Neural Networks

Transfer Function	Form	Operation
Hard limit	 $a = \text{hardlim}(n)$	$a = +1 \text{ if } n > 0$ $a = 0 \text{ if } n < 0$
Linear	 $a = \text{purelin}(n)$	$a = n$
Log-Sigmoid	 $a = \text{logsig}(n)$	$a = \frac{1}{1 + e^{-n}}$
Positive linear (a.k.a. rectified linear or ReLU)	 $a = \text{poslin}(n)$	$a = n \text{ if } n > 0$ $a = 0 \text{ if } n < 0$

neural network textbooks provide and elaborate on those guidelines at different layers in a neural network with some consistency and much disagreement, suggesting that the best practices should (and usually does) come from experience.

Typically, a neuron has more than a single input. In that case, each individual input p_i can be shown as an element of the input vector $\mathbf{p}$. Each of the individual input values would have its own adjustable weight w_i of the weight vector $\mathbf{W}$. Figure 7.6 shows a multiple-input neuron with R individual inputs.

For this neuron, the net input n can be expressed as:

$$n = w_{1,2}p_1 + w_{1,2}p_2 + w_{1,3}p_3 + \dots + w_{1,R}p_R + b$$

Considering the input vector $\mathbf{p}$ as a $R \times 1$ vector and the weight vector $\mathbf{W}$ as a $1 \times R$ vector, then n can be written in matrix form as:

$$n = \mathbf{W_p} + b$$

where $\mathbf{W_p}$ is a scalar (i.e., 1×1 vector).

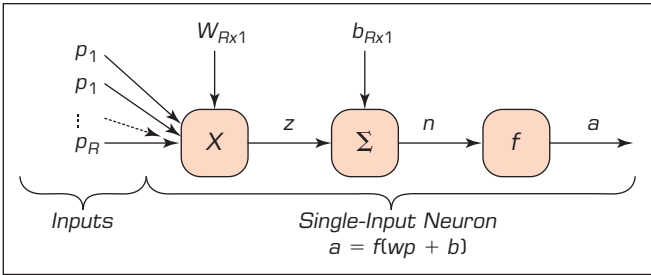


FIGURE 7.6 Typical Multiple-Input Neuron with R Individual Inputs.

Moreover, each neural network is typically composed of multiple neurons connected to each other and structured in consecutive *layers* so that the outputs of a layer work as the inputs to the next layer. Figure 7.7 shows a typical neural network with four neurons at the input (i.e., first) layer, four neurons at the hidden (i.e., middle) layer, and a single neuron at the output (i.e., last) layer. Each of the neurons has its own weight, weighting function, bias, and transfer function and processes its own input(s) as described.

While the inputs, weighting functions, and transfer functions in a given network are fixed, the values of the weights and biases are adjustable. The process of adjusting weights and biases in a neural network is what is commonly called *training*. In fact, in practice, a neural network cannot be used effectively for a prediction problem unless it is well trained by a sufficient number of examples with known *actual outputs* (a.k.a. *targets*). The goal of the training process is to adjust network weights and biases such that the network output for each set of inputs (i.e., each sample) is adequately close to its corresponding target value.

Analytics in Action 7.1 provides a case where computer gaming companies are using advanced analytics to better understand and engage with their customers.

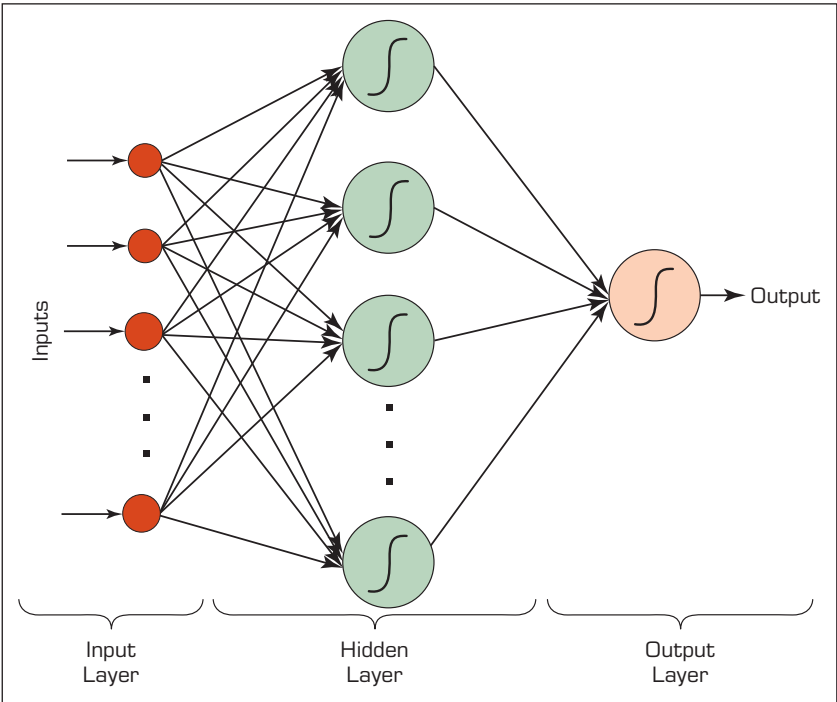


FIGURE 7.7 Typical Neural Network Architecture with Three Layers (Input, Hidden, and Output).

ANALYTICS IN ACTION 7.1 Gaming Companies Use Data Analytics to Score Points with Players

Video gamers are a special breed. Sure, they spend a lot of time playing games, but they're also building social networks. Like sports athletes, video game players thrive on competition. They play against other gamers online. Those who earn first place, or even second or third place, have bragging rights. And like athletes who invest a lot of time training, video gamers take pride in the number of hours they spend playing. Furthermore, as games increase in complexity, gamers take pride in developing unique skills to best their compatriots.

Video game companies can tap into this environment and learn valuable information about their customers, especially their behaviors and the underlying motivations. These customer data enable companies to improve the gaming experience and better engage players.

Traditionally, the gaming industry appealed to its customers—the gamers—by offering striking graphics and captivating visualizations. As technology advanced, the graphics became more vivid with hi-def renditions. Companies have continued to use technology in highly creative ways to develop games that attract customers and capture their interests, which results in more time spent playing and higher affinity levels. What video game companies have not done as well is to fully utilize technology to understand the contextual factors that drive sustained brand engagement.

Know the Players

In today's gaming world, creating an exciting product is no longer enough. Games must strongly appeal to the visual and auditory senses in an era when people expect cool graphics and cutting-edge sound effects. Games must also be properly marketed to reach highly targeted player groups. There are also opportunities to monetize gaming characters in the form of commercially available merchandise (e.g., toy store characters) or movie rights. Making a game successful requires programmers, designers, scenarists, musicians, and marketers to work together and share information. That is where gamer and gaming data come into play.

For example, the size of a gamer's network—the number and types of people a gamer plays with or against—usually correlates with more time spent playing and more money that is spent. The more relationships gamers have, the higher the likelihood they will play more games with more people because they enjoy the experience. Network effects amplify engagement volumes.

A New Level of Gaming

Video gaming has evolved from the days of PAC-MAN and arcades. The widespread availability of the Internet has fueled the popularity of video games by bringing them into people's homes via a wide range of electronics such as the personal computer and mobile devices. The world of computer games is now a powerful and profitable business.



According to NewZoo's *Global Games Market Report* from April 2017, the global games market in 2017 saw:

- \$109 billion in revenues.
- 7.8 percent increase from the previous year.
- 2.2 billion gamers globally.
- 42 percent of the market being mobile.

These data also help companies better understand the types of games each individual likes to play. These insights enable the company to recommend additional games across other genres that will likely exert a positive impact on player engagement and satisfaction. Companies can also use these data in marketing campaigns to target new gamers or entice existing gamers to upgrade their memberships, for example, to premium levels.

Monetize Player Behaviors

Collaborative filtering (cFilter) is an advanced analytic function that makes automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). The cFilter function supposes that if User A has the same opinion as User B on one issue, then User A is more likely to have User B's opinion on a different issue when compared to a random user. This shows that predictions are specific to a gamer based on data from many other gamers.

Filtering systems are often used by online retailers to make product recommendations. The analytics can determine products that a customer will like based on what other shoppers who made similar purchases also bought, liked, or rated highly. There are many examples across other industries such as healthcare, finance, manufacturing, and telecommunication.

The cFilter analytic function offers several benefits to online video game companies:

- **Marketers can run more effective campaigns.** Connections between gamers naturally form to create clusters. Marketers can isolate common player characteristics and leverage those insights for campaigns. Conversely, they can isolate players who do not belong to a cluster and determine what unique characteristics contribute to their nonconforming behaviors.
- **Companies can improve player retention.** A strong membership in a community of gamers decreases the chances of churn. The greater the incentives for gamers to belong to a group of active participants, the more desire they have to engage in competitions. This increases the “stickiness” of customers and can lead to more game subscriptions.
- **Data insights lead to improved customer satisfaction.** Clusters indicate a desire for certain types of games that correspond to distinct gamer interests and behaviors. Companies can create gaming experiences that are unique to each player. Enticing more people to play and play longer enhances gamer satisfaction.

Once they understand why customers want to play games and uncover their relationships with other gamers, companies can create the right incentives for players to keep returning. This ensures a sustained customer base and stable revenue streams.

Boost Loyalty and Revenue

Regardless of the genre, each video game has passionate players who seek each other for competitions. The thrill of a conquest attracts avid engagement. Over time, distinct networks of gamers are formed, with each participant constructing social relationships that often lead to more frequent and intense gaming interactions.

The gaming industry is now utilizing data analytics and visualizations to discern customer behaviors better and uncover player motivations. Looking at customer segments is no longer enough. Companies are now looking at microsegments that go beyond traditional demographics like age or geographic location to understand customer preferences such as favorite games, preferred levels of difficulty, or game genres.

By gaining analytic insights into gamer strategies and behaviors, companies can create unique gaming experiences that are attuned to these behaviors. By engaging players with the games and features they desire, video game companies gain a devoted following, grow profits, and develop new revenue streams through merchandising ventures.

For a visual treat, watch a short video (<https://www.teradata.com/Resources/Customer-Videos/Art-of-Analytics-The-Sword>) to see how the companies can use analytics to decipher gamer relationships that drive user behaviors and lead to better games.

Source: Teradata Case Study. <https://www.teradata.com/Resources/Case-Studies/Gaming-Companies-Use-Data-Analytics> (accessed July 2022).

Technology Insights 7.1 briefly describes the common components (or elements) of a typical artificial neural network along with their functional relationships.

TECHNOLOGY INSIGHTS 7.1 Elements of an Artificial Neural Network

A neural network is composed of processing elements that are organized in different ways to form the network's structure. The basic processing unit in a neural network is the neuron. A number of neurons are then organized to establish a network of neurons. Neurons can be organized in a number of different ways; these various network patterns are referred to as *topologies* or *network architectures*. Some of the most common ANN architectures are explained in detail in Chapter 5 of another textbook of the same authors (Sharda et al., 2021, pp. 259–261). One of the most popular approaches, known as the *feedforward-multilayered perceptron*, allows all neurons to link the output in one layer to the input of the next layer, but it does not allow any feedback linkage (Haykin, 2009).

Processing Element (PE)

The PE of an ANN is an artificial neuron. Each neuron receives inputs, processes them, and delivers a single output as shown in Figure 7.5. The input can be raw input data or the output of other processing elements. The output can be the final result (e.g., 1 means yes, 0 means no), or it can be input to other neurons.

Network Structure

Each ANN is composed of a collection of neurons that are grouped into layers, often called as **network structure**. A typical structure is shown in Figure 7.8. Note the three layers: input, intermediate (called *the hidden layer*), and output. A **hidden layer** is a layer of neurons that takes input from the previous layer and converts those inputs into outputs for further processing. Several hidden layers can be placed between the input and output layers, although it is common to use only one hidden layer. In that case, the hidden layer simply converts inputs

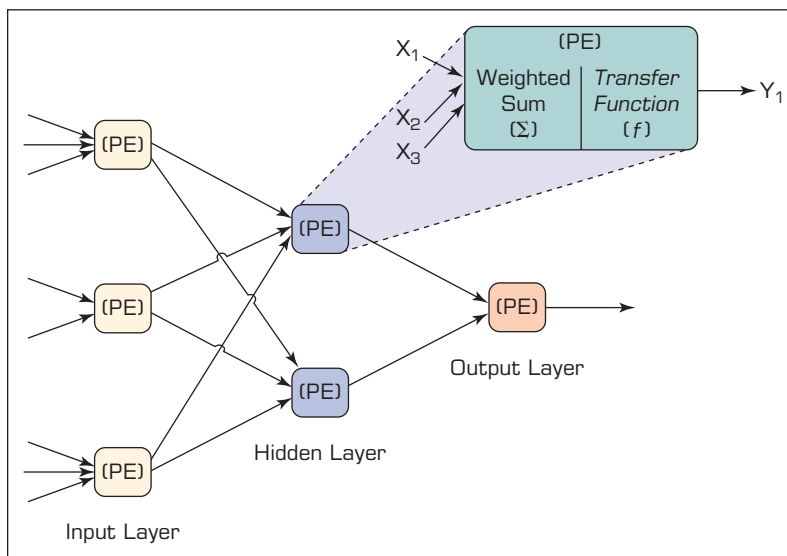


FIGURE 7.8 Neural Network with One Hidden Layer.

PE: processing element (an artificial representation of a biological neuron); X_i : inputs to a PE; y : output generated by a PE; Σ : summation function; and f : activation/transfer function.

into a nonlinear combination and passes the transformed inputs to the output layer. The most common interpretation of the hidden layer is as a feature-extraction mechanism; that is, the hidden layer converts the original inputs in the problem into a higher-level combination of such inputs.

In ANN, when information is processed, many of the processing elements perform their computations at the same time. This parallel processing resembles the way the human brain works, and it differs from the serial processing of conventional computing.

Inputs

Each input corresponds to a single attribute. For example, if the problem is to decide on approval or disapproval of a loan, attributes could include the applicant's income level, age, and home ownership status. The numeric value, or the numeric representation of non-numeric value, of an attribute is the input to the network. Several types of data, such as text, picture, and voice, can be used as inputs. Preprocessing may be needed to convert the data into meaningful inputs from symbolic/non-numeric data or to numeric/scale data.

Outputs

The output of a network contains the solution to a problem. For example, in the case of a loan application, the output can be “yes” or “no.” The ANN assigns numeric values to the output, which may then need to be converted into categorical output using a **threshold value** so that the results would be 1 for “yes” and 0 for “no.”

Connection Weights

Connection weights are the key elements of an ANN. They express the relative strength (or mathematical value) of the input data or the many connections that transfer data from layer to layer. In other words, weights express the relative importance of each input to a processing element and, ultimately, to the output. Weights are crucial in that they store learned patterns of information. It is through repeated adjustments of weights that a network learns.

Summation Function

The **summation function** computes the weighted sums of all input elements entering each processing element. A summation function multiplies each input value by its weight and totals the values for a weighted sum. The formula for n inputs (represented with X) in one processing element is shown in Figure 7.9a, and for several processing elements, the summation function formulas are shown in Figure 7.9b.

Transfer Function

The summation function computes the internal stimulation, or activation level, of the neuron. Based on this level, the neuron may or may not produce an output. The relationship between the internal activation level and the output can be linear or nonlinear. The relationship is expressed by one of several types of transformation (transfer) functions (see Table 7.1 for a list of commonly used activation functions). Selection of the specific **activation function** affects the network's operation. Figure 7.10 shows the calculation for a simple sigmoid-type activation function example.

The transformation modifies the output levels to fit within a reasonable range of values (typically between 0 and 1). This transformation is performed before the output reaches the next level. Without such a transformation, the value of the output becomes very large, especially when there are several layers of neurons. Sometimes a threshold value is used instead of a transformation function. A threshold value is a hurdle value for the output of a neuron to trigger the next level of neurons. If an output value is smaller than the threshold value, it will not be passed to the next level of neurons. For example, any value of 0.5 or less becomes 0, and any value above 0.5 becomes 1. A transformation can occur at the output of each processing element, or it can be performed only at the final output nodes. See Figure 7.10 for simple calculations of summation and transfer functions.

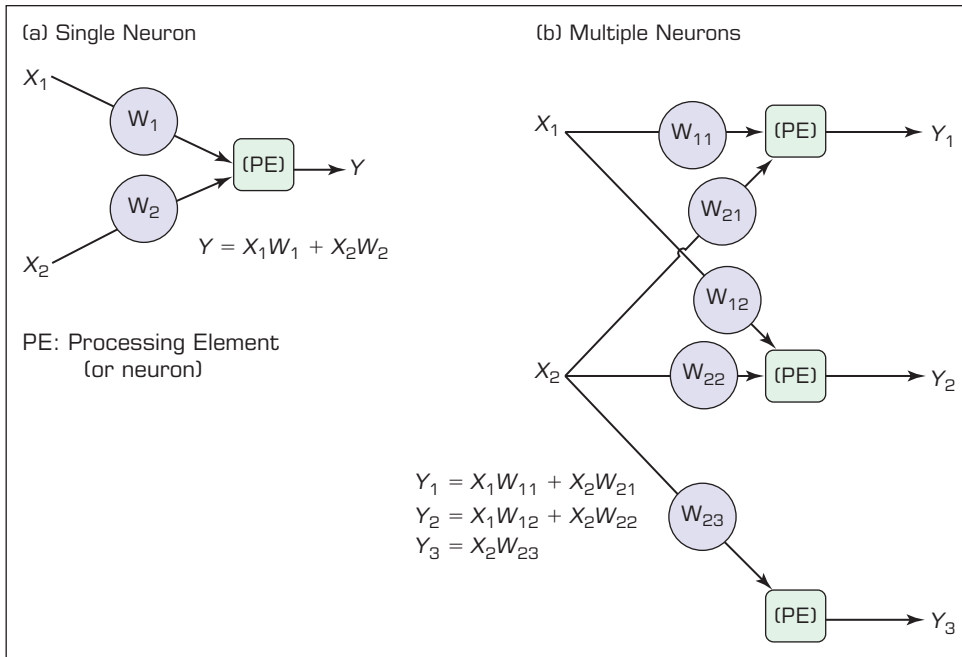


FIGURE 7.9 Summation Function for (a) a Single Neuron/PE and (b) Several Neurons/PEs.

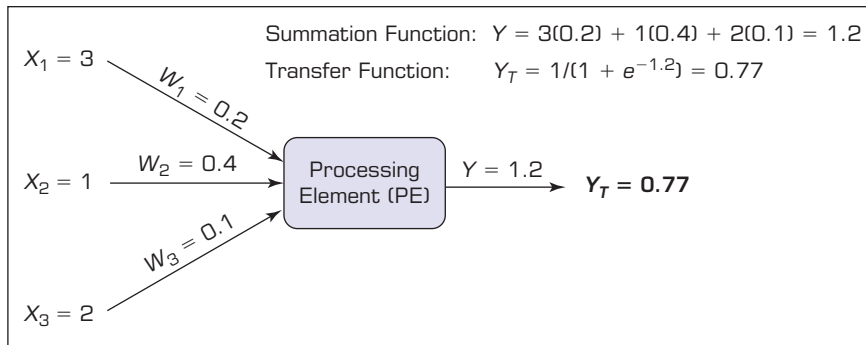


FIGURE 7.10 Example of ANN Transfer Function.

SECTION 7.3 REVIEW QUESTIONS

1. How does a single artificial neuron (i.e., PE) work?
2. List and briefly describe the most commonly used ANN activation functions.
3. What is MLP, and how does it work?
4. Explain the function of weights in ANN.
5. Describe the summation and activation functions in MLP-type ANN architecture.

7.4 PROCESS OF DEVELOPING NEURAL NETWORK-BASED SYSTEMS

Although the development process of ANN is similar to the structured design methodologies of traditional computer-based information systems, some phases are unique or have some unique aspects. In the process described here, we assume that the preliminary steps

of system development, such as determining information requirements, conducting a feasibility analysis, and gaining a champion in top management for the project, have been completed successfully. Such steps are generic to any information system.

As shown in Figure 7.11, the development process for an ANN application includes nine steps. In step 1, the data to be used for training and testing the network are collected. Important considerations are that the particular problem is amenable to a neural network solution and that adequate data exist and can be obtained. In step 2, training data must be identified, and a plan must be made for testing the performance of the network.

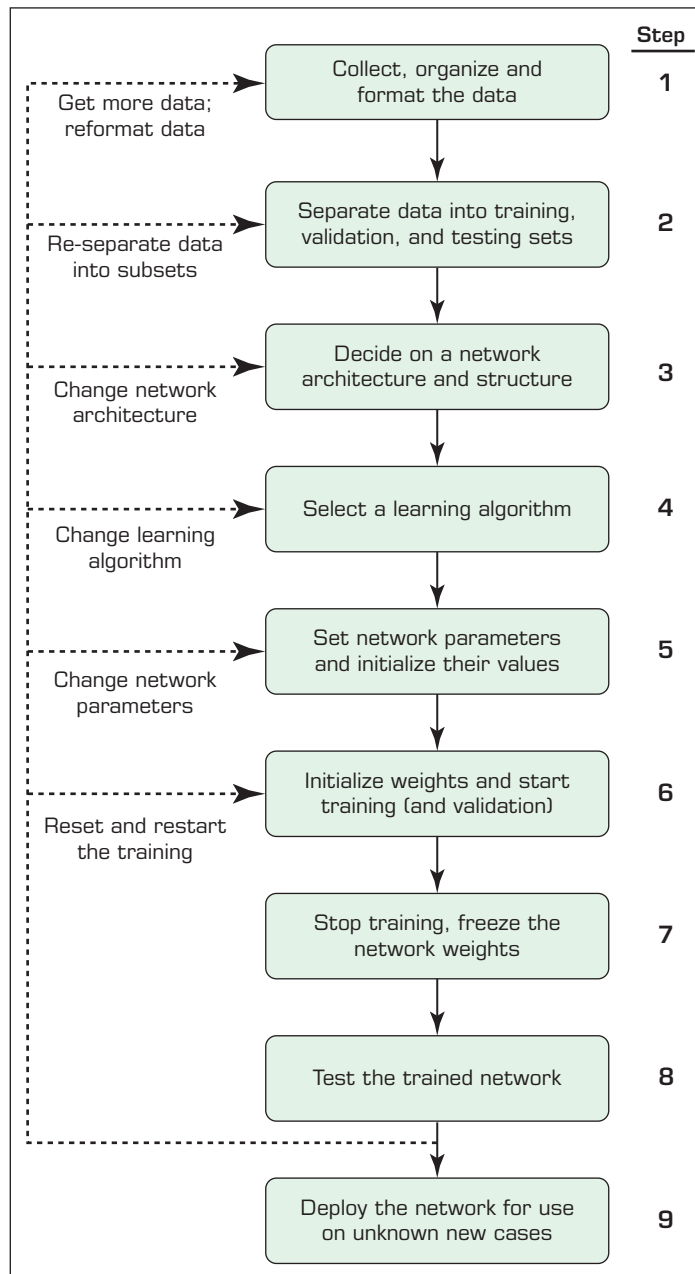


FIGURE 7.11 Development Process of an ANN Model.

In steps 3 and 4, a network architecture and a learning method are selected. The availability of a particular development tool or the capabilities of the development personnel may determine the type of neural network to be constructed. Also, certain problem types have demonstrated high success rates with certain configurations (e.g., multilayer feedforward neural networks for bankruptcy prediction [Altman (1968), Wilson and Sharda (1994), and Olson, Delen, and Meng (2012)]). Important considerations are the exact number of neurons and the number of layers. Some packages use genetic algorithms to select the network design.

There are several parameters for tuning the network to the desired learning performance level. Part of the process in step 5 is the initialization of the network weights and parameters followed by the modification of the parameters as training performance feedback is received. Often, the initial values are important in determining the efficiency and length of training. Some methods change the parameters during training to enhance performance.

Step 6 transforms the application data into the type and format required by the neural network. This may require writing software to preprocess the data or performing these operations directly in an ANN package. Data storage and manipulation techniques and processes must be designed for conveniently and efficiently retraining the neural network when needed. The application data representation and ordering often influence the efficiency and possibly the accuracy of the results.

In steps 7 and 8, training and testing are conducted iteratively by presenting input and desired or known output data to the network. The network computes the outputs and adjusts the weights until the computed outputs are within an acceptable tolerance of the known outputs for the input cases. The desired outputs and their relationships to input data are derived from historical data (i.e., a portion of the data collected in step 1).

In step 9, a stable set of weights is obtained. Then the network can reproduce the desired outputs given inputs such as those in the training set. The network is ready for use as a stand-alone system or as part of another software system where new input data will be presented to it and its output will be a recommended decision.

Learning Process in ANN

In **supervised learning**, the learning process is inductive; that is, connection weights are derived from existing cases. The usual process of learning involves three tasks (see Figure 7.12):

1. Compute temporary outputs.
2. Compare outputs with desired targets.
3. Adjust the weights and repeat the process.

Like any other supervised machine-learning technique, neural network training is usually done by defining a **performance function** (F) (a.k.a. *cost function* or *loss function*) and optimizing (minimizing) that function by changing model parameters. Usually, the performance function is nothing but a measure of error (i.e., the difference between the actual input and the target) across all inputs of a network. There are several types of error measures (e.g., sum square errors, mean square errors, cross entropy, or even custom measures) all of which are designed to capture the difference between the network outputs and the actual outputs.

The training process begins by calculating outputs for a given set of inputs using some random weights and biases. Once the network outputs are on hand, the performance function can be computed. The difference between the actual output (Y or Y_T) and the desired output (Z) for a given set of inputs is an error called *delta* (in calculus, the Greek symbol Δ , means “difference”).

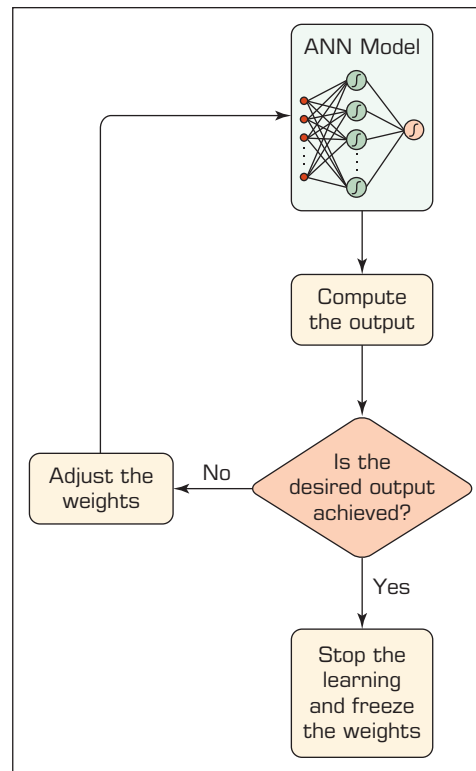


FIGURE 7.12 Supervised Learning Process of an ANN.

The objective is to minimize delta (i.e., reduce it to 0 if possible), which is done by adjusting the network's weights. The key is to change the weights in the proper direction, making changes that reduce delta (i.e., error). Different ANNs compute delta in different ways, depending on the learning algorithm being used. Hundreds of learning algorithms are available for various situations and configurations of ANN.

Backpropagation for ANN Training

The optimization of performance (i.e., minimization of the error or delta) in the neural network is usually done by an algorithm called **stochastic gradient descent (SGD)**, which is an iterative gradient-based optimizer used for finding the minimum (i.e., the lowest point) in performance functions, as in the case of neural networks. The idea behind the SGD algorithm is that the derivative of the performance function with respect to each current weight or bias indicates the amount of change in the error measure by each unit of change in that weight or bias element. These derivatives are referred to as *network gradients*. Calculation of network gradients in the neural networks requires application of an algorithm called **backpropagation**, which is the most popular neural network learning algorithm, that applies *the chain rule of calculus* to compute the derivatives of functions formed by composing other functions whose derivatives are known [more on the mathematical details of this algorithm can be found in Rumelhart, Hinton, and Williams (1986)].

Backpropagation (short for *back-error propagation*) is the most widely used supervised learning algorithm in neural computing (Principe, Euliano, and Lefebvre, 2000). By using the SGD mentioned previously, the implementation of backpropagation algorithms is relatively straightforward. A neural network with backpropagation learning includes

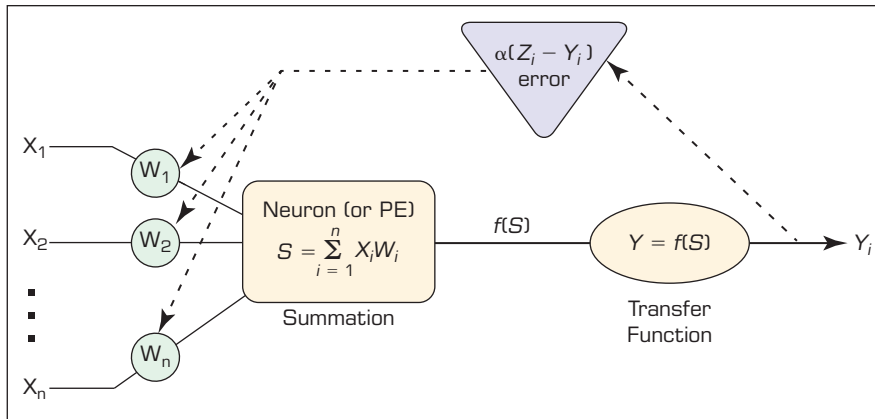


FIGURE 7.13 Backpropagation of Error for a Single Neuron.

one or more hidden layers. This type of network is considered feedforward because there are no interconnections between the output of a processing element and the input of a node in the same layer or in a preceding layer. Externally provided correct patterns are compared with the neural network's output during (supervised) training, and feedback is used to adjust the weights until the network has categorized all training patterns as correctly as possible (the error tolerance is set in advance).

Starting with the output layer, errors between network-generated actual output and the desired outputs are used to correct/adjust the weights for the connections between the neurons (see Figure 7.13). For any output neuron j , the error (delta) $= (Z_j - Y_j) (df/dx)$, where Z and Y are the desired and actual outputs, respectively. Using the sigmoid function, $f = [1 + \exp(-x)]^{-1}$, where x is proportional to the sum of the weighted inputs to the neuron, is an effective way to compute the output of a neuron in practice. With this function, the derivative of the sigmoid function $df/dx = f(1 - f)$ and of the error is a simple function of the desired and actual outputs. The factor $f(1 - f)$ is the logistic function, which serves to keep the error correction well bounded. The weight of each input to the j th neuron is then changed in proportion to this calculated error. A more complicated expression can be derived to work backward in a similar way from the output neurons through the hidden layers to calculate the corrections to the associated weights of the inner neurons. This complicated method is an iterative approach to solving a nonlinear optimization problem that is very similar in meaning to the one characterizing multiple linear regression.

In backpropagation, the learning algorithm includes the following procedures:

1. Initialize weights with random values and set other parameters.
2. Read in the input vector and the desired output.
3. Compute the actual output via the calculations, working forward through the layers.
4. Compute the error.
5. Change the weights by working backward from the output layer through the hidden layers.

This procedure is repeated for the entire set of input vectors until the desired output and the actual output agree within some predetermined tolerance. Given the calculation requirements for one iteration, training a large network can take a very long time; therefore, in one variation, a set of cases is run forward, and an aggregated error is fed backward to speed the learning. Sometimes, depending on the initial random weights

and network parameters, the network does not converge to a satisfactory performance level. When this is the case, new random weights must be generated, and the network parameters, or even its structure, may have to be modified before another attempt is made. Current research is aimed at developing algorithms and using parallel computers to improve this process. For example, genetic algorithms (GA) can be used to guide the selection of the network parameters to maximize the performance of the desired output. In fact, most commercial ANN software tools are now using GA to help users “optimize” the network parameters in a semiautomated manner.

A central concern in the training of any type of machine-learning model is **overfitting**. It happens when the trained model is highly fitted to the training data set but performs poorly with regard to external data sets. Overfitting causes serious issues with respect to the generalizability of the model. A large group of strategies known as *regularization* strategies is designed to prevent models from overfitting by making changes or defining constraints for the model parameters or the performance function.

In the classic ANN models of small size, a common regularization strategy to avoid overfitting is to assess the performance function for a separate validation data set as well as the training data set after each iteration. Whenever the performance stopped improving for the validation data, the training process would be stopped. Figure 7.14 shows a typical graph of the error measure by the number of iterations of training. As shown, in the beginning, the error decreases in both training and validation data by running more and more iterations, but from a specific point (shown by the dashed line), the error starts increasing in the validation set while still decreasing in the training set. It means that beyond that number of iterations, the model becomes overfitted to the data set with which it is trained and cannot necessarily perform well when it is fed with some external data. That point actually represents the recommended number of iterations for training a given neural network.

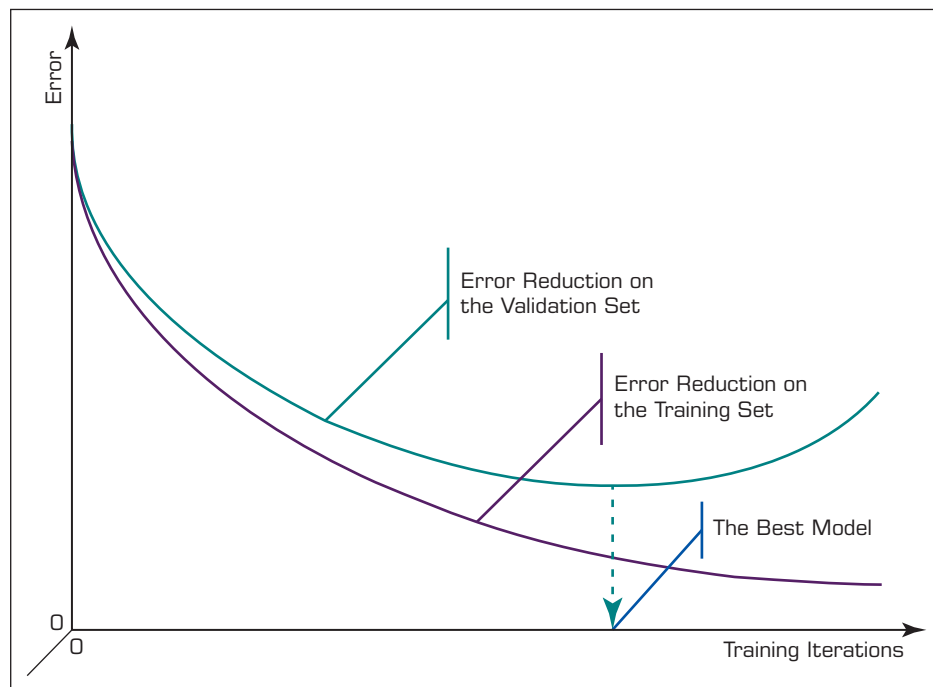


FIGURE 7.14 Illustration of the Overfitting in ANN—Gradually Changing Error Rates in the Training and Validation Data Sets as the Number of Iterations Increases.

► SECTION 7.4 REVIEW QUESTIONS

1. List the nine steps in conducting a neural network project.
2. What are some of the design parameters for developing a neural network?
3. Draw and briefly explain the three-step process of learning in ANN.
4. How does backpropagation learning work?
5. What is overfitting in ANN learning? How does it happen, and how can it be mitigated?
6. Describe the different types of neural network software available today.

7.5 ILLUMINATING THE BLACK BOX OF ANN

Machine learning methods in general, and artificial neural networks in specific, have been used as effective tool for solving highly complex real-world problems in a wide range of application areas. Even though these tools have been proven to be more superior predictors and/or cluster identifiers in many problem scenarios (compared to their traditional counterparts), in some applications, there exists an additional need to know “how the model does what it does.” Most machine learning techniques, led by ANN, are typically known as black boxes, capable of solving complex problems but lacking the explanation of their capabilities. This lack of transparency situation is commonly referred to as the “black box” syndrome.

It is important to be able to explain a model’s “inner being”; such an explanation offers assurance that the network has been properly trained and will behave as desired once deployed in a business analytics environment. Such a need to “look under the hood” might be attributable to a relatively small training set (as a result of the high cost of data acquisition) or a very high liability in case of a system error. One example of such an application is the deployment of airbags in vehicles. Here, both the cost of data acquisition (crashing vehicles) and the liability concerns (danger to human lives) are rather significant. Another representative example for the importance of explanation is loan-application processing. If applicants are refused a loan, they have the right to know why. Having a prediction system that does a good job on differentiating good and bad applications may not be sufficient if it does not also provide the justification of its predictions.

Technology Insights 7.2 provides an overview of prediction model explainability/transparency (nowadays, also called as explainable AI (XAI, in short), explainable machine learning (XML, in short)) techniques, along with their advantage and disadvantages.

TECHNOLOGY INSIGHTS 7.2 Explainability and Transparency in Machine Learning Models

In predictive analytics and machine learning, there has been a trade-off between model complexity and model performance—more complex the machine learning models (e.g., deep neural networks, random forest, and gradient boosted machines) more accurate is the prediction outcomes. As the model gets simpler (move toward more parsimony), less predictive performance is observed, but at the same time, more interpretable and transparent the model becomes. The seminal research paper by Ribiero et al. (2016) entitled “Why Should I Trust You,” appropriately highlights the issue with machine learning models and their infamous nature of being black boxes. With the new and enhanced interest in machine learning methods, especially since the advent of model ensembles and deep learning, model interpretability has become a fast-growing field of research; nowadays this trend is appropriately called as Explainable AI and Human Interpretable Machine Learning.

As you have seen in this book, predictive analytics methods and their underlying machine learning algorithms are really good at capturing complex relationships between input

and output variables (producing very accurate prediction models), but are not nearly as good at explaining how they do what they do (i.e., model transparency/model explainability). In order to mitigate this deficiency (also called the **black-box syndrome**), machine learning community proposed several methods, most of which are characterized as sensitivity analysis. Some of these methods are global (providing explanation based on average score of all data samples) and some are local (providing single sample level explanations). In the context of predictive modeling, **sensitivity analysis** usually refers to an exclusive experimentation process designed and executed to discover the cause and effect relationship between the input variables and output variable. Some of the variable importance methods are model/algorithm specific (i.e., applied to decision trees, neural networks, or random forest) and some are model/algorithm agnostic (i.e., applied to any and every predictive model). Here are the most commonly used variable importance methods employed in machine learning and predictive modeling:

1. Developing and observing a well-trained decision tree model to see the relative discernibility of the input variables—closer to the root of the tree a variable is used to split, greater is its importance/relative-contribution to the prediction model.
2. Developing and observing a rich and large random forest model and assessing the variable split statistics. If the ratio of a given variable's selection into candidate counts (i.e., number of times a variable selected as the level-0 splitter divided by number of time it was picked randomly as one of split candidates) is larger, then its importance/relative-contribution is also greater. This process can be extended to the top three layers of the trees to generate a weights average of the split statistics and can be used as a measure of variable importance for random forest models.
3. Sensitivity analysis based on input value perturbation, where the input variables are gradually and systematically changed/perturbed one at a time and the relative change in the output is observed—larger the change in the output, greater the importance of the perturbed variable. This method is often used in trained feed-forward neural network modeling where all of the input variables are numeric and standardized/normalized.
4. Sensitivity analysis based on leave-one-out methodology. This method can be applied to any type of predictive analytics method (i.e., predictive model agnostic). Because of its general applicability and ease of implementation, this sensitivity analysis method is used in several commercial and free/open source analytics tools as default, and therefore is further explained below.
5. Sensitivity analysis based on developing a surrogate model to assess the variable importance of a single record/sample using LIME or SHAP methodologies. While the previous methods are considered global interpreters, LIME and SHAP are called local because they explain the importance of the variables at the sample level (as opposed to the average of all samples). Because of the increased interest in these latest methods, they (LIME and SHAP) will be explained below.

Sensitivity analysis based on leave-one-out methodology.

The sensitivity analysis based on leave-one-out methodology relies on the experimental process of systematically removing input variables, one at a time, from the input variable set, developing and testing a model, and observing the impact of the absence of this variable on the predictive performance of the machine learning model (see Figure 7.15 for a graphical illustration of the sensitivity analysis process). Specifically, in the process of leave-one-out sensitivity analysis, first, the model is trained and tested (often using a k -fold cross validation) with all input variables, and the best predictive accuracy is recorded as the baseline. Then, the same model with the exact same settings is trained and tested with all but one input variable, and the best prediction accuracy is recorded, and this process is repeated as many times as the number of variables, each time omitting/excluding a different input variable. Then the degradation from the baseline is measured for the omission of each variable, and these measures are used to create a tabular and graphical illustration of the variable importance.

This sensitivity analysis method can be used for any predictive analytics and machine learning method, and historically, it has often been used for support vector machines, decision trees, logistic regression as well as for artificial neural networks. Saltelli (2008), in his

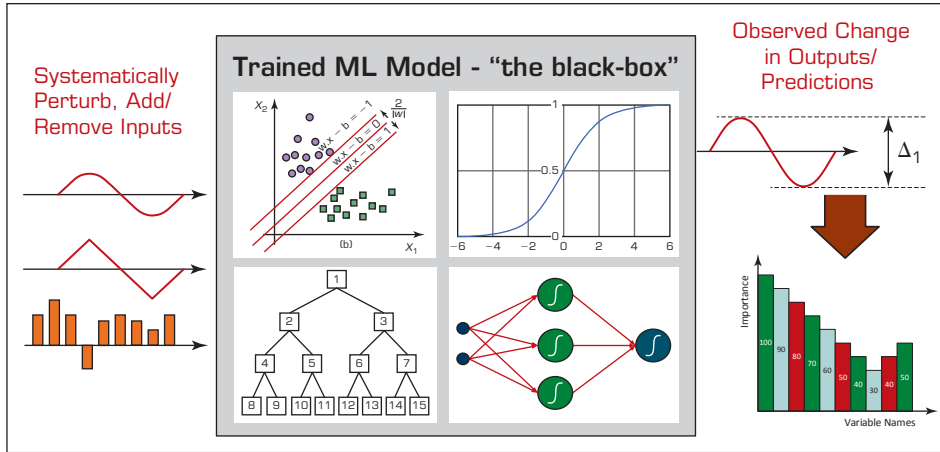


FIGURE 7.15 A Graphical Depiction of the Sensitivity Analysis Process.

sensitivity analysis book, formalized the algebraic representation of this measurement process with the following equation.

$$S_i = \frac{V_i}{V(F_t)} = \frac{V(E(F_t|X_i))}{V(F_t)}$$

In this equation, the denominator, $V(F_t)$, refers to the variance in the output variable. In the numerator, $V(E(F_t|X_i))$, E is the expectation operator to call for an integral over parameter X_i ; that is, inclusive of all input variables except X_i , the V , the variance operator applies a further integral over X_i . The variable contribution (i.e., importance), represented as S_i , for the i^{th} variable, is calculated as the normalized sensitivity measure. In a later study, Saltelli et al. (2019) proved that this equation is the most probable measure of model sensitivity that is capable of ranking input variables (i.e., the predictors) in the order of importance for any combination of interactions including the non-orthogonal relationships amongst the input variables.

Using this method of sensitivity analysis may produce slightly different importance measures for different model types for the same dataset. In such situation, we can either select and use the variable importance measured of the most predictive model type while ignoring the ones generated by the other model types, or we can use some type combination (i.e., ensemble) of all importance measures produced by all model types. In order to properly combine the sensitivity analysis results for all prediction model types, we can use an information fusion-based methodology—a weighted average where the weights are determined based on the prediction power of the individual model types. Particularly, by modifying the above equation in such a way that the sensitivity measure of an input variable n obtained based on the information combined (i.e., fused) from m number of prediction model types. The following equation represents this waited summation function.

$$S_{n(\text{fused})} = \sum_{i=1}^m \omega_i S_{in} = \omega_1 S_{1n} + \omega_2 S_{2n} + \dots + \omega_m S_{mn}$$

In this equation ω_i represents the normalized contribution/weight for each prediction model where the level of contribution/weight of a model is calculated as a function of its relative predictive power—larger the prediction power (i.e., accuracy) greater the value of ω .

Local Interpretability with LIME and SHAP

In linear regression models, *beta* coefficients explain the prediction for all the data points (i.e., if a variable value increases by one unit, prediction value increases by beta for every data point). This is commonly called as global interpretability. In a causal modeling, this is called average causal analysis. However, this does not specifically explain individual differences—the effect

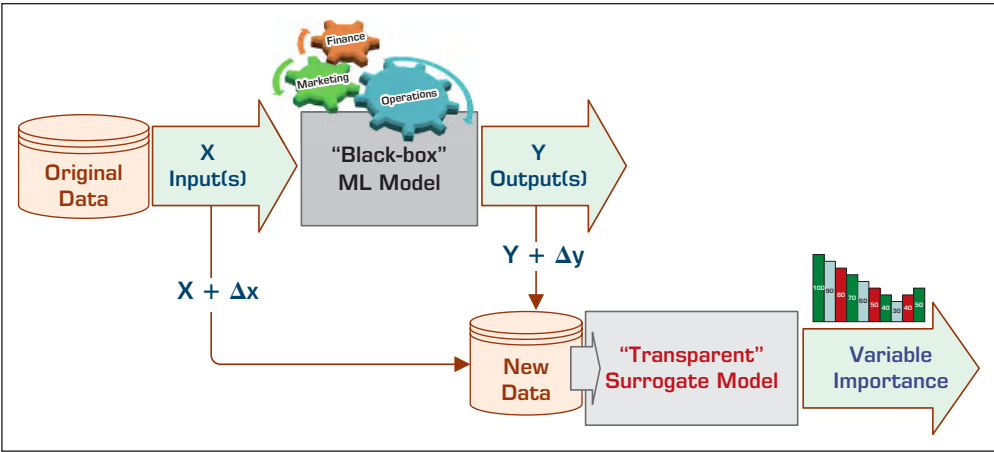


FIGURE 7.16 From Global Black-Box Model to Local Surrogate Model for Variable Importance.

of variable values for rejecting the loan application of one applicant could be different than another. This is local often called local interpretability, which is the explanation for an individual data point or local subsection of the joint distribution of independent variables. LIME and SHAP are the two recent and popular methods that aim to provide local explainability by building surrogate models to the black-box. They tweak the input slightly (like we do in sensitivity analysis) and test the changes in prediction through a surrogate model representation (Figure 7.16). Since these surrogate models still treat the machine learning models as a black box, these methodologies are considered as model agnostic.

Local Interpretable Model-Agnostic Explanations (LIME) is a relatively new variable assessment method developed to explain the prediction of any prediction model (i.e., classifier) in a human interpretable manner by developing/learning a surrogate model locally based on the specifics of the prediction (the predicted case/sample). The main idea behind LIME is to simulate the relationships between inputs and outputs of a complex machine learning model by developing a surrogate (similar but much simpler and explainable) model for explaining the prediction results of a single record (i.e., a specific customer). To do so, the surrogate model utilizes a sample of synthetically generated records closely resembling the actual predicted record itself. What have made LIME a popular technique includes it being model agnostic (works with any type of prediction model), has the ability to generate human interpretable outcomes, providing the explanation at the local (single record) level, being additive (the sum of importance values of all the variables for a data point is equal to the final prediction), and being a computationally very efficient algorithm. More on LIME and its algorithmic details can be found in Ribeiro et al. (2016).

Shapley Additive Explanations (SHAP) is another recently proposed technique for model interpretation. It is widely considered as the best local model interpretation technique, superior to LIME in several dimensions. Although SHAP employs a similar concept to LIME, it provides theoretical guarantees based on the game theory concept of Shapley values; it is capable of capturing and representing complex relationships at the local level beyond just linear relationships, and produces more accurate, robust, and reliable interpretation outcomes. SHAP can also be used effectively for variable selection—we can select a subset of the variables with higher variable importance, drop other variables (if we want to reduce the number of variables); because of its consistency property, the order of variable importance does not change, hence less important variables can be ignored. More detailed on SHAP can be found in Lundberg and Lee (2017).

Which technique is the best? As is the case almost always, there is no magic answer or solution to this question that applies to all problems and situations. If you are interested in model agnostic global explainability, perhaps the leave-one-out type sensitivity analysis could be used. If a model specific interpretation is desired, then the appropriate sensitivity analysis for

the specific model type could be utilized. If you desire local interpretation, then either LIME or SHAP can be used; out of these two, SHAP is considered as a better method but it is computationally more expensive than LIME. If a quick explanation at the single record level is needed for a large and complex problem space, perhaps LIME could be the method of choice.

In ANNs, for model explainability and/or variable importance identification, value perturbation type sensitivity analysis has been a widely practiced technique. Perhaps this is because the input variables in ANN are all numerically represented, and numerical perturbation of variables is feasible, relatively easy, and hence, a popular manipulation technique. In this case, value perturbation-based sensitivity analysis is a method for extracting the cause-and-effect relationships among the inputs and the outputs of a trained neural network model. In the process of performing sensitivity analysis, the trained neural network's learning capability is disabled so that the network weights are not changed/affected. The basic procedure behind this type of sensitivity analysis is that the inputs to the network are systematically changed/perturbed within the allowable value ranges, and the corresponding change in the output is recorded for each and every input variable (Principe et al., 2000). The top portion of Figure 7.17 shows a graphical illustration of this process. The first input is varied between its mean plus and minus of a user-defined number of standard deviations (or for categorical variables, all of its possible values are used) while all other input variables are fixed at their respective means (or modes). The network output is computed for a user-defined number of steps above and below the mean. This process is repeated for each input. As a result, a report is generated to summarize the variation of each output with respect to the variation in each input. The generated report often contains a column plot (along with numeric values presented on the x -axis), reporting the relative sensitivity values for each input variable. A representative example of sensitivity analysis on ANN models is provided in Analytics in Action 7.2.

ANALYTICS IN ACTION 7.2 Sensitivity Analysis Reveals Injury Severity Factors in Traffic Accidents

According to the National Highway Traffic Safety Administration (NHTSA), over 6 million traffic accidents claim more than 40,000 lives each year in the United States. The causes of accidents and related injury severity are of special interest to traffic safety researchers. Such research is aimed at reducing not only the number of accidents but also the severity of injury. One way to accomplish the latter is to identify the most profound factors that affect injury severity. Understanding the circumstances under which drivers and passengers are more likely to be severely injured (or killed) in a vehicle accident can help improve the overall driving safety situation. Factors that potentially elevate the risk of injury severity of vehicle occupants in the event of an accident include demographic and/or behavioral characteristics of the person (e.g., age, gender, seatbelt usage, use of drugs or alcohol while driving), environmental factors, and/or roadway conditions at the time of the accident (e.g., surface conditions, weather or light conditions, direction of the impact, vehicle orientation in the crash, occurrence of a rollover), as well as technical characteristics of the vehicle itself (e.g., age, body type) (Topuz and Delen, 2021).

In one of the exploratory data mining studies, Delen et al. (2006) used a large sample of data—30,358 police-reported accident records obtained from the General Estimates System of NHTSA—to identify which factors become increasingly more important in escalating the probability of injury severity during a traffic crash. Accidents examined in this study included a geographically representative sample of multiple-vehicle collision accidents, single-vehicle fixed-object collisions, and single-vehicle noncollision (rollover) crashes.

Contrary to many of the previous studies conducted in this domain, which have primarily used regression-type generalized linear models where the functional relationships between injury severity and crash-related factors are assumed to be linear (which is an oversimplification of the reality in most real-world situations), Delen and his colleagues (2006) decided to go in a different direction.

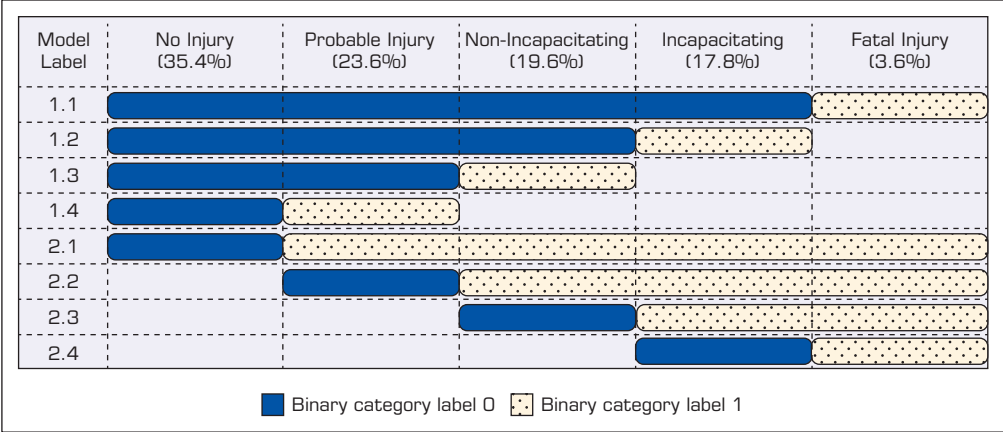


FIGURE 7.17 Graphical Representation of the Eight Binary ANN Model Configurations.

Sources: Topuz, K., & Delen, D. (2021). A probabilistic Bayesian inference model to investigate injury severity in automobile crashes. *Decision Support Systems*, 150, 113557; Delen, D., R. Sharda, & M. Bessonov. (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. *Accident Analysis and Prevention*, 38(3).

Because NN are known to be superior in capturing highly nonlinear complex relationships between the predictor variables (crash factors) and the target variable (severity level of the injuries), they decided to use a series of ANN models to estimate the significance of the crash factors on the level of injury severity sustained by the driver.

From a methodological standpoint, Delen et al. (2006) followed a two-step process. In the first step, they developed a series of prediction models (one for each injury severity level) to capture the in-depth relationships between the crash-related factors and a specific level of injury severity. In the second step, they conducted sensitivity analysis on the trained neural network models to identify the prioritized importance of crash-related factors as they relate to different injury severity levels. In the formulation of the study, the five-class prediction problem was decomposed into a number of binary classification models to obtain the granularity of information needed to identify the “true” cause-and-effect relationships between the crash-related factors and different levels of injury severity. As shown in Figure 7.18, eight different neural network models have been developed and used in the sensitivity analysis to identify the key determinants of increased injury severity levels.

The results revealed considerable differences among the models built for different injury severity levels. This implies that the most influential factors in prediction models highly depend on the level of injury severity. For example, the study revealed that the variable seat-belt use was the most important determinant for predicting higher levels of injury severity (such as incapacitating injury or fatality), but it was one of the least significant predictors for lower levels of injury severity (such as non-incapacitating injury and minor injury). Another interesting finding involved gender: The driver’s gender was among the significant predictors

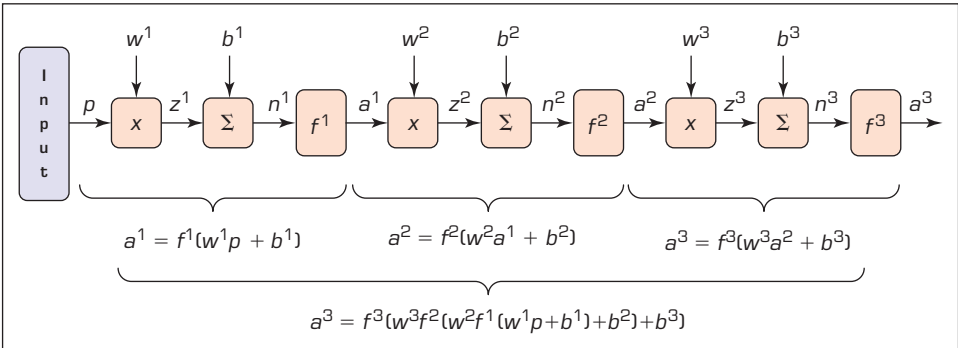


FIGURE 7.18 Vector Representation of the First Three Layers in a Typical MLP Network.

for lower levels of injury severity, but it was not among the significant factors for higher levels of injury severity, indicating that more serious injuries do not depend on the driver's gender. Another interesting and somewhat intuitive finding of the study indicated that age becomes an increasingly more significant factor as the level of injury severity increases, implying that older people are more likely to incur severe injuries (and fatalities) in serious vehicle crashes than younger people.

► SECTION 7.5 REVIEW QUESTIONS

1. What is the so-called black-box syndrome?
2. Why is it important to be able to explain an ANN's model structure?
3. How does sensitivity analysis work in ANN?
4. Search the Internet to find other methods to explain ANN methods. Report the results.

7.6 DEEP NEURAL NETWORKS

Until recently (before the advent of deep learning phenomenon), most neural network applications involved network architectures with only a few hidden layers and a limited number of neurons in each layer. Even in relatively complex business applications of neural networks, the number of neurons in networks hardly exceeded a few thousands. In fact, the processing capability of computers at the time was such a limiting factor that central processing units (CPU) were hardly able to run networks involving more than a couple of layers in a reasonable time. In recent years, development of graphics processing units (GPUs) along with the associated programming languages (e.g., CUDA by NVIDIA) that enable people to use them for data analysis purposes has led to more advanced applications of neural networks. GPU technology has enabled us to successfully run neural networks with over a million neurons. These larger networks are able to go deeper into the data features and extract more sophisticated patterns that could not be detected otherwise.

While deep networks can handle a considerably larger number of input variables, they also need relatively larger data sets to be trained satisfactorily; using small data sets for training deep networks typically leads to overfitting of the model to the training data and poor and unreliable results in case of applying to external data. Thanks to the Internet-and Internet of Things (IoT)-based data-capturing tools and technologies, larger data sets are now available in many application domains for deeper neural network training.

The input to a regular ANN model is typically an array of size $R \times 1$, where R is the number of input variables. In the deep networks, however, we are able to use *tensors* (i.e., N -dimensional arrays) as input. For example, in image recognition networks, each input (i.e., image) can be represented by a matrix indicating the color codes used in the image pixels; or for video processing purposes, each video can be represented by several matrices (i.e., a 3D tensor), each representing an image involved in the video. In other words, tensors provide us with the ability to include additional dimensions (e.g., time, location) in analyzing the data sets.

Except for these general differences, the different types of deep networks involve various modifications to the architecture of standard neural networks that equip them with distinct capabilities of dealing with particular data types for advanced purposes. In the following section, we discuss some of these special network types and their characteristics.

Feedforward Multilayer Perceptron (MLP)-Type Deep Networks

Feedforward **multilayer perceptron (MLP)** deep networks, also known as *deep feedforward networks*, are the most general type of deep networks. These networks are simply large-scale neural networks that can contain many layers of neurons and handle tensors as their input. The types and characteristics of the network elements (i.e., weight functions, transfer functions) are pretty much the same as in the standard ANN models. These models are called *feedforward* because the flow of information that goes through them is always forwarding and no feedback connections (i.e., connections in which outputs of a model are fed back to itself) are allowed. The neural networks in which feedback connections are allowed are called *recurrent neural networks (RNN)*. General RNN architectures, as well as a specific variation of RNNs called **long short-term memory (LSTM) networks**, are discussed in later sections of this chapter.

Generally, a sequential order of layers has to be held between the input and the output layers in the MLP-type network architecture. This means that the input vector has to pass through all layers sequentially and cannot skip any of them; moreover, it cannot be directly connected to any layer except for the very first one; the output of each layer is the input to the subsequent layer. Figure 7.18 demonstrates a vector representation of the first three layers of a typical MLP network. As shown, there is only one vector going into each layer, which is either the original input vector (p for the first layer) or the output vector from the previous hidden layer in the network architecture (a^{i-1} for the i^{th} layer). There are, however, some special variations of MLP network architectures designed for specialized purposes in which these principles can be violated.

Impact of Random Weights in Deep MLP

Optimization of the performance (loss) function in many real applications of deep MLPs is a challenging issue. The problem is that applying the common gradient-based training algorithms with random initialization of weights and biases that is very efficient for finding the optimal set of parameters in shallow neural networks most of the time could lead to getting stuck in the locally optimal solutions rather than catching the global optimum values for the parameters. As the depth of network increases, chances of reaching a global optimum using random initializations with the gradient-based algorithms decrease. In such cases, usually pretraining the network parameters using some *unsupervised* deep learning methods such as **deep belief networks (DBNs)** can be helpful (Hinton, Osindero, and Teh, 2006). DBNs are a type of a large class of deep neural networks called *generative models*. Introduction of DBNs in 2006 is considered as the beginning of the current deep learning renaissance (Goodfellow et al., 2016), since prior to that, deep models were considered too difficult to optimize. In fact, the primary application of DBNs today is to improve classification models by pretraining of their parameters.

Using these unsupervised learning methods, we can train the MLP layers, one at a time, starting from the first layer, and use the output of each layer as the input to the subsequent layer and initialize that layer with an unsupervised learning algorithm. At the end, we will have a set of initialized values for the parameters across the whole network. Those pretrained parameters, instead of random initialized parameters, then can be used as the initial values in the supervised learning of the MLP. This pretraining procedure has been shown to cause significant improvements to the deep classification applications. Figure 7.19 illustrates the classification errors that resulted from training a deep MLP network with (blue circles) and without (black triangles) pretraining of parameters (Bengio, 2009). In this example, the blue line represents the observed error rates of testing a classification model (on 1000 heldout examples) trained using a purely supervised approach with 10 million examples, whereas the black line indicates the error rates on the same

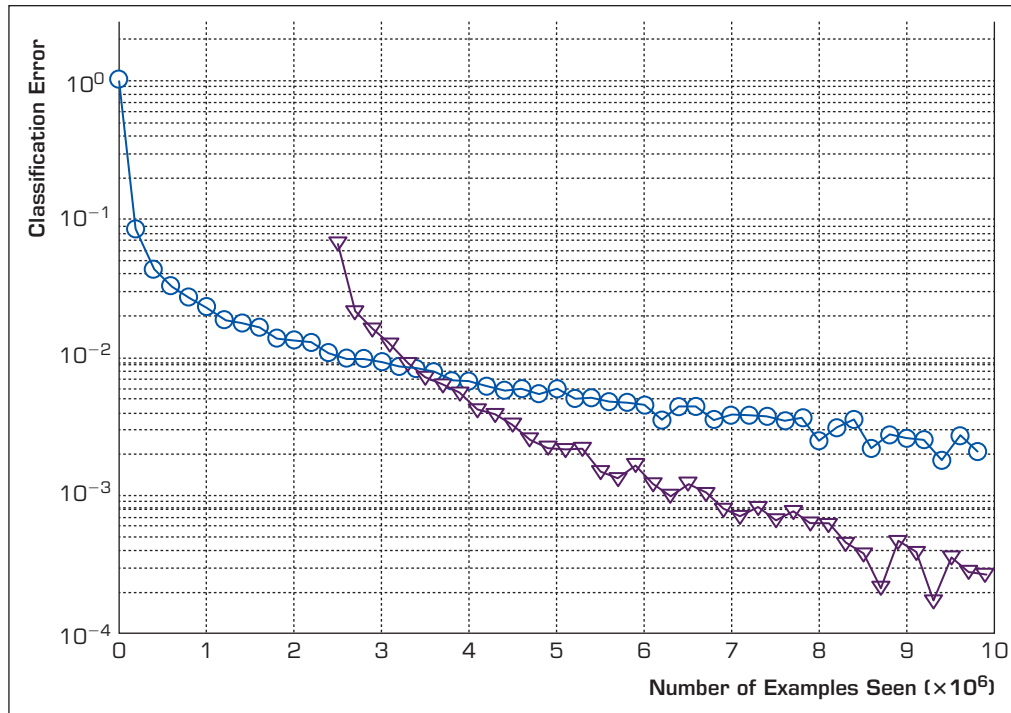


FIGURE 7.19 The Effect of Pretraining Network Parameters on Improving Results of a Classification-Type Deep Neural Network.

testing data set when 2.5 million examples were initially used for unsupervised training of network parameters (using DBN) and then the other 7.5 million examples along with the initialized parameters were used to train a supervised classification model. The diagrams clearly show a significant improvement in terms of the classification error rate in the model pretrained by a deep belief network.

More Hidden Layers Versus More Neurons?

An important question regarding the deep MLP models is “Would it make sense (and produce better results) to restructure such networks with only a few layers, but many neurons in each?” In other words, the question is why do we need deep MLP networks with many layers when we can include the same number of neurons in just a few layers (i.e., wide networks instead of deep networks). According to the universal approximation theorem (Cybenko, 1989; Hornik, 1991), a sufficiently large single-layer MLP network will be able to approximate any function. Although theoretically founded, such a layer with many neurons may be prohibitively large and hence may fail to learn the underlying patterns correctly. A deeper network can reduce the number of neurons required at each layer and hence decrease the generalization error. Whereas theoretically it is still an open research question, practically using more layers in a network seems to be more effective and computationally more efficient than using many neurons in a few layers.

Like typical artificial neural networks, multilayer perceptron networks can also be used for various prediction, classification, and clustering purposes. Especially when a large number of input variables are involved or in cases that the nature of input has to be an N -dimensional array, a deep multilayer network design needs to be employed.

Analytics in Action 7.3 provides an excellent case for the use of advanced analytics to better manage traffic flows in crowded cities.

ANALYTICS IN ACTION 7.3 Georgia DOT Variable Speed Limit Analytics Help Solve Traffic Congestions

The Background

When the Georgia Department of Transportation (GDOT) wanted to optimize the use of Big Data and advanced analytics to gain insight into transportation, it worked with Teradata to develop a proof of concept evaluation of GDOT's variable speed limit (VSL) pilot project.

The VSL concept has been adopted in many parts of the world, but it is still relatively new in the United States. As GDOT explains,

VSL are speed limits that change based on road, traffic, and weather conditions. Electronic signs slow down traffic ahead of congestion or bad weather to smooth out flow, diminish stop-and-go conditions, and reduce crashes. This low-cost, cutting edge technology alerts drivers in real time to speed changes due to conditions down the road. More consistent speeds improve safety by helping to prevent rear-end and lane changing collisions due to sudden stops.

Quantifying the customer service, safety, and efficiency benefits of VSL is extremely important to GDOT. This fits within a wider need to understand the effects of investments in intelligent transportation systems as well as other transportation systems and infrastructures.

VSL Pilot Project on I-285 in Atlanta

GDOT conducted a VSL pilot project on the northern half, or “top end,” of I-285 that encircles Atlanta. This 36-mile stretch of highway was equipped with 88 electronic speed limit signs that adjusted speed limits in 10 mph increments from 65 miles per hour (mph) to the minimum of 35 mph. The objectives were twofold:

1. Analyze speeds on the highway before versus after implementation of VSL.
2. Measure the impact of VSL on driving conditions.

To obtain an initial view of the traffic, the Teradata data science solution identified the locations and durations of “persistent slowdowns.” If highway speeds are above “reference speed,” then traffic is considered freely flowing. Falling below the reference speed at any point on the highway is considered a slowdown. When slowdowns persist across multiple consecutive minutes, a persistent slowdown can be defined.

By creating an analytic definition of slowdowns, it is possible to convert voluminous and highly variable speed data into patterns to support closer investigation. The early analyses of the data revealed that the clockwise and counterclockwise directions of the same highway may show significantly different frequency and duration of slowdowns. To better understand how slowdowns affect highway traffic, it is useful to take our new definition and zoom in on a specific situation. Figure 7.20 shows a specific but typical Atlanta afternoon on I-285, at a section of highway where traffic is moving clockwise, from west to east, between mile markers MM10 in the west to the east end at MM46.

The first significant slowdown occurred at 3:00 P.M. near MM32. The size of the circles represents duration (measured in minutes). The slowdown at MM32 was nearly four hours long. As the slowdown “persisted,” traffic speed diminished behind it. The slowdown formed on MM32 became a bottleneck that caused traffic behind it to slow down as well. The “comet trail” of backed-up traffic at the top left of Figure 7.21 illustrates the sequential formation of slowdowns at MM32 and then farther west, each starting later in the afternoon and not lasting as long.

Measuring Highway Speed Variability

The patterns of slowdowns on the highway as well as their different timings and locations led us to question their impact on drivers. If VSL could help drivers better anticipate the stop-and-go nature of the slowdowns, then being able to quantify the impact would be of interest to GDOT. GDOT was particularly concerned about what happens when a driver first encounters a slowdown. “While we do not know what causes the slowdown, we do know that drivers have made

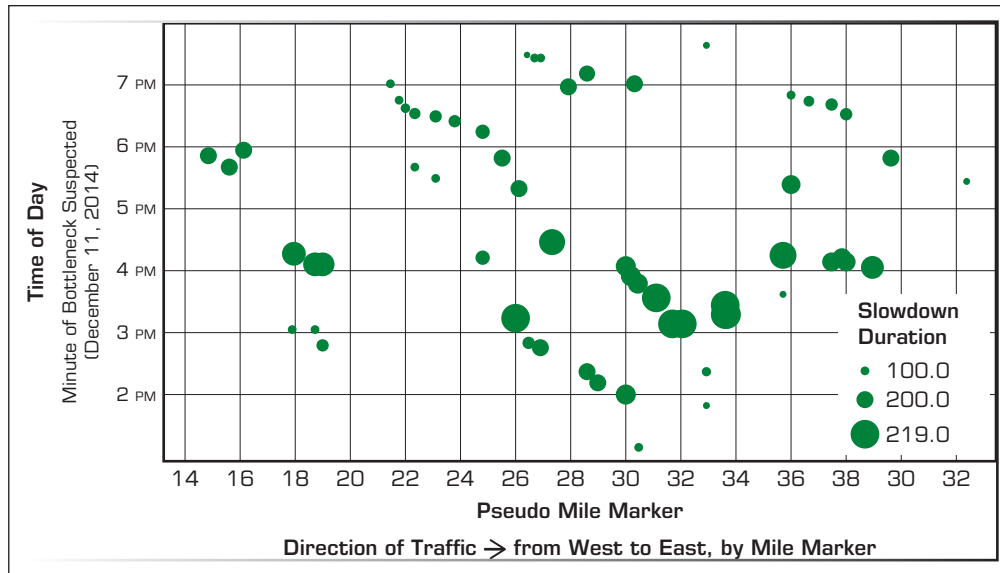


FIGURE 7.20 Traffic Moving Clockwise during the Afternoon.

speed adjustments. If the slowdown was caused by an accident, then the speed reduction could be quite sudden; alternatively, if the slowdown was just caused by growing volumes of traffic, then the speed reduction might be much more gradual.”

Identifying Bottlenecks and Traffic Turbulence

A bottleneck starts as a slowdown at a particular location. Something like a “pinch point” occurs on the highway. Then, over a period of time, traffic slows down behind the original pinch point. A bottleneck is a length of highway where traffic falls below 60 percent of reference speed and can stay at that level for miles. Figure 7.21 shows a conceptual representation of a bottleneck.

While bottlenecks are initiated by a pinch point, or slowdown, that forms the head of the queue, it is the end of the queue that is the most interesting. The area at the back of a queue is where traffic encounters a transition from free flow to slowly moving congested conditions. In the worst conditions, the end of the queue can experience a rapid transition. Drivers moving at

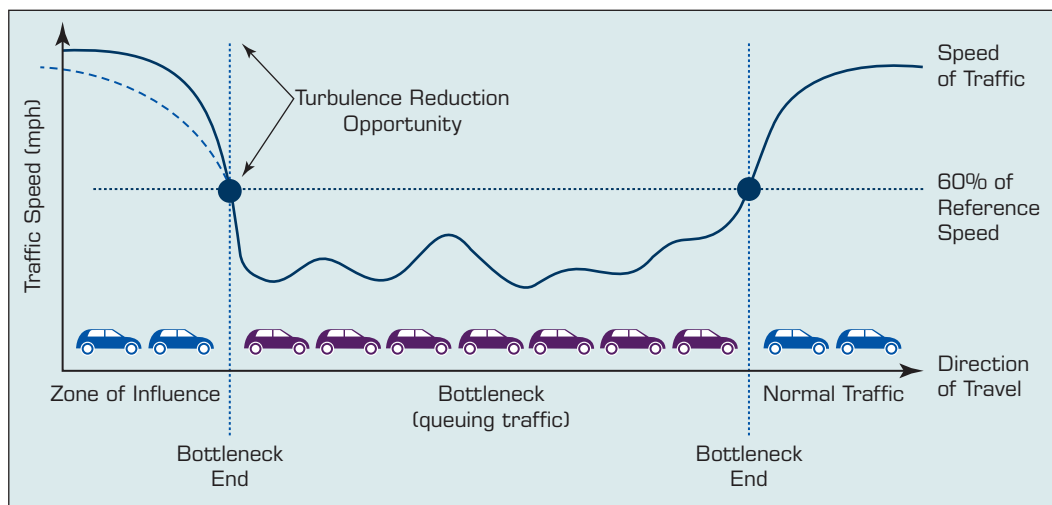


FIGURE 7.21 Graphical Depiction of a Bottleneck on a Highway.

highway speed may unexpectedly encounter slower traffic. This condition is ripe for accidents and is the place where VSL can deliver real value.

Powerful New Insight on Highway Congestion

The availability of new Big Data sources that describe the “ground truth” of traffic conditions on highways provides rich new opportunities for developing and analyzing highway performance metrics. Using just a single data source on detailed highway speeds, we produced two new and distinctive metrics using Teradata advanced data science capabilities.

First, by defining and measuring persistent slowdowns, we helped traffic engineers understand the frequency and duration of slow speed locations on a highway. The distinction of measuring a persistent slowdown versus a fleeting one is uniquely challenging and requires data science. It provides the ability to compare the number, duration, and location of slowdowns in a way that is more informative and compelling than simple averages, variances, and outliers in highway speeds.

The second metric was the ability to measure turbulence caused by bottlenecks. By identifying where bottlenecks occur and then narrowing in on their very critical zones of influence, we can make measurements of speeds and traffic deceleration turbulence within those zones. Data science and analytics capabilities demonstrated reduced turbulence when VSL is active in the critical zone of a bottleneck.

There is much more that could be explored within this context. For example, it is natural to assume that because most traffic is on the road during rush hours, VSL provides the most benefits during these high-traffic periods. However, the opposite may be true, which could provide a very important benefit of the VSL program.

Although this project was small in size and was just a proof of concept, a combination of similar projects beyond just transportation under the name of “smart cities” is underway around the United States and abroad. The goal is to use a variety of data from sensors to multimedia, rare event reports to satellite images along with advanced analytics that include deep learning and cognitive computing to transform the dynamic nature of cities toward better to best for all stakeholders.

Source: Teradata Case Study. “Georgia DOT Variable Speed Limit Analytics Help Solve Traffic Congestion.” <https://preview.teradata.com/Resources/Case-Studies/Georgia-DOT-Variable-Speed-Limit-Analytics-Help-Solve-Traffic-Congestion> (accessed July 2022); “Georgia DOT Variable Speed Limits.” www.dot.ga.gov/DriveSmart/SafetyOperation/Pages/VSL.aspx (accessed August 2021). Used with permission from Teradata.

In the next section, we discuss a very popular variation of deep MLP architecture called **convolutional neural network (CNN)** specifically designed for computer vision applications (e.g., image recognition, handwritten text processing).

SECTION 7.6 REVIEW QUESTIONS

1. What is meant by “deep” in deep neural networks? Compare deep neural networks to shallow neural networks.
2. What is GPU? How does it relate to deep neural networks?
3. How does a feedforward multilayer perceptron-type deep network work?
4. Comment on the impact of random weights in developing deep MLP.
5. Which strategy is better: more hidden layers versus more neurons?

7.7 CONVOLUTIONAL NEURAL NETWORKS

CNNs (LeCun et al., 1989) are among the most popular types of deep learning methods. CNNs are in essence variations of the deep MLP architecture, initially designed for computer vision applications (e.g., image processing, video processing, text recognition) but are also applicable to nonimage data sets.

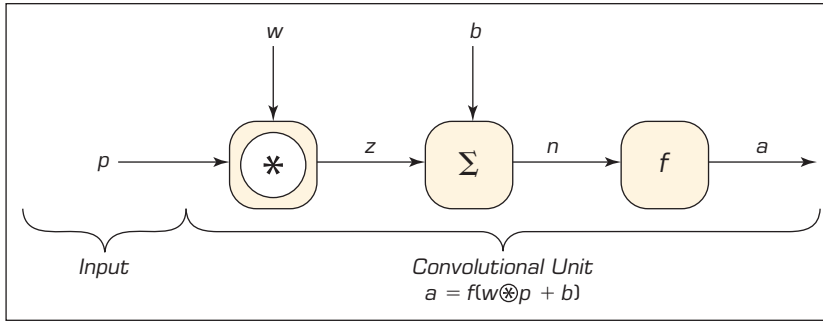


FIGURE 7.22 Typical Convolutional Network Unit.

The main characteristic of the convolutional networks is having at least one layer involving a *convolution weight function* instead of general matrix multiplication. Figure 7.22 illustrates a typical convolutional unit.

Convolution, typically shown by the $\otimes$ symbol, is a linear operation that essentially aims at extracting simple patterns from sophisticated data patterns. For instance, in processing an image containing several objects and colors, convolution functions can extract simple patterns like the existence of horizontal or vertical lines or edges in different parts of the picture. We discuss convolution functions in more detail in the next section.

A layer containing a **convolution function** in a CNN is called a *convolution layer*. This layer is often followed by a **pooling** (a.k.a. *subsampling*) layer. Pooling layers are in charge of consolidating the large tensors to one with a smaller size and reducing the number of model parameters while keeping their important features. Different types of pooling layers are also discussed in the following sections.

Convolution Function

In the description of MLP networks, it was said that the weight function is generally a matrix manipulation function that multiplies the weight vector into the input vector to produce the output vector in each layer. Having a very large input vector/tensor, which is the case in most deep learning applications, we need a large number of weight parameters so that each single input to each neuron could be assigned a single weight parameter. For instance, in an image-processing task using a neural network for images of size 150×150 pixels, each input matrix will contain 22,500 (i.e., 150 times 150) integers, each of which should be assigned its own weight parameter per each neuron it goes into throughout the network. Therefore, having even only a single layer requires thousands of weight parameters to be defined and trained. As one might guess, this fact would dramatically increase the required time and processing power to train a network, since in each training iteration, all of those weight parameters have to be updated by the SGD algorithm. The solution to this problem is the convolution function.

The convolution function can be thought of as a trick to address the issue defined in the previous paragraph. The trick is called *parameter sharing*, which in addition to computational efficiency provides additional benefits. Specifically, in a convolution layer, instead of having a weight for each input, there is a set of weights referred to as the *convolution kernel* or *filter*, which is shared between inputs and moves around the input matrix to produce the outputs. The kernel is typically represented as a small matrix of size $W_{r \times c}$; for a given input matrix V , then, the convolution function can be stated as:

$$Z_{i,j} = \sum_{k=1}^r \sum_{l=1}^c w_{k,l} v_{i+k-1, j+l-1}$$

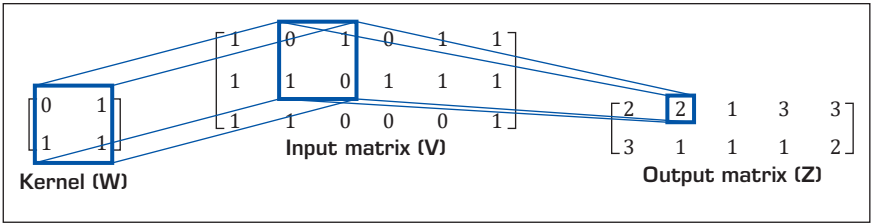


FIGURE 7.23 Convolution of a 2×2 Kernel by a 3×6 Input Matrix.

For example, assume that the input matrix to a layer and the convolution kernel is

$$V = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \quad W = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

Figure 7.23 illustrates how the convolution output can be computed. As shown, each element of the output matrix results from summing up the one-by-one point multiplications of the kernel elements into a corresponding $r \times c$ (in this example, 2×2 because the kernel is 2×2) subset of the input matrix elements. So, in the example shown, the element at the second column of the first row of the output matrix is in fact $0(0) + 1(1) + 1(1) + 1(0) = 2$.

It can be seen that the magnitude of each element in the output matrix directly depends on how the matched kernel (with the 2×2 matrix) and the input matrix are involved in calculation of that element. For example, the element at the fourth column of the first row of the output matrix is the result of convoluting the kernel by a part of the input matrix, which is exactly the same as the kernel. This suggests that by applying the convolution operation, we actually are converting the input matrix into an output in which the parts that have a particular feature (reflected by the kernel) are placed in the square box.

This characteristic of convolution functions is especially useful in practical image-processing applications. For instance, if the input matrix represents the pixels of an image, a particular kernel representing a specific shape (e.g., a diagonal line) may be convoluted into that image to extract parts of the image involving that specific shape. Figure 7.24, for example, shows the result of applying a 3×3 horizontal line kernel to a 15×15 image of a square. Clearly, the horizontal kernel produces an output in which the location of horizontal lines (as a feature) in the original input image is identified.

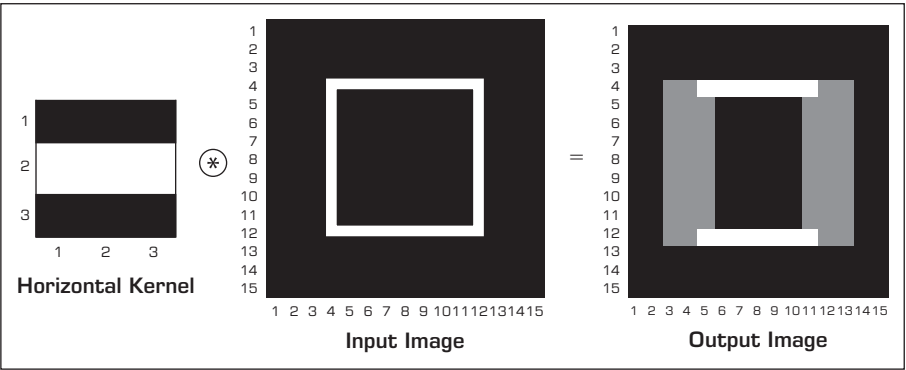


FIGURE 7.24 Example of Using Convolution for Extracting Features (Horizontal Lines in This Example) from Images.

Convolution using a kernel of size $r \times c$ will reduce the number of rows and columns in the output by $r - 1$ and $c - 1$, respectively. In the recent case, for example, using a 2×2 kernel for convolution, the output matrix has 1 row and 1 column less than the input matrix. To prevent this change of size, we can *pad* the outside of the input matrix with zeros before convolving, that is, to add $r - 1$ rows and $c - 1$ columns of zeros to the input matrix. On the other hand, if we want the output matrix to be even smaller, we can have the kernel to take larger *strides*, or kernel movements. Normally, the kernel is moved one step at a time (i.e., $\text{stride} = 1$) when performing the convolution. By increasing this stride to 2, the size of the output matrix is reduced by a factor of 2.

Although the main benefit of employing convolution in the deep networks is parameter sharing, which effectively reduces the required time and processing power to train the network by reducing the number of weight parameters, it involves some other benefits as well. A convolution layer in a network will have a property called *equivariance* for translation purposes (Goodfellow et al., 2016). It simply means that any changes in the input will lead to a change in the output in the same way. For instance, moving an object in the input image by 10 pixels in a particular direction will lead to moving its representation in the output image by 10 pixels in the same direction. Apart from image-processing applications, this feature is especially useful for analyzing time-series data using convolutional networks where convolution can produce a kind of timeline that shows when each feature appears in the input.

It should be noted that in almost all of the practical applications of convolutional networks, many convolution operations are used in parallel to extract various kinds of features from the data, because a single feature is hardly enough to fully describe the inputs for the classification or recognition purposes. Also, as noted before, in most real-world applications, we have to represent the inputs as multi-dimensional tensors. For instance, in the processing of color images as opposed to gray scale pictures, instead of having 2D tensors (i.e., matrices) that represent the color of pixels (i.e., black or white), one will have to use 3D tensors because each pixel should be defined using the intensity of red, blue, and green colors.

Pooling

Most of the times, a convolution layer is followed by another layer known as the *pooling* (a.k.a. *subsampling*) layer. The purpose of a pooling layer is to consolidate elements in the input matrix to produce a smaller output matrix while maintaining the important features. Normally, a pooling function involves an $r \times c$ consolidation window (similar to a kernel in the convolution function) that moves around the input matrix and in each move calculates some summary statistics of the elements involved in the consolidation window so that it can be put in the output image. For example, a particular type of pooling function called *average pooling* takes the average of the input matrix elements involved in the consolidation window and puts that average value as an element of the output matrix in the corresponding location. Similarly, the *max pooling* function (Zhou et al.) takes the maximum of the values in the window as the output element. Unlike convolution, for the pooling function, given the size of the consolidation window (i.e., r and c), stride should be carefully selected so that there would be no overlaps in the consolidations. The pooling operation using an $r \times c$ consolidation window reduces the number of rows and columns of the input matrix by a factor of r and c , respectively. For example, using a 3×3 consolidation window, a 15×15 matrix will be consolidated to a 5×5 matrix.

Pooling, in addition to reducing the number of parameters, is especially useful in the image-processing applications of deep learning in which the critical task is to determine whether a feature (e.g., a particular animal) is present in an image while the exact spatial location of the same in the picture is not important. However, if the location of

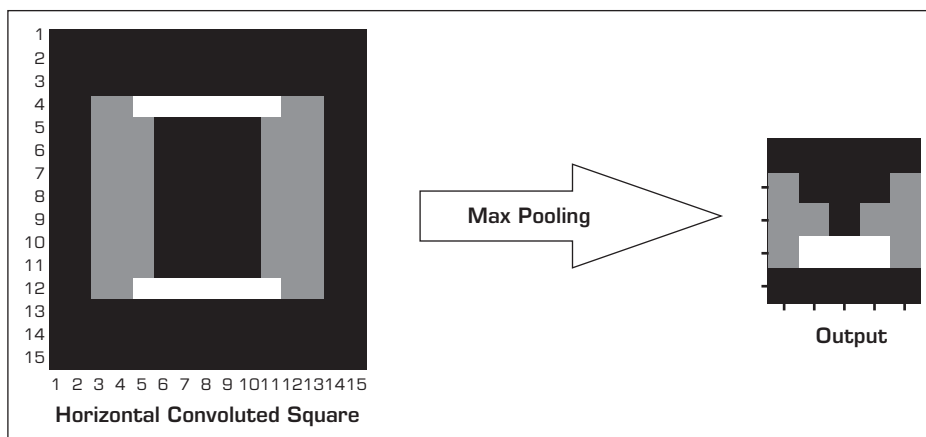


FIGURE 7.25 An Example of Applying Max Pooling on an Output Image to Reduce Its Size.

features is important in a particular context, applying a pooling function could potentially be misleading.

You can think of pooling as an operation that summarizes large inputs whose features are already extracted by the convolution layer and shows us just the important parts (i.e., features) in each small neighborhood in the input space. For instance, in the case of the image-processing example shown in Figure 7.24, if we place a max pooling layer after the convolution layer using a 3×3 consolidation window, the output will be like what is shown in Figure 7.25. As shown, the 15×15 already convoluted image is consolidated in a 5×5 image while the main features (i.e., horizontal lines) are maintained therein.

Sometimes pooling is used just to modify the size of matrices coming from the previous layer and convert them to a specified size required by the following layer in the network.

There are various types of pooling operations such as max pooling, average pooling, the L^2 norm of a rectangular neighborhood, and weighted average pooling. The choice of proper pooling operation as well as the decision to include a pooling layer in the network at all depends highly on the context and properties of the problem that the network is solving. There are some guidelines in the literature to help the network designers in making such decisions (Boureau et al., 2011; Boureau, Ponce, and LeCun, 2010; Scherer, Müller, and Behnke, 2010).

Image Processing Using Convolutional Networks

Real applications of deep learning in general and CNNs in particular highly depend on the availability of large, annotated data sets. Theoretically, CNNs can be applied to many practical problems, and today there are many large and feature-rich databases for such applications available. Nevertheless, the biggest challenge is that in supervised learning applications, one needs an already annotated (i.e., labeled) data set to train the model before we can use it for prediction/identification of other unknown cases. Whereas extracting features of data sets using CNN layers is an unsupervised task, the extracted features will not be of much use without having labeled cases to develop a classification network in a supervised learning fashion. That is why image classification networks traditionally involve two pipelines: visual feature extraction and image classification.

ImageNet (<http://www.image-net.org>) is an ongoing research project that provides researchers with a large database of images, each linked to a set of synonym words (known as *synset*) from WordNet (a word hierarchy database). Each *synset* represents

a particular concept in the WordNet. Currently, WordNet includes more than 100,000 synsets, each of which is supposed to be illustrated by an average of 1000 images in the ImageNet. ImageNet is a huge database for developing image processing-type deep networks. It contains more than 15 million labeled images in 22,000 categories. Because of its sheer size and proper categorization, ImageNet is by far the most widely used benchmarking data set to assess the efficiency and accuracy of deep networks designed by deep learning researchers.

One of the first convolutional networks designed for image classification using the ImageNet data set was AlexNet (Krizhevsky, Sutskever, and Hinton, 2012). It was composed of five convolution layers followed by three fully connected (a.k.a. dense) layers (see Figure 7.26 for a schematic representation of AlexNet). One of the contributions of this relatively simple architecture that made its training remarkably faster and computationally efficient was the use of rectified linear unit (ReLU) transfer functions in the convolution layers instead of the traditional sigmoid functions. By doing so, the designers addressed the issue called the *vanishing gradient problem* caused by very small derivatives of sigmoid functions in some regions of the images. The other important contribution of this network that has a dramatic role in improving the efficiency of deep networks was the introduction of the concept of dropout layers to the CNNs as a regularization technique to reduce overfitting. A dropout layer typically comes after the fully connected layers and applies a random probability to the neurons to switch off some of them and make the network sparser.

In the recent years, in addition to a large number of data scientists who showcase their deep learning capabilities, a number of well-known industry-leading companies such as Microsoft, Google, and Facebook have participated in the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The goal in the ILSVRC classification task is to design and train networks that are capable of classifying 1.2 million input images into one of the 1000 image categories. For instance, **GoogLeNet** (a.k.a. Inception), a deep convolutional network architecture designed by Google researchers, was the winning architecture of ILSVRC 2014 with a 22-layer network and only a 6.66 percent classification error rate, only slightly (5.1%) worse than the human-level classification error (Russakovsky et al., 2015). The main contribution of the GoogLeNet architecture was to introduce a module called *Inception*. The idea of Inception is that because one would have no idea of the size of convolution kernel that would perform best on a particular data set, it is better to include multiple convolutions and let the network decide which one to use. Therefore, as shown in Figure 7.27, in each convolution layer, the data

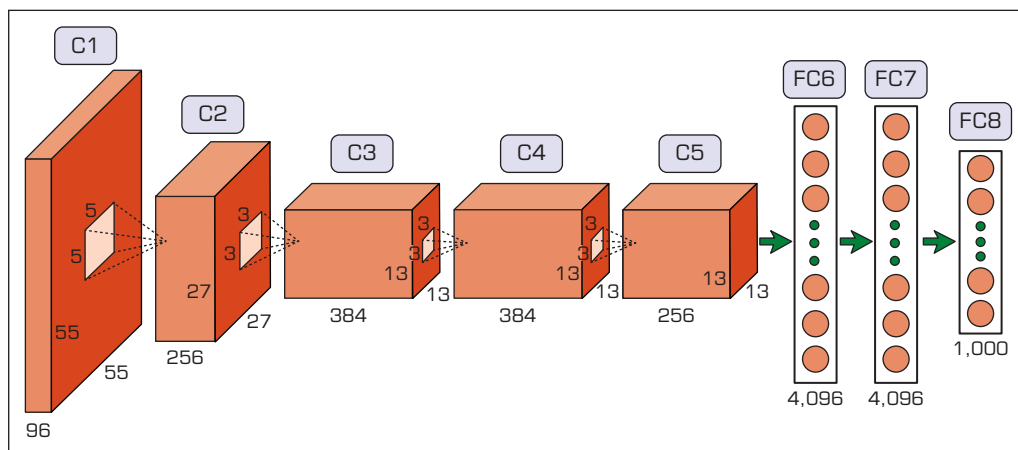


FIGURE 7.26 Architecture of AlexNet, a Convolutional Network for Image Classification.

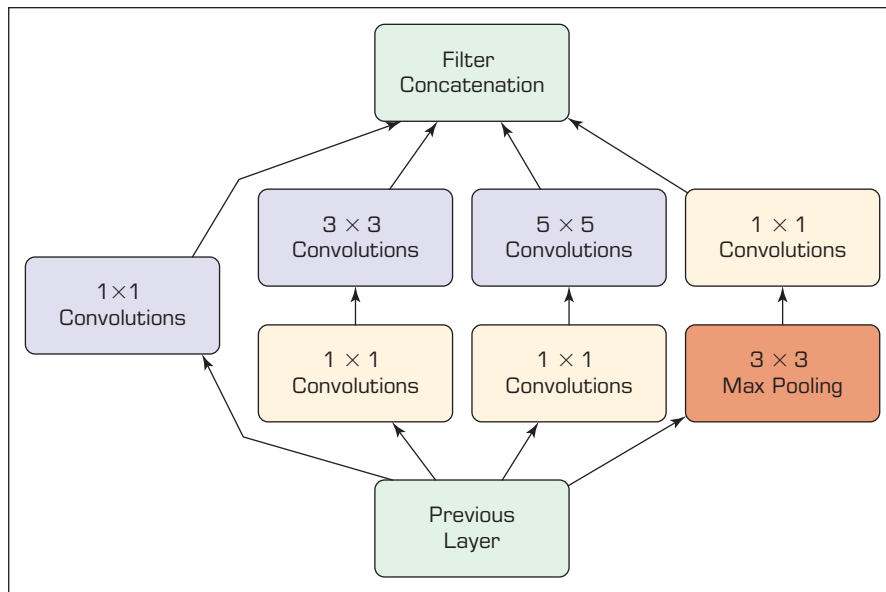


FIGURE 7.27 Conceptual Representation of the Inception Feature in GoogLeNet.

coming from the previous layer is passed through multiple types of convolution and the outputs are concatenated before going to the next layer. Such architecture allows the model to take into account both local features via smaller convolutions and high abstracted features via larger ones.

Google recently launched a new service, **Google Lens**, that uses deep learning artificial neural network algorithms (along with other AI techniques) to deliver information about the images captured by users from their nearby objects. This involves identifying the objects, products, plants, animals, and locations and providing information about them on the Internet. Some other features of this service are the capability of saving contact information from a business card image on the phone, identifying type of plants and breed of animals, identifying books and movies from their cover photos, and providing information (e.g., stores, theaters, shopping, reservations) about them. Figure 7.28 shows two examples of using the Google Lens app on an Android mobile device.

Even though later more accurate networks have been developed (e.g., He, Zhang, Ren, & Sun, 2015) in terms of efficiency and processing requirements (i.e., smaller number of layers and parameters), GoogLeNet is considered to be one of the best architectures to date. Apart from AlexNet and GoogLeNet, several other convolutional network architectures such as Residual Networks (ResNet), VGGNet, and Xception have been developed and contributed to the image-processing area, all relying on the ImageNet database.

In a May 2018 effort to address the labor-intensive task of labeling images on a large scale, Facebook published a *weakly supervised training* image recognition deep learning project (Mahajan et al., 2018). This project used hashtags made by the users on the images posted on Instagram as labels and trained a deep learning image recognition model based on that. The model was trained using 3.5 billion Instagram images labeled with around 17,000 hashtags using 336 GPUs working in parallel; the training procedure took a few weeks to be accomplished. A preliminary version of the model (trained using only 1 billion images and 1500 hashtags) was then tested on the ImageNet benchmark data set and is reported to have outperformed the state-of-the-art models in terms of accuracy by more than 2 percent. This big achievement by Facebook surely will open doors to a new world of image processing using deep learning since it can dramatically increase the size of available image data sets that are labeled for training purposes.

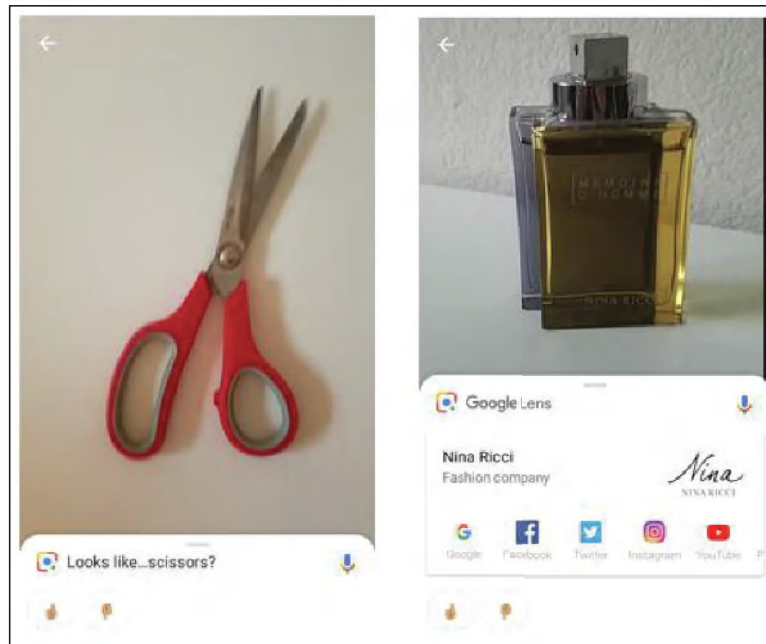


FIGURE 7.28 Two Examples of Using the Google Lens, a Service Based on Convolutional Deep Networks for Image Recognition. Source: ©2018 Google LLC, used with permission. Google and the Google logo are registered trademarks of Google LLC.

Use of deep learning and advanced analytics methods to classify images has evolved into the recognition of human faces and has become a very popular application for a variety of purposes. It is discussed in Analytics in Action 7.4.

ANALYTICS IN ACTION 7.4 From Image Recognition to Face Recognition

Face recognition, although seemingly similar to image recognition, is a much more complicated undertaking. The goal of face recognition is to identify the individual as opposed to the class it belongs to (human), and this identification task needs to be performed on a nonstatic (i.e., moving person) 3D environment. Face recognition has been an active research field in AI for many decades with limited success until recently. Thanks to the new generation of algorithms (i.e., deep learning) coupled with large data sets and computational power, face recognition technology is starting to make a significant impact on real-world applications. From security to marketing, face recognition and the variety of applications/use cases of this technology are increasing at an astounding pace.

Some of the premier examples of face recognition (both in advancements in technology and in the creative use of the technology perspectives) come from China. Today in China, face recognition is a very hot topic both from business development and from application development perspectives. Face recognition has become a fruitful ecosystem with hundreds of start-ups in China. In personal and/or business settings, people in China are widely using and relying on devices whose security is based on automatic recognition of their faces.

As perhaps the largest scale practical application case of deep learning and face recognition in the world today, the Chinese government recently started a project known as “Sharp Eyes” that aims at establishing a nationwide surveillance system based on face recognition. The project plans to integrate security cameras already installed in public places with private cameras on buildings and to utilize AI and deep learning to analyze the videos from those cameras. With millions of cameras and billions of lines of code, China is building a high-tech authoritarian future. With this system, cameras in some cities can scan train and bus stations as well as airports

to identify and catch China's most wanted suspected criminals. Billboard-size displays can show the faces of jaywalkers and list the names and pictures of people who do not pay their debts. Facial recognition scanners guard the entrances to housing complexes.

An interesting example of this surveillance system is the “shame game” (Mozur, 2018). An intersection south of Changhong Bridge in the city of Xiangyang previously was a nightmare. Cars drove fast, and jaywalkers darted into the street. Then, in the summer of 2017, the police put up cameras linked to facial recognition technology and a big outdoor screen. Photos of law-breakers were displayed alongside their names and government identification numbers. People were initially excited to see their faces on the screen until propaganda outlets told them that this was a form of punishment. Using this, citizens not only became a subject of this shame game but also were assigned negative citizenship points. Conversely, on the positive side, if people are caught on camera showing good behavior, like picking up a piece of trash from the road and putting it into a trash can or helping an elderly person cross an intersection, they get positive citizenship points that can be used for a variety of small awards.

China already has an estimated 200 million surveillance cameras—four times as many as the United States. The system is mainly intended to be used for tracking suspects, spotting suspicious behavior, and predicting crimes. For instance, to find a criminal, the image of a suspect can be uploaded to the system, matching it against millions of faces recognized from videos of millions of active security cameras across the country. This can find individuals with a high degree of similarity. The system also is merged with a huge database of information on medical records, travel bookings, online purchases, and even social media activities of every citizen and can monitor practically everyone in the country (with 1.4 billion people), tracking where they are and what they are doing each moment (Denyer, 2018). Going beyond narrowly defined security purposes, the government expects Sharp Eyes to ultimately assign every individual in the country a “social credit score” that specifies to what extent an individual is trustworthy.

While such an unrestricted application of deep learning (i.e., spying on citizens) is against the privacy and ethical norms and regulations of many western countries, including the United States, it is becoming a common practice in countries with less restrictive privacy laws and concerns as in China. Even western countries have begun to plan on employing similar technologies in limited scales only for security and crime prevention purposes. The FBI's Next Generation Identification System, for instance, is a lawful application of facial recognition and deep learning that compares images from crime scenes with a national database of mug shots to identify potential suspects.

Sources: Mozur, P. (2018, June 8). “Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras.” *The New York Times*. <https://www.nytimes.com/2018/07/08/business/china-surveillance-technology.html>; Denyer, S. (2018, January). “Beijing Bets on Facial Recognition in a Big Drive for Total Surveillance.” *The Washington Post*. https://www.washingtonpost.com/news/world/wp/2018/01/07/feature/in-china-facial-recognition-is-sharp-end-of-a-drive-for-total-surveillance/?noredirect=on&utm_term=.e73091681b31 (Accessed July 2022).

Text Processing Using Convolutional Networks

In addition to image processing, which was in fact the main reason for the popularity and development of convolutional networks, they have been shown to be useful in some large-scale text mining tasks as well. Especially since 2013, when Google published its **word2vec** project (Mikolov et al., 2013a; Mikolov et al., 2013b), the applications of deep learning for text mining have increased remarkably.

Word2vec is a two-layer neural network that gets a large text corpus as the input and converts each word in the corpus to a numeric vector of any given size (typically ranging from 100 to 1000) with very interesting features. Although word2vec itself is not a deep learning algorithm, its outputs (word vectors also known as **word embeddings**) already have been widely used in many deep learning research and commercial projects as inputs.

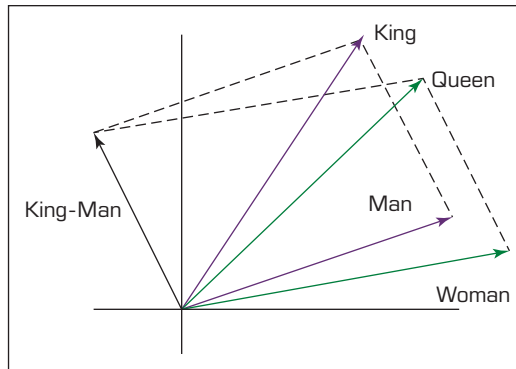


FIGURE 7.29 Typical Vector Representation of Word Embeddings in a Two-Dimensional Space.

One of the most interesting properties of word vectors created by the word2vec algorithm is maintaining the words' relative associations. For example, vector operations

$$\text{vector}(\text{'King'}) - \text{vector}(\text{'Man'}) + \text{vector}(\text{'Woman'})$$

and

$$\text{vector}(\text{'London'}) - \text{vector}(\text{'England'}) + \text{vector}(\text{'France'})$$

will result in a vector very close to $\text{vector}(\text{'Queen'})$ and $\text{vector}(\text{'Paris'})$, respectively. Figure 7.29 shows a simple vector representation of the first example in a two-dimensional vector space.

Moreover, the vectors are specified in such a way that those of a similar context are placed very close to each other in the n -dimensional vector space. For instance, in the word2vec model pretrained by Google using a corpus including about 100 billion words (taken from Google News), the closest vectors to the $\text{vector}(\text{'Sweden'})$ in terms of cosine distance, as shown in Table 7.2, identify European country names near the Scandinavian region, the same region in which Sweden is located.

Additionally, since word2vec takes into account the contexts in which a word has been used and the frequency of using it in each context in guessing the meaning of the word, it enables us to represent each term with its semantic context instead of just the syntactic/symbolic term itself. As a result, word2vec addresses several word variation

TABLE 7.2 Example of the word2vec Project Indicating the Closest Word Vectors to the Word "Sweden"

Word	Cosine Distance
Norway	0.760124
Denmark	0.715460
Finland	0.620022
Switzerland	0.588132
Belgium	0.585635
Netherlands	0.574631
Iceland	0.562368
Estonia	0.547621
Slovenia	0.531408

issues that used to be problematic in traditional text mining activities. In other words, word2vec is able to handle and correctly represent words including typos, abbreviations, and informal conversations. For instance, the words *Frnce*, *Franse*, and *Frans* would all get roughly the same word embeddings as their original counterpart *France*. Word embeddings are also able to determine other interesting types of associations such as distinction of entities (e.g., $\text{vector}[\text{'human'}] - \text{vector}[\text{'animal'}] \sim \text{vector}[\text{'ethics'}]$) or geopolitical associations (e.g., $\text{vector}[\text{'Iraq'}] - \text{vector}[\text{'violence'}] \sim \text{vector}[\text{'Jordan'}]$).

By providing such a meaningful representation of textual data, in recent years, word2vec has driven many deep learning-based text mining projects in a wide range of contexts (e.g., medical, computer science, social media, marketing), and various types of deep networks have been applied to the word embeddings created by this algorithm to accomplish different objectives. Particularly, a large group of studies had developed convolutional networks applied to the word embeddings with the aim of *relation extraction* from textual data sets. Relation extraction is one of the subtasks of natural language processing (NLP) that focuses on determining whether two or more named entities recognized in the text form specific relationships (e.g., “A *causes* B”; “B *is caused by* A”). For instance, Zeng et al. (2014) developed a deep convolutional network (see Figure 7.30) to classify relations between specified entities in sentences. To this end, these researchers used a matrix format to represent each sentence. Each column of the input matrices is in fact the word embedding (i.e., vector) associated with one of the words involved in the sentence. Zeng et al. then used a convolutional network, shown in the right box in Figure 7.30, to automatically learn the sentence-level features and concatenate those features (i.e., the output vector of the CNN) with some basic lexical features (e.g., the order of the two words of interest within the sentence and the left and right tokens for each of them). The concatenated feature vector then is fed into a classification layer with a *softmax* transfer function, which determines the type of relationship between the two words of interest among multiple predefined types. The softmax transfer function is the most common type of function to be used for classification layers, especially when the number of classes is more than two. For classification problems with only two outcome categories, log-sigmoid transfer functions are also very popular. The proposed approach by Zeng et al. was shown to correctly classify the relation between the marked terms in sentences of a sample data set with an 82.7 percent accuracy.

In a similar study, Nguyen and Grishman (2015) used a four-layer convolutional network with multiple kernel sizes in each convolution layer fed by the real-valued

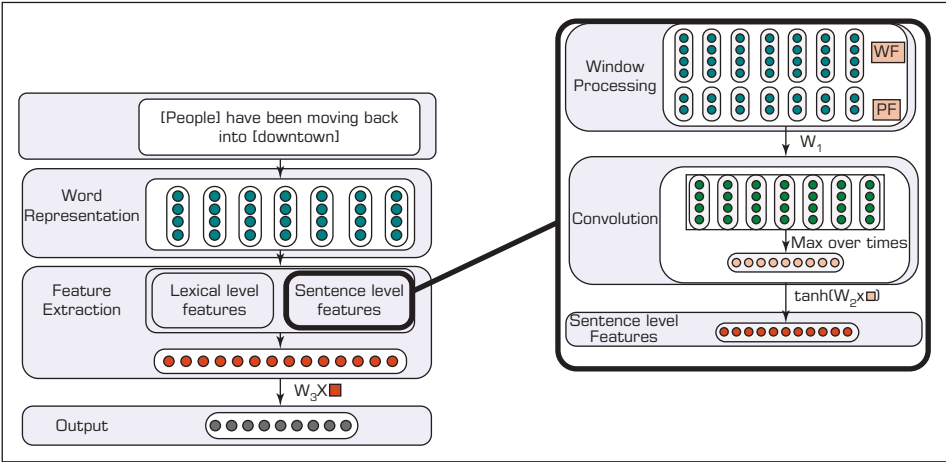


FIGURE 7.30 CNN Architecture for Relation Extraction Task in Text Mining.

vectors of words included in sentences to classify the type of relationship between the two marked words in each sentence. In the input matrix, each row was the word embedding associated with a word in the same sequence in the sentence as the row number. In addition, these researchers included two more columns to the input matrices to represent the relative position of each word (either positive or negative) with regard to each of the marked terms. The automatically extracted features then were passed through a classification layer with softmax function for the type of relationship to be determined. Nguyen and Grishman trained their model using 8000 annotated examples (with 19 predefined classes of relationships) and tested the trained model on a set of 2717 validation data sets and achieved a classification accuracy of 61.32 percent (i.e., more than 11 times better performance than guessing).

Such text mining approaches using convolutional deep networks can be extended to various practical contexts. Again, the big challenge here, just as in image processing, is lack of sufficient large annotated data sets for supervised training of deep networks. A *distant supervision* method of training has been proposed (Mintz et al., 2009) to address this challenge. It suggests that large amounts of training data can be produced by aligning knowledge base (KB) facts with texts. In fact, this approach is based on the assumption that if a particular type of relation exists between an entity pair (e.g., “A” *is a component of* “B”) in the KB, then every text document containing the mention of the entity pair would express that relation. However, since this assumption was not very realistic, Riedel, Yao, and McCallum (2010) later relaxed it by modeling the problem as a *multi-instance learning* problem. They suggest assigning labels to a bag of instances rather than a single instance that can reduce the noise of the distant supervision method and create more realistic labeled training data sets (Kumar, 2017).

► SECTION 7.7 REVIEW QUESTIONS

1. What is CNN?
2. For what type of applications can CNN be used?
3. What is convolution function in CNN and how does it work?
4. What is pooling in CNN? How does it work?
5. What is ImageNet and how does it relate to deep learning?
6. What is the significance of AlexNet? Draw and describe its architecture.
7. What is GoogLeNet? How does it work?
8. How does CNN process text? What are word embeddings, and how do they work?
9. What is word2vec, and what does it add to traditional text mining?

7.8 RECURRENT NETWORKS AND LONG SHORT-TERM MEMORY NETWORKS

Human thinking and understanding to a great extent relies on *context*. It is crucial for us, for example, to know that a particular speaker uses very sarcastic language (based on her previous speeches) to fully catch all the jokes that she makes. Or to understand the real meaning of the word *fall* (i.e., either *the season* or *to collapse*) in the sentence “It is a nice day of fall” without knowledge about the other words in the surrounding sentences would only be guessing, not necessarily understanding. Knowledge of context is typically formed based on observing events that happened in the past. In fact, human thoughts are persistent, and we use every piece of information we previously acquired about an event in the process of analyzing it rather than throwing away our past knowledge and thinking from scratch every time we face similar events or situations. Hence, there seems to be a recurrence in the way humans process information.

While deep MLP and convolutional networks are specialized for processing a *static* grid of values like an image or a matrix of word embeddings, sometimes the *sequence* of input values is also important to the operation of the network to accomplish a given task and hence should be taken into account. Another popular type of neural networks is **recurrent neural network (RNN)** (Rumelhart et al., 1986), which is specifically designed to process sequential inputs. An RNN basically models a *dynamic* system where (at least in one of its hidden neurons) the state of the system (i.e., output of a hidden neuron) at each time point t depends on both the inputs to the system at that time and its state at the previous time point $t - 1$. In other words, RNNs are the type of neural networks that have memory and that apply that memory to determine their future outputs. For instance, in designing a neural network to play chess, it is important to take into account several previous moves while training the network, because a wrong move by a player can lead to the eventual loss of the game in the subsequent 10–15 plays. Also, to understand the real meaning of a sentence in an essay, sometimes we need to rely on the information portrayed in the previous several sentences or paragraphs. That is, for a true understanding, we need the context built sequentially and collectively over time. Therefore, it is crucial to consider a memory element for the neural network that takes into account the effect of prior moves (in the chess example) and prior sentences and paragraphs (in the essay example) to determine the best output. This memory portrays and creates the context required for the learning and understanding.

In static networks like MLP-type CNNs, we are trying to find some functions (i.e., network weights and biases) that map the inputs to some outputs that are as close as possible to the actual target. In dynamic networks like RNNs, on the other hand, both inputs and outputs are sequences (patterns). Therefore, a dynamic network is a dynamic system rather than a function because its output depends not only on the input but also on the previous outputs. Most of the RNNs use the following general equation to define the values of their hidden units (Goodfellow et al., 2016).

$$a^{(t)} = f(a^{(t-1)}, p^{(t)}, \theta)$$

In this equation, $a^{(t)}$ represents the state of the system at time t , and $p^{(t)}$ and θ represent the input to the unit at time t and the parameters, respectively. Applying the same general equation for calculating the state of system at time $t - 1$, we will have:

$$a^{(t-1)} = f(a^{(t-2)}, p^{(t-1)}, \theta)$$

In other words:

$$a^{(t)} = f(fa^{(t-2)}, p^{(t-1)}, \theta, p^{(t)}, \theta)$$

And this equation can be extended multiple times for any given sequence length. Graphically, a recurrent unit in a network can be depicted in a circuit diagram like the one shown in Figure 7.31. In this figure, D represents the *tap delay lines*, or simply the *delay* element of the network that, at each time point t , contains $a^{(t)}$, the previous output value of the unit. Sometimes instead of just one value, we store several previous output values in D to account for the effect of all of them. Also *iw* and *lw* represent the weight vectors applied to the input and the delay, respectively.

Technically speaking, any network with feedback can actually be called a *deep network*, because even with a single layer, the loop created by the feedback can be thought of as a static MLP-type network with many layers (see Figure 7.32 for a graphical illustration of this structure). However, in practice, each recurrent neural network would involve dozens of layers, each with feedback to itself, or even to the previous layers, which makes a recurrent neural network even deeper and more complicated.

Because of the feedbacks, computation of gradients in the recurrent neural networks would be somewhat different from the general backpropagation algorithm used

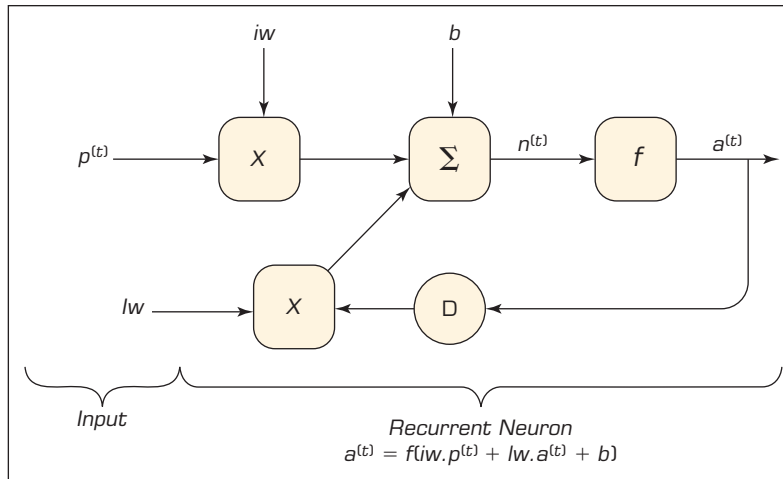


FIGURE 7.31 Typical Recurrent Unit.

for the static MLP networks. There are two alternative approaches for computing the gradients in the RNNs, namely real-time recurrent learning (RTRL) and backpropagation through time (BTT), whose explanation is beyond the scope of this chapter. Nevertheless, the general purpose remains the same; once the gradients have been computed, the same procedures are applied to optimize the learning of the network parameters.

The LSTM networks (Hochreiter & Schmidhuber, 1997) are variations of recurrent neural networks that today are known as the most effective sequence modeling technique and are the base of many practical applications. In a dynamic network, the weights are called the *long-term memory* while the feedbacks role is the *short-term memory*.

In essence, only the short-term memory (i.e., feedbacks; previous events) provides a network with the context. In a typical RNN, the information in the short-term memory is continuously replaced as new information is fed back into the network over time. That is why RNNs perform well when the gap between the relevant information and the place that is needed is small. For instance, for predicting the last word in the sentence “The referee blew her whistle,” we just need to know a few words back (i.e., the referee) to correctly predict. Since in this case the gap between the relevant information (i.e., the referee) and where it is needed (i.e., to predict whistle) is small, an RNN network can easily perform this learning and prediction task.

However, sometimes the relevant information required to perform a task is far away from where it is needed (i.e., the gap is large). Therefore, it is quite likely that it would have already been replaced by other information in the short-term memory by the time it is needed for the creation of the proper context. For instance, to predict the last word in “I went to a carwash yesterday. It cost \$5 to wash my *car*,” there is a relatively larger gap between the relevant information (i.e., carwash) and where it is

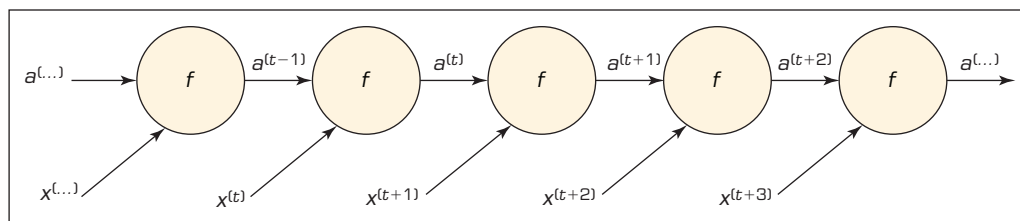


FIGURE 7.32 Unfolded View of a Typical Recurrent Network.

needed. Sometimes we may even need to refer to the previous paragraphs to reach the relevant information for predicting the true meaning of a word. In such cases, RNNs usually do not perform well since they cannot keep the information in their short-term memory for a long enough time. Fortunately, LSTM networks do not have such a short-coming. The term *long short-term memory network* then refers to a network in which we are trying to remember what happened in the past (i.e., feedbacks; previous outputs of the layers) for a long enough time so that it can be used/leveraged in accomplishing the task when needed.

From an architectural viewpoint, the memory concept (i.e., remembering “what happened in the past”) is incorporated in LSTM networks by incorporating four additional layers into the typical recurrent network architecture: three gate layers, namely *input gate*, *forget* (a.k.a. *feedback*) *gate*, and *output gate*, and an additional layer called **Constant Error Carousel (CEC)**, also known as *the state unit* that integrates those gates and interacts them with the other layers. Each gate is nothing but a layer with two inputs, one from the network input and the other a feedback from the final output of the whole network. The gates involve log-sigmoid transfer functions. Therefore, their outputs will be between 0 and 1 and describe how much of each component (either input, feedback, or output) should be let through the network. Also, CEC is a layer that falls between the input and the output layers in a recurrent network architecture and applies the gates outputs to make the short-term memory long.

To have a long short-term memory means that we want to keep the effect of previous outputs for a longer time. However, we typically do not want to indiscriminately remember everything that has happened in the past. Therefore, gating provides us with the capability of remembering prior outputs selectively. The input gate will allow selective inputs to the CEC; the forget gate will clear the CEC from the unwanted previous feedbacks; and the output gate will allow selective outputs from the CEC. Figure 7.33 shows a simple depiction of a typical LSTM architecture.

In summary, the gates in the LSTM are in charge of controlling the flow of information through the network and dynamically change the time scale of integration based on the input sequence. As a result, LSTM networks are able to learn long-term dependencies among the sequence of inputs more easily than the regular RNNs.

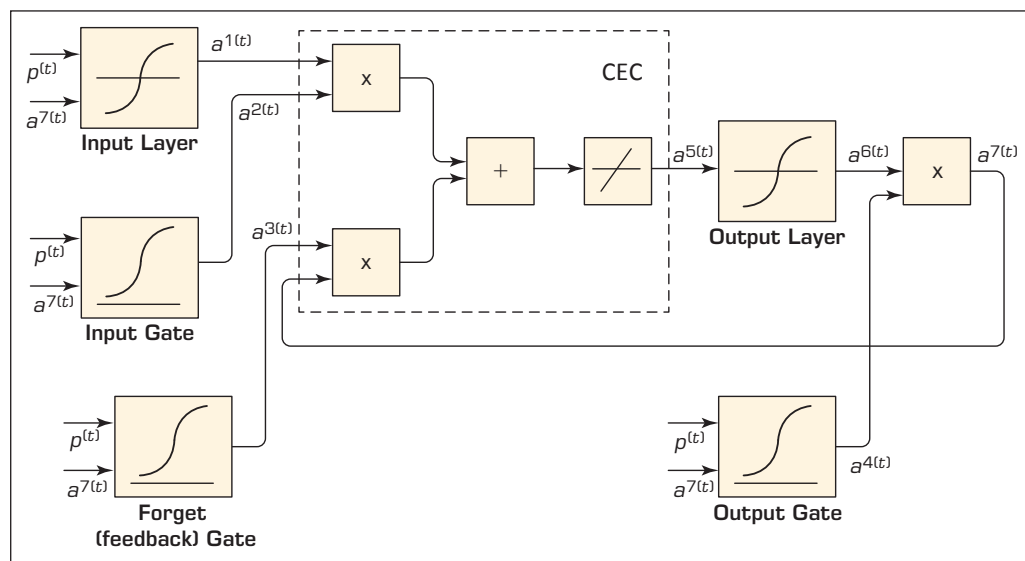


FIGURE 7.33 Typical Long Short-Term Memory (LSTM) Network Architecture.

LSTM Networks Applications

Since their emergence in the late 1990s (Hochreiter & Schmidhuber, 1997), LSTM networks have been widely used in many sequence modeling applications, including image captioning (i.e., automatically describing the content of images) (Vinyals, Toshev, Bengio, and Erhan, 2017, 2015; Xu et al., 2015), handwriting recognition and generation (Graves, 2013; Graves and Schmidhuber, 2009; Keyzers et al. 2017), parsing (Liang et al. 2016; Vinyals, Kaiser, et al., 2015), speech recognition (Graves and Jaitly, 2014; Graves, Jaitly, and Mohamed, 2013; Graves, Mohamed, and Hinton, 2013), and machine translation (Bahdanau, Cho, and Bengio, 2014; Sutskever, Vinyals, and Le, 2014).

Currently, we are surrounded by multiple deep learning solutions working on the basis of speech recognition, such as Apple's Siri, Google Now, Microsoft's Cortana, and Amazon's Alexa, several of which we deal with on a daily basis (e.g., checking on the weather, asking for a Web search, calling a friend, and asking for directions on the map). Note taking is not a difficult, frustrating task anymore since we can easily record a speech or lecture, upload the digital recording on one of the several cloud-based speech-to-text service providers' platforms, and download the transcript in a few seconds. The Google cloud-based speech-to-text service, for example, supports 120 languages and their variants and has the ability to convert speech to text either in real time or using recorded audios. The Google service automatically handles the noise in the audio; accurately punctuates the transcripts with commas, question marks, and periods; and can be customized by the user to a specific context by getting a set of terms and phrases that are very likely to be used in a speech and recognizing them appropriately.

Machine translation refers to a subfield of AI that employs computer programs to translate speech or text from one language to another. One of the most comprehensive machine translation systems is the **Google's Neural Machine Translation (GNMT)** platform. GNMT is basically an LSTM network with eight encoder and eight decoder layers designed by a group of Google researchers in 2016 (Wu et al., 2016). GNMT is specialized for translating whole sentences at a time as opposed to the former version of Google Translate platform, which was a phrase-based translator. This network is capable of naturally handling the translation of rare words (that previously was a challenge in machine translation) by dividing the words into a set of common subword units. GNMT currently supports automatic sentence translations between more than 100 languages. Figure 7.34 shows how a sample sentence was translated from French to English by GNMT and a human translator. It also indicates how closely the GNMT translations between different language pairs were ranked by the human speakers compared with translations made by humans.

Although machine translation has been revolutionized by the virtue of LSTMs, it encounters challenges that make it far from a fully automated high-quality translation. Like image-processing applications, there is a lack of sufficient training data (manually translated by humans) for many language pairs on which the network can be trained. As a result, translations between rare languages are usually done through a bridging language (mostly English) that may result in higher chances of error.

In 2014, Microsoft launched its **Skype Translator** service, a free voice translation service involving both speech recognition and machine translation with the ability of translating real-time conversations in 10 languages. Using this service, people speaking different languages can talk to each other in their own languages via a Skype voice or video call, and the system recognizes their voices and translates their every sentence through a translator bot in near real time for the other party. To provide more accurate translations, the deep networks used in the backend of this system were trained using conversational language (i.e., using materials such as translated Web pages, movie subtitles, and casual phrases taken from people's conversations in social networking Web sites) rather than the formal language commonly used in documents. The output of the speech recognition module of the system then goes through TrueText, a Microsoft technology for normalizing text that is capable of identifying mistakes and disfluencies (e.g., pauses during the speech or repeating some parts of speech, or adding fillers like "um" and "ah" when speaking)

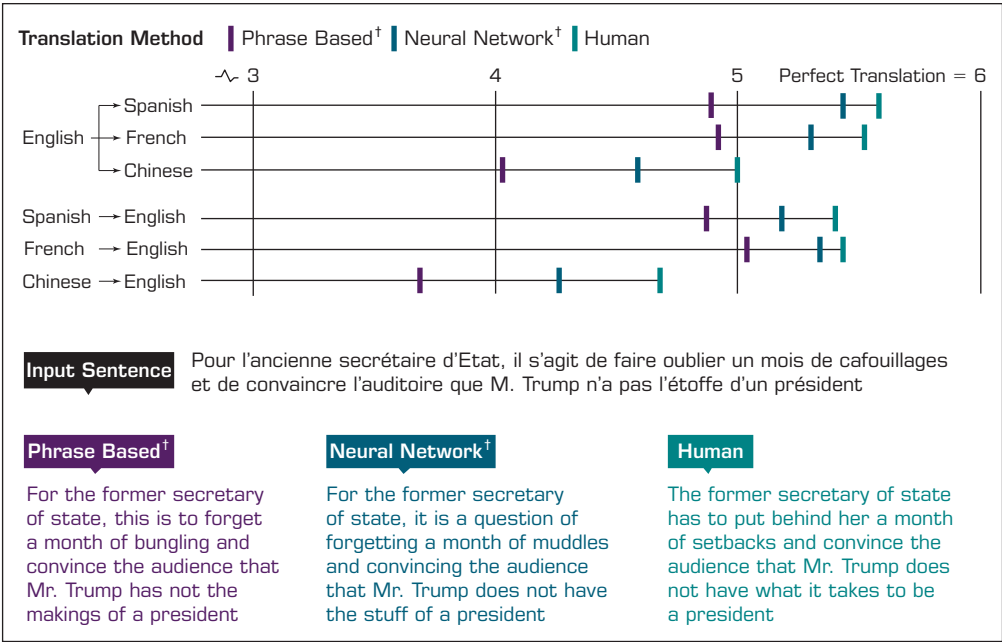


FIGURE 7.34 Example Indicating the Close-to-Human Performance of the Google Neural Machine Translator (GNMT).

that people commonly conduct in their conversations and account for them for making better translations. Figure 7.35 shows the four-step process involved in the Skype Translator by Microsoft, each of which relies on the LSTM type of deep neural networks.

CHATGPT

What Is ChatGPT and How Does It Work?

In November 2022, OpenAI¹ released their AI language model known as Chat-based Generative Pretrained Transformer (ChatGPT),² which rapidly garnered significant public attention and media coverage due to its impressive language processing abilities. These include question-answering, language translation, language generation, summarization, and even programming in various common computer languages.

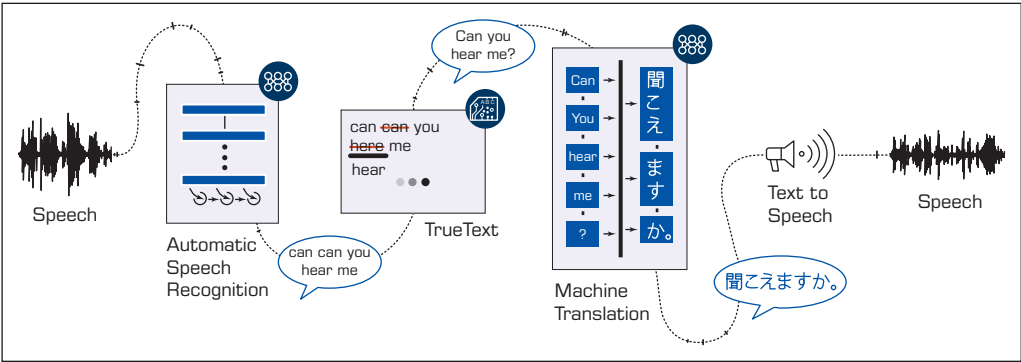


FIGURE 7.35 Four-Step Process of Translating Speech Using Deep Networks in the Microsoft Skype Translator.

¹ <https://openai.com/>
² <https://chat.openai.com/chat>

The development process of ChatGPT involved training it on an enormous corpus of text data from books, websites, articles, and a variety of other human language sources. The exact amount of text used for the initial training of ChatGPT (and its ancestors, GPT-1, GPT-2, and GPT-3) is not publicly disclosed, but it is estimated to be in the order of hundreds of billions of words.

ChatGPT falls in a class of machine learning NLP models known as Large Language Models (LLMs). These models operate by (1) being pretrained on huge corpora of text data using cutting-edge deep learning algorithms and (2) inferring a response for a given user input by breaking it down into its parts and identifying the key concepts and the context, analyzing those components using NLP techniques to understand the meaning, and ultimately generating a response (considering the meaning and context of the input) using deep learning-based text generation algorithms.

Technically speaking, ChatGPT is created and fine-tuned using a combination of supervised machine learning and reinforcement learning, while the latter plays a more important role in making it different from other LLMs. Like other LLMs, the core pretraining approaches for ChatGPT are Next Token Prediction and Masked Language Modeling. Through these approaches, the model learns the statistical structure of language to predict the next word in a sequence of words (you probably have seen this feature in the language models incorporated in smartphone keyboards) or a missing word from a sequence, based on the likelihood it estimates for each possible word in its vocabulary.

However, what makes ChatGPT unique and its generated responses more aligned with the language context than other LLMs is a special reinforcement learning approach called Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022). The original notion of reinforcement learning (RL), which is one of the three basic machine learning paradigms, is based on a reward system in which an agent perceives its environment through repetitive trials and errors and adjusts the model parameters by analyzing rewards and punishments it receives for its desired and undesired actions. The RLHF approach extends RL by developing a special reward policy in three steps based on inputs from human annotators (OpenAI, 2023) (see Figure 7.35a).

In the **first step**, known as *Supervised Fine-Tuning (SFT)*, human labelers are employed to generate written responses to prompts sampled from a prompt data set. For instance, the prompt could be “What is a smartphone?” to which a labeler might respond like “It is a pocket-size portable computer that can be used for making phone calls as well

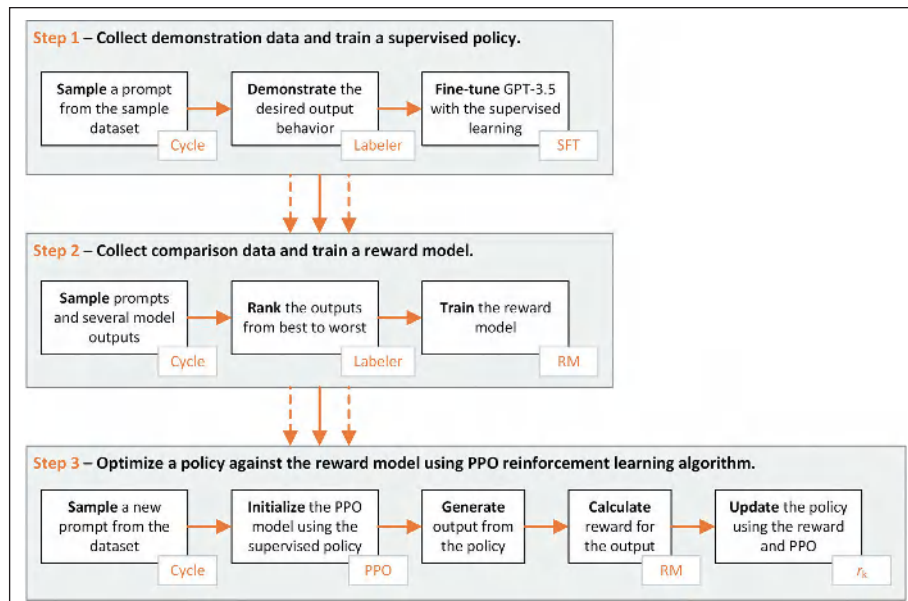


FIGURE 7.35a Three Steps of Reinforcement Learning from Human Feedback (RLHF).

as for *photography, navigation, and many other daily activities*.” These responses are then used to fine-tune the pretrained LLM by learning a baseline supervised policy that enables it to generate responses to prompts selected from a certain list. Once the SFT model is trained, it is used to generate multiple responses for each prompt in a limited list of prompts. These model-generated responses are the raw ingredients for the second step of RLHF.

The **second step** involves learning a *Reward Model (RM)*. To this end, the large number of responses generated by the fine-tuned SFT model in the previous step are voted by human labelers to determine a human preference score for each response. For example, assume that for the abovementioned prompt (i.e., smartphone), the SFT model generated the following answers:

- a) It is a portable device that can be used to make calls, take pictures, and do web browsing.
- b) A smartphone is a cell phone that has extra features such as Wi-Fi connectivity and a digital camera.
- c) It is a phone typically with a touch screen, a graphical user interface, and many useful applications.
- d) It is a minicomputer that comes in different sizes and prices that has a slot for a wireless SIM card.

Human labelers are then asked to rank answers generated for each prompt, leading to the creation of a new comparison data set with the rankings as the labels (e.g., A > B > D > C), which subsequently is used to train the reward model. In essence, the role of this reward model in ChatGPT is to mimic human preferences such that the model can choose the most human-favorable answer to a prompt among various possible answers it can generate.

The **third step** is where reinforcement learning comes into play to fine-tune the baseline SFT model (developed in step 1) by allowing it to optimize the Reward Model. The process is accomplished using a specific algorithm known as *Proximal Policy Optimization (PPO)*, which is designed to train agents in the reinforcement learning process.

Throughout the repetitive PPO process, first, an output is generated for a prompt by the baseline SFT policy, then it is sent through the trained Reward Model and a reward is determined for that (representing the extent to which humans prefer that output), which is fed back into the SFT model to update the policy (see step 3 in Figure 7.35a).

Applications of ChatGPT

A major concern raised among practitioners from different areas since the emergence of ChatGPT was whether it is going to replace them and take their jobs. While that might be a valid concern to some extent for some professions, thinking about the opportunities that proper use of this technology can open before people and businesses to resolve their problems more efficiently and create a competitive edge for themselves can alleviate such concerns. In this section, we review some of the potential breakthroughs we may expect to see in various areas of business as a result of using ChatGPT.

CUSTOMER SERVICE Chatbots have been used extensively during the past few years by businesses as virtual assistants to serve their customers, mostly for answering some basic frequent questions or connecting them to a relevant human expert for further assistance. However, their language capabilities are limited and many times this leads to even more confusion and dissatisfaction among customers as they might see those bots as a barrier between themselves and the human customer service experts. The ChatGPT technology, given its extraordinary potential to interact with human subjects, can be a huge game changer in this area. A customized chatbot relying on this technology can resolve a considerably wider variety of customer issues, decrease the workload on the human representatives and make them involved only when they are needed, and reduce the customer service expenditure for the business owners.

HEALTHCARE ChatGPT can offer benefits to all parties in the healthcare industry. It can be incorporated into the telehealth platforms as a virtual assistant to help patients with scheduling appointments, answering their general questions (e.g., drug information), and managing their medical records. Also, being trained by millions of medical articles on the web, it can be employed as a clinical decision support tool for doctors by providing consultation for complex cases or alarming about potential drug-drug interactions. Moreover, healthcare providers can use this technology to improve and automate processes such as patient triage and maintaining and summarizing patients' medical records.

MARKETING One of the major use cases of ChatGPT in marketing, which already has been begun by many businesses, is content creation. With the growth of social media platforms in the past decades and the high reliance of digital marketing campaigns on such channels for promoting products and services, it is now crucial for businesses to constantly provide their customers with informative and entertaining content. These include blog posts about their products or services, social media posts, or even podcast narratives, which all can be facilitated by ChatGPT.

COMPUTER PROGRAMMING While it is still too soon to ask ChatGPT to perform large-scale programming projects (e.g., developing a mobile application), it is capable of accomplishing simple to moderate programming tasks in specific languages such as Python, JavaScript, Java, and C++. This feature is especially useful for people with minimal programming knowledge who may need small helper programs to automate their routine daily operations. Additionally, ChatGPT provides step-by-step instructions along with each program it generates, which can be extremely helpful for learners to expand their coding skills.

Limitations of ChatGPT

Most of the major drawbacks of ChatGPT are related to the data set used for the initial training of its model. Following are a few of the most mentioned limitations of ChatGPT.

BIASES First, the training data is full of biases since it contains people's conversations and viewpoints in public forums, social media, personal blogs, and other publicly available channels. The extraordinarily large volume of the training data makes its auditing very difficult. Hence, following the well-known "garbage in, garbage out" principle, the generated answers might sometimes reflect gender, racial, societal, or historical biases. GPT-3, the ChatGPT's predecessor, for instance, is shown to have biases indicating gender stereotypes in its responses (e.g., describing women as less powerful than men) (Lucy and Bamman 2021). Also, in around %70 of the answers to the prompts about "Muslims," GPT-3 is shown to include references to "violence."³

FALSE ANSWERS Again, since the data set used for training ChatGPT is not audited, there are chances that the bot generates logically or historically false answers to some prompts. Especially, keep in mind that ChatGPT is a language model, not a calculator; so, while its responses are grammatically correct, its earlier version was struggling with solving even basic math problems, such as giving the question "If a banana weighs 0.5 lbs and I have 7 lbs of bananas and 9 oranges, how many pieces of fruits do I have?" received the answer of "You have 16 pieces of fruits, 7 bananas and 9 oranges," which is clearly not correct as the correct answer should be 23⁴. When even more complex question are posed to the latest version, the answer seems to be more accurate and properly articulated (see Figure 7.35b). Azaria reported that even sometimes its answer is just a random number based on human preference for numbers (Azaria 2022).

³ <https://hai.stanford.edu/news/rooting-out-anti-muslim-bias-popular-language-model-gpt-3>

⁴ See <https://www.wsj.com/articles/ai-bot-chatgpt-needs-some-help-with-math-assignments-11675390552> for a detailed discussion on this limitation of ChatGPT.

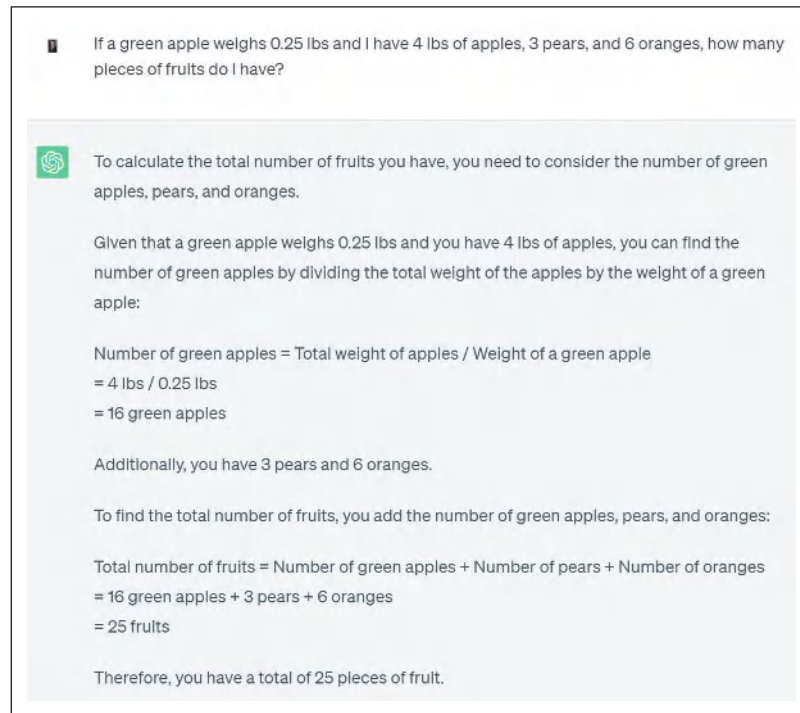


FIGURE 7.35b An Example of an Incorrect Answer by ChatGPT to a Simple Math Problem; the Correct Answer Is 23.

In addition, we still must be cautious about relying on even the factual responses generated by ChatGPT, since it uses numerous sources to compile an answer to a given prompt and it is not clear what sources and to what extent were involved in creating a given answer and how credible they are.

OBSOLETE DATA Another major limitation of the current version of ChatGPT is that it is pretrained using a snapshot of the internet taken in September 2021. This means that it may not be able to provide accurate output (if any) to the prompts related to the recent updates in the world. However, this does not seem to be a long-lasting shortcoming, and it is expected that the next generations of the chatbot would not suffer from that.

► SECTION 7.8 REVIEW QUESTIONS

1. What is RNN? How does it differ from CNN?
2. What is the significance of “context,” “sequence,” and “memory” in RNN?
3. Draw and explain the functioning of a typical recurrent neural network unit.
4. What is the LSTM network, and how does it differ from RNNs?
5. List and briefly describe three different types of LSTM applications.
6. How do Google’s Neural Machine Translation and Microsoft Skype Translator work?

7.9 COMPUTER FRAMEWORKS FOR IMPLEMENTATION OF DEEP LEARNING

Advances in deep learning owe its recent popularity, to a great extent, to advances in the software and hardware infrastructure required for its implementation. In the past few decades, GPUs have been revolutionized to support the playing of high-resolution videos as well as advanced video games and virtual reality applications. However, GPUs’ huge processing potential had not been effectively utilized for purposes other than graphics processing up until

a few years ago. Thanks to software libraries such as Theano (Bergstra et al., 2010), **Torch** (Collobert, Kavukcuoglu, and Farabet, 2011), **Caffe** (Jia et al., 2014), PyLearn2 (Goodfellow et al., 2013), **Tensorflow** (Abadi et al., 2016), and MXNet (Chen et al., 2015) developed with the purpose of programming GPUs for general-purpose processing (just as CPUs), and particularly for deep learning and analysis of Big Data, GPUs have become a critical enabler for the modern-day analytics. The operation of these libraries mostly relies on a parallel computing platform and application programming interface (API) developed by NVIDIA called *Compute Unified Device Architecture (CUDA)*, which enables software developers to use GPUs made by NVIDIA for general-purpose processing. In fact, each deep learning framework consists of a high-level scripting language (e.g., Python, R, Lua) and a library of deep learning routines usually written in C (for using CPUs) or CUDA (for using GPUs).

We next introduce some of the most popular software libraries used for deep learning by researchers and practitioners, including Torch, Caffe, Tensorflow, Theano, and Keras, and discuss some of their specific properties.

Torch

Torch (Collobert et al., 2011) is an open-source scientific computing framework (available at www.torch.ch) for implementing machine-learning algorithms using GPUs. The Torch framework is a library based on LuaJIT, a compiled version of the popular Lua programming language (www.lua.org). In fact, Torch adds a number of valuable features to Lua that make deep learning analyses possible; it enables supporting n -dimensional arrays (i.e., tensors), whereas tables (i.e., two-dimensional arrays) normally are the only data-structuring method used by Lua. Additionally, Torch includes routine libraries for manipulating (i.e., indexing, slicing, transposing) tensors, linear algebra, neural network functions, and optimization. More importantly, while Lua by default uses CPU to run the programs, Torch enables use of GPUs for running programs written in the Lua language.

The easy and extremely fast scripting properties of LuaJIT along with its flexibility have made Torch a very popular framework for practical deep learning applications such that today its latest version, Torch7, is widely used by a number of big companies in the deep learning area, including Facebook, Google, and IBM, in their research labs, as well as for their commercial applications.

Caffe

Caffe is another open-source deep learning framework (available at <http://caffe.berkeleyvision.org>) created by Yangqing Jia (2013), a PhD student at the University of California–Berkeley, which the Berkeley AI Research (BAIR) then further developed. Caffe has multiple options to be used as a high-level scripting language, including the command line, Python, and MATLAB interfaces. The deep learning libraries in Caffe are written in the C++ programming language.

In Caffe, everything is done using text files instead of code. That is, to implement a network, generally we need to prepare two text files with the *.prototxt* extension that are communicated by the Caffe engine via JavaScript Object Notation (JSON) format. The first text file, known as the *architecture* file, defines the architecture of the network layer by layer, where each layer is defined by a name, a type (e.g., data, convolution, output), the names of its previous (bottom) and next (top) layers in the architecture, and some required parameters (e.g., kernel size and stride for a convolutional layer). The second text file, known as the *solver* file, specifies the properties of the training algorithm, including the learning rate, maximum number of iterations, and processing unit (CPU or GPU) to be used for training the network.

While Caffe supports multiple types of deep network architectures like CNN and LSTM, it is particularly known to be an efficient framework for image processing due to its incredible speed in processing image files. According to its developers, it is able to process over 60 million images per day (i.e., 1 ms/image) using a single NVIDIA K40 GPU. In 2017, Facebook released an improved version of Caffe called Caffe2 (www.caffe2.ai) with the aim of improving the original framework to be effectively used for deep learning

architectures other than CNN and with a special emphasis on portability for performing cloud and mobile computations while maintaining scalability and performance.

TensorFlow

Another popular open-source deep learning framework is TensorFlow. It was originally developed and written in Python and C++ by the Google Brain Group in 2011 as *DistBelief*, but it was further developed into TensorFlow in 2015. TensorFlow at this time is the only deep learning framework that, in addition to CPUs and GPUs, supports Tensor Processing Units (TPUs), a type of processor developed by Google in 2016 for the specific purpose of neural network machine learning. In fact, TPUs were specifically designed by Google for the TensorFlow framework.

Although Google has not yet made TPUs available to the market, it is reported that it has used them in a number of its commercial services such as Google search, Street View, Google Photos, and Google Translate with significant improvements reported. A detailed study performed by Google shows that TPUs deliver 30 to 80 times higher performance per watt than contemporary CPUs and GPUs (Sato, Young, and Patterson, 2017). For example, it has been reported (Ung, 2016) that in Google Photos, an individual TPU can process over 100 million images per day (i.e., 0.86 ms/image). Such a unique feature will probably put TensorFlow way ahead of the other alternative frameworks in the near future as soon as Google makes TPUs commercially available.

Another interesting feature of TensorFlow is its visualization module, TensorBoard. Implementing a deep neural network is a complex and confusing task. TensorBoard refers to a Web application involving a handful of visualization tools to visualize network graphs and plot quantitative network metrics with the aim of helping users to better understand what is going on during training procedures and to debug possible issues.

Theano

In 2007, the Deep Learning Group at the University of Montreal developed the initial version of a Python library, **Theano** (<http://deeplearning.net/software/theano>), to define, optimize, and evaluate mathematical expressions involving multi-dimensional arrays (i.e., tensors) on CPU or GPU platforms. Theano was one of the first deep learning frameworks but later became a source of inspiration for the developers of TensorFlow. Theano and TensorFlow both pursue a similar procedure in the sense that in both a typical network implementation involves two sections: in the first section, a computational graph is built by defining the network variables and operations to be done on them; and the second section runs that graph (in Theano by compiling the graph into a *function* and in TensorFlow by creating a *session*). In fact, what happens in these libraries is that the user defines the structure of the network by providing some simple and symbolic syntax understandable even for beginners in programming, and the library automatically generates appropriate codes in either C (for processing on CPU) or CUDA (for processing on GPU) to implement the defined network. Hence, users without any knowledge of programming in C or CUDA and with just a minimum knowledge of Python are able to efficiently design and implement deep learning networks on the GPU platforms.

Theano also includes some built-in functions to visualize computational graphs as well as to plot the network performance metrics even though its visualization features are not comparable to TensorBoard.

Keras: An Application Programming Interface

While all described deep learning frameworks require users to be familiar with their own syntax (through reading their documentations) to be able to successfully train a network, fortunately there are some easier, more user-friendly ways to do so. **Keras** (<https://keras.io/>) is an open-source neural network library written in Python that functions as a high-level application programming interface (API) and is able to run on top of various deep learning frameworks including Theano and TensorFlow. In essence, Keras just by getting the key properties

of network building blocks (i.e., type of layers, transfer functions, and optimizers) via an extremely simple syntax automatically generates syntax in one of the deep learning frameworks and runs that framework in the backend. While Keras is efficient enough to build and run general deep learning models in just a few minutes, it does not provide several advanced operations provided by TensorFlow or Theano. Therefore, in dealing with special deep network models that require advanced settings, one still needs to directly use those frameworks instead of Keras (or other APIs such as *Lasagne*) as a proxy.

Analytics in Action 7.5 is an excellent example to the synergistic use deep learning methods and free, open-source tools to develop clinical decision support systems for the clinicians to minimize and potentially eliminate patients' readmissions to the emergency rooms.

ANALYTICS IN ACTION 7.5 A Deep Learning Approach to Predicting Early Bounce-Backs to the Emergency Departments

Motivation

Patients who are deemed safe for discharge from the emergency department (ED) but then are subsequently admitted to the hospital, die, or are transferred within a short time window represent a subset of patients who may have had clinically important errors or adverse events on the initial visit. As such, focusing on the review of patients who return to the ED within 72 hours after discharge (a.k.a. bounce-back) is a common quality assurance process for hospital systems. A bounce-back is defined as a patient's return to the ED within a 72-hour window after discharge (from ED), regardless of whether the new visit leads to hospitalization or just requires outpatient treatment. Bounce-backs may be due to various reasons: some returns are planned (e.g., wound check), while other bounce-backs may represent substandard quality of care, including missed diagnoses, improper disposition (i.e., discharge versus admission), or inadequate follow-up after discharge.

Methodology

Emergency department visits to the network of four major hospitals in New York City between January 2015 and April 2019 from the Emergency Physicians Integrated Care (EPIC) database were used in this study. We excluded visits by non-adult (< 18 years of age) patients as well as by "frequent flyers" (patients with more than 4 ED visits per year). The resulting data set contains 657,752 visits (by 382,624 unique patients), from which around 5.5% (38,630 visits) were patient returns within 72 hours after discharge. Each visit record in the data set included initial vital signs (blood pressure, temperature, heart rate, respiratory rate, peripheral oxygen saturation), demographics (age and gender), and ED diagnosis.

In addition, each visit contains an ED physician documentation note, which includes a free-text narrative describing the patient's medical history, presenting illness, physical exam, and medical decision-making process. After performing regular text preprocessing operations (converting words to lower-case, lemmatization, removing punctuations, etc.) on the physicians' notes, we used the Doc2Vec approach to embed each note into a vector in a 300-dimensional space.

In the end, a rich combination of feature space is constructed using both structured and unstructured data elements. Figure 7.36 illustrates the steps involved in the data preprocessing phase.

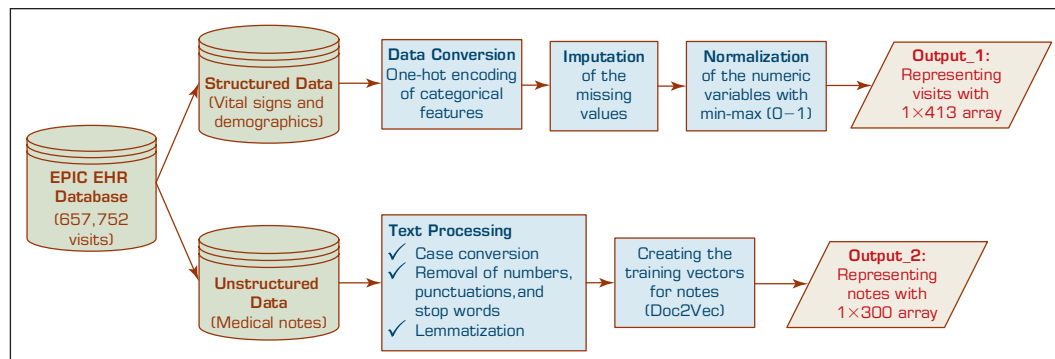


FIGURE 7.36 A Graphical Summary of the Data Preprocessing Phase.

TABLE 7.3 Prediction Model Performance

Metric	Vitals, History, and Demographics only	Physician Notes Included	Improvement
Accuracy	0.668	0.748	0.080
Sensitivity/Recall	0.615	0.756	0.141
Specificity	0.671	0.747	0.076
AUROC	0.674	0.766	0.092

Results

The deep-learning neural network was trained using 80% of the data and tested on the remaining 20%, which is not involved in training (a random stratified sampling approach was used for partitioning). We used Tensorflow (with Keras backend) via the KNIME Analytics Platform extension to perform the model training. A best-practice driven optimization methodology is employed to identify the initial learning rate, avoid overfitting, and systematically adjust the learning process.

The final deep learning model was able to predict 74.8% of the bounce-backs (AUROC = 0.766). These results show that leveraging state-of-the-art deep learning techniques to extract features from unstructured ED physician narrative notes and incorporating them (along with structured vitals and demographics data) in the prediction task can remarkably improve the outcome. Table 7.3 shows the changes on the prediction metrics instigated by inclusion of the unstructured data.

In addition, a sensitivity analysis is performed on the structured features of the model revealed that factors such as patients’ age, type of medical insurance, chronic conditions

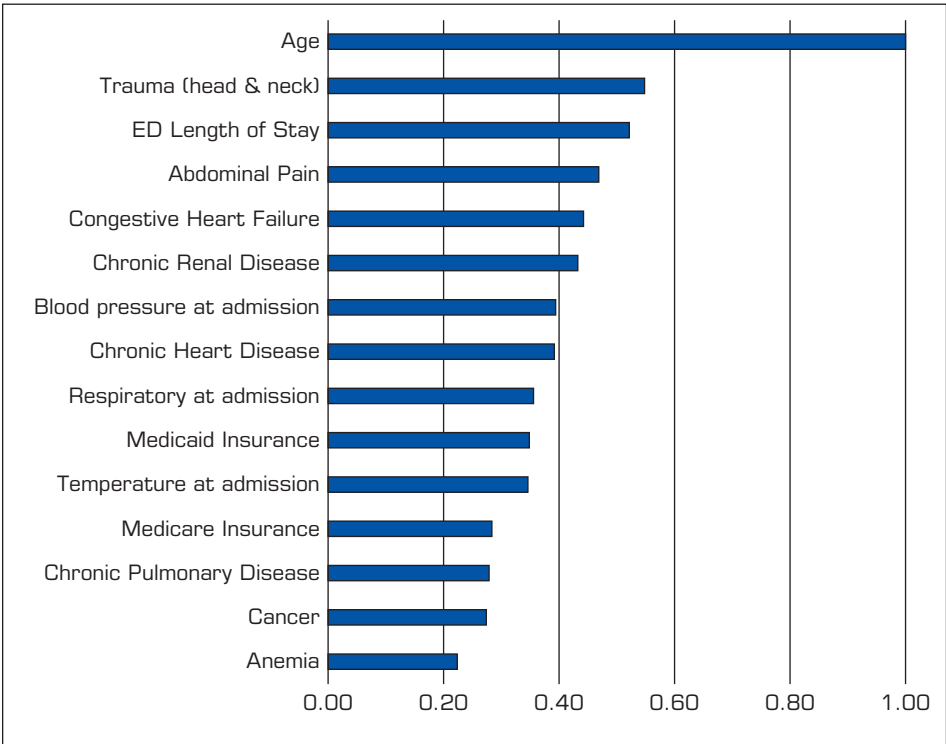


FIGURE 7.37 Top 15 Most Influential Factors Based on the Sensitivity Analysis.

(especially renal, heart, and respiratory), heart failure, and head or neck trauma are among the top factors increasing the chances of patients' unplanned return to the ED. See Figure 7.37 for the top 15 risk factors along with their relative importance score obtained via the sensitivity analysis.

The proposed framework may be used as a decision support tool to assist emergency care clinicians in the early identification of the high-risk patients who are likely to bounce-back and provide them with timely and appropriate care.

Sources: Davazdahemami, B., Peng, P., & Delen, D. (2022). A deep learning approach for predicting early bounce-backs to the emergency departments. *Healthcare Analytics*, 2, 100018; Peng, P., Davazdahemami, B., Delen, D., Shapiro, J., & Manini, A. F. (2019). 288 Predicting Potential Bouncebacks to the Emergency Department: A Machine Learning Approach. *Annals of Emergency Medicine*, 74(4), S114.

SECTION 7.9 REVIEW QUESTIONS

1. Despite the short tenure of deep learning implementation, why do you think there are several different computing frameworks for it?
2. Define *CPU*, *NVIDIA*, *CUDA*, and *deep learning*, and comment on the relationship between them.
3. List and briefly define the characteristics of different deep learning frameworks.
4. What is Keras, and how is it different from the other frameworks?

7.10 COGNITIVE COMPUTING

We are witnessing a significant increase in the way technology is evolving. Things that once took decades are now taking months, and the things that we see only in sci-fi movies are becoming reality, one after another. Therefore, it is safe to say that in the next decade or two, technological advancements will transform how people live, learn, and work in a rather dramatic fashion. The interactions between humans and technology will become intuitive, seamless, and perhaps transparent. Cognitive computing will have a significant role to play in this transformation. Generally speaking, cognitive computing refers to the computing systems that use mathematical models to emulate (or partially simulate) the human cognition process to find solutions to complex problems and situations where the potential answers can be imprecise. While the term **cognitive computing** is often used interchangeably with AI and smart search engines, the phrase itself is closely associated with IBM's cognitive computer system Watson and its success on the television show *Jeopardy!* Details on Watson's success on *Jeopardy!* can be found in Opening Vignette in Chapter 6.

According to Cognitive Computing Consortium (2018), cognitive computing makes a new class of problems computable. It addresses highly complex situations that are characterized by ambiguity and uncertainty; in other words, it handles the kinds of problems that are thought to be solvable by human ingenuity and creativity. In today's dynamic, information-rich, and unstable situations, data tend to change frequently, and they often conflict. The goals of users evolve as they learn more and redefine their objectives. To respond to the fluid nature of users' understanding of their problems, the cognitive computing system offers a synthesis not just of information sources but also of influences, contexts, and insights. To achieve such a high-level of performance, cognitive systems often need to weigh conflicting evidence and suggest an answer that is "best" rather than "right." Figure 7.38 illustrates a general framework for cognitive computing where data and AI technologies are used to solve complex real-world problems.

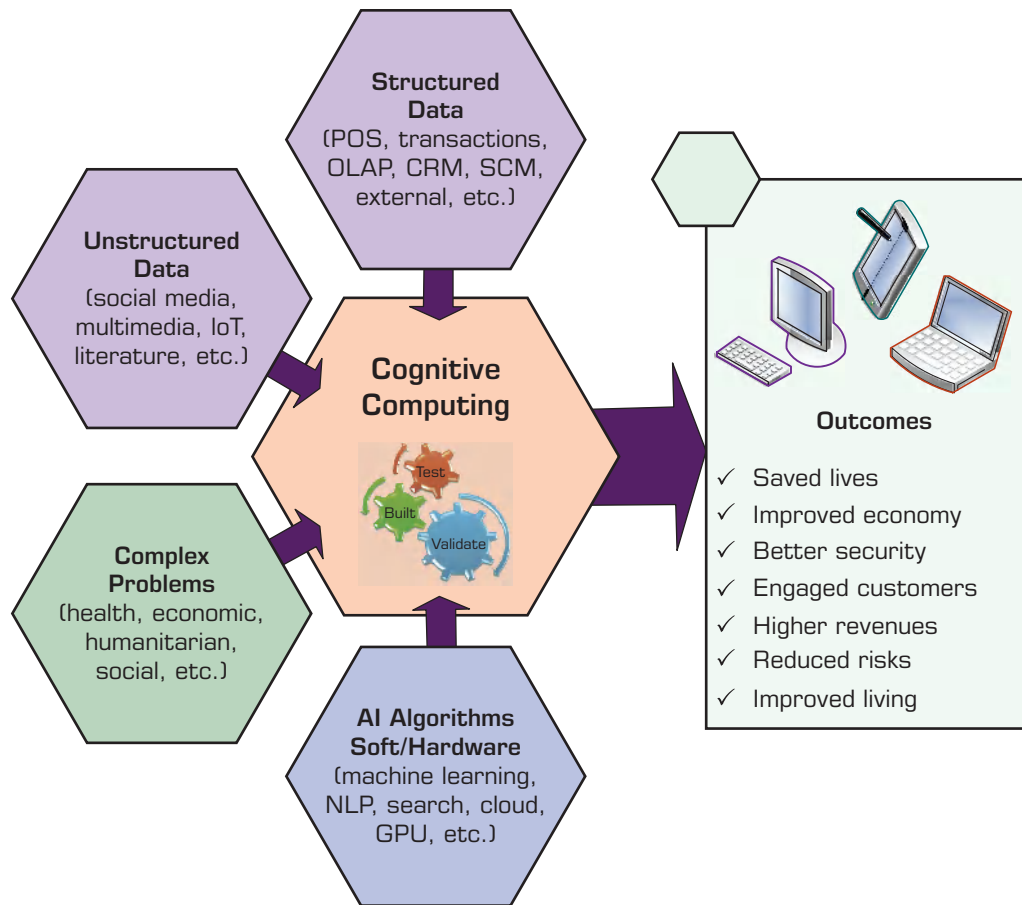


FIGURE 7.38 Conceptual Framework for Cognitive Computing and Its Promises.

How Does Cognitive Computing Work?

As one would guess from the name, cognitive computing works much like a human thought process, reasoning mechanism, and cognitive system. These cutting-edge computation systems can find and synthesize data from various information sources and weigh context and conflicting evidence inherent in the data to provide the best possible answers to a given question or problem. To achieve this, cognitive systems include self-learning technologies that use data mining, pattern recognition, deep learning, and NLP to mimic the way the human brain works.

Using computer systems to solve the types of problems that humans are typically tasked with requires vast amounts of structured and unstructured data fed to machine-learning algorithms. Over time, cognitive systems are able to refine the way in which they learn and recognize patterns and the way they process data to become capable of anticipating new problems and modeling and proposing possible solutions.

To achieve those capabilities, cognitive computing systems must have the following key attributes as defined by the Cognitive Computing Consortium (2018):

- **Adaptive:** Cognitive systems must be flexible enough to learn as information changes and goals evolve. The systems must be able to digest dynamic data in real time and make adjustments as the data and environment change.

- **Interactive:** Human-computer interaction (HCI) is a critical component in cognitive systems. Users must be able to interact with cognitive machines and define their needs as those needs change. The technologies must also be able to interact with other processors, devices, and cloud platforms.
- **Iterative and stateful:** Cognitive computing technologies can also identify problems by asking questions or pulling in additional data if a stated problem is vague or incomplete. The systems do this by maintaining information about similar situations that have previously occurred.
- **Contextual:** Understanding context is critical in thought processes, so cognitive systems must understand, identify, and mine contextual data, such as syntax, time, location, domain, requirements, and a specific user's profile, tasks, or goals. Cognitive systems may draw on multiple sources of information, including structured and unstructured data and visual, auditory, or sensor data.

How Does Cognitive Computing Differ from AI?

Cognitive computing is often used interchangeably with AI, the umbrella term used for technologies that rely on data and scientific methods/computations to make (or help/support in making) decisions. But there are differences between the two terms, which can largely be found within their purposes and applications. AI technologies include—but are not limited to—machine learning, neural computing, NLP, and, most recently, deep learning. With AI systems, especially in machine-learning systems, data are fed into the algorithm for processing (an iterative and time-demanding process that is often called *training*) so that the systems “learn” variables and interrelationships among those variables so that it can produce predictions (or characterizations) about a given complex problem or situation. Applications based on AI and cognitive computing include intelligent assistants, such as Amazon’s Alexa, Google Home, and Apple’s Siri. A simple comparison between cognitive computing and AI is given in Table 7.4 (Reynolds and Feldman, 2014; CCC, 2018).

As can be seen in Table 7.4, the differences between AI and cognitive computing are rather marginal. This is expected because cognitive computing is often characterized as a subcomponent of AI or an application of AI technologies tailored for a specific

TABLE 7.4 Cognitive Computing versus Artificial Intelligence (AI)

Characteristic	Cognitive Computing	Artificial Intelligence (AI)
Technologies used	• Machine learning • Natural language processing • Neural networks • Deep learning • Text mining • Sentiment analysis	• Machine learning • Natural language processing • Neural networks • Deep learning
Capabilities offered	Simulate human thought processes to assist humans in finding solutions to complex problems	Find hidden patterns in a variety of data sources to identify problems and provide potential solutions
Purpose	Augment human capability	Automate complex processes by acting like a human in certain situations
Industries	Customer service, marketing, healthcare, entertainment, service sector	Manufacturing, finance, healthcare, banking, securities, retail, government

purpose. AI and cognitive computing both utilize similar technologies and are applied to similar industry segments and verticals. The main difference between the two is the purpose: while cognitive computing is aimed at helping humans to solve complex problems, AI is aimed at automating processes that are performed by humans; at the extreme, AI is striving to replace humans with machines for tasks requiring “intelligence,” one at a time.

In recent years, cognitive computing typically has been used to describe AI systems that aim to simulate human thought process. Human cognition involves real-time analysis of environment, context, and intent among many other variables that inform a person’s ability to solve problems. A number of AI technologies are required for a computer system to build cognitive models that mimic human thought processes, including machine learning, deep learning, neural networks, NLP, text mining, and sentiment analysis.

In general, cognitive computing is used to assist humans in their decision-making process. Some examples of cognitive computing applications include supporting medical doctors in their treatment of disease. IBM Watson for Oncology, for example, has been used at Memorial Sloan Kettering Cancer Center to provide oncologists evidence-based treatment options for cancer patients. When medical staff input questions, Watson generates a list of hypotheses and offers treatment options for doctors to consider. Whereas AI relies on algorithms to solve a problem or to identify patterns hidden in data, cognitive computing systems have the loftier goal of creating algorithms that mimic the human brain’s reasoning process to help humans solve an array of problems as the data and the problems constantly change.

In dealing with complex situations, context is important, and cognitive computing systems make context computable. They identify and extract context features such as time, location, task, history, or profile to present a specific set of information that is appropriate for an individual or for a dependent application engaged in a specific process at a specific time and place. According to the Cognitive Computing Consortium, they provide machine-aided serendipity by wading through massive collections of diverse information to find patterns and then apply those patterns to respond to the needs of the user at a particular moment. In a sense, cognitive computing systems aim at redefining the nature of the relationship between people and their increasingly pervasive digital environment. They may play the role of assistant or coach for the user, and they may act virtually autonomously in many problem-solving situations. The boundaries of the processes and domains these systems can affect are still elastic and emergent. Their output may be prescriptive, suggestive, instructive, or simply entertaining.

In the short time of its existence, cognitive computing has proved to be useful in many domain and complex situations and is evolving into many more. The typical use cases for cognitive computing include the following:

- Development of smart and adaptive search engines
- Effective use of natural language processing
- Speech recognition
- Language translation
- Context-based sentiment analysis
- Face recognition and facial emotion detection
- Risk assessment and mitigation
- Fraud detection and mitigation
- Behavioral assessment and recommendations

Cognitive analytics is a term that refers to cognitive computing–branded technology platforms, such as IBM Watson, that specialize in processing and analyzing large, unstructured data sets. Typically, word processing documents, e-mails, videos, images, audio files, presentations, Web pages, social media, and many other data formats need to be manually tagged with metadata before they can be fed into a traditional analytics

engine and Big Data tools for computational analyses and insight generation. The principal benefit of utilizing cognitive analytics over those traditional Big Data analytics tools is that for cognitive analytics such data sets do not need to be pretagged. Cognitive analytics systems can use machine learning to adapt to different contexts with minimal human supervision. These systems can be equipped with a chatbot or search assistant that understands queries, explains data insights, and interacts with humans in human languages.

Cognitive Search

Cognitive search is the new generation search method that uses AI (advanced indexing, NLP, and machine learning) to return results that are much more relevant to users. Forrester defines cognitive search and knowledge discovery solutions as “a new generation of enterprise search solutions that employ AI technologies such as natural language processing and machine learning to ingest, understand, organize, and query digital content from multiple data sources” (Gualtieri, 2017). Cognitive search creates searchable information out of nonsearchable content by leveraging cognitive computing algorithms to create an indexing platform.

Searching for information is a tedious task. Although current search engines do a very good job in finding relevant information in a timely manner, their sources are limited to publically available data over the Internet. Cognitive search proposes the next generation of search tailored for use in enterprises. It is different from traditional search because, according to Gualtieri (2017), it:

- **Can handle a variety of data types.** Search is no longer just about unstructured text contained in documents and in Web pages. Cognitive search solutions can also accommodate structured data contained in databases and even nontraditional enterprise data such as images, video, audio, and machine-/sensor-generated logs from IoT devices.
- **Can contextualize the search space.** In information retrieval, the context is important. Context takes the traditional syntax-/symbol-driven search to a new level where it is defined by semantics and meaning.
- **Employ advanced AI technologies.** The distinguishing characteristic of cognitive search solutions is that they use NLP and machine learning to understand and organize data, predict the intent of the search query, improve the relevancy of results, and automatically tune the relevancy of results over time.
- **Enable developers to build enterprise-specific search applications.** Search is not just about a text box on an enterprise portal. Enterprises build search applications that embed search in customer 360 applications, pharma research tools, and many other business process applications. Virtual digital assistants such as Amazon Alexa, Google Now, and Siri would be useless without powerful searches behind the scenes. Enterprises wishing to build similar applications for their customers will also benefit from cognitive search solutions. Cognitive search solutions provide software development kits (SDKs), APIs, and/or visual design tools that allow developers to embed the power of the search engine in other applications.

Figure 7.39 shows the progressive evolution of search methods from good old keyword search to modern-day cognitive search on two dimensions—ease of use and value proposition.

An Example of Cognitive Computing: IBM Watson

IBM Watson is perhaps the smartest computer system built to date. Since the emergence of computers and subsequently AI in the late 1940s, scientists have compared the performance of these “smart” machines with human minds. Accordingly, in the mid- to late-1990s,

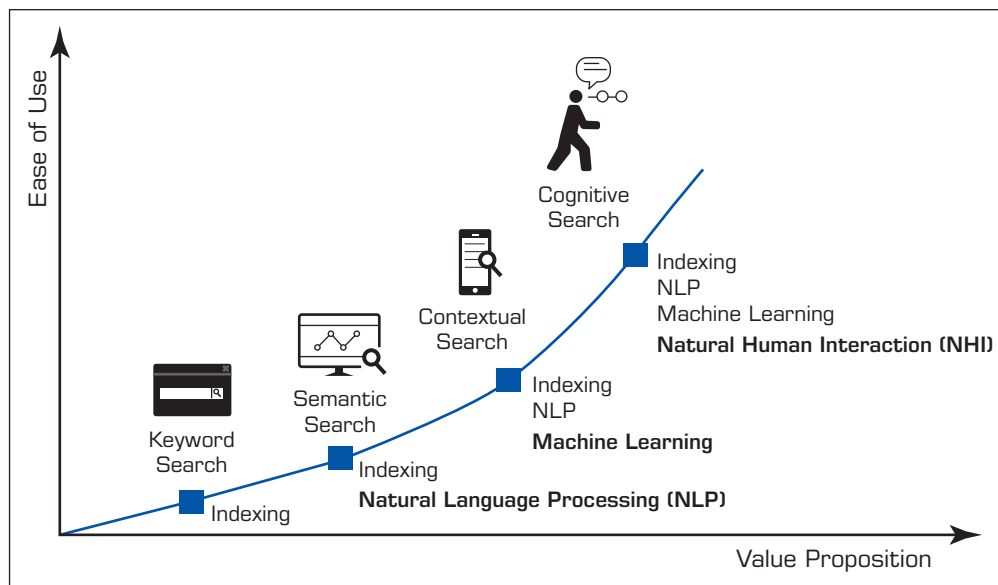


FIGURE 7.39 Progressive Evolution of Search Methods.

IBM researchers built a smart machine and used the game of chess (generally credited as the game of smart humans) to test its ability against the best of human players. On May 11, 1997, an IBM computer called Deep Blue beat the world chess grandmaster after a six-game match series: two wins for Deep Blue, one for the champion, and three draws. The match lasted several days and received massive media coverage around the world. It was the classic plot line of human versus machine. Beyond the chess contest, the intention of developing this kind of computer intelligence was to make computers able to handle the kinds of complex calculations needed to help discover new drugs and to do the broad financial modeling needed to identify trends and do risk analysis, handle large database searches, and perform massive calculations needed in advanced fields of science.

After a couple of decades, IBM researchers came up with another idea that was perhaps more challenging: a machine that could not only play the American TV quiz show *Jeopardy!* but also beat the best of the best. Compared to chess, *Jeopardy!* is much more challenging. While chess is well structured and has very simple rules and therefore is a very good match for computer processing, *Jeopardy!* is neither simple nor structured. *Jeopardy!* is a game designed to test human intelligence and creativity. Therefore, a computer designed to play the game needed to be a cognitive computing system that can work and think like a human. Making sense of imprecision inherent in human language was the key to success.

In 2010, an IBM research team developed Watson, an extraordinary computer system—a novel combination of advanced hardware and software—designed to answer questions posed in natural human language. The team built Watson as part of the **DeepQA** project and named it after IBM's first president, Thomas J. Watson. The team that built Watson was looking for a major research challenge: one that could rival the scientific and popular interest of Deep Blue and would have clear relevance to IBM's business interests. The goal was to advance computational science by exploring new ways for computer technology to affect science, business, and society at large. Accordingly, IBM research undertook a challenge to build Watson as a computer system that could compete at the human champion level in real time on *Jeopardy!* The team wanted to create a real-time automatic contestant on the show capable of listening, understanding, and responding, not merely a laboratory exercise.

After the show, the big question was “So what now?” Was developing Watson all for a quiz show? Absolutely not! Showing the rest of the world what Watson (and the cognitive system behind it) could do became an inspiration for the next generation of intelligent information systems. For IBM, it was a demonstration of what is possible with cutting-edge analytics and computational sciences. The message is clear: If a smart machine can beat the best of the best in humans at what they are the best at, think about what it can do for your organizational problems.

The innovative and futuristic technologies that made Watson one of the most acclaimed technological advances of this decade are being leveraged as computational foundation for several tools to analyze and characterize unstructured data for prediction-type problems. These experimental tools include Tone Analyzer and Personality Insights. Using textual content, these tools have shown the ability to predict outcomes of complex social events and globally popular competitions.

WATSON PREDICTS THE WINNER OF EUROVISION SONG CONTEST A tool developed on the foundations of IBM Watson, Watson Tone Analyzer, uses computational linguistics to identify *tone* in written text. Its broader goal is to have business managers use the Tone Analyzer to understand posts, conversations, and communications of target customer populations and to respond to their needs and wants in a timely manner. One could, for example, use this tool to monitor social media and other Web-based content, including wall posts, tweets, product reviews, and discussion boards as well as longer documents such as articles and blog posts. Or one could use it to monitor customer service interactions and support-related conversations. Although it sounds as if any other text-based detection system can build on sentiment analysis, Tone Analyzer differs from these systems in that it analyzes and characterizes textual content. Watson Tone Analyzer measures social tendencies and opinions, using a version of the Big-5, the five categories of personality traits (i.e., openness, agreeableness, conscientiousness, extroversion, and neuroticism), along with other emotional categories to detect the tone in a given textual content. As an example, Slowey (2017b) used IBM’s Watson Tone Analyzer to predict the winner of the 2017 Eurovision Songs Contest. Using nothing but the lyrics of the previous years’ competitions, Slowey discovered a pattern that suggested most winners had high levels of agreeableness and conscientiousness. The results (produced before the contest) indicated that Portugal would win the contest, and that is exactly what happened. Try it out yourself:

- Go to Watson Tone Analyzer (<https://tone-analyzer-demo.ng.bluemix.net>).
- Copy and paste your own text in the provided text entry field.
- Click “Analyze.”
- Observe the summary results as well as the specific sentences where specific tones are the strongest

Another tool built on the linguistic foundations of IBM Watson is Watson Personality Insight, which seems to work quite similar to Watson Tone Analyzer. In another fun application case, Slowey (2017a) used Watson Personality Insight to predict the winner of the best picture category at the 2017 Oscar Academy Awards. Using the scripts of the movies from the past years, Slowey developed a generalized profile for winners and then compared that profile to those of the newly nominated movies to identify the upcoming winner. Although in this case, Slowey incorrectly predicted *Hidden Figures* as the winner, the methodology she followed was unique and innovative and hence deserves credit.

One of the worthiest endeavors for Watson (or Watson-like large-scale cognitive computing systems) is to help doctors and other medical professionals to diagnose diseases and identify the best treatment options that would work for an individual patient. Although Watson is new, this very novel and worthy task is not new to the world of

computing. In the early 1970s, several researchers at Stanford University developed a computer system, **MYCIN**, to identify bacteria causing severe infections, such as bacteremia and meningitis, and to recommend antibiotics with the dosage adjusted for the specifics of an individual patient (Buchanan and Shortliffe, 1984). This six-year effort relied on a rule-based expert system, a type of AI system, where the diagnoses and treatment knowledge nuggets/rules were elicited from a large number of experts (i.e., doctors with ample experience in the specific medical domain). The resulting system was then tested on new patients, and its performance was compared to those of the experienced doctors used as the knowledge sources/experts. The results favored MYCIN, providing a clear indication that properly designed and implemented AI-based computer systems can meet and often exceed the effectiveness and efficiency of even the best medical experts. After more than four decades, Watson is now trying to pick up where MYCIN left the mission of using smart computer systems to improve the health and well-being of humans by helping doctors with the contextual information that they need to better and more quickly diagnose and treat their patients.

The first industry targeted to utilize Watson was healthcare, followed by security, finance, retail, education, public services, and research. The following sections provide short descriptions of what Watson can do (and, in many cases, is doing) for these industries.

HEALTHCARE AND MEDICINE The challenges that healthcare is facing today are rather big and multifaceted. With the aging U.S. population, which may be partially attributed to better living conditions and advanced medical discoveries fueled by a variety of technological innovations, the demand for healthcare services is increasing faster than the supply of resources. As we all know, when there is an imbalance between demand and supply, prices go up and quality suffers. Therefore, we need cognitive systems like Watson to help decision makers optimize the use of their resources in both clinical and managerial settings.

According to healthcare experts, only 20 percent of the knowledge that physicians use to diagnose and treat patients is evidence based. Considering that the amount of medical information available is doubling every five years and that much of these data are unstructured, physicians simply do not have time to read every journal that can help them keep up to date with the latest advances. Given the growing demand for services and the complexity of medical decision making, how can healthcare providers address these problems? The answer could be to use Watson or similar cognitive systems that have the ability to help physicians in diagnosing and treating patients by analyzing large amounts of data—both structured data coming from electronic medical record databases and unstructured text coming from physician notes and published literature—to provide evidence for faster and better decision making. First, the physician and the patient can describe symptoms and other related factors to the system in natural language. Watson can then identify the key pieces of information and mine the patient's data to find relevant facts about family history, current medications, and other existing conditions. It can then combine that information with current findings from tests and then can form and test hypotheses for potential diagnoses by examining a variety of data sources—treatment guidelines, electronic medical record data, doctors' and nurses' notes, and peer-reviewed research and clinical studies. Next, Watson can suggest potential diagnostics and treatment options with a confidence rating for each suggestion.

Watson also has the potential to transform healthcare by intelligently synthesizing fragmented research findings published in a variety of outlets. It can dramatically change the way medical students learn. It can help healthcare managers to be proactive about upcoming demand patterns, optimally allocate resources, and improve processing of payments. Early examples of leading healthcare providers that use Watson-like cognitive systems include MD Anderson, The Cleveland Clinic, and Memorial Sloan Kettering.

SECURITY As the Internet expands into every facet of our lives—e-commerce, e-business, smart grids for energy, smart homes for remote control of residential gadgets and appliances—to make things easier to manage, it also opens up the potential for ill-intended people to intrude in our lives. We need smart systems like Watson that are capable of constantly monitoring for abnormal behavior and, when it is identified, preventing people from accessing our lives and harming us. This could be at the corporate or even national security system level; it could also be at the personal level. Such a smart system could learn who we are and become a digital guardian that could make inferences about activities related to our life and alert us whenever abnormal things happen.

FINANCE The financial services industry faces complex challenges. Regulatory measures as well as social and governmental pressures for financial institutions to be more inclusive have increased. And the customers the industry serves are more empowered, demanding, and sophisticated than ever before. With so much financial information generated each day, it is difficult to properly harness the appropriate information on which to act. Perhaps the solution is to create smarter client engagement by better understanding risk profiles and the operating environment. Major financial institutions are already working with Watson to infuse intelligence into their business processes. Watson is tackling data-intensive challenges across the financial services sector, including banking, financial planning, and investing.

RETAIL The retail industry is rapidly changing according to customers' needs and wants. Empowered by mobile devices and social networks that give them easier access to more information faster than ever before, customers have high expectations for products and services. While retailers are using analytics to keep up with those expectations, their bigger challenge is efficiently and effectively analyzing the growing mountain of real-time insights that could give them a competitive advantage. Watson's cognitive computing capabilities related to analyzing massive amounts of unstructured data can help retailers reinvent their decision-making processes around pricing, purchasing, distribution, and staffing. Because of Watson's ability to understand and answer questions in natural language, Watson is an effective and scalable solution for analyzing and responding to social sentiment based on data obtained from social interactions, blogs, and customer reviews.

EDUCATION With the rapidly changing characteristics of students—who are more visually oriented/stimulated, constantly connected to social media and social networks, and with increasingly shorter attention spans—what should the future of education and the classroom look like? The next generation of educational systems should be tailored to fit the needs of the new generation with customized learning plans, personalized textbooks (digital ones with integrated multimedia—audio, video, animated graphs/charts, etc.), dynamically adjusted curriculum, and perhaps smart digital tutors and 24/7 personal advisors. Watson seems to have what it takes to make all this happen. With its NLP capability, students can converse with it just as they do with their teachers, advisors, and friends. This smart assistant can answer students' questions, satisfy their curiosity, and help them keep up with the endeavors of the educational journey.

GOVERNMENT For local, regional, and national governments, the exponential rise of Big Data presents an enormous dilemma. Today's citizens are more informed and empowered than ever before, and that means they have high expectations for the value of the public sector serving them. And government organizations can now gather enormous volumes of unstructured, unverified data that could serve their citizens, but only if those data can be analyzed efficiently and effectively. IBM Watson's cognitive computing may help make sense of this data deluge, speeding governments' decision-making processes and helping public employees to focus on innovation and discovery.

RESEARCH Every year, hundreds of billions of dollars are spent on research and development, most of it documented in patents and publications, creating an enormous amount of unstructured data. To contribute to the extant body of knowledge, one needs to sift through these data sources to find the outer boundaries of research in a particular field. This is very difficult, if not impossible, work if it is done with traditional means, but Watson can act as a research assistant to help collect and synthesize information to keep people updated on recent findings and insights. For instance, the New York Genome Center is using the IBM Watson cognitive computing system to analyze the genomic data of patients diagnosed with a highly aggressive and malignant brain cancer and to more rapidly deliver personalized, life-saving treatment to patients with this disease (Royyuru, 2014).

► SECTION 7.10 REVIEW QUESTIONS

1. What is cognitive computing, and how does it differ from other computing paradigms?
2. Draw a diagram and explain the conceptual framework of cognitive computing. Make sure to include inputs, enablers, and expected outcomes in your framework.
3. List and briefly define the key attributes of cognitive computing.
4. How does cognitive computing differ from ordinary AI techniques?
5. What are the typical use cases for cognitive analytics?
6. Explain what the terms *cognitive analytics* and *cognitive search* mean.
7. What is IBM Watson and what is its significance to the world of computing?
8. How does Watson work?
9. List and briefly explain five use cases for IBM Watson.

Chapter Highlights

- Deep learning is among the latest trends in AI that come with great expectations.
- The goal of deep learning is similar to those of the other machine-learning methods, which is to use sophisticated mathematical algorithms to learn from data similar to the way that humans learn.
- What deep learning has added to the classic machine-learning methods is the ability to automatically acquire the features required to accomplish highly complex and unstructured tasks.
- Deep learning belongs to the representation learning within the AI learning family of methods.
- The recent emergence and popularity of deep learning can largely be attributed to very large data sets and rapidly advancing computing infrastructures.
- Artificial neural networks emulate the way the human brain works. The basic processing unit is a neuron. Multiple neurons are grouped into layers and linked together.
- In a neural network, knowledge is stored in the weight associated with the connections between neurons.
- Backpropagation is the most popular learning paradigm of feedforward neural networks.
- An MLP-type neural network consists of an input layer, an output layer, and a number of hidden layers. The nodes in one layer are connected to the nodes in the next layer.
- Each node at the input layer typically represents a single attribute that may affect the prediction.
- The usual process of learning in a neural network involves three steps: (1) compute temporary outputs based on inputs and random weights, (2) compute outputs with desired targets, and (3) adjust the weights and repeat the process.
- Developing neural network-based systems requires a step-by-step process. It includes data preparation and preprocessing, training and testing, and conversion of the trained model into a production system.

- Neural network software allows for easy experimentation with many models. Although neural network modules are included in all major data mining software tools, specific neural network packages are also available.
- Neural network applications abound in almost all business disciplines as well as in virtually all other functional areas.
- Overfitting occurs when neural networks are trained for a large number of iterations with relatively small data sets. To prevent overfitting, the training process is controlled by an assessment process using a separate validation data set.
- Neural networks are known as *black-box models*. Sensitivity analysis is often used to shed light into the black box to assess the relative importance of input features.
- Deep neural networks broke the generally accepted notion of “no more than two hidden layers are needed to formulate complex prediction problems.” They promote increasing the hidden layer to arbitrarily large numbers to better represent the complexity in the data set.
- MLP deep networks, also known as *deep feedforward networks*, are the most general type of deep networks.
- The impact of random weights in the learning process of deep MLP is shown to be a significant issue. Nonrandom assignment of the initial weights seems to significantly improve the learning process in deep MLP.
- Although there is no generally accepted theoretical basis for this, it is believed and empirically shown that in deep MLP networks, multiple layers perform better and converge faster than few layers with many neurons.
- CNNs are arguably the most popular and most successful deep learning methods.
- CNNs were initially designed for computer vision applications (e.g., image processing, video processing, text recognition) but also have been shown to be applicable to nonimage or non-text data sets.
- The main characteristic of the convolutional networks is having at least one layer involving a convolution weight function instead of general matrix multiplication.
- The convolution function is a method to address the issue of having too many network weight parameters by introducing the notion of parameter sharing.
- In CNN, a convolution layer is often followed by another layer known as the *pooling* (a.k.a. *subsampling*) layer. The purpose of a pooling layer is to consolidate elements in the input matrix in order to produce a smaller output matrix while maintaining the important features.
- ImageNet is an ongoing research project that provides researchers with a large database of images, each linked to a set of synonym words (known as *synset*) from WordNet (a word hierarchy database).
- AlexNet is one of the first convolutional networks designed for image classification using the ImageNet data set. Its success rapidly popularized the use and reputation of CNNs.
- GoogLeNet (a.k.a. *Inception*), a deep convolutional network architecture designed by Google researchers, was the winning architecture at ILSVRC 2014.
- Google Lens is an app that uses deep learning artificial neural network algorithms to deliver information about the images captured by users from their nearby objects.
- Google’s word2vec project remarkably increased the use of CNN-type deep learning for text mining applications.
- RNN is another deep learning architecture designed to process sequential inputs.
- RNNs have memory to remember previous information in determining context-specific, time-dependent outcomes.
- A variation of RNN, the LSTM network is today known as the most effective sequence modeling technique and is the base of many practical applications.
- Two emerging LSTM applications are Google Neural Machine Translator and Microsoft Skype Translator.
- Deep learning implementation frameworks include Torch, Caffe, TensorFlow, Theano, and Keras.
- Cognitive computing makes a new class of problems computable by addressing highly complex situations that are characterized by ambiguity and uncertainty; in other words, it handles the kinds of problems that are thought to be solvable by human ingenuity and creativity.
- Cognitive computing finds and synthesizes data from various information sources and weighs the context and conflicting evidence inherent in the data in order to provide the best possible answers to a given question or problem.
- The key attributes of cognitive computing include adaptability, interactivity, being iterative, stateful, and contextual.

- *Cognitive analytics* is a term that refers to cognitive computing–branded technology platforms, such as IBM Watson, that specialize in the processing and analysis of large unstructured data sets.
- Cognitive search is the new generation of search method that uses AI (advanced indexing, NLP, and machine learning) to return results that are much more relevant to the user than traditional search methods.
- IBM Watson is perhaps the smartest computer system built to date. It has coined and popularized the term *cognitive computing*.
- IBM Watson beat the best of men (the two most winning competitors) at the quiz game *Jeopardy!*, showcasing the ability of computers to do tasks that are designed for human intelligence.
- Watson and systems like it are now in use in many application areas including healthcare, finance, security, and retail.

Key Terms

activation function	deep belief network (DBN)	machine learning	representation learning
artificial intelligence (AI)	deep learning	Microsoft Skype	sensitivity analysis
artificial neural networks (ANN)	deep neural network	Translator	stochastic gradient descent (SGD)
backpropagation	DeepQA	multilayer perceptron (MLP)	summation function
black-box syndrome	Google Lens	MYCIN	supervised learning
Caffe	GoogLeNet	network structure	TensorFlow
cognitive analytics	Google Neural Machine Translator (GNMT)	neural network	Theano
cognitive computing	graphics processing unit (GPU)	neuron	threshold value
cognitive search	hidden layer	overfitting	Torch
connection weight	IBM Watson	perceptron	transfer function
constant error carousel (CEC)	ImageNet	performance function	word embeddings
convolution function	Keras	pooling	word2vec
convolutional neural network (CNN)	long short-term memory (LSTM)	processing element (PE)	
		recurrent neural network (RNN)	

Questions for Discussion

1. What is deep learning? What can deep learning do that traditional machine-learning methods cannot?
2. List and briefly explain different learning paradigms/methods in AI.
3. What is representation learning, and how does it relate to machine learning and deep learning?
4. List and briefly describe the most commonly used ANN activation functions.
5. What is MLP, and how does it work? Explain the function of summation and activation weights in MLP-type ANN.
6. List and briefly describe the nine-step process in conducting a neural network project.
7. Draw and briefly explain the three-step process of learning in ANN.
8. How does the backpropagation learning algorithm work?
9. What is overfitting in ANN learning? How does it happen, and how can it be prevented?
10. What is the so-called black-box syndrome? Why is it important to be able to explain an ANN's model structure?
11. How does sensitivity analysis work in ANN? Search the Internet to find other methods to explain ANN methods.
12. What is meant by "deep" in deep neural networks? Compare deep neural network to shallow neural network.
13. What is GPU? How does it relate to deep neural networks?
14. How does a feedforward multilayer perceptron-type deep network work?
15. Comment on the impact of random weights in developing deep MLP.
16. Which strategy is better: more hidden layers versus more neurons?
17. What is CNN?
18. For what type of applications can CNN be used?

19. What is the convolution function in CNN, and how does it work?
20. What is pooling in CNN? How does it work?
21. What is ImageNet, and how does it relate to deep learning?
22. What is the significance of AlexNet? Draw and describe its architecture.
23. What is GoogLeNet? How does it work?
24. How does CNN process text? What is word embeddings, and how does it work?
25. What is word2vec, and what does it add to the traditional text mining?
26. What is RNN? How does it differ from CNN?
27. What is the significance of *context*, *sequence*, and *memory* in RNN?
28. Draw and explain the functioning of a typical recurrent neural network unit.
29. What is LSTM network, and how does it differ from RNNs?
30. List and briefly describe three different types of LSTM applications.
31. How do Google's Neural Machine Translation and Microsoft Skype Translator work?
32. Despite its short tenure, why do you think deep learning implementation has several different computing frameworks?
33. Define and comment on the relationship between CPU, NVIDIA, CUDA, and deep learning.
34. List and briefly define the characteristics of different deep learning frameworks.
35. What is Keras, and how does it differ from other frameworks?
36. What is cognitive computing and how does it differ from other computing paradigms?
37. Draw a diagram and explain the conceptual framework of cognitive computing. Make sure to include inputs, enablers, and expected outcomes in your framework.
38. List and briefly define the key attributes of cognitive computing.
39. How does cognitive computing differ from ordinary AI techniques?
40. What are the typical use cases for cognitive computing/analytics?
41. What is cognitive computing/analytics? What is cognitive search?
42. What is IBM Watson, and what is its significance to the world of computing?
43. How does IBM Watson work, what are the underlying enablers and their functions?
44. List and briefly explain five of the potential use cases for IBM Watson.

Exercises

Teradata University Network (TUN) and Other Hands-On and Internet Exercises

1. Go to the Teradata University Network Web site (teradatauniversitynetwork.com). Search for teaching and learning materials (e.g., articles, application cases, white papers, videos, exercises) on deep learning, cognitive computing, and IBM Watson. Read the material you have found. If needed, also conduct a search on the Web to enhance your findings. Write a report on your findings.
2. Deep learning is relatively new to the world of analytics. Its application cases and success stories are just starting to emerge in the Web. Conduct a comprehensive search on your school's digital library resources to identify at least five journal articles where interesting deep learning applications are described. Write a report on your findings.
3. Most of the applications of deep learning today are developed using R- and/or Python-based open-source computing resources. Identify those resources (frameworks such as Torch, Caffe, TensorFlow, Theano, Keras) available for building deep learning models and applications. Compare and contrast their capabilities and limitations. Based on your findings and understanding of these resources, if you were to develop a deep learning application, which one would you choose to employ? Explain and justify/defend your choice.
4. *Cognitive computing* has become a popular term to define and characterize the extent of the ability of machines/computers to show "intelligent" behavior. Thanks to IBM Watson and its success on *Jeopardy!*, cognitive computing and cognitive analytics are now part of many real-world intelligent systems. In this exercise, identify at least three application cases where cognitive computing was used to solve complex real-world problems. Summarize your findings in a professionally organized report.
5. Download KNIME analytics platform, one of the most popular free/open-source software tools from knime.org. Identify the deep learning examples (where Keras is used to build some exemplary prediction/classification models) in its example folder. Study the models in detail. Understand what it does and how exactly it does it. Then, using a different but similar data set, build and test your own deep learning prediction model. Report your findings and experiences in a written document.
6. Search for articles related to "cognitive search." Identify at least five pieces of written material (a combination of journal articles, white papers, blog posts, application cases, etc.). Read and summarize your findings. Explain your understanding of cognitive search and how it differs from regular search methods.
7. Go to Teradata.com. Search and find application case studies and white papers on deep learning and/or

cognitive computing. Write a report to summarize your findings, and comment on the capabilities and limitations (based on your understanding) of these technologies.

8. Go to SAS.com. Search and find application case studies and white papers on deep learning and/or cognitive computing. Write a report to summarize your findings, and comment on the capabilities and limitations (based on your understanding) of these technologies.
9. Go to IBM.com. Search and find application case studies and white papers on deep learning and/or cognitive

computing. Write a report to summarize your findings, and comment on the capabilities and limitations (based on your understanding) of these technologies.

10. Go to TIBCO.com or some other advanced analytics company Web site. Search and find application case studies and white papers on deep learning and/or cognitive computing. Write a report to summarize your findings, and comment on the capabilities and limitations (based on your understanding) of these technologies.

References

- Abad, M., P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, ... M. Isard. (2016). TensorFlow: a system for large-scale machine learning. *OSDI*, 16, pp. 265–283.
- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), pp. 589–609.
- Azaria, A. 2022. *ChatGPT Usage and Limitations*
- Bahdanau, D., K. Cho, & Y. Bengio. (2014). “Neural Machine Translation by Jointly Learning to Align and Translate.” ArXiv Preprint ArXiv:1409.0473.
- Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, 2(1), pp. 1–127.
- Bergstra, J., O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, ... Y. Bengio. (2010). “Theano: A CPU and GPU Math Compiler in Python.” *Proceedings of the Ninth Python in Science Conference*, Vol. 1.
- Boureau, Y.-L., N. Le Roux, F. Bach, J. Ponce, & Y. LeCun (2011). “Ask the Locals: Multi-Way Local Pooling for Image Recognition.” *Proceedings of the International Computer Vision (ICCV’11) IEEE International Conference*, pp. 2651–2658.
- Boureau, Y.-L., J. Ponce, & Y. LeCun. (2010). “A Theoretical Analysis of Feature Pooling in Visual Recognition.” *Proceedings of International Conference on Machine Learning (ICML’10)*, pp. 111–118.
- Buchanan, B. G., & E. H. Shortliffe. (1984). *Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison-Wesley.
- Cognitive Computing Consortium. (2018). <https://cognitivecomputingconsortium.com/resources/cognitive-computing-defined/#1467829079735-c0934399-599a> (accessed July 2018).
- Chen, T., M. Li, Y. Li, M. Lin, N. Wang, M. Wang, ... Z. Zhang. (2015). “Mxnet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems.” ArXiv Preprint ArXiv:1512.01274.
- Collobert, R., K. Kavukcuoglu, & C. Farabet. (2011). “Torch7: A Matlab-like Environment for Machine Learning.” Big-Learn, NIPS workshop.
- Cybenko, G. (1989). Approximation by superpositions of a Sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4), 303–314.
- Delen, D., R. Sharda, & M. Bessonov, M. (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. *Accident Analysis & Prevention*, 38(3), 434–444.
- Denyer, S. (2018, January). “Beijing Bets on Facial Recognition in a Big Drive for Total Surveillance.” *The Washington Post*. https://www.washingtonpost.com/news/world/wp/2018/01/07/feature/in-china-facial-recognition-is-sharp-end-of-a-drive-for-total-surveillance/?noredirect=on&utm_term=.e73091681b31.
- Goodfellow, I., Y. Bengio, & A. Courville. (2016). *Deep Learning*. Cambridge, MA: MIT Press.
- Goodfellow, I. J., D. Warde-Farley, P. Lamblin, V. Dumoulin, M. Mirza, R. Pascanu, ... Y. Bengio. (2013). “Pylearn2: A Machine Learning Research Library.” ArXiv Preprint ArXiv: 1308.4214.
- Graves, A. (2013). “Generating Sequences with Recurrent Neural Networks.” ArXiv Preprint ArXiv: 1308.0850.
- Graves, A., & N. Jaitly. (2014). “Towards End-to-End Speech Recognition with Recurrent Neural Networks.” *Proceedings on International Conference on Machine Learning*, pp. 1764–1772.
- Graves, A., N. Jaitly, & A. Mohamed. (2013). “Hybrid Speech Recognition with Deep Bidirectional LSTM.” IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 273–278.
- Graves, A., A. Mohamed, & G. Hinton. (2013). “Speech Recognition with Deep Recurrent Neural Networks.” IEEE Acoustics, Speech and Signal Processing (ICASSP) International Conference, pp. 6645–6649.
- Graves, A., & J. Schmidhuber. (2009). “Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks.” *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, pp. 545–552.
- Gualtieri, M. (2017). “Cognitive Search Is the AI Version of Enterprise Search, Forrester.” go.forrester.com/blogs/17-06-12-cognitive_search_is_the_ai_version_of_enterprise_search/ (accessed July 2018).
- Haykin, S. S. (2009). *Neural Networks and Learning Machines*, 3rd ed. Upper Saddle River, NJ: Prentice Hall.
- He, K., X. Zhang, S. Ren, & J. Sun. (2015). “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.” *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034.

- Hinton, G. E., S. Osindero, & Y.-W. Teh. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7), 1527–1554.
- Hochreiter, S., & J. Schmidhuber (1997) Long short-term memory. *Neural Computation*, 5(8), 1735–1780.
- Hornik, K. (1991). Approximation capabilities of multi-layer feedforward networks. *Neural Networks*, 4(2), 251–257.
- IBM. (2011). “IBM Watson.” www.ibm.com/watson/ (accessed July 2017).
- Jia, Y. (2013) “Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding.” <http://Goo.Gl/Fo9Y08> (accessed June 2018).
- Jia, Y., E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, . . . T. Darrell, T. (2014). “Caffe: Convolutional Architecture for Fast Feature Embedding.” *Proceedings of the ACM International Conference on Multimedia*, pp. 675–678.
- Keyzers, D., T. Deselaers, H. A. Rowley, L.-L. Wang, & V. Carbone. (2017). Multi-language online handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1180–1194.
- Krizhevsky, A., I. Sutskever, & G. Hinton. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097–1105S.
- Kumar, S. (2017). “A survey of deep learning methods for relation extraction.” <http://arxiv.org/abs/1705.03645>. (accessed June 2018)
- LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, & L. D. Jackel. (1989). Backpropagation applied to handwritten ZIP code recognition. *Neural Computation*, 1(4), 541–551.
- Liang, X., X. Shen, J. Feng, L. Lin, & S. Yan. (2016). “Semantic object parsing with graph LSTM.” *European Conference on Computer Vision*. New York, NY: Springer, pp. 125–143.
- Lucy, L., and Bamman, D. 2021. “Gender and Representation Bias in GPT-3 Generated Stories,” in *Proceedings of the Third Workshop on Narrative Understanding*, pp. 48–55.
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Mahajan, D., R. Girshick, V. Ramanathan, M. Paluri, & L. van der Maaten. (2018). Advancing state-of-the-art image recognition with deep learning on hashtags. <https://code.facebook.com/posts/1700437286678763/advancing-state-of-the-art-image-recognition-with-deep-learning-on-hashtags/> (accessed June 2022).
- Mikolov, T., K. Chen, G. Corrado, & J. Dean. (2013). “Efficient Estimation of Word Representations in Vector Space.” ArXiv Preprint ArXiv:13013781.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. ArXiv Preprint ArXiv:1301.3781.
- Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, & J. Dean. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.
- Mintz, M., S. Bills, R. Snow, & D. Jurafsky. (2009) “Distant Supervision for Relation Extraction Without Labeled Data.” *Proceedings of the joint Conference of the Forty-Seventh Annual Meeting of the Association for Computational Linguistics and the Fourth International joint Conference on Natural Language Processing of the AFNLP*, Vol. 2, pp. 1003–1011
- Mozur, P. (2018, June 8). “Inside China ‘s Dystopian Dreams: A.I., Shame and Lots of Cameras.” *The New York Times*, issue June 8, <yr>2018.
- Nguyen, T. H., & R. Grishman. (2015). “Relation Extraction: Perspective from Convolutional Neural Networks.” *Proceedings of the First Workshop on Vector Space Modeling for Natural Language Processing*, pp. 39–48.
- Olson, D. L., D. Delen, and Y. Meng. (2012) Comparative analysis of data mining models for bankruptcy prediction. *Decision Support Systems*, 52(2), pp. 464–473.
- OpenAI (2023). OpenAIBlog post on ChatGP. Accessed at <https://openai.com/blog/chatgpt> (accessed March 2023).
- Principe, J. C., N. R. Euliano, and W. C. Lefebvre. (2000). *Neural and Adaptive Systems: Fundamentals Through Simulations*. New York: Wiley.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., and Ray, A. 2022. “Training Language Models to Follow Instructions with Human Feedback,” *ArXiv Preprint ArXiv:2203.02155*.
- Reynolds, H., & S. Feldman. (2014, July/August). “Cognitive Computing: Beyond the Hype.” *KM World*, 23(7), p. 21.
- Riedel, S., L. Yao, & A. McCallum. (2010). “Modeling Relations and Their Mentions Without Labeled Text.” *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. New York, NY: Springer, pp. 148–163.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). “Why should I trust you?” Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.
- Royyuru, A (2014). “IBM ‘s Watson Takes on Brain Cancer: Analyzing Genomes to Accelerate and Help Clinicians Personalize Treatments.” Thomas J. Watson Research Center, www.research.ibm.com/articles/genomics.shtml (accessed September 2014).
- Rumelhart, D. E., G. E. Hinton, & R. J. Williams. (1986). “Learning Representations by Back-Propagating Errors.” *Nature*, 323(6088), pp. 533
- Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, . . . M. Bernstein. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252.
- Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., . . . & Tarantola, S. (2008). *Global Sensitivity Analysis: The Primer*. John Wiley & Sons.
- Saltelli, A., Aleksankina, K., Becker, W., Fennell, P., Ferretti, F., Holst, N., . . . & Wu, Q. (2019). Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental Modelling & Software*, 114, 29–39.

- Sato, K., C. Young, & D. Patterson. (2017). "An In-Depth Look at Google 's First Tensor Processing Unit (TPU)." <https://cloud.google.com/blog/big-data/2017/05/an-indepth-look-at-googles-first-tensor-processing-unit-tpu>. (accessed June 2018)
- Scherer, D., A. Muller, & S. Behnke. (2010). "Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition." *International Conference on Artificial Neural Networks*. New York, NY: Springer, 92–101.
- Sharda, R., Delen, D., & Turban, E. (2021). *Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support*. New York, NY: Pearson Education Limited.
- Slowey, L. (2017a, January 25). "Winning the Best Picture Oscar: IBM Watson and Winning Predictions." <https://www.ibm.com/blogs/internet-of-things/best-picture-oscar-watson-predicts/> (accessed August 2018).
- Slowey, L. (2017b, May 10). Watson predicts the winners: Eurovision 2017. <https://www.ibm.com/blogs/internet-of-things/eurovision-watson-tone-predictions/> (accessed August 2018).
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). "Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems*, pp. 3104–3112.
- Topuz, K., & Delen, D. (2021). A probabilistic Bayesian inference model to investigate injury severity in automobile crashes. *Decision Support Systems*, 150, 113557.
- Ung, G. M. (2016, May). Google's tensor processing unit could advance Moore 's Law 7 Years into the future. *PC World*. <https://www.pcworld.com/article/3072256/google-io/googles-tensor-processing-unit-said-to-advance-moores-law-seven-years-into-the-future.html> (accessed July 2018).
- Vinyals, O., Kaiser, L., Koo, T. Petrov, S., Sutskever, I., & Hinton, G. (2015) "Grammar As a Foreign Language." *Advances in Neural Information Processing Systems*, pp. 2773–2781.
- Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). "Show and Tell: A Neural Image Caption Generator." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3156–3164
- Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2017). "Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge." *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), 652–663.
- Wilson, R. L., & R. Sharda. (1994). "Bankruptcy Prediction Using Neural Networks." *Decision Support Systems*, 11(5), 545–557. .
- Wu, Y., M. Schuster, Z. Chen, Q. V Le, M. Norouzi, W. Macherey, & K. Macherey. (2016). "Google 's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation." ArXiv Preprint ArXiv:1609.08144
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., & Bengio, Y. (2015). "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." *Proceedings of the Thirty-Second International Conference on Machine Learning*, pp. 2048–2057.
- Zeng, D., Liu, K., Lai, S., Zhou, G., & Zhao (2014). Relation classification via convolutional deep neural network. <http://doi.org/http://aclweb.org/anthology/JC/C14/C14-1220.pdf>. (accessed June 2018).

Prescriptive Analytics: Optimization and Simulation

LEARNING OBJECTIVES

- Understand the applications of prescriptive analytics techniques in combination with reporting and predictive analytics
- Understand the basic concepts of analytical decision modeling
- Understand the concepts of analytical models for selected decision problems, including linear programming and simulation models for decision support
- Describe how spreadsheets can be used for analytical modeling and solutions
- Explain the basic concepts of optimization and when to use them
- Describe how to structure a linear programming model
- Explain what is meant by sensitivity analysis, what-if analysis, and goal seeking
- Understand the concepts and applications of different types of simulation
- Understand potential applications of discrete event simulation
- Understand basic concepts and applications of genetic algorithms

This chapter extends the analytics applications beyond reporting and predictive analytics. It includes coverage of selected techniques that can be employed in combination with predictive models to help support decision-making. We focus on techniques that can be implemented relatively easily using either spreadsheet tools or by using stand-alone software tools. Of course, there is much additional detail to be learned about management science models, but the objective of this chapter is to simply illustrate what is possible and how it has been implemented in real settings.

We present this material with a note of caution: Modeling can be a difficult topic and is as much an art as it is a science. The purpose of this chapter is not necessarily for you to *master the topics* of modeling and analysis. Rather, the material is geared toward *gaining familiarity* with the important concepts as they relate to prescriptive analytics and their use in decision-making. It is important to recognize that the modeling we discuss here is only cursorily related to the concepts of data modeling. You should not confuse

the two. We walk through some basic concepts and definitions of decision modeling. We next introduce the idea of modeling directly in spreadsheets. We then discuss the structure and application of two successful time-proven models and methodologies: linear programming and simulation. As noted earlier, one could take multiple courses just in these two topics, but our goal is to give you a sense of what is possible. This chapter includes the following sections:

- 8.1** Opening Vignette: Balancing Delivery Routes, Production Schedules, and Inventory 442
- 8.2** Model-Based Decision-Making 444
- 8.3** Structure of Mathematical Models for Decision Support 450
- 8.4** Certainty, Uncertainty, and Risk 452
- 8.5** Decision Modeling with Spreadsheets 453
- 8.6** Mathematical Programming Optimization 458
- 8.7** Multiple Goals, Sensitivity Analysis, What-If Analysis, and Goal Seeking 467
- 8.8** Decision Analysis with Decision Tables and Decision Trees 471
- 8.9** Introduction to Simulation 473
- 8.10** Genetic Algorithms and Developing GA Applications 480

8.1 OPENING VIGNETTE: Balancing Delivery Routes, Production Schedules, and Inventory

PROBLEM

Praxair/Linde is one of the world's largest industrial gas and engineering companies. They have supplied a wide variety of gas and chemical products to manufacturers for more than a century. Their products are distributed all over the world across countless industries, from hospitals to semiconductor manufacturing to clean fuel vehicles. Their distribution network and product inventory balancing decisions must be made regularly.

White Martins, a subsidiary of Linde is the largest industrial gas company in Brazil and is part of a worldwide team managing various cryogenics, like argon, oxygen, nitrogen, and carbon dioxide. All their product data are maintained in a data lake with two primary concerns: transactional data and inventory management.

Marcos Guimaraes is a mechanical engineer who has worked with industrial gas and oil for more than 40 years. He manages the bulk distribution operations in South America for White Martins. As the Logistics Director, he is accountable for all bulk distribution operations, managing customer inventory, and assuring on-time product delivery without any product outages.

Guimaraes must make product allocation, truck routing, and cost of service calculations. His operation includes a fleet of 650 vehicles with more than 1,000 drivers who distribute 150,000 tons of product per month to thousands of customers across Brazil. This leads to large complex models that include uncertain demand. This decision led him to schedule and coordinate 18,000 delivery trips per month.

He had been using an optimization solution for product distribution, but it managed product allocations without any routing capabilities. The dimensions of his data set grew so large that the models were too complex for his tools. He wanted to build a new model to grow and perform very well, managing orders, product allocation, and distribution in one place.

SOLUTION APPROACH

Guimaraes has a goal of running White Martins' logistics as a world-class operation, bringing value to customers in their supply chain using state of the art tools. Though he has access to industry-leading products for data visualization, he needed a different tool for making optimization decisions based on real data. He chose Analytic Solver, a larger extension of the Solver built into Microsoft Excel that we will learn about in this chapter. Guimaraes notes, "Analytic Solver's optimization capabilities provide routing and allocation of product to each customer for each plant. The system performs daily and allows us to forecast three months into the future. It helps us understand how to balance inventory across nearby plants."

Though most client product demand can stay within a predictable variance, there are times when clients forecast a need for significant increase soon.

Using a large data model built on historical data, "Analytic Solver shows how our inventory will look in the future. We use demand and consumption as uncertain variables to understand the minimum and maximum levels of inventory we'd like to have. Analytic Solver will calculate how much and whether to transfer product from one plant to another."

Industrial gas plants are made possible by mechanical operations. The costs of running a plant are not static because there are realities that need to be considered in terms of maintenance, wear and tear, and runtime.

Guimaraes and his team use the model to understand these costs and risks. "Once we have a model in place, it's shared with the business executives so they can understand the costs of running the plants and delivering product. These are mechanical systems, so they can't realistically run 100% of the time. We use Analytic Solver to run a risk calculation of the plant having breakages."

Aspects of the models employed by Guimaraes, and his team are kept simple due to historical certainty in the data they use. Analytic Solver manages uncertain variables for reliability, demand, and other human factors. "Demand is a human uncertainty, and we could see failure with a 1% variance. We need calibration to be as close to perfect as possible." Running such a large operation requires operational rigor and the capacity for balancing inventory. Guimaraes relies on Analytic Solver every day for his role.

RESULTS

This model is essential to White Martins for managing product supply, understanding production performance, and tracking merchant pickup. All these factors play into models Guimaraes' team solve to guide plant operations managers' decision-making. "To increase inventory, we need to know whether to vaporize product to accommodate a need or physically bring product to a plant."

► QUESTIONS FOR DISCUSSION

1. What types of objectives and constraints one might face in planning timely and profitable delivery of industrial chemicals to customers?
2. Which other companies might employ similar industrial decision models?

LESSONS LEARNED FROM THIS VIGNETTE

Modeling supply chain operations was one of the early applications of linear programming models. As you will see in this chapter, such models can help us optimize many decisions where we have an objective to maximize (such as profits) or minimize (such as costs) but we face many constraints. Solving these models of different sizes in a reasonable amount of time has been a major area of research since the Second World War.

Technology has evolved to a point where many problems can be easily solved using the built-in solver in Microsoft Excel that is provided by Frontline Systems. Even larger problems can be solved by employing their extended Analytic Solver.

Source: Adapted from a contribution by Daniel Fylstra, FrontLine Systems., Included with permission.

8.2 MODEL-BASED DECISION-MAKING

As the preceding vignette indicates, making decisions using analytical model is what we call prescriptive analytics. In the last several chapters, we have learned the value and the process of knowing the history of what has been going on and use that information to also predict what is likely to happen. However, we go through that exercise to determine what we should do next. This might entail deciding which customers are likely to buy from us and making an offer or giving a price point that will maximize the likelihood that they would buy, and our profit would be optimized. Conversely, it might involve being able to predict which customer is likely to go somewhere else and making a promotion offer to retain them as a customer and optimize our value. We may need to make decisions on awarding contracts to our vendors to make sure all our needs are covered, and the costs are minimized. We could be facing a situation of deciding which prospective customers should receive what promotional campaign material so that our cost of promotion is not outrageous, and we maximize the response rate while managing within a budget. We may be deciding how much to pay for different paid search keywords to maximize the return on investment of our advertising budget. In another setting, we may have to study the history of our customers' arrival patterns and use that information to predict future arrival rates and apply that to schedule an appropriate number of store employees to maximize customer responses and optimize our labor costs. We could be deciding where to locate our warehouses based on our analysis and prediction of demand for our products and the supply chain costs. We could be setting daily delivery routes based on product volumes to be delivered at various locations and the delivery costs and vehicle availability. One can find hundreds of examples of situations where data-based decisions are valuable. Indeed, the biggest opportunity for the growing analytics profession is the ability to use descriptive and predictive insights to help a decision maker make better decisions. Although there are situations where one can use experience and intuition to make decisions, it is more likely that a decision supported by a model will help a decision maker make better decisions. In addition, it also provides decision makers with justification for what they are recommending. Thus, prescriptive analytics has emerged as the next frontier for analytics. It essentially involves using an analytical model to help guide a decision maker in planning or automating the decision process so that a model can make recommendations or decisions. Because the focus of prescriptive analytics is on making recommendations or making decisions, some call this category of analytics decision analytics.

INFORMS publications, such as *Interfaces*, *ORMS Today*, and *Analytics* magazine, all include stories that illustrate successful applications of decision models in real settings. This chapter includes many examples of such prescriptive analytic applications. Applying models to real-world situations can save millions of dollars or generate millions of dollars in revenue. Christiansen et al. (2009) describe the applications of such models in shipping company operations using TurboRouter, a decision support system (DSS) for ship routing and scheduling. They claim that over the course of a 3-week period, a company used this model to better utilize its fleet, generating additional profit of \$1 – 2 million in such a short time. We provide another example of a model application in Analytics in Action 8.1 that illustrates a sports application.

ANALYTICS IN ACTION 8.1 Canadian Football League Optimizes Game Schedule

Canadian Football League (CFL) is Canada's equivalent of the U.S. National Football League (NFL). It had a challenge of organizing 81 football games for 9 teams over a period of 5 months optimally while stabilizing matching priorities for sales revenue, television ratings, and the team rest days. Other considerations include organizing games over different time zones and the main rivalry games to be held on major public holidays. For any league, a robust schedule is a driving force for a variety of business collaborations, such as coordinating with broadcasting channels and organizing ground ticket sales. If the schedule is not optimized, it would directly hamper the promotions thus resulting in a huge loss of revenue and bad channel ratings. CFL used to create match schedules manually and hence had to figure out finer ways to improve their schedules, taking all the constraints into account. They had tried to work with a consultant to build a comprehensive model for scheduling, but the implementation remained a challenge. The League decided to tackle the issue with the Solver available within Microsoft Excel. Some of the matching priorities to be balanced while optimizing the schedule were:

1. Sales Revenue—Setting a schedule with matches and time slots to those clubs that generate more revenue.
2. Channel Ratings—Setting a schedule with games that would improve channel ratings for the broadcasting company.
3. Team Rest Days—Setting a schedule with the two teams playing against each other having enough rest days.

The league decided to improve the match schedules by giving the player rest days as a higher priority, followed by sales revenue and channel scores for the broadcasting company. This is mainly because the sales revenue and channel scores are a byproduct of team players' performance on the field, which is directly related to the rest days taken by the teams.

Methodology/Solution

Initially, organizing schedules was a huge task to perform on Excel through the built-in Solver feature. Frontline systems provided a premium version for Solver, which allowed the model size to grow from about 200 decisions to 8,000 decisions. The League had to even add in more industry-specific constraints such as telecasting across different time zones, double-header games cannot be overlapped, and arch rival games to be scheduled on Labor Day. Added limitations were never simple until the Frontline Systems consultants stepped up to help CFL turn this nonlinear problem into a linear problem. The linear programming “engine” got the model running. Premium Solver software turned out to be of great help to get an improved schedule.

Results/Benefits

Using the optimized schedule would lead to increased revenue through higher ticket sales and higher TV scores for the broadcasting channels. This was achieved because the tool was able to support added constraints of the vendors with great ease. The optimized schedule pleased most of the league's stakeholders. This is a repetitive process, but those match schedules were CFL's most advanced season match schedules to date.

QUESTIONS FOR DISCUSSION

1. List three ways in which Solver-based scheduling of games could result in more revenue as compared to the manual scheduling.
2. In what other ways can CFL leverage the Solver software to expand and enhance their other business operations?
3. What other considerations could be important in scheduling such games?

What Can We Learn from This Analytics In Action?

By using the Solver add-in for Excel, the CFL made better decisions in scheduling their games by taking stakeholders and industry constraints into consideration, leading to revenue generation and good channel ratings. Thus, an optimized schedule, a purview of prescriptive analytics,

derived significant value. According to the case study, the modeler, Mr Trevor Hardy, was an expert Excel user, but not an expert in modeling. However, the ease of use of Excel permitted him to develop a practical application of prescriptive analytics.

Compiled from “Canadian Football League Uses Frontline Solvers to Optimize Scheduling in 2016.” Solver, September 7, 2016, www.solver.com/news/canadian-football-league-uses-frontline-solvers-optimize-scheduling-2016 (accessed July 2022); Kostuk, Kent J., and Keith A. Willoughby. “A Decision Support System for Scheduling the Canadian Football League.” *Interfaces*, vol. 42, no. 3, 2012, pp. 286–295; Dilkina, Bistra N., and William S. Havens. The U.S. National Football League Scheduling Problem. Intelligent Systems Lab, www.cs.cornell.edu/~bistra/papers/NFLsched1.pdf (accessed July 2022).

Prescriptive Analytics Model Examples

Modeling is a key element for prescriptive analytics. In the examples mentioned earlier in the introduction and analytics in actions, one must employ a mathematical model to be able to recommend a decision for any realistic problem. For example, deciding which customers (among potentially millions) will receive what offer to maximize the overall response value but staying within a budget is not something you can do manually. Building a probability-based response maximization model with the budget as a constraint would give us the information we are seeking. Depending on the problem we are addressing, there are many classes of models, and there are often many specialized techniques for solving each one. We will learn about two different modeling methods in this chapter. Most universities have multiple courses that cover these topics under titles such as Operations Research, Management Science, Decision Support Systems, and Simulation that can help you build more expertise in these topics. Because prescriptive analytics typically involves the application of mathematical models, sometimes the term *data science* is more commonly associated with the application of such mathematical models. Before we learn about mathematical modeling support in prescriptive analytics, let us understand some modeling issues first.

Identification of the Problem and Environmental Analysis

No decision is made in a vacuum. It is important to analyze the scope of the domain and the forces and dynamics of the environment. A decision maker needs to identify the organizational culture and the corporate decision-making processes (e.g., who makes decisions, degree of centralization). It is entirely possible that environmental factors have created the current problem. This can formally be called **environmental scanning and analysis**, which is the monitoring, scanning, and interpretation of collected information. Business intelligence/business analytics (BI/BA) tools can help identify problems by scanning for them. The problem must be understood, and everyone involved should share the same frame of understanding because the problem will ultimately be represented by the model in one form or another. Otherwise, the model will not help the decision maker.

VARIABLE IDENTIFICATION Identification of a model’s variables (e.g., decision, result, uncontrollable) is critical, as are the relationships among the variables. Influence diagrams, which are graphical models of mathematical models, can facilitate the identification process. A more general form of an influence diagram, a cognitive map, can help a decision maker develop a better understanding of a problem, especially of variables and their interactions.

FORECASTING (PREDICTIVE ANALYTICS) As we have noted previously, an important prerequisite of prescriptive analytics is knowing what has happened and what is likely to happen. This form of predictive analytics is essential for construction and manipulating

models because when a decision is implemented, the results usually occur in the future. There is no point in running a what-if (sensitivity) analysis on the past because decisions made then have no impact on the future. Online commerce and communication have created an immense need for **forecasting** and an abundance of available information for performing it. These activities occur quickly, yet information about such purchases is gathered and should be analyzed to produce forecasts. Part of the analysis involves simply predicting demand; however, forecasting models can use product life cycle needs and information about the marketplace and consumers to analyze the entire situation, ideally leading to additional sales of products and services.

We describe an effective example of such forecasting and its use in decision-making in making refinery purchase applications in Analytics in Action 8.2.

ANALYTICS IN ACTION 8.2 Refinery Modeling

Problem

Downstream Advisors (DA) is a professional consulting firm specializing in the technical, commercial, and financial aspects of the downstream petroleum industries, such as refining, marketing, transportation, petrochemicals, and electric power. When companies are interested in buying, selling, or optimizing their oil and gas refineries, Downstream Advisors and its expert consultants are asked to evaluate and simulate various refinery functions. DA has built a refinery model called Smart Ref, which can be customized to build a model of a refinery a company already has or is considering acquiring. The most important aspect of their analysis is the flow rate between various processes and determining what needs to go where. For example, they may determine how and where to route crude products depending on fractionation and processing requirements. Consultant William Peltier remarks, “With more than 50 interlocked process units, determining which products need to move from one place to another gets pretty complicated.”

It took a decade to build Smart Ref, and it has been used on client work for several years. Consultants at DA can run multiple use cases and model outcomes based on a range of probabilities. These types of modeling exercises are exactly what investors need when they consider purchasing refineries and pipelines. Smart Ref enables these investors to understand their investment during the due diligence process. Downstream Advisors also has the capability of comparing existing refinery performance to other refineries for competitive analysis, or for understanding tax scenarios.

Modeling and Solution Approach

Downstream Advisors use Analytic Solver developed by Frontline Systems to solve the refinery models. Frontline System’s Solver is built-in Microsoft Excel. In this chapter, we will learn how to model and solve some optimization problems using this Solver. Analytic Solver has allowed DA to build and automate data models they have worked on for decades, even for small projects like gasoline blenders. The Solver allows DA consultants to simulate the refinery operations and make recommendations to an investor. Consultant Peltier remarks:

“They may want to spend more or less money for different benefits, or they may have a different level of the approval process depending on where they’re investing.”

Though the free Solver for Excel is useful for DA consultants, there are much larger problems to solve using Analytic Solver. “There are two kinds of models commonly used in the refinery industry: a linear programming model that’s very specific and tries to do what we do, and a process model that’s much more detailed,” notes Peltier. “It’s hard to change the linear model and we have a lot of nonlinearity—sometimes in ways that we don’t know how to automate.”

The process model operates at the molecular level and is used for designing new equipment. Smart Ref’s use of Analytic Solver enables this type of work and at high speeds. All

refinery modeling is run on each consultant's own computers. The solver can provide an optimal solution for their problems in seconds.

The biggest application with the most variables is the distribution decision-making. The biggest application with the most variables is the distribution decision-making. Consultant Davis continues, "Solver figures out what goes where. We know what products will meet which specifications, like viscosities or boiling points."

Solver Enables Downstream Advisors to Do Business

Downstream Advisors has a significant competitive advantage with its innovative Smart Ref application. Peltier says, "There was nothing like this before we developed it. We wondered about 20-25 years ago if it could even be built. So we experimented with a process to build a model and got serious about making it useful for many things. And Solver is what makes that product possible." It allows Smart Ref to produce a robust simulation of real-world physical processes with optimized results delivered in seconds.

QUESTIONS FOR DISCUSSION

1. Discuss how a refinery's operation may involve building a model to simulate its operation. Conduct an online search to find some papers that discuss this aspect.
2. Why would prospective investors be interested in studying the results of such a model?

Lessons Learned from This Story

Modeling refinery operations was one of the early applications of linear programming models. As you will see in this chapter, such models can help us optimize many decisions where we have an objective to maximize (such as profits) or minimize (such as costs) but we face many constraints. Solving these models of different sizes in a reasonable amount of time has been a major area of research since World War II. Technology has evolved to a point where many problems can be easily solved using the built-in solver in Microsoft Excel that is provided by Frontline Systems. Even larger problems can be solved by employing their Analytic Solver.

Source: Adapted from a contribution by Daniel Fylstra, Front Line Systems. <https://www.frontlinesystems.com/news/downstream-advisors-analytic-solver-case-study> (accessed July 2022)

Model Categories

Table 8.1 classifies some decision models into seven groups and lists several representative techniques for each category. Each technique can be applied to either a **static** or a **dynamic model**, which can be constructed under assumed environments of certainty, uncertainty, or risk. To expedite model construction, we can use special decision analysis systems that have modeling languages and capabilities embedded in them. These include spreadsheets, data mining systems, online analytic processing (OLAP) systems, and modeling languages that help an analyst build a model. We will introduce one of these systems later in the chapter.

MODEL MANAGEMENT Models, like data, must be managed to maintain their integrity, and thus their applicability. Such management is done with the aid of model-based management systems, which are analogous to database management systems (DBMS).

KNOWLEDGE-BASED MODELING DSS uses mostly quantitative models, whereas expert systems use qualitative, knowledge-based models in their applications. Some knowledge is necessary to construct solvable (and therefore usable) models. Many of the predictive analytics techniques, such as classification and clustering, can be used in building knowledge-based models.

TABLE 8.1 Categories of Models

Category	Process and Objective	Representative Techniques
Optimization of problems with few alternatives	Find the best solution from a small number of alternatives	Decision tables, decision trees, analytic hierarchy process
Optimization via algorithm	Find the best solution from a large number of alternatives, using a step-by-step improvement process	Linear and other mathematical programming models, network models
Optimization via an analytic formula	Find the best solution in one step, using a formula	Some inventory models
Simulation	Find a good enough solution or the best among the alternatives checked, using experimentation	Several types of simulation
Heuristics	Find a good enough solution, using rules	Heuristic programming, expert systems
Predictive models	Predict the future for a given scenario	Forecasting models, Markov analysis
Other models	Solve a what-if case, using a formula	Financial modeling, waiting lines

CURRENT TRENDS IN MODELING One recent trend in modeling involves the development of model libraries and solution technique libraries. Some of these codes can be run directly on the owner's Web server for free, and others can be downloaded and run on a local computer. The availability of these codes means that powerful optimization and simulation packages are available to decision makers who may have only experienced these tools from the perspective of classroom problems. For example, the Wisconsin Institute for Discovery at the University of Wisconsin-Madison maintains the NEOS Server for Optimization at <https://neos-server.org/neos/index.html> (accessed July 2022). You can find links to other sites by clicking the Resources link at [informs.org](https://www.informs.org), the Web site of the Institute for Operations Research and the Management Sciences (INFORMS). A wealth of modeling and solution information is also available from INFORMS.

There is a clear trend toward developing and using cloud-based tools and software to access and even run software to perform modeling, optimization, simulation, and so on. This has, in many ways, simplified the application of many models to real-world problems. However, to use models and solution techniques effectively, it is necessary to truly gain experience through developing and solving simple ones. This aspect is often overlooked. Organizations that have key analysts who understand how to apply models indeed apply them very effectively. This is most notably occurring in the revenue management area, which has moved from the province of airlines, hotels, and automobile rentals to retail, insurance, entertainment, and many other areas. CRM also uses models, but they are often transparent to the user. With management models, the amount of data and model sizes are quite large, necessitating the use of data warehouses to supply the data and parallel computing hardware to obtain solutions in a reasonable time frame.

There is a continuing trend toward making analytics models completely transparent to the decision maker. For example, **multidimensional analysis (modeling)** involves data analysis in several dimensions. In multidimensional analysis (modeling), data are generally shown in a spreadsheet format, with which most decision makers are familiar. Many decision makers accustomed to slicing and dicing data cubes are now using OLAP systems that access data warehouses. Although these methods may make modeling palatable, they also eliminate many important and applicable model classes from consideration, and they eliminate some important and subtle solution interpretation aspects. Modeling involves much more than data analysis with trend lines and establishing relationships with statistical methods.

There is also a trend to build a model of a model to help in its analysis. An **influence diagram** is a graphical representation of a model; that is, a model of a model. Some influence diagram software packages can generate and solving the resultant model.

SECTION 8.2 REVIEW QUESTIONS

1. List three lessons learned from modeling.
2. List and describe the major issues in modeling.
3. What are the major types of models used in DSS?
4. Why are models not used in industry as frequently as they should or could be?
5. What are the current trends in modeling?

8.3 STRUCTURE OF MATHEMATICAL MODELS FOR DECISION SUPPORT

In the following sections, we present the topics of analytical mathematical models (e.g., mathematical, financial, and engineering). These include the components and the structure of models.

The Components of Decision Support Mathematical Models

All **quantitative models** are typically made up of four basic components (see Figure 8.1): result (or outcome) variables, decision variables, uncontrollable variables (and/or parameters), and intermediate result variables. Mathematical relationships link these components together. In nonquantitative models, the relationships are symbolic or qualitative. The results of decisions are determined based on the decision made (i.e., the values of the decision variables), the factors that cannot be controlled by the decision maker (in the environment), and the relationships among the variables. The modeling process involves identifying the variables and relationships among them. Solving a model determines the values of these and the result variable(s).

RESULT (OUTCOME) VARIABLES **Result (outcome) variables** reflect the level of effectiveness of a system; that is, they indicate how well the system performs or attains its goal(s). These variables are outputs. Examples of result variables are shown in Table 8.2. Result variables are considered *dependent variables*. Intermediate result variables are sometimes used in modeling to identify intermediate outcomes. In the case of a dependent variable, another event must occur first before the event described by the variable can occur. Result variables depend on the occurrence of the decision variables and the uncontrollable variables.

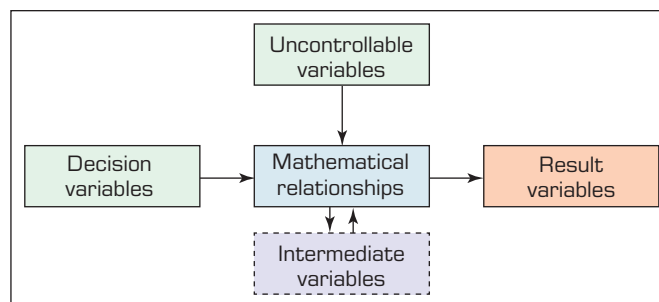


FIGURE 8.1 The General Structure of a Quantitative Model.

TABLE 8.2 Examples of Components of Models

Area	Decision Variables	Result Variables	Uncontrollable Variables and Parameters
Financial investment	Investment alternatives and amounts	Total profit, risk Rate of return on investment (ROI) Earnings per share Liquidity level	Inflation rate Prime rate Competition
Marketing	Advertising budget Where to advertise	Market share Customer satisfaction	Customer's income Competitor's actions
Manufacturing	What and how much to produce Inventory levels Compensation programs	Total cost Quality level Employee satisfaction	Machine capacity Technology Materials prices
Accounting	Use of computers Audit schedule	Data processing cost Error rate	Computer technology Tax rates Legal requirements
Transportation	Shipments schedule Use of smart cards	Total transport cost Payment float time	Delivery distance Regulations
Services	Staffing levels	Customer satisfaction	Demand for services

DECISION VARIABLES **Decision variables** describe alternative courses of action. The decision maker controls the decision variables. For example, for an investment problem, the amount to invest in bonds is a decision variable. In a scheduling problem, the decision variables are people, times, and schedules. Other examples are listed in Table 8.2.

UNCONTROLLABLE VARIABLES, OR PARAMETERS In any decision-making situation, there are factors that affect the result variables but are not under the control of the decision maker. Either these factors can be fixed, in which case they are called **uncontrollable variables**, or **parameters**, or they can vary, in which case they are called *variables*. Examples of factors are the prime interest rate, a city's building code, tax regulations, and utilities costs. Most of these factors are uncontrollable because they are in and determined by elements of the system environment in which the decision maker works. Some of these variables limit the decision maker and therefore form what are called *constraints* of the problem.

INTERMEDIATE RESULT VARIABLES **Intermediate result variables** reflect intermediate outcomes in mathematical models. For example, in determining machine scheduling, spoilage is an intermediate result variable, and total profit is the result variable (i.e., spoilage is one determinant of total profit). Another example is employee salaries. This constitutes a decision variable for management: It determines employee satisfaction (i.e., intermediate outcome), which, in turn, determines the productivity level (i.e., final result).

The Structure of Mathematical Models

The components of a quantitative model are linked by mathematical (algebraic) expressions—equations or inequalities.

A very simple financial model is

$$P = R - C$$

where P = profit, R = revenue, and C = cost. This equation describes the relationship among the variables. Another well-known financial model is the simple present-value cash flow model, where P = present value, F = a future single payment in dollars, i = interest rate (percentage), and n = number of years. With this model, it is possible to determine the present value of a payment of \$100,000 to be made 5 years from today, at a 10% (0.1) interest rate, as follows:

$$P = 100,000 / (1 + 0.1)^5 = 62,092$$

We present more interesting and complex mathematical models in the following sections.

► SECTION 8.3 REVIEW QUESTIONS

- 1. What is a decision variable?
- 2. List and briefly discuss the major components of a quantitative model.
- 3. Explain the role of intermediate result variables.

8.4 CERTAINTY, UNCERTAINTY, AND RISK

The¹ decision-making process involves evaluating and comparing alternatives. During this process, it is necessary to predict the future outcome of each proposed alternative. Decision situations are often classified based on what the decision maker knows (or believes) about the forecasted results. We customarily classify this knowledge into three categories (see Figure 8.2), ranging from complete knowledge to complete ignorance:

- Certainty
- Uncertainty
- Risk

When we develop models, any of these conditions can occur, and different kinds of models are appropriate for each case. Next, we discuss both the basic definitions of these terms and some important modeling issues for each condition.

Decision-Making under Certainty

In decision-making under **certainty**, it is *assumed* that complete knowledge is available so that the decision maker knows exactly what the outcome of *each course of action* will be (as in a deterministic environment). It may not be true that the outcomes are 100% known, nor is it necessary to really evaluate *all* the outcomes, but often this assumption simplifies the model and makes it tractable. The decision maker is viewed as a perfect predictor of the future because it is assumed that there is only one outcome for each alternative.

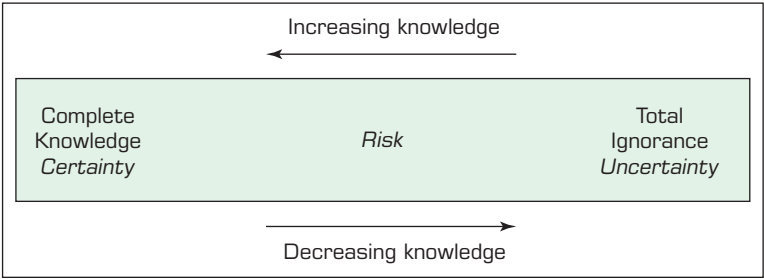


FIGURE 8.2 The Zones of Decision-Making.

¹Some parts of the original versions of these sections were adapted from Turban and Meredith (1994).

For example, the alternative of investing in U.S. Treasury bills is one for which there is complete availability of information about the future return on investment if it is held to maturity. A situation involving decision-making under certainty occurs most often with structured problems and short time horizons (up to 1 year). Certainty models are relatively easy to develop and solve, and they can yield optimal solutions. Many financial models are constructed under assumed certainty, even though the market is anything but 100% certain.

Decision-Making under Uncertainty

In decision-making under **uncertainty**, the decision maker considers situations in which several outcomes are possible for each course of action. In contrast to the risk situation, in this case, the decision maker does not know, or cannot estimate, the probability of occurrence of the possible outcomes. Decision-making under uncertainty is more difficult than decision-making under certainty because there is insufficient information. Modeling of such situations involves assessment of the decision maker's (or the organization's) attitude toward risk.

Managers attempt to avoid uncertainty as much as possible, even to the point of assuming it away. Instead of dealing with uncertainty, they attempt to obtain more information so that the problem can be treated under certainty (because it can be “almost” certain) or under calculated (i.e., assumed) risk. If more information is not available, the problem must be treated under a condition of uncertainty, which is less definitive than the other categories.

Decision-Making under Risk (Risk Analysis)

A decision made under **risk**² (also known as a *probabilistic* or *stochastic* decision-making situation) is one in which the decision maker must consider several possible outcomes for each alternative, each with a given probability of occurrence. The long-run probabilities that the given outcomes will occur are assumed to be known or can be estimated. Under these assumptions, the decision maker can assess the degree of risk associated with each alternative (called *calculated* risk). Most major business decisions are made under assumed risk. **Risk analysis** (i.e., calculated risk) is a decision-making method that analyzes the risk (based on assumed known probabilities) associated with different alternatives. Risk analysis can be performed by calculating the expected value of each alternative and selecting the one with the best expected value.

SECTION 8.4 REVIEW QUESTIONS

1. Define what it means to perform decision-making under assumed certainty, risk, and uncertainty.
2. How can decision-making problems under assumed certainty be handled?
3. How can decision-making problems under assumed uncertainty be handled?
4. How can decision-making problems under assumed risk be handled?

8.5 DECISION MODELING WITH SPREADSHEETS

Models can be developed and implemented in a variety of programming languages and systems. We focus primarily on *spreadsheets* (with their add-ins), modeling languages, and transparent data analysis tools. With their strength and flexibility, spreadsheet packages were quickly recognized as easy-to-use implementation software for the development of a wide range of applications in business, engineering, mathematics, and science.

²Our definitions of the terms *risk* and *uncertainty* were formulated by F. H. Knight of the University of Chicago in 1933. Other, comparable definitions also are in use.

Spreadsheets include extensive statistical, forecasting, and other modeling and database management capabilities, functions, and routines. As spreadsheet packages evolved, add-ins were developed for structuring and solving specific model classes. Among the add-in packages, many were developed for DSS development. These DSS-related add-ins include Solver (Frontline Systems Inc., solver.com) (accessed July 2022) and *What'sBest!* (a version of Lindo, from Lindo Systems, Inc., lindo.com) (accessed July 2022) for performing linear and nonlinear optimization; NeuralTools for artificial neural networks; Evolver for genetic algorithms; and @RISK (Palisade Corp., palisade.com) (accessed July 2022) for performing simulation studies. Comparable add-ins are available for free or at a very low cost. (Conduct a Web search to find them; new ones are added to the marketplace on a regular basis.)

The spreadsheet is clearly the most popular *end-user modeling tool* because it incorporates many powerful financial, statistical, mathematical, and other functions. Spreadsheets can perform model solution tasks such as linear programming and regression analysis. The spreadsheet has evolved into an important tool for analysis, planning, and modeling (see Farasyn, Perkoz, & Van de Velde, 2008; Hurley & Balez, 2008; Ovchinnikov & Milner, 2008). Analytics in Actions 8.3 and 8.4 describe interesting applications of spreadsheet-based models in a nonprofit setting.

ANALYTICS IN ACTION 8.3 Pennsylvania Adoption Exchange Uses Spreadsheet Model to Better Match Children with Families

The Pennsylvania Adoption Exchange (PAE) was established in 1979 by the State of Pennsylvania to help county and nonprofit agencies find prospective families for children without parents who had not been adopted due to age or special needs. The PAE keeps detailed records about children and preferences of families who may adopt them. The exchange looks for families for the children across all 67 counties of Pennsylvania.

The Pennsylvania Statewide Adoption and Permanency Network is responsible for finding permanent homes for orphans. If after a few attempts the network fails to place a child with a family, they then get help from the PAE. The PAE uses an automated assessment tool to match children to families. This tool gives matching recommendations by calculating a score between 0% and 100% for a child on 78 pairs of the child's attribute values and family preferences. For some years now, the PAE has struggled to give adoption match recommendations to caseworkers for children. They are finding it difficult to manage a vast database of children collected over time for all 67 counties. The basic search algorithm produced match recommendations that were proving unfruitful for caseworkers. As a result, the number of children who have not been adopted has increased significantly, and there is a growing urgency to find families for them.

Methodology/Solution

The PAE started collecting information about the children and families through online surveys that include a new set of questions. These questions collect information about hobbies of the child, child–caseworker preferences for families, and preference of the age range of children by families. The PAE and consultants created a spreadsheet-matching tool that included additional features compared to the previously used automated tool. In this model, caseworkers can specify the weight of the attributes for selecting a family for a child. For example, if a family had a narrow set of preferences regarding gender, age, and race, then those factors can receive a higher weight. Also, caseworkers can give preference about the family's county of residence, as community relationship is an important factor for a child. Using this tool, the matching committee can compare a child and family on each attribute, thus making a more accurate match decision between a family and a child.

Results/Benefits

Since the PAE started using the new spreadsheet model for matching a family with a child, they have been able to make better matching decisions. As a result, the percentage of children getting a permanent home has increased.

This short case is one of the many examples of using spreadsheets as a decision support tool. By creating a simple scoring system for a family's desire and a child's attribute, a better matching system is produced so that fewer rejections are reported on either side.

QUESTIONS FOR DISCUSSION

1. What were the challenges faced by PAE while making adoption matching decisions?
2. What features of the new spreadsheet tool helped PAE solve their issues of matching a family with a child?

Source: Based on Slaugh, V. W., Akan, M., Kesten, O., & Unver, M. U. (2016). The Pennsylvania Adoption Exchange improves its matching process. *Interfaces*, 46(2), 133–154.

ANALYTICS IN ACTION 8.4 Metro Meals on Wheels Treasure Valley Uses Excel to Find Optimal Delivery Routes

Meals on Wheels Association of America (now Meals on Wheels America) is a not-for-profit organization that delivers approximately one million meals to homes of older people in need across the United States. Metro Meals on Wheels Treasure Valley is a local branch of Meals on Wheels America operating in Idaho. This branch has a team of volunteer drivers that drive their personal vehicles each day to deliver meals to 800 clients along 21 routes and cover an area of 2,745 square kilometers.

The Meals on Wheels Treasure Valley organization was facing many issues. First, they were looking to minimize the delivery time as the cooked food was temperature sensitive and could perish easily. They wanted to deliver the cooked food within 90 minutes after a driver left for the delivery. Second, the scheduling process was very time consuming. Two employees spent much of their time developing scheduled routes for delivery. A route coordinator determined the stops according to the number of meal recipients for a given day. After determining the stops, the coordinator made a sequence of stops that minimized the travel time of volunteers. This routing schedule was then entered into an online tool to determine turn-by-turn driving instructions for drivers. The whole process of manually deciding routes was taking a lot of extra time. Metro Meals on Wheels wanted a routing tool that could improve their delivery system and generate routing solutions for both one-way and round-trip directions for delivering meals. Those who drive regularly could deliver the warmers or coolers the next day. Others who drive only occasionally would need to come back to the kitchen to drop off the warmers/coolers.

Methodology/Solution

To solve the routing problem, a spreadsheet-based tool was developed. This tool had an interface to easily input information about the recipient such as his/her name, meal requirements, and delivery address. This information needed to be filled in the spreadsheet for each stop in the route. Next, Excel's Visual Basic for Applications functionality was used to access a developer's networking map application programming interface (API) called MapQuest. This API was used to create a travel matrix that calculated time and distance needed for delivery of the meal. This tool gave time and distance information for 5,000 location pairs a day without any cost.

When the program starts, the MapQuest API first validates the entered addresses of meal recipients. Then the program uses the API to retrieve driving distance, estimated driving time, and turn-by-turn instructions for driving between all stops in the route. The tool can then find the optimal route for up to 30 stops within a feasible time limit.

Results/Benefits

As a result of using this tool, the total annual driving distance decreased by 10,000 miles, while travel time was reduced by 530 hours. Metro Meals on Wheels Treasure Valley saved \$5,800 in 2015, based on an estimated savings rate of \$0.58 per mile (for a midsize sedan). This tool also reduced the time spent on route planning for meal deliveries. Other benefits included increased volunteer satisfaction and more retention of volunteers.

QUESTIONS FOR DISCUSSION

1. What were the challenges faced by Metro Meals on Wheels Treasure Valley related to meal delivery before adoption of the spreadsheet-based tool?
2. Explain the design of the spreadsheet-based model.
3. What are the intangible benefits of using the Excel-based model to Metro Meals on Wheels?

Source: Based on Manikas, A. S., Kroes, J. R., & Gattiker, T. F. (2016). Metro Meals on Wheels Treasure Valley employs a low-cost routing tool to improve deliveries. *Interfaces*, 46(2), 154–167.

Other important spreadsheet features include what-if analysis, goal seeking, data management, and programmability (i.e., macros). With a spreadsheet, it is easy to change a cell's value and immediately see the result. Goal seeking is performed by indicating a target cell, its desired value, and a changing cell. Extensive database management can be performed with small data sets, or parts of a database can be imported for analysis (which is essentially how OLAP works with multidimensional data cubes; in fact, most OLAP systems have the look and feel of advanced spreadsheet software after the data are loaded). Templates, macros, and other tools enhance the productivity of building DSS.

Most spreadsheet packages provide seamless integration because they read and write common file structures and easily interface with databases and other tools. Microsoft Excel is the most popular spreadsheet package. In Figure 8.3, we show a simple loan calculation model in which the boxes on the spreadsheet describe the contents of the cells, which contain formulas. A change in the interest rate in cell E7 is immediately reflected in the monthly payment in cell E13. The results can be observed and analyzed immediately. If we require a specific monthly payment, we can use goal seeking to determine an appropriate interest rate or loan amount.

Static or dynamic models can be built in a spreadsheet. For example, the monthly loan calculation spreadsheet shown in Figure 8.3 is static. Although the problem affects the borrower over time, the model indicates a single month's performance, which is replicated. A dynamic model, in contrast, represents behavior over time. The loan calculations in the spreadsheet shown in Figure 8.4 indicate the effect of prepayment on the principal over time. Risk analysis can be incorporated into spreadsheets by using built-in random-number generators to develop simulation models (see the next chapter).

Spreadsheet applications for models are reported regularly. We will learn how to use a spreadsheet-based optimization model in the next section.

SECTION 8.5 REVIEW QUESTIONS

1. What is a spreadsheet?
2. What is a spreadsheet add-in? How can add-ins help in DSS creation and use?
3. Explain why a spreadsheet is so conducive to the development of DSS.

	A	B	C	D	E	F	G	H
1								
2								
3		Simple Loan Calculation Model in Excel						
4								
5								
6		Loan Amount			\$150,000			
7		Interest Rate			8.00%			
8		Number of Years			30			
9						=E8*12		
10		Number of Months			360			
11		Interest Rate/Month			0.67%	=E7/12		
12								
13		Monthly Loan Payment			\$1,100.65	=PMT (E11, E10, E6, 0)		
14								
15								
16								
17		Excel Spreadsheet Static Model Example of a Simple Loan						
18								
19								
20								
21								
22								

FIGURE 8.3 Excel Spreadsheet Static Model Example of a Simple Loan Calculation of Monthly Payments.

	A	B	C	D	E	F	G	H	I	J	K
1											
2											
3		Dynamic Loan Calculation Model with Prepayment in Excel									
4											
5											
6		Loan Amount			\$150,000						
7		Interest Rate			8.00%						
8		Number of Years			30						
9						=E8*12					
10		Number of Months			360						
11		Interest Rate/Month			0.67%	=E7/12					
12											
13		Monthly Loan Payment			\$1,100.65	=PMT (E11, E10, E6, 0)					
14											
15											
16											
17		Excel Spreadsheet Dynamic Model Example of a Simple Loan									
18											
19		=E\$13	=C\$20								
20			\$100.00								
21		Normal	Prepay	Total	Principle						
22		Payment	Amount	Payment	Owed						
23	0				\$150,000						
24	1	\$1,100.65	\$100.00	\$1,200.65	\$149,795	=E23*(1+E\$11)-D24					
25	2	\$1,100.65	\$100.00	\$1,200.65	\$149,597						
26	3	\$1,100.65	\$100.00	\$1,200.65	\$149,394						
27	4	\$1,100.65	\$100.00	\$1,200.65	\$149,189						
28	5	\$1,100.65	\$100.00	\$1,200.65	\$148,983						
29											
30											

A \$100 Prepayment every Month—Loan is paid off in Month 270

Copy the Cells in Row 24 into Rows 25 through Row 383 to get 360 Months of Results

FIGURE 8.4 Excel Spreadsheet Dynamic Model Example of a Simple Loan Calculation of Monthly Payments and the Effects of Prepayment.

8.6 MATHEMATICAL PROGRAMMING OPTIMIZATION

Mathematical programming is a family of tools designed to help solve managerial problems in which the decision maker must allocate scarce resources among competing activities to optimize a measurable goal. For example, the distribution of machine time (the resource) among various products (the activities) is a typical allocation problem.

Linear programming (LP) is the best-known technique in a family of optimization tools called *mathematical programming*; in LP, all relationships among the variables are linear. It is used extensively in DSS (see Analytics in Action 8.6). LP models have many important applications in practice. These include supply chain management, product mix decisions, routing, and so on. Special forms of the models can be used for specific applications. For example, Analytics in Action 8.5 describes a spreadsheet model that was used to create a schedule for physicians.

LP allocation problems usually display the following characteristics:

- A limited quantity of economic resources is available for allocation.
- The resources are used in the production of products or services.
- There are two or more ways in which the resources can be used. Each is called a *solution* or a *program*.
- Each activity (product or service) in which the resources are used yields a return in terms of the stated goal.
- The allocation is usually restricted by several limitations and requirements, called *constraints*.

ANALYTICS IN ACTION 8.5 Mixed-Integer Programming Model Helps the University of Tennessee Medical Center with Scheduling Physicians

Regional Neonatal Associates is a nine-physician group working for the Neonatal Intensive Care Unit (NICU) at the University of Tennessee Medical Center in Knoxville, Tennessee. The group also serves two local hospitals in the Knoxville area for emergency purposes. For many years, one member of the group would schedule physicians manually. However, as his retirement approached, there was a need for a more automatic system to schedule physicians. The physicians wanted this system to balance their workload, as the previous schedules did not properly balance workload among them. In addition, the schedule needed to ensure there would be 24–7 NICU coverage by the physicians, and if possible, accommodate individual preferences of physicians for shift types. To address this problem, the physicians contacted the faculty of Management Science at the University of Tennessee.

The problem of scheduling physicians to shifts was characterized by constraints based on workload and lifestyle choices. The first step in solving the scheduling issue was to group shifts according to their types (day and night). The next step was determining constraints for the problem. The model needed to cover a 9-week period with nine physicians, with two physicians working weekdays and one physician overnight and on weekends. In addition, one physician had to be assigned exclusively for 24–7 coverage to the two local hospitals. Other obvious constraints also needed to be considered. For example, a day shift could not be assigned to a physician just after a night shift.

Methodology/Solution

The problem was formulated by creating a binary, mixed-integer optimization model. The first model divided workload equally among the nine physicians. But it could not assign an equal number of day and night shifts among them. This created a question of fair distribution. In addition, the physicians had differing opinions of the assigned workload. Six physicians wanted a schedule in which an equal number of day and night shifts would be assigned to each physician in the 9-week schedule, while the others wanted a schedule based on individual preference

of shifts. To satisfy requirements of both groups of physicians, a new model was formed and named the Hybrid Preference Scheduling Model (HPSM). For satisfying the equality requirement of six physicians, the model first calculated 1 week's workload and divided it for 9 weeks for them. This way, the work was divided equally for all six physicians. The workload for the three remaining physicians was distributed in the 9-week schedule according to their preference. The resulting schedule was reviewed by the physicians, and they found the schedule more acceptable.

Results/Benefits

The HPSM method accommodated both the equality and individual preference requirements of the physicians. In addition, the schedules from this model provided better rest times for the physicians compared to the previous manual schedules, and vacation requests could also be accommodated in the schedules. The HPSM model can solve similar scheduling problems demanding relative preferences among shift types.

Techniques such as mixed-integer programming models can build optimal schedules and help in operations. These techniques have been used in large organizations for a long time. Now it is possible to implement such prescriptive analytic models in spreadsheets and other easily available software.

QUESTIONS FOR DISCUSSION

1. What was the issue faced by the Regional Neonatal Associates group?
2. How did the HPSM model solve all the physician's requirements?

Source: Adapted from Bowers, M. R., Noon, C. E., Wu, W., & Bass, J. K. (2016). Neonatal physician scheduling at the University of Tennessee Medical Center. *Interfaces*, 46(2), 168–182.

The LP allocation model is based on the following rational economic assumptions:

- Returns from different allocations can be compared; that is, they can be measured by a common unit (e.g., dollars, utility).
- The return from any allocation is independent of other allocations.
- The total return is the sum of the returns yielded by the different activities.
- All data are known with certainty.
- The resources are to be used in the most economical manner.

Allocation problems typically have many possible solutions. Depending on the underlying assumptions, the number of solutions can be either infinite or finite. Usually, different solutions yield different rewards. Of the available solutions, at least one is the best, in the sense that the degree of goal attainment associated with it is the highest (i.e., the total reward is maximized). This is called an **optimal solution**, and it can be found by using a special algorithm.

Linear Programming Model

Every LP model is composed of *decision variables* (whose values are unknown and are searched for), an *objective function* (a linear mathematical function that relates the decision variables to the goal, measures goal attainment, and is to be optimized), *objective function coefficients* (unit profit or cost coefficients indicating the contribution to the objective of one unit of a decision variable), *constraints* (expressed in the form of linear inequalities or equalities that limit resources and/or requirements; these relate the variables through linear relationships), *capacities* (which describe the upper and sometimes lower limits on the constraints and variables), and *input/output (technology) coefficients* (which indicate resource utilization for a decision variable).

TECHNOLOGY INSIGHTS 8.1 Linear Programming

LP is perhaps the best-known optimization model. It deals with the optimal allocation of resources among competing activities. The allocation problem is represented by the model described here.

The problem is to find the values of the decision variables X_1, X_2 , and so on, such that the value of the result variable Z is maximized, subject to a set of linear constraints that express the technology, market conditions, and other uncontrollable variables. The mathematical relationships are all linear equations and inequalities. Theoretically, any allocation problem of this type has an infinite number of possible solutions. Using special mathematical procedures, the LP approach applies a unique computerized search procedure that finds the best solution(s) in a matter of seconds. Furthermore, the solution approach provides automatic sensitivity analysis.

Let us look at an example. MBI Corporation, which manufactures special-purpose computers, needs to make a decision: How many computers should it produce next month at the Boston plant? MBI is considering two types of computers: the CC-7, which requires 300 days of labor and \$10,000 in materials, and the CC-8, which requires 500 days of labor and \$15,000 in materials. The profit contribution of each CC-7 is \$8,000, whereas that of each CC-8 is \$12,000. The plant has a capacity of 200,000 working days per month, and the material budget is \$8 million per month. Marketing requires that at least 100 units of the CC-7 and at least 200 units of the CC-8 be produced each month. The problem is to maximize the company's profits by determining how many units of the CC-7 and how many units of the CC-8 should be produced each month. Note that in a real-world environment, it could possibly take months to obtain the data in the problem statement, and while gathering the data the decision maker would no doubt uncover facts about how to structure the model to be solved. Web-based tools for gathering data can help.

Modeling in LP: An Example

A standard LP model can be developed for the MBI Corporation problem just described. As discussed in Technology Insights 8.1, the LP model has three components: decision variables, result variables, and uncontrollable variables (constraints).

The decision variables are as follows:

$$X_1 = \text{units of CC-7 to be produced}$$

$$X_2 = \text{units of CC-8 to be produced}$$

The result variable is as follows:

$$\text{Total profit} = Z$$

The objective is to maximize total profit:

$$Z = 8,000X_1 + 12,000X_2$$

The uncontrollable variables (constraints) are as follows:

$$\text{Labor constraint: } 300X_1 + 500X_2 \leq 200,000 \text{ (in days)}$$

$$\text{Budget constraint: } 10,000X_1 + 15,000X_2 \leq 8,000,000 \text{ (in dollars)}$$

$$\text{Marketing requirement for CC-7: } X_1 \geq 100 \text{ (in units)}$$

$$\text{Marketing requirement for CC-8: } X_2 \geq 200 \text{ (in units)}$$

This information is summarized in Figure 8.5.

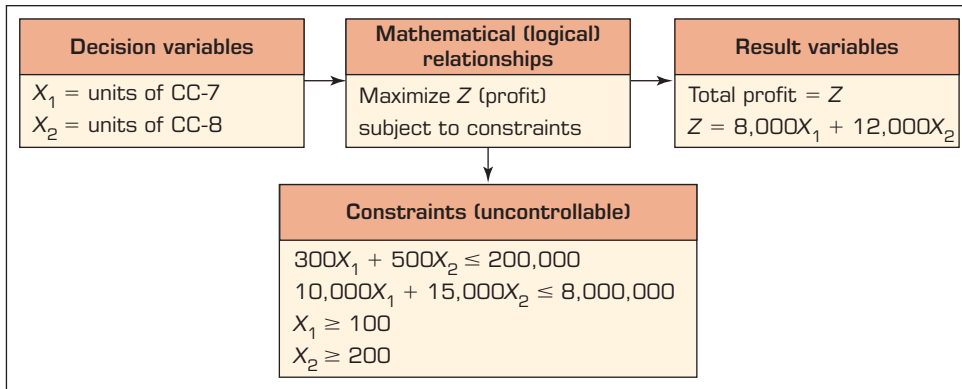


FIGURE 8.5 Mathematical Model of a Product-Mix Example.

The model also has a fourth, hidden component. Every LP model has some internal intermediate variables that are not explicitly stated. The labor and budget constraints may each have some slack in them when the left-hand side is strictly less than the right-hand side. This slack is represented internally by slack variables that indicate excess resources available. The marketing requirement constraints may each have some surplus in them when the left-hand side is strictly greater than the right-hand side. This surplus is represented internally by surplus variables indicating that there is some room to adjust the right-hand sides of these constraints. These slack and surplus variables are intermediate. They can be of great value to a decision maker because LP solution methods use them in establishing sensitivity parameters for economic what-if analyses.

The product-mix model has an infinite number of possible solutions. Assuming that a production plan is not restricted to whole numbers—which is a reasonable assumption in a monthly production plan—we want a solution that maximizes total profit: an optimal solution. Fortunately, Excel comes with the add-in Solver, which can readily obtain an optimal (best) solution to this problem. Although the location of Solver add-in has moved from one version of Excel to another, it is still available as a free add-in. Look for it under the Data tab and on the Analysis ribbon. If it is not there, you should be able to enable it by going to Excel's Options Menu and selecting Add-ins.

We enter these data directly into an Excel spreadsheet, activate Solver, and identify the goal (by setting Target Cell equal to Max), decision variables (by setting By Changing Cells), and constraints (by ensuring that Total Consumed elements is less than or equal to Limit for the first two rows and is greater than or equal to Limit for the third and fourth rows). Cells C7 and D7 constitute the decision variable cells. Results in these cells will be filled after running the Solver Add-in. Target Cell is Cell E7, which is also the result variable, representing a product of decision variable cells and their per unit profit coefficients (in Cells C8 and D8). Note that all the numbers have been divided by 1,000 to make it easier to type (except the decision variables). Rows 9–12 describe the constraints of the problem: the constraints on labor capacity, budget, and the desired minimum production of the two products X_1 and X_2 . Columns C and D define the coefficients of these constraints. Column E includes the formulae that multiply the decision variables (Cells C7 and D7) with their respective coefficients in each row. Column F defines the right-hand side value of these constraints. Excel's matrix multiplication capabilities (e.g., SUMPRODUCT function) can be used to develop such row and column multiplications easily.

After the model's calculations have been set up in Excel, it is time to invoke the Solver Add-in. Clicking on the Solver Add-in (again under the Analysis group under Data Tab) opens a dialog box (window) that lets you specify the cells or ranges that define the objective function cell, decision/changing variables (cells), and the constraints. Also,

in Options, we select the solution method (usually Simplex LP), and then we solve the problem. Next, we select all three reports—Answer, Sensitivity, and Limits—to obtain an optimal solution of $X_1 = 333.33$, $X_2 = 200$, Profit = \$5,066,667, as shown in Figure 8.6. Solver produces three useful reports about the solution. Try it. Solver now also includes the ability to solve nonlinear programming problems and integer programming problems by using other solution methods available within it.

The following example was created by Professor Rick Wilson of Oklahoma State University to further illustrate the power of spreadsheet modeling for decision support.

The table in Figure 8.7 describes some hypothetical data and attributes of nine “swing states” for the 2016 election. Attributes of the nine states include their number of electoral votes, two regional descriptors (note that three states are classified as neither North nor South), and an estimated “influence function,” which relates to increased candidate support per unit of campaign financial investment in that state.

For instance, influence function F1 shows that for every financial unit invested in that state, there will be a total of a 10-unit increase in voter support (let units stay general here), made up of an increase in young men support by 3 units, older men support by 1 unit, and young and older women each by 3 units.

The campaign has 1,050 financial units to invest in the nine states. It must invest at least 5% in each state of the total overall invested, but no more than 25% of the overall total invested can be in any one state. All 1,050 units do not have to be invested (your model must correctly deal with this).

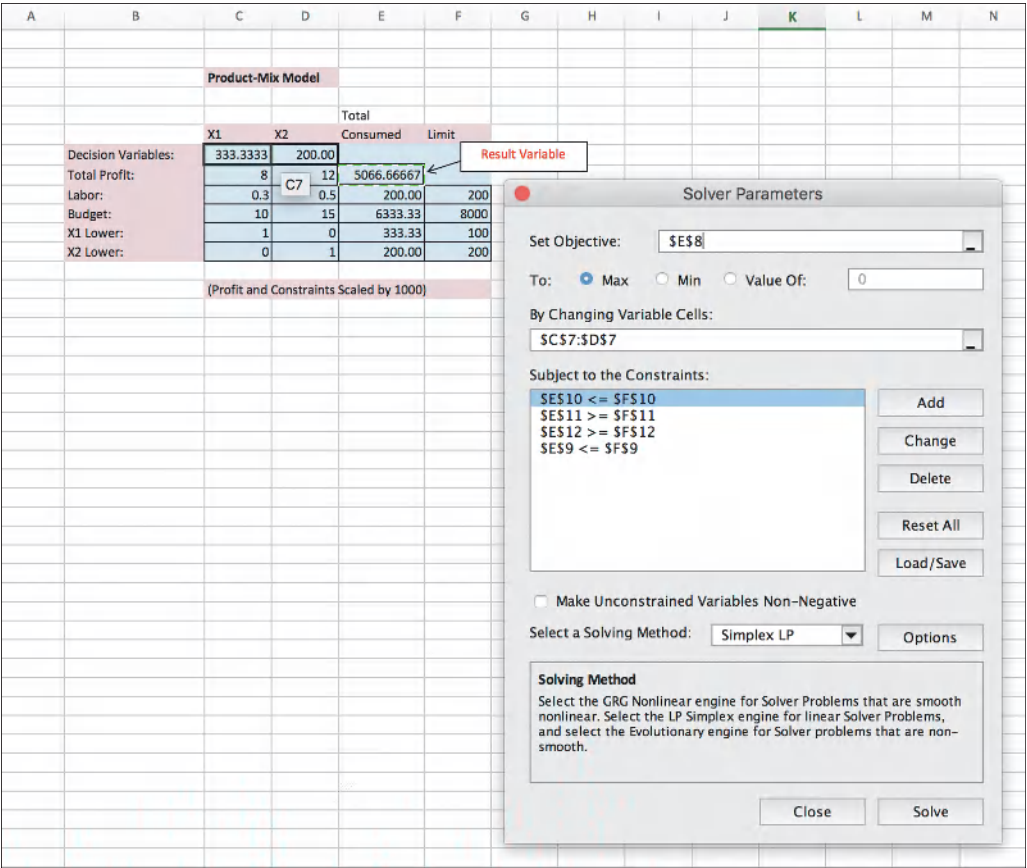


FIGURE 8.6 Excel Solver Solution to the Product-Mix Example.

	A	B	C	D	E	F	G	H
1								
2			Electoral			Influence		
3			State	Votes	W/E	N/S	Function	
4			NV	6	West		F1	
5			CO	9	West		F2	
6			IA	6	West	North	F3	
7			WI	10	West	North	F1	
8			OH	18	East	North	F2	
9			VA	13	East	South	F2	
10			NC	15	East	South	F1	
11			FL	29	East	South	F3	
12			NH	4	East		F3	
13								
14		F1	Young	Old				
15		Men	3	1	4			
16		Women	3	3	6			
17			6	4	10	Total		
18								
19		F2	Young	Old				
20		Men	1.5	2.5	4			
21		Women	2.5	1	3.5			
22			4	3.5	7.5	Total		
23								
24		F3	Young	Old				
25		Men	2.5	2.5	5			
26		Women	1	2	3			
27			3.5	4.5	8	Total		
28								

FIGURE 8.7 Data for Election Resource Allocation Example.

The campaign has some other restrictions as well. From a financial investment standpoint, the West states (in total) must have campaign investments at levels that are at least 60% of the total invested in East states. In terms of people influenced, the decision to allocate financial investments to states must lead to at least 9,200 total people influenced. Overall, the total number of females influenced must be greater than or equal to the total number of males influenced. (We humbly recognize the presence of nonbinary individuals, but the intent here is simply to illustrate a model-building exercise.) Also, at least 46% of all people influenced must be “older.”

Our task is to create an appropriate integer programming model that determines the optimal integer (i.e., whole number) allocation of financial units to states that maximizes the sum of the products of the electoral votes times units invested subject to the other restrictions. (Thus, indirectly, this model is giving preference to states with higher numbers of electoral votes.) Note that for ease of implementation by the campaign staff, all decisions for allocation in the model should lead to integer values.

The three aspects of the models can be categorized based on the following questions that they answer:

- 1. What do we control?** The amount invested in advertisements across the nine states, Nevada, Colorado, Iowa, Wisconsin, Ohio, Virginia, North Carolina, Florida, and New Hampshire, which are represented by the nine decision variables, NV, CO, IA, WI, OH, VA, NC, FL, and NH.
- 2. What do we want to achieve?** We want to maximize the total number of electoral votes gains. We know the value of each electoral vote in each state (EV), so this amounts to $EV \times \text{Investments}$ aggregated over the nine states, that is,

$$\text{Max } (6\text{NV} + 9\text{CO} + 6\text{IA} + 10\text{WI} + 18\text{OH} + 13\text{VA} + 15\text{NC} + 29\text{FL} + 4\text{NH})$$

- 3. What constrains us?** Following are the constraints as given in the problem description:

- a.** No more than 1,050 financial units to invest into, that is, $\text{NV} + \text{CO} + \text{IA} + \text{WI} + \text{OH} + \text{VA} + \text{NC} + \text{FL} + \text{NH} \leq 1,050$.

- b.** Invest at least 5% of the total in each state, that is,

$$NV \geq 0.05 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

$$CO \geq 0.05 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

$$IA \geq 0.05 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

$$WI \geq 0.05 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

$$OH \geq 0.05 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

$$VA \geq 0.05 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

$$NC \geq 0.05 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

$$FL \geq 0.05 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

$$NH \geq 0.05 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

We can implement these nine constraints in a variety of ways using Excel.

- c.** Invest no more than 25% of the total in each state.

As with (b) we need nine individual constraints again because we do not know how much of the 1,050 we will invest. We must write the constraints in “general” terms.

$$NV \leq 0.25 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

$$CO \leq 0.25 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

$$IA \leq 0.25 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

$$WI \leq 0.25 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

$$OH \leq 0.25 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

$$VA \leq 0.25 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

$$NC \leq 0.25 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

$$FL \leq 0.25 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

$$NH \leq 0.25 (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

- d.** Western states must have investment levels that are at least 60% of the Eastern states.

$$\text{West States} = NV + CO + IA + WI$$

$$\text{East States} = OH + VA + NC + FL + NH$$

$$\text{So, } (NV + CO + IA + WI) \geq 0.60 (OH + VA + NC + FL + NH).$$

Again, we can implement this constraint in a variety of ways using Excel.

- e.** Influence at least 9,200 total people, that is,

$$(10NV + 7.5CO + 8IA + 10WI + 7.5OH + 7.5VA + 10NC + 8FL + 8NH) \geq 9,200$$

- f.** Influence at least as many females as males. This requires transition of influence functions.

$$F1 = 6 \text{ women influenced, } F2 = 3.5 \text{ women}$$

$$F3 = 3 \text{ women influenced}$$

$$F1 = 4 \text{ men influenced, } F2 = 4 \text{ men}$$

$$F3 = 5 \text{ men influenced}$$

So, implementing females $\geq$ meals, we get:

$$(6NV + 3.5CO + 3IA + 6WI + 3.5OH + 3.5VA + 6NC + 3FL + 3NH) \\ \geq (4NV + 4CO + 5IA + 4WI + 4OH + 4VA + 4NC + 5FL + 5NH)$$

As before, we can implement this in Excel in a couple of different ways.

- g. At least 46% of all people influenced must be old. Note that old is subjective here. The reader can assign any age to call a group “old” or older.

All people influenced were on the left-hand side of the constraint (e). So, old people influenced would be:

$$(4NV + 3.5CO + 4.5IA + 4WI + 3.5OH + 3.5VA + 4NC + 4.5FL + 4.5NH)$$

This would be set $\geq 0.46^*$ the left-hand side of constraint (e). $(10NV + 7.5CO + 8IA + 10WI + 7.5OH + 7.5VA + 10NC + 8FL + 8NH)$, which would give a right-hand side of $(0.46NV + 3.45CO + 3.68IA + 4.6WI + 3.45OH + 3.45VA + 4.6NC + 3.68FL + 3.68NH)$

This is the last constraint other than to force all variables to be integers.

All told in algebraic terms, this integer programming model would have 9 decision variables and 24 constraints (one constraint for integer requirements).

Implementation

One approach would be to implement the model in strict “standard form,” or a row-column form, where all constraints are written with decision variables on the left-hand side, a number on the right-hand side. Figure 8.8 shows such an implementation and displays the solved model.

Alternatively, we could use the spreadsheet to calculate different parts of the model in a less rigid manner, as well as uniquely implementing the repetitive constraints (b) and (c), and have a much more concise (but not as transparent) spreadsheet. This is shown in Figure 8.9.

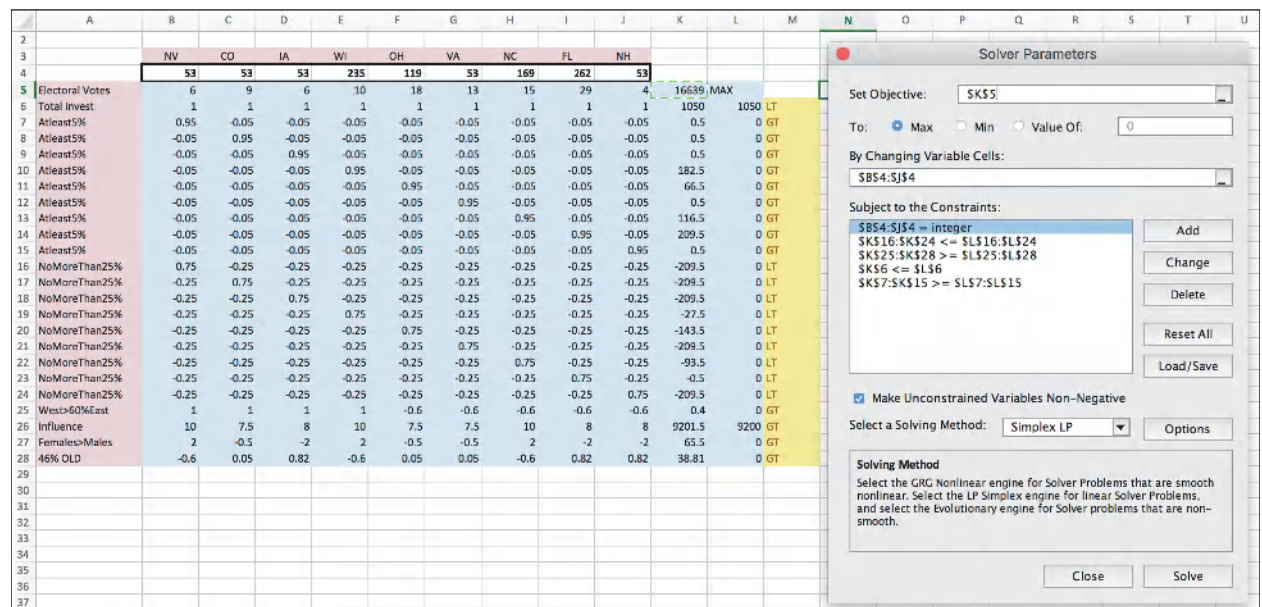


FIGURE 8.8 Model for Election Resource Allocation—Standard Version.

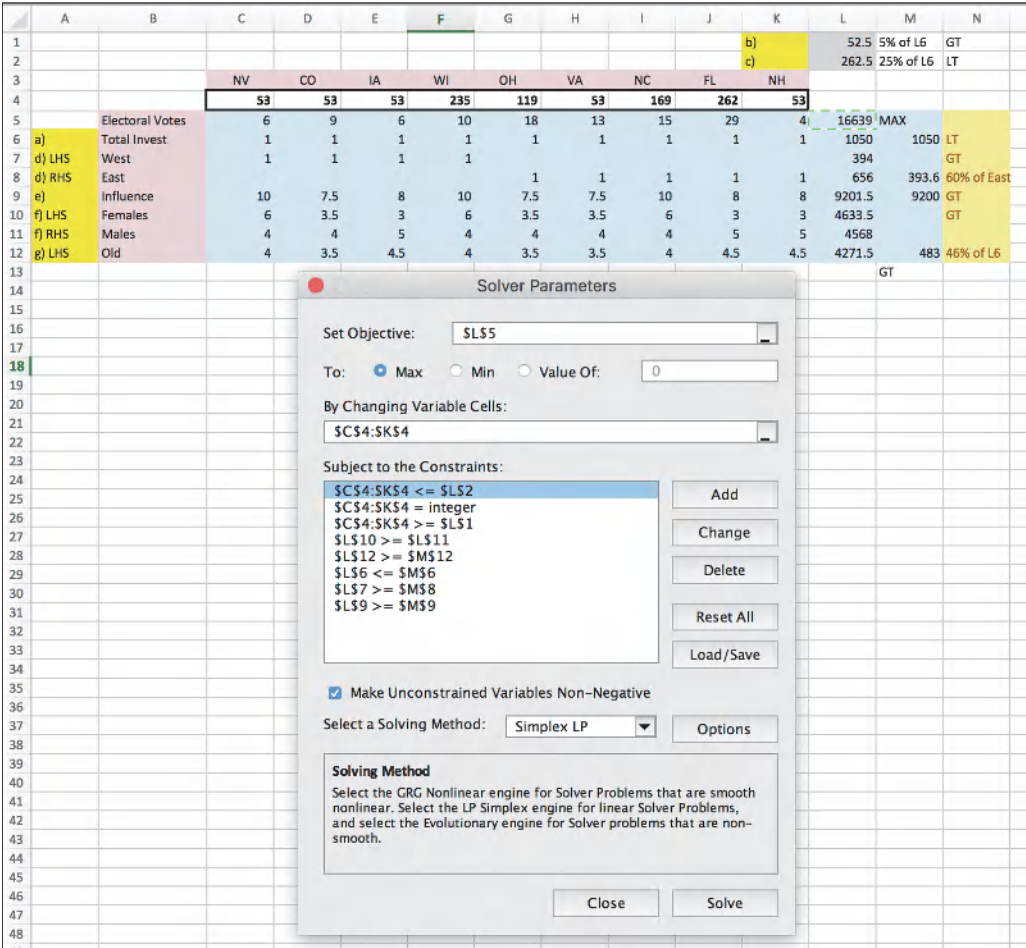


FIGURE 8.9 A Compact Formulation for Election Resource Allocation.

LP models (and their specializations and generalizations) can also be specified directly in a number of other user-friendly modeling systems. Two of the best known are Lindo and Lingo (Lindo Systems, Inc., lindo.com [accessed July 2022]; demos are available). Lindo is an LP and integer programming system. Models are specified in essentially the same way that they are defined algebraically. Based on the success of Lindo, the company developed Lingo, a modeling language that includes the powerful Lindo optimizer and extensions for solving nonlinear problems. Many other modeling languages such as AMPL, AIMMS, MPL, XPRESS, and others are available.

The most common optimization models can be solved by a variety of mathematical programming methods, including the following:

- Assignment (best matching of objects)
- Dynamic programming
- Goal programming
- Investment (maximizing rate of return)
- Linear and integer programming
- Network models for planning and scheduling
- Nonlinear programming
- Replacement (capital budgeting)
- Simple inventory models (e.g., economic order quantity)
- Transportation (minimize cost of shipments)

SECTION 8.6 REVIEW QUESTIONS

1. List and explain the assumptions involved in LP.
2. List and explain the characteristics of LP.
3. Describe an allocation problem.
4. Define the product-mix problem.
5. Define the blending problem.
6. List several common optimization models.

8.7 MULTIPLE GOALS, SENSITIVITY ANALYSIS, WHAT-IF ANALYSIS, AND GOAL SEEKING

Many, if not most, decision situations involve juggling between competing goals and alternatives. In addition, there is significant uncertainty about the assumptions and predictions being used in building a prescriptive analytics model. The following paragraphs simply recognize that these are also addressed in prescriptive analytics software and techniques. Coverage of these techniques is usually common in prescriptive analytics or operations research/management science courses.

Multiple Goals

The analysis of management decisions aims at evaluating, to the greatest possible extent, how far each alternative advances managers toward their goals. Unfortunately, managerial problems are seldom evaluated with a single simple goal, such as profit maximization. Today's management systems are much more complex, and one with a single goal is rare. Instead, managers want to attain *simultaneous goals*, some of which may conflict. Different stakeholders have different goals. Therefore, it is often necessary to analyze each alternative considering its determination of each of several goals (see Koksalan & Zions, 2001).

For example, consider a profit-making firm. In addition to earning money, the company wants to grow, develop its products and employees, provide job security to its workers, and serve the community. Managers want to satisfy the shareholders and at the same time enjoy high salaries and expense accounts, and employees want to increase their take-home pay and benefits. When a decision is to be made—say, about an investment project—some of these goals complement each other, whereas others conflict. Kearns (2004) described how the analytic hierarchy process (AHP) combined with integer programming, addresses multiple goals in evaluating information technology (IT) investments.

Many quantitative models of decision theory are based on comparing a single measure of effectiveness, generally some form of utility to the decision maker. Therefore, it is usually necessary to transform a multiple-goal problem into a single-measure-of-effectiveness problem before comparing the effects of the solutions. This is a common method for handling multiple goals in an LP model.

Certain difficulties may arise when analyzing multiple goals:

- It is usually difficult to obtain an explicit statement of the organization's goals.
- The decision maker may change the importance assigned to specific goals over time or for different decision scenarios.
- Goals and subgoals are viewed differently at various levels of the organization and within different departments.
- Goals change in response to changes in the organization and its environment.
- The relationship between alternatives and their role in determining goals may be difficult to quantify.

- Complex problems are solved by groups of decision makers, each of whom has a personal agenda.
- Participants assess the importance (priorities) of the various goals differently.

Several methods of handling multiple goals can be used when working with such situations. The most common ones are

- Utility theory
- Goal programming
- Expression of goals as constraints, using LP
- A points system

Sensitivity Analysis

A model builder makes predictions and assumptions regarding input data, many of which deal with the assessment of uncertain futures. When the model is solved, the results depend on these data. **Sensitivity analysis** attempts to assess the impact of a change in the input data or parameters on the proposed solution (i.e., the result variable).

Sensitivity analysis is extremely important in prescriptive analytics because it allows flexibility and adaptation to changing conditions and to the requirements of different decision-making situations, provides a better understanding of the model and the decision-making situation it attempts to describe, and permits the manager to input data to increase the confidence in the model. Sensitivity analysis tests relationships such as the following:

- The impact of changes in external (uncontrollable) variables and parameters on the outcome variable(s)
- The impact of changes in decision variables on the outcome variable(s)
- The effect of uncertainty in estimating external variables
- The effects of different dependent interactions among variables
- The robustness of decisions under changing conditions

Sensitivity analyses are used for:

- Revising models to eliminate too-large sensitivities
- Adding details about sensitive variables or scenarios
- Obtaining better estimates of sensitive external variables
- Altering a real-world system to reduce actual sensitivities
- Accepting and using the sensitive (and hence vulnerable) real world, leading to the continuous and close monitoring of actual results

The two types of sensitivity analyses are automatic and trial and error.

AUTOMATIC SENSITIVITY ANALYSIS Automatic sensitivity analysis is performed in standard quantitative model implementations such as LP. For example, it reports the range within which a certain input variable or parameter value (e.g., unit cost) can vary without having any significant impact on the proposed solution. Automatic sensitivity analysis is usually limited to one change at a time, and only for certain variables. However, it is powerful because of its ability to establish ranges and limits very fast (and with little or no additional computational effort). Sensitivity analysis is provided by Solver and almost all other software packages such as Lingo. Consider the MBI Corporation example introduced previously. Sensitivity analysis could be used to determine that if the right-hand side of the marketing constraint on CC-8 could be decreased by one unit, then the net profit would increase by \$1,333.33. This is valid for the right-hand side decreasing to zero. Significant additional analysis is possible along these lines.

TRIAL-AND-ERROR SENSITIVITY ANALYSIS The impact of changes in any variable, or in several variables, can be determined through a simple trial-and-error approach. You change some input data and solve the problem again. When the changes are repeated several times, better and better solutions may be discovered. Such experimentation, which is easy to conduct when using appropriate modeling software, such as Excel, has two approaches: what-if analysis and goal seeking.

What-If Analysis

What-if analysis is structured as *What will happen to the solution if an input variable, an assumption, or a parameter value is changed?* Here are some examples:

- What will happen to the total inventory cost if the cost of carrying inventories increases by 10%?
- What will be the market share if the advertising budget increases by 5%?

With the appropriate user interface, it is easy for managers to ask a computer to model these types of questions and get immediate answers. Furthermore, they can perform multiple cases and thereby change the percentage, or any other data in the question, as desired. The decision maker does all this directly, without a computer programmer.

Figure 8.10 shows a spreadsheet example of a what-if query for a cash flow problem. When the user changes the cells containing the initial sales (from 100 to 120) and the sales growth rate (from 3% to 4% per quarter), the program immediately recomputes the value of the annual net profit cell (from \$127 to \$182). At first, initial sales were 100, growing at 3% per quarter, yielding an annual net profit of \$127. Changing the initial sales cell to 120 and the sales growth rate to 4% causes the annual net profit to rise to \$182. What-if analysis is common in many decision systems. Users are given the opportunity to change their answers to some of the system's questions, and a revised recommendation is found.

4									
5									
6									
7	Unit revenue	\$	1.20						
8	Unit cost	\$	0.60						
9									
10	Initial sales		120						
11	Sales growth rate		0.04						
12									
13	Annual net profit	\$	182						
14									
15									
16									
17	Cash Flow Model for 1996								
18									
19			Qtr1	Qtr2	Qtr3	Qtr4		Annual	
20	Sales		120	125	130	135		510	
21	Revenue	\$	144	\$ 150	\$ 156	\$ 162	\$	611	
22	Variable cost	\$	72	\$ 75	\$ 78	\$ 81	\$	306	
23	Fixed cost	\$	30	\$ 31	\$ 31	\$ 32	\$	124	
24	Net profit	\$	42	\$ 44	\$ 47	\$ 49	\$	182	
25									

FIGURE 8.10 Example of a What-If Analysis Done in an Excel Worksheet.

5																																										
6																																										
7	Investment Problem				Initial Investment:		\$	1,000.00																																		
8	Example of GoalSeeking				Interest Rate:			10%																																		
9																																										
10	Find the Interest Rate				<table><thead><tr><th>Year</th><th>Annual Returns</th></tr></thead><tbody><tr><td>1</td><td>\$ 120.00</td></tr><tr><td>2</td><td>\$ 130.00</td></tr><tr><td>3</td><td>\$ 140.00</td></tr><tr><td>4</td><td>\$ 150.00</td></tr><tr><td>5</td><td>\$ 160.00</td></tr><tr><td>6</td><td>\$ 152.00</td></tr><tr><td>7</td><td>\$ 144.40</td></tr><tr><td>8</td><td>\$ 137.18</td></tr><tr><td>9</td><td>\$ 130.32</td></tr><tr><td>10</td><td>\$ 123.80</td></tr></tbody></table>		Year	Annual Returns	1	\$ 120.00	2	\$ 130.00	3	\$ 140.00	4	\$ 150.00	5	\$ 160.00	6	\$ 152.00	7	\$ 144.40	8	\$ 137.18	9	\$ 130.32	10	\$ 123.80	<table><thead><tr><th>NPV Calculations</th></tr></thead><tbody><tr><td>\$109.09</td></tr><tr><td>\$118.18</td></tr><tr><td>\$127.27</td></tr><tr><td>\$136.36</td></tr><tr><td>\$145.45</td></tr><tr><td>\$138.18</td></tr><tr><td>\$131.27</td></tr><tr><td>\$124.71</td></tr><tr><td>\$118.47</td></tr><tr><td>\$112.55</td></tr></tbody></table>			NPV Calculations	\$109.09	\$118.18	\$127.27	\$136.36	\$145.45	\$138.18	\$131.27	\$124.71	\$118.47	\$112.55
Year	Annual Returns																																									
1	\$ 120.00																																									
2	\$ 130.00																																									
3	\$ 140.00																																									
4	\$ 150.00																																									
5	\$ 160.00																																									
6	\$ 152.00																																									
7	\$ 144.40																																									
8	\$ 137.18																																									
9	\$ 130.32																																									
10	\$ 123.80																																									
NPV Calculations																																										
\$109.09																																										
\$118.18																																										
\$127.27																																										
\$136.36																																										
\$145.45																																										
\$138.18																																										
\$131.27																																										
\$124.71																																										
\$118.47																																										
\$112.55																																										
11	(the Internal Rate of																																									
12	Return-IRR)																																									
13	that yields an NPV																																									
14	of \$0																																									
15																																										
16																																										
17																																										
18																																										
19																																										
20																																										
21																																										
22																																										
23				The NPV Solutions:		\$261.55																																				
24																																										

FIGURE 8.11 Goal-Seeking Analysis.

Goal Seeking

Goal seeking calculates the values of the inputs necessary to achieve a desired level of an output (goal). It represents a backward solution approach. The following are some examples of goal seeking:

- What annual R&D budget is needed for an annual growth rate of 15% by 2018?
- How many nurses are needed to reduce the average waiting time of a patient in the emergency room to less than 10 minutes?

An example of goal seeking is shown in Figure 8.11. For example, in a financial planning model in Excel, the internal rate of return (IRR) is the interest rate that produces a net present value (NPV) of zero. Given a stream of annual returns in Column E, we can compute the NPV of planned investment. By applying goal seeking, we can determine the internal rate of return where the NPV is zero. The goal to be achieved is NPV equal to zero, which determines the internal rate of return of this cash flow, including the investment. We set the NPV cell to the value 0 by changing the interest rate cell. The answer is 38.77059%.

COMPUTING A BREAK-EVEN POINT BY USING GOAL SEEKING Some modeling software packages can directly compute break-even points, which is an important application of goal seeking. This involves determining the value of the decision variables (e.g., quantity to produce) that generate zero profit.

In many general applications programs, it can be difficult to conduct sensitivity analysis because the prewritten routines usually present only a limited opportunity for asking what-if questions. In a DSS, the what-if and the goal-seeking options must be easy to perform.

SECTION 8.7 REVIEW QUESTIONS

1. List some difficulties that may arise when analyzing multiple goals.
2. List the reasons for performing sensitivity analysis.
3. Explain why a manager might perform what-if analysis.
4. Explain why a manager might use goal seeking.

8.8 DECISION ANALYSIS WITH DECISION TABLES AND DECISION TREES

Decision situations that involve a finite and usually not too large number of alternatives are modeled through an approach called **decision analysis** (see Arsham, 2006a,b; Decision Analysis Society, decision-analysis.society.informs.org). Using this approach, the alternatives are listed in a table or a graph, with their forecasted contributions to the goal(s) and the probability of obtaining the contribution. These can be evaluated to select the best alternative.

Single-goal situations can be modeled with *decision tables* or *decision trees*. Multiple goals (criteria) can be modeled with several other techniques, described later in this chapter.

Decision Tables

Decision tables conveniently organize information and knowledge in a systematic, tabular manner to prepare it for analysis. For example, say that an investment company is considering investing in one of three alternatives: bonds, stocks, or certificates of deposit (CDs). The company is interested in one goal: maximizing the yield on the investment after 1 year. If it were interested in other goals, such as safety or liquidity, the problem would be classified as one of *multicriteria decision analysis* (see Koksalan & Zions, 2001).

The yield depends on the state of the economy sometime in the future (often called the *state of nature*), which can be in solid growth, stagnation, or inflation. Experts estimated the following annual yields:

- If there is solid growth in the economy, bonds will yield 12%, stocks 15%, and time deposits 6.5%.
- If stagnation prevails, bonds will yield 6%, stocks 3%, and time deposits 6.5%.
- If inflation prevails, bonds will yield 3%, stocks will bring a loss of 2%, and time deposits will yield 6.5%.

The problem is to select the one best investment alternative. These are assumed to be discrete alternatives. Combinations such as investing 50% in bonds and 50% in stocks must be treated as new alternatives.

The investment decision-making problem can be viewed as a *two-person game* (see Kelly, 2002). The investor makes a choice (i.e., a move), and then a state of nature occurs (i.e., makes a move). Table 8.3 shows the payoff of a mathematical model. The table includes *decision variables* (the alternatives), *uncontrollable variables* (the states of the economy, e.g., the environment), and *result variables* (the projected yield, e.g., outcomes). All the models in this section are structured in a spreadsheet framework.

If this were a decision-making problem under certainty, we would know what the economy would be and could easily choose the best investment. But that is not the case, so we must consider the two situations of uncertainty and risk. For uncertainty, we do not know the probabilities of each state of nature. For risk, we assume that we know the probabilities with which each state of nature will occur.

TABLE 8.3 Investment Problem Decision Table Model

Alternative	State of Nature (Uncontrollable Variables)		
	Solid Growth (%)	Stagnation (%)	Inflation (%)
Bonds	12.0	6.0	3.0
Stocks	15.0	3.0	-2.0
CDs	6.5	6.5	6.5

TREATING UNCERTAINTY Several methods are available for handling uncertainty. For example, the *optimistic approach* assumes that the best possible outcome of each alternative will occur and then selects the best of the best (i.e., stocks). The *pessimistic approach* assumes that the worst possible outcome for each alternative will occur and selects the best of these (i.e., CDs). Another approach simply assumes that all states of nature are equally possible (see Clemen & Reilly, 2000; Goodwin & Wright, 2000; Kontoghiorghe, Rustem, & Siokos, 2002). Every approach for handling uncertainty has serious problems. Whenever possible, the analyst should attempt to gather enough information so that the problem can be treated under assumed certainty or risk.

TREATING RISK The most common method for solving this risk analysis problem is to select the alternative with the greatest expected value. Assume that experts estimate the chance of solid growth at 50%, the chance of stagnation at 30%, and the chance of inflation at 20%. The decision table is then rewritten with the known probabilities (see Table 8.3). An expected value is computed by multiplying the results (i.e., outcomes) by their respective probabilities and adding them. For example, investing in bonds yields an expected return of $12(0.5) + 6(0.3) + 3(0.2) = 8.4\%$.

This approach can sometimes be a dangerous strategy because the utility of each potential outcome may be different from the value. Even if there is an infinitesimal chance of a catastrophic loss, the expected value may seem reasonable, but the investor may not be willing to cover the loss. For example, suppose a financial advisor presents you with an “almost sure” investment of \$1,000 that can double your money in one day, and then the advisor says, “Well, there is a 0.9999 probability that you will double your money, but unfortunately there is a 0.0001 probability that you will be liable for a \$500,000 out-of-pocket loss.” The expected value of this investment is as follows:

$$\begin{aligned} 0.9999(\$2,000 - \$1,000) + .0001(-\$500,000 - \$1,000) &= \$999.90 - \$50.10 \\ &= \$949.80 \end{aligned}$$

The potential loss could be catastrophic for any investor who is not a billionaire. Depending on the investor’s ability to cover the loss, an investment has different expected utilities. Remember that the investor makes the decision only *once*.

Decision Trees

An alternative representation of the decision table is a decision tree. A **decision tree** shows the relationships of the problem graphically and can handle complex situations in a compact form. However, a decision tree can be cumbersome if there are many alternatives or states of nature. TreeAge Pro (TreeAge Software Inc., treeage.com) (accessed July 2022) and PrecisionTree (Palisade Corp., palisade.com) (accessed July 2022) include powerful, intuitive, and sophisticated decision tree analysis systems. These vendors also provide excellent examples of decision trees used in practice. Note that the phrase *decision tree* has been used to describe two different types of models and algorithms. In the current context, decision trees refer to scenario analysis. On the other hand, some classification algorithms in predictive analysis (see Chapters 4 and 5) are also called decision tree algorithms. The reader is advised to note the difference between two different uses of the same name—decision tree.

A simplified investment case of **multiple goals** (a decision situation in which alternatives are evaluated with several, sometimes conflicting, goals) is shown in Table 8.4. The three goals (criteria) are yield, safety, and liquidity. This situation is under assumed certainty; that is, only one possible consequence is projected for each alternative; the more complex cases of risk or uncertainty could be considered. Some of the results are qualitative (e.g., low, high) rather than numeric.

TABLE 8.4 Multiple Goals

Alternative	Yield (%)	Safety	Liquidity
Bonds	8.4	High	High
Stocks	8.0	Low	High
CDs	6.5	Very high	High

See Clemen and Reilly (2000), Goodwin and Wright (2000), and Decision Analysis Society (<https://connect.informs.org/das/home>) (accessed July 2022) for more on decision analysis. Although doing so is quite complex, it is possible to apply mathematical programming directly to decision-making situations under risk. We discuss several other methods of treating risk later in the book. These include simulation, certainty factors, and fuzzy logic.

► SECTION 8.8 REVIEW QUESTIONS

1. What is a decision table?
2. What is a decision tree?
3. How can a decision tree be used in decision-making?
4. Describe what it means to have multiple goals.

8.9 INTRODUCTION TO SIMULATION

In this section and the next, we introduce a category of techniques that are used for supporting decision-making. Very broadly, these methods fall under the umbrella of simulation. **Simulation** is the appearance of reality. In decision systems, simulation is a technique for conducting experiments (e.g., what-if analyses) with a computer on a model of a management system. Strictly speaking, simulation is a *descriptive* rather than a *prescriptive* method. There is no automatic search for an optimal solution. Instead, a simulation model describes or predicts the characteristics of a given system under different conditions. When the values of the characteristics are computed, the best of several alternatives can be selected. The simulation process usually repeats an experiment many times to obtain an estimate (and a variance) of the overall effect of certain actions. For most situations, a computer simulation is appropriate, but there are some well-known manual simulations (e.g., a city police department simulated its patrol car scheduling with a carnival game wheel).

Typically, real decision-making situations involve some randomness. Because many decision situations deal with semistructured or unstructured situations, reality is complex, which may not be easily represented by optimization or other models but can often be handled by simulation. Simulation is one of the most used decision support methods. Analytics in Action 8.6 illustrates the value of simulation in a setting where the problem complexity does not permit building a traditional optimization model.

Major Characteristics of Simulation

Simulation typically involves building a model of reality to the extent practical. Simulation models may suffer from fewer assumptions about the decision situation as compared to other prescriptive analytic models. In addition, simulation is a technique for *conducting experiments*. Therefore, it involves testing specific values of the decision or uncontrollable variables in the model and observing the impact on the output variables.

Finally, simulation is normally used only when a problem is too complex to be treated using numerical optimization techniques. Complexity in this situation means either that the problem cannot be formulated for optimization (e.g., because the assumptions do

ANALYTICS IN ACTION 8.6 Steel Tubing Manufacturer Uses a Simulation-Based Production Scheduling System

A steel manufacturing plant produces rolled-steel tubes for different industries across the country. They build tubes based on a customer's requirements and specifications. Maintaining high-quality norms and timely delivery of products are two of the foremost important criteria for this steel tubing plant. The plant views its manufacturing system as a sequence of operations where it unrolls steel from one reel and rolls it onto a different reel. This happens once the forming, welding, editing, or inspecting operation is finished. The ultimate product would be a reel of rolled steel tubing that weighs about 20 tons. The reel is then shipped to the customer.

A key challenge for management is to be able to predict the appropriate delivery date for an order, and its impact on the currently planned production schedule. Given the complexity of the production process, it is not easy to develop an optimization model in Excel or other software to build a production schedule (see Analytics in Action 8.1). The issue is that these tools fail to capture key planning issues such as employee schedules and qualifications, material accessibility, material allocation complication, and random aspects of the operation.

Methodology/Solution

When traditional modeling methods do not capture the problem subtleties or complexities, a simulation model could perhaps be built. The predictive analysis approach uses a versatile Simio simulation model that takes into consideration all the operational complexity, manufacturing material matching algorithms, and deadline considerations. Also, Simio's service offering, known as risk-based planning and scheduling (RPS), provides some user interfaces and reports simply designed for production management. This gives the client the ability to explore the impact of a new order on their production plan and schedule within about 10 minutes.

Results/Benefits

Such models provide significant visibility into the production schedule. The risk-based planning and scheduling system should be able to warn the master scheduler that a specific order has a chance of being delivered late. Changes could also be made sooner to rectify issues with an order. Success for this steel tubing manufacturer is directly tied to product quality and on-time delivery. By exploitation of Simio's predictive RPS offering, the plant expects improved market share.

QUESTIONS FOR DISCUSSION

1. Explain the advantages of using Simio's simulation model over traditional methods.
2. In what ways has the predictive analysis approach helped management achieve the goals of analyzing the production schedules?
3. Besides the steel manufacturing industry, in what other industries could such a modeling approach help improve quality and service?

What Can We Learn from This Analytics In Action?

By using Simio's simulation model, the manufacturing plant made better decisions in assessment of operations, taking all of the problem issues into consideration. Thus, a simulation-based production scheduling system could derive higher returns and market share for the steel tubing manufacturer. Simulation is an important technique for prescriptive analytics.

Compiled from Arthur, Molly. "Simulation-Based Production Scheduling System." *www.simio.com*, Simio LLC, 2014; "Risk-Based Planning and Scheduling (RPS) with Simio." *www.simio.com*, Simio LLC, *www.simio.com/about-simio/why-simio/simio-RPS-risk-based-planning-and-scheduling.php* (accessed July 2022).

not hold), that the formulation is too large, that there are too many interactions among the variables, or that the problem is stochastic in nature (i.e., exhibits risk or uncertainty).

Advantages of Simulation

Simulation is used in decision support modeling for the following reasons:

- The theory is fairly straightforward.
- A great amount of *time compression* can be attained, quickly giving a manager some feel as to the long-term (1- to 10-year) effects of many policies.
- Simulation is descriptive rather than normative. This allows the manager to pose what-if questions. Managers can use a trial-and-error approach to problem solving and can do so faster, at less expense, more accurately, and with less risk.
- A manager can experiment to determine which decision variables and which parts of the environment are important, and with different alternatives.
- An accurate simulation model requires an intimate knowledge of the problem, thus forcing the model builder to constantly interact with the manager. This is desirable for DSS development because the developer and manager both gain a better understanding of the problem and the potential decisions available.
- The model is built from the manager's perspective.
- The simulation model is built for one problem and typically cannot solve any other problem. Thus, no generalized understanding is required of the manager; every component in the model corresponds to part of the real system.
- Simulation can handle an extremely wide variety of problem types, such as inventory and staffing, as well as higher-level managerial functions, such as long-range planning.
- Simulation generally can include the real complexities of problems; simplifications are not necessary. For example, simulation can use real probability distributions rather than approximate theoretical distributions.
- Simulation automatically produces many important performance measures.
- Simulation is often the only DSS modeling method that can readily handle relatively unstructured problems.
- Some relatively easy-to-use simulation packages (e.g., Monte Carlo simulation) are available. These include add-in spreadsheet packages (e.g., @RISK), influence diagram software, Java-based (and other Web development) packages, and the visual interactive simulation systems to be discussed shortly.

Disadvantages of Simulation

The primary disadvantages of simulation are as follows:

An optimal solution cannot be guaranteed, but relatively good ones are generally found. Simulation model construction can be a slow and costly process, although newer modeling systems are easier to use than ever.

Solutions and inferences from a simulation study are usually not transferable to other problems because the model incorporates unique problem factors.

Simulation is sometimes so easy to explain to managers that analytic methods are often overlooked.

Simulation software sometimes requires special skills because of the complexity of the formal solution method.

The Methodology of Simulation

Simulation involves setting up a model of a real system and conducting repetitive experiments on it. The methodology consists of the following steps, as shown in Figure 8.12:

1. **Define the problem.** We examine and classify the real-world problem, specifying why a simulation approach is appropriate. The system's boundaries, environment, and other such aspects of problem clarification are handled here.

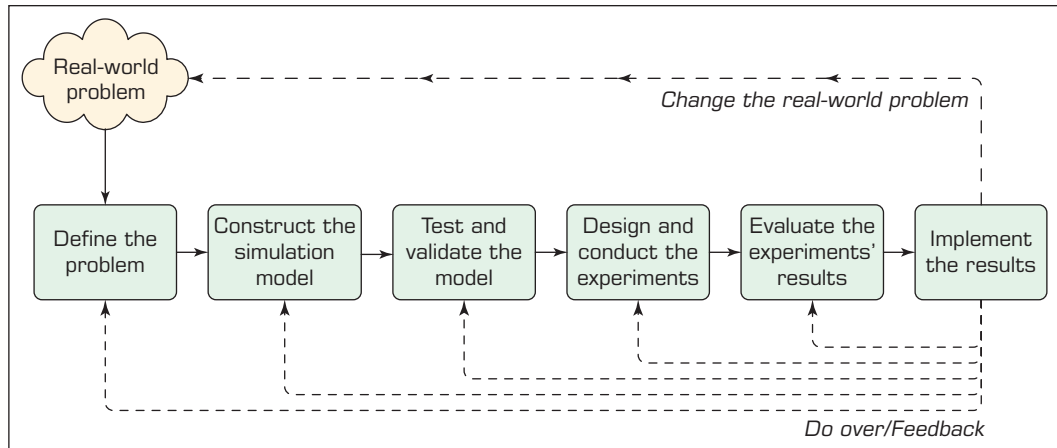


FIGURE 8.12 The Process of Simulation.

2. **Construct the simulation model.** This step involves determination of the variables and their relationships, as well as data gathering. Often the process is described by using a flowchart, and then a computer program is written.
3. **Test and validate the model.** The simulation model must properly represent the system being studied. Testing and validation ensure this.
4. **Design the experiment.** When the model has been proven valid, an experiment is designed. Determining how long to run the simulation is part of this step. There are two important and conflicting objectives: accuracy and cost. It is also prudent to identify typical (e.g., mean and median cases for random variables), best-case (e.g., low-cost, high-revenue), and worst-case (e.g., high-cost, low-revenue) scenarios. These help establish the ranges of the decision variables and environment in which to work and also assist in debugging the simulation model.
5. **Conduct the experiment.** Conducting the experiment involves issues ranging from random-number generation to result presentation.
6. **Evaluate the results.** The results must be interpreted. In addition to standard statistical tools, sensitivity analyses can also be used.
7. **Implement the results.** The implementation of simulation results involves the same issues as any other implementation. However, the chances of success are better because the manager is usually more involved with the simulation process than with other models. Higher levels of managerial involvement generally led to higher levels of implementation success.

Banks and Gibson (2009) presented some useful advice about simulation practices. For example, they list the following seven issues as the common mistakes committed by simulation modelers. The list, though not exhaustive, provides general directions for professionals working on simulation projects.

- Focusing more on the model than on the problem
- Providing point estimates
- Not knowing when to stop
- Reporting what the client wants to hear rather than what the model results say
- Lack of understanding of statistics
- Confusing cause and effect
- Failure to replicate reality

Simulation Types

As we have seen, simulation and modeling are used when pilot studies and experimenting with real systems are expensive or sometimes impossible. Simulation models allow us to investigate various interesting scenarios before making any investment. In fact, in simulations, the real-world operations are mapped into the simulation model. The model consists of relationships and, consequently, equations that all together present the real-world operations. The results of a simulation model, then, depend on the set of parameters given to the model as inputs.

There are various simulation paradigms such as Monte Carlo simulation, discrete event, agent based, or system dynamics. One of the factors that determine the type of simulation technique is the level of abstraction in the problem. Discrete events and agent-based models are usually used for middle or low levels of abstraction. They usually consider individual elements such as people, parts, and products in the simulation models, whereas systems dynamics is more appropriate for aggregate analysis.

In the following section, we introduce the major types of simulation: probabilistic simulation, time-dependent and time-independent simulation, and visual simulation. There are many other simulation techniques such as system dynamics modeling, and agent-based modeling. As has been noted before, the goal here is to make you aware of the potential of some of these techniques as opposed to make you an expert in using them.

PROBABILISTIC SIMULATION In probabilistic simulation, one or more of the independent variables (e.g., the demand in an inventory problem) are probabilistic. They follow certain probability distributions, which can be either discrete distributions or continuous distributions:

- *Discrete distributions* involve a situation with a limited number of events (or variables) that can take on only a finite number of values.
- *Continuous distributions* are situations with unlimited numbers of possible events that follow density functions, such as the normal distribution.

The two types of distributions are shown in Table 8.5.

TIME-DEPENDENT VERSUS TIME-INDEPENDENT SIMULATION *Time-independent* refers to a situation in which it is not important to know exactly when the event occurred. For example, we may know that the demand for a certain product is three units per day, but we do not care *when* during the day the item is demanded. In some situations, time may not be a factor in the simulation at all, such as in steady-state plant control design. However, in waiting-line problems applicable to e-commerce, it is important to know the precise time of arrival (to know whether the customer will have to wait). This is a *time-dependent* situation.

TABLE 8.5 Discrete versus Continuous Probability Distributions

Daily Demand	Discrete Probability	Continuous Probability
5	0.10	Daily demand is normally distributed with a mean of 7 and a standard deviation of 1.2
6	0.15	
7	0.30	
8	0.25	
9	0.20	

Monte Carlo Simulation

In most business decision problems, we usually employ one of the following two types of probabilistic simulations. The most common simulation method for business decision problems is the **Monte Carlo simulation**. This method usually begins with building a model of the decision problem without having to consider the uncertainty of any variables. Then we recognize that certain parameters or variables are uncertain or follow an assumed or estimated probability distribution. This estimation is based on analysis of past data. Then we begin running sampling experiments. Running sampling experiments consists of generating random values of uncertain parameters and then computing values of the variables that are impacted by such parameters or variables. These sampling experiments essentially amount to solving the same model hundreds or thousands of times. We can then analyze the behavior of these dependent or performance variables by examining their statistical distributions. This method has been used in simulations of physical as well as business systems. A good public tutorial on the Monte Carlo simulation method is available on Palisade.com (http://www.palisade.com/risk/monte_carlo_simulation.asp) (accessed July 2022). Palisade markets a tool called @RISK, a popular spreadsheet-based Monte Carlo simulation software. Front Line Systems also offer an Analytic Solver Simulation tool. Of course, it is also possible to build and run Monte Carlo experiments within an Excel spreadsheet without using any add-on software. But these tools make it more convenient to run such experiments in Excel-based models, particularly in terms of setting up the parameters for a simulation model, running the model, and visualizing its results. Monte Carlo simulation models have been used in many commercial applications. Examples include Procter & Gamble using these models to determine hedging foreign-exchange risks; Lilly using the model for deciding optimal plant capacity; Abu Dhabi Water and Electricity Company using @Risk for forecasting water demand in Abu Dhabi; and literally thousands of other actual case studies. Each of the simulation software companies' Web sites include many such success stories.

Discrete Event Simulation

Discrete event simulation refers to building a model of a system where the interaction between different entities is studied. The simplest example of this is a shop consisting of a server and customers. By modeling the customers arriving at various rates and the server serving at various rates, we can estimate the average performance of the system, waiting time, the number of waiting customers, and so on. Such systems are viewed as collections of customers, queues, and servers. There are thousands of documented applications of discrete event simulation models in engineering, business, and so on. Tools for building discrete event simulation models have been around for a long time, but these have evolved to take advantage of developments in graphical capabilities for building and understanding the results of such simulation models.

Conventional Simulation Inadequacies

Simulation is a well-established, useful, descriptive, mathematics-based method for gaining insight into complex decision-making situations. However, simulation does not usually allow decision makers to see how a solution to a complex problem evolves over (compressed) time, nor can decision makers interact with the simulation (which would be useful for training purposes and teaching). Simulation generally reports statistical results at the end of a set of experiments. Decision makers are thus not an integral part of simulation development and experimentation, and their experience and judgment cannot be used directly. If the simulation results do not match the intuition or judgment of the decision maker, a *confidence gap* in the results can occur.

Visual Interactive Simulation

Visual interactive simulation (VIS), also known as **visual interactive modeling (VIM)** and *visual interactive problem solving*, is a simulation method that lets decision makers see what the model is doing and how it interacts with the decisions made, as they are made. This technique has been used with great success in operations analysis in many fields such as supply chain and healthcare. The user can employ their knowledge to determine and try different decision strategies while interacting with the model. Enhanced learning, about both the problem and the impact of the alternatives tested, can and does occur. Decision makers also contribute to model validation. Decision makers who use VIS generally support and trust their results.

VIS uses animated computer graphic displays to present the impact of different managerial decisions. It differs from regular graphics in that the user can adjust the decision-making process and see results of the intervention. A visual model is a graphic used as an integral part of decision-making or problem solving, not just as a communication device. Some people respond better than others to graphical displays, and this type of interaction can help managers learn about the decision-making situation.

VIS can represent static or dynamic systems. Static models display a visual image of the result of one decision alternative at a time. Dynamic models display systems that evolve over time, and the evolution is represented by animation. The latest visual simulation technology has been coupled with the concept of virtual reality, where an artificial world is created for a number of purposes, from training to entertainment to viewing data in an artificial landscape. For example, the U.S. military uses VIS systems so that ground troops can gain familiarity with terrain or a city to orient themselves very quickly. Pilots also use VIS to gain familiarity with targets by simulating attack runs. The VIS software can also include GIS coordinates.

Visual Interactive Models and DSS

VIM in DSS has been used in several operations management decisions. The method consists of priming (like priming a water pump) a visual interactive model of a plant (or company) with its current status. The model then runs rapidly on a computer, allowing managers to observe how a plant is likely to operate in the future.

Waiting-line management (queuing) is a good example of VIM. Such a DSS usually computes several measures of performance for the various decision alternatives (e.g., waiting time in the system). Complex waiting-line problems require simulation. VIM can display the size of the waiting line as it changes during the simulation runs and can also graphically present the answers to what-if questions regarding changes in input variables.

The VIM approach can also be used in conjunction with artificial intelligence. Integration of the two techniques adds several capabilities that range from the ability to build systems graphically to learning about the dynamics of the system. These systems, especially those developed for the military and the video-game industry, have “thinking” characters who can behave with a relatively high level of intelligence in their interactions with users.

Simulation Software

Hundreds of simulation packages are available for a variety of decision-making situations. Many run as Web-based systems. *ORMS Today*, a publication of INFORMS, publishes a periodic review of simulation software. One recent review (current as of July 2022) is located at <https://pubsonline.informs.org/magazine/orms-today/2021-simulation-software-survey> (accessed July 2022).

As noted at the outset, modeling for optimization or simulation can require significantly more expertise than a chapter like this can offer as one topic in an introductory book.

The reader is urged to consult resources mentioned in this section for learning more about simulation and play with some basic capabilities in Excel.

► SECTION 8.9 REVIEW QUESTIONS

1. List the characteristics of simulation.
2. List the advantages and disadvantages of simulation.
3. List and describe the steps in the methodology of simulation.
4. List and describe the types of simulation.
5. Define *visual simulation* and compare it to conventional simulation.
6. Describe the features of VIS (i.e., VIM) that make it attractive for decision makers.

8.10 GENETIC ALGORITHMS AND DEVELOPING GA APPLICATIONS

(Note: This section is excerpted from an earlier edition of our sister book on Decision Support Systems). In this section, we introduce another technique that is commonly used in analytics. As noted in the context of linear programming or simulation, this section can also easily take one or more semesters of coursework just for this topic.

Genetic algorithms (GA) are a part of global search techniques used to find approximate solutions to optimization-type problems that are too complex to be solved with traditional optimization methods (which are guaranteed to produce the best solution to a specific problem). Genetic algorithms have been successfully applied to a wide range of highly complex real-world problems, including vehicle routing, bankruptcy prediction, and Web searching.

Genetic algorithms are a part of the machine-learning family of methods under artificial intelligence. Because they cannot guarantee the truly optimal solution, genetic algorithms are considered to be heuristic methods. Genetic algorithms are sets of computational procedures that conceptually follow the steps of the biological process of evolution. That is, better and better solutions evolve from the previous generation of solutions until an optimal or near-optimal solution is obtained.

Genetic algorithms (also known as **evolutionary algorithms**) demonstrate self-organization and adaptation in much the same way that biological organisms do by following the chief rule of evolution, *survival of the fittest*. The method improves the solutions by producing offspring (i.e., a new collection of feasible solutions) using the best solutions of the current generation as “parents.” The generation of offspring is achieved by a process modeled after biological reproduction whereby mutation and crossover operators are used to manipulate genes in constructing newer and “better” chromosomes. Notice that a simple analogy between genes and decision variables and between chromosomes and potential solutions underlies the genetic algorithm terminology.

Example: The Vector Game

To illustrate how genetic algorithms work, we describe the classical Vector game. This game is like MasterMind. As your opponent gives you clues about how good your guess is (i.e., the outcome of the fitness function), you create a new solution, using the knowledge gained from the recently proposed solutions and their quality.

DESCRIPTION OF THE VECTOR GAME Vector is played against an opponent who secretly writes down a string of six digits (in a genetic algorithm, this string consists of a *chromosome*). Each digit is a decision variable that can take the value of either 0 or 1. For example, say that the secret number that you are to figure out is 001010. You must try to guess this number as quickly as possible (with the least number of trials).

You present a sequence of digits (a guess) to your opponent, and he or she tells you how many of the digits (but not which ones) you guessed are correct (i.e., the fitness function or quality of your guess). For example, the guess 110101 has no correct digits (i.e., the score = 0). The guess 111101 has only one correct digit (the has no correct digits (i.e., the score = 0). The guess 111101 has only one correct digit (the third one, and hence the score = 1).

DEFAULT STRATEGY: RANDOM TRIAL AND ERROR There are 64 possible six-digit strings of binary numbers. If you pick numbers at random, you will need, on average, 32 guesses to obtain the right answer. Can you do it faster? Yes, if you can interpret the feedback provided to you by your opponent (a measure of the goodness or fitness of your guess). This is how a genetic algorithm works.

IMPROVED STRATEGY: USE OF GENETIC ALGORITHMS The following are the steps in solving the Vector game with genetic algorithms:

1. Present to your opponent four strings, selected at random. (Select four arbitrarily. Through experimentation, you may find that five or six would be better.) Assume that you have selected these four:

(A) 110100; score = 1 (i.e., one digit guessed correctly)
 (B) 111101; score = 1
 (C) 011011; score = 4
 (D) 101100; score = 3

2. Because none of the strings is entirely correct, continue.
3. Delete (A) and (B) because of their low scores. Call (C) and (D) parents.
4. “Mate” the parents by splitting each number as shown here between the second and third digits (the position of the split is randomly selected):

(C) 01:1011
 (D) 10:1100

Now combine the first two digits of (C) with the last four of (D) (this is called crossover). The result is (E), the first offspring:

(E) 011100; score = 3

Similarly, combine the first two digits of (D) with the last four of (C). The result is (F), the second offspring:

(F) 101011; score = 4

It looks as though the offspring are not doing much better than the parents.

5. Now copy the original (C) and (D).
6. Mate and crossover the new parents, but use a different split. Now you have two new offspring, (G) and (H):

(C) 0110:11
 (D) 1011:00
 (G) 0110:00; score = 4
 (H) 1011:11; score = 3

Next, repeat step 2: Select the best “couple” from all the previous solutions to reproduce. You have several options, such as (G) and (C). Select (G) and (F). Now duplicate and crossover. Here are the results:

(F) 1:01011
 (G) 0:11000
 (I) 111000; score = 3

(J) 001011; score = 5

You can also generate more offspring:

(F) 101:011

(G) 011:000

(K) 101000; score = 4

(L) 011011; score = 4

Now repeat the processes with (J) and (K) as parents, and duplicate the crossover:

(J) 00101:1

(K) 10100:0

(M) 001010; score = 6

That's it! You have reached the solution after 13 guesses. Not bad compared to the expected average of 32 for a random-guess strategy.

Terminology of Genetic Algorithms

A genetic algorithm is an iterative procedure that represents its candidate solutions as strings of genes called **chromosomes** and measures their viability with a fitness function. The fitness function is a measure of the objective to be obtained (i.e., maximum, or minimum). As in biological systems, candidate solutions combine to produce offspring in each algorithmic iteration, called a *generation*. The offspring themselves can become candidate solutions. From the generation of parents and children, a set of the fittest survive to become parents that produce offspring in the next generation. Offspring are produced using a specific genetic reproduction process that involves the application of crossover and mutation operators. Along with the offspring, some of the best solutions are also migrated to the next generation (a concept called **elitism**) in order to preserve the best solution achieved up until the current iteration. Following are brief definitions of these key terms:

- **Reproduction.** Through **reproduction**, genetic algorithms produce new generations of potentially improved solutions by selecting parents with higher fitness ratings or by giving such parents a greater probability of being selected to contribute to the reproduction process.
- **Crossover.** Many genetic algorithms use a string of binary symbols (each corresponding to a decision variable) to represent chromosomes (potential solutions), as was the case in the Vector game described earlier. **Crossover** means choosing a random position in the string (e.g., after the first two digits) and exchanging the segments either to the right or the left of that point with those of another string's segments (generated using the same splitting schema) to produce two new offspring.
- **Mutation.** This genetic operator was not shown in the Vector game example. **Mutation** is an arbitrary (and minimal) change in the representation of a chromosome. It is often used to prevent the algorithm from getting stuck in a local optimum. The procedure randomly selects a chromosome (giving more probability to the ones with better fitness value) and randomly identifies a gene in the chromosome and inverses its value (from 0 to 1 or from 1 to 0), thus generating one new chromosome for the next generation. The occurrence of mutation is usually set to a very low probability (0.1%).
- **Elitism.** An important aspect in genetic algorithms is to preserve a few of the best solutions to evolve through the generations. That way, you are guaranteed to end up with the best possible solution for the current application of the algorithm. In practice, a few of the best solutions are migrated to the next generation.

How Do Genetic Algorithms Work?

Figure 8.13 is a flow diagram of a typical genetic algorithm process. The problem to be solved must be described and represented in a manner amenable to a genetic algorithm. Typically, this means that a string of 1s and 0s (or other more recently proposed complex representations) are used to represent the decision variables, the collection of which represents a potential solution to the problem. Next, the decision variables are mathematically and/or symbolically pooled into a *fitness function* (or *objective function*). The fitness function can be one of two types: maximization (something that is more is better, such as profit) or minimization (something that is less is better, such as cost). Along with the fitness function, all the constraints on decision variables that collectively dictate whether a solution is a feasible one should be demonstrated. Remember that only feasible solutions can be a part of the solution population. Infeasible ones are filtered out before finalizing a generation of solutions in the iterations process. Once the representation is complete, an initial set of solutions is generated (i.e., the initial population). All infeasible solutions are eliminated, and fitness functions are computed for the feasible ones. The solutions are rank ordered based on their fitness values; those with better fitness values are given more probability (proportional to their relative fitness value) in the random selection process.

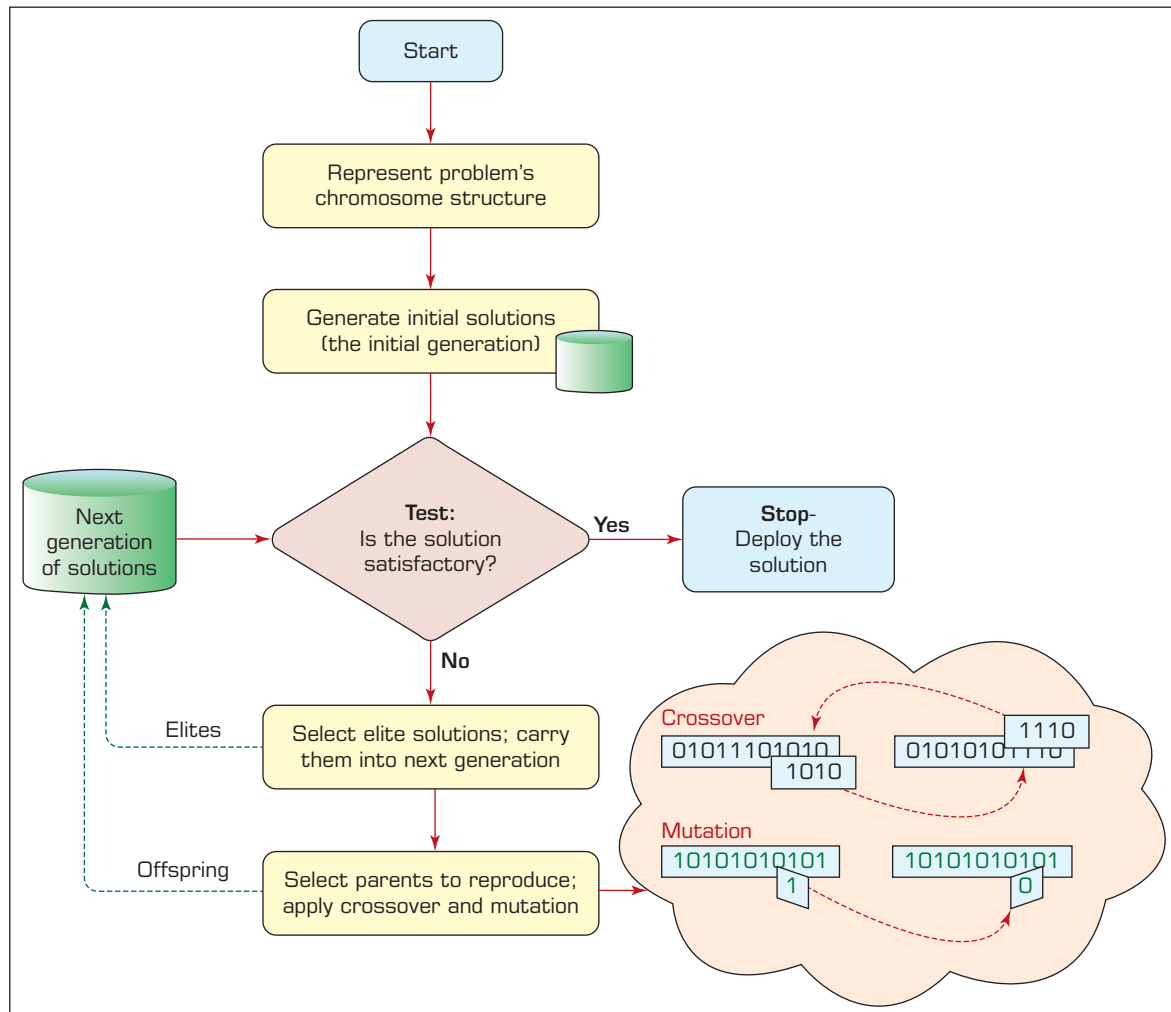


FIGURE 8.13 Genetic Algorithm Process.

A few of the best solutions are migrated to the next generation. Using a random process, several sets of parents are identified to take part in the generation of offspring. Using the randomly selected parents and the genetic operators (i.e., crossover and mutation), offspring are generated. The number of potential solutions to generate is determined by the population size, which is an arbitrary parameter set prior to the evolution of solutions. Once the next generation is constructed, the solutions go through the evaluation and generation of new populations for several iterations. This iterative process continues until a good enough solution is obtained (an optimum is not guaranteed), no improvement occurs over several generations, or the time/iteration limit is reached.

As mentioned, a few parameters must be set prior to the execution of the genetic algorithm. Their values are dependent on the problem being solved and are usually determined through trial and error:

- Number of initial solutions to generate (i.e., the initial population)
- Number of offspring to generate (i.e., the population size)
- Number of parents to keep for the next generation (i.e., elitism)
- Mutation probability (usually a very low number, such as 0.1%)
- Probability distribution of crossover point occurrence (generally equally weighted)
- Stopping criteria (time/iteration based or improvement based)
- The maximum number of iterations (if the stopping criteria are time/iteration based)

Sometimes these parameters are set and frozen beforehand, or they can be varied systematically while the algorithm is running for better performance. Many other tweaks can be made to the algorithm.

Example: The Knapsack Problem

The knapsack problem is a conceptually simple optimization problem that can be solved directly using analytical methods. Even so, it is ideal for illustrating a genetic algorithm approach. Say that you are going on an overnight hike and have a number of items that you could take along. Each item has a weight (in pounds) and a benefit or value to you on the hike (say, in U.S. dollars), and you can take one, at most, of each item (sorry, no partial items allowed—it’s all or nothing). There is a capacity limit on the weight you can carry (only one constraint, but there can be several measures and capacities, including volume, time, etc.). The knapsack problem has many important applications, including determining what items to carry on a space shuttle mission. For our example, there are seven items, numbered 1 through 7, with respective benefits and weights as follows:

Item	1	2	3	4	5	6	7
Benefit	5	8	3	2	7	9	4
Weight	7	8	4	10	4	6	4

The knapsack holds a maximum of 22 pounds. The string 1010100, with a total benefit or fitness of $7 = 4 + 4 = 15$, can represent a solution of items 1, 3, and 5.

We can set up the problem in an Excel worksheet, where we represent a solution as a string of seven 1s and 0s and the fitness function as the total benefit, which is the sum of the gene values in a string solution multiplied by their respective benefit coefficients. The method generates a set of random solutions (i.e., initial parents), uses the objective function (i.e., total benefit) for the fitness function, and selects parents randomly to create generations of offspring through crossover and mutation operations. Selection is statistically based on the parents’ fitness values. Higher values are more likely to be selected than lower ones. Palisade.com offers a package called Evolver, an easy-to-use Excel add-in genetic algorithm package (from Palisade Corp., palisade.com; demo available online)

to explore this further. It uses an innovative genetic algorithm to quickly solve complex optimization problems in finance, scheduling, manufacturing, and so on. Their website offers many case studies of applications in various industries.

Genetic Algorithm Applications

Genetic algorithms are a type of machine learning for representing and solving complex problems. They provide a set of efficient, domain-independent search heuristics for a broad spectrum of applications, including the following:

- Dynamic process control
- Induction of optimization of rules
- Discovery of new connectivity topologies (e.g., neural computing connections, neural network design)
- Simulation of biological models of behavior and evolution
- Complex design of engineering structures
- Pattern recognition
- Scheduling
- Transportation and routing
- Layout and circuit design
- Telecommunication
- Graph-based problems

A genetic algorithm interprets information that enables it to reject inferior solutions and accumulate good ones, and thus it learns about its universe. Genetic algorithms are also suitable for parallel processing.

Genetic algorithms are often used to improve the performance of other artificial intelligence methods, such as ES or neural networks. In neural networks, genetic algorithms dynamically adjust to find the optimal network weights.

► SECTION 8.10 REVIEW QUESTIONS

1. Define *genetic algorithm*.
2. Describe the evolution process in genetic algorithms. How is it similar to biological evolution?
3. Describe the major genetic algorithm operators
4. List major areas of genetic algorithm application
5. Conduct an online search to find three recent genetic algorithm applications. Discuss.

Chapter Highlights

- Models play a major role in DSS because they are used to describe real decision-making situations. There are several types of models.
- Models can be static (i.e., a single snapshot of a situation) or dynamic (i.e., multiperiod).
- Analysis is conducted under assumed certainty (which is most desirable), risk, or uncertainty (which is least desirable).
- Influence diagrams graphically show the interrelationships of a model. They can be used to enhance the use of spreadsheet technology.
- Spreadsheets have many capabilities, including what-if analysis, goal seeking, programming, database management, optimization, and simulation.
- Decision tables and decision trees can model and solve simple decision-making problems.
- Mathematical programming is an important optimization method.
- LP is the most common mathematical programming method. It attempts to find an optimal allocation of limited resources under organizational constraints.

- The major parts of an LP model are the objective function, the decision variables, and the constraints.
- Multicriteria decision-making problems are difficult but not impossible to solve.
- What-if and goal seeking are the two most common methods of sensitivity analysis.
- Many DSS development tools include built-in quantitative models (e.g., financial, statistical) or can easily interface with such models.
- Simulation is a widely used DSS approach that involves experimentation with a model that represents the real decision-making situation.
- Simulation can deal with more complex situations than optimization, but it does not guarantee an optimal solution.
- There are many different simulation methods. Some that are important for decision-making include Monte Carlo simulation and discrete event simulation.
- Genetic algorithms are search techniques that emulate the natural process of biological evolution. They utilize three basic operations: reproduction, crossover, and mutation.
- Reproduction is a process that creates the next-generation population based on the performance of different cases in the current population.
- Crossover is a process that allows elements in different cases to be exchanged to search for a better solution.
- Mutation is a process that changes an element in a case to search for a better solution.

Key Terms

certainty	goal seeking	multiple goals	sensitivity analysis
decision analysis	influence diagram	Mutation	simulation
decision table	intermediate result	reproduction	uncertainty
decision tree	variable	optimal solution	uncontrollable
decision variable	linear programming (LP)	parameter	variable
discrete event simulation	mathematical	quantitative model	visual interactive
environmental scanning	programming	result (outcome)	modeling (VIM)
and analysis	Monte Carlo simulation	variable	visual interactive
forecasting	multidimensional analysis	risk	simulation (VIS)
genetic algorithm	(modeling)	risk analysis	what-if analysis

Questions for Discussion

1. How does prescriptive analytics relate to descriptive and predictive analytics?
2. Explain the differences between static and dynamic models. How can one evolve into the other?
3. What is the difference between an optimistic approach and a pessimistic approach to decision-making under assumed uncertainty?
4. Explain why solving problems under uncertainty sometimes involves assuming that the problem is to be solved under conditions of risk.
5. Excel is probably the most popular spreadsheet software for PCs. Why? What can we do with this package that makes it so attractive for modeling efforts?
6. Explain how decision trees work. How can a complex problem be solved by using a decision tree?
7. Explain how LP can solve allocation problems.
8. What are the advantages of using a spreadsheet package to create and solve LP models? What are the disadvantages?
9. What are the advantages of using an LP package to create and solve LP models? What are the disadvantages?
10. What is the difference between decision analysis with a single goal and decision analysis with multiple goals (i.e., criteria)? Explain the difficulties that may arise when analyzing multiple goals.
11. Explain how multiple goals can arise in practice.
12. Compare and contrast what-if analysis and goal seeking.
13. Describe the general process of simulation.
14. List some of the major advantages of simulation over optimization and vice versa.
15. Many computer games can be considered visual simulation. Explain why.
16. Explain why VIS is particularly helpful in implementing recommendations derived by computers.

Exercises

Hands-on Exercises

- Create the spreadsheet models shown in Figures 8.3 and 8.4.
 - What is the effect of a change in the interest rate from 8% to 10% in the spreadsheet model shown in Figure 8.3?
 - For the original model in Figure 8.3, what interest rate is required to decrease the monthly payments by 20%? What change in the loan amount would have the same effect?
 - In the spreadsheet shown in Figure 8.4, what is the effect of a prepayment of \$200 per month? What prepayment would be necessary to pay off the loan in 25 years instead of 30 years?
- Solve the MBI product-mix problem described in this chapter, using either Excel's Solver or a student version of an LP solver, such as LINDO. LINDO is available from LINDO Systems, Inc., at lindo.com; others are also available—search the Web. Examine the solution (output) reports for the answers and sensitivity report. Did you get the same results as reported in this chapter? Try the sensitivity analysis outlined in the chapter; that is, lower the right-hand side

of the CC-8 marketing constraint by one unit, from 200 to 199. What happens to the solution when you solve this modified problem? Eliminate the CC-8 lower-bound constraint entirely (this can be done easily by either deleting it in Solver or setting the lower limit to zero) and re-solve the problem. What happens? Using the original formulation, try modifying the objective function coefficients and see what happens.

- Investigate via a Web search how models and their solutions are used by the U.S. Department of Homeland Security in the “war against terrorism.” Also investigate how other governments or government agencies are using models in their missions.
- This problem was contributed by Dr. Rick Wilson of Oklahoma State University.

The recent drought has hit farmers hard. Cows are eating candy corn! You are interested in creating a feed plan for the next week for your cattle using the following seven nontraditional feeding products: Chocolate Lucky Charms cereal, Butterfinger bars, Milk Duds, vanilla ice cream, Cap'n Crunch cereal, candy corn (because the real corn is all dead), and Chips Ahoy cookies.

	Choc Lucky Charms	Butterfinger	Milk Duds	Vanilla Ice Cream	Cap'n Crunch	Candy Corn	Chips Ahoy
\$/lb	2.15	7	4.25	6.35	5.25	4	6.75
Choc	YES	YES	YES	NO	NO	NO	YES
Protein	75	80	45	65	72	26	62
TDN	12	20	18	6	11	8	12
Calcium	3	4	4.5	12	2	1	5

Their per pound cost is shown, as is the protein units per pound they contribute, the total digestible nutrients (TDN) they contribute per pound, and the calcium units per pound.

You estimate that the total amount of nontraditional feeding products contributes the following amount of nutrients: at least 20,000 units of protein, at least 4,025 units of TDN, at least 1,000 but no more than 1,200 units of calcium.

There are some other miscellaneous requirements as well.

- The chocolate in your overall feed plan (in pounds) cannot exceed the amount of nonchocolate poundage. Whether a product is considered chocolate or not is shown in the table (YES = chocolate, NO = not chocolate).
- No one feeding product can make up more than 25% of the total pounds needed to create an acceptable feed mix.

- There are two cereals (Chocolate Lucky Charms and Cap'n Crunch). Combined, they can be no more than 40% (in pounds) of the total mix required to meet the mix requirements.

Determine the optimal levels of the seven products to create your weekly feed plan that minimizes cost. Note that all amounts of products must *not* have fractional values (whole numbered pounds only).

- This exercise was also contributed by Dr. Rick Wilson of Oklahoma State University to illustrate the modeling capabilities of Excel Solver.

National signing day for rugby recruiting season 2022 has been completed. Now, as the recruiting coordinator for the San Diego State University Aztec rugby team, it is time to analyze the results and plan for 2023.

You've developed complex analytics and data collection processes and applied them for the past few recruiting seasons to help you develop a plan for 2023. Basically, you have divided the area in which you

actively recruit rugby players into eight different regions. Each region has a per-target cost, a “star rating” (average recruit “star” ranking, from 0 to 5, like what Rivals uses for football), a yield or acceptance rate percentage (the percentage of targeted recruits who come to SDSU), and a visibility measure, which represents a measure of how much publicity SDSU gets for recruiting in that region, measured per target (increased visibility will enhance future recruiting efforts).

	Cost/target	Avg star rating	Acceptance rate %	Visibility per target
Region1	125	3	40	0
Region2	89	2.5	42	0
Region3	234	3.25	25	2
Region4	148	3.1	30	3
Region5	321	3.5	22	7
Region6	274	3.45	20	4
Region7	412	3.76	17	5
Region8	326	3.2	18	5.5

Your goal is to create a LINEAR mathematical model that determines *the number of target recruits you should pursue in each region* in order to have an estimated yield (expected number) of *at least* 25 rugby recruits for next year while minimizing cost. (Region 1 with yield of 40%: if we target 10 people, the expected number that will come is $0.4 \times 10 = 40$.)

In determining the optimal number of targets in each region (which, not surprisingly, should be integer values), you must also satisfy the following conditions:

- No more than 20% of the total targets (not the expected number of recruits) should be from any one region.
- Each region should have at least 4% of the total targets (again, not the expected number of recruits, but the number of targets).
- The average star rating of the targets must be at least equal to 3.3.
- The average visibility value of the targets must be at least equal to 3.5.
- Off on the recruiting trail you go!

6. This exercise was also contributed by Dr. Rick Wilson of Oklahoma State University.

You are the Water Resources Manager for Thirstiville, OK, and are working out the details for next year's contracts with three different entities to supply water to your town. Each water source (A, B, C) provides water of different quality. The quality assessment is aggregated together in two values P1 and P2, representing a composite of contaminants, such as THMs, HAAs, and so on. The sources each have a maximum of water that they can provide (measured in thousands of gallons), a minimum that we must purchase from them, and a per-thousand-gallon cost.

	MIN	MAX	P1	P2	COST
Source A	400	1,000	4	1	0.25
Source B	1,000	2,500	3.5	3	0.175
Source C	0	775	5	2.5	0.20

On the product end, you must procure water such that you can provide three distinct water products for next year (this is all being done at the aggregate “city” level). You must provide drinking water to the city, and then water to two different wholesale clients (this is commonly done by municipalities). The table below shows requirements for these three products, and the “sales” or revenue that you get from each customer (by thousand gallons, same scale as the earlier cost).

For each of the three water products/customers, MIN is the minimum that we have to provide to each, MAX is the maximum that we can provide (it is reasonable to be provided with a targeted range of product to provide to our customers), the maximum P1 and P2 weighted average for the water blended together for each quality “category” (the contaminants) per customer, and the sales price.

	MIN	MAX	P1	P2	SALES
Drinking	1,500	1,700	3.75	2.25	0.35
WSale 1	250	325	No Req.	2.75	0.4
WSale 2	No limit	No limit	4	2	0.425

Yes, the second wholesale customer (WSale 2) will take as much water as you can blend for them.

Obviously, water from all three sources will need to be blended to meet the Thirstiville customer requirements. There is one more requirement: for each of the three products (drinking water and the two wholesale clients), Source A and Source B both individually (yes, separately) must make up at least 20% of the total amount of the production of that water type. We do not have such a requirement for Source C.

Create an appropriate LP model that determines how to meet customer water demand for next year *while maximizing profit (sales less costs)*. Summarize your results (something more than telepathy—say, some sort of table of data beyond the model solution?) It must use words ☺ and indicate how much water we should promise to buy from our three sources. Integers are not required.

7. This problem was contributed by Dr. Rick Wilson of Oklahoma State University after a recent ice storm.

The state of Oklahoma has a backup plan in place when the grid infrastructure is threatened, and black-outs might result due to a lack of MW's. It has standing agreements from 7 MW-producing power plant facilities spread across 5 adjoining states. The agreements prescribe the per hour cost of using the power plant to generate MW's sent to OK, the rate of MW production at the plant per hour, and the maximum number of hours

of power generation that can be used for OK MW production. This information is provided in the table below.

Also in the table: an independently determined assessment of the quality of service provided by the power plant (this would be related to MW generation) and an assessment of rate stability in advent of a crisis

(this would be related to hours used, as there is a clause in the contract that allows the power plant to alter costs under certain emergency conditions).

An additional notation in the table indicates whether that power plant would serve the west section of Oklahoma's grid, or the east section.

Plant	CO1	CO2	NM	KS1	KS2	MO	AR
Cost/hr.	6.8	9.4	8.1	12.2	17.2	16	13.2
MW Rate/hr.	1	1.3	1.1	2.3	2.7	2.5	2.1
MAX hours	72	36	90	24	48	60	45
Region	W	W	W	W	E	E	E
QoS (MW)	6	6.5	7	8	5	5.75	7.5
Rate stability (Hrs)	2	7	4	6.75	4	3.5	5

CO () - Colorado;

NM - New Mexico;

KS () - Kansas;

MO - Missouri;

AR - Arkansas.

Create a plan (using an IP model) where you determine how many hours each plant should be operated for the benefit of Oklahoma to generate at least 500 MW from the 7 power plants that minimizes cost. Plan requirements:

- 1) Cannot exceed the maximum hours for each plant.
 - 2) Besides the 500 min requirement (you can exceed this), the plan should provide at least 150 MW to the West region, and at least 200 MW to the East region.
 - 3) The stability measure must average less than or equal to 3.85 (using hours as your metric).
 - 4) The QoS measure must average at least (or equal) to 6.4 (using MW's as your metric).
 - 5) Each state (not plant) must account for at least 15% of the total number of hours used.
 - 6) Integer hours only please.
8. This problem was contributed by Dr Rick Wilson of Oklahoma State University.

The Fallin Oil Company (FOC) owns 6 oil-producing wells in a region of Oklahoma that has been beset by a lot of seismic activity (specifically, earthquakes). While the scientific evidence is limited with regard to whether the well operation enhances the seismic activity, FOC is a good environmental citizen and has offered these six wells to be operated based upon a USGS-determined optimization model meant to simultaneously consider factors such as production, cost and risk to the environment (specifically, enhancing the likelihood for earthquakes). The model's use is experimental, and thus, FOC gets a sizeable grant from the State of Oklahoma to trial its use.

Each of the six wells (which have female nicknames: Abigail, Delana, Evelyn, Kelsey, Lucinda, and Marissa—we recognize that the wells are genderless, but it has been a common practice in this industry) possess operational parameters relevant to the study:

- Flow per hour operated (in barrels)
- Cost of operation per hour operated (Units)
- EQ—An EarthQuake enhancement measure—higher value means the more likely that the continued operation of this will cause enhanced earthquakes.
- MAX—The maximum hours that this well can be operated during the next operational period.
- Region—The sub region where the wells are located.
- Safety—the environmental “safety” of the well operation. (Higher is better than lower, scale of 1-100).

Create a LP model that determines the optimal number of hours to run each well that minimizes cost while providing at least 1,100 barrels of oil. Other requirements of the well production plan:

- Each well cannot operate more than its predefined MAX hours as stated.
- Additionally, each well ALONE cannot account for more than 30% of the total well production hours.
- Based on time (hours), the average EQ across the six wells cannot exceed 6.44. (This is a weighted average based upon hours used times the EQ measure).
- Based on time (hours), the average “safety” of oil produced must be at least 89. (Similar to above).
- Finally, each region cannot account for more than 55% of the total hours necessary to meet the production output requirements.

Hours MUST be whole numbers.

Wells (Code names)	Abigail	Delana	Evelyn	Kelsey	Lucinda	Marissa
flow - in barrels - per hour	2.7	3.4	4.1	3.1	5	3.9
cost - per hour	4.2	4.5	5.3	4.1	4.4	6
EQ - per hour	7	6	8.5	4.25	8	5
MAX hours	100	80	75	120	95	120
Region	1	1	1	2	2	3
Safety per hour	85	80	78	90	95	93

9. This problem was contributed by Dr. Rick Wilson of Oklahoma State University.

After the initial rush of Covid-19 vaccines, the domestic demand in the United States dropped significantly. The U.S. government started sharing the surplus vaccines

with other countries. Here is a fictitious example of this scenario. Find the least cost way to ship surplus vaccines from 4 supply locations in the United States (NY, LA, KC, ATL) to 7 geographic regions around the world (see table). These geographic regions have attributes of:

	EU	Middle East	Africa	SE Asia	Canada	Mexico	S America	MAX	VQUAL
NY	500	950	1200	1500	300	750	1250	2200	3
LA	900	1250	1400	975	355	450	1100	3140	3.5
KC	950	1150	1275	1255	275	475	1220	1150	3,25
Atlanta	750	1200	1325	1600	575	500	1200	2550	4
Need	4100	1100	3200	4700	2250	3600	2400		
POL	80	75	60	62	85	90	59		
PCT	24	15	5	9	22	11	7		

Need—maximum number of vaccines needed (in 000).
POL—political measure of importance to serve the region with vaccines.
PCT—percentage of population already vaccinated.
The 4 supply locations have attributes of:
MAX—maximum number of vaccines that can be shipped (in 000).
VQUAL—quality of a vaccine shipped from that location.
The 4 × 7 table shown (outlined) is the per vaccine (in 000) cost measure.
Note: The (000) is just for context and integer reasons—do NOT alter the numbers in your model.

Requirements for the shipping plan:

- a) Cannot exceed Need for a region.
- b) Cannot exceed MAX for a vaccine supply location, though you can ship less than MAX.
- c) Must send at least 20% of Need to each region.
- d) The overall distribution plan must have an average POL measure of at least 65 per unit shipped. This is an overall plan measure to at least meet a political threshold.
- e) The overall distribution plan must have an average PCT measure of no more than 9.5 per unit shipped. This is an overall plan measure to make sure the

supplies go to regions that have a lower existing vaccination rate.

- f) Each region must have vaccines shipped to them that average at least 3.5 in VQUAL. Again, this is each region individually.
- g) No more than 1500 units can be sent from any one supply location to any one geographic region.
- h) At the very end, force solutions to have integer values.

10. This problem was contributed by Dr Rick Wilson of Oklahoma State University.

Bless You, Inc. (BYI) is a nonprofit company that, among other things, provides medical treatment packs to overseas mission groups (consider this a simple example of how one could use analytics in a social or nonprofit setting—even though we’re going to assume that the WHO will reimburse you for sending these Medical Treatment Packs [MTP] to international locations). We are illustrating a proof-of-concept supply chain model for one of its products called MTP-A which as the name implies, represents a bundle of medicine and supplies that can treat 1,000 patients.

BYI produces MTP-A at four locations in the United States—Tulsa, Chicago, Atlanta and Dallas. It sends packets to 4 warehouses (Miami, Houston, LA, and Baltimore) where they are readied to be sent overseas. The

three overseas locations of interest in this application is North Africa (NAFR), West Africa (WAFR), and Central America (CAMER). Our goal is to find the optimal way to

supply the three overseas locations such that the WHO reimbursement less our costs are maximized (Revenue minus shipping costs).

QUALITY		Miami	Houston	LA	Baltimore	MAX MTP-A
4	Tulsa	22	17	23	20	1200
3.5	Chicago	23	30	31	18	1500
3.25	Atlanta	13	15	32	19	1350
3.75	Dallas	21	10	22	28	1175

CAPACITY	1500	1200	1600	1000
50%	750	600	800	500

	Miami	Houston	LA	Baltimore	REQUEST	REIMBURSE
NAFR	28	35	62	27	1252	50
WAFR	26	38	55	30	2000	52
CAMER	20	18	22	42	1476	38

Model parameters (shown in the table and outlined below).

Top cost matrix—per MTP-A unit from production facility to warehouse.

Bottom cost matrix—per MTP-A unit from warehouse to international location.

MAX MTP-A—maximum amount of MTP-A that can be produced at that facility.

Quality—average quality of an MTP-A unit produced at the facility.

Capacity—maximum MTP-A units that the specific warehouse can handle.

REQUEST—maximum number of MTP-A's we can ship to the specific international location.

REIMBURSE—reimbursement received from the WHO for each MTP-A unit sent to that international location.

Other model requirements:

Besides the supply, demand, and capacity requirements, there are other restrictions that we must place on our supply chain solution.

- Each warehouse must use at least 50% of its maximum capacity (shown in the table). Obviously, it can also not use MORE than 100% of the capacity.
- Each international location must be shipped at least 75% of its maximum requested numbers of MTP-A's.
- The quality of the MTP-A's sent to each warehouse must have a weighted average value of at least 3.55. Note that is at EACH warehouse.

D) For the production facility to warehouse portion of the supply chain, no more than 1,000 MTP-A's can be sent from any one production facility to any one warehouse.

E) For the warehouse to international portion of the supply chain, there are limits on how many MTP-A's can be sent from a specific warehouse to a specific international location. The limits depend on the warehouse—Miami = 900, Houston = 700, LA = 900, Baltimore = 800. Again, this is a specific warehouse to specific international location limit, not a limit on an entire warehouse.

When you have a working model, force integers.

Create an appropriate LP model to determine the optimal way to distribute MTP-A's from where they are produced, to where they are warehoused to where they are needed. The objective is to maximize the NET returns—REIMBURSEMENT less the shipping costs. All requirements must be met as described above, and no inventory is allowed in the warehouses. The usual smell test must be maintained as well (products cannot magically appear in the supply chain).

Summarize your model solution with either a picture or some other visual way that describes the flow of MTP-A's in your supply chain. Nothing too elaborate—just a way to “convey the solution” to upper management.

**This problem is inspired by the real company Blessings International (blessing.org). Bless them! **

11. Canadian Cannabis Consortium Company (CCCC) wishes to develop its initial product planning design as it takes advantage of new legalization laws in Canada and the United States for cannabis products. It expects to expand and modify their cannabis distribution in the future, but they wish to maximize their profitability initially prior to the new market hitting equilibrium.

CCCC has 5 production facilities/farms (Code named F1 through F5) where raw cannabis is grown. The Farms' (Fx) cannabis attributes include the maximum crop available (lbs), the per lb. growing cost, the THC measure per lb., and the CBD measure per lb. As a policy, the minimum amount of cannabis that can be used from each Farm is 60% of the maximum. And don't forget to model the maximum as well. The data are shown on the next page.

DATA				
	MAX	Cost	THC	CBD
F1	350	27	13	6
F2	425	31	9	8
F3	525	35	10	9
F4	800	26	15	10
F5	400	37	11	8.8

	Sales	Manu cost
PAIN	46	9
RELAX	40	8
FOCUS	54	13

CCCC will take cannabis from the five farms and blend them together to make three products—PAIN, RELAX and FOCUS. Each product has a different sales price (per lb.) and a different cost to manufacture (also per lb.). Additionally, each product has different requirements in terms of the weighted average value of THC and CBD.

PAIN: THC levels must average less than or equal to 11.2.

RELAX: THC levels must average greater than or equal to 12.5

FOCUS: THC levels must average less than or equal to 10.1

CBD: All three products must INDIVIDUALLY average at least 8.7.

PAIN: an individual farm cannot provide more than 35% of the total cannabis used in PAIN.

RELAX: an individual farm cannot provide more than 32.5% of the total cannabis used in RELAX.

FOCUS: does not have this requirement.

Finally, each of the three products must make up at least 25% of total production (measured in lbs.). Additionally, you cannot make more than 1,000 lbs. of any one product.

CCCC wishes to determine how to optimally design the production of its 3 products using the cannabis grown at its 5 farms. All costs and revenues are measured on a per pound basis. Costs are production costs and growing costs (see data provided). Revenues come from total product made times the sales price. The goal of your model is to maximize revenue less costs.

At the very end, when your model is working well, save the solution without integers and then force the variables to be integers. If it takes longer than 3 minutes, kill and save that solution. Summarize your solution with some sort of rudimentary picture, graphic or brief discussion.

12. The problem was also contributed by Dr Rick Wilson of Oklahoma State University.

Each year, there are many complaints about how the NCAA Men's Basketball Committee seeds and schedules teams in the 68-team single elimination tournament to determine the National Champion. It is inevitable that somebody is unhappy. We will approach a portion of the task by using our modeling skills to generate an alternative schedule that will likely differ from the actual assignments. Your task is twofold: 1) generate a model that "solves" the situation using OUR parameters (which is basic scheduling common sense IMO) and then 2) do a thorough comparison of the actual tournament assignments and the results of your model.

Create an LP model that will assign teams to regions. You will be using only the top 24 teams (the #1 thru #6 seeds) and all 4 regions. The objective of your model for assigning teams will be to minimize the sum of distance from the location of each team to their assigned region. Distance data are provided on a separate spreadsheet.

The restrictions for team assignment to regions follow below. These are in the spirit of common sense and the NCAA tournament committee "rules," but more much holistic than their "micro" approaches.

- Each of the four regions (West—San Francisco, South—San Antonio, East—Philadelphia, Midwest—Chicago) will have exactly one #1 seed assigned, exactly one #2 seed assigned, exactly one #3 assigned, exactly one #4 assigned, exactly one #5 assigned and exactly one #6 assigned.
- Teams from the same conference cannot be assigned to the same region (unless there are more than 4 teams from the same conference—the SEC in our case). Conferences are shown on the data file (it is a team attribute). Also, don't worry about conferences that do not have multiple teams (i.e., IGNORE THEM!!!).

For the SEC—make sure each region has at least 1 SEC team, but no more than 2 SEC teams assigned.

- 3) The “too-close” rule—a loose adaptation of a true NCAA rule. Do not allow St. Mary’s to be assigned to the West region. But it is okay to allow (not force) Villanova to be assigned to the East region. I can explain further if this bothers you in any way.
- 4) “Marquee” Value—“Q”—There are 9 teams that have a “Q” factor—they are either marquee teams or teams that have captured the interest of the basketball fanbase because of some unique aspect of the team. This is scaled on a value of 1-10.

There are two different requirements that must be met BY REGION regarding the Q factor.

- a. Each Region must have at least 2 teams that have a Q factor.
- b. For the teams assigned to the region that have a Q factor, their average value must be ≥ 4.2 . Note, the 15 teams that do not have a Q factor do NOT factor into this constraint.

Part A—THE MODEL—Implement an appropriate linear programming model that assigns the 24 teams to Regions, minimizing the sum of overall distances subject to the items listed above. Suggestion: Attack modularly AND model efficiently. What do we mean “efficiently”? If you are a little sloppy with your constraints, or include unnecessary or duplicative constraints, you may exceed the 100-constraint limit imposed by the Solver. Model carefully.

Part B—THE COMPARISON—Compare your solution to the NCAA’s assignments. Keep in mind neither

model is necessarily “better” and each approach is using different criteria. NOTE: Don’t just forget this part. At least 20% of your grade will be based on a thorough comparison of your model solution to the NCAA actual bracket.

Specifically, measure the following for both your solution and the actual assignments (obviously, only the 24 teams of interest).

- Miles overall and individually for the #1 seeds, #2 seeds, #3 seeds, #4 seeds, #5 seeds, and #6 seeds.
- The number of regions where multiple teams from the same conference are assigned (do not consider the SEC). (Your solution should have a measure of “0”).
- The number of regions where the SEC assignments (at least 1 but no more than 2) are violated (Your solution should have a measure of “0”).
- The number of regions that do NOT satisfy the Weighted Q requirements (Your solution should have a measure of “0”).
- The number of regions that do NOT satisfy the minimum number of marquee teams’ requirement (Your solution should have a measure of “0”).

A simple summary of the team assignments in a nice understandable format would also be helpful.

NOTE: The data file shows the actual NCAA assignments—in yellow.

13. Solve the knapsack problem from Section 8.10 manually, and then solve it using Evolver. Try another code (find one on the Web).
14. Search online to find vendors of genetic algorithms and investigate the business applications of their products. What kinds of applications are most prevalent?

References

- Arsham, H. (2006a). “Modeling and Simulation Resources.” *home.ubalt.edu/ntsbarsh/Business-stat/RefSim.htm* (accessed July 2022).
- Arsham, H. (2006b). “Decision Science Resources.” *home.ubalt.edu/ntsbarsh/Business-stat/Refop.htm* (accessed July 2022).
- Banks, J., & Gibson, R. R. (2009). “Seven Sins of Simulation Practice.” *INFORMS Analytics*, 24–27. *www.analyticsmagazine.org/summer-2009/193-strategic-problems-modeling-the-market-space* (accessed July 2022).
- Christiansen, M., K. Fagerholt, G. Hasle, A. Minsaa, & B. Nygreen. (2009, April). “Maritime Transport Optimization: An Ocean of Opportunities.” *OR/MS Today*, 36(2), 26–31.
- Clemen, R. T., & Reilly, T. (2000). *Making Hard Decisions with Decision Tools Suite*. Belmont, MA: Duxbury Press.
- Farasyn, I., K. Perkoz, & W. Van de Velde. (2008, July/ August). “Spreadsheet Models for Inventory Target Setting at Procter & Gamble.” *Interfaces*, 38(4), 241–250.
- Goodwin, P., & Wright, G. (2000). *Decision Analysis for Management Judgment*, 2nd ed. New York: Wiley.
- Hurley, W. J., & M. Balez. (2008, July/August). “A Spreadsheet Implementation of an Ammunition Requirements Planning Model for the Canadian Army.” *Interfaces*, 38(4), 271–280. *ingrammicrocommerce.com*, “CUSTOMERS,” *https://www.ingrammicrocommerce.com/customers/* (accessed July 2022).
- Kearns, G. S. (2004, January–March). “A Multi-Objective, Multicriteria Approach for Evaluating IT Investments: Results from Two Case Studies.” *Information Resources Management Journal*, 17(1), 37–62.
- Kelly, A. (2002). *Decision Making Using Game Theory: An Introduction for Managers*. Cambridge, UK: Cambridge University Press.
- Koksalan, M., & S. Zionts. (Eds.). (2001). *Multiple Criteria Decision Making in the New Millennium*. Berlin: Springer-Verlag.
- Kontoghiorghe, E. J., B. Rustem, & S. Siokos. (2002). *Computational Methods in Decision Making, Economics, and Finance*. Boston: Kluwer.
- Ovchinnikov, A., & J. Milner. (2008, July/August). “Spreadsheet Model Helps to Assign Medical Residents at the University of Vermont’s College of Medicine.” *Interfaces*, 38(4), 311–323.

Landscape of Business Analytics Tools

LEARNING OBJECTIVES

- Become familiar with the landscape of analytics and data science tools and platforms
- Learn about the simple classification of the analytics tools
- Know the advantages and disadvantages of free versus paid analytics tools
- Learn how to objectively select the best tool for a given analytics project
- Understand the nuances between open-source and closed-source/proprietary analytics tools
- Know the advantages of cloud-based analytics platforms over the desktop tools
- Become familiar with free-open-source analytics and data science tools
- Develop better understanding of a variety of analytics tools through tutorials
- Know the most popular proprietary analytics tools in the market

Business analytics (or data science, or data analytics, etc., the name changes based on one's preference) has been the name of the game quite some time in both businesses and academic circles, and there is no sign of slowing down anytime soon. To keep up with this fast growing/evolving technology, universities are rushing to create new degree programs while businesses are creatively coming up with new departments and roles/titles to take advantage of the momentum. Because of this tremendous interest in business analytics and data science, myriad of software tools and computing platforms are being developed (by startups as well as the well-established IT companies) and serviced to satisfy the needs and wants of the variety of analytics communities. This chapter provides an overview of the landscape of the existing analytics and data science tools and platforms. Although, we aim to have a comprehensive coverage, because there exists a very large number of analytics and data science tools and platforms in the marketplace, some of the tools may be unintentionally omitted. Also, perhaps because of the ease and convenience in accessing the necessary tools and resources, this chapter portrays a positive bias toward free and **open-source software** tools by providing a more detailed characterization along with step-by-step tutorials, in most cases.

- 9.1 Opening Vignette: How Seagate Is Using KNIME to Tackle the Digital Transformation 495
- 9.2 Importance of Analytics Tools 497
- 9.3 Free and Open-Source Analytics' Programming Languages 505
- 9.4 Free and Open-Source Analytics' Visual Tools 527
- 9.5 Commercial Analytics Tools 552

9.1 OPENING VIGNETTE: How Seagate Is Using Knime to Tackle the Digital Transformation

FROM DESCRIPTIVE TO PREDICTIVE ANALYTICS

Seagate—a global leader in design and manufacturing of disk drives, flash storage, and other data storage and management solutions—has recognized that the digital transformation is something that cannot be ignored. Every day, at one of Seagate's many global sites, people are interacting with or requiring results out of data. Moving up the analytics maturity curve—from descriptive to predictive analytics—was an important strategic objective. To achieve this, a tool that could easily be rolled out across the entire organization, was simple for others to learn, and could slot in with existing tools and infrastructure such as Excel, JMP, Minitab, **Tableau**, Matlab, **Python**, and R was needed.

In 2017, KNIME was selected as the tool of choice for Seagate's data needs. The company implemented a tailored and unique Citizen Data Scientist (CDS) training program to train employees in using not only **KNIME Analytics Platform**, but also other complementary data science tools. The program consisted of a mix of online and onsite trainings as well as workshops to teach users and advocate for the CDS program. Some of those attendees are now trainers and advocates themselves.

To help excite and encourage Seagate employees, the main messages shared at these trainings and workshops were the importance of (1) moving with the digital transformation to stay relevant in the industry and (2) upskilling employees to become citizen data scientists and to enable them to get the most out of their data, independently.

CREATING DEMOS IN THREE DAYS

With new products or product ideas, there is always an issue with the data. Some ideas are researched for many years, with different reasons for delay. For a long time, the US R&D engineers had a long wait time to get the data they needed. KNIME Analytics Platform was used to demo the idea of pulling the data from Asia with a multithreading scheme. It took approximately two weeks to fine-tune everything, but this action is what really started to highlight the benefit of the CDS program and, more importantly, the value of using KNIME Analytics Platform. What people had been struggling with for almost a year, had taken just a few days to develop and demo as a solution. The magic behind it is that KNIME allows for multithread parallel queries by simply dragging and dropping nodes onto workflow canvas that would otherwise need sophisticated coding skills.

ENABLING TEAMS TO BE MORE FLEXIBLE AND CREATIVE

In another case, KNIME is enabling the Research and Product Development team to be more creative and productive. The team's goal is to provide creative product solutions. There is no such thing as routine because there's always the need to change and adapt to the way data is analyzed. Furthermore, there is always a new way of doing things, new data to read, new structures, new formats, and more.

KNIME Analytics Platform is the perfect solution for prototyping new ideas and presenting these ideas and results quickly. “I would describe prototyping in KNIME as similar to building with Lego,” says Debin Wang, Staff Engineer. “There are many different building blocks to choose from.” KNIME, just like Lego, encourages individuals to be creative and imaginative when building something. And the best part: when R&D teams have a new requirement or change in the data structure, the KNIME workflow can quickly and easily adapt to it. In R&D, not everything works out—especially not the first time. KNIME helps to significantly shorten the exploratory cycle. What previously took an engineer weeks in line-by-line coding, can now be done in a few days. This is because KNIME, compared to other data science tools, is more flexible and intuitive.

ENABLING SAVINGS OF OVER \$1,000,000

KNIME is used by the Recording Head Engineering Group for the dynamic modeling of downstream metrics wafers to sliders. Recording head manufacturing is very complex. The sequential layering process includes more than 1,500 steps, which form patterns of electrical conductors and magnetic material on a ceramic disk (wafer). A single wafer takes more than four months to complete. Even then, it can still require subsequent processing downstream in the supply chain before testing takes place, which could highlight processing issues. One of these 1,500 steps previously had an average duration of six months between processing and testing. That meant six months of potentially at-risk material before a fault was detected. Using KNIME, an advanced modeling workflow was created to accurately predict expected results. The user-friendly nature of KNIME also enabled Seagate to integrate the model into the existing wafer fab control system. This reduced the feedback loop from six months to four weeks, and enabled savings of over \$1,000,000 for this single business area.

In another example, a similar methodology was used. Here the team created a KNIME workflow, which resulted in saving \$300,000 worth of scrap materials. This workflow was able to predict the materials required for future processes in a different process area, which would have otherwise been scrapped under the existing univariate process control (SPC) system.

In a third example, the team reduced the time spent in ensuring that the two wafer fabrication facilities (Minnesota and Northern Ireland), were completely in sync. Existing systems required a significant number of monitoring hours per week. A KNIME workflow was built to cater for the high number of false positives that the team was dealing with—eliminating them from the review process and saving valuable time each week.

OVER 150 USERS WORKING INDEPENDENTLY WITH DATA

The Citizen Data Scientist program has been extremely effective. Since launching in 2017, there are over 100 general users (learners, practitioners, and analysts) and almost 50 power users (KNIME/Seagate evangelists). Employees are working more independently with their data and getting better insights, faster. They are also able to generate significant business savings—in terms of both time and money—by developing workflows and solutions to overcome business challenges or pain points. Moreover, the feeling of empowerment has been a significant motivator for Seagate employees. “KNIME has empowered people who previously may have not considered the discovery and application of machine learning techniques to dip their toes into the world of data science” says Brendan Doherty, Staff Data Scientist, Seagate Technology.

In 2019, Seagate purchased a KNIME Server, and it is predicted that the number of both general and power users will continue to increase. Looking forward, the CDS program will likely involve **AWS** so that learners have access to KNIME Server applications on AWS in order to do sandbox or development work. Seagate is already getting into the next level of maturity with predictive analytics and is starting to see tremendous business impact.

► QUESTIONS FOR THE OPENING VIGNETTE

1. How does KNIME help in digital transformation efforts at Seagate?
2. What is citizen data scientist? Why is it important?
3. In this application case, what do you think is the most prevailing advantage of KNIME Analytics Platform? Discuss your answer.
4. What were the three cost-savings mentioned in the case? What were the contributions of a visual tool like KNIME in these savings?
5. List and briefly discuss three different advantages of visual programming tools (like KNIME) in analytically motivated organization (like Seagate).

WHAT WE CAN LEARN FROM THIS VIGNETTE

This chapter is about the broad and varied landscape of analytics and data science tools. As you will see herein, although there is a significant overlap among all of these tools in terms of analytics functionality, each software tool/platform comes with its differentiating characteristics, advantages and limitations. Regardless of the task, let it be digital transformation, better customer engagement, pricing and revenue maximization, or fraud detection, the goal is to use one of more of these analytics and data science tools to turn big data into big insight to achieve the desirable outcomes. Complex projects often require the use not just one but several analytics and data science tools and platforms to successfully produce the much-needed actionable outcomes within the feasible time window. That is, an optimal blend of analytics tools would lead to faster and better decision support. The opening vignette summarized how an advanced technology company can use a mix of free and open-source analytics tool and commercial cloud-based IT platforms to streamline its digital transformation initiatives.

Source: KNIME Customer Success Stories. “How Seagate is using KNIME to Tackle the Digital Transformation” retrieved from <https://www.knime.com/solutions/success-story/how-seagate-is-using-knime-to-tackle-the-digital-transformation> (accessed July 2022); and <https://www.knime.com/about>.

9.2 IMPORTANCE OF ANALYTICS TOOLS

Because of the tremendous growth in the popularity of business analytics and data science that we have been experiencing in recent years (both in academia and industry), the landscape of tools used to implement for these technologies are also evolving and expanding at an unprecedented pace. As can be seen in the rest of this chapter, there already are numerous analytics software tools, ranging from commercial/paid to free/open-source, from graphical workflow-type to language-based programming platforms. Some of these tools are cloud-based, some are individual local installations, and yet, some are both—providing local tools and complementary cloud-based platforms for computational enhanced efficiency and ease of model deployment. Some of these tools may be better at data preprocessing and/or flexible and powerful model building, while others may be superior in terms of ease of learning, use and deployment. As evidenced from the continually expanding landscape of tool offerings, there is no one tool that is best at everything. Most business analytics and data science professionals tend to use a portfolio of tools and programming languages collectively to produce the best possible outcomes for their data analytics projects.

The success of data scientists at work depends largely on the tool (or perhaps more accurately, the toolbox) they rely on. The mathematical and statistical knowledge of the underlying algorithms, the experience with the most effective analytics platforms, the domain knowledge/wisdom, team/people management skill, are all expected, important, necessary components for a prolific data scientist and a successful data science project.

However, there are other more contingent and complementary factors that also influence the final outcome and the impression left by the data science project on the stakeholders.

Clearly, every project has a deadline to comply, a list of requirements to meet, and a budget to operate within. Most data analytics projects are designed so stringently that it does not leave much room for sandboxing or an extended trial-and-error type exploration. Data scientist is expected to identify and implement a solution in a short time while making sure that it complies with all requirements and constraints—it meets and desirably exceeds the needs and wants of the stakeholder (often a customer/manager within the company). To do so, he/she may need to quickly experiment with different techniques to identify and adopt the best possible roadmap and repeated mechanisms for the project. Of course, every project has a budget too. The rapid implementation of the right solution is often additionally constrained by a limited budget.

Some projects are quite complex and require specialized algorithms beyond just the classic general machine learning algorithms. Sometimes, data scientists are forced to learn new techniques and new algorithms—that are required by the specific project—on the spot, and given the deadline, the learning curve must develop within a very short time. In this case, the more time data scientist has to dedicate to theory and math, the faster the learning will proceed. Agility and creativity must be blended for a “perfect” solution.

Finally, let us consider the deployment phase of the analytics continuum. Despite the fact that this is where the “perfect” solution that we built is put in use so that it could make a difference in the intended parts of the organization, traditionally, in academia, in our analytics courses, we don’t allocate enough time and effort to cover this phase in due details. This lack of coverage may be attributed to the fact that every system has its own specific technical structure, deployment procedures, integration mechanisms, and security policies. The passage from prototype to production must be precise, rapid, and as secure. Data scientist cannot degrade the parts of the solution while moving it into production or take the risk of exposing it to less secure execution, considering that the solution will be consumed by a different set of less expert users. If and when required, a different set of solutions, with more or less degrees of interactions and sophistication levels, with more secure framework, can and should be made available for the consumption of the larger user base.

As can be deduced, many other contingent factors, such as ease of learning, speed of prototype implementation, debugging and testing options to ensure the correctness of the solution, flexibility to experiment with different approaches, availability of help from external contributors and experts, and finally automation and security capabilities, contribute to the success of the data science project, beyond the math/stat/AI, the analytics experience, and the depth of domain knowledge. All of these contingent factors depend heavily on the tools and the data science platforms the data scientists choose to use.

A Multidimensional Categorization of Analytics Tools

1. **Free versus Paid/Commercial.** In recent years, we have been witnessing tremendous interest in developing free, community supported analytics projects that rely heavily on crowdsourcing. Spearheaded by R and Python, this trend also spread to other nonprogramming-based free-of-charge analytics tools and platforms (e.g., **Weka**, KNIME, Orange). Historically, we have always had access to free, but scaled-down, time limited, or capability constrained, versions of commercial analytics tools. The free tools that we are considering here are not of this type; they are free for full version with no restriction on time of capabilities, and free for all purposes (educational or commercial). Well, as the saying goes, “there is no such thing as a free lunch!” may also apply here, if one thinks every free tool is in fact truly free.

Some of these tools may be overly complicated and technical with very little or no learning/educational/training material, which may lead to prohibitively huge spendings, killing the free nature of the tool.

2. **Open versus Closed Source.** Open-source software refers to the computer software whose source code is open to the general public to access, validate, modify, improve, and extend. Within the specified conditions of the open-source license, one can reuse what is openly available to modify, extend, and recreate his/her own software product for a specific purpose. Often, the software created from open-source code is made open source too (either voluntarily or as the requirement of the open use license). On the other hand, with closed source software (also known as **proprietary software**), the public is not allowed with access to the source code, so they cannot see or modify it in any way. **Closed source software** is usually comes in compiled/binary representation that hides not only its logical characteristics but also the programming code used to create it, rendering to a complete black box. Also, open source may not mean 100% everything is open to the lowest level of granularity. That is, an open-source software may have precompiled, binary components that with no source code availability.

There is a difference between open source and open architecture. Open architecture, it may or may not be open source, intentionally designed to allow to interact with other software tools. This interaction can change from a simple exchange of inputs/outputs to a full consumption of the other tool's modeling capabilities in an embedded framework. For instance, KNIME Analytics Platform is designed produce specially formatted outputs files that can easily be consumes by cloud platforms of Tableau and **PowerBI** and it also replicates/consumes all of machine learning functionalities of H2O and Weka.

Recently, in analytics and data science tools marketplace, perhaps because of the popularity of R and Python, there seems to be a positive trend for open source and/or open architecture tools and platforms. Most of the proprietary tools are embracing the open-source momentum in data science by making their tools receptive and friendly to the consumption of other open source platforms, especially to R and Python.

3. **Local versus Server/Cloud.** Converting high volumes of variety of data sources into information and knowledge within a reasonable time frame requires highly capable computational resources. Therefore, in the age of big data that we are living, real-world business analytics and data science projects must be executed on the cloud—some type of private/public/hybrid cloud/server platform that can intelligently pool storage and processing capabilities to create fast and accurate actionable insight. Often, the design phase of a business analytics project take place on a local workstation, and once validated on a similar but smaller/simpler data source, it is moved and fully executed on the cloud-based analytics platform. Cloud platform can also serve as the deployment and management/versioning place for analytics models.

Analytics tools can be classified as (1) completely local for everything (e.g., Orange, Weka, JMP), (2) local for model design and development and cloud for deployment (e.g., KNIME, **RapidMiner**, Tableau, PowerBI), and (3) cloud for everything (e.g., SAS Viya, AWS Analytics, Microsoft Azure, **Google Colab**). Based on the needs of the organization, any one of these three options can be adopted and successfully used.

4. **Visual versus Syntactic.** The tools in business analytics and data science are evolving in two directions: programming-based tools (reminding us of the good old days where the analytic is done using low level codes like C/C++ or high-level codes like SAS code) and workflow-based visual tools. In the current analytics tools landscape, the two most popular analytics tools, namely R and Python, are programming based

while many of the popular commercial tools are visual, workflow (aka, pipeline, logic-flow) (e.g., KNIME, RapidMiner, Orange, **SAS Enterprise Miner**, **IBM SPSS Modeler**, etc.).

While R is a high-level statistical analysis and data visualization language, Python (despite its reputation as the number one tool for analytics and data science) is a general-purpose programming language that can be used for any software development task. While performing data science, one can use either of these tools to read the data, process the data, create visualizations, build and test models, using readily available, virtually unlimited functions created and made available in these programming platforms. That is, when you are writing R or Python code, you are just connecting the already written codes/functions in a logical manner with a small number of lines. What matter here is to be aware of the large list of optional codes/functions and their syntactic details to make smart decision.

Visual, workflow-type tools are also using the same or similar functions and low-level codes underneath those widgets and nodes. The main difference is that, in a programming language you are to make sure of the accurate selection and implementation (paying attention to the syntactic details) of the functions you are consuming, while in visual tool you are consuming the same functions via simple drag-and-drop gestures. While in programming languages, you have more control over the details of which function to use and how to use it, in visual tools you are sparing yourself from syntactic details, and instead, spending your time and mental energy to better modeling and experimentation.

Table 9.1 provides a comparison of some of the analytics tools in the above listed four comparative dimensions. As can be seen, some of the tool in some columns possess both characteristics.

TABLE 9.1 Comparison of Some Analytics Tools in Four Different Dimensions

	Free vs. Paid	Open vs. Close	Local vs. Server	Visual vs. Program
Python	F	O	L+S	P
R	F	O	L	P
KNIME	F	O	L+S	V
RapidMiner	F/P	O	L+S	V
Orange	F	O	L	V
Weka	F	O	L	V
Tableau	P	C	L+S	V
PowerBI	F/P	C	L+S	V
SAS Viya	P	C	S	V/P
SAS EM	P	C	L	V
JMP	P	C	L	V
IBM Modeler	P	C	L	V
IBM Watson	P	C	S	V
TIBCO Statistica	P	C	L+S	V
Teradata Vantage	P	C	S	V

Popularity of the Analytics Tools

So, which tools and their creator companies are the best and/or most popular among the data scientists? There is no generally accepted answer to this question. The answer depends on you ask, what criteria you use, and what weights you assign to each criterion. Following is just three of many reputable sources one can find and use to answer these questions. The first two are the leading IT research and consultancy companies, and the last one, is a pervasive web-based portal (i.e., an information repository) for everything related to data mining, business analytics, data science, and machine learning.

GARTNER Gartner, Inc is a technological research and consulting firm based in Stamford, Connecticut that conducts research on technology and shares this research both through private consulting as well as executive programs and conferences. **Magic Quadrant** is a series of market research reports published by Gartner that rely on proprietary qualitative data analysis methods to demonstrate market trends, such as direction, maturity and participants. Figure 9.1 shows one of Gartner's Magic Quadrant for Advanced Analytics Platforms. As can be seen, SAS, IBM and Dell (on the commercial and proprietary tools side) along with KNIME and RapidMiner (on free and open-source tool side) are listed as leaders.



FIGURE 9.1 Gartner's Magic Quadrant for Advanced Analytics Platforms.

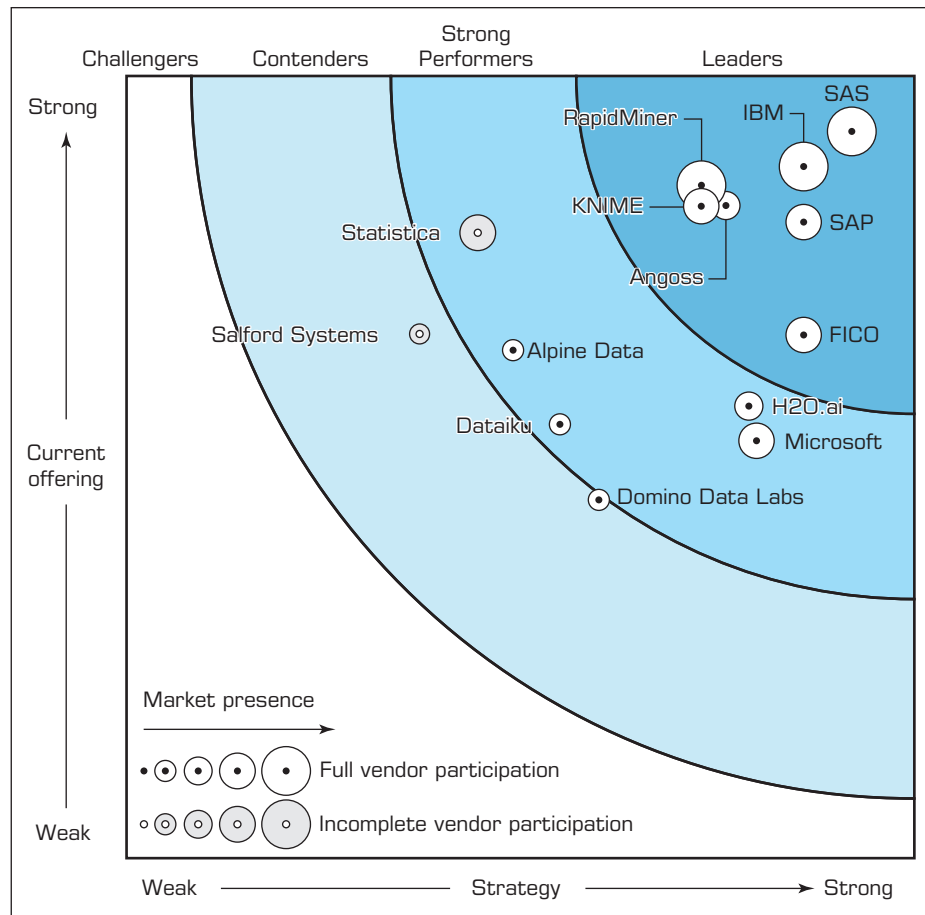


FIGURE 9.2 Forrester Wave for Predictive Analytics and Machine Learning Solutions.

FORRESTER Forrester is a research and consultancy company in the technology space that offers a variety of services including market research, business consulting, and educational events. As part of its dissemination of market research findings, similar to Garner, Forrester also provides reports that analyses the competitive position of companies and software tools in the business analytics and data science area. A summary of their report, named “**Forrester Wave**” for Predictive Analytics and Machine Learning Solutions, is presented in Figure 9.2. In this figure, the x-axis shows the strength of each company’s strategy, while the y-axis shows the strength of their current offerings. The size and shading of the circles around each data point indicate the strength of each vendor in the marketplace. As was the case in the Gartner Magic Quadrant, Forrester Wave shows SAS, IBM, KNIME, and RapidMiner as the leaders in the marketplace.

The reason for not seeing Python and R among the popular tools in the Gartner and Forrester reports is because these reports only include companies and their product, excluding free and open-source community-type projects and software tools.

KDNUGGETS.COM KDNuggetst.com is well-known portal for analytics and data science. It started a repository of interesting and useful data mining links (for data sets, articles, software tools, and training classes) and period newsletters, nowadays reformed itself into analytics and data science knowledge repository with interesting articles, and blog posts

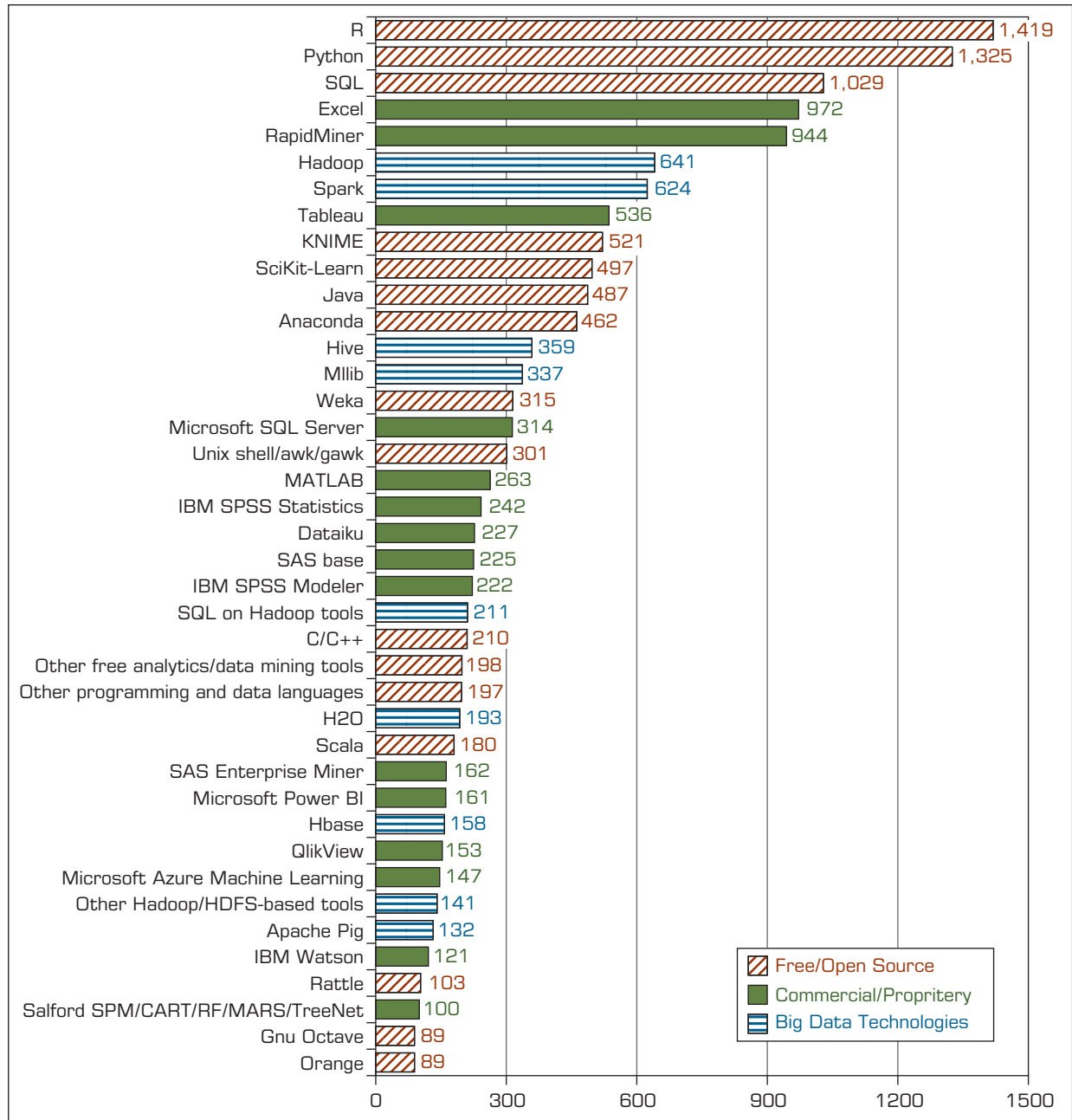


FIGURE 9.3 List of Popular Analytics Tools (as per a KDnugget.com survey).

from industry experts, along with the usual useful links, polls, and newsletters. A list of most popular analytics tools (according to a poll administered by KDnugget.com) is presented earlier in Chapter 4 and is also included here in the form of a horizontal bar chart (Figure 9.3). In this chart, the analytics tools are categorized into (1) free/open source, (2) commercial/proprietary, and (3) big data technologies. Most of the popular tools in this list are further described in this chapter, and short tutorials are provided for most of them to help the reader in evaluating them.

The ranking of analytics tools provided in Figure 9.3 is by no means an objective/optimal outcome; rather it is just a snapshot, obtained from a survey of voluntary data scientist,

at a particular instance in time. If we were to replicate this survey now, using our own sample of potential participants (i.e., analytics professionals), quite likely that we would obtain a slightly different result. That is, popularity of these tools' changes over time, and change based on the cohort you are able to survey. For instance, if we were to do this now, the most popular tool would most likely be Python as opposed to R, although these two tools would still occupy the top two spots. Analytics in Action 9.1 provides an interesting application of analytics to identify patterns of terrorist behaviors (e.g., the use of business analytics and machine learning to discover and track funding of terrorists' activities).

ANALYTICS IN ACTION 9.1 Predictive Analytic and Data Mining Help Stop Terrorist Funding

The terrorist attack on the World Trade Center on September 11, 2001, underlined the importance of open-source intelligence. The USA PATRIOT Act and the creation of the U.S. Department of Homeland Security heralded the potential application of information technology and data mining techniques to detect money laundering and other forms of terrorist financing. Law enforcement agencies had been focusing on money laundering activities via normal transactions through banks and other financial service organizations.

Law enforcement agencies are now focusing on international trade pricing as a terrorism funding tool. International trade has been used by money launderers to move money silently out of a country without attracting government attention. This transfer is achieved by overvaluing imports and undervaluing exports. For example, a domestic importer and foreign exporter could form a partnership and overvalue imports, thereby transferring money from the home country, resulting in crimes related to customs fraud, income tax evasion, and money laundering. The foreign exporter could be a member of a terrorist organization.

Data mining techniques focus on analysis of data on import and export transactions from the U.S. Department of Commerce and commerce-related entities. Import prices that exceed the upper quartile import prices and export prices that are lower than the lower quartile export prices are tracked. The focus is on abnormal transfer prices between corporations that may result in shifting taxable income and taxes out of the United States. An observed price deviation may be related to income tax avoidance/evasion, money laundering, or terrorist financing. The observed price deviation may also be due to an error in the U.S. trade database.

Data mining will result in efficient evaluation of data, which, in turn, will aid in the fight against terrorism. The application of information technology and data mining techniques to financial transactions can contribute to better intelligence information.

QUESTIONS FOR DISCUSSION

1. How can data mining be used to fight terrorism? Comment on what else can be done beyond what is covered in this short application case.
2. Do you think data mining, although essential for fighting terrorist cells, also jeopardizes individuals' rights of privacy?

Sources: Zdanowic, J. S. (2004). "Detecting money laundering and terrorist financing via data mining." *Communications of the ACM*, 47(5), 53; Bolton, R. J. (2002). "Statistical fraud detection: A review." *Statistical Science*, 17(3), 235.

SECTION 9.2 REVIEW QUESTIONS

1. Why is it so critical to choose the right tool for analytics projects?
2. What are the reasons for the existence of such a large and varied analytics tools landscape?
3. What are the four common dimensions we can use to compare analytics tools?
4. Which analytics tool is your most favorite? Why?

9.3 FREE AND OPEN-SOURCE ANALYTICS' PROGRAMMING LANGUAGES

As we can see in Figure 9.3, there seems to be a powerful momentum toward the development and use of free and open-source tools in analytics and data science projects. Some of these tools are programming based while others are visual modeling oriented. In this section, we will cover the two most popular programming-based analytics tools, namely Python and R, providing how to get started with them and explaining their use through simple step-by-step tutorials.

R



R is one of the most popular analytics tools. It is commonly known as a system for statistical computation and stunning graphics. The core of R is an interpreted computer language, which allows branching and looping as well as modular programming using functions. Most of the user-visible functions in R are written in R. However, it is possible for the user to consume functions and procedures written in legacy languages like C, C++, or FORTRAN. The R distribution contains functionality for a large number of statistical and machine learning procedures.

R was initially written by Ross Ihaka and Robert Gentleman at the Department of Statistics of the University of Auckland in Auckland, New Zealand. Most believe that the name “R” stems from the first letter of these two initial creators of the language. R has become a community project where many people contribute and consume new and improved statistics, analytics, and graphical capabilities.

R is managed and distributed by a global network of servers called CRAN (Comprehensive R Archive Network). For more detailed on the history and current structure of R, one should go to R-Project site at <https://www.r-project.org/>.

How to Get Start With R?

There are numerous ways to acquire, install, and use R for business analytics. Here are some of the most common ones, followed by our recommendations.

1. **R-Project.** This is the default option. With this option you get a CLI (command line interface) to write and execute your code. It does not provide a lot of bells and whistles that some of other options listed below may, such as IntelliSense (automatic code completion, instant parameter info, etc.), color coding (of the section of the code for ease of recognition), environment management (for libraries and the development options), among others.
2. **RStudio.** This is perhaps the most popular option to develop R programs today. **RStudio** is an integrated development environment (IDE) for R programming. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging, and workspace management. One can download and install the desktop version of RStudio from <https://www.rstudio>.

com/products/rstudio/download/ for his/her operating system (Windows, MacOS, or Linux) free of charge. RStudio.com also offers commercial desktop, server, and cloud versions of the R development platform.

3. **RStudio Cloud.** RStudio Cloud is a lightweight, cloud-based solution that allows anyone to develop, share, teach, and learn R programming online. With RStudio Cloud, one can analyze data using the RStudio IDE, directly from within one's favorite Web browser. Because it is cloud-based, it allows easy sharing of projects with other team members, students in a class, or participants in workshop, or the community at large anywhere in the world. Although a commercial product, there is a free version of RStudio Cloud that one can sign up and use (<https://rstudio.cloud/>).
4. **Anaconda.** **Anaconda** simplifies the downloading and installing programming tools for R and Python. With its intuitive graphical user interface (e.g., Anaconda Navigator), it centralizes and simplifies the program installation, package management, and deployment. Anaconda installation includes most commonly used data-science application and packages suitable for Windows, Linux, and macOS. Because Anaconda comes with quite a bit of preinstalled tools and packages, most of which a newbie may not need, some prefer to install the smaller slimdown version of Anaconda called **Miniconda**. It is a tradeoff between a slimdown installation with some additional installation of packages afterward and a more complete installation that may have some nonneeded packages. For convenience's sake, if you are pursuing this option, we recommend Anaconda(<https://www.anaconda.com/>) over Miniconda (<https://conda.io/>). Figure 9.4 shows the Anaconda.Navigator with several programming tool installation options. For RStudio, one can execute the RStudio installation module in the navigator.
5. **Visual Studio Code.** **Visual Studio Code** has become one of the most popular IDE for R, Python, C/C#, and several other programming languages. It is a free tool that can run on Windows, MacOS, or Linux. The R extension for Visual Studio Code supports extended syntax highlighting, code completion, linting, formatting, interacting with R terminals, viewing data, plots, workspace variables, help pages, managing packages and working with R Markdown documents.
6. **Jupyter Notebook.** Although **Jupyter notebook** is well-known browser-based programming environment for Python, it can be used for other languages by installing and configuring its environment the proper language kernels. The R kernel for the Jupyter Notebook environment is called IRkernel and can be downloaded from <https://cran.r-project.org/web/packages/IRkernel/>.
7. **Google Colab.** One can use Google Colab cloud-based development platforms for R programming. Because Colab is originally designed for Python, it would need some setup procedures. Colaboratory, or "Colab" in short, is a product provide by Google Research. Colab allows anybody to write and execute arbitrary Python (or R) code through the browser, and is especially well suited for machine learning and data analysis and education (<https://research.google.com/colaboratory/>). While Jupyter is the open-source project on which Colab is based, Colab allows use and share of Jupyter notebooks with others without having to download, install, or run anything. Colab notebooks are stored in Google Drive, or can be loaded from GitHub. Colab notebooks can be shared just as you would with Google Docs or Sheets. Technically, in Colab, one can run Python and R within the same notebook by switching back and forth with pry2 package.

These are some of the most commonly practiced development environments for R programming. From these six options, without any local or cloud-based constraints, the two that stands out are RStudio (which has been the most popular option in recent history) and Visual Studio Code (which is the fastest growing option among analytics educators and professions).

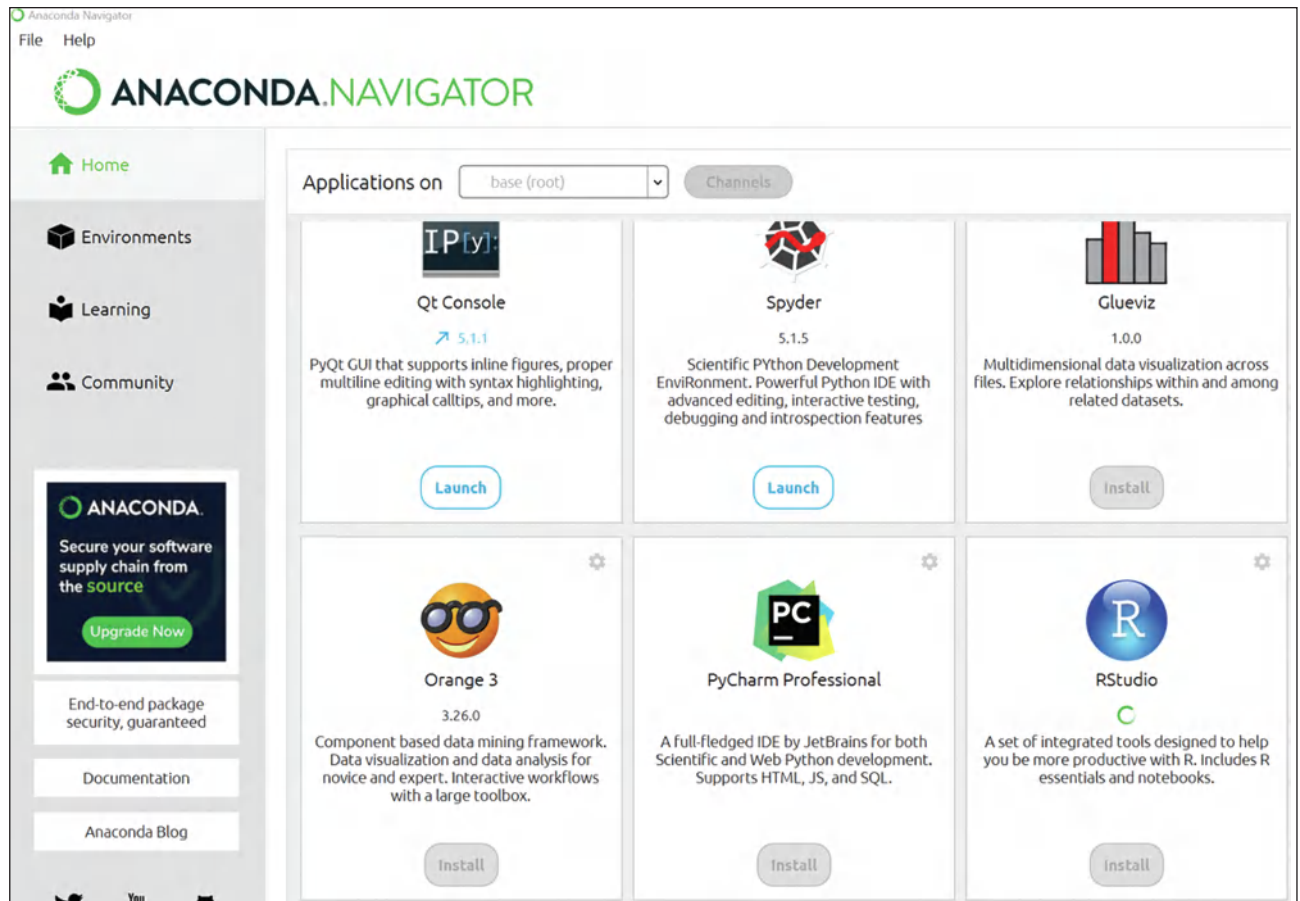


FIGURE 9.4 Anaconda.Navigator User Interface.

TUTORIAL

R for Analytics Application Tutorial—Predicting Employee Attrition

This simple application is built to demonstrate (1) descriptive analytics (i.e., data visualization) capabilities of R language, (2) show how to develop and test predictive models in R, and (3) illustrate the Explainable AI function in R for the employee churn prediction problem.

The data set used in this tutorial is donated to Kaggle by IBM, and is publicly available at (<https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset>). The data set can be used to uncover the factors that lead to employee attrition.

Let us get started with the R programming of this interesting problem using RStudio, one of the most popular integrated development environments (IDE) for R programming. When you are replicating the tutorial code, make sure to be precise in its syntax because R programming is case sensitive (i.e., *X* and *x* are not the same variables and *randomForest* and *RandomForest* are not the same libraries).

First, we need to install the required libraries. Here is a handy reusable R code that will install needed libraries (in this case, “ggplot2,” “randomForest,” “caret,” “lime”) if they are not already installed, and their dependencies (if any).

```

libs = c("ggplot2", "randomForest", "caret", "lime")
for (i in libs)
{
  if(!is.element(i, .packages(all.available = TRUE)))
  {
    install.packages(i, dependencies = T)
  }
  library(i, character.only = TRUE)
}
lapply(libs, require, character.only = TRUE)

```

A brief definition of the installed libraries: *ggplot2* is an R library and is widely regarded as the most advanced visualization library; *randomForest* is the library for the Random Forest machine learning algorithm; *caret* is the library used for machine learning in R; and *lime* is a library used for explainable AI.

Let us set the workspace to the current directory, clear the workspace, and read the data:

```

setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
rm(list=ls())
HR = read.csv(file=" ../data//IBM-HR-Employee-Attrition.csv",
header=T)

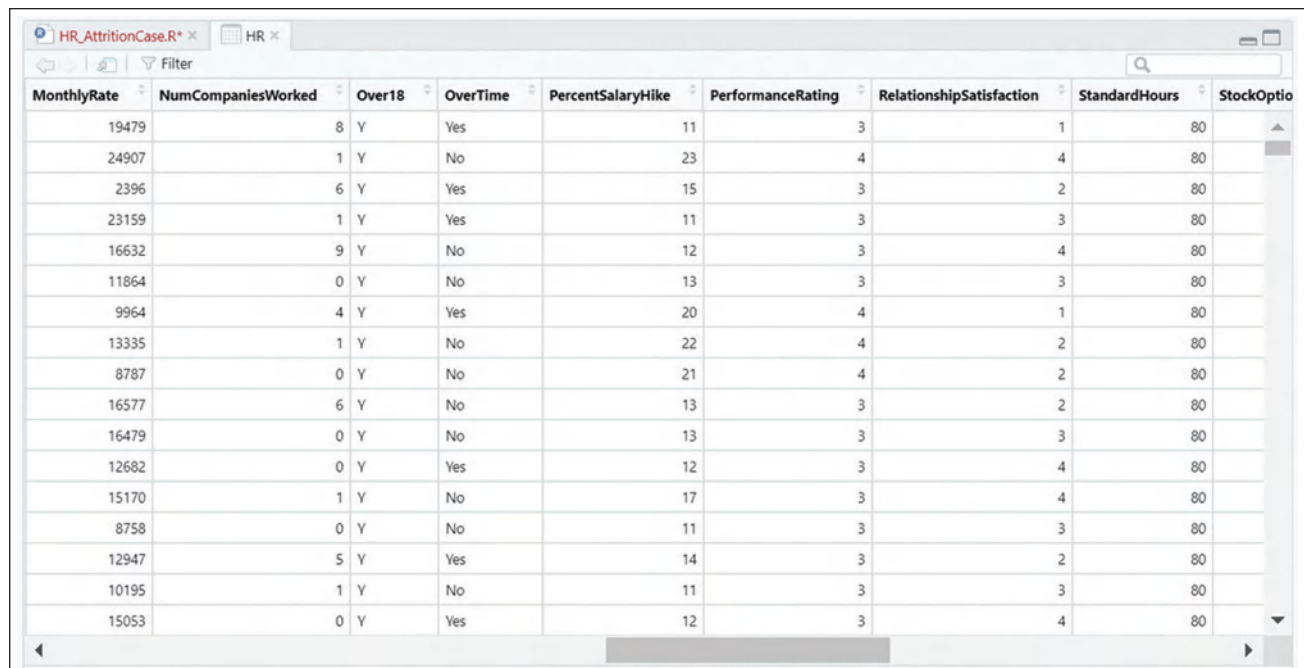
```

Let us visually inspect the first few rows of the data. For this, we can use `head(HR)` and see the first six rows of the data (which is the default setting in R), or we can specify how many of the rows we would like to see. The following code will show the top 10 rows.

```
head(HR, 10)
```

We can also view the entire data set in a Table format using the following function and produce the output shown in Figure 9.5:

```
View(HR)
```



MonthlyRate	NumCompaniesWorked	Over18	OverTime	PercentSalaryHike	PerformanceRating	RelationshipSatisfaction	StandardHours	StockOptions
19479	8	Y	Yes	11	3	1	80	
24907	1	Y	No	23	4	4	80	
2396	6	Y	Yes	15	3	2	80	
23159	1	Y	Yes	11	3	3	80	
16632	9	Y	No	12	3	4	80	
11864	0	Y	No	13	3	3	80	
9964	4	Y	Yes	20	4	1	80	
13335	1	Y	No	22	4	2	80	
8787	0	Y	No	21	4	2	80	
16577	6	Y	No	13	3	2	80	
16479	0	Y	No	13	3	3	80	
12682	0	Y	Yes	12	3	4	80	
15170	1	Y	No	17	3	4	80	
8758	0	Y	No	11	3	3	80	
12947	5	Y	Yes	14	3	2	80	
10195	1	Y	No	11	3	3	80	
15053	0	Y	Yes	12	3	4	80	

FIGURE 9.5 Output of the View(HR) Function.

As we inspect the first 10 rows of the data table visually, we notice that some of the columns (i.e., “Over18,” “EmployeeCount,” and “StandardHours”) have no variance (repeating the same value for all 10 rows). Before deleting these columns from the data table, we want to confirm our observation with simple charts for these attributes using the following functions:

```
plot(as.factor(HR$Over18))
plot(as.factor(HR$StandardHours))
plot(as.factor(HR$EmployeeCount))
```

Now that it is confirmed our observation, we use the following lines to delete these three attributes:

```
HR$Over18=NULL
HR$StandardHours = NULL
HR$EmployeeCount = NULL
```

Let us now check the attributes and their data types: This is an excellent function in R to see what are the attributes in the given data table and what are their data types and first few values. Figure 9.6 shows the output of this function.

```
str(HR)
```

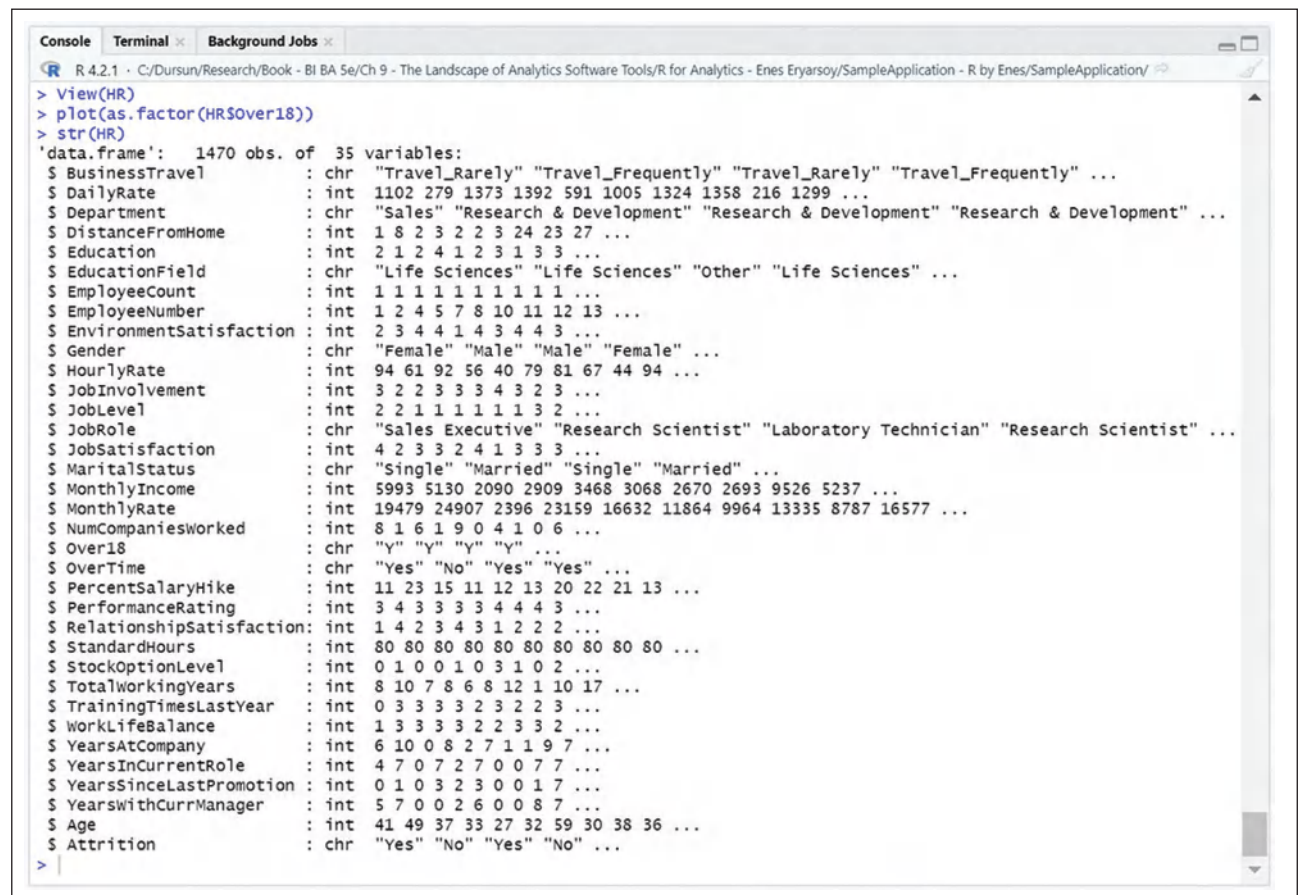


FIGURE 9.6 List of All Attributes, Their Data Types and First Few Values.

Notice that there are some attributes that are stored as “chr” meaning they are of character type. Since these are in fact categorical variables, we should convert these “chr” type variables into “factor” type variables in R, so that we can properly use them in our data analysis using the following code. Once we do that, we can use `str(HR)` function to confirm the output.

```
HR[sapply(HR, is.character)] = lapply(HR[sapply(HR, is.character)],
  as.factor)
str(HR)
```

Now we will see that all of the attributes seem to be of R processable types. Before moving forward, let us also check for any missing values. The easiest way to do this is by using the “`summary()`” function in R. The function summarizes each column and indicates number of missing values for each column, if there is any.

```
summary(HR)
```

It looks like, in this case, we have no attributes that contain missing values in the data set.

Data Visualization

It is time to do some visualization. Let’s first look at the relationship between age, and income.

```
r=ggplot(data=HR, aes(x=Age, y=MonthlyIncome))
```

Add a geometry to the plot.

```
r + geom_point()
```

It seems like there is a slight relationship. Let’s check the correlation.

```
cor(HR$Age, HR$MonthlyRate)
```

Let us beautify the graph with some colors.

```
r + geom_point(colour="Blue")
```

But could we color the graph according to gender. The answer is yes, with the following function.

```
r + geom_point(aes(colour=Gender))
```

Increase the size for better visuals

```
r + geom_point(aes(colour=Gender), size=2)
```

What about setting bubble sizes according to years in current role?

```
r + geom_point(aes(size=YearsInCurrentRole, colour=Gender))
```

The visualization that we created so far is shown in Figure 9.7.

Now, let’s plot some a Histogram and a Density maps. For the histogram, let us look at total salaries paid by each department.

```
s = ggplot(data=HR, aes(x=MonthlyIncome))
s + geom_histogram(binwidth=1000)
```

Add color:

```
s + geom_histogram(binwidth=1000, colour="Red", fill="Orange")
```

Add different color for each department:

```
s + geom_histogram(binwidth=1000, aes(fill=Department),
  colour="Black")
```

The histogram we just created is shown in Figure 9.8.

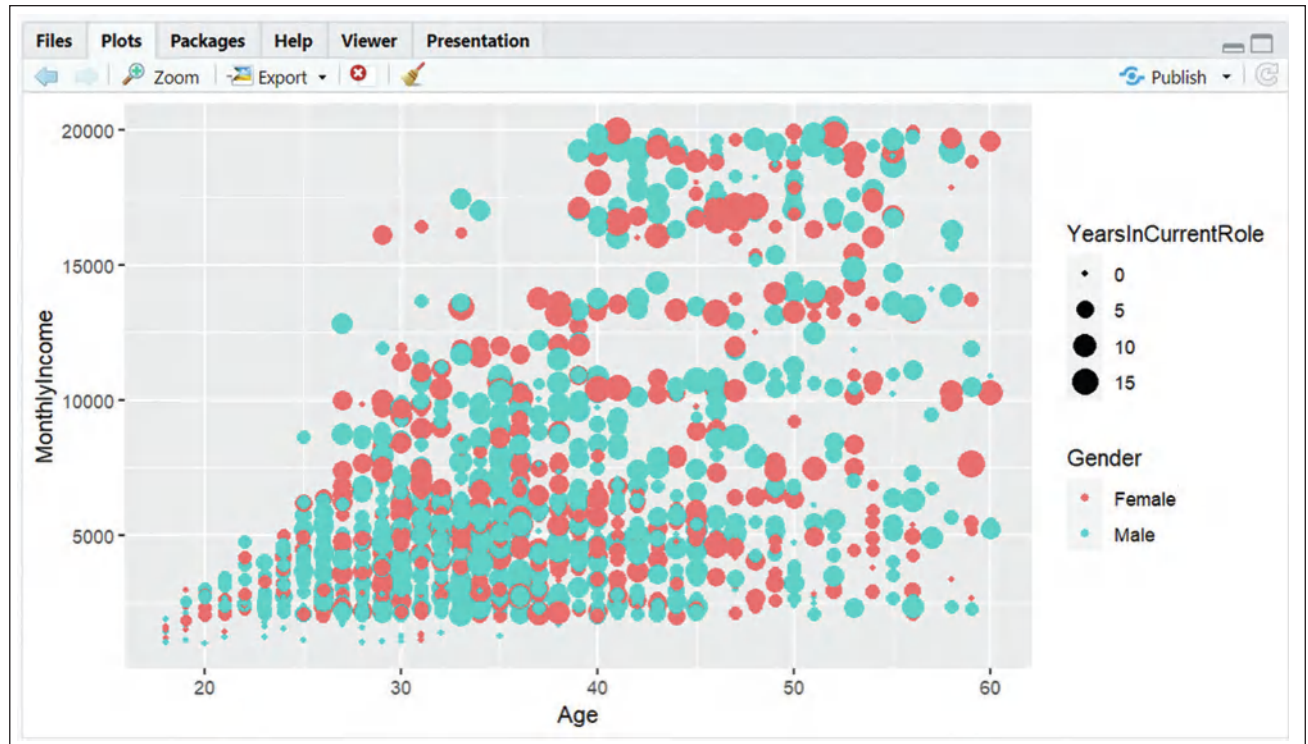


FIGURE 9.7 The Visualization of the Relationship between Age and MonthlyIncome in a Bubble Chart.

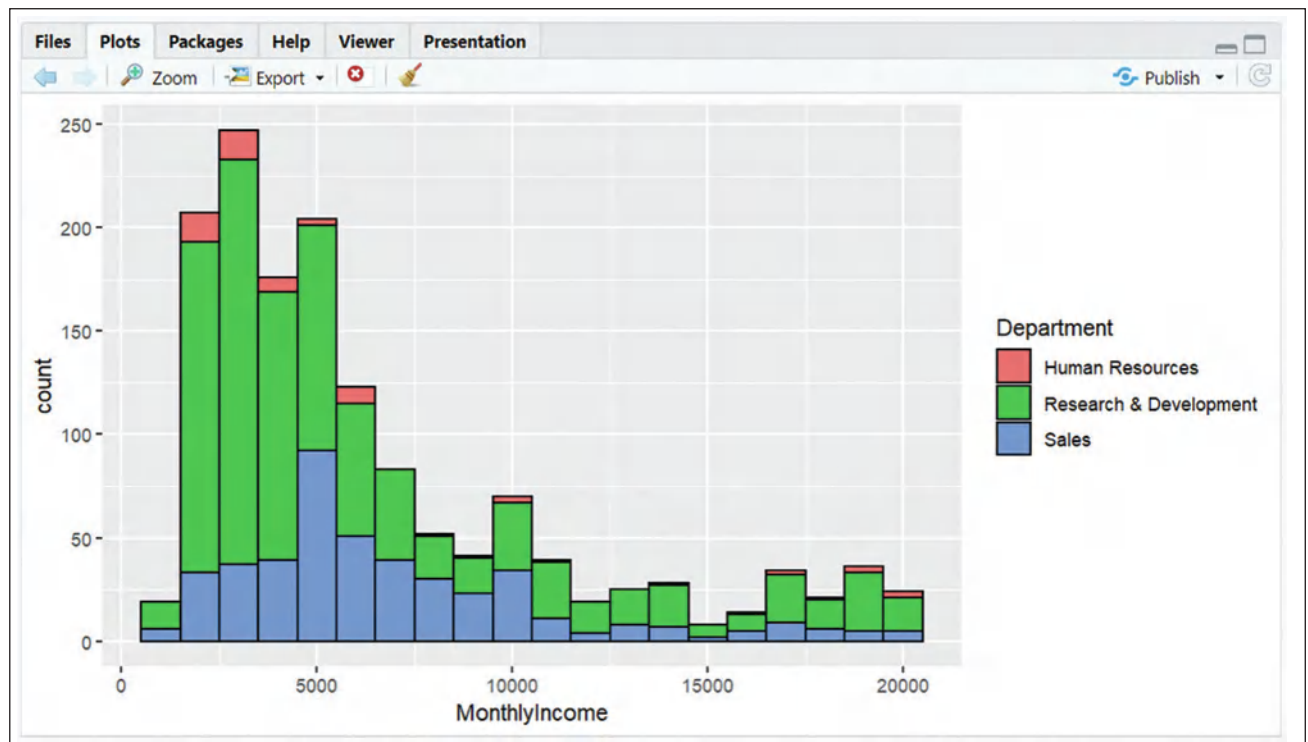


FIGURE 9.8 A Color Enhanced Histogram.

Is there a linkage between employee churn and salary? First, let us try a Histogram to answer this question:

```
plot = ggplot(data=HR, aes(x=MonthlyIncome))
plot + geom_histogram(binwidth=1000, aes(fill=Attrition),
  colour="Black")
```

Let us now try to visualize this relationship with a Density plot. The output of the following code is shown in Figure 9.9.

```
s + geom_density()
s + geom_density(aes(fill=Attrition))
s + geom_density(aes(fill=Attrition), position="Stack")
```

These are just a few simple examples of the unlimited visualization one can create with R and **ggplot**.

Machine Learning for Predictive Modeling

In this section, we will write R code to predict the target variable (“attrition” of employees being Yes or No). Let us see how the values in Attrition attribute are distributed.

```
table(HR$Attrition)
```

With this function, we get No = 1233 and Yes = 237. It looks like we have an imbalanced data set at hand; that is, we have significantly more records with No values (nonchurners) than Yes values (churners). Well, we will go with what we have for this

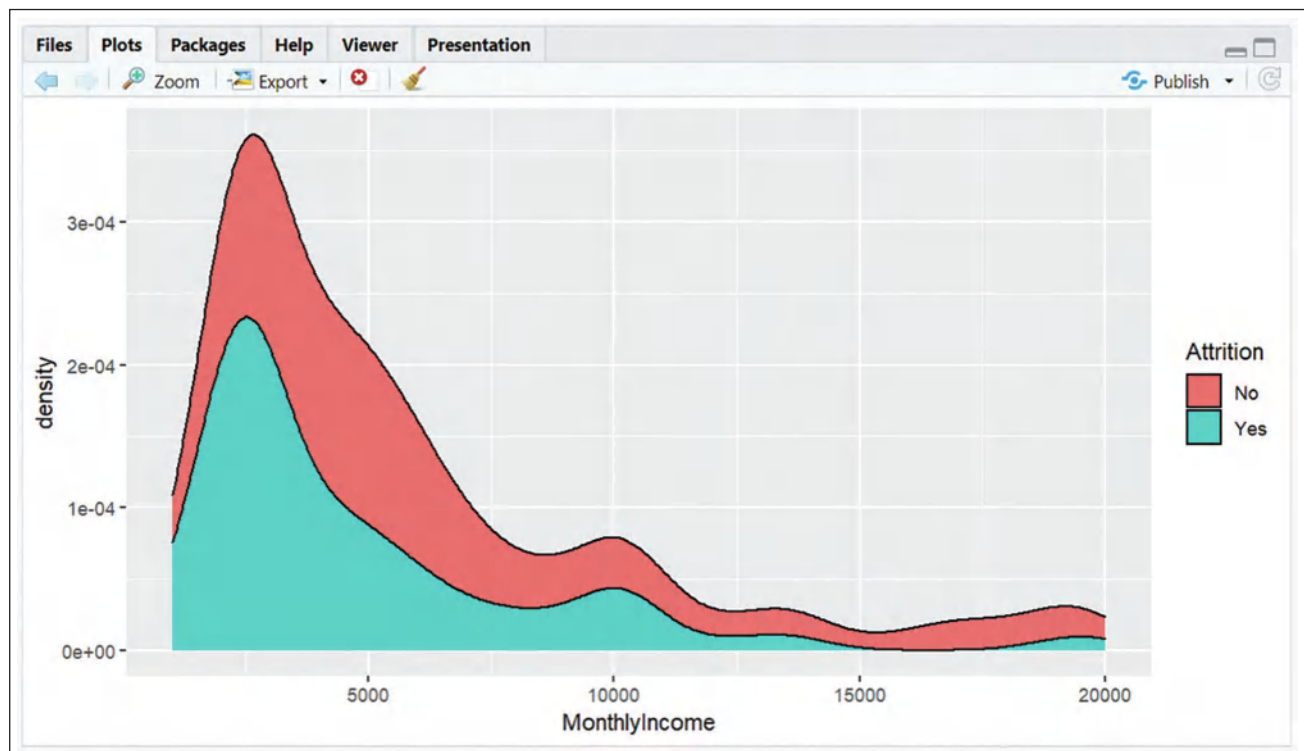


FIGURE 9.9 Illustration of the Density Plot.

exercise. In reality, we look for ways to treat the imbalance nature of the data before developing accurate and informative predictive models.

As part of the simple experimental design (using a simple split procedure), we are going to set one quarter of our data aside for testing.

```
ncol(HR)
set.seed(1453) #using a set random number seed ensures
  repeatability of our results
sampledf = sample(nrow(HR), nrow(HR)*.75)
train = HR[sampledf,]
test = HR[-sampledf,]
dim(train)
dim(test)
```

Next, we initialize the randomForest with a set random seed for reproducibility of our results, and build/induce the model using the train data set.

```
set.seed(1453)
rf = randomForest(Attrition ~., data=train)
```

The character “~.” means all attributes/columns (except the target column) should be used as input variables to build/train our model. Now that we train a model called rf, let’s predict the accuracy of this model using the “test” data set.

```
pred = predict(rf, newdata=test)
```

Let us get the confusion matrix for the test data set results, and compute the accuracy metrics. The results are shown in Figure 9.10.

```
confusionMatrix(test$Attrition, pred)
confMatSplit = table(test$Attrition, pred)
accSplit = sum(diag(confMatSplit))/sum(confMatSplit)
```

Let’s develop a random forest model using Cross Validation. The result of the rfModel development process is shown in Figure 9.11.

```
rfModel = train(Attrition ~., # Attrition is a function of
  all other variables
data = train, # Use the train data as the training data
method = "rf", # Use the 'random forest' algorithm
trControl = trainControl(method = "cv", # Use cross-validation
number = 10, # Use 10 folds for cross-validation
verboseIter = TRUE)) # show progress
rfModel
```

Let us calculate the confusion matrix, and accuracy for the cross-validation process. The result of this following code is shown in Figure 9.12.

```
confMatCV = table(test$Attrition, predictions)
confusionMatrix(test$Attrition, predictions)
accCV = sum(diag(confMatCV))/sum(confMatCV)
accCV
paste0("Accuracy using train/test split is:", round(accSplit, 2))
paste0("Accuracy using CV is:", round(accCV, 2))
```

It looks like cross-validation slightly outperformed 25/75 single-split procedure (0.8614 versus 0.8560).

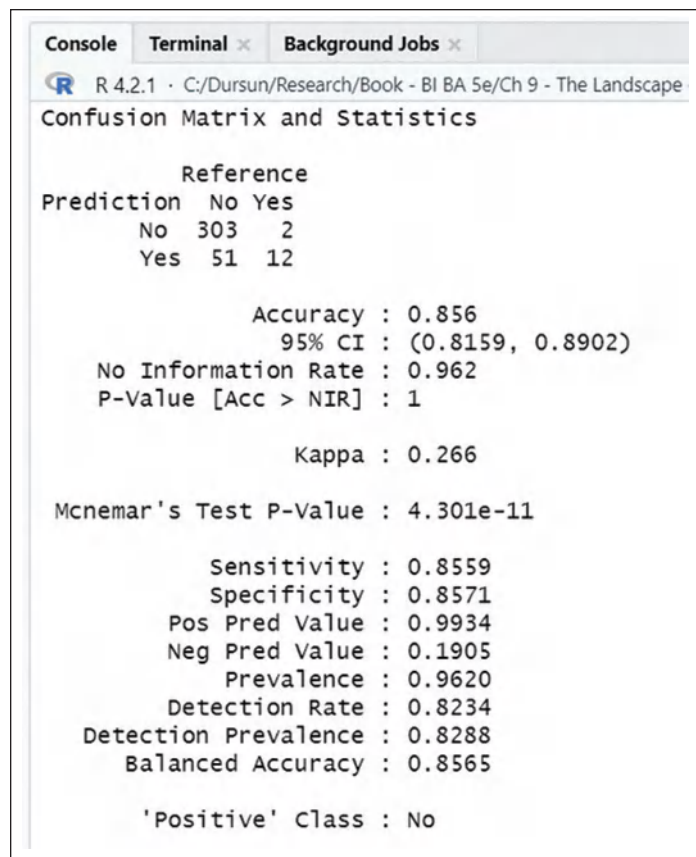


FIGURE 9.10 The Confusion Matrix and the Accuracy Metrics of the Developed rfModel on Test Data.

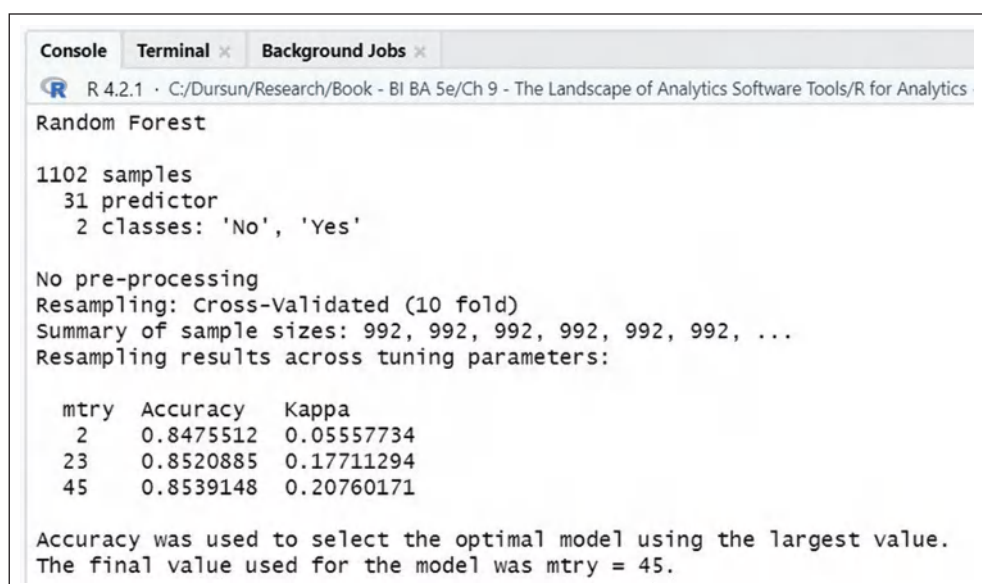


FIGURE 9.11 The Summary Information about the rfModel Development Process.

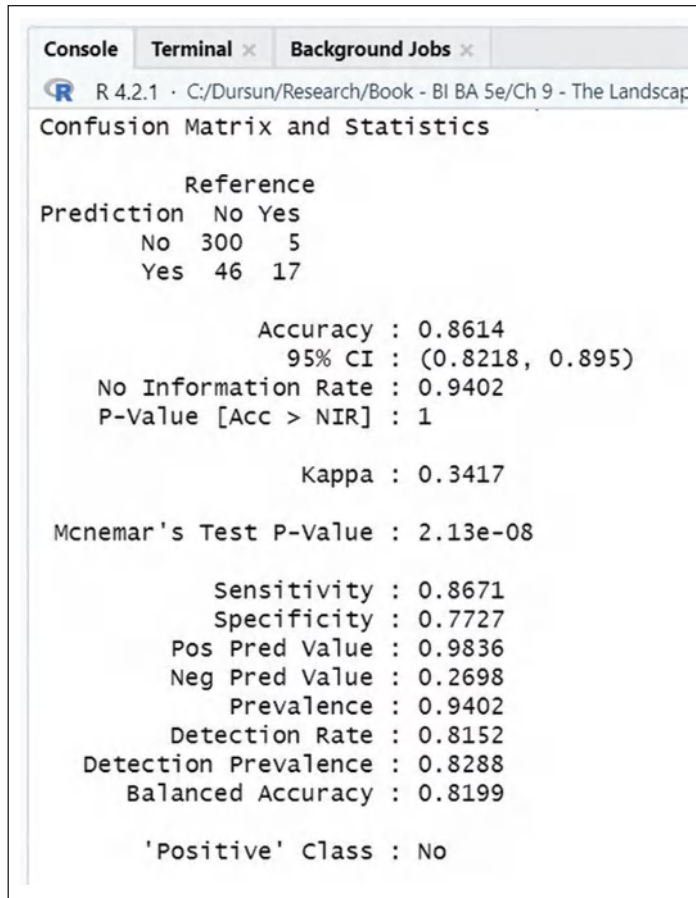


FIGURE 9.12 The Confusion Matrix and the Accuracy Metrics of the Cross-Validation Process.

Explainable AI

To find out which variables played more discerning role in the employee churn prediction problem, let's perform variable importance on the rf model.

```

varImps = varImp(rf)
varImps
rownames(varImps)[which(varImps == max(varImps))]

```

As we can see in Figure 9.13, the MonthlyIncome seems to be the most influential factor followed by Age and DailyRate.

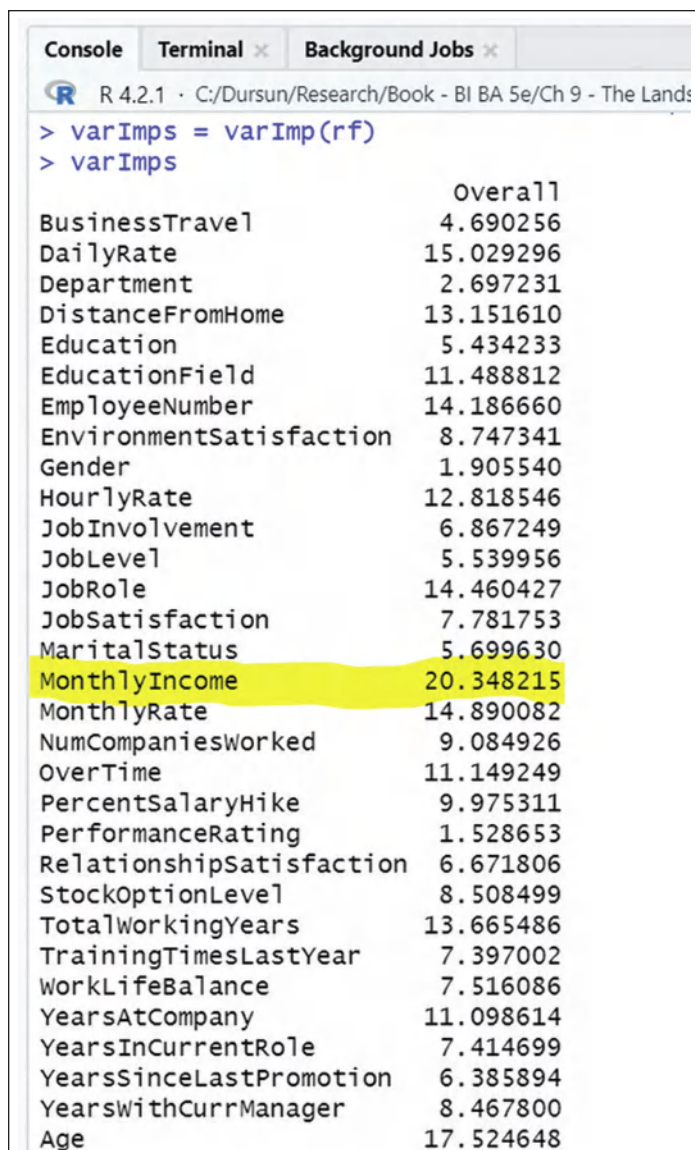
Now, let's look at how these variables contribute to the determination of an individual prediction. Given an employee (picked randomly in the following code), we are to determine the contribution (in terms of magnitude and direction) of each variable to the predicted outcome.

```

set.seed(1453) # set the random number generation seed for
replicability
selEmp = sample(1:nrow(HR),1) # randomly select an employee
selEmpData = HR[selEmp,]
selEmpData$Attrition #determine the employees churn value
(Atrition = Yes)

```

Why? Let us prepare the explainer. The graphical outcome of the following LIME (local interpretable model-agnostic explanations) execution code is shown in Figure 9.14.



	Overall
BusinessTravel	4.690256
DailyRate	15.029296
Department	2.697231
DistanceFromHome	13.151610
Education	5.434233
EducationField	11.488812
EmployeeNumber	14.186660
EnvironmentSatisfaction	8.747341
Gender	1.905540
HourlyRate	12.818546
JobInvolvement	6.867249
JobLevel	5.539956
JobRole	14.460427
JobSatisfaction	7.781753
MaritalStatus	5.699630
MonthlyIncome	20.348215
MonthlyRate	14.890082
NumCompaniesWorked	9.084926
OverTime	11.149249
PercentSalaryHike	9.975311
PerformanceRating	1.528653
RelationshipSatisfaction	6.671806
StockOptionLevel	8.508499
TotalWorkingYears	13.665486
TrainingTimesLastYear	7.397002
WorkLifeBalance	7.516086
YearsAtCompany	11.098614
YearsInCurrentRole	7.414699
YearsSinceLastPromotion	6.385894
YearsWithCurrManager	8.467800
Age	17.524648

FIGURE 9.13 The Output of the Variable Importance Calculation Procedure.

```
explainer = lime(HR, rfModel)
limeExp = explain(selEmpData, explainer, n_labels = 2,
  n_features = 10)
plot_features(limeExp)
```

For the case 569 (who is a churner), the results indicate that OverTime contributed the most (larger the value more likely to produce a Yes/Churn value).

A screenshot of the RStudio desktop tool used for this tutorial is shown in Figure 9.15.

Rattle (for visual analytics programming in R)

Rattle is a popular GUI for data mining using R. It presents statistical description and visual summaries of data, transforms data so that it can be readily for modeling, builds both unsupervised and supervised machine learning models from the data, presents the performance of models graphically, and scores new data sets for deployment into production. An exciting feature of Rattle is that all of your interactions through the graphical user interface are captured as an R script, which can be copied and executed in R independent of the Rattle interface.

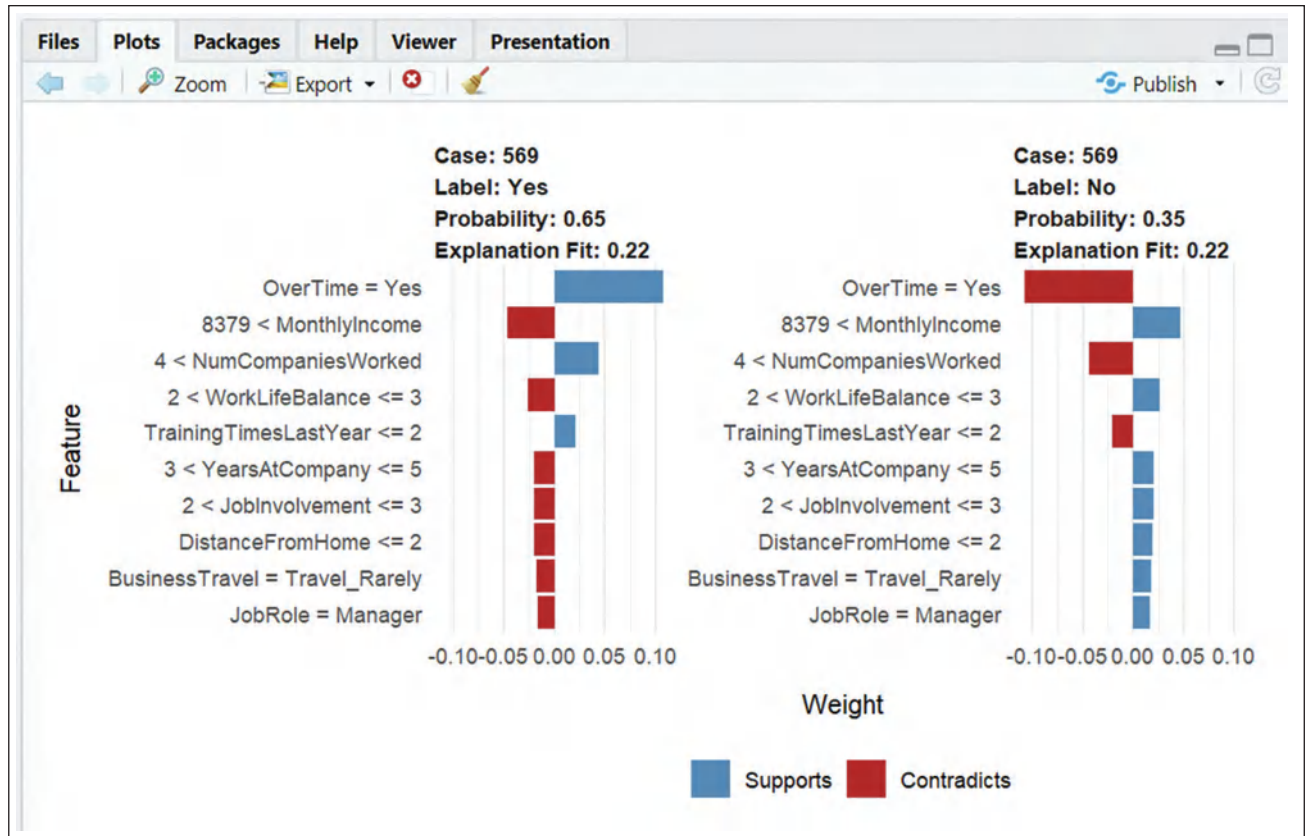


FIGURE 9.14 The Graphical Output of the LIME Execution Code.

Rattle uses **RGtk2** as the GUI Toolkit. It can be installed from a Togaware as shown below:

- On Linux:

```
> install.packages("https://access.togaware.com/RGtk2_2.20.36.2.tar.gz", repos=NULL)
> install.packages("https://access.togaware.com/cairoDevice_2.28.tar.gz", repos=NULL)
```

- On Windows:

```
> install.packages("https://access.togaware.com/RGtk2_2.20.36.2.zip", repos=NULL)
> install.packages("https://access.togaware.com/cairoDevice_2.28.zip", repos=NULL)
```

- On Mac:

```
> install.packages("https://access.togaware.com/RGtk2_2.20.36.2.tgz", repos=NULL)
> install.packages("https://access.togaware.com/cairoDevice_2.28.tgz", repos=NULL)
```

Once installed, executing the following two R commands should start Rattle interface. More on Rattle can be found at <https://r4stats.com/articles/software-reviews/rattle/> (a screen capture of the Rattle is also shown in Figure 9.16).

```
> library(rattle)
> rattle()
```

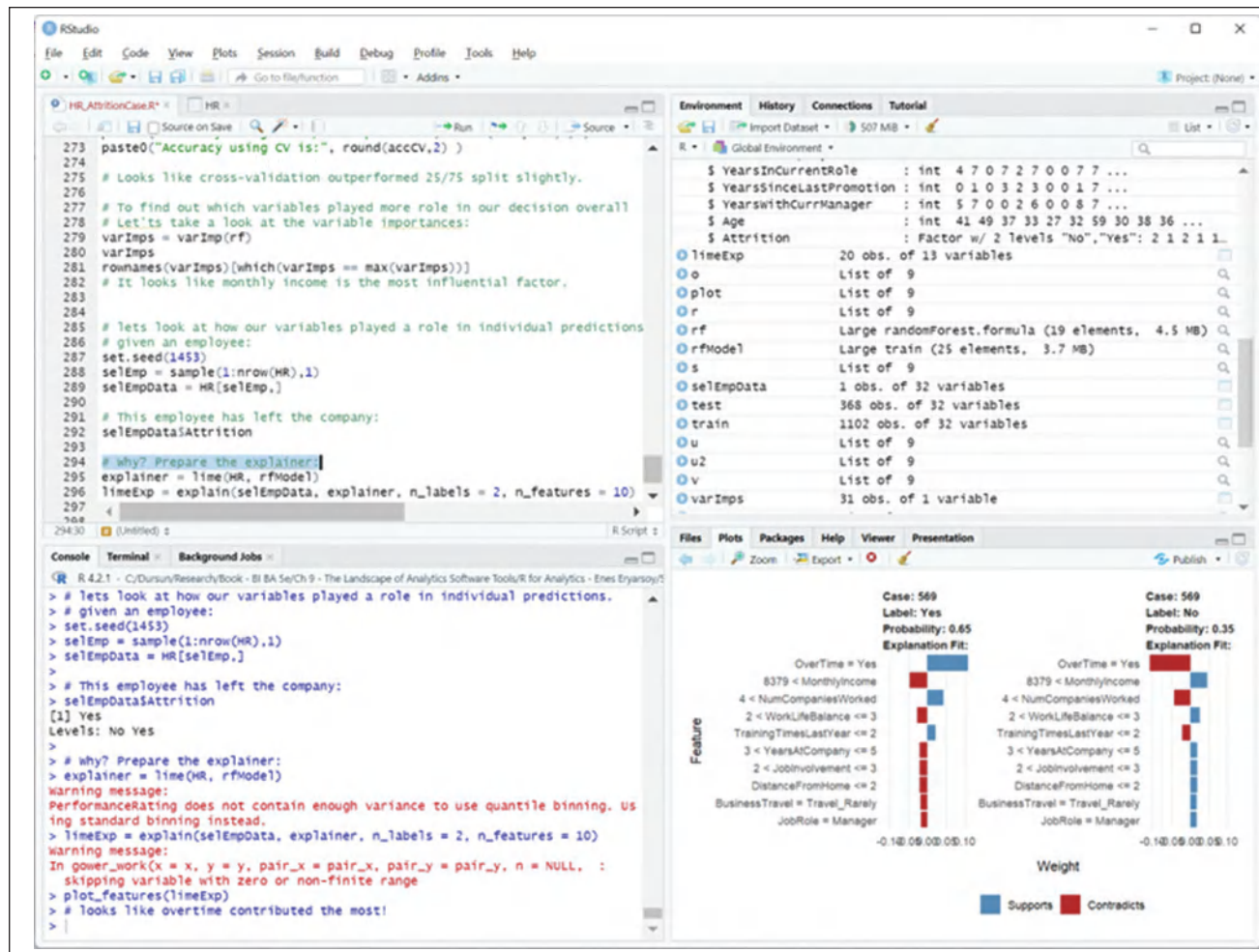


FIGURE 9.15 A Screenshot of the RStudio Desktop Tool Used for This Tutorial.

Python



Python is one of the most popular computer programming languages used for a multitude of computational tasks that include (1) building websites and ecommerce applications, (2) developing specialized software for a variety of business applications, (3) automate computational tasks, (4) automate data acquisition and storage, and (5) conduct data analytics (e.g., business intelligence, business analytics, data science, data mining, machine learning, deep learning, among others). As can be seen, unlike R, Python is not just for analytic, but is a general-purpose programming language that can be used to create a wide variety of different computer applications.

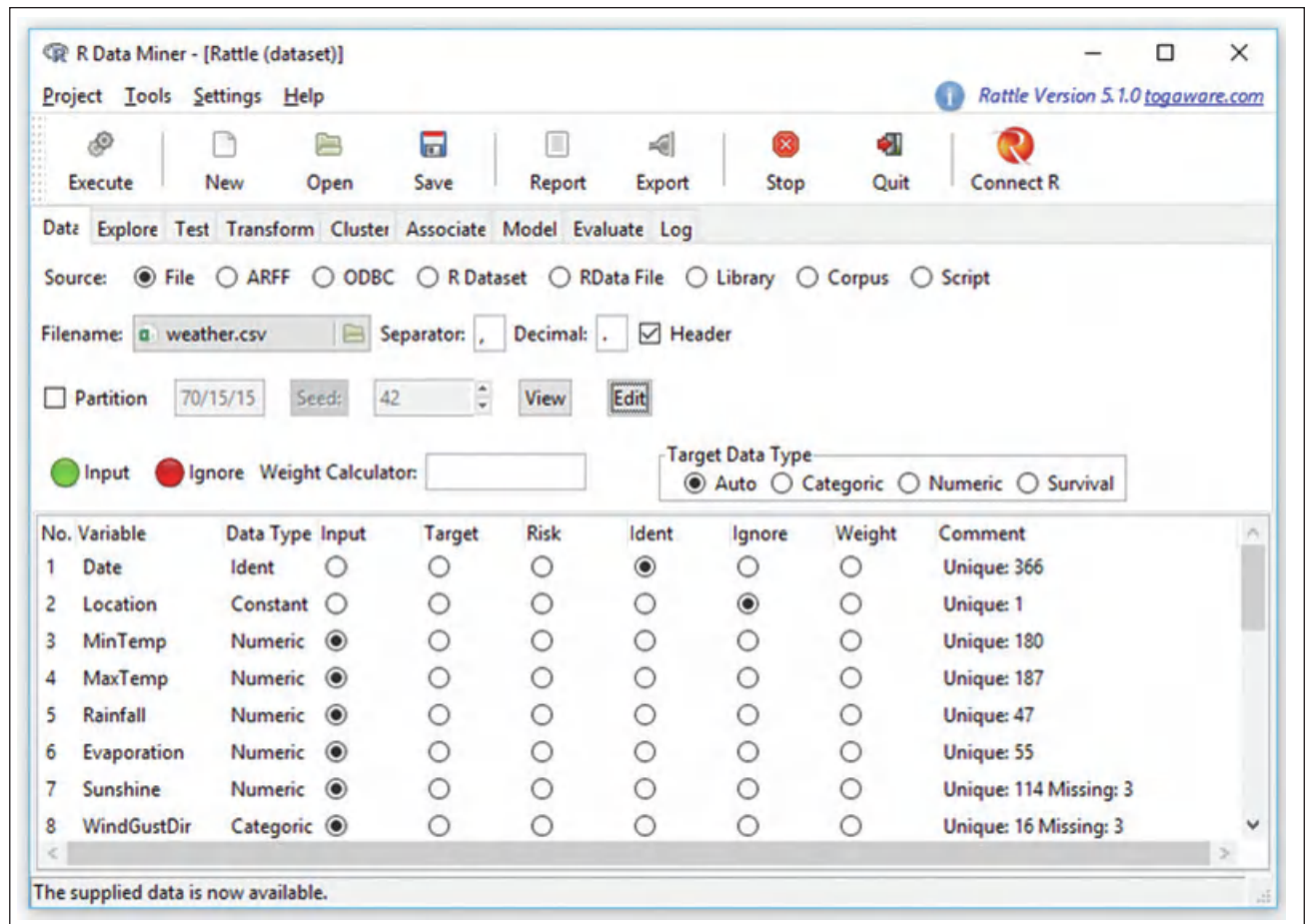


FIGURE 9.16 A Screen Capture of the Rattle Graphical User Interface.

Technically speaking, Python is a dynamic, interpreted (bytecode-compiled) computer programming language. There are no type declarations of variables, parameters, functions, or methods in source code. This makes the code brief and flexible, and efficient (i.e., saves time for developer). An excellent way to see how Python code works is to run the Python interpreter and type code right into it. If you ever have a question like, “What happens if I add an int to a list?” Just typing it into the Python interpreter is a fast and likely the best way to see what happens. (See below to see what really happens!)

Because of its versatility and ease of use, many developers around the world have embraced Python, not only as a developer and consumer of the existing libraries and other sources, but also contributor to the Python repository. Resultantly, Python has become a crowdsourced programming tool for masses, from academics to industry practitioners. Giant internet companies have been investing greatly to the development of Python resources and making them available to the broader community free of charge.

Python is freely distributed by the Python Software Foundation (<https://www.python.org/>), and the global Python community welcomes and encourages anyone for participation as a consumer of the existing resources or creators/contributor of the new ones. The Python community is based on mutual respect, tolerance, encouragement, and creativity, and they organize workshops and conferences to help each other live up to these principles.

Because of its popularity, largely fueled by its vast number of reusable functions, algorithms, and libraries created by highly reputed scholars and industry leaders, in recent years, Python has become the indisputably number one tool for business analytics and data science (see Figure 1 [popular data science tools]). Data scientist around the world are using Python to attack problems that were previously deemed unsolvable by creatively weaving data and functions toward reusable and repeatable solution architectures.

How to get started with Python?

There are numerous ways to acquire, install, and use Python for business analytics and data science. Here are some of the most common ones, followed by our recommendations.

1. **Python Software Foundation.** This is the default option. With this option you can get the executable version of Python for Windows, MacOS, Linux, and several other less popular operating systems, as well as the Python source code for the latest or any legacy versions (<https://www.python.org/downloads/>). The installation from this source comes with a default development environment called Integrated Development and Learning Environment (IDLE).
2. **PyCharm, Spyder, PyDev, Atom, etc.** These are some of the most popular IDEs to code Python programs nowadays. Most of these are commercial tools, but almost all of them also provide free, often restricted or scaled-down, community versions.
3. **Anaconda.** Anaconda simplifies the downloading and installing programming tools for Python and R. Anaconda installation also provides other tools for data visualization and analytics. With its intuitive graphical user interface (e.g., Anaconda Navigator), it centralizes and simplifies the program installation, package management, and deployment. Anaconda installation includes most commonly used data-science application and packages/libraries suitable for Windows, Linux, and macOS. If you are new to Python programming, we recommend Anaconda (<https://www.anaconda.com/>) installation, although it may include more than what you will be using. Figure 9.4 shows the Anaconda.Navigator with several programming tool installation options, that includes **PyCharm, Spyder**, and Jupyter Notebook for Python programming.
4. **Visual Studio Code.** Visual Studio Code has become one of the most popular IDE for Python, and some of the other programming languages like R, C/C#, Java, PHP, Julia. It is a free tool that can run on Windows, MacOS, or Linux. The Python extension for Visual Studio Code supports extended syntax highlighting, code completion, linting, and formatting.
5. **Jupyter Notebook.** Jupyter notebook is well-known browser-based programming environment for Python. Because it runs within a browser, Jupyter Notebook is widely known to be the most user-friendly Python programming environment. It can be individually installed (via <https://jupyter.org/>), or it can be installed via Anaconda.
6. **Google Colab.** Colab is originally designed to make Python programming easy for teams. Colaboratory, or “Colab” in short, is a product provide by Google Research. Colab allows anybody to write and execute arbitrary Python code through the browser, and is especially well suited for machine learning, data analysis, and business analytics education (<https://research.google.com/colaboratory/>). While Jupyter is the open-source project on which Colab is based, Colab allows use and share of Jupyter notebooks with others without having to download, install, or run anything. Colab notebooks are stored in Google Drive, or can be loaded from GitHub. Colab notebooks can be shared just as you would with Google Docs or Sheets. Technically, in Colab, one can run Python and R within the same notebook by switching back and forth with `pry2` package.

These are some of the most commonly practiced development environments for Python programming. From these six options, without any local

or cloud-based constraints, the two that stand out are Jupyter Notebook (which has been the most popular option in recent history) and Visual Studio Code (which is the fastest growing option among analytics educators and professions).

TUTORIAL

Python for Analytics, Application Tutorial—Predicting Movie Financial Success

This simple application is built to demonstrate how to (1) read/inject a data set, (2) inspect prepare the data for predictive analytics, and (3) develop and test predictive models.

Two versions of the IMDB box office data set will be used for this tutorial. Each record in the data represents features of a movie (title, year of production, budget, etc.) as well as a binary feature that shows whether the movie box office exceeded \$100M. The “IMDB-classification.csv” is a balanced data set (50% of movies sold less/more than \$100M), while the “IMDB-imbalanced.csv” is an imbalanced version of the same data with only 22% of the movies sold more than \$100M.

The publicly available full version of the data set can be accessed at <https://www.kaggle.com/code/saurav9786/imdb-score-prediction-for-movies/data>. In this tutorial, we used a subset of movies, and derived/calculated the binary target variable from the “gross” column.

IMPORTING DATA INTO PYTHON A Comma Separated Values (CSV) data file can be imported into Python in several ways. The two libraries that are most commonly used are NumPy and Pandas. In this tutorial, we use the NumPy library for data ingestion, manipulation, and preprocessing. That is, the data table will be imported as a NumPy array of size $m \times n$ where m is the number of records and n is the number of features. This procedure requires us to first import the NumPy package into the Python kernel using its common alias (i.e., `np`) as follows:

```
>>> import numpy as np
```

In order to import a csv file, we may use the `genfromtxt` function of the `numpy` library.

```
>>> datac= np.genfromtxt('fname', delimiter=',', dtype= None,
    encoding=None, names=True)
```

`fname` must be replaced with the complete csv file name (case sensitive). Make sure to put the csv file in the main python directory before running the code. Using this method for importing a data file, each column is identified by its header title and each row is identified by its index number:

```
>>> datac['budget']          # all the values(rows) in column
    titled 'budget'
>>> datac[3]                # all the values(columns) in the
    row indexed 3 (4th row)
>>> datac['budget'][0:5]    # only the first 5 rows from the
    'budget' column
```

Classification using a balanced data set. Here we will be training an artificial neural network classification model to predict movies box office (being less/more than \$100M) given their characteristics. For demonstration, we are going to use only “budget” and “duration” of the movies as predictors of box office sales (i.e., the ‘100M_sales’ variable). First, we import the ‘IMDB-classification.csv’ data set:

```
>>> data= np.genfromtxt('IMDB-classification.csv', delimiter=
    ',', dtype=None, encoding= None, names=True)
```

Data preparation. After importing, each column is identified as a 1-dimensional NumPy array. In order to use the features in model building, we need to reshape those arrays and convert them to 2D arrays. For simplicity, we rename the reshaped arrays as `x1` and `x2`:

```
>>> x1= np.reshape(data["budget"], (-1,1))
>>> x2= np.reshape(data["duration"], (-1,1))
```

Now that we have extracted and reshaped the predictor variable arrays, we can concatenate them into a single array (we call it `X`) and use that from this point on to refer to the set of all predictor features:

```
>>> X= np.concatenate((x1,x2), axis=1)
```

Again, for simplicity, we specify the target column and rename it to 'y'. No array reshaping is needed.

```
>>> y= data['100M_sales']
```

Data Normalization. Since variables were measured in different scales, it is crucial to normalize their scale of measurement before using them for training a ANN model. We use the min-max normalization approach for this purpose. This approach can be implemented using the *MinMaxScaler* function from the preprocessing submodule of the sklearn library of Python:

```
>>> from sklearn.preprocessing import MinMaxScaler
```

After importing the required function, we may create a scaler object, fit the predictor variables (`X`) array to that to adjust its parameters, and use it to transform the original `X` array values to normalized values (in the range 0-1). We call the transformed array, `X_trans`.

```
>>> scaler= MinMaxScaler()
>>> scaler.fit(X)
>>> X_trans= scaler.transform(X)
```

Data Splitting. Another essential preprocessing step is to split the data set into training and test subsets to ensure that the same records used for training will not be used for evaluation of the classification model. This can be implemented using the `train_test_split` function from the `model_selection` submodule of the sklearn library in python:

```
>>> from sklearn.model_selection import train_test_split
```

After importing the required function, we can apply it to the transformed `X` array as well as the `y` array and randomly split each into two subsets, according to a split proportion that we determine for the test subset:

```
>>> X_train, X_test, y_train, y_test= train_test_split(X_trans,
    y, random_state=0, test_size=0.20)
```

Training a neural network classifier. Now that we have the normalized training and test data sets separated, we can use the training arrays to build and tune any classifier algorithm. For demonstration, we use artificial neural network algorithm in this tutorial, but the procedure is almost the same for other types of classifiers. The `neural_network` submodule of the sklearn library can be used to implement this approach in python. We first import that submodule using an alias (i.e., `ann`) for ease of referring down the road:

```
>>> import sklearn.neural_network as ann
```

Now, from the `ann` submodule we call the `MLPClassifier` function to set up the model parameters and create a ANN classifier object which we call `ann_model0`:

```
>>> ann_model0 = ann.MLPClassifier((4), activation= 'logistic',
    max_iter= 1500)
```

The classifier created can then fit to the training arrays for the network weights to be adjusted:

```
>>> ann_model0.fit(X_train, y_train)
```

Once the model was trained, we can apply it to the `X_test` array (using the `predict` function) to make binary predictions (0 or 1) for the test instances. We store the predictions in a new array called `y_pred` to be compared later to the actual classes (i.e., `y_test`) for model performance evaluation:

```
>>> y_pred = ann_model0.predict(X_test)
```

Model Evaluation. Variety of metrics are used to evaluate the performance of a classification model. In python, all those measures are incorporated in the metrics submodule of the sklearn library. To evaluate a model, we first need to import those functions, then feed them with the arrays containing actual as well as predicted classes (i.e., `y_test` and `y_pred`):

```
>>> from sklearn.metrics import accuracy_score,
    precision_score, recall_score, f1_score
>>> accuracy_score(y_test, y_pred)
0.9
>>> precision_score(y_test, y_pred)
0.96
>>> recall_score(y_test, y_pred)
0.8275862
>>> f1_score(y_test, y_pred)
0.8888889
```

Due to randomness involved and the relatively small data set used for training, it is likely that you obtain somewhat worse performance metrics. To obtain better results you may modify the code and include more predictor features (reshape them and concatenate all to form the `X` array). Also, you may try other network settings by changing the hyperparameters (number of layers, number of neurons, etc.) and evaluate the model performance in each case.

Another approach for evaluating a classifier is using the Receiver Operating Characteristics (ROC) curve. To this end, we first need to obtain the predicted probabilities of belonging to each class for each instance (using the `predict_proba` function). Those probabilities will then be used to calculate the false positive rate (fpr) and true positive rate (tpr), which are needed to plot the ROC curve.

```
>>> y_pred_proba = ann_model0.predict_proba(X_test)
```

The output of this stage would be a `nx2` array where `n` is the number of instances in the test data set. For each instance, two probabilities (adding up to 1) are generated that indicate probabilities of belonging each instance to classes 0 and 1, respectively. For calculating the fpr and tpr measures, however, we only need the second column of that array (i.e., `y_pred_proba[:, 1]`).

```
>>> from sklearn.metrics import roc_curve
>>> fpr, tpr, cut_offs = roc_curve(y_test, y_pred_proba[:, 1])
```

Now that we have fpr and tpr values associated with different cutoff values for classification, we can simply plot them against each other (using the `plot` function from the `pyplot` submodule of the `matplotlib` library).

```
>>> import matplotlib.pyplot as plt
>>> plt.plot(fpr, tpr)
>>> plt.show()
```

The output of these code is shown in Figure 9.17.

Alternatively, we can directly calculate the area under the ROC curve without necessarily plotting it. For this purpose, we need the `roc_auc_score` function from the metrics submodule of the sklearn library:

```
>>> from sklearn.metrics import roc_auc_score
>>> roc_auc_score(y_test, y_pred)
0.8976640711902113
```



FIGURE 9.17 A Screenshot of the Python Code On.

Please note that, because of the randomness in the process, your accuracy results may be significantly different than the ones presented here. To test this, you can rerun the whole notebook again and observe the differences between different runs. These differences are especially significant when the data set is heterogeneous and relatively small in number of records. One can eliminate the difference by setting random number seeds at any/every function that involves randomness.

CLASSIFICATION USING AN IMBALANCED DATA SET In this part, we are going to use an imbalanced version of the IMDB data set to train and test the same classification model. With an imbalanced data set there is always the risk of training a biased model toward predicting the majority class, when we use the regular classification algorithms. Various approaches are suggested to cancel out such bias; in this tutorial, we demonstrate random over sampling and random under sampling as two common approaches to deal with this problem.

First, as a benchmark, try to replicate all the above steps with the imbalanced IMDB data set provided and note the performance metrics. For us, the benchmark model led to the following performance:

```
Accuracy: 0.8208955223880597
Precision: 0.8181818181818182
Recall: 0.2903225806451613
F1-score: 0.4285714285714286
Area under ROC: 0.635452552458503
```

The results clearly show that the trained model is biased toward predicting more negative instances (i.e., low recall score). We are going to improve this model by adding a few more steps to the procedure to balance the training portion of the data (the test portion must remain intact to reflect the actual performance of model on a data set with real proportions of different classes).

To perform the training using a resampled balanced data set we use the `RandomOverSampler` function from the `over_sampling` submodule of the `imblearn` library in Python. Undersampling can also be done in a similar manner, but using the `RandomUnderSampler` from the `under_sampling` submodule of the `imblearn` library. All the steps from data import until data split are identical to the benchmark model. However, before feeding the training data into the `MLPClassifier` function we are to perform a few additional steps. First, import the required function from the library:

```
>>> from imblearn.over_sampling import RandomOverSampler
```

Then create a `RandomOverSampler` object (we call it `ros`) using a random state value and fit the training data arrays to that to that to adjust it:

```
>>> ros = RandomOverSampler(random_state=100)
>>> ros.fit(X_train, y_train)
```

Finally, apply the adjusted `ros` object to the training arrays (using the `fit_sample` function) to resample them and create balanced training arrays (we call them `X_train_resampled` and `y_train_resampled`):

```
>>> X_train_resampled, y_train_resampled = ros.fit_sample(
    X_train, y_train)
```

Now we can proceed the model training and evaluation steps using the resampled train arrays and the original test arrays.

```
>>> import sklearn.neural_network as ann
>>> ann_model2 = ann.MLPClassifier((4), activation=
'logistic', max_iter= 800)
>>> ann_model2.fit(X_train_resampled, y_train_resampled)
>>> y_pred = ann_model2.predict(X_test)
>>> accuracy_score(y_test, y_pred)
0.820895
>>> precision_score(y_test, y_pred)
0.620689
>>> recall_score(y_test, y_pred)
0.580645
>>> f1_score(y_test, y_pred)
0.6
>>> y_pred_proba = ann_model2.predict_proba(X_test)
>>> from sklearn.metrics import roc_curve
>>> fpr, tpr, cut_offs = roc_curve(y_test, y_pred_proba[:,1])
>>> import matplotlib.pyplot as plt
>>> plt.plot(fpr, tpr)
>>> plt.show()
```

The output of this code is shown in Figure 9.18.

```
>>> from sklearn.metrics import roc_auc_score
>>> roc_auc_score(y_test, y_pred)
0.736924
```

Compared to the benchmark model, clearly the oversampling strategy led to improving the model performance. Of course, even the current performance may not look ideal which is most probably due to employing only 2 predictor features. You may modify the code and include more and more features that you think are relevant to predicting the sales and observe changes in the model performances.

Similarly, keeping all the data preprocessing steps unchanged, you may try other classification algorithms available through the sklearn library and see if they perform better in predicting the target variable. Below you may find the official documentation helping you for implementing other popular classification algorithms through this library (e.g., k-nearest neighbors, logistic regression, decision tree, random forest, and many more).

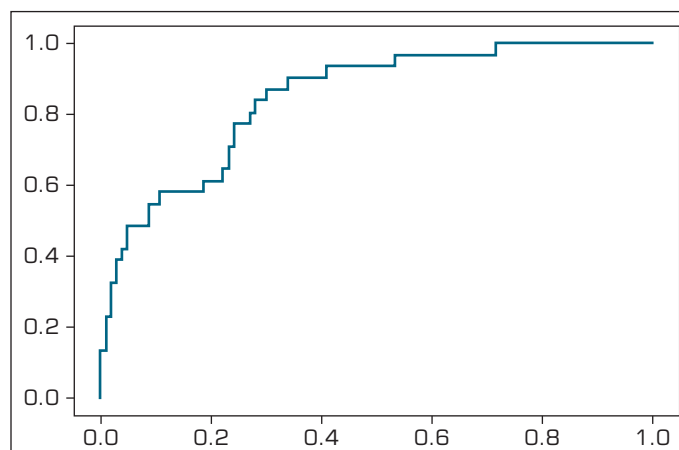


FIGURE 9.18 Visualization of the ROC Curve.

In addition, the MLP neural network classifier in sklearn involves several more settings that provides you with more flexibility to optimize your models (see https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html)

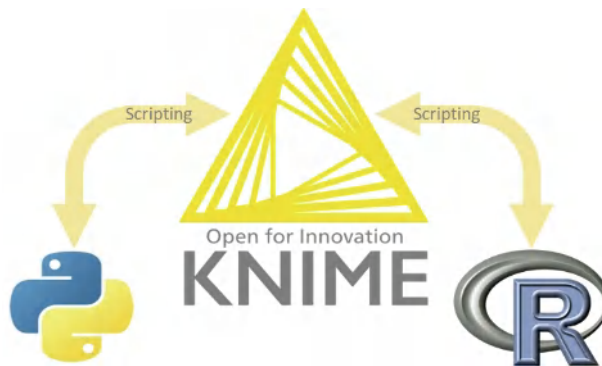
► SECTION 9.3 REVIEW QUESTIONS

1. What are the advantages of using a programming language like Python or R for analytics?
2. What are the commonalities and differences between Python and R for analytics?
3. What are the most popular machine learning libraries in R used for analytics projects?
4. What are the most popular machine learning libraries in Python used for analytics projects?

9.4 FREE AND OPEN-SOURCE ANALYTICS' VISUAL TOOLS

There seems to be a misconception (a myopic view) about free and open-source tools being programming based and commercial and proprietary tools being visually oriented. Although it is true that the most popular programming-based analytics tools (i.e., R and Python) are free and open-source but there are several, highly capable visually oriented analytics tools that are also free and open source. In this section, we will review and exemplify some of the most commonly used visually oriented (also called **visual programming** based) analytics tools.

KNIME



KNIME (Konstanz Information Miner) is a free and open-source data analytics, visualization, reporting, and data bending/integration software platform. The development of KNIME was started in early 2004 by a team of software engineers at University of Konstanz, in Germany. The original developer team headed by Michael Berthold came from a company in Silicon Valley providing software for the pharmaceutical industry (Berthold et al., 2020). The initial goal was to create a modular, highly scalable, and open data processing platform that allowed for the easy integration of different data loading, processing, transformation, analysis, and visual exploration modules without the focus on any particular application area. The platform was intended to be a collaboration and research platform and also serve as an integration platform for various other data analysis projects. In its early days, mid-to-late 2000s, KNIME was developed to be used primarily in pharmaceutical analytics research. Since then, it has evolved to a generic tool that can be used in all types of analytics applications including CRM for customer data analysis, business intelligence, **text mining** and financial data analysis. Recently attempts were made to use KNIME as robotic process automation (RPA) tool. Currently,

KNIME's headquarters are based in Zurich, with additional offices in Konstanz and Berlin (Germany), and Austin (USA).

I, Dursun Delen, one of the coauthors of this book, in my over 30 years of analytics experience, have used many of the analytics tools and platforms, commercial and free/open-source for educational (teaching business analytics and data science courses at every level, undergraduate, masters, doctoral, and executive education) and commercial (consulting to for-profit companies, nonprofit organization, as well as government agencies) purposes. Although I still use a variety of analytics tools and data science platforms (because I am a true believer of the toolbox approach, as opposed to a single tool for everything), for the last several years, I have been using KNIME as my core analytics platform. With KNIME, I have been able to create and deploy data science solutions for complex problems that resulted value creating outcomes for my clients, high-impact publications in reputable journals, and easy to understand and follow type case studies and tutorials for my professional books and textbooks. One particular feature of KNIME that makes it very attractive for me is its open and expandable architecture that enables me to incorporate the capabilities of popular data science tools like Python, R, Spark, H2O, among other into the workflow in a seamless manner. The following are a few of the main criteria that make a strong case for KNIME as an analytics platform (Silipo, 2020).

PROJECT CONSTRAINTS—TIME AND MONEY As mentioned above, KNIME Analytics Platform is one of those free and open-source software for data analytics. The implications of being open source and free-of-charge are self-explanatory, reducing the licensing legal headaches and the impact on the project budget. The impact of visual programming (using a workflow-type graphical user interface) might require a few more words of explanation. Visual programming has become quite popular in recent times, and it aims at substituting, partially or completely, the coding practice. In visual programming, a Graphical User Interface (GUI) guides you through all necessary steps to build a pipeline (workflow) of dedicated blocks (nodes). Each node implements a given task; each workflow of nodes takes your data from the beginning till the end of the designed journey. A workflow substitutes a script; a node substitutes one or more script lines.

In KNIME Analytics Platform, nodes are created by drag-and-drop (or double-click) from the Node Repository into the workflow editor in the central part of the KNIME workbench. Node after node, in a logical order, the pipeline is quickly built, configured, executed, inspected, and documented. See a screenshot of the KNIME Analytics Platform graphical user interface in Figure 9.19.

Visual programming is a key feature of KNIME Analytics Platform for quick prototyping, as well as ease of explaining and understanding the underlying logic of the project. It makes the tool very easy to learn and use by less-expert consumers of business analytics. Producing a few different experimental prototypes, before deciding the final direction of the project, is fast and quite straightforward. The ease of implementation frees up time to think deeper of possible theoretical alternatives to the current solution. I, as an instructor, like the fact that it makes my job easier to explain the complete process of converting data into actionable insight on a logical/orderly fashion with the help of an intuitive graphical workflow.

THE LEARNING CURVE Visual Programming also makes the learning curve much shorter and faster than code-based programming tools. Data science is now used in more or less all disciplines, including the humanities, languages, life sciences, economics, social sciences, engineering, and the rest of the other unsusceptible areas of the human wisdom. Not all scientists (or managers/decision-makers) are expert programmers and not all of them have enough spare time to become expert programmers. A GUI-based visual tool can be learned and applied in much less time than a code-based tool, freeing up again precious time and resources for more important investigations, conceptualizations, and problem solving.

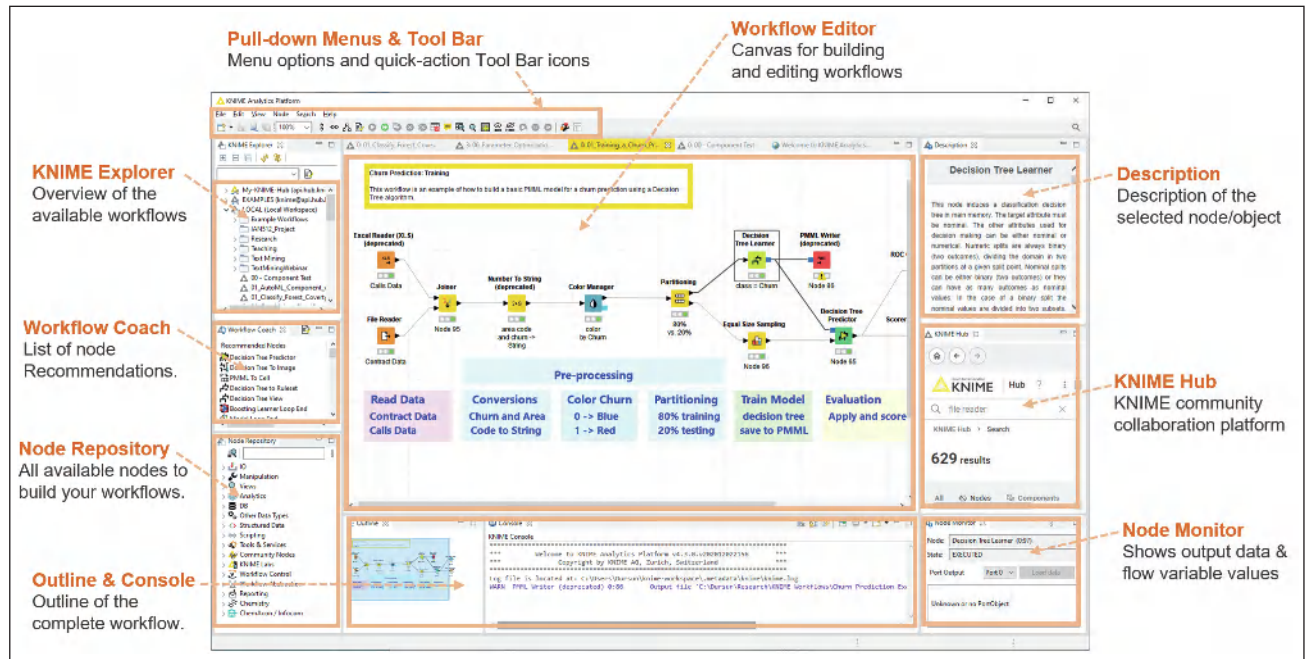


FIGURE 9.19 KNIME Analytics Platform User Interface.

Also, when preparing to become future scientists in the humanities, languages, life sciences, economics, or other disciplines, a GUI-based tool might free up more time to study and research the underlying concepts and theories. The idea is to shorten the time spent on syntactic details of programming in order to spend more quality time on the semantics of model building and problem solving. I have seen too often entire months dedicated to learning the syntactic nature of the coding practices, before even approaching and implementing any data analysis technique. With KNIME Analytics Platform, in just a few weeks, you can assemble quite capable workflows for data transformation and for the training and testing of a wide variety of machine learning algorithms. That said, for uniquely complex and novel projects, data scientist may need to incorporate some unique and advanced functionalities that only exists in analytics programming languages, like Python and R, into the model building and problem-solving process. Having built on an open and extendable architecture, KNIME Analytics Platform also allows for those types of extensions through a selection of scripting and wrapper nodes.

THE KNIME COMMUNITY Another great help for beginners comes from the **KNIME Hub** (<https://hub.knime.com/>). With the KNIME Hub, the creators of the platform have expanded the already existing set of resource for KNIME Analytics Platform within the KNIME Community. The KNIME Hub is the public repository for the KNIME community. Here, you can share your workflows and download workflows created by other KNIME users. Once you locate the workflow you are interested in by a simple search in the Hub, you can simply drag-and-drop it from the Hub page into your KNIME Analytics Platform workflow pane. Once in the local workspace, you can start adapting it to your data and your needs. Of course, you can also share your own work on the KNIME Hub for the public good. Just copy the workflows to share from your local workspace into the *My-KNIME-Hub/Public* folder in the KNIME Explorer panel within the KNIME workbench (see Figure 9.20).

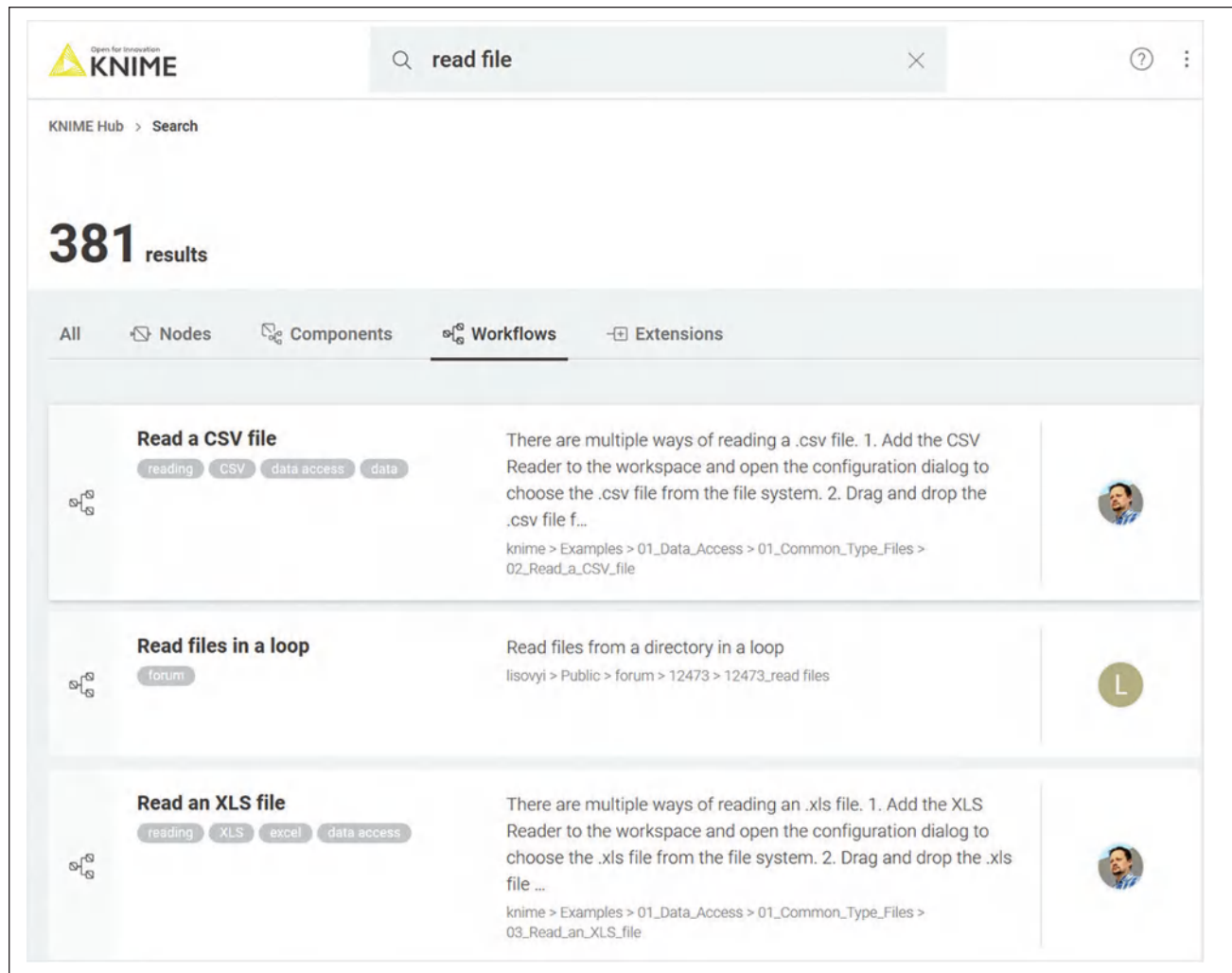


FIGURE 9.20 Resulting List of Workflows from Search “read file” on the KNIME Hub.

The KNIME community engagement does not stop at the KNIME Hub. It is indeed very active with tips and tricks on the KNIME Forum (<https://forum.knime.com/>) as well. Here, you can ask questions or search for previous answers. The community is very active, and it is highly likely that somebody has already asked your question. Finally, contributions by the community are available as posts on the KNIME Blog (<https://www.knime.com/blog>), as books in the KNIME Press (<https://www.knime.com/knimepress>), as videos on the KNIME TV (<https://www.youtube.com/user/KNIMETV>) channel on YouTube.

CORRECTNESS AND FLEXIBILITY Easy is good, but can I also ensure it is correct? Is it flexible enough to experiment with alternative methods and procedures? This is indeed a key question, because for many of the software tools available nowadays “easy” comes with the tradeoff of “control” for correctness and of “flexibility” for alternative options, often manifests as a canned, black-box type solution that dictates “trust me, I know what I am doing.” For data scientist, these types of magical solutions have always been questionable—they want to have the full control of the semiexperimental process and its underlying options toward finding the most “optimal” solutions.

Automated Machine Learning (also called AutoML or AML) has become very popular in recent times. It carries the promise to get your data, spit out some results, without you even lifting a finger. As fascinating as this option might sound, together with the promise it also carries some risk. First of all, it works as a black-box. Its decision process is not transparent. When I feed the box with my data, I have to completely trust the machine that the data analysis process is correct, fits the data characteristics and distributions, and is tuned on the problem I wish to solve. It is probably a trusting exercise, like in these psychology support meetings. I, personally, prefer to make sure that all steps in the analysis are implemented as per the best practices, under my control, according to the original design of the application. Although AutoML is a desired final stage in data science, because of the artistic and creative nature of the solution development, it is still a work in progress, and for real data scientist, nothing more than an early experimental step in piloting. KNIME Analytics Platform provides a very modular and detailed workflow development environment tooling with an extensive collection of data wrangling and machine learning algorithms, allowing for ample flexibility to systematically and easily optimize the modeling parameters.

EXTENSIVE COVERAGE OF DATA SCIENCE TECHNIQUES Another necessary complement to the ease of use is the coverage of data science techniques. Without an extensive coverage of the commonly and less commonly used data wrangling techniques, machine learning algorithms, wide range of data types and formats, and lack of integration with the most commonly used database software and data sources, reporting tools, other scripting and languages, ease of use would be of limited convenience. A capable analytics platform is expected to keep up with the latest trends and technologies through native implementations and/or seamless external connections. Since no tool can do it all, integrating into and playing nice with other tools and technologies in the fast-evolving business analytics and data science landscape is not just a “nice to have capability” but is a requirement for a comprehensive analytics platform. KNIME Analytics Platform offers one of the most comprehensive coverage of machine learning algorithms—from traditional ones (e.g., linear and logistic regression, decision trees, neural networks, support vector machines, naïve Bayes, k -nearest neighbor) to more recent ensembles (e.g., random forest, gradient boosted trees) and deep learning algorithms. A large number of nodes is also available in KNIME to implement a myriad of data wrangling techniques. KNIME Analytics Platform also connects to most every data source from databases to cloud repositories, from big data platforms to individual flat files (Delen et al., 2021).

KNIME Analytics Platform has a seamless integration with several popular reporting platforms including BIRT, Tableau, Qlik, PowerBI, and Spotfire. In most cases, it turns out that it is not even necessary to move to an external reporting software tool to visualize your data and your analysis results. Several JavaScript based nodes are already available within KNIME, in the Views/JavaScript category in the Node Repository. These nodes implement data visualization using a variety of plots and charts: from a simple scatter plot to a more complex sunburst chart, from a simple histogram to a parallel coordinate plot, and much more. If you combine these visuals within a component (an integrated modeling construct in KNIME), the resultant visual becomes an interactive, integrated, and informative information dashboard (see Figure 9.21).

DATA SCIENCE IN THE ENTERPRISE One last step is to deploy the outputs of the modeling efforts into production so that the insight generated by these models can be consumed for better and faster decision-making. In case of an enterprise, this deployment must be quick, easy/intuitive, seamless, and secure. This is the last and perhaps the most

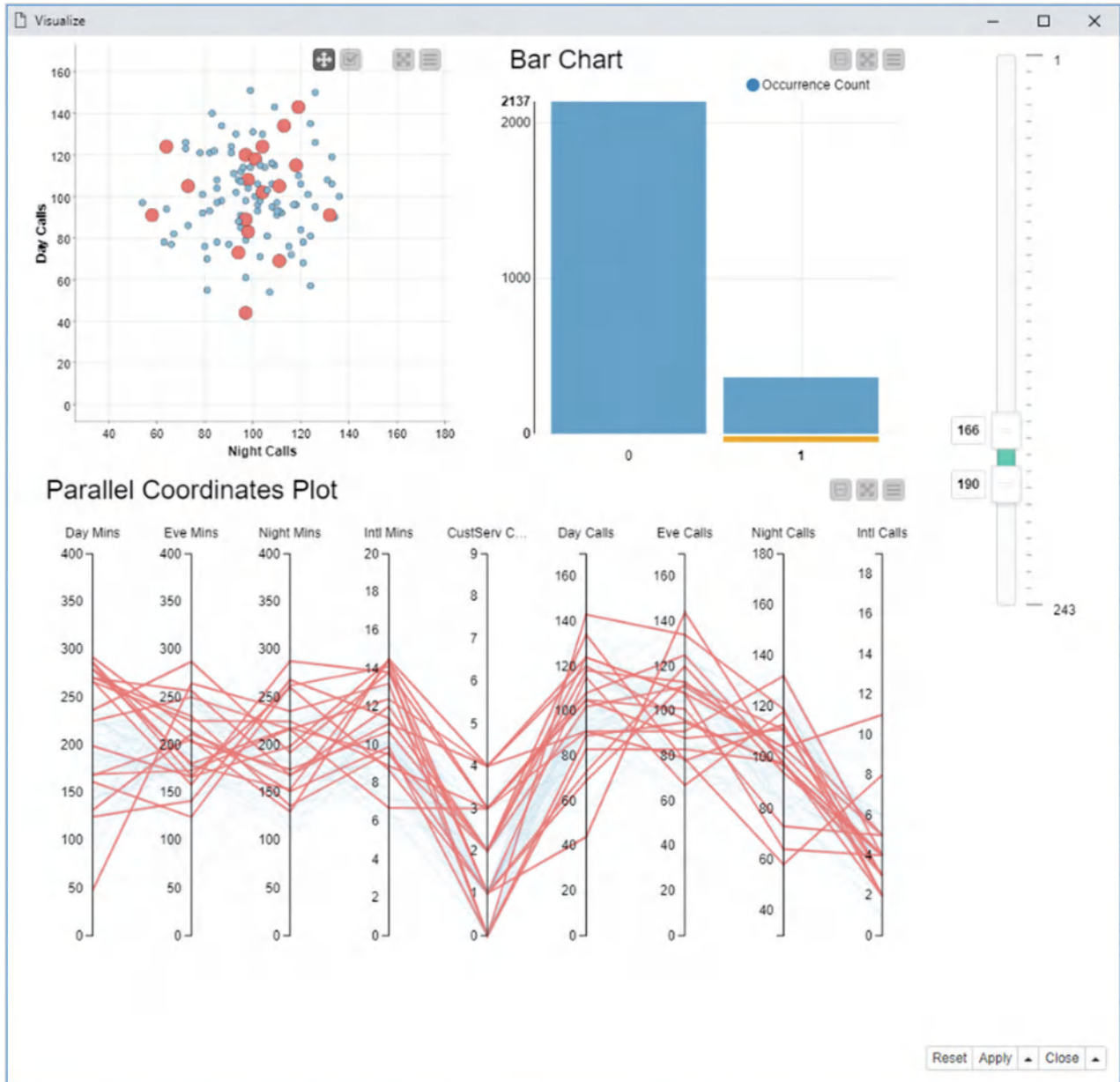


FIGURE 9.21 The Composite View from a Component Containing a Few Charts and Plots.

critical step in the chain of actions in a data science project. This process of moving the application into the real world is often called “moving into production” or “deployment.” KNIME Analytics Platform works seamlessly with KNIME Server to make the transition from design and development to productizing and deployment an easy and quick process (see Figure 9.22 and Table 9.2).

Table summarizes some of the most prevailing characteristics that make KNIME Analytics Platform an excellent analytics tool for business analytics and data science learning, teaching, and practicing (Delen, 2021).

Analytics in Action 9.2 is about the use of KNIME along with R programming for better management of medical supply inventories to identify and eliminate waste.

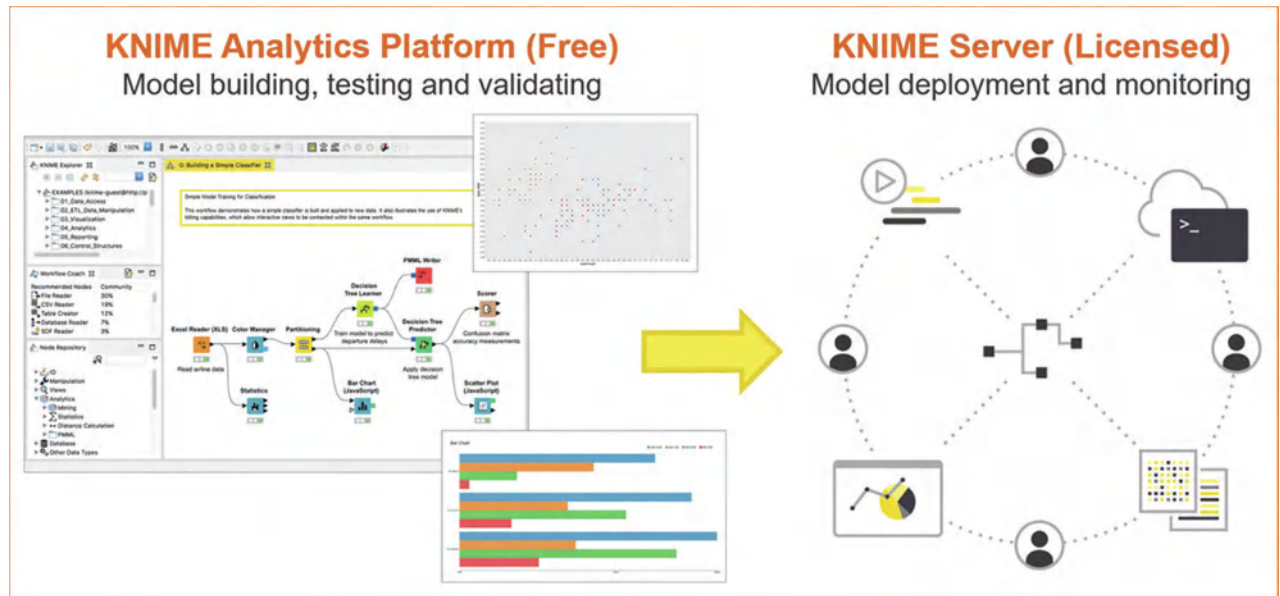


FIGURE 9.22 Integration between KNIME Analytics Platform and KNIME Server.

TABLE 9.2 The Top 10 List for “Why Should You Consider KNIME?”

1	Easy to Use	Graphical user interface with an intuitive workflow type model building logic makes the analytics platform very easy and intuitive to learn, teach, and use
2	Open Source	The source of the analytics platform is open to everyone for exploring and innovating.
3	Cost	It is free! Free for everything Not a scaled-down, time-limited, trial, or community version, but the full functional version is free for everyone (educators, researchers, and practitioners) for everything—learning, teaching, and consulting.
4	Community Support	Access to KNIME gurus as well as the large and highly active community of users for help on questions, modeling hints, and example workflows is just a few clicks away on KNIME website.
5	Rich Functionality	With over 2000 native nodes and many more from the third parties (via extensions), KNIME offers perhaps the richest collection of data science functions.
6	Platform Agnostic	Natively runs on all three popular operating systems—Windows, MacOS, and Linux.
7	Connectivity	Connects to most every data source (local, Web-based, or cloud) and consumes any and every data type (structured, semistructured, unstructured)
8	Expandability	Allows for building your own nodes and use the nodes build by others. Also, you can expand the functionality and expandability via JavaScript, Python, and R integration nodes.
9	Modularity	Allows for creation of reusable modeling artefacts at varying levels of granularity via the use of meta-nodes and components.
10	Deployability	Putting the workflows in perpetual use via deployment to the cloud (on KNIME server or on your own space in AWS or Azure) is a quick and straightforward process.

ANALYTICS IN ACTION 9.2 Leveraging Predictive Analytics Prevents \$1.3 Million Worth of Medical Supply Waste

Closing the Never-Ending Waste Cycle

In the United States alone, \$5 billion worth of medical supplies are thrown away each year. This epidemic of waste in health care facilities raises the cost of care and takes up space in landfills—both of which have significant social and environmental consequences. This waste can be attributed to poor inventory management. Without accurate inventory tracking and analysis, stockpiles up and sits, unused, until it expires.

This project, which was run by Z5 Inventory—a full-service supply chain lifecycle solution company for health care providers—consisted of two main phases. Phase one: clean out all the excess products (those not purchased recently or above the Periodic Automatic Replenishment levels) from 30 health care facilities and send them to a Z5-run warehouse. Phase two: use predictive analytics to reallocate these products to health care facilities in need.

The chief goal was to help clients reduce product overstock, gain shelf space for newer products, improve staff efficiency, reduce inventory supply costs, and reduce waste in the health care supply chain to as close to zero as possible. If this process were to be implemented in all US health care facilities, billions could be saved annually. This would ultimately save valuable taxpayer money, reduce landfill levels, and give at-risk communities supplies they might not have otherwise had access to.

Forecasting Demand Using KNIME and R

After determining the project objectives with key stakeholders (Z5 and client C-Suite), data scientists drafted out the components of the project (i.e., demand forecast, reallocation, best cycle selector, etc.), and how they should interact.

The first step was processing the data for analysis, checking for and handling errors such as missing values, invalid entries, and date range errors. Then, using the KNIME Integration with R, data scientists could begin the demand forecasting (based on the purchase history of the health care facility), and create a list of products for each facility. R was used to split excess products into different destination facilities based on a ranking. This ranking was calculated using purchase frequency, average purchase quantity, standard deviation of purchase quantity, as well as the predicted quantity on the wish list. Additionally, since the goal was to reallocate products to health care facilities as much as possible, a “best cycle selector” was created to ensure that for a given product, its use was maximized in the destination facility.

Throughout the process, data scientists checked back in with key stakeholders, presenting interim results, and making necessary changes to the KNIME workflow. Robustness of models is continually evaluated based on customer feedback. Until now, there have been no complaints of over-shipment, which indicates that the model is successful in helping to solve the challenge of over-stocked health care facilities.

Saving of \$1.3 Million

This project was carried out in the Mid-Atlantic region of one of the largest health care networks in the United States, where 32 health care facilities participated. USD 1.3 million dollars' worth of medical and surgical product was moved to prevent expiration and disposal within the project execution. This is a phenomenal achievement. It indicates that, if all US health care providers employed a similar strategy, nearly all the USD 5 billion that goes to waste would actually be used.

Ideally, by utilizing this process, the central warehouse that handles distribution to health care facilities will become perfectly efficient. To be precise, it would be empty. The current projection is being able to distribute all product within the following month. In a perfect world, waste in the health care supply chain would be reduced to zero. Realistically, given the number of uncontrollable variables in the industry such as clinician preference, supplier and manufacturer changes, and the increasing rate of industry consolidation, this is not possible. However, what can be controlled is reducing products going to waste on shelves and being thrown away.

Advantages of Using KNIME

KNIME nodes are visual and self-documenting, which saves a significant amount of time and makes results easier to understand for the noncoder. Sometimes, they can even reproduce the results without knowing how to code. Previously, when coding directly in R or Python, a lot of time was spent on documentation to ensure the scripts were easy to read and results were reproducible. This is now time, which can be spent modifying and improving the project processes.

Because the data scientists all come from a statistics background and code in R daily, they remain happy because KNIME gives them the flexibility to still work in R whilst remaining in one uniform platform. They can therefore take advantage of the best of both worlds. KNIME Analytics Platform is a great tool with many powerful capabilities such as comprehensive ETL nodes, an easy-to-use drag-and-drop interface, and integrations with many other popular data science and data mining tools. It simply makes doing analytics and pre-processing tasks simpler and faster.

Source: KNIME Customer Success Stories. “Leveraging Predictive Analytics Prevents \$1.3 Million Worth of Medical Supply Waste” retrieved from <https://www.knime.com/solutions/success-story/leveraging-predictive-analytics-prevents-millions-of-dollars-worth-of-medical-supply-waste> (accessed August 2022); and <https://www.knime.com/about>.

TUTORIAL

Predicting Customer Churn

The data set includes 1,000 rows/customers and 38 columns, a mix of numeric and nominal variables. The target variable is named Churn with values of Y (yes) and N (no). The data set also has a unique row identifier, named CutID. Follow these steps to create a workflow for customer churn.

Once you start the KNIME Analytics Platform (KAP), you will see the Welcome screen with some latest developments on software updates, training classes, and exiting news from the KNIME community. The following steps will help you create your first complete workflow.

- Step 1. Create the workflow group and the workflow.** Before reading the data into the platform, it is a good practice to create a *Workflow group* to collect and organize your data sets and workflows experiments. To do so, from the File menu, select File→New, and from the selection window, select New KNIME Workflow Group (see Figure 1). Name your Workflow group as Customer Churn Project. Then, using the same window, create a New KNIME Workflow and name it Customer Churn Model. See the results in Figure 9.23.
- Step 2. Read the data.** In KAP, you can get the data into your Workflow in three different ways: (1) you can drag-and-drop the proper data reader node (in this case, CSV Reader node) from the IO folder in Node Repository, (2) drag it from your file explorer directly into the workflow (in this case, KAP will sense the data type and use the appropriate reader node to ingest the data), and (3) drag and drop it into your workflow group, so that it is in the same place where your workflows will be, and then drag-and-drop it into the workflow designer. The best practice suggests following the option #3 (see Figure 9.24). Once the data are imported in the workflow, run it and inspect the File Table (via the last option in the executed node’s context/right-click menu).
- Step 3. Inspect the Variables.** You can do this by using either a Statistics node or Data Explorer node. Once you connect and execute one of these nodes (via context menu option, Execute and Open the View). In the Data Explorer View, you will see your variables are groups in two main tabs, Numeric and Nominal. Go through each and every variable in these two nodes, paying attention to

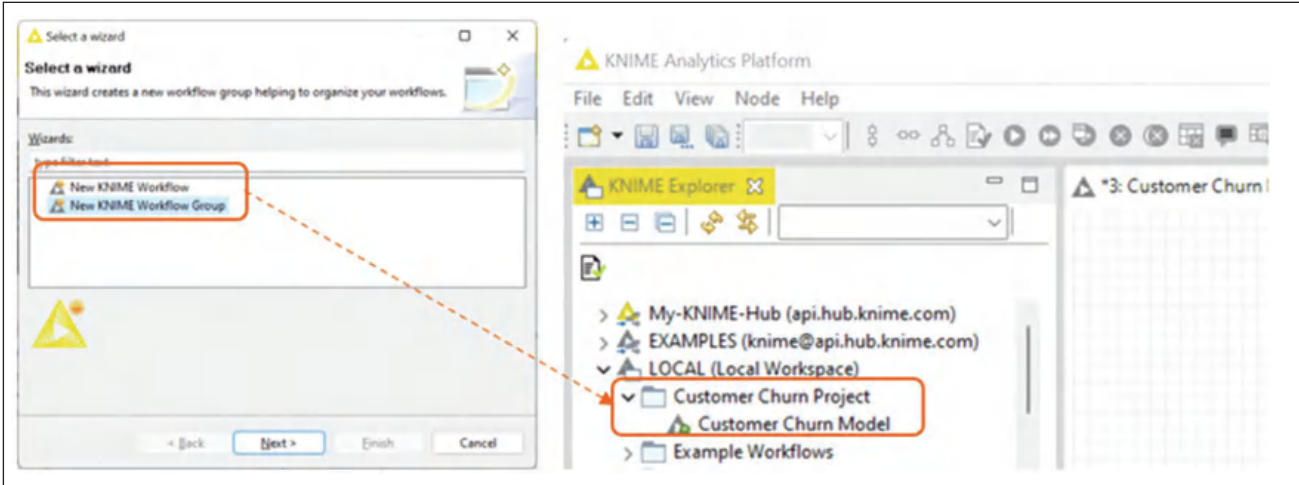


FIGURE 9.23 Creating a New Workflow Group and then a Workflow in KNIME.

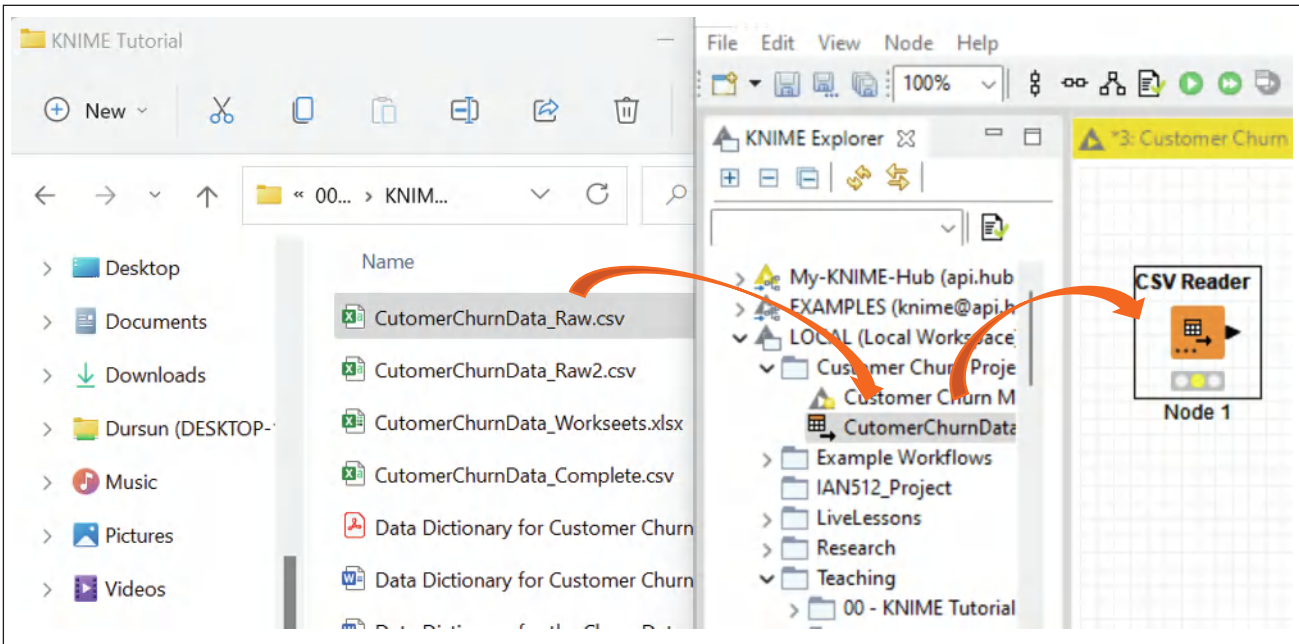


FIGURE 9.24 Imputing Data into the Workflow Group and then Dropping It into Designer.

anomalies and missing values (notice that Tenure in numeric and Confer and Ebill in Nominal tabs have some missing values).

Step 4. Impute the missing values. Using the Missing node, we will impute the missing values in the data set. Missing Value node has a Default imputation procedure that applies the same rule to all variables of the same type (in this case, we will use Median for all numerical variables, and Most Frequent Value for all nominal/String variables) (see Figure 9.25). One can also select and impute each variable with different method/rule using the Column Setting tab on this node description.

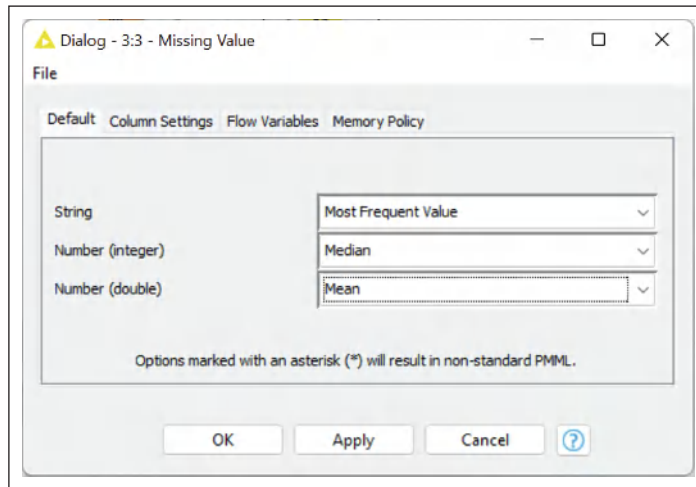


FIGURE 9.25 Configuration Interface for Missing Value Imputation Node.

- Step 5. Preprocess the data.** Here we can filter columns and rows (using Column Filter and Row Filter nodes) that are not useful for the intended task. We can also transform any data item into different data types/values/formats, and also create new columns using a variety of nodes like Rule Engine, Math Formula, Numeric Binner, among others. In this case, we will use a Column Filter to filter out the unique row identifier, CustID. Here, we can also use Color Manager to assign different colors to the target variable values for more appealing visualization down the road.
- Step 6. Split the data.** Here we will use a Partitioning node to split the data into training and testing using a 30/70, stratified random sampling procedure (see Figure 9.26). We also use a static random number seed for replication.

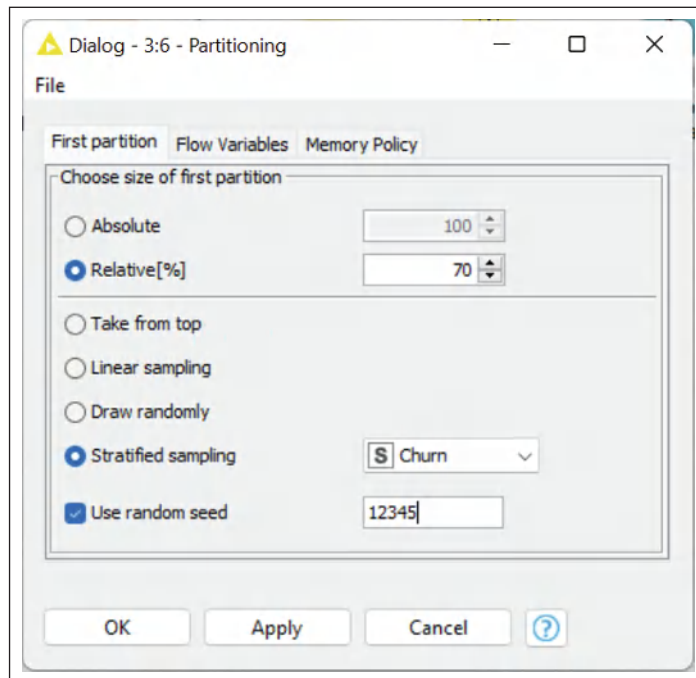


FIGURE 9.26 Configuration of the Data Partitioning Node.

Step 7. Train and test the model. Here we will use decision tree. We will feed 70% of the data (i.e., the training portion) to the Decision Tree Learner node and remaining 30% to the Decision Tree Predictor node. Then, we will connect the model output from the learner node to the model input of the predictor node. In the Decision Tree configuration window, make sure to select the target class variable properly, and for the visual simplicity, also select Binary nominal splits (see Figure 9.27). You can also experiment with other decision tree creation hyperparameters to obtain better performance. From the context menu, clicking on the Execute and Open View will show the decision tree in Figure 9.28. The colored bar charts are because of the Color Manager we added in the workflow in Step 5.

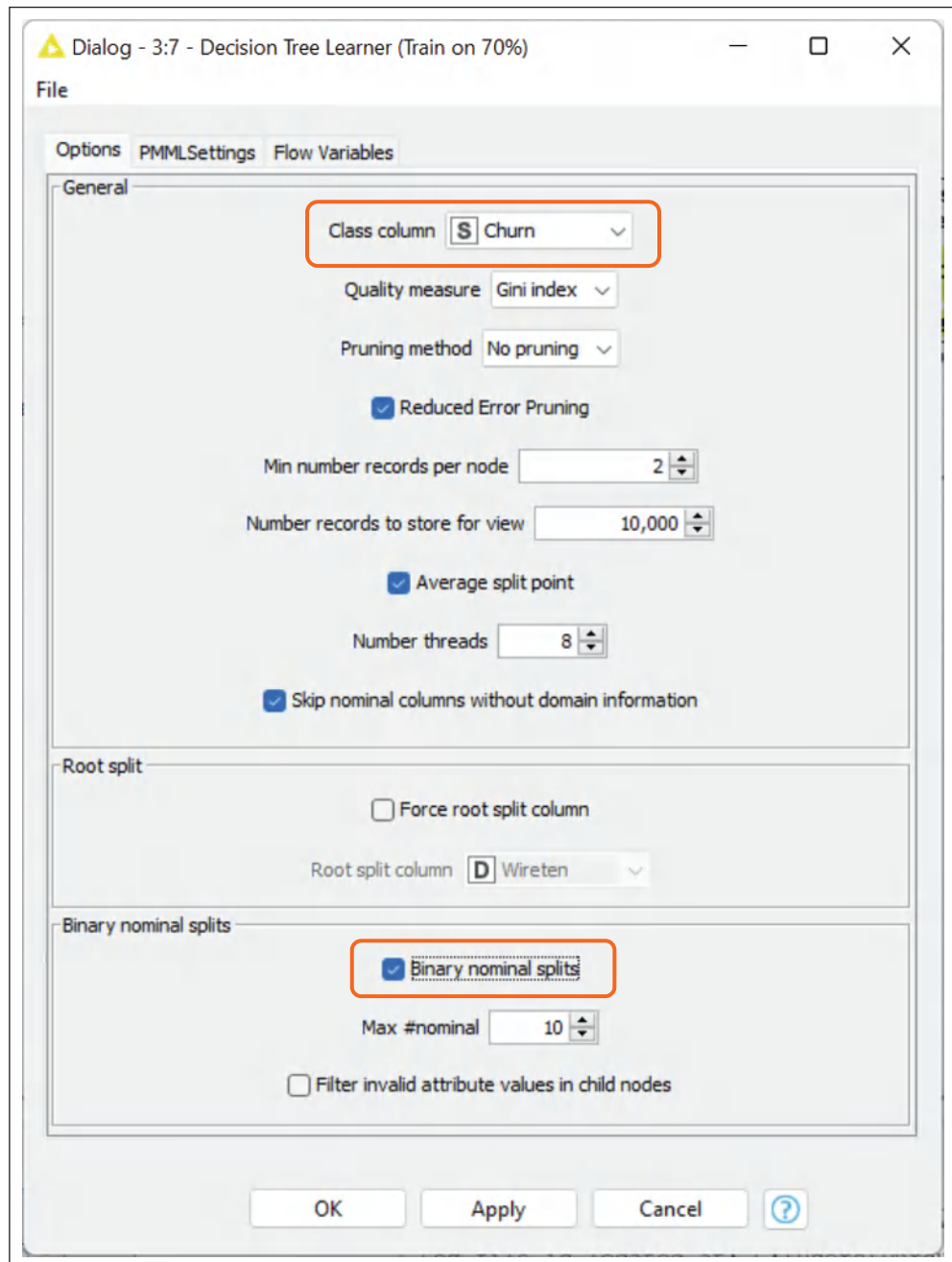


FIGURE 9.27 Configuration Specifics of the Decision Tree Lerner Node.

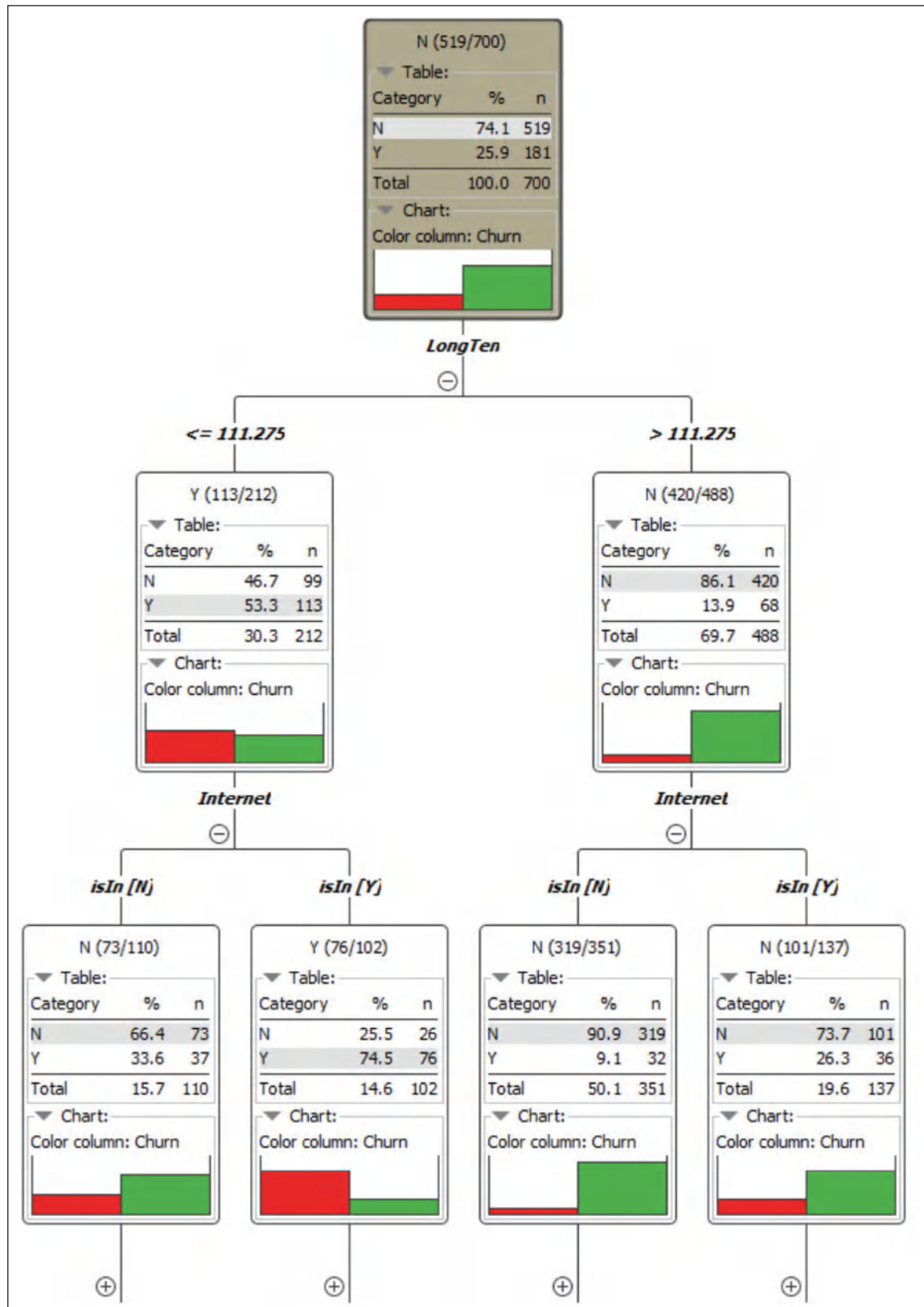


FIGURE 9.28 A Snapshot of the Induced Decision Tree (the Output of the Decision Tree Learner).

In Decision Tree Predictor node description window, we check “Append column with normalized class distributions” and enter “_DT” in the Suffix for probability columns” text field. We will need these probability columns to be able to create ROC curve in Step 8.

Step 8. Score the model. Here we will score the prediction performance of the developed decision tree model on the test data set. To do so, we will use Scorer node, where we will set the First Column to Churn and Second Column to Prediction (Churn). Then, in the context menu, we click on Execute and Open Views option to create the output shown in Figure 9.29. To create an ROC curve, we use ROC Curve node, and configure its values as shown in Figure 9.30a and Figure 9.30b shows the resultant ROC curve.

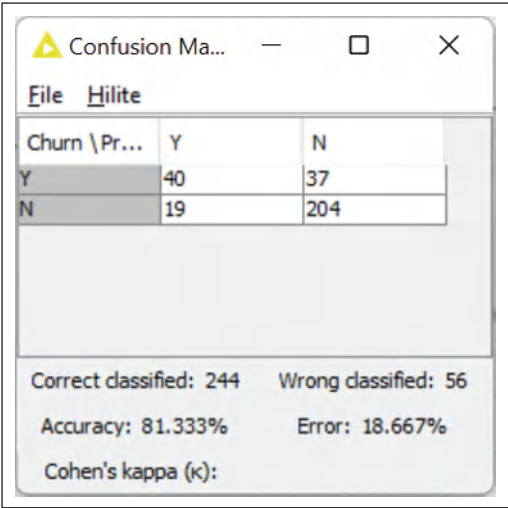


FIGURE 9.29 Output of the Confusion Matrix Node.

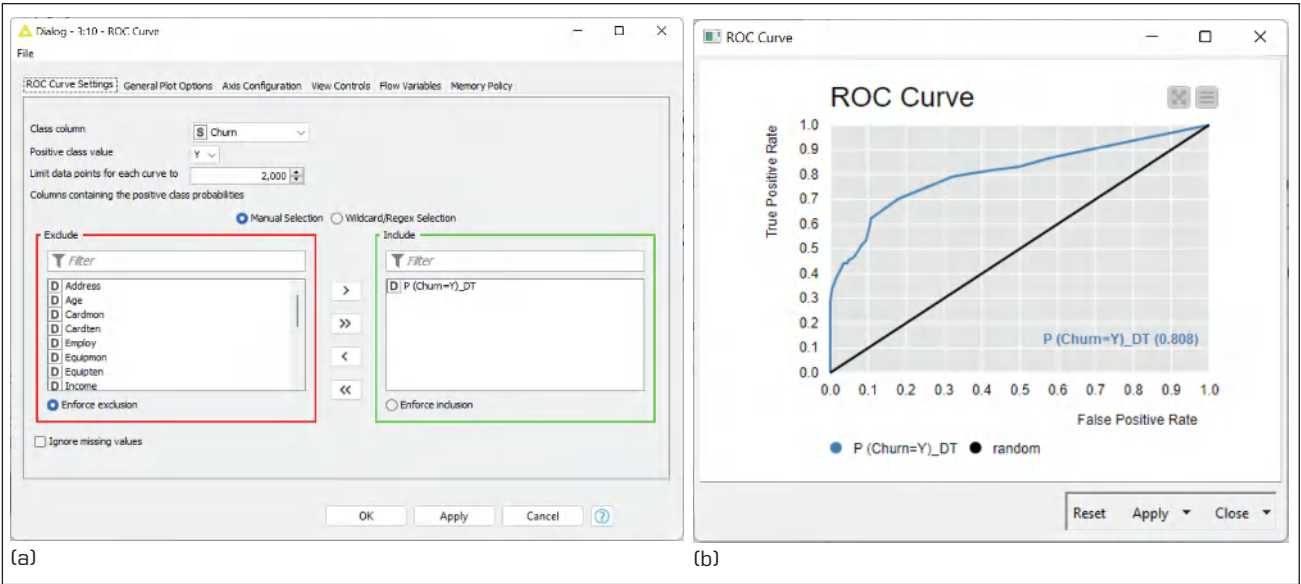


FIGURE 9.30 (a) Configuration of the ROC Curve Node. (b) Output of the ROC Curve Node.

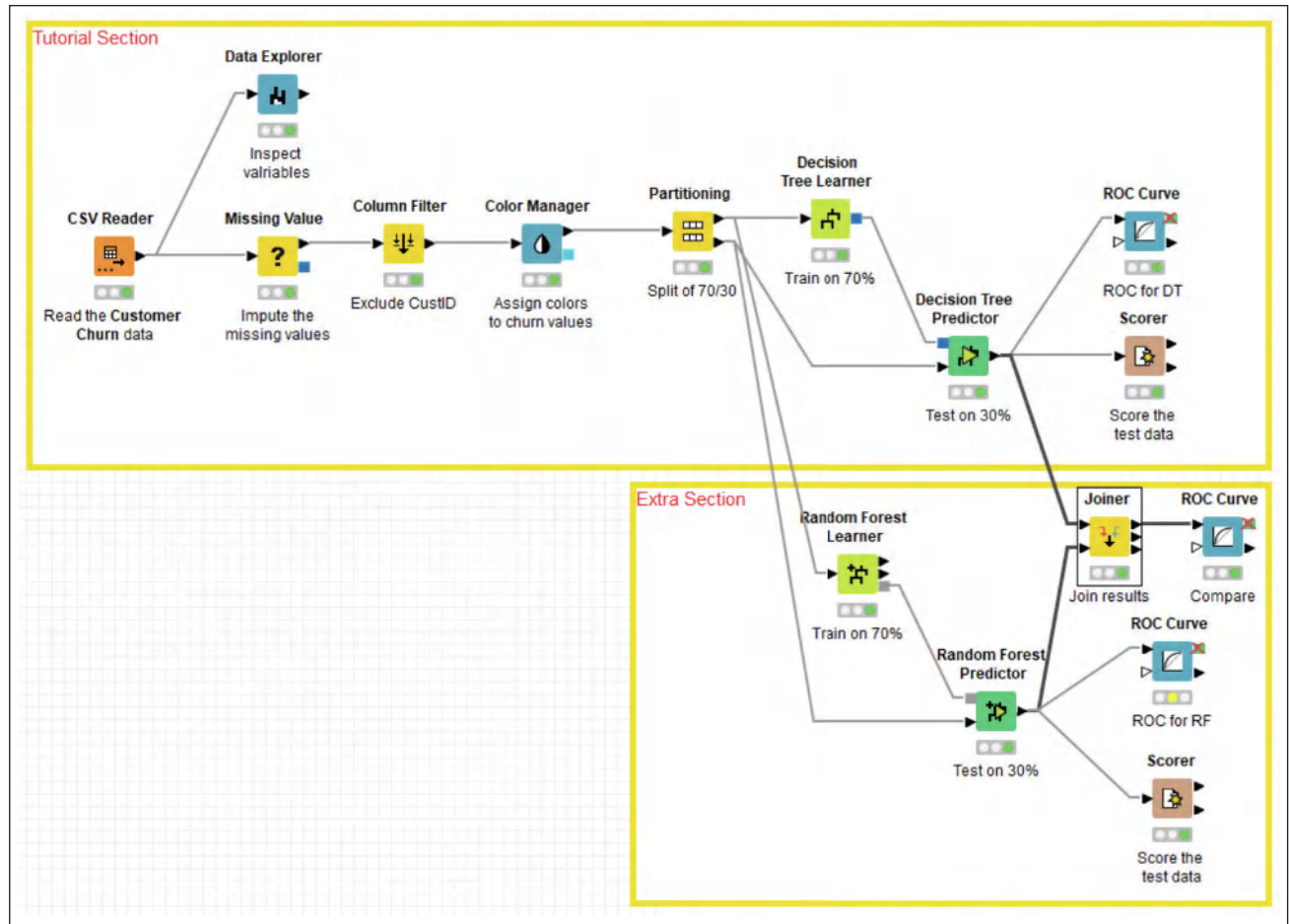


FIGURE 9.31 The Complete Workflow for Tutorial and an Extra Section for Random Forest.

The complete workflow for this tutorial is shown in Figure 9.31. To expand on this, we can use the same training and testing data partitions to easily create a comparative analysis between decision tree and random forest methods (or for any number of machine learning methods) within the same workflow (see the Extra Section in Figure 9.31).

Orange



Orange is a free, python-based, open-source software developed by a group of academics at the University of Ljubljana (<https://orangedatamining.com>). Orange has a component-based data mining software architecture that includes data visualization, exploration, preprocessing, and modeling techniques. Orange also offers text analytics module called the Text Mining widget for textual data analysis. To illustrate the functionality of Orange, in this section we present a tutorial on text mining of published literature. We will use the same data to conduct the similar text mining using JMP, a commercial tool, in the section 9.5, and make some high-level comparative analysis between Orange and **JMP Pro**.

TUTORIAL

Text Mining of Published Literature

The increasing interest in text analytics and text mining has led to development and offering of many software tools. The increase in textual content coming from a wide variety of sources coupled with exponential increase in social media content in different languages have effectively resulted in a plethora of text mining software tools. In this chapter, we will examine Orange.

To streamline the text mining process, we present a generalized text mining framework (see Figure 9.32), which can be executed with any analytics software with text

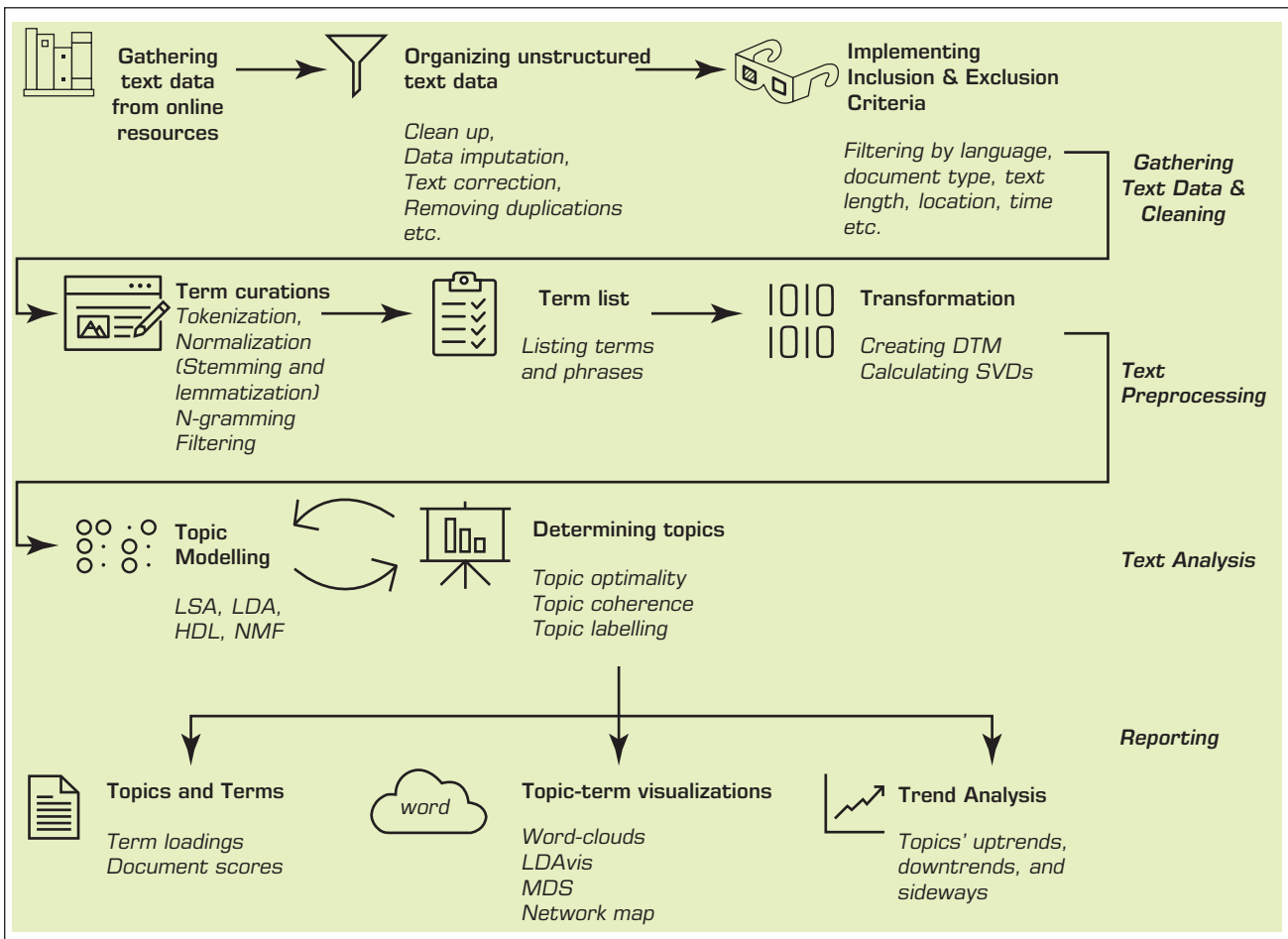


FIGURE 9.32 The Four-Step Text Mining Framework.

mining capabilities. This process can be summarized in 4 main steps: (1) Gathering and cleaning of textual data, (2) Preprocessing of the consolidated data, (3) Analysis of the preprocessed data, and (4) Reporting of the findings.

In the first step, the textual data, on which the analysis is considered, is collected from relevant sources in commonly accepted file types (.xls, .txt, .CSV, etc.). Then, the obvious anomalies in this textual content are identified and corrected. Finally, if there are any inclusion or exclusion criteria as per the analytics problem being addressed, a proper subset of the data is created by filtration.

In the second step, document-based list of terms and phrases are obtained by performing text processing tasks that include tokenization, normalization, n-gram generation, stop-word filtering, among others. This resultant data is then converted into a document-term-matrix (DTM). Finally, a dimensional reduction is applied to the large and sparse table (i.e., DTM) using singular value decomposition (SVD) method to reduce the dimensionality and obtain more interpretable results.

In the third step, the analytics ready, structured textual data are used to discover new knowledge. Specifically, at this stage, the modern topic modeling methods, Latent Semantic Analysis (LSA), **Latent Dirichlet Allocation (LDA)**, Hierarchical Dirichlet Process (HDP), and Negative Matrix Factorization (NMF), as well as some of the classical models like Latent Cluster Analysis (LCA) and Discriminant Analysis are used to find natural patterns and themes/topics are calculated.

In the last stage, the discovered findings are interpreted, contextualized, and visually presented as reports. The artifacts may include document-topic tables and visuals, topic-terms tables and visuals, dynamic word clouds, or longitudinal or topic-based trend graphs.

The textual data used in this tutorial is obtained from digital online publication databases, which consists of all articles published in the Journal Information of Systems, up until the end of 2021.

Orange features a visual programming frontend for explorative rapid data analysis and interactive data visualization. The general view of its user interface is shown in Figure 9.33. In addition to the default functions provided with the initial install, (can be accessed via the toolbars), many add-ons created and provided by orange developers community can be added to the environment. The Widget Dock on the left side of the screen contains groups of widgets/functions to be used in analyses: Data manipulation functions can be found under the Data group, data transformation and manipulation functions (e.g., “filtering,” “select columns/rows,” “transpose,” and “preprocess”) can be found under the Transform group. The Visualize group offers many charting widgets. The Model, Evaluate, and Unsupervised groups provide supervised/unsupervised machine learning functions. Finally, the Orange provides a separate widget group under add-ons for Text Mining. Models in the Orange are built and run on the workflow bench. Widgets in groups are added to the workflow bench with the drag-and-drop method. The associated widgets are then linked/connected to each other.

In order to install and activate Orange’s Text Mining widget group, user need to select Add-ons under the Options tab, and activate/check the Text option. Once activates, Text Mining functions/widgets can be used for text inputting, manipulation, preprocessing, mining, and visualizations purposes.

Step 1. Gathering and cleaning the textual data

In Orange, textual data can be ingested in several ways. Cooked/processed or raw texts can be read using the “Corpus,” “Import Documents,” or “Create Corpus” widgets. Orange also allows users to extract text from some popular databases automatically. These online databases include The Guardian, NY Times, PubMed, Twitter, and Wikipedia. The

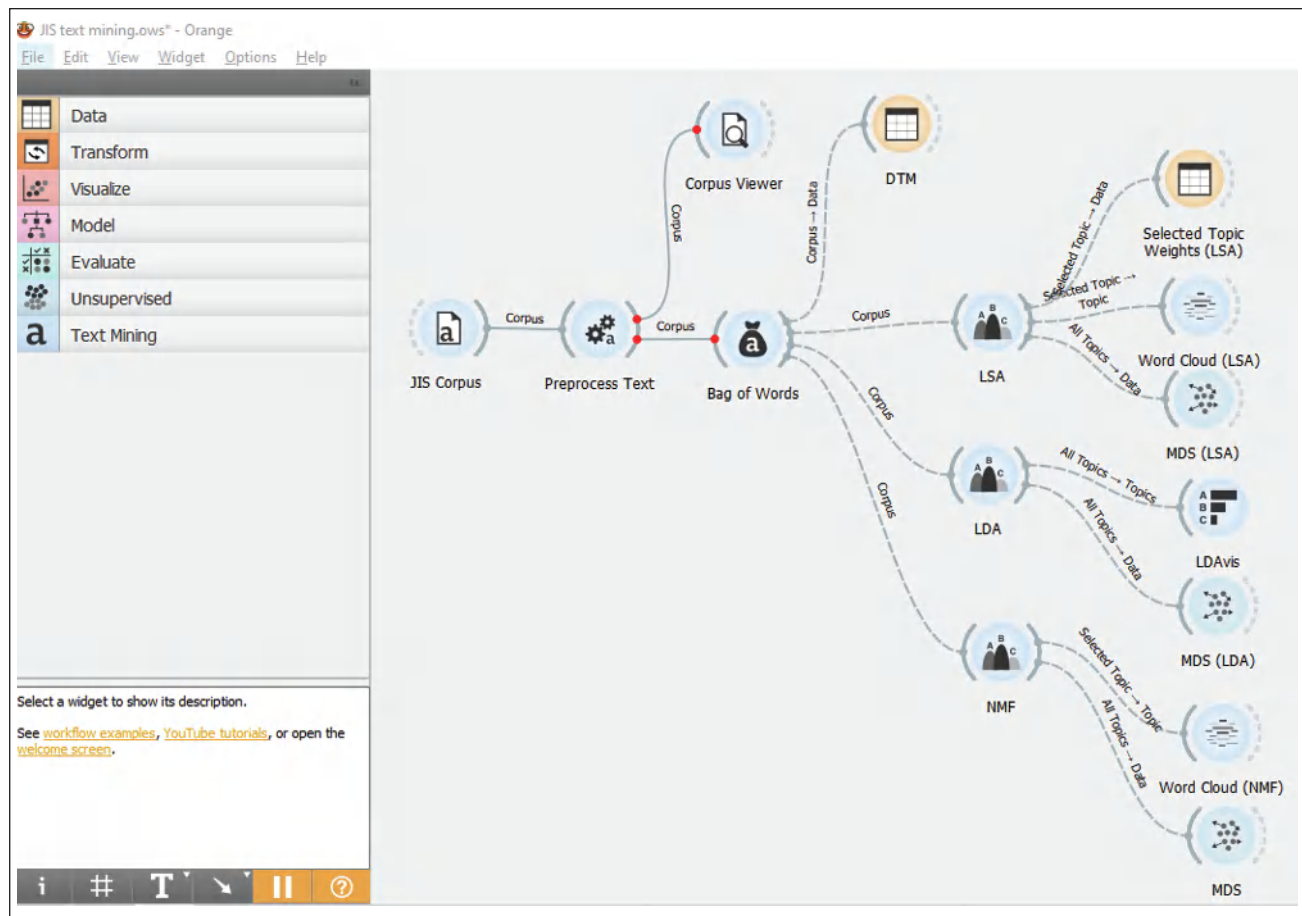


FIGURE 9.33 Orange Mainboard View.

“Corpus” widget can read textual data stored in Excel (.xlsx), comma-separated (.csv), and native tab-delimited (.tab) formats. The first operation to model the text data is to drag the “Corpus” widget and drop it onto the workflow. Then, click on the widget and use the browse tab in the Corpus widget to locate and read the text files. In case of duplications, users can use the “Duplicate Detection” widget to detect and remove duplicated texts before starting text preprocessing.

Step 2. Preprocessing the textual data

“Preprocess Text” widget is used for Text Preprocessing in the Orange. This widget can be dragged and dropped from the Text Mining module onto the workflow, or it can be added by clicking on the right side of the “Corpus” widget and selecting it from the widget menu that comes up during the link creation phase (see Figure 9.34a). Text operations can also be viewed in the corpus viewer widget (see Figure 9.34b).

The Preprocess Text widget has the necessary functions to perform the preprocessing tasks. (see Figure 9.35). At this stage, Texts can be split into smaller units (tokens), standardized, normalized (via stemming and lemmatization), and filtered based on stop-words, lexicons, Regex, and frequency. Also, users can create n-grams and part-of-speech (POS) tag markers.

After the preprocessing stage, DTM is obtained using the Bag of Words widget. DTMs can be modified under a variety of available options (e.g., Term Frequency, Document Frequency, Regularization). Finally, DTM can be examined and confirmed with a Data Table connected to the Bag of Words widget.

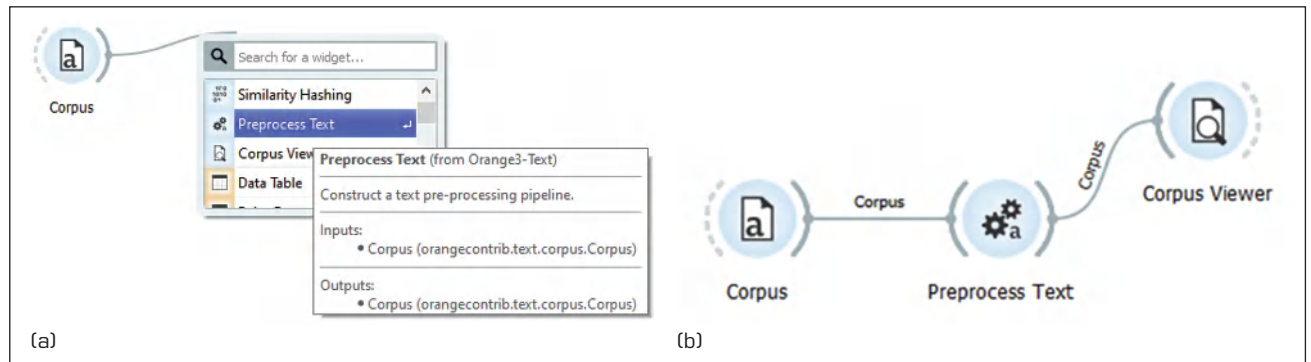


FIGURE 9.34 (a) Orange Linking the Corpus to Preprocess Text Widget. (b) Orange Linked the Corpus File with the Preprocess Text.

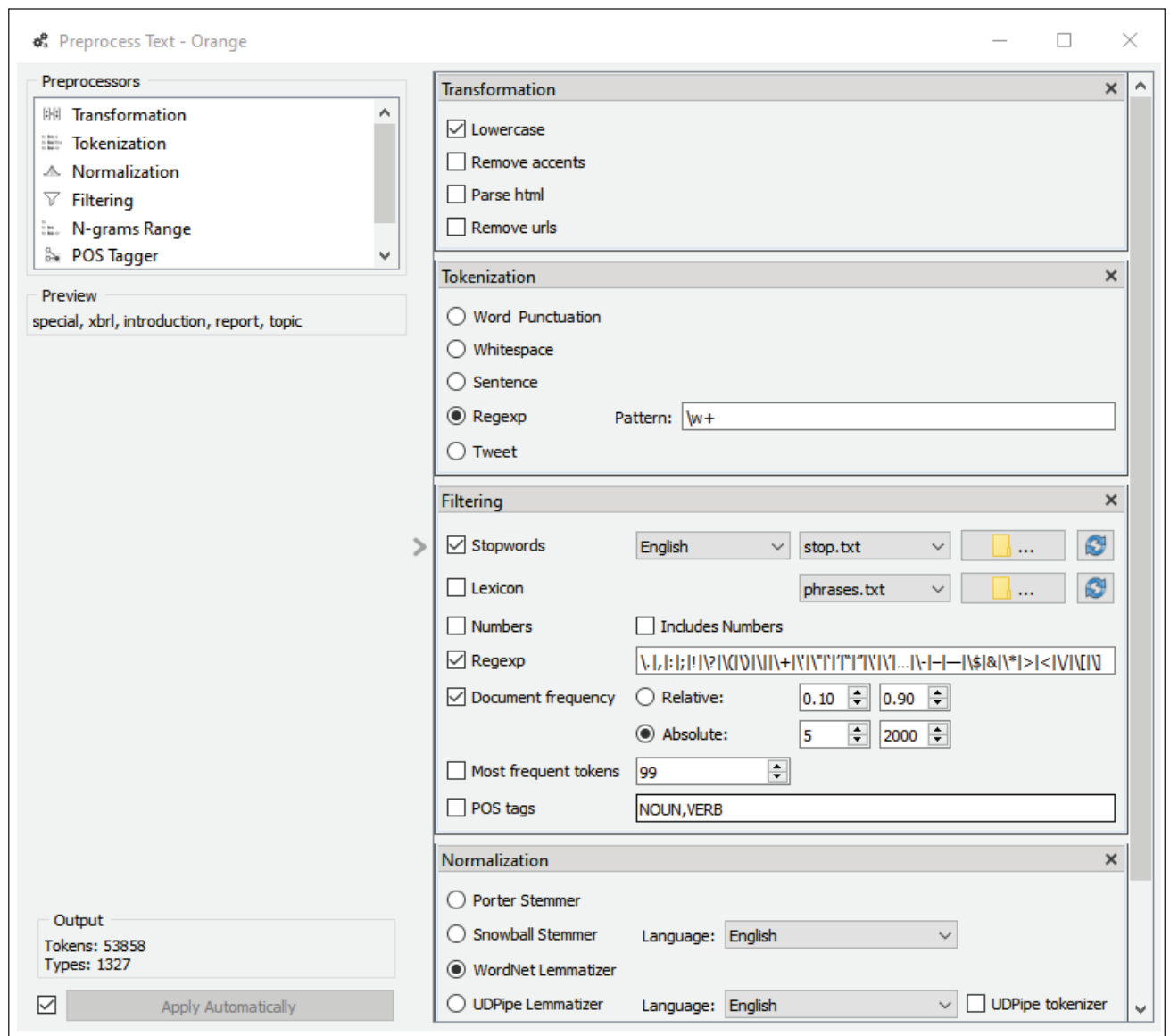


FIGURE 9.35 Orange Preprocess Text Dialog Box.

Step 3. Performing the text analysis

After obtaining the DTM, user can perform topic modeling and sentiment analysis through text mining widgets. Orange provides four types of topic modeling methods: Latent Semantic Indexing/Analysis (LSI or LSA), Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP), and Negative Matrix Factorization (NMF) (see Figure 9.36), all of which can be executed/run simultaneously.

Step 4. Reporting of the obtained results

Orange provides many ways to report on the text mining results. The most preferred reporting methods are the topic-term table, the word cloud widget, the multidimensional scaling (MDS) graph, and the LDAvis widget. Users can create topic-term tables by connecting a data table to the topic modeling widget. The connection type should be a “Selected Topic—Data” pairing to obtain the topic-term data table. After this process, the selected topic and the weights of these terms can be seen in the data table. Topics and terms can also be easily visualized using the Word Cloud widget.

The MDS widget creatively and intuitively illustrates the relationships among the topics. The MDS widget can be found in the Unsupervised Widgets group and can be executed by connecting it to a topic modeling widget. User can choose among the provided options to plot the desired MDS visualization (see Figure 9.37). MDS widget also offers optimization options: besides using the PCA (Torgerson) algorithm, the overlapping between topics (dots) can be manipulated through the Jitter and randomization can be achieved with Randomization.

The final reporting widget is the LDAvis., which lists terms within each topic along with the topic’s relevance score. The frequencies of each word in the overall and related topic can also be displayed together via a horizontal stacked bar chart.

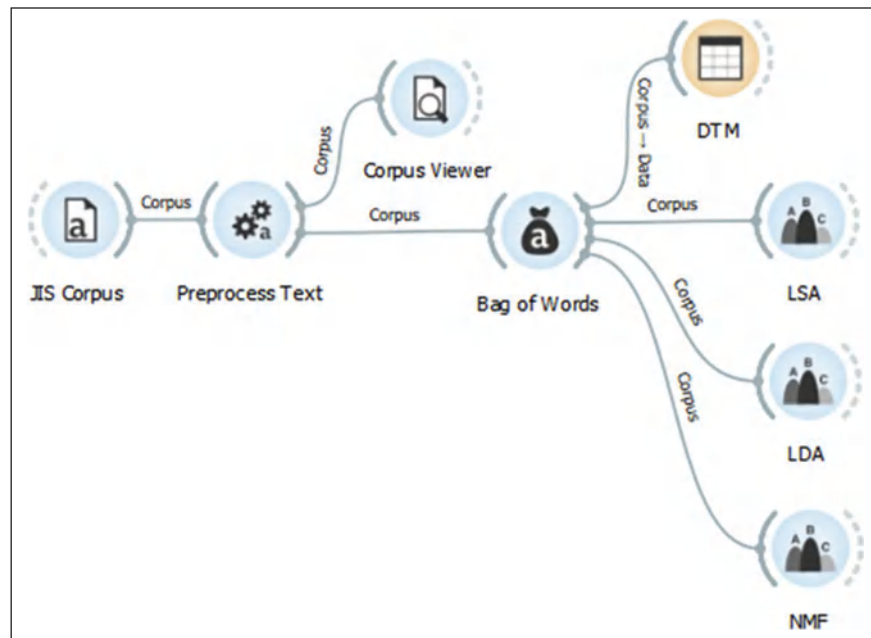


FIGURE 9.36 Orange Linking Topic Models.

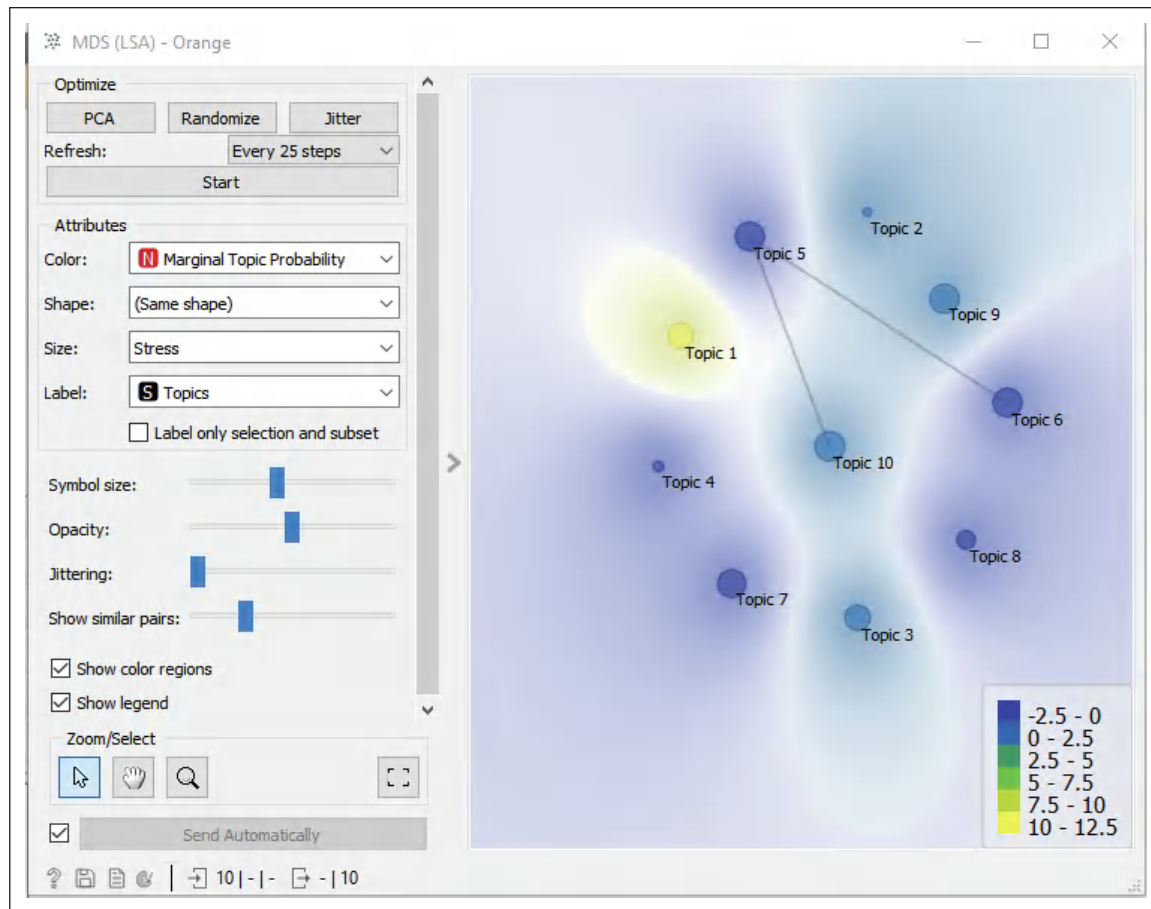


FIGURE 9.37 Orange MDS Optimization.

Weka



Weka (Waikato Environment for Knowledge Analysis), developed at the University of Waikato, New Zealand, is a free and open-source software for data mining. In addition to descriptive statistics and data visualization, Weka also supports most of the standard data mining tasks, such as data preprocessing, feature selection, association, clustering, classification, regression. It offers a rich collection of machine learning algorithms (both supervised and unsupervised).

The initial development efforts of Weka dates all the way back to 1993 in variety of procedural programming languages. In 1997, the developers of the tool decided to redesign and redevelop Weka from scratch in Java, including the implementations of modeling algorithms. Since its development, it has been used many universities around the world for teaching data mining.

Weka can be downloaded from Weka Wiki (https://waikato.github.io/weka-wiki/downloading_weka/) and installed on machines running on Windows, MacOS, or Linux. Once installed, user can develop data mining models using one of three options (see Figure 9.38): (1) Explorer (good old fashion menu and tab driven execution interface), (2) KnowledgeFlow (the workflow type model development, similar to other visual programming tools), (3) CLI (executing everything syntactically via command line interface).

In addition to these three, Experimenter can be used for designing controlled experiments, running those experiments, then analyzing the collected results for further model building. And the last but not the least, Workbench can be used combine all of the other applications (GUI interfaces) into a single interface, which is especially useful if you are switching between two or more different interfaces (e.g., going back and forth between Explorer and Experimenter) to try out a lot of what if scenarios.

The most commonly used application type (of its GUI options) in WEKA is perhaps the Explorer. Figure 9.39 shows a snapshot of Explorer with different panes (sub windows) and selection options while Figure 9.40 shows the decision tree development function along with Weka's decision tree viewer.



FIGURE 9.38 Weka's Data Mining Application (i.e., GUI selection) Menu.

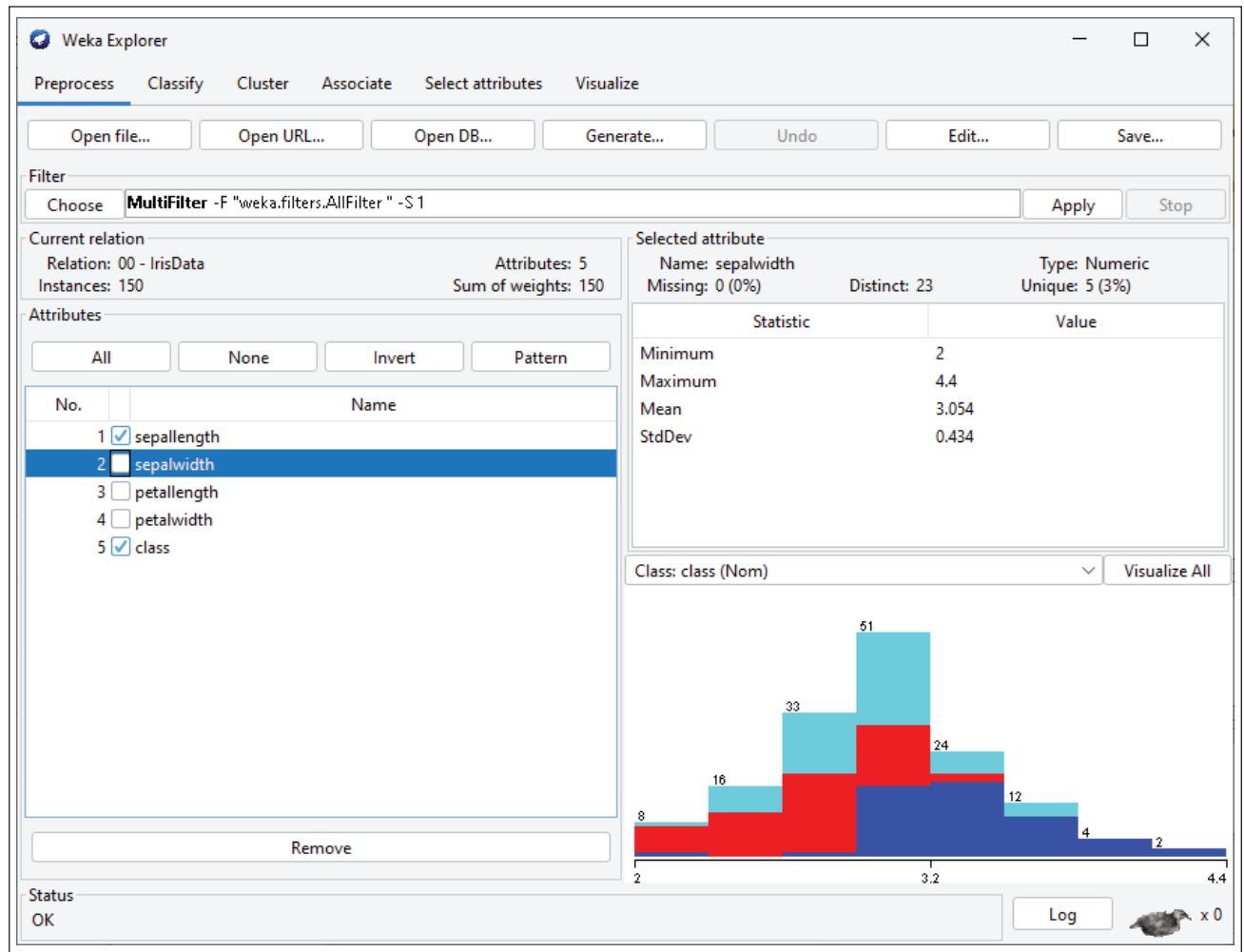


FIGURE 9.39 Weka Explorer User Interface with Menu Driven Modeling Options.

RapidMiner



RapidMiner, formerly known as YALE (yet another learning environment), is a comprehensive data science platform with visual workflow design capabilities. It was developed in 2001 by a group of researchers at the Artificial Intelligence Unit of the Technical University of Dortmund.

RapidMiner software tool can be downloaded from their official website (<https://rapidminer.com/>). It has a free version with limited functionality—use of a maximum of 10,000 data rows and only one processor. RapidMiner also offer an educational program where students and instructors can have access to a full educational license free of charge.

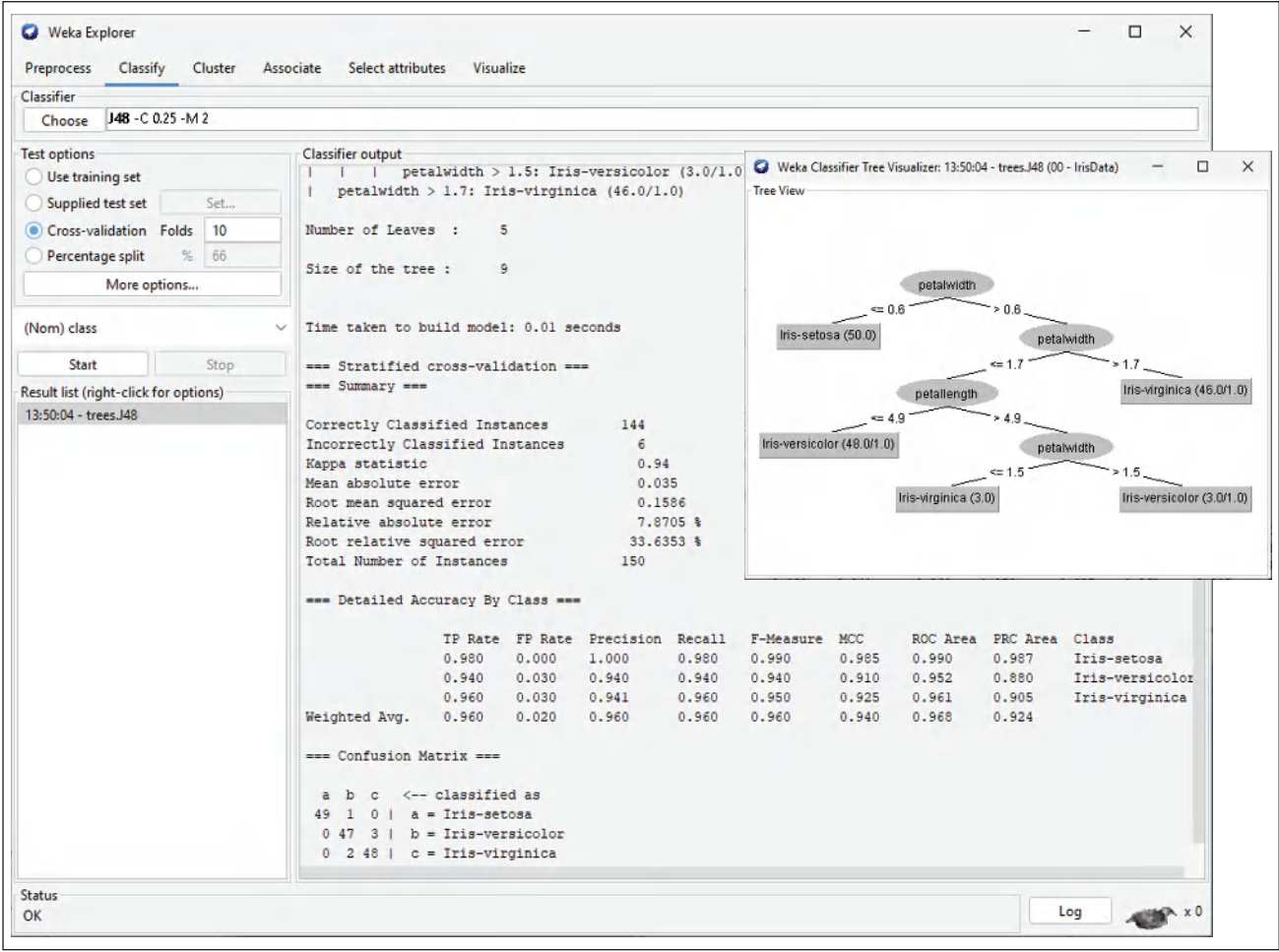


FIGURE 9.40 Developing a Decision Tree Model for the Iris Data Set.

The overview of the use interface is shown in Figure 9.41. As is the case in most visual programming tools, the interface provides a workflow development pane, surrounded by smaller windows with for file repository, operations/function library, parameter editing, and function level description/help.

TUTORIAL

Predicting Survival using the Titanic Data Set

In this tutorial, we will perform predictive modeling using the one of the built-in data set, Titanic. The goal is to classify survival of the passengers based on the provided multivariate features using a variety of classification methods.

When you start RapidMiner Studio, first interface you get is the Welcome screen with plenty of options to get a quick start on an analytics project (Figure 9.42). If you are an experienced user, you would perhaps start with a Blank Process and develop the solution completely from an empty workflow space. For this tutorial, we will select Auto Model, which is a recipe/wizard type user interface that help the user to easily and quickly develop an analytics application/workflow.

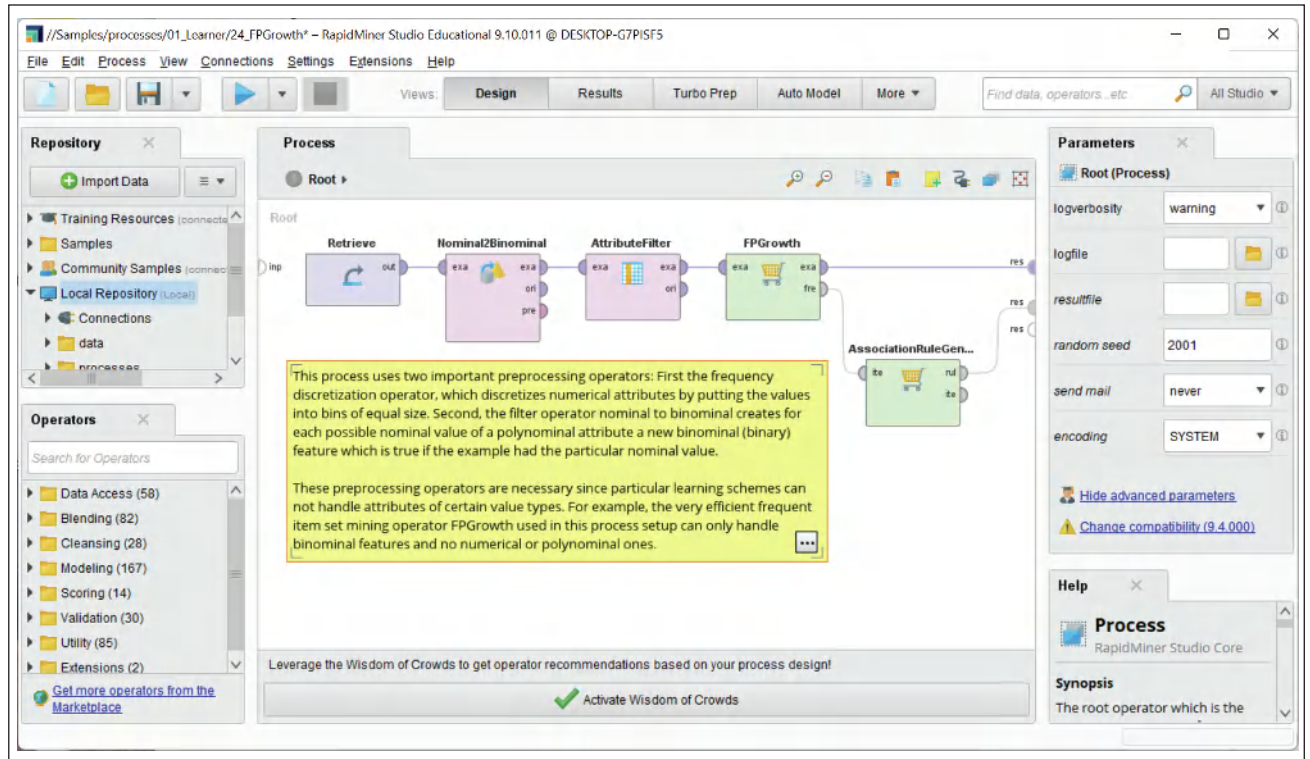


FIGURE 9.41 Overview of RapidMiner Studio User Interface.

Then we select the data set (Samples→Data→Titanic) and click NEXT to move on to the next step. In this Select Task step, we select Predict, and click on the Survived column as our selection of the target variable, and click NEXT. In the Prepare Data step, we are given the option of making changes to the default options on the cost/benefit values. We will keep the default values, and move on by clicking on NEXT. In this Select Input step, we can include or exclude columns from our input variable list. Here, we will again take the default suggestions and move on to the next step. This step, named Model Types, allows us to include or exclude some of the prediction model from the model building/texting/creating process. Here we will select all of the model types and accept their suggested default parameters, and click RUN. This step may take some time to complete because the tool will *training and test* all of the model types, while trying to optimize their learning hyperparameters. Once completed, we can see the comparative analysis of all the models in the results window (see Figure 9.43).

As can be seen, several model type produced very accurate models (classification error < 3%), deep learning being the most accurate and decision tree being the fastest. At this point, you can export the results to a third-party tool for better visualization and reporting (see Figure 9.44 for the available alternatives).

► SECTION 9.4 REVIEW QUESTIONS

1. What are the advantages of using visual programming tools like the ones covered in in this section?
2. Which of these tools you like the most? Why?

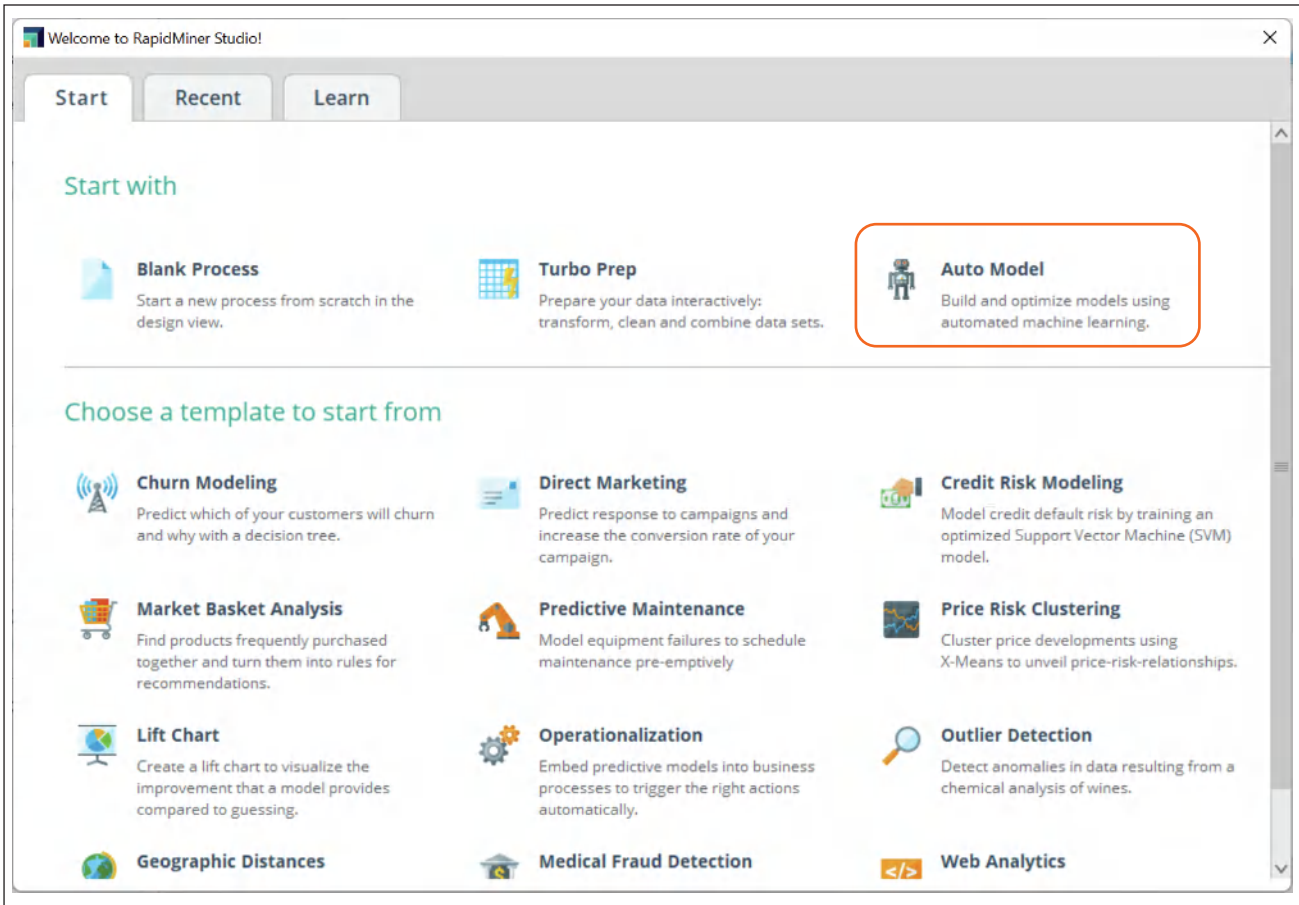


FIGURE 9.42 Starting Interface for RapidMiner Studio.

3. Do you know any other visual, free, open-source analytics tools? If not, you can just search for them and see what you can find out. Since the tools landscape is highly dynamic and still expanding, there is a good chance that you can find some interesting ones.

9.5 COMMERCIAL ANALYTICS TOOLS

The popularity of analytics along with its machine learning capabilities in business world dates back to 1980s with the strong emergence of data mining and knowledge discovery tools and technologies. Back then, there were not widespread use of Python or R as analytics programming languages (although R was started to be developed as a statistics library in mid 1950s and Python as a general-purpose programming language in early 1990s), or KNIME and Orange as free and open-source visual programming interfaces in practice. There was Weka, however, as a free and open-source tool, but mostly to be used for data mining education and academic research, rather than in real-world analytics applications. Most every other analytics (data mining) tools were commercial and proprietary. Some of the noteworthy examples of such tools include **Clementine**, PolyAnalyst, Statistica, IBM Intelligent Miner, SAS Enterprise Miner, **KXEN**, XLMiner, among others. The movement of free and open source has emerged after the turn of the century, and gained significant momentum after 2010. Creating, sharing, and consuming have become

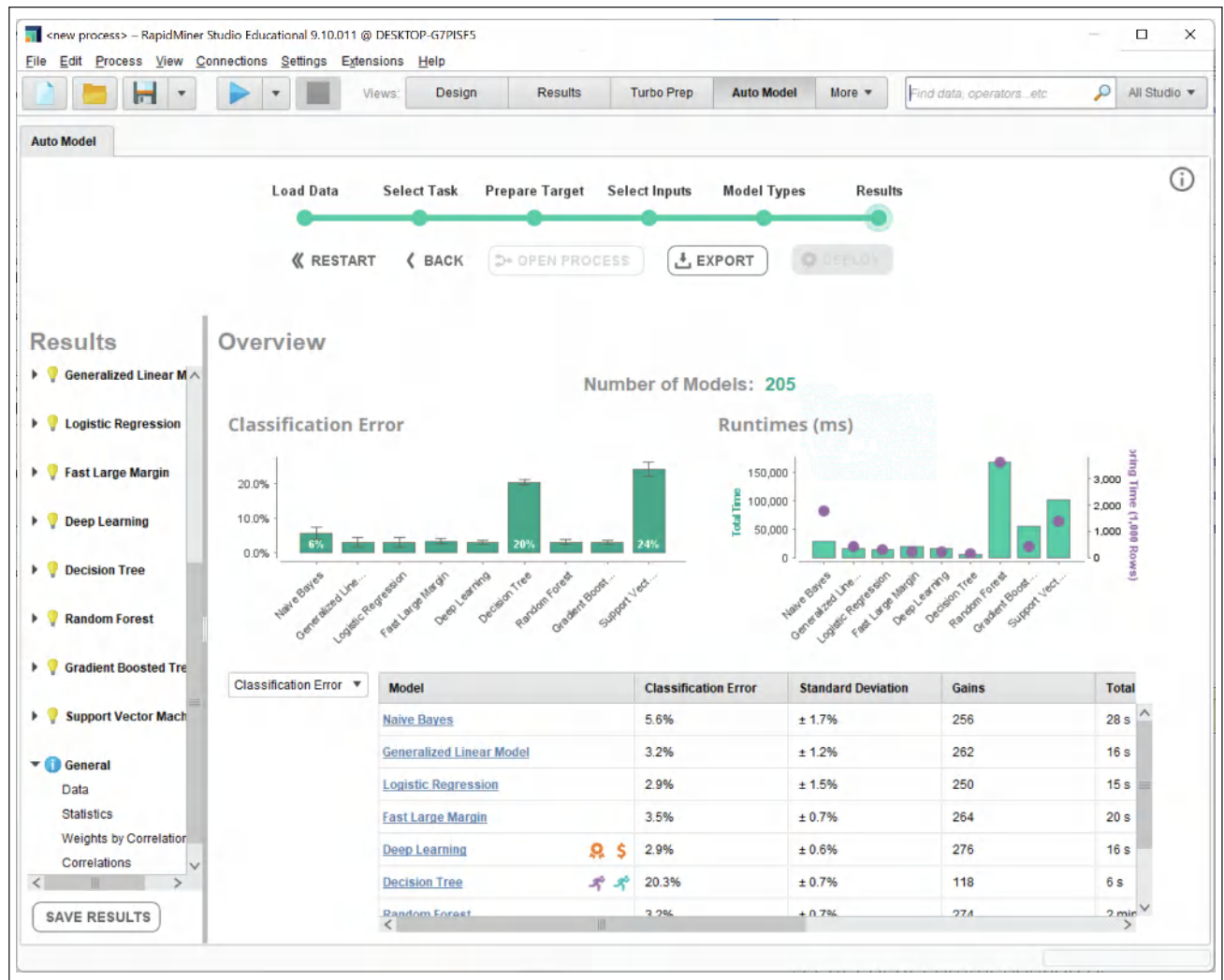


FIGURE 9.43 Auto Model Creating Interface.

a recipe for crowdsourcing and community building. Nowadays, this community-based crowdsourced trends seem to be the driving force for defining the future of analytics and data science. Even the largest commercial players (e.g., Google, Microsoft, Yahoo, and others) in the analytics and data science market are embracing this development and becoming a part of this sharing structure. What follows are brief description of some of the most popular commercial (as opposed to free) and proprietary (as opposed to open source) analytics and data science tools.

Alteryx

Alteryx is a relatively new analytics software company, founded in 1997 as SRC and renamed as Alteryx in 2006. Bases in in Irvine, California with a development center in Broomfield, Colorado, Alteryx went public in 2017.

Alteryx Designer, the main development tool for Alteryx analytics platform, is an easy to use, drag-and-drop user interface that allows quick and easy creation of workflow type analytics solutions to business problems. The software is purposefully

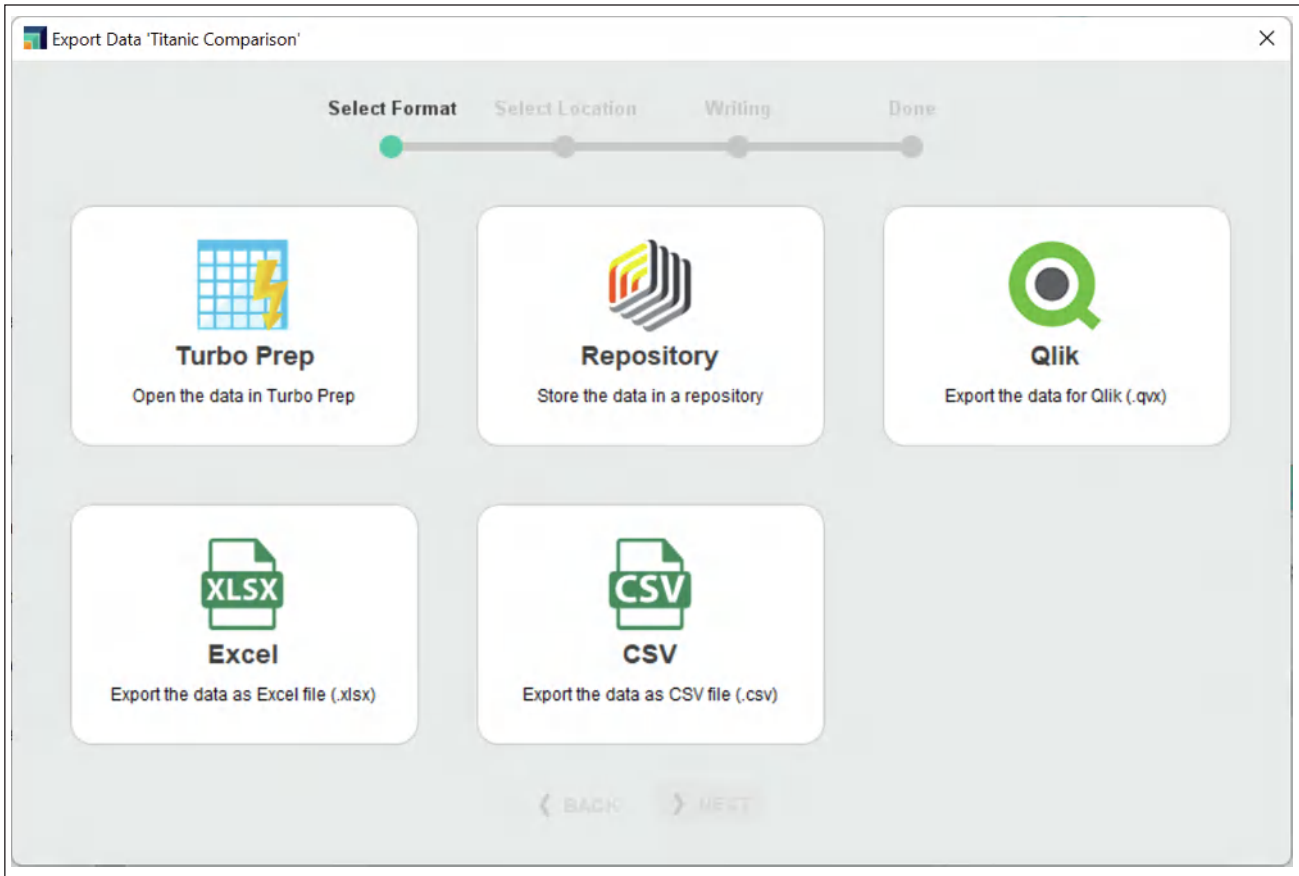


FIGURE 9.44 Export Option Interface.

designed to make analytics and data science easily accessible to any data consumer and decision maker. Like other visual programming tools (SAS Enterprise Miner, IBM SPSS Modeler, KNIME, Orange, etc.), Alteryx provides plenty of native functions to import data from a wide variety of sources; assess, visualize, and preprocess the data; build, test, and validate a wide range of machine learning models; and deploy and report innovative analytic solutions. In addition to its native functionalities/capabilities, Alteryx also provides an open architecture to integrate and utilize other analytics tools and programming languages.

As a rapidly popularized analytics tool, Alteryx is often compared to other established analytics tools like Excel, Tableau, and KNIME. Each of these tools has their advantages and disadvantages as they are compared to Alteryx. Excel is a general-purpose spreadsheet modeling and data manipulation tool (not really a business intelligence and data analytics tool), and Tableau primarily is a data visualization tool. Both Excel and Tableau provide some advanced analytics capabilities via add-ons and functional integrations/scripting, compared to Alteryx, they are not full-fledged business analytics and data science tools.

Perhaps the most appropriate, all-inclusive, head-to-head comparison of Alteryx can be made against KNIME. They are very similar in broad data ingestion and manipulation functionalities, appealing workflow-type drag-and-drop model development process, utilization of other tools and programming languages, and seamless cloud-based deployment options. The main difference are as follows: (1) Alteryx is a commercial and proprietary

tool while KNIME is a free and open-source platform, (2) Alteryx Designer runs only on Windows platform while KNIME Analytics Platform runs natively on Windows, macOS, and Linux, (3) Alteryx provides a very limited-version of their tool to educational institutions (including only the data preprocessing and visualization function, not including any of the key functions like machine learning, text mining, computer vision), while KNIME provides all of its functions without any limitations to educational institutions and commercial users, and (4) Alteryx is designed to have fewer building blocks (for ease of model development by nontechnical users) and hence has a shorter learning curve, while KNIME has thousands of nodes for any low-level data manipulation and model building tasks, making it suitable to more technical users with slightly longer learning curve.

Alteryx also has a cloud-based development options, provided by Trifacta—a privately owned software company headquartered in San Francisco with offices in Bengaluru, Boston, Berlin and London. This cloud offering of Alteryx provides all of analytics capabilities in a cloud-based infrastructure, including ingestion of data, model building and testing, deployment and analysis.

IBM

As one of the most prominent players in the analytics space, IBM offers a variety of analytics tools, some local some cloud based. One of the local tools that IBM office, which is well known in academic world, is IBM SPSS. IBM SPSS includes two primary products: IBM SPSS Statistics, a statistical analysis, data visualization and reporting tool, and IBM SPSS Modeler, a data science and predictive analytics platform with a drag-and-drop GUI and plenty of machine learning capabilities. IBM SPSS Modeler was originally developed by Integral Solutions Ltd, out of Basingstoke, United Kingdom, and named Clementine. After SPSS's acquisition of the product, the name changed to SPSS Clementine, and then to PASW Modeler. Following IBM's 2009 acquisition of SPSS, the product was renamed IBM SPSS Modeler, its current name. A snippet of its workflow is shown in Analytic in Action 5.3 in Chapter 5.

WATSON ANALYTICS IS THE CLOUD-BASED ANALYTICS OFFERING OF IBM Following the sensational entrance to the world of analytics with a win over the two best human competitors at the TV gameshow Jeopardy! IBM Watson has become one of the most popular, cloud-based analytics platforms. Not only it serves to the commercial clients in a wide variety of industries, IBM Watson also offers educational licenses and programs along with teaching curriculum to higher education institutions free of charge (<https://www.ibm.com/academic/home>).

SAS

Like IBM, SAS is one of the largest and highly acclaimed companies in the analytics and data science marketplace, both as a tool vendor and service/solution/consultancy provider. SAS offers a wide variety of analytics tools from basic statistics (that is where they started several decades ago, and named the company SAS—Statistical Analysis Software) to time series forecasting, data mining to advanced machine learning. While some of these tools have been developed as local/desktop software products (e.g., SAS Base, SAS Enterprise Miner, JMP), SAS' recent developments are mostly focusing on cloud-based infrastructures under the name of SAS Viya. Their flagship analytics tool is called SAS Enterprise Miner, which is a visual, workflow type, user-friendly analytics software for development, testing, and deployment of data models. **SAS Viya**, on the other hand, is the server/cloud-based recreation/transformation of SAS' analytics capabilities. Because it is hosted on the cloud, users of SAS Viya can have access to the latest and most efficient analytics capabilities via a browser. A snapshot of SAS Viya interface showcasing a data visualization dashboard is shown in Figure 9.45.

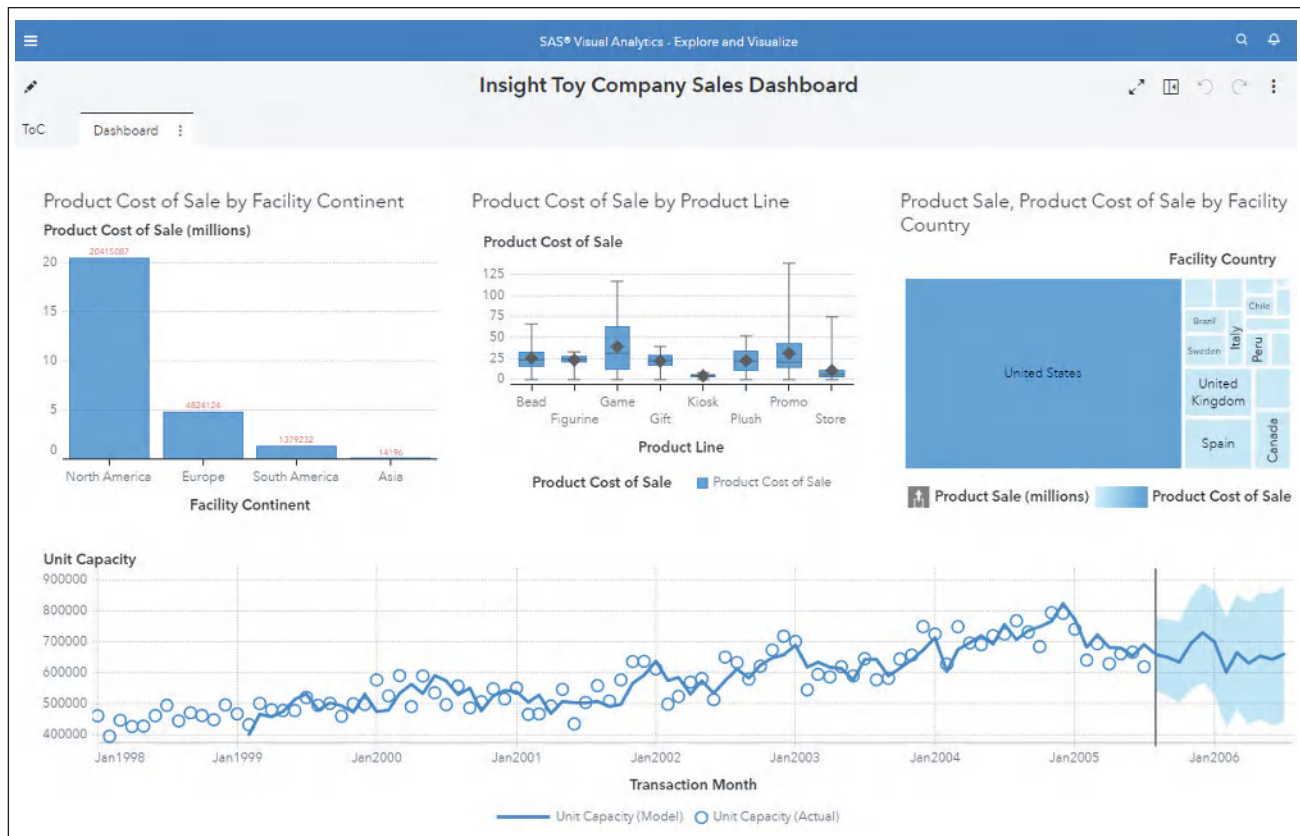


FIGURE 9.45 SAS Viya Screenshot with a Dashboard.

The advantages SAS Viya over the traditional analytics tools include:

- **Execution speed.** In analytics, accuracy is important but speedy access to insights is also critical. With ever-increasing global competition, the ability to make quick, well-informed decisions is critical to success of mere survival of companies. SAS Viya's inherently cloud-based in-memory processing ensures creation of insight generated in near real time, avoiding delays to your essential business decisions.
- **Scalability.** One of the most appealing advantages of cloud-hosted analytics platforms is that they're naturally flexible. SAS Viya can be scaled up and down as your business requirements shrink and grow. Whether that variation occurs over a number of months or over the course of a single day, by scaling your platform and processing requirements you make sure you only pay for what you use. This can offer the potential for real value and efficiencies within your organization.
- **Central management.** SAS Viya provides a centralized base from which all analytics activities managed. SAS Viya's web-based environment ensures analytics teams to have all the tools they will need to manage, explore, process, interact and model large and variant data sources. Such a single base can serve for all business users, no matter what their role is in using the data for their own needs, processing it in a secure, governed environment.
- **Openness to other languages.** SAS Viya provides data scientists with the opportunity to work using their preferred programming languages, including SAS, R, Python, Lua, or Java. This openness ensures analytics managers to obtain maximum value from their teams of data analysts without the need for them all to be trained in SAS.

- **Access to internet data.** Once again, because of its inherently cloud-based presence, SAS Viya can access the native sources of Big Data directly without any delays or transformations. For instance, analysts can have direct access to social media and IoT data in real time for real insight to take advantage of the rapidly changing market conditions and customer needs/wants.
- **Ease of use.** Use of analytics within an internet browser (with no need to install anything locally), via a drag-and-drop interface for model development make it easy to turn big data into big insights. Such a system can empower everyone within the organization to make faster, better decisions by democratizing data and analytics.

SAS Viya is a platform that hosts many tools and functionalities for real-world and real-time business analytic and data science. SAS Viya offering range from visual analytics to statistics, machine learning to data science programming. Figure 9.46 shows a tabular

Capabilities	SAS Visual Analytics	SAS Visual Statistics	SAS Visual Machine Learning	SAS Visual Data Science	SAS Data Science Programming*	SAS Visual Data Science Decisioning	SAS Visual Forecasting	SAS Visual Text Analytics	SAS Intelligent Decisioning
Data access; data preparation, data quality & information cataloging	✓	✓	✓	✓	✓	✓	✓	✓	✓
Advanced flow steps & information governance	^	^	✓	✓		✓			
Visualization & reporting	✓	✓	✓	✓		✓	✓	✓	✓
Conversational AI & chatbots	✓	✓	✓	✓		✓	✓	✓	✓
Statistics		✓	✓	✓	✓	✓			
Matrix programming				✓	✓	✓			
Machine learning & deep learning			✓	✓	✓	✓			
Model deployment & management	^	^	✓	✓		✓			
Forecasting	^	^	^	✓	✓	✓	✓		
Text analytics	^	^	^	✓	✓	✓		✓	
Optimization	^	^	^	✓	✓	✓			
Econometrics	^	^	^	✓	✓	✓			
Digital decisioning	^	^	^	^	^	✓			✓
Event streaming analytics	^	^	^	^	^	✓			
Advanced workload management	^	^	^	^	^	^	^	^	^

* Programmatic interface only; no visual interface.
^ Optional.

FIGURE 9.46 SAS Viya Tools and Their Range of Capabilities.

illustration of the range of SAS Viya tools and their analytics capabilities. The next section provides a description along with a text mining tutorial for JMP Pro, which is a local/desktop analytics tool developed by SAS Institute.

JMP

JMP Pro (an advanced version of JMP) is a commercial software tool developed by SAS Institute (https://www.jmp.com/en_gb/software/predictive-analytics-software.html). JMP Pro, a user-friendly desktop tool, supports both Windows and macOS operating systems. In addition to ample statistical and data mining capabilities, JMP Pro also offers a text mining extension called Text Explorer, which is a comprehensive and highly capable text processing engine. In this section, we will use the textual data that we used in Orange tutorial, conduct very similar text mining procedures, and draw some comparisons between Orange and JMP Pro.

TUTORIAL

Text Mining of Published Literature

Again, as we did with Orange, to streamline the text mining process, we will follow the generic framework presented in Figure 9.32. What follows are the execution of the four steps in this framework using JMP Pro analytics software using the same data set obtained from digital online publication databases, which consist of the articles published in the Journal Information of Systems, up until the end of 2021.

Step 1. Gathering and cleaning the textual data

JMP Pro allows uploading many different types of files. Files can be opened, saved, or exported in the desired format with the relevant tabs under the File menu on the toolbar (see Figure 9.47). “Edit,” “Tables,” “Rows,” and “Cols” tabs provide several data manipulation options on the JMP Pro menu. All analysis types are listed under the “Analyze” tab. The “Graph” tab offers various graph types. Users also can create tailored graphs using “graph builder” under the Graph tab. The Table Panel, located in the upper left corner of the data table, allows users to create scripts and edit them. Column Panel and Row Panel, respectively, under the Table Panel, allow following the Column and Row operations (see Figure 9.47).

In JMP Pro, the texts can be analyzed solo or combined with formulas, for example, title, abstract, and keywords. Users create formulas by the Formula item on the selected column. The “Recode” item in the column menu can regulate text rows. Recode item allows several text manipulation specifications such as grouping, splitting, editing, converting, trimming, or removing text spaces on the text. Duplications between articles can be found and removed by the “Select Duplicate Rows” item under the Row tab in the menu. A new subset can be created by selecting the relevant rows based on the Inclusion & Exclusion Criteria using the “Data filter” under the Row tab.

Step 2. Preprocessing the textual data

In JMP Pro, text mining analysis is named “Text Explorer,” which can be found under the Analyze tab in the main menu. After “Text Explorer” is clicked and opened, the relevant text column should be selected from the “Text Columns” box. The Text Explorer provides analysis in seven languages: English, German, Spanish, French, Italian, Japanese, and Chinese (simplified/traditional). At the same time, some filters may apply to phrases, words, and characters in this menu. In addition to filters, with the “Stemming” option, the texts can be parsed without any stemming (No Stemming), by combining similar stems (Stem for Combining), or by stemming entirely (Stem All Terms). In addition, Regex or Basic Words options are offered in the Tokenizing item. While “Basic Words” only

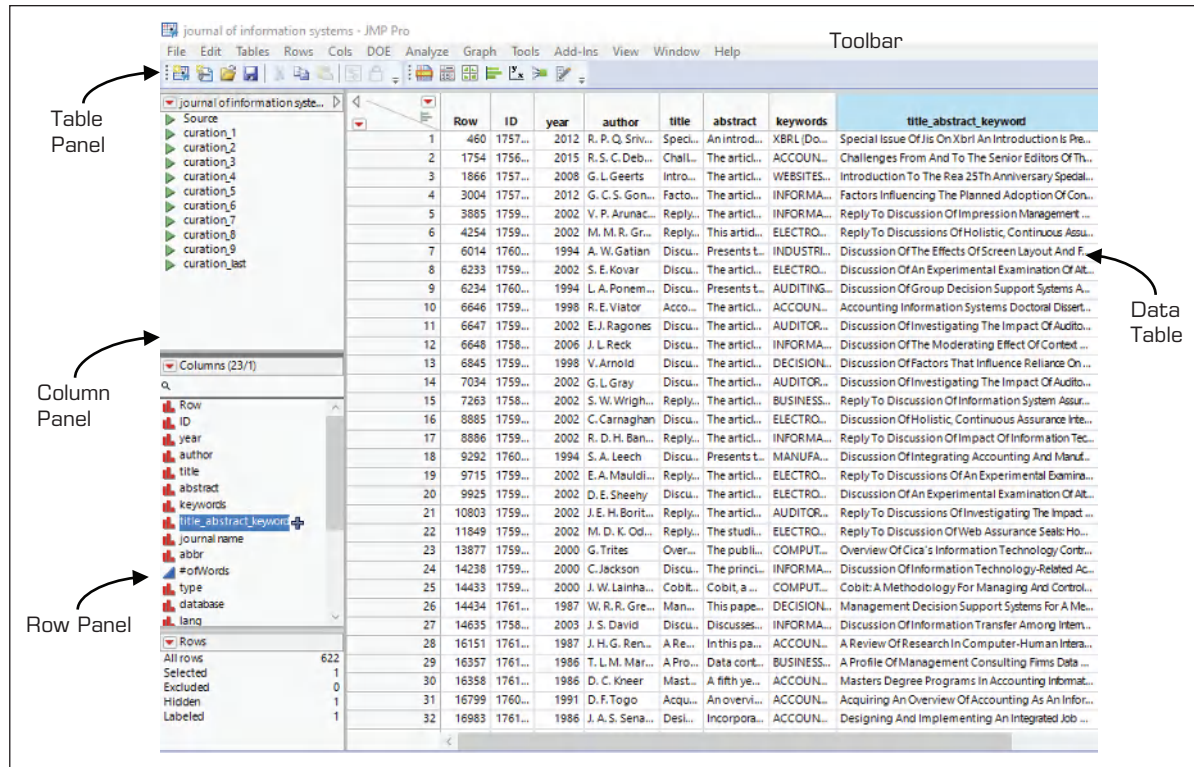


FIGURE 9.47 JMP Pro Main Page View.

tokenizes words, “Regex” provides advanced control in texts. Users can remove unwanted characters from textual data on the regex screen using different types of Regex commands, from words to phrasing punctuation, on sample or current text data (see Figure 9.48).

After the regex process, Text Explorer will show the initial analysis results. Besides reporting some basic statistics, Term and Phrase Lists will be listed. It facilitates users to review the lists manually. The Text Explorer allows for manually examining the phrases and terms to decide which ones to include in the analysis. Thus, phrases considered necessary can be included, and terms considered redundant can be excluded from the model. Using the “Show text” item can be beneficial to see selected words or phrases in the text data. Phrases are added to the term list based on the user’s preference and then examine the term and phrase collectively. What follows are stemming, recoding, and stop-word eliminations to homogenize and simplify the feature space.

All of the preprocessing operations performed on the Text Explorer can be examined, and the operations performed can be reversed or revised undone. Users can manage the preprocessing operations by “Manage Stop Words,” “Manage Recodes,” “Manage Phrases,” and “Manage Stem Exceptions” options under the “Term Options” tab.

After the term list is prepared, the DTM matrix is created, and then, the corpus dimension is reduced with the SVD method. DTM can be obtained with the “Save Document Term Matrix” under the text explorer tab. DTM can be optimized using specific criteria and weights using the “Latent Semantic Analysis, SVD” option on the text explorer tab (see Figure 9.49a). Under this option, “the maximum number of terms” and “minimum term frequency criteria” help to mitigate outlier terms. Weighting methods handle imbalances arising from excessive use of words. Although TF IDF is commonly

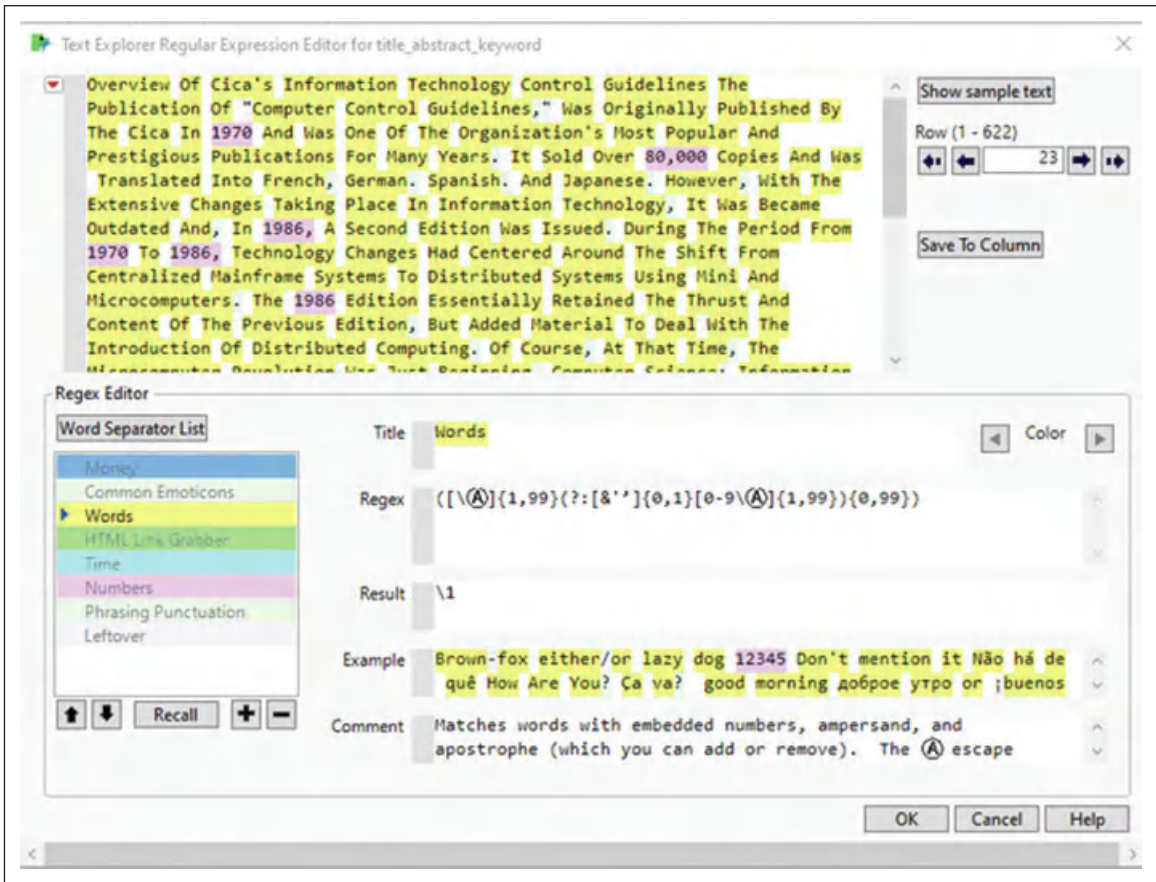


FIGURE 9.48 JMP Pro Text Explorer Regex Menu.

used as a weighting method, JMP Pro also provides Binary, Ternary, Frequency, and Log Freq methods. “Number of Singular Vectors” defines the number of dimensions wanted to reduce. JMP Pro uses 100 dimensions by default. If desired, centering and scaling are normalized in the DTM (see Figure 9.49b).

Step 3. Performing the text analysis

After the DTM is normalized and optimized, text analysis can be performed. JMP Pro offers five different analysis types: Latent Class Analysis (LCA), Latent Semantic Analysis (LSA), Discriminant Analysis (DA), Term Selection (TS), and Sentiment Analysis (SA). Apart from these, Topic Analysis (TA) can also be used, which is the rotated version of LSA.

LCA creates term clusters with cluster mixture probabilities using DTM. LSA does a similar unsupervised clustering by applying weighting and SVD on the DTM. Discriminant analysis, as a supervised method, attempts to predict the membership of each document in a group based on the DTM. Term Selection, another supervised method, determines which terms best describe a specified output variable using generalized regression. Finally, TA is the most popular JMP Pro tool; the VARIMAX method rotated the version of the LSA analysis. TA results will be provided at this stage.

Since TA is an unsupervised method, the user must determine the number of topics to analyze. Several methods can determine the optimum number of topics. The most interpretable method is to plot scree plots of eigenvalues obtained from SVD. The Singular Values table can be transferred to a data table with the “Make into Data Table” tab. Scree

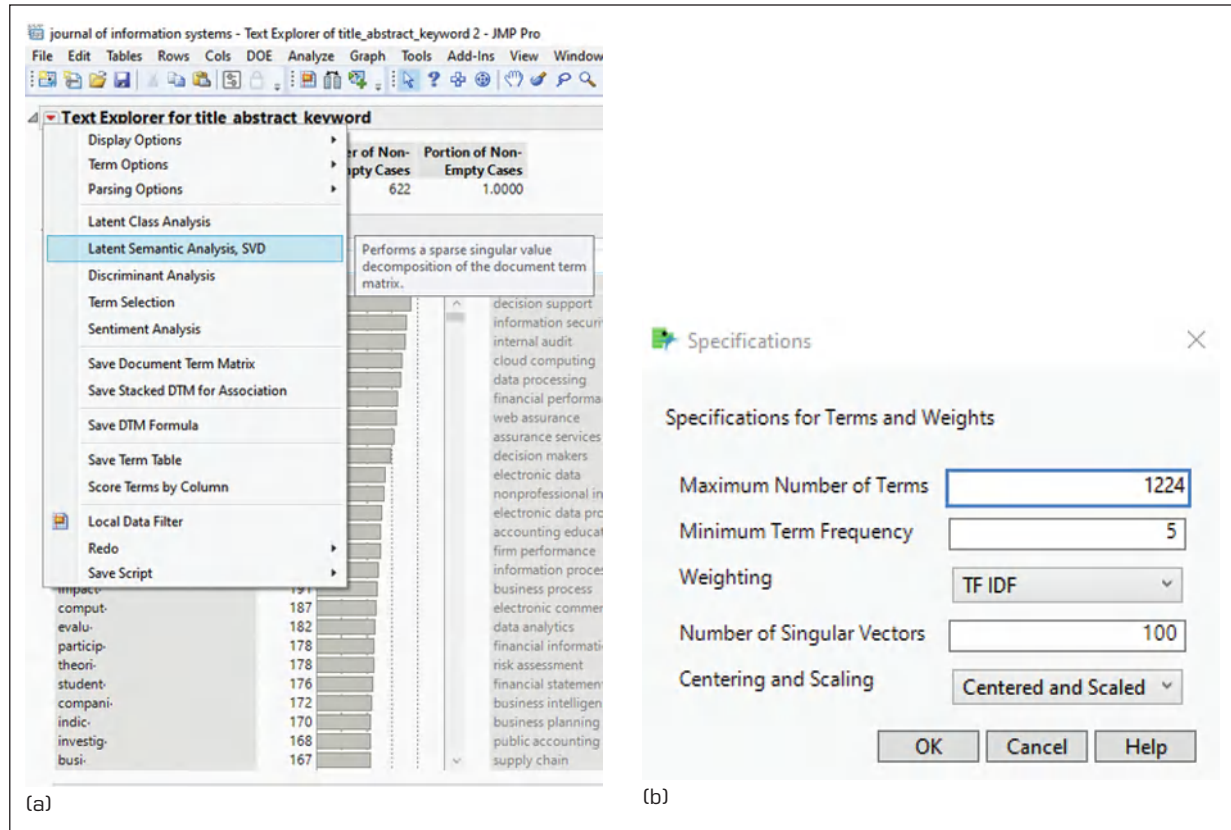


FIGURE 9.49 (a) JMP Pro Text Explorer Creating DTM Pathway. (b) JMP Pro Text Explorer Creating DTM Specifications.

Plot can then be plotted using Graph Builder. The Scree plot is a line graph between the eigenvalues and the number of SVDs. The optimal number of topics can be found at the elbow point where the curve significantly flattens. In this case, The optimal number of topics can be valued between 10 topics and 18 topics (see Figure 9.50a and 9.50b).

At this point, JMP Pro reports the topics and their related terms in a topic-term table. This topic-term table also provides loading values. Loadings are ordered from largest to smallest for each topic. This way, the most relevant word for each topic will be on the top line (see Figure 9.51).

Step 4. Reporting of the obtained results

The topic analysis results can be reported in a variety of ways in JMP Pro. The most common types are topic-term tables, word clouds, and trend graphs. Topic-terms tables can be copied and pasted into any other tool like Word or Excel for further formatting. Word clouds can be quickly plotted in JMP Pro by selecting “Word Clouds by Topic” from the “Display Options” in the topic analysis menu. The terms in the word cloud are sized and bolded based on the term frequency values.

Trend graphs can help monitor the changes in topics over time. The most representative topics can be assigned quickly to each document using the data table screen. After the representative topics are determined, the last action is to plot the topic trends. Users can use a line graph to plot the trends. First, a line graph is drawn by selecting the “TopicID” and “year” columns on the Graph Builder tab. Then, the topic trends can be shown using colored lines (see Figure 9.52).

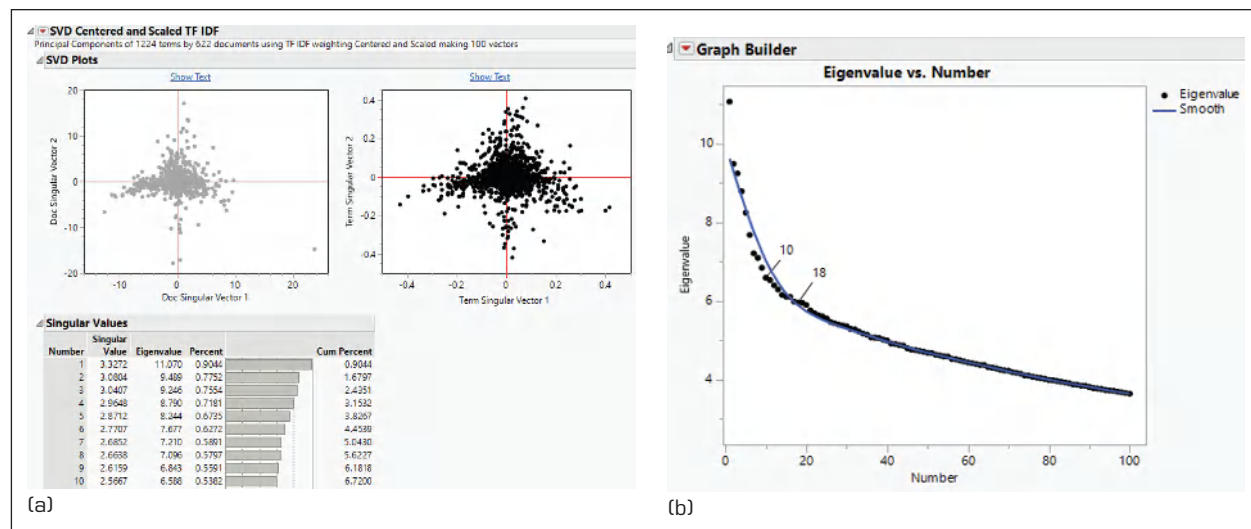


FIGURE 9.50 (a) JMP Pro Text Explorer SVD Results View. (b) JMP Pro Scree Plot.

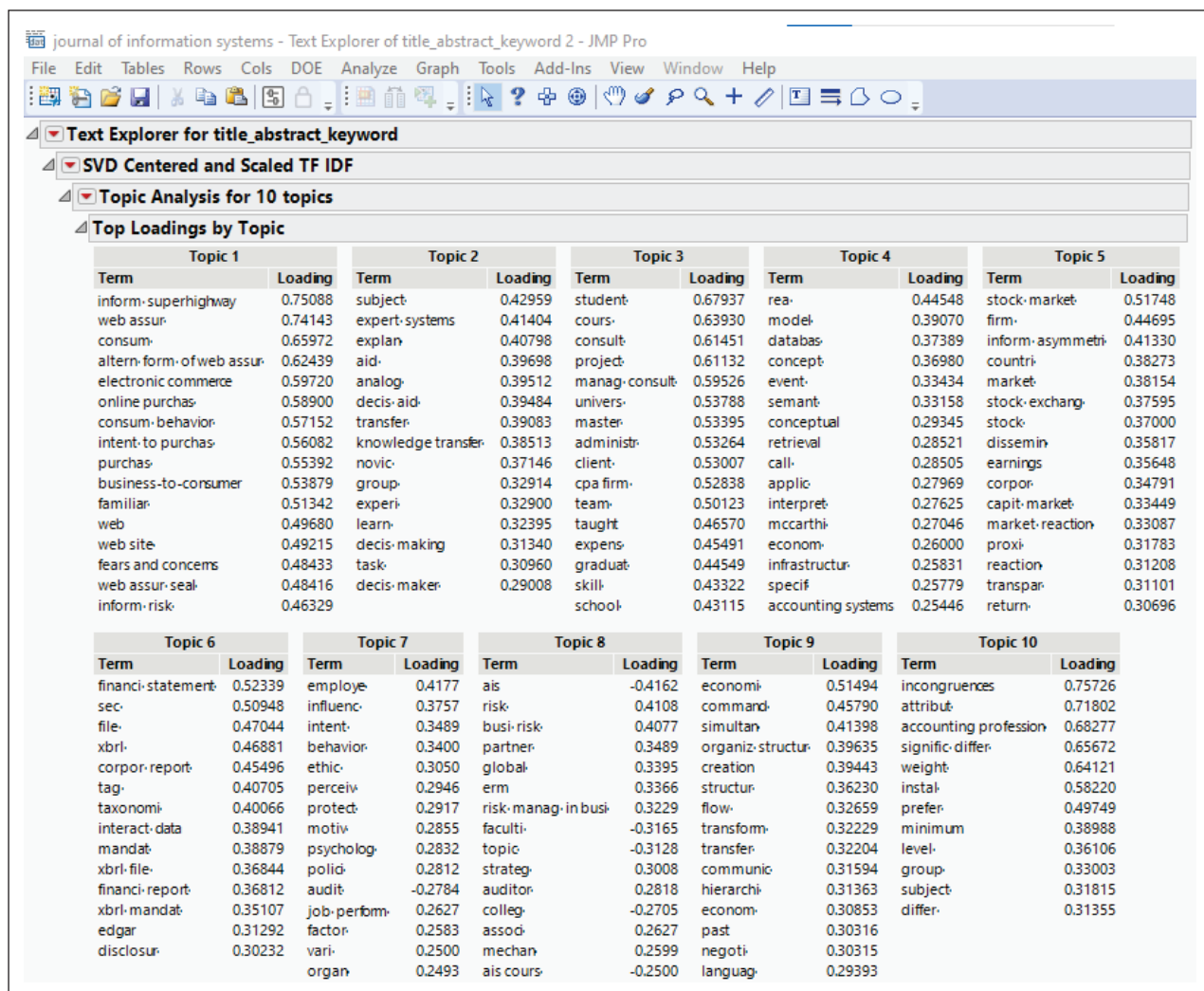


FIGURE 9.51 JMP Pro Text Explorer Topics.

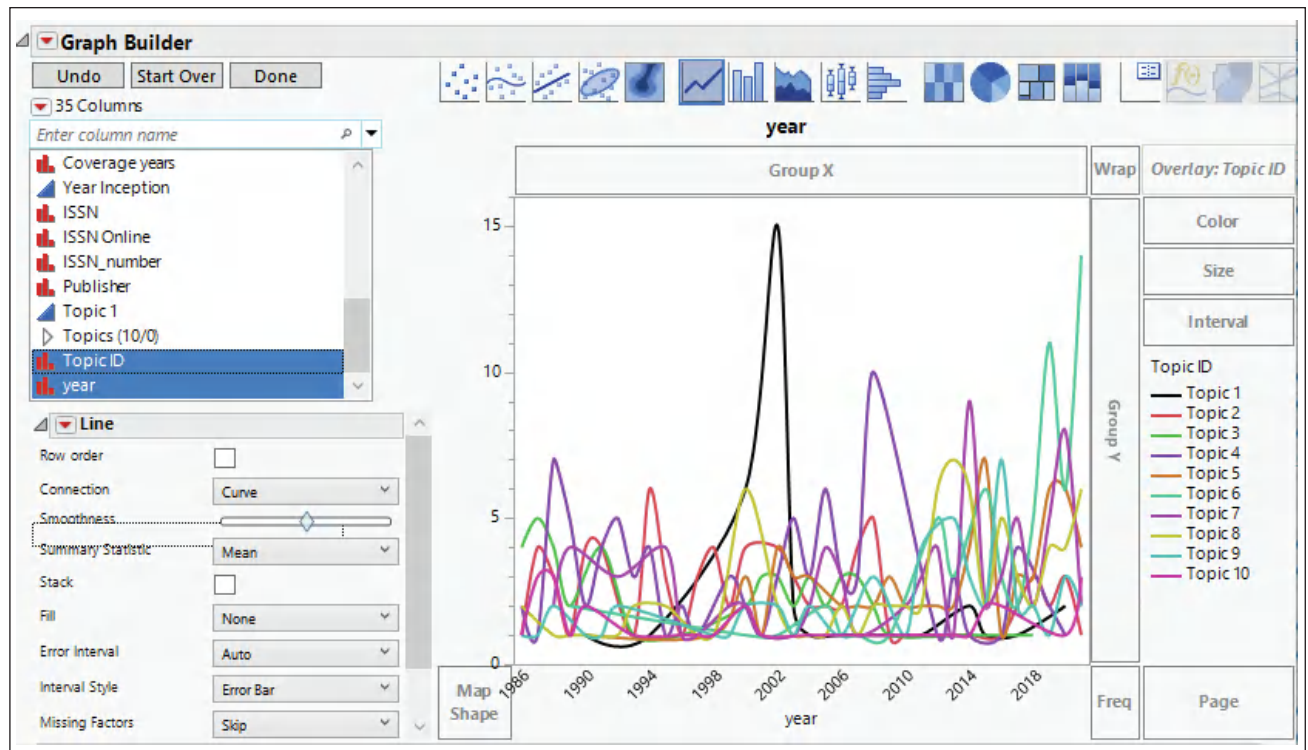


FIGURE 9.52 JMP Pro Trend Graphs Using Graph Builder.

A Comparison of JMP Pro and Orange

Both of these software tools have strengths and weaknesses in terms of performing text mining. The comparison is also organized around the four steps in the text mining framework (Figure 9.50).

- Gathering and cleaning the textual data. Both applications can read various types of files. However, JMP Pro can read broader range of file type. With respect to reading speeds, it can be said that JMP Pro is faster than the Orange. JMP Pro also provides more functional tools for data clean-up, imputation, text correction, and duplication detection.
- Preprocessing the textual data. Many preprocessing functions are offered in both of these software tools. However, Orange preprocessing tools are better organized compared to JMP Pro's. Although Orange offers more preprocessing algorithms, it seems to lack on the capability for editing of terms. Therefore, Monitoring terms and phrases in Orange is not as easily doable. The most crucial feature of JMP Pro is its ease of monitoring and editing terms and phrases in the text. In addition, preprocessing operations can also be trackable in JMP Pro.
- Performing the text analysis. Regarding text analysis, both of these software tools provide various analysis types. However, while JMP Pro provides only one topic modeling algorithm (i.e., LSA) under the text explorer, Orange offers four topic modeling options (i.e., LSA, LDA, HDL, NMF). Although in both of these tools the application models can be designed and developed quickly and easily, JMP Pro performs better in terms of analysis speed.
- Reporting of the obtained results. Finally, both software tools offer a number of options in terms of reporting. Development of Graphs seems to be easier in JMP Pro. Both software tools provide word cloud visualization, but Orange seems to

have more appealing options in terms of data visualization. In both software, trend analysis cannot be performed directly from text mining results, hence, additional design and development is often needed.

Overall, these are only two of many commercial and free/open-source tools that you can obtain and use for conducting text mining projects. Both of these analytics tools are highly capable for conducting all of the main phases in text mining. Based on the project requirements and the specifics of the available data, one tool may be more advantages than the other.

Teradata

Teradata is one of the well-known software companies that provides cloud-based data management and business analytics software, products, and services. Teradata is well known for their data warehousing tools and solution, an IT market they shaped and transformed for a few decades until the start of analytic and data science era. The company was formed in 1979 in Brentwood, California, as a collaboration between researchers at Caltech and Citibank's advanced technology group. Along the way, it was acquired by other companies, including NCR in 1991, only to be separated from NCR in 2007, to become an independent public company.

With acquisitions and internal restructuring, Teradata had rapidly transformed itself into a big data analytics company. Teradata has always been very supportive of the university programs in the era of business intelligence and analytics. They were the primary sponsor and host of Teradata University Network consortium where several analytics companies (e.g., MicroStrategy, Tableau, SAS, Tableau, KXEN), provided their software tools and training materials to educational institutions to teach BI and analytics courses. Two of the coauthors of this book had served on the advisory board of this consortium for a long time, until it was decided to be ended.

In October 2018, Teradata started calling the cloud analytics software product line Vantage. **Teradata Vantage** is an advanced analytics platform embedded with analytic engines and functions, which can be implemented with preferred data science languages (e.g., SQL, Python, R) and tools (e.g., Teradata Studio, Teradata AppCenter, R Studio, Jupyter Notebook) on any data volume of any type by diverse analytics personas (e.g., Data Scientist, Citizen Data Scientist, Business Analyst) across multiple environments (On-Premises, Private Cloud, Public Cloud Marketplaces). There are five important conceptual pieces central to understanding Vantage: Analytics Engines and Functions, Data Storage and Access, Analytic Languages and Tools, Deployment, and Usage. Figure 9.53 illustrates the general architecture of Vantage and its interrelationships with other tools.

Analytic Engines and Functions

An analytic engine is a comprehensive framework that includes all the software components that are well integrated into a container (e.g., Docker) to deliver advanced analytics functionality that can be implemented by a well-defined set of user personas. An analytic engine's components include:

- Advanced Analytics functions
- Access points to data storage that can ingest multiple data types
- Integration into visualization and analytic workflow tools
- Built in management and monitoring tools
- Highly scalable and performant environment with established thresholds

It is advantageous to have an analytic engine as it delivers a containerized compute environment that can be separated from data storage. Furthermore, analytic engines can be tailored for access and use by specific personas (e.g., DS, Business Analyst).

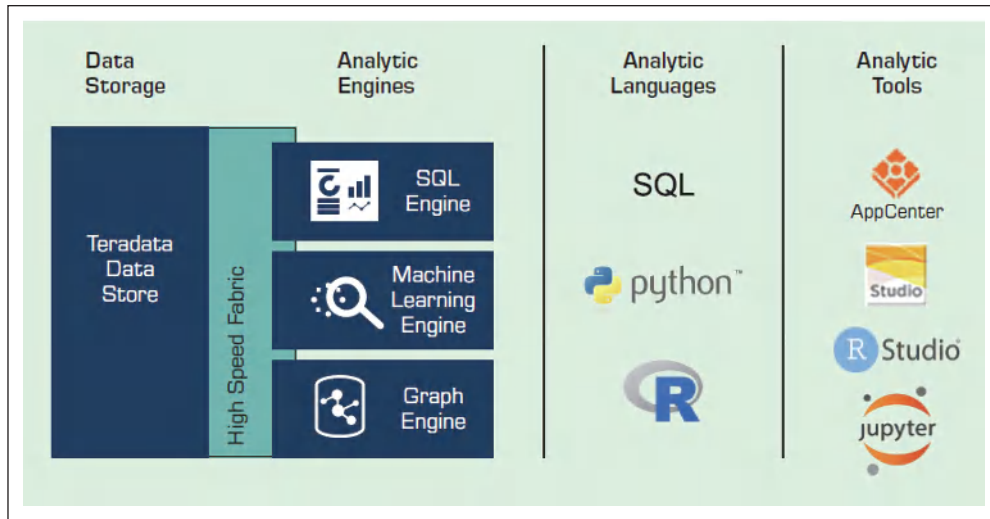


FIGURE 9.53 Teradata Vantage Architecture.

There are three analytic engines in the first release of Vantage. These are NewSQL Engine, Machine Learning Engine, and Graph Engine.

The NewSQL engine includes embedded analytic functions. Teradata will continue to add more functions for the high-speed analytics processing required to operationalize analytics. New functions within the NewSQL engine include:

- nPath
- Sessionization
- Attribution
- Time series
- 4D analytics
- Scoring functions (e.g., Naïve Bayes, GLM, Decision Forests)

The Machine Learning engine delivers more than 120 prebuilt analytic functions for path, pattern, statistical, and text analytics to solve a range of business problems. Functions range from understanding sentiment to predictive part failure analysis.

The Graph engine provides a set of functions that discover relationships between people, products, and processes within a network. Graph analytics solve complex problems such as social network connections, influencer relationships, fraud detection, and threat identification.

Vantage embeds analytic engines close to the data, which eliminates the need to move data, allowing users to run their analytics against larger data sets without sampling and execute models with greater speed and frequency. This is made possible through the use of containers managed by Kubernetes, which allow businesses to easily manage and deploy new cutting-edge analytic engines, such as Spark and TensorFlow, both of which will be available in the near future. Another benefit of containers is the ability to scale out the engines.

From a user's perspective, Vantage is a unified analytic and data framework. Under the covers, it contains a cross-engine orchestration layer that pipelines the right data and analytic request to the right analytic engine across a high-speed data fabric. This enables a business analyst or data scientist, for example, to invoke analytic functions from different engines in a single application, such as Jupyter Notebook, without enduring the trouble of hopping from one analytic server or application to another. The result is a tightly integrated analytic implementation that's not restrained by functional or data silos.

- *Data Storage and Access:* Teradata Vantage comes with a natively embedded Teradata MPP Database. Furthermore, a high-speed data fabric (Teradata QueryGrid™ and Presto™) connects the platform to external data sources that include third-party enterprise data warehouses (e.g., Oracle), open source data platforms (e.g., Hadoop), no-SQL databases (e.g., Cassandra), and others. Data support ranges from relational, spatial, and temporal to XML, JSON, Avro, and time-series formats.
- *Analytic Languages and Tools:* Teradata Vantage was built out of the recognition that analytics professionals such as Data Scientists and Business Analysts require a diverse set of languages and tools to process large data volumes to deliver analytic insights. Vantage includes languages such as SQL, R, and Python on which analytics functions can be executed through Teradata Studio, R Studio, and Jupyter Notebooks.
- *Deployment:* Vantage platform provides the same analytic processing across deployment options, including the Teradata Cloud and public cloud, as well as on-premises installations on Teradata hardware or commodity hardware. It is also available as a service.
- *Usage:* Teradata Vantage is intended to be used by multiple analytic personas. The ease of SQL ensures that citizen data scientists and business analysts can implement prebuilt analytic functions integrated into the analytic engines. The ability to invoke Teradata-supported packages such as dplyr and teradataml ensures that Data Scientists familiar with R and Python can execute analytic packages through R Studio and Jupyter notebooks, respectively, on the platform. Users who are not proficient at executing programs can invoke analytic functions codified into Apps built into Teradata AppCenter, an app building framework available in Vantage, to deliver compelling visualizations such as Sankey, Tree, Sigma diagrams, or word clouds.
- *Example Usage:* A global retailer had a website that suboptimally delivered search results to potential buyers. With online purchases accounting for 25% of total sales, inaccurate search results negatively impacted the customer experience and the bottom line. The retailer implemented Teradata machine learning algorithms, available in Teradata Vantage, to accumulate, parse, and classify search terms and phrases. The algorithms delivered the answers needed to identify search results that closely matched online customer needs. This led to more than \$1.3 million in incremental revenue from high-value customers, as measured by purchase volumes, over a two-month holiday period.

Analytics in Action 9.3 illustrates another application of Teradata Vantage where its advanced network analytics capabilities were deployed to analyze data from a large electronic medical records data warehouse.

ANALYTICS IN ACTION 9.3 Analyzing Disease Patterns from an Electronic Medical Records Data Warehouse

The Center for Health Systems Innovation at Oklahoma State University has been given a massive data warehouse by Cerner Corporation, a major electronic medical records (EMRs) provider, to help develop analytic applications. The data warehouse contains EMRs on the visits of more than 50 million unique patients across US hospitals (2000–2015). It is the largest and the industry's only relational database that includes comprehensive records with pharmacy, laboratory, clinical events, admissions, and billing data. The database also includes more than 2.4 billion laboratory results and more than 295 million orders for nearly 4,500 drugs by name and brand. It is one of the largest compilations of deidentified, real-world, HIPAA-compliant data of its type.

The EMRs can be used to develop multiple analytics applications. One application is to understand the relationships between diseases based on the information about the simultaneous diseases developed in the patients. When multiple diseases are present in a patient, the condition is called comorbidity. The comorbidities can be different across population groups. In an application (Kalgotra, Sharda, & Croff, 2017), the authors studied health disparities in terms of comorbidities by gender.

To compare the comorbidities, a network analysis approach was applied. A network is comprised of a defined set of items called nodes, which are linked to each other through edges. An edge represents a defined relationship between the nodes. A very common example of network is a friendship network in which individuals are connected to each other if they are friends. Other common networks are computer networks, Web page networks, road networks, and airport networks. To compare the comorbidities, networks of the diagnoses developed by men and women were created. The information about the diseases developed by each patient in the lifetime history was used to create a comorbidity network. For the analysis, 12 million female patients and 9.9 million male patients were used. To manage such a huge data set, Teradata Aster Big Data platform was used. To extract and prepare the network data, SQL, SQL-MR, and SQL-GR frameworks supported by Aster were utilized. To visualize the networks, Aster AppCenter and Gephi were used.

Figure 9.54 presents the female and male comorbidity networks. In these networks, nodes represent different diseases classified as the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), aggregated at the three-digit level. Two diseases are linked based on the similarity calculated using Salton Cosine Index. The larger the size of a node, the greater the comorbidity of that disease. The female comorbidity network is denser than the male network. The number of nodes and edges in the female network are 899 and 14,810, respectively, whereas the number of nodes and edges in the male network are 839 and 12,498, respectively. The visualizations present a difference between the pattern of diseases developed in male and female patients. Specifically, females have more comorbidities of mental disorders than males. On the other hand, the strength of some disease associations between lipid metabolism and chronic heart disorders is stronger in males than females. Such health disparities present questions for biological, behavioral, clinical, and policy research.

The traditional database systems would be taxed in efficiently processing such a huge data set. The Teradata Aster made the analysis of data containing information on millions of records fairly fast and easy. Network analysis is often suggested as one method to analyze big data sets. It helps understand the data in one picture. In this application, the comorbidity network explains the relationship between diseases at one place.

Source: Kalgotra, P., Sharda, R., & Croff, J. M. (2017). Examining health disparities by gender: A multi-morbidity network analysis of electronic medical record. *International Journal of Medical Informatics*, 108, 22–28.

TIBCO

TIBCO Software Inc. is known to be a business intelligence software company founded in 1997 in Palo Alto, California. Its early offering of a data visualization tool, Spotfire, is well known in business intelligence marketplace since 1990s. Evolution of business intelligence into business analytics and data science forced TIBCO to go beyond data visualization and to become more like a full-service business analytics and data science tool and service provider. This thinking led to TIBCO acquiring Statistica, a comprehensive predictive analytics tool developed and marketed by StatSoft, and integrating it into its TIBCO Data Science platform.

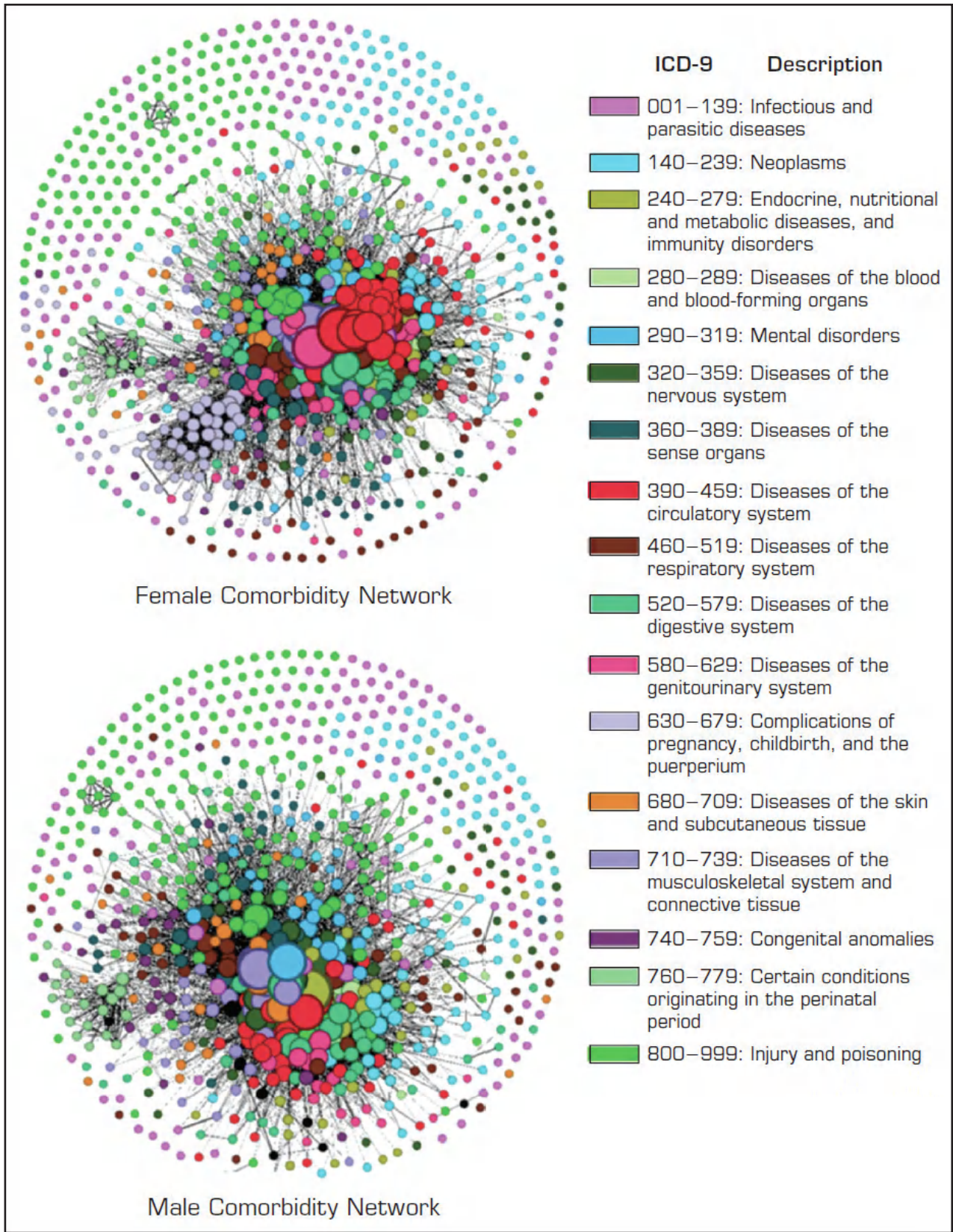


FIGURE 9.54 Female and Male Comorbidity Networks.

Other Analytics Tools

Because of its overwhelming popularity in practice and in academia, analytics is being used rather broadly to define any data driven decision analysis. Commercial tool providers are using this popular label to enhance the marketing appeal of their offerings. Although most every tool in the market claim to cover all three phases in analytics continuum (i.e., descriptive, predictive, and prescriptive), some are strong in data acquisition and preparation, some are superior on the latest automated machine learning modeling procedures, and some are very good at seamless deployment and integration. In this last section, because of the space limitation for the chapter, we categorize, list, and briefly define these analytics tools that excel in certain parts of the analytics continuum.

BUSINESS INTELLIGENCE AND DATA VISUALIZATION TOOLS Some of the tools in the analytics market are originally designed and offered as data visualization tools in the midst of the business intelligence era. Although, most of them possess some predictive and prescriptive capabilities, they are commonly known as business intelligence and data visualization software platforms. The tools that fall into this category are led by Tableau, arguably the most data visualization tool in the market with the largest market share. Tableau, a successful startup created by PhD candidate Chris Stolte under Professor Pat Hanrahan at Stanford University, is recently acquired by Salesforce in 2019 for price of \$15.7 billion. PowerBI, another hugely popular data visualization tool, is Microsoft's direct competition to Tableau. Both of these tools provide local/desktop tools for creation of visuals and dashboards, and server for deployment and utilization of those visuals. PowerBI provides the desktop tool for free while Tableau does not. Qlik (<https://www.qlik.com/>) and Spotfire (<https://www.tibco.com/products/tibco-spotfire>) are the other two well-established data visualization tools that have significant penetration in certain industries and application domains. MicroStrategy (<https://www.microstrategy.com/>) is yet another data visualization and business intelligence tools that has been around for a long time. Another noteworthy effort was to build data visualization and descriptive analytics tools that work within Microsoft Excel. In addition to Microsoft's offerings like Pivot Charts, Power Query, and Power Pivot, among the many third-party add-on offerors, the one that survived the time is XLMiner, now owned by FrontlineSolver (<https://www.solver.com/xlminer>).

Some of the successful data visualization tools are acquired by the large software and consultancy companies during the upward era of business intelligence, for instance, IBM acquired Cognos, SAP acquired Business Objects, and Oracle acquired Hyperion. Other noteworthy companies in the data visualization space include Lumira (another ingrown dataviz tool by SAP), Google Charts and Google Data Studio, Dundas, Zoho Analytics, and Grafana (a free and open-source data visualization tool).

SOFTWARE AS A SERVICE PROVIDERS Big data requires ample computational resources. Creating a big data capable computational environment locally has proven to be inefficient at best and infeasible in most cases. Instead, many companies are turning to **cloud-based software, software-as-a-service (SaaS)**, business models to perform real-world big data analytics. In this space, the frontrunner of the top-ten is Amazon's AWS, followed by Microsoft's Azure, Google Cloud, Alibaba Cloud, IBM Cloud, Salesforce, Oracle Cloud, SAP, Snowflake Cloud, and VMware Cloud.

BIG DATA HANDLERS Big data comes with big promises and big challenges. Handling big data is challenging, not only because of the sheer size/volume of the data but also the variety and velocity that the big data brings alone. Although people in analytics like big data, because of its potential value-add, proper handling of it requires a significantly

different computational infrastructure. Recognizing the need, and the potential size of such a market, several new companies have emerged in this space as “big data analytics” service and tool/platform providers. Some of the noteworthy ones include Databricks, Datameer, Dataiku, Datarobot, and Domo.

► SECTION 9.5 REVIEW QUESTIONS

1. What are the advantages of using commercial and proprietary tools. Do your own search to go beyond what is mentioned in this section.
2. What are the commonalities and differences between IBM and SAS analytics offerings?
3. Why the market in commercial and proprietary analytics tools is so diverse?
4. If you were to be given the opportunity to choose and use any tool from the ones described in this section, which one would you choose? Why?

Chapter Highlights

- Business analytics has been the name of the game quite some time in both businesses and academic circles.
- To keep up with this fast growing/evolving interest in business analytics and data science universities are rushing to create new degree programs while businesses are creatively coming up with new departments and roles/titles to take advantage of the momentum.
- Business analytics and data science, myriad of software tools and computing platforms are being developed (by startups as well as the well-established IT companies) and serviced to satisfy the needs and wants of the variety of analytics communities.
- A wide variety of analytics software tools exist in the market, ranging from commercial/paid to free/open-source, from graphical workflow-type to language-based programming platforms.
- There is no one tool that is the best of every aspect of data analytics.
- As evidenced from the continually expending landscape of tool offerings, there is no one tool that is best at everything.
- As opposed to one tool, what is more practical is to possess a tool kit with several tools.
- The success of data scientists at work depends largely on the tools (or perhaps more accurately, the toolbox) they rely on.
- Analytics tools can be classified as free (of charge) versus paid/commercial.
- Analytics tools can be classified as open-source versus closed-source/proprietary.
- Analytics tools can be classified as local versus server/cloud based.
- Analytics tools can be classified as visual versus programmatic.
- Programming-based analytics tools include R and Python.
- Visual programming tools include KNIME, Orange, Alteryx, IBM SPSS Modeler, and SAS Enterprise Miner.
- R is one of the most popular analytics tools despite the fact that is commonly known for statistical computation and stunning graphics.
- RStudio is the most popular IDE for developing R programming-based analytics solutions.
- Anaconda is an integrates navigation tool for downloading and installing R and Python IDEs.
- There are several IDEs for Python coding, but the one that leads the pack is Jupyter Notebook.
- Google Colab is a cloud-based platforms developed to make team-based Python programming easy.
- KNIME is a free and open-source analytics platform for learning, teaching, and practicing data science.
- Orange is a free and open-source data mining tools with a visually appealing user interface.
- Weka (Waikato Environment for Knowledge Analysis), developed at the University of Waikato, New Zealand, is a free and open-source software for data mining.
- RapidMiner, formerly known as YALE (yet another learning environment), is a comprehensive data science platform with visual workflow design capabilities.

- Alteryx Designer, the main development tool for Alteryx analytics platform, is an easy to use, drag-and-drop user interface that allows quick and easy creation of workflow type analytics solutions to business problems.
- IBM is one of the most prominent players in the analytics market space, offering a variety of analytics tools, some local (e.g., IBM SPSS Modeler, Cognos) some cloud based (e.g., **Watson Analytics**).
- SAS is one of the largest and highly acclaimed companies in the analytics and data science marketplace, both as a tool vendor and service/solution/consultancy provider.
- SAS offers a wide variety of local (e.g., SAS Enterprise Miner, JMP) and cloud-based (SAS Viya) analytics tools from basic statistics to time series forecasting, data mining to advanced machine learning.
- JMP Pro is a user-friendly local/desktop analytics tool that supports both Windows and macOS operating systems.
- Teradata is one of the well-known software companies that provides cloud-based data management and business analytics software, products, and services.
- Teradata Vantage is an advanced analytics platform embedded with analytic engines and functions, which can be implemented with preferred data science languages (e.g., SQL, Python, R) and tools.
- Statistica is a comprehensive predictive analytics tool integrated into TIBCO Data Science platform.
- There are several analytic tools that fall under the Business Intelligence and Data Visualization Tools category, which includes Tableau, PowerBI, Qlik, Spotfire, and MicroStrategy.
- The top-ten cloud-based analytics platform providers include Amazon's AWS, Microsoft's Azure, Google Cloud, Alibaba Cloud, IBM Cloud, Salesforce, Oracle Cloud, SAP, Snowflake Cloud, and VMware Cloud
- The newly emerged, well-known big data analytics companies include Databricks, Datameer, Dataiku, Datarobot, and Domo.
- Cognitive computing makes a new class of problems computable by addressing highly complex situations that are characterized by ambiguity and uncertainty; in other words, it handles the kinds of problems that are thought to be solvable by human ingenuity and creativity.
- Cognitive computing finds and synthesizes data from various information sources and weighs the context and conflicting evidence inherent in the data in order to provide the best possible answers to a given question or problem.
- The key attributes of cognitive computing include adaptability, interactivity, being iterative, stateful, and contextual.
- *Cognitive analytics* is a term that refers to cognitive computing-branded technology platforms, such as IBM Watson, that specialize in the processing and analysis of large unstructured data sets.
- Cognitive search is the new generation of search method that uses AI (advanced indexing, NLP, and machine learning) to return results that are much more relevant to the user than traditional search methods.
- IBM Watson is perhaps the smartest computer system built to date. It has coined and popularized the term *cognitive computing*.
- IBM Watson beat the best of men (the two most winning competitors) at the quiz game *Jeopardy!*, showcasing the ability of computers to do tasks that are designed for human intelligence.
- Watson and systems like it are now in use in many application areas including healthcare, finance, security, and retail.

Key Terms

Anaconda
AWS
Clementine
Closed-source software
Cloud-based software
Forrester
Forrester Wave
Gartner
Google Colab

IBM SPSS Modeler
JMP Pro
Jupyter Notebook
KDnuggets
KNIME Analytics Platform
KNIME Hub
KXEN
LDA
Magic Quadrant

Miniconda
 Open-source software
 Orange
 PowerBI
 Proprietary software
 PyCharm
 Python
 R
 RapidMiner
 Rattle
 RStudio
 SAS Enterprise Miner

SAS Viya
 Software-as-a-service (SaaS)
 Spyder
 Tableau
 Teradata Vantage
 Text Mining
 TIBCO
 Visual programming
 Visual Studio Code
 Watson Analytics
 Weka

Questions for Discussion

1. Why is it so critical to choose the right tool for analytics projects?
2. What are the reasons for the existence of such a large and varied analytics tools landscape?
3. What are the four common dimensions we can use to compare analytics tools?
4. Which analytics tool is your most favorite? Why?
5. What are the advantages of using a programming language like Python or R for analytics?
6. What are the commonalities and differences between Python and R for analytics?
7. What are the most popular machine learning libraries in R used for analytics projects?
8. What are the most popular machine learning libraries in Python used for analytics projects?
9. What are the advantages of using visual programming tools like the ones covered in in this section?
10. Which of the visual programming software tools you like the most? Why?
11. Do you know any other visual, free, open-source analytics tools? If not, you can just search for them and see what you can find out. Since the tools landscape is highly dynamic and still expanding, there is a good chance that you can find some interesting ones.
12. What are the advantages of using commercial and proprietary tools. Do your own search to go beyond what is mentioned in this section.
13. What are the commonalities and differences between IBM and SAS analytics offerings?
14. Why the market in commercial and proprietary analytics tools is so diverse?
15. If you were to be given the opportunity to choose and use any tool from the ones described in this section, which one would you choose? Why?

Exercises

Hands-On and Internet Exercises

1. Download and install R programming language and RStudio IDE using Anaconda installation (<https://www.anaconda.com/>), or any other installation methods of your choosing. Identify a classification type data set from UCI (<https://archive.ics.uci.edu/ml/index.php>). Develop a model using at least two classification methods, such as decision trees and logistic regression. Compare the results, and write a 5-page report summarizing your findings and experiences.
2. Download and install Python programming language and Jupyter Notebook IDE using Anaconda installation (<https://www.anaconda.com/>), or any other installation methods of your choosing. Identify a classification type data set from UCI (<https://archive.ics.uci.edu/ml/index.php>). Develop a model using at least two classification methods, such as decision trees and logistic regression. Compare the results, and write a 5-page report summarizing your findings and experiences.
3. Download and install both R and Python (<https://www.anaconda.com/>). Identify a regression type data set from UCI (<https://archive.ics.uci.edu/ml/index.php>). Develop two different types of machine learning models (i.e., linear regression and support vector machines) with both of these analytics tools. Compare the results, and write a 5-page report summarizing your findings and experiences.
4. Download and install KNIME Analytics Platform (<https://www.knime.com/>). Identify a classification type data set from UCI (<https://archive.ics.uci.edu/ml/index.php>). Develop a model using at least two classification methods, such as decision trees and logistic regression.

Compare the results, and write a 5-page report summarizing your findings and experiences.

5. Download and install Orange (<https://orangedatamining.com/>). Identify a classification type data set from UCI (<https://archive.ics.uci.edu/ml/index.php>). Develop a model using at least two classification methods, such as decision trees and logistic regression. Compare the results, and write a 5-page report summarizing your findings and experiences.
6. Download and install Weka (<https://www.cs.waikato.ac.nz/ml/weka/>). Identify a classification type data set from UCI (<https://archive.ics.uci.edu/ml/index.php>). Develop a model using at least two classification methods, such as decision trees and logistic regression. Compare the results, and write a 5-page report summarizing your findings and experiences.
7. Download and install both Python (<https://www.anaconda.com/>) and KNIME Analytics Platform (<https://www.knime.com/>). Identify a large classification type data set from Kaggle (<https://www.kaggle.com/datasets>). Develop two different types of machine learning models (i.e., decision trees and artificial neural networks) with both of these analytics tools. Compare the results, and write a 5-page report summarizing your findings and experiences.
8. Download and install both R (<https://www.anaconda.com/>) and Orange (<https://orangedatamining.com/>). Identify a large regression type data set from Kaggle (<https://www.kaggle.com/datasets>). Develop two

different types of machine learning models (i.e., artificial neural networks and support vector machines) with both of these analytics tools. Compare the results, and write a 5-page report summarizing your findings and experiences.

9. Go to JMP.com. Download a 30-day evaluation copy of JMP. Identify a large classification type data set from Kaggle (<https://www.kaggle.com/datasets>). Develop two different types of machine learning models (i.e., artificial neural networks and support vector machines). Write a 5-page report summarizing your findings and experiences.
10. Go to Alteryx.com. Download a 30-day evaluation copy of Alteryx Designer. Identify a large classification type data set from Kaggle (<https://www.kaggle.com/datasets>). Develop two different types of machine learning models (i.e., artificial neural networks and support vector machines). Write a 5-page report summarizing your findings and experiences.
11. Go to TIBCO download site (<https://www.tibco.com/resources/product-download/tibco-statistica-trial-download-for-windows>), download and install a 30-day evaluation copy of Statistica. Identify a large classification type data set from Kaggle (<https://www.kaggle.com/datasets>). Develop two different types of machine learning models (i.e., artificial neural networks and support vector machines). Write a 5-page report summarizing your findings and experiences.

References

- Berthold, M. R., Borgelt, C., Höppner, F., Klawonn, F. and Silipo, R., 2020, "Guide to Intelligent Data Science: How to Intelligently Make Use of Real Data," New York: Springer International Publishing.
- Delen, D. (2021, August). Better Practices for Teaching Business Analytics Taxonomy for academic institutions' program and course offerings. *INFORMS OR/MS Today*, August 2021, pp. 38–43.
- Delen, D. (2020). *Predictive Analytics: Data Mining, Machine Learning and Data Science for Practitioners*. Upper Saddle River, NJ, USA: FT Press (A Pearson Company).
- Delen, D. (2021, October). KNIME Analytics Platform: Open-source business analytics and data science tool provides comprehensive capabilities in the classroom. *INFORMS OR/MS Today*, October 2021, pp. 22–23.
- Delen, D., Helfrich, S., & Silipo, R. (2021, March). KNIME Analytics Platform for Visual Data Science and Business Analytics Teaching. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (pp. 1373–1373).
- Silipo, R., (2020). Why KNIME? Medium, July 14, <https://medium.com/swlh/why-knime-98c835afc186>.
- Kalgotra, P., Sharda, R., & Croff, J. M. (2017). Examining health disparities by gender: A multimorbidity network analysis of electronic medical record. *International Journal of Medical Informatics*, 108, 22–28.

AI-Based Trends in Analytics and Data Science

LEARNING OBJECTIVES

- Describe the current and future use of cloud computing in business analytics
- Describe how geospatial and location-based analytics are assisting organizations
- Describe image analytics and use of images as an alternative source of data
- Describe the IoT and its characteristics
- Discuss the benefits and drivers of IoT
- Describe typical IoT applications in a diversity of fields
- Describe smart appliances and homes
- Understand the concept of smart cities, their content, and their benefits
- Understand the potential of using network properties in analytics
- Describe other emerging applications in AI: **genomics**, health, etc.

Data Science and AI are fast growing. Virtually every day, one can read about new applications that are emerging in different disciplines, industries, and locations. In our attempt to excite you to learn more about these technologies and their potential, this chapter covers additional areas where there is a lot of new activity happening. We hope that you will review these sections as a foundational step but can read about other similar applications in your own interest areas. This chapter includes the following sections:

- 10.1** Application Vignette: Discover Foods Employs IoT and Machine Learning to Ensure Food Quality 575
- 10.2** Cloud-Based Analytics 576
- 10.3** Location-Based Analytics 589
- 10.4** Image Analytics/Alternative Data 596
- 10.5** IoT Essentials 599
- 10.6** IoT Applications 608
- 10.7** 5G Technologies and Impact on AI 614
- 10.8** Other Emerging AI Topics: Robotic Process Automation (RPA) 616
- 10.9** Bioinformatics and Health Network Science 618
- 10.10** Other Recent Developments 622

10.1 APPLICATION VIGNETTE: Discover Foods Employs IoT and Machine Learning to Ensure Food Quality

Discover Foods is one of market leaders in the food processing industry. For Discover Foods, it is very important to predict the temperature of the food for maintaining its safety. A significant factor in assessing the food quality based on temperature is the nature of the food itself. Consider a food that must be maintained at a subzero temperature for optimal quality. The quality of food gets affected when the temperature rises and falls several times as in a typical temperature time series graph. Any time interval the item spends above its optimal temperature, it can impact the food's quality and safety. Now consider a hot food item, soup. Each time interval it spends below its piping hot optimal temperature, its quality negatively gets affected. Recent advances in Remote Sensing, **Cloud Computing** and **Machine Learning** hold the potential to revolutionize Food Safety. Now more than ever, it is possible to automate the maintenance of temperature, microorganism growth, humidity and other parameters pertaining to the upkeep of food standards in real time. The implementation of **Artificial Intelligence** in Food Safety should yield greater food quality, lower operational costs, and minimal losses from the spoilage of edibles. Discover Foods has come up with an algorithm that learns from historic temperature fluctuations in each food sample and predicts the same for a specified time interval which prevent or at least minimize losses due to depreciation of food quality.

Continuous temperature readings are gathered using IoT sensors for a total duration of three months. Since the temperature sensor readings are continuous, the 1-minute averages of said temperature readings along with their respective timestamps are gathered as individual data points and saved in .csv format. In the restaurant's pantry multiple sensors have been placed in different food samples. Each food sample is isolated from the other; however, they are all influenced by room temperature. The data has been converted into two columns, namely, a temperature column with respective date-time column and then temperature. These are normalized. The outliers were left untreated because the rise and fall in temperature are genuine occurrences due to the opening and closing of pantry doors which could cause the room temperature to briefly change.

Discover Foods have used statistical models such as auto-regressive integrated moving average (ARIMA), seasonal auto-regressive integrated moving average (SARIMA), and machine learning models like linear regression, tree-based models, support vector machines, and neural network models. ARIMA model optimal parameters ($p=1$, $q=0$, $d=2$) are obtained using auto-correlation function (ACF) and partial auto-correlation function. For the ARIMA model, mean absolute percentage error (MAPE) for fitted values was 0.87 and mean squared error (MSE) for fitted values was 0.38. Similarly for SARIMA model these parameters stood at MAPE: 0.726 and MSE: 776. Detailed performance of the optimal ARIMA performance is provided in the below table:

Forecasting in Future: Minutes	Test MSE	Test RMSE	Test MAPE
1	0.38	0.62	1.42602622
5	0.14	0.37	0.669931299
10	0.89	0.95	1.992238442
15	18.72	4.33	9.438480259
30	11.35	3.37	6.717362951
60	3.8	1.95	3.389628978
90	3.82	1.95	4.624790902
120	2.42	1.56	3.338980725
150	2.39	1.54	3.408343084
180	3.65	1.91	4.464382412

CLASSICAL MACHINE LEARNING AND DEEP LEARNING MODELS

As seen in Table 10.1, the performance of linear regression is on par with a neural network model LSTM; however, LSTM is better at predicting both outliers and central values simultaneously. This is important because the preservation of food quality relies on being able to predict suboptimal outlier temperatures that are too hot or cold. Hence **Discover Foods** algorithm of choice for solving this problem is LSTM. Table 10.2 exhibits LSTM model's performance at a longer time window—600 minutes. These models are computationally intensive during model building. It would be beneficial to explore several hyper parameter values and model architectures that can train the models over a distributed system.

As can be observed from Tables 10.1 and 10.2, the machine learning and **deep learning** models supersede the benchmark ARIMA model in terms of both prediction accuracy and time complexity. Of the given algorithms SVM takes an unusually long time during the training phase. The reason why SVM takes very long time to train is that Kernelized SVMs require the computation of a distance function between each point in the data set, which is the dominating cost of $O(n \text{ features} \times n^2 \text{ observations})$. Though CART models perform well they are likely to overfit significantly, despite pruning. Thus, they were discarded after noting that they would not generalize well in real time.

This AI platform was deployed in real time to validate the model performance. The temperature predictions were able to minimize the losses from the spoilage of edibles. Right tail outlier prediction is of higher priority than overall accuracy. Since the model is intended to predict and warn restaurant staff when the food quality is likely to get compromised (i.e., when stored food warms up), extra emphasis must be given to high temperature Outlier Prediction.

DISCUSSIONS FOR QUESTIONS

1. Why is the temperature check critical in food industry?
2. Review some of these algorithm names from previous chapters and discuss why LSTM might do better, or CART might overfit the data.
3. Can you think of other sensor applications along these lines?

WHAT CAN WE LEARN FROM THIS VIGNETTE?

By combining data from sensors and machine learning algorithms, we can derive significant value from the data being collected to make more optimal decisions. One of the many emerging trends in Data Science/AI, Internet of Things is becoming a big industry. As we will see in this chapter, IoT has many diverse applications, with many more yet to be developed. You will also learn about other recent and emerging trends in AI/data science techniques and applications.

Contributed by: Prajwal Shetty, Dhinakar ABN, D. Narayana and Anwesh Reddy of Great Learning, India.

10.2 CLOUD-BASED ANALYTICS

Another emerging technology trend that business analytics users should be aware of is cloud computing. The National Institute of Standards and Technology (NIST) defines **cloud computing** as “a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, and services) that can be rapidly provisioned and released with minimal management effort or service-provider interaction.” Wikipedia (Cloud Computing, n.d.) defines cloud computing as “a style of computing in which dynamically scalable and often virtualized resources are provided over the Internet. Users need not have knowledge of, experience in, or

TABLE 10.1 Comparison of Different Classical ML Models and LSTM Model Predictions for 60 Min, 120 Min, and 180 Min Time Period

Window Size	Model	Train MAE	Test MAE	Train MSE	Test MSE	Train RMSE	Test RMSE	Train R2 Score	Test R2 Score
60	LR	0.0562	0.0494	0.0143	0.0088	0.1195	0.0936	0.9878	0.9853
60	KNN	0.0475	0.1103	0.0119	0.0289	0.1091	0.1701	0.9898	0.9516
60	SVM	0.0515	0.0485	0.0125	0.0082	0.1118	0.0908	0.9893	0.9862
60	DT	0.047	0.0554	0.0103	0.0098	0.1016	0.0992	0.9912	0.9835
60	RF	0.045	0.0536	0.0098	0.0093	0.0988	0.0962	0.9917	0.9845
60	LSTM	0.0585	0.0575	0.0139	0.0093	0.1178	0.0964	0.9881	0.9845
60	RNN	0.0549	0.0523	0.015	0.0095	0.1224	0.0974	0.9872	0.9841
60	GRU	0.0563	0.0519	0.0134	0.0084	0.1159	0.0919	0.9885	0.9859
120	LR	0.0562	0.0495	0.0143	0.0088	0.1194	0.0936	0.9878	0.9853
120	KNN	0.0482	0.1451	0.0115	0.0489	0.107	0.2211	0.9902	0.9182
120	SVM	0.0527	0.0504	0.0128	0.0089	0.1132	0.0942	0.989	0.9852
120	DT	0.0472	0.0559	0.0103	0.0111	0.1012	0.1056	0.9912	0.9813
120	RF	0.0445	0.0533	0.0093	0.0091	0.0964	0.0954	0.992	0.9848
120	LSTM	0.0545	0.0482	0.0138	0.0085	0.1174	0.0921	0.9882	0.9858
120	RNN	0.0667	0.061	0.0157	0.0101	0.1253	0.1006	0.9865	0.9831
120	GRU	0.0534	0.0511	0.0135	0.0089	0.1161	0.0941	0.9884	0.9852
180	LR	0.0562	0.0495	0.0142	0.0088	0.1193	0.0936	0.9878	0.9853
180	KNN	0.047	0.1693	0.0107	0.067	0.1034	0.2588	0.9908	0.8879
180	SVM	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
180	DT	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
180	RF	0.0445	0.0533	0.0092	0.0091	0.096	0.0955	0.9921	0.9847
180	LSTM	0.0579	0.0523	0.014	0.0088	0.1182	0.0938	0.988	0.9853
180	RNN	0.0557	0.0578	0.0145	0.0098	0.1204	0.0988	0.9875	0.9837
180	GRU	0.056	0.0521	0.0139	0.0087	0.1178	0.0934	0.9881	0.9854

TABLE 10.2 LSTM Model Actual vs. Predicted for Both Train and Test for 600 Min Time

Window Size	Model Name	Train MAE	Test MAE	Train MSE	Test MSE	Train RMSE	Test RMSE	Train R2 Score	Test R2 Score
600	LSTM	0.0538	0.0508	0.0133	0.0087	0.1152	0.093	0.9885	0.9885

control over the technology infrastructures in the cloud that supports them.” This definition is broad and comprehensive. In some ways, cloud computing is a new name for many previous, related trends: utility computing, application service provider grid computing, on-demand computing, software as a service (SaaS), and even older, centralized computing with dumb terminals. But the term *cloud computing* originates from a reference to the Internet as a “cloud” and represents an evolution of all the previously shared/centralized computing trends. The Wikipedia entry also recognizes that cloud computing is a combination of several IT (information technology) components as services. For example, *infrastructure as a service* (IaaS) refers to providing computing *platforms as a service* (PaaS), as well as all the basic platform provisioning, such as management administration, security, and so on. It also includes SaaS, which includes applications to be delivered through a Web browser, whereas the data and the application programs are on some other server.

Although we do not typically look at Web-based e-mail as a prime example of cloud computing, it can be considered a basic cloud application. Typically, the e-mail application stores the data (e-mail messages) and the software (e-mail programs that let us process and manage e-mails). The e-mail provider also supplies the hardware/software and the entire basic infrastructure. If the Internet is available, one can access the e-mail application from anywhere in the cloud. When the application is updated by the e-mail provider (e.g., when Gmail updates its e-mail application), it becomes available to all customers. Social networking Web sites like Facebook, Twitter, and LinkedIn are also examples of cloud computing. Thus, any Web-based application is in a way an example of a cloud application. Another example of a general cloud application is Google Docs and Spreadsheets. This application allows a user to create text documents or spreadsheets that are stored on Google’s servers and are available to the users anywhere they have access to the Internet. Again, no programs need to be installed as “the application is in the cloud.” The storage space is also “in the cloud.” Even Microsoft’s popular office applications (all available as Office 365) are all available in the cloud, with the user not needing to download any software. In general, as companies move to selling subscriptions to their software/services rather than sell the software, the migration to cloud is becoming more of a norm. Thus, this section’s goal is just to introduce you to the terminology of the cloud and of course, illustrate how analytics/data science capabilities are available through the cloud.

A good general business example of cloud computing is Amazon.com’s Web services. Amazon.com has developed an impressive technology infrastructure for e-commerce as well as for BI, customer relationship management, and supply-chain management. It has built major data centers to manage its own operations. Through Amazon.com’s cloud services, many other companies can employ these very same facilities to gain advantages of these technologies without having to make a similar investment. Like other cloud-computing services, a user can subscribe to any of the facilities on a pay-as-you-go basis. This model of letting someone else own the hardware and software but making use of the facilities on a pay-per-use basis is the cornerstone of cloud computing. Several companies offer cloud-computing services, including Salesforce.com, IBM Cloud, Microsoft Azure, Google, Adobe, and many others.

Cloud computing, like many other IT trends, has resulted in new offerings in analytics. These options permit an organization to scale up its data warehouse and pay only for what it uses. The end user of a cloud-based analytics service may use one organization for analysis applications that, in turn, uses another firm for the platform or infrastructure. The next several paragraphs summarize the latest trends in the interface of cloud computing and BI/business analytics. A few of these statements are adapted from an early paper written by Haluk Demirkan and one of the coauthors of this book (Demirkan & Delen, 2013).

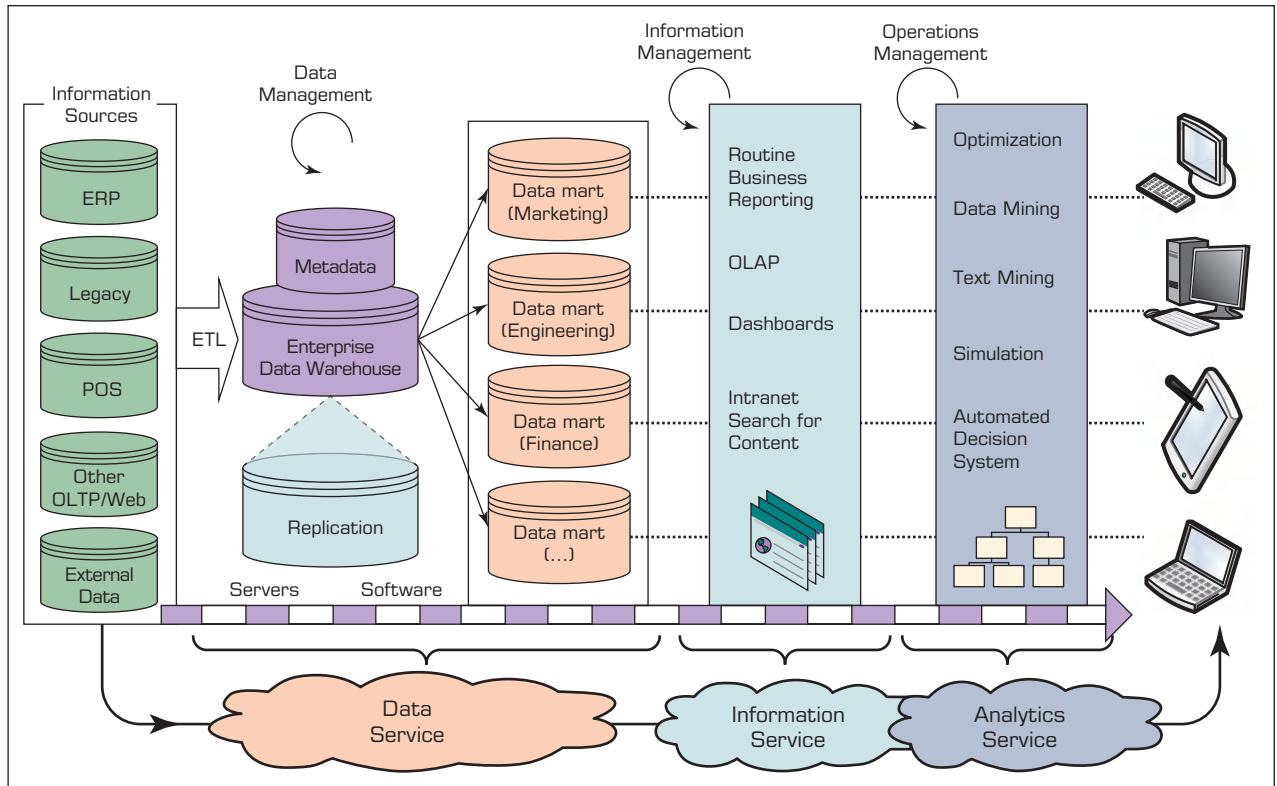


FIGURE 10.1 Conceptual Architecture of a Cloud-Oriented Support System. *Source:* Based on Demirkan, H., & Delen, D. (2013, April). Leveraging the Capabilities of Service-Oriented Decision Support Systems: Putting Analytics and Big Data in Cloud. *Decision Support Systems*, 55(1), 412–421.

Figure 10.1 illustrates a conceptual architecture of a service-oriented decision support environment, that is, a cloud-based analytics system. This figure superimposes the cloud-based services on the general analytics architecture presented in previous chapters.

In service-oriented decision support solutions, (1) operational systems, (2) data warehouses, (3) online analytic processing, and (4) end-user components can be obtained individually or bundled and provided to the users as service. Any or all these services can be obtained through the cloud. Because the field of cloud computing is fast evolving and growing at a rapid pace, there is much confusion about the terminology being used by various vendors and users. The labels vary from Infrastructure, Platform, Software, Data, Information, and **Analytics as a Service**. In the following, we define these services. Then we summarize the current technology platforms and highlight applications of each through Analytics in Action.

Data as a Service (DaaS)

The concept of data as a service advocates the view that “where data lives”—the actual platform on which the data resides—does not matter. Data can reside in a local computer or in a server at a server farm inside a cloud-computing environment. With DaaS, any business process can access data wherever it resides. Data as a service began with the notion that data quality could happen in a centralized place, cleansing and enriching data and offering it to different systems, applications, or users, irrespective of where they were in the organization, computers, or on the network. This has now been replaced with master data management and customer data integration solutions,

where the record of the customer (or product, or asset, etc.) may reside anywhere and is available as a service to any application that has the services allowing access to it. By applying a standard set of transformations to the various sources of data (for example, storing date/time information in one format but making it available in a native format needed by an application) and then enabling applications to access the data via open standards such as SQL, XQuery, and XML, service requestors can access the data regardless of vendor or system.

With DaaS, customers can quickly move thanks to the simplicity of the data access and the fact that they do not need extensive knowledge of the underlying data. If customers require a slightly different data structure or have location-specific requirements, the implementation is easy because the changes are minimal (agility). Second, providers can build the base with the data experts and outsource the analysis or presentation layers (which allows for very cost-effective user interfaces and makes change requests at the presentation layer much more feasible), and access to the data is controlled through the data services. It tends to improve data quality because there is a single point for updates.

Desktop as a Service (DaaS)

In Desktop as a Service, as the name suggests, service providers offer virtual desktop with all the required applications for the users to perform their tasks without having to worry about the software, hardware, and security of the desktop. It is also referred as PC as a Service (PCaaS). This service enables the users to connect and work remotely from anywhere around the world. Amazon WorkSpaces (<https://aws.amazon.com/workspaces/>) (accessed August 2022) is a DaaS service provided by Amazon Web Services (AWS). We can build our own virtual desktop as per our need with the available compute, memory, and storage options. It is an on-demand service with pay-as-you-go pricing. We can scale up/down (elasticity) the resources as per our requirements and thus only pay for the resources we use. Examples of companies using Amazon Workspaces service are Grubhub, Fox, Endemol, and Maximus. There are many other third-party service providers offering DaaS like services. For example, Dell (<https://www.dell.com/en-us/dt/services/pc-as-a-service.htm>) (accessed August 2022) claims to offer a predictable and more flexible monthly pricing.

Software as a Service (SaaS)

This model allows consumers to use applications and software that run on distant computers in the cloud infrastructure. Consumers need not worry about managing underlying cloud infrastructure and must pay for the use of software only. All we need is a Web browser or an app on a mobile device to connect to the cloud. Gmail is an example of SaaS. Many analytics providers are moving to this model. For example, SAS is available as SAS Viya.

Platform as a Service (PaaS)

Using this model, companies can deploy their software and applications in the cloud so that their customers can use them. Companies do not have to manage resources needed to manage their applications in cloud-like networks, servers, storage, or operating systems. This reduces the cost of maintaining underlying infrastructure for running their software and saves time for setting up this infrastructure. Now, users can focus on their business rather than focusing on managing infrastructure for running their software. Examples of PaaS are Microsoft Azure, Amazon EC2, and Google App Engine.

Infrastructure as a Service (IaaS)

In this model, infrastructure resources like networks, storage, servers, and other computing resources are provided to client companies. Clients can run their application and have administrative rights to use these resources but do not manage the underlying

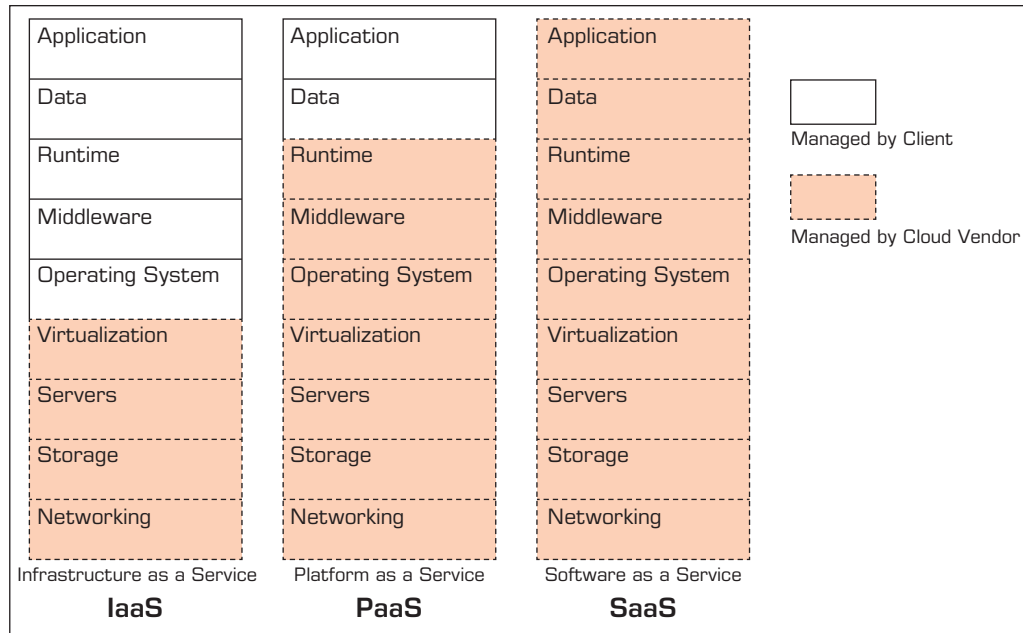


FIGURE 10.2 Technology Stack as a Service for Different Types of Cloud Offerings.

infrastructure. Clients must pay for usage of infrastructure. A good example of this is Amazon.com's Web services. Amazon.com has developed an impressive technological infrastructure that includes data centers. Other companies can use Amazon.com's cloud services on a pay-per-use-basis without having to make similar investments. Similar services are offered by all major cloud providers such as IBM, Microsoft, Google, and so on.

We should note that there is considerable confusion and overlap in the use of cloud terminology. For example, some vendors also add information as a service (IaaS), which is an extension of DaaS. This IaaS is different from infrastructure as a service described earlier. Our goal here is to just recognize that there are varying degrees of services that an organization can subscribe to manage their IT infrastructure in general and the analytics applications in particular. Figure 10.2 highlights the level of service subscriptions a client uses in each of the three major types of cloud offerings. SaaS is clearly the highest level of cloud service that a client may get. For example, in using Office 365, an organization is using the software as a service. The client is only responsible for bringing in the data. Many of the analytics as a service application fall in this category as well. Furthermore, several analytics as a service provider may in turn use clouds such as Amazon's AWS or Microsoft Azure to provide their services to the end users. We will see examples of such services shortly.

Essential Technologies for Cloud Computing

VIRTUALIZATION **Virtualization** is the creation of a virtual version of something like an operating system or server. A simple example of virtualization is the logical division of a hard drive to create two separate hard drives in a computer. Virtualization can be in all three areas of computing:

- **Network virtualization:** It is the splitting of available bandwidth into channels, which disguises complexity of the network by dividing it into manageable parts. Then each bandwidth can be allocated to a particular server or device in real time.
- **Storage virtualization:** It is the pooling of physical storage from multiple network storage devices into a single storage device that can be managed from a central console.

- **Server virtualization:** It is the masking of physical servers from server users. Users do not have to manage the actual servers or understand complicated details of server resources.

This difference in the level of virtualization relates to which cloud service one employs.

Analytics in Action 10.1 illustrates an application of cloud technologies that enable a mobile application and allow for significant reduction in information miscommunication.

ANALYTICS IN ACTION 10.1 Major West Coast Utility Uses Cloud-Mobile Technology to Provide Real-Time Incident Reporting

Historical communication between utilities and first responders has been by phone calls or two-way radios. Some of these are with first responders on the scene, and some with dispatch or other units of the first responder organization. When a member of the public sees an incident on the field, they usually just call 911, which is routed to first responders. Dispatch centers route the closest first responder to the field, who then call back to the center either on their radios or cell phones to let them know the actual status. The dispatch centers then call the incident in to the appropriate utility, who then send their own team to the field for further resolution. This also requires that the exact location be conveyed to the dispatch center from the field, and from the former to utility—particularly challenging if the incident location is not at a specific address (e.g., along a freeway, across open land). The utility also needs to let the dispatch center know the status of their own crew. This information must also be relayed to the first responders on the field. Much of this process relies on information being communicated orally and then forwarded to one or more recipients, with information also flowing back and forth along the same chain. All of this can result in garbled communication and/or incomplete messages, which can eat away precious minutes or even hours in emergencies.

A major West Coast Utility, a leader in using technology to address traditional problems, determined that many of these challenges can be addressed through better information sharing in a timelier manner using cloud-mobile technology. Their territory encompassed densely populated cities to far-flung rural communities with intervening miles of desert, national parks, and more.

Recognizing that most first responders have a smartphone or tablet, the utility selected Connixt's iMarq™ mobile suite to provide a simple-to-use mobile app that allows first responders to advise the utility of any incident in the field. The technology also keeps the first responders apprised of the utility's response status with respect to the incident.

With a targeted base of over 20,000 first responders spread across the entire territory, lowering barriers to adoption was an especially key factor. "Improving communication with groups that are outside your organization is historically difficult," says G. Satish, cofounder, and CEO, Connixt. "For this deployment, the focus on simplicity is the key to its success."

First responders are invited to download and self-register the app, and once the utility grants access rights, they can report incidents using their own tablets or smartphones. The first responder simply uses a drop-down menu to pick from a list of preconfigured incidents, taps an option to indicate if they will wait at the scene, and attach photographs with annotations—all with a few touches on their device. The utility receives notification of the incident, reviews the time and geo stamped information (no more mixed-up addresses), and updates their response. This response (which may be a truck roll) is sent to the first responders and maintained in the app.

The simplicity of the solution makes it easy for the first responders. They use their own phone or tablet, communicate in a way they are used to, and provide needed information simply and effectively. They can see the utility updates (such as the status of the truck that was sent). Missed or garbled phone messages are minimized. Options such as recording voice memos, using speech-to-text and more, are also available.

Cloud technology has been particularly useful in this case—deployment is faster without issues related to hardware procurement, installation, and appropriate backups. Connixt's cloud-based **Mobile Extension Framework (MXFTM)** is architected for rapid configuration

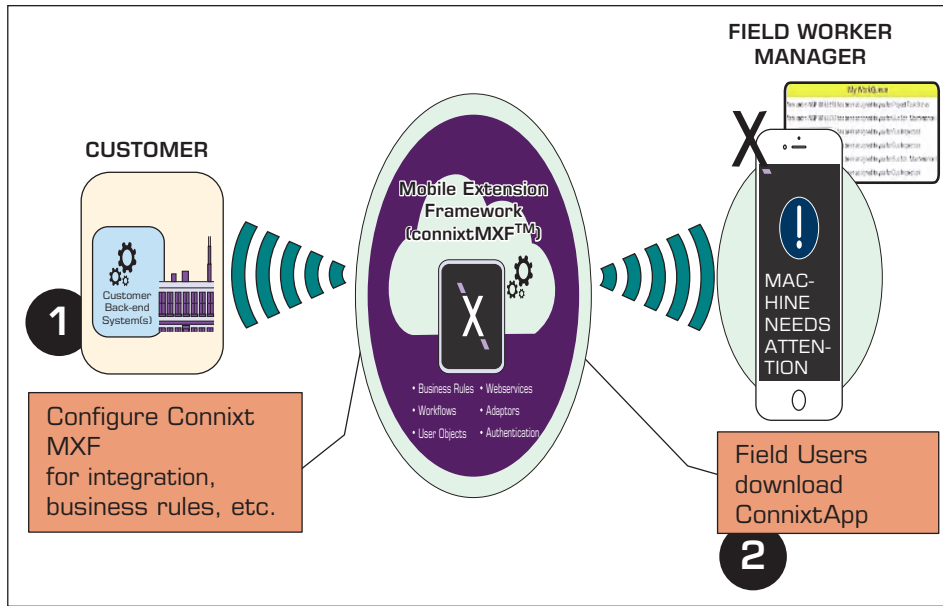


FIGURE 10.3 Interconnections between Workers and Technology in a Cloud Analytics Application.

and deployment—configuration is completed in the cloud, and once configured, the apps are ready for download and deployment. More importantly, MXF enables easy modifications to form and processes—for example, if the utility needs to add additional options to the incident drop-down, they simply add this once in MXF. Within minutes the option is now available on the field for all users. Figure 10.3 illustrates this architecture.

There are further benefits from a system that leverages ubiquitous cloud and mobile technologies. Because all the business logic and configurations are stored in the cloud, the solution itself can act as a stand-alone system for customers who have no back-end systems—important in the context of small and medium businesses (SMBs). And for those with back-end systems, the connectivity is seamless through Web services and the back-end system serves as the system of record. This additionally helps businesses adopt technology in a phased manner—starting with a noninvasive, stand-alone system with minimal internal IT impact while automating field operations, and then moving toward back-end system integration.

On the other hand, the mobile apps are themselves system agnostic—they communicate using standard Web services and the end device can be Android or iOS and smartphone or tablet. Thus, irrespective of the device used, all communication, business logic, and algorithms are standardized across platforms/devices. As native apps across all devices, iMarq leverages standard technology that is provided by the device manufacturers and the OS vendors. For example, using native maps applications allows the apps to benefit from improvements made by the platform vendors; thus, as maps become more accurate, the end users of the mobile apps also benefit from these advances.

Finally, for successful deployments, enterprise cloud-mobile technology must be heavily user-centric. The look and feel must be geared to user-comfort, much as users expect from any mobile app they use. Treating the business user as an app consumer meets their standard expectations of an intuitive app that immediately saves them time and effort. This approach is essential to ensuring successful adoption.

The utility now has better information from first responders, as information is directly shared from the field (not through a dispatcher or other third party), pictures are available, and there is geo- and time-stamping. Garbled phone messages are avoided. The two-way communication between utility and the first responder in the field is improved. Historical records of the incidents are kept.

The utility and the first responders are now more unified in their quick and complete responses to incidents, improving service to the public. By tightening ties with first responders (police and fire department personnel), the public is served with a better coordinated and superior response for incidents that are discovered by first responders.

QUESTIONS FOR DISCUSSION

1. How does cloud technology impact enterprise software for small and mid-size businesses?
2. What are some of the areas where businesses can use mobile technology?
3. What types of businesses are likely to be the forerunners in adopting cloud-mobile technology?
4. What are the advantages of cloud-based enterprise software instead of the traditional on-premises model?
5. What are the risks of cloud versus traditional on-premises applications?

Source: Used with permission from G. Satish, Connixit, Inc.

Cloud Deployment Models

Cloud services can be acquired in several ways, from building an entirely private infrastructure to sharing with others. The following three models are the most common.

- **Private cloud:** This can also be called internal cloud or corporate cloud. It is a more secure form of cloud service than public clouds like Microsoft Azure and Amazon Web Services. It is operated solely for a single organization having a mission critical workload and security concerns. It provides the same benefits as a public cloud-like service, scalability, changing computing resources on demand, and so on. Companies that have a private cloud have direct control over their data and applications. The disadvantage of having a private cloud is the cost of maintaining and managing the cloud because on-premises IT staff are responsible for managing it.
- **Public cloud:** In this model the subscriber uses the resources offered by service providers over the Internet. The cloud infrastructure is managed by the service provider. The main advantage of this public cloud model is saving time and money in setting up hardware and software required to run their business. Examples of public clouds are Microsoft Azure, Google Cloud Platform, and Amazon AWS.
- **Hybrid cloud:** The hybrid cloud gives businesses great flexibility by moving workloads between private and public clouds. For example, a company can use hybrid cloud storage to store its sales and marketing data, and then use a public cloud platform like Amazon Redshift to run analytical queries to analyze its data. The main requirement is network connectivity and API (application program interface) compatibility between the private and public cloud.

Major Cloud Platform Providers in App Development and Deployment

- **Amazon Elastic Beanstalk:** Amazon Elastic Beanstalk is a service offered by Amazon Web Services. It can deploy, manage, and scale Web applications. It supports the following programming languages: Java, Ruby, Python, PHP, Go, Docker, and .NET on servers like Apache, Nginx, Passenger, and IIS. A user must upload the code for the application, and Elastic Beanstalk handles the deployment of the

application, load balancing, and autoscaling and monitors the health of the application. So, the user can focus on building Web sites, mobile applications, API back-end, content management systems, SaaS, and so on, while the applications and infrastructure to manage them are taken care of by Elastic Beanstalk. A user can use Amazon Web Services or an integrated development environment like Eclipse or Visual Studio to upload their application. A user must pay for AWS resources needed to store and run the applications.

- **IBM Cloud:** IBM Cloud is a cloud platform that allows a user to build apps using many open-source computer technologies. Users can also deploy and manage hybrid applications using the software. With IBM Analytics Engine (a reincarnation of **IBM Watson**), whose services are available on IBM Cloud, users can now create next-generation cognitive applications that can discover, innovate, and make decisions. IBM Analytics Engine services can be used for analyzing emotions and synthesizing natural-sounding speech from text. It uses the concept of cognitive computing to analyze text, video, and images. It supports many programming languages like Java, Go, PHP, Ruby, and Python.
- **Microsoft Azure:** Azure is a cloud platform created by Microsoft to build, deploy, and manage applications and services through a network of Microsoft data centers. It serves as both PaaS and IaaS and offers many solutions such as analytics, data warehousing, remote monitoring, and predictive maintenance.
- **Google App Engine:** Google App Engine is Google's Cloud computing platform used for developing and hosting applications. Managed by Google's data centers, it supports developing apps in Python, Java, Ruby, and PHP programming languages. The big query environment offers data warehouse services through the cloud.
- **OpenShift:** OpenShift is Red Hat's cloud application platform based on a PaaS model. Through this model, application developers can deploy their applications on the cloud. There are two different models available for OpenShift. One serves as a public PaaS and the other serves as a private PaaS. OpenShift Online is Red Hat's public PaaS that offers development, build, hosting, and deployment of applications in the cloud. The private PaaS, OpenShift Enterprise, allows development, build, and deployment of applications on an internal server or a private cloud platform. RedHat was acquired by IBM in 2019.

Analytics as a Service (AaaS)

Analytics and data-based managerial solutions—the applications that query data for use in business planning, problem solving, and decision support—are evolving rapidly and being used by every organization. Enterprises are being flooded with information and getting insights from this data is a big challenge for them. Along with that, there are challenges related to data security, data quality, and compliance. AaaS is an extensible analytical platform using a cloud-based delivery model where various BI and data analytics tools can help companies in better decision making and get insights from their huge amount of data. The platform covers all functionality aspects from collecting data from physical devices to data visualization. AaaS provides an agile model for reporting and analytics to businesses so they can focus on what they do best. Customers can either run their own analytical applications in the cloud or they can put their data on the cloud and receive useful insights.

AaaS combines aspects of cloud computing with Big Data analytics and empowers data scientists and analysts by allowing them to access centrally managed information data sets. They can now explore information data sets more interactively and discover

richer insights more rapidly, thus erasing many of the delays that they may face while discovering data trends. For example, a provider might offer access to a remote analytics platform for a fee. This allows the client to use analytics software for as long as it is needed. AaaS is a part of SaaS, PaaS, and IaaS, thus helping IT significantly reduce costs and compliance risk, while increasing productivity of users.

AaaS in the cloud has economies of scale and scope by providing many virtual analytical applications with better scalability and higher cost savings. With growing data volumes and dozens of virtual analytical applications, chances are that more of them leverage processing at contrasting times, usage patterns, and frequencies.

Data and text mining is another very promising application of AaaS. The capabilities that a service orientation (along with cloud computing, pooled resources, and parallel processing) brings to the analytics world can also be used for large-scale optimization, complex multicriteria decision problems, and distributed simulation models. Next, we identify selected cloud-based analytics offerings.

Representative Analytics as a Service Offerings

IBM CLOUD IBM is making all its analytics offerings available through its cloud. IBM Cloud offers several categories of analytics and AI (Artificial Intelligence). For example, IBM Analytics Engine integrates most of the analytics features and capabilities that can be built and deployed through their cloud. In addition, it employs text mining and deep learning at a remarkably important level. It was introduced earlier in the context of text mining.

AMAZON QUICKSIGHT Amazon QuickSight (<https://aws.amazon.com/quicksight/>) (accessed August 2022) is a cloud-based business intelligence tool which helps us to analyze, create customized visualizations/Dashboards and get insights from our data by seamlessly integrating with data sources. It can also perform advanced machine learning-based analytics like Natural Language Processing (NLP).

GOOGLE DATA STUDIO Google Data Studio (<https://datastudio.google.com/>) (accessed August 2022) is a cloud-based analytics platform by Google using which we can connect to data virtually and process the data quickly to create dashboards and reports. It also enables collaboration between teams/individuals by sharing insights real time. The other analytics offerings by Google are Looker, Big Query.

MINEMYTEXT.COM One of the areas of major growth in analytics is text mining. Text mining identifies high-level topics of documents, infers sentiments from reviews, and visualizes the document or term/concept relationships, as covered in the text mining chapter. A start-up called MineMyText.com (accessed August 2022) offers these capabilities in the cloud through their Web site.

SAS VIYA SAS Institute is making its analytics software offering available on demand through the cloud. Currently, SAS Visual Statistics is only available as a cloud service and is a competitor of Tableau.

TABLEAU Tableau, a major visualization software that was introduced in the context of descriptive analytics, is also available through the cloud as Tableau Cloud.

SNOWFLAKE Snowflake is a cloud-based data warehouse solution. Users can bring together their data from multiple sources as one source and analyze it using Snowflake.

ANALYTICS IN ACTION 10.2 GO-JEK Employs Google Cloud Platform for Forecasting and Pricing

The majority of Indonesians experience automobile traffic congestion. The country's highways as well as infrastructure struggle to serve its 260 million individuals, approximately ten million of whom are living in Jakarta, the capital city. Indonesians primarily depend on motorcycles, as well as motorbike taxis, called ojek, for their commuting or travel needs.

GO-JEK was founded in 2010 and is based in Jakarta. It started as a phone call facility for ojek bookings. The company has used the need for the program to be among the few “unicorn”—privately held startup companies estimated at over USD one billion—within Southeast Asia.

Starting 2010, GO JEK collected information to learn consumer preferences and in 2015 released a mobile app which bundled ride hailing, food shipping and delivery, and then shopping for groceries. “Demand for the application was insatiable and we hit hypergrowth very quickly,” says Willem Pienaar, Data Science Platform Lead at GO-JEK.”

GO-JEK runs its applications and data on Google Cloud Platform. GCP plays a significant role in helping GO-JEK to get insights into the data of 1 million drivers (about the population of Delaware) in about 167 locations involving 2.5 million customers (about twice the population of Hawaii). The average amount of data generated per day is about 4TB to 5TB. The GCP technologies used by GO-JEK include Cloud Dataflow, Cloud Bigtable, BigQuery, Cloud Pub/Sub and Cloud Storage. These services played a vital role in driving the success of the mobile app.

“We opted for BigQuery to store our training data, because of its high scalability and the fact it is a completely cloud-based service means you don’t have to manage any infrastructure,” says Pienaar, Data Science Platform Lead at GO-JEK.

Overall, GCP has helped GO-JEK mobile app in achieving the below results.

- Supports more than one million motorbike owners with quick access to riders as well as enhanced routes.
- Enables on-demand forecasting as well as rates changes.
- Helps GO-JEK in achieving rapid scaling and international expansion.

QUESTIONS FOR DISCUSSION

1. How did GCP play an instrumental role in the success of GO-JEK Mobile application?
2. How does GCP save time for GO-JEK by managing the underlying infrastructure to focus on exploring data and features?
3. What are the GCP services GO-JEK used to overcome and manage the increasing demand of customers?

Source: Compiled from GO-JEK Case Study | Google Cloud, <https://cloud.google.com/customers/go-jek/> (accessed August 2022).

Illustrative Analytics Applications Employing the Cloud Infrastructure

In this section we highlight a couple other cloud analytics applications. We present them as one section as opposed to individual Analytics in Action.

Using Azure IoT (Internet of Things), Stream Analytics, and Machine Learning to Improve Mobile Healthcare Services

People are increasingly using mobile applications to keep track of the amount of exercise they do every day and maintain their health history as well. **Zion China**, which is a provider of mobile healthcare services, has produced an innovative health monitoring tool that gathers data about health problems such as glucose levels, blood pressure, diet, medication, and exercise of their users and help them improve their quality of life by giving them suggestions on how to manage their health and prevent or cure illness daily.

The huge volume of real-time data presented scalability and data management problems, so the company collaborated with Microsoft to take advantage of **Stream Analytics**, Machine Learning, IOT solution and Power BI, which also improved data security and analysis. Zion China was completely dependent on traditional BI with data being collected from various devices or cloud. Using a cloud-based analytics architecture, Zion was able to add several features, speed, and security. They added an IoT (Internet of Things) hub to the front end for better transmission of data from device to cloud. The data is first transferred from the device to a mobile application via Bluetooth and then to an IoT hub via HTTPS and AMQP. Stream Analytics helps in processing the real-time data gathered in the IoT hub and generates insights and useful information, which is further streamed to an SQL database. They use Azure Machine Learning to generate predictive models on diabetes patient data and improve the analysis and prediction levels. Power BI provides simple and easy visualization of data insights achieved from analysis to the users.

Sources: “Zion China Uses Azure IoT, Stream Analytics, and Machine Learning to Evolve Its Intelligent Diabetes Management Solution” at <https://advocacypublic.cloudamppe.microsoft.com/en-au/story/zionchina> (accessed August 2022) and <https://microsoft.github.io/techcasestudies/iot/2016/12/02/IoT-ZionChina.html> (accessed August 2022).

Chime Enhances Customer Experience Using Snowflake

Chime, a banking option, offers a Visa debit card, FDIC-insured spending and savings account, and a mobile application app that makes banking easier for people. Chime wanted to learn about their customer engagement. They wanted to analyze data across their mobile, Web, and backend platforms to help enhance the user experience. However, pulling and aggregating data from multiple sources such as ad services from Facebook and Google and events from other third-party analytics tools like JSON (JavaScript Object Notation) docs was a laborious task. They wanted a solution that could aggregate data from these multiple sources and analyze the data set. Chime needed a solution that could process JSON data sources and query them using standard SQL database tables.

Chime started using Snowflake Elastic Data Warehouse solution. Snowflake pulled data from all 14 data sources of chime, including data like JSON docs from applications. Snowflake helped Chime analyze JSON data quickly to enhance member services and provide a more personalized banking experience to customers.

Source: Based on **Snowflake.net**. (n.d.). Chime Delivers Personalized Customer Experience Using Chime. <http://www.snowflake.net/product> (accessed August 2022).

We are entering the “petabyte age,” and traditional data and analytics approaches are beginning to show their limits. Cloud analytics is an emerging alternative solution for large-scale data analysis. Data-oriented cloud systems include storage and computing in a distributed and virtualized environment. A major advantage of these offerings is the rapid diffusion of advanced analysis tools among the users, without significant investment in technology acquisition. These solutions also come with many challenges, such as security, service level, and data governance. Several concerns have been raised about cloud computing, including loss of control and privacy, legal liabilities, cross-border political issues, and so on. According to **Cloud Security Alliance**, the top three security threats in the cloud are data loss and leakage, hardware failure of equipment, and an insecure interface. All the data in the cloud is accessible by the service provider, so the service provider can unknowingly or deliberately alter the data or can pass the data to a third party for purposes of law without asking the company. Research is still limited in this area. As a result, there are many opportunities to bring analytical, computational, and conceptual modeling into the context of service science, service orientation, and cloud

intelligence. Nonetheless, cloud computing is an important initiative for analytics professionals to watch as it is a fast-growing area.

SECTION 10.2 REVIEW QUESTIONS

1. Define *cloud computing*. How does it relate to PaaS, SaaS, and IaaS?
2. Give examples of companies offering cloud services.
3. How does cloud computing affect BI?
4. How does DaaS change the way data is handled?
5. What are the several types of cloud platforms?
6. Why is AaaS cost-effective?
7. Name at least three major cloud service providers.
8. Give at least three examples of analytics-as-a-service providers.

10.3 LOCATION-BASED ANALYTICS

Thus far, we have seen many examples of organizations employing analytical techniques to gain insights into their existing processes through informative reporting, predictive analytics, forecasting, and optimization techniques. In this section, we learn about a critical emerging trend—incorporation of location data in analytics. Figure 10.4 gives our classification of location-based analytic applications. We first review applications that make use of static location data that is usually called *geospatial data*. We then examine the explosive growth of applications that take advantage of all the location data being generated by today's devices. This section first focuses on analytics applications that are being developed by organizations to make better decisions in managing operations, targeting customers, promotions, and so forth. Then we will also explore analytics applications that are being developed to be used directly by a consumer, some of which also take advantage of the location data.

Geospatial Analytics

A consolidated view of the overall performance of an organization is usually represented through visualization tools that provide actionable information. The information may include current and forecasted values of various business factors and

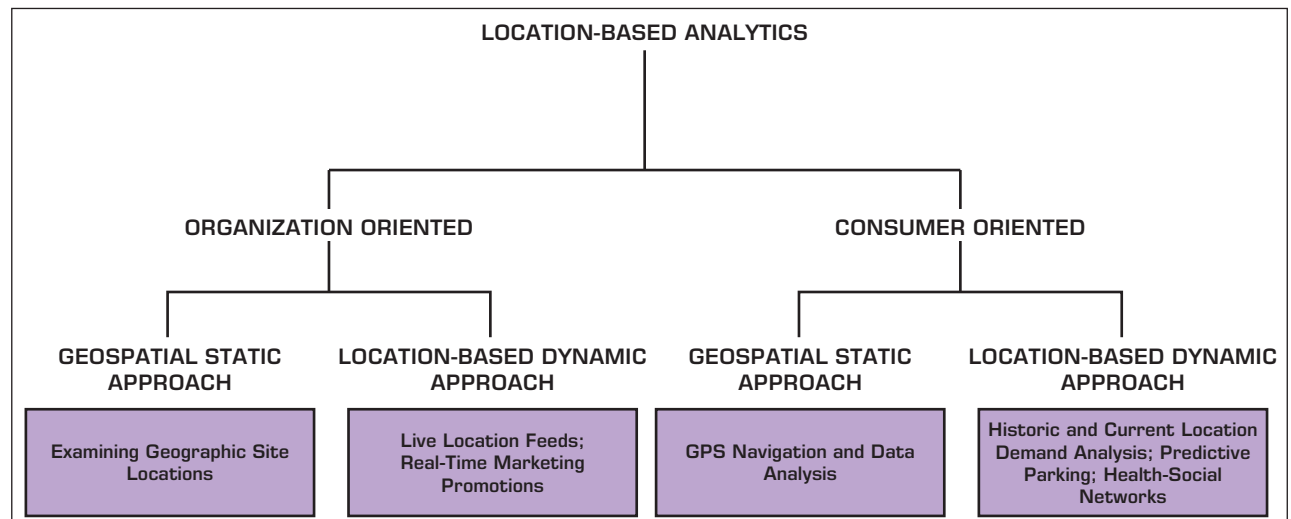


FIGURE 10.4 Classification of Location-Based Analytics Applications.

key performance indicators (KPIs). Looking at the KPIs as overall numbers via various graphs and charts can be overwhelming. There is a substantial risk of missing potential growth opportunities or not identifying the problematic areas. As an alternative to simply viewing reports, organizations employ visual maps that are geographically mapped and based on the traditional location data, usually grouped by postal codes. These map-based visualizations have been used by organizations to view the aggregated data and get more meaningful location-based insights. The traditional location-based analytic techniques using geocoding of organizational locations and consumers hamper the organizations in understanding “true location-based” impacts. Locations based on postal codes offer an aggregate view of a large geographic area. This poor granularity may not help pinpoint the growth opportunities within a region, as the location of target customers can change rapidly. Thus, an organization’s promotional campaigns may not target the right customers if it is based on postal codes. To address these concerns, organizations are embracing location and spatial extensions to analytics. The addition of location components based on latitudinal and longitudinal attributes to the traditional analytical techniques enables organizations to add a new dimension of “where” to their traditional business analyses, which currently answers the questions of “who,” “what,” “when,” and “how much.”

Location-based data are now readily available from **geographic information systems (GIS)**. These are used to capture, store, analyze, and manage data linked to a location using integrated sensor technologies, global positioning systems installed in smartphones, or through RFID (Radio Frequency Identification) deployments in the retail and healthcare industries.

By integrating information about the location with other critical business data, organizations are now creating location intelligence. Location intelligence is enabling organizations to gain critical insights and make better decisions by optimizing important processes and applications. Organizations now create interactive maps that further drill down to details about any location, offering analysts the ability to investigate latest trends and correlate location-specific factors across multiple KPIs. Analysts can now pinpoint trends and patterns in revenue, sales, and profitability across geographical areas.

By incorporating demographic details into locations, retailers can determine how sales vary by population level and proximity to other competitors; they can assess the demand and efficiency of supply-chain operations. Consumer product companies can identify the specific needs of customers and customer complaint locations and easily trace them back to the products. Sales reps can better target their prospects by analyzing their geography.

A company that is the market leader in providing GIS data is ESRI (esri.com). ESRI licenses its **ArcGIS** software to thousands of customers including commercial, government, and the military. It would take a book or more to highlight applications of ESRI’s GIS database and software! Another company grindgis.com identifies over 60 categories of GIS applications (<http://grindgis.com/blog/gis-applications-uses>, accessed August 2022). A few examples that have not been mentioned yet include the following:

- **Agricultural applications:** By combining location, weather, soil, and crop-related data, very precise irrigation and fertilizer applications can be planned. Examples include companies such as proagrica.com and ageagle.com (they combine GIS and the latest information collected through drones, another emerging technology).
- **Crime analysis:** Superimposition of crime data including date, time, and type of crime onto the GIS data can provide significant insights into crime patterns and police staffing.
- **Disease spread prediction:** One of the first known examples of descriptive analytics is the analysis of the cholera outbreak in London in 1854. Dr. John Snow plotted the cases of cholera on a map and was able to refute the theory that the cholera

outbreak was being caused by bad air. The map helped him pinpoint the outbreak to a bad water well (**Tulchinsky, 2018**). We have come a long way from needing to plot maps manually, but the idea of being able to track and then predict outbreaks of diseases, such as the flu, using GIS and other data has become a major field. Of course, the COVID-19 pandemic led to incorporation of analytics in many health applications, including display of various COVID-19 statistics (number of cases, death rates, hospitalization rates, ICU utilization, positivity rates, vaccination rates, etc.) to be displayed at local, county, state, and national levels, all using GIS data.

In addition, with location intelligence, organizations can quickly overlay weather and environmental effects and forecast the level of impact on critical business operations. With technological advancements, geospatial data is now being directly incorporated in enterprise data warehouses. Location-based in-database analytics enable organizations to perform complex calculations with increased efficiency and get a single view of all the spatially oriented data, revealing hidden trends and new opportunities. For example, Teradata's data warehouse supports the geospatial data feature based on the SQL/MM standard. The geospatial feature is captured as a new geometric data type called ST_GEOMETRY. It supports a large spectrum of shapes, from simple points, lines, and curves to complex polygons in representing the geographic areas. They are converting the nonspatial data of their operating business locations by incorporating the latitude and longitude coordinates. This process of geocoding is readily supported by service companies like NAVTEQ and Tele Atlas, which maintain worldwide databases of addresses with geospatial features and make use of address-cleansing tools like Informatica and Trillium, which support mapping of spatial coordinates to the addresses as part of extract, transform, and load functions.

Organizations across a variety of business sectors are employing **geospatial analytics**. We will review some examples next. Analytics in Action 10.3 provides an example of how location-based information was used in making loan approval decisions. Loan approval decisions by estimating the value of a property has been one of the early application areas of analytics. The case below illustrates how to utilize a business intelligence framework and a geographic information system (GIS) to aid in the appraisal decision by analyzing data-related to the appraised property in real time.

ANALYTICS IN ACTION 10.3 Improving Home Loan Appraisal Process Using BI and Geographic Information Systems

Current Appraisal Problem

Danuja, the appraisal manager of ABC Bank (Bank name disguised) of Thailand was handed over the responsibility of managing the current appraisal process. The appraisal procedure is time consuming. Danuja's team can only manage just two to three applications per day, or maybe a maximum of four instances when they are within one region. The bank receives an average of 7,000–8,000 applications cases per month. Also, the appraisal values are extremely different between external and internal appraisers. For example, for the exact same categories of homes with very much the same qualities, there was an inconsistency within the appraisal values of as much as 257 %, which is undesirable. This impacted the overall loan approval process. In evaluating the problem, Danuja worked with Jong, senior certified analytics professional from the **Business Intelligence (BI)** Team, and Kampu, a GIS solutions expert to find a solution for this issue of appraisal discrepancy.

Solution

Their focus is on the home loan services, which comprises about 52% share of the total loans offered by the bank. They realized that GIS integrated with other pertinent loan information can provide more complete insights for the organizations to make clearer decisions with respect to geographic regions. After their meeting, they decided to work on the two most crucial

factors that had greater impact on the appraisal process, the % growth of the properties and the weighted quality scores.

Kampu analyzed the 30,000 data points of the home loans provided by the bank and built machine learning models to perform predictions. The machine learning models were fed data from the bank's GIS and MIS database. She explored models like decision tree, support vector machine, neural network, and regression analysis. The variables used in the models are total area of the house, region, Appraisal price, % Predicted growth and appraisal price after growth. For example, the models with all the related factors and attributes revealed that the area under application H10001 has the fastest growth rate among the real estate properties with 9% predicted growth. The trained model can now be used to assess the price of a new home loan application that had been appraised by the team.

Using the H10001 situation being a good example, Danuja can insist upon making use of the initial appraisal worth of 7,000,000 baht suggested by their appraisal staff to home mortgage committee and have the agreement be approved. Also, they can utilize this GIS-based BI integration to collect details of the properties, change details, and try various scenarios prior to making the pricing call.

QUESTIONS FOR DISCUSSION

1. What exactly are the components to think about with respect to property appraisal?
2. Put together a GIS-based business intelligence design with the house appraisal procedure as well as determine the elements and their related features.
3. Assuming being part of appraisal team for H10001 application, what factors can be suggested to the house loan committee?

Source: Contributed by Dr. Jongswas Chongwatpol, NIDA, Thailand.

ANALYTICS IN ACTION 10.4 Starbucks Exploits GIS and Analytics to Grow Worldwide

One of the key challenges for any organization that is trying to grow its presence is deciding the location of its next store. Starbucks faces the same question. To identify new store locations, more than 700 Starbucks employees (referred to as partners) in 15 countries use an ArcGIS-based market planning and BI solution called Atlas. Atlas provides partners with workflows, analysis, and store performance information so that local partners in the field can make decisions when identifying new business opportunities.

As reported in multiple sources, Atlas is employed by local decision makers to understand the population trends and demand. For example, in China, there are over 1,200 Starbucks stores, and the company is opening a new store every day. Information such as trade areas, retail clusters and generators, traffic, and demographics is important in deciding the next store's location. After analyzing a new market and neighborhood, a manager can look at specific locations by zooming into an area in the city and identifying where three new office towers may be completed over the next two months, for example. After viewing this area on the map, a workflow window can be created that will help the manager move the new site through approval, permitting, construction, and eventually opening.

By integrating weather and other local data, one can also better manage demand and supply-chain operations. Starbucks is integrating its enterprise business systems with its GIS solutions in Web services to see the world and its business in new ways. For example, Starbucks integrates AccuWeather's forecasted real-feel temperature data. This forecasted temperature data can help localize marketing efforts. If a hot week in Memphis is forthcoming, Starbucks analysts can select a group of coffee houses and get detailed information on past and future weather patterns, as well as store characteristics. This knowledge can be used to design a localized promotion for Frappuccino's, for example, helping Starbucks anticipate what its customers will be wanting a week in advance.

Major events also have an impact on coffee houses. When 150,000 people descended on San Diego for the Pride Parade, local baristas served a lot of customers. To ensure the best possible customer experience, Starbucks used this local event knowledge to plan staffing and inventory at locations near the parade.

QUESTIONS FOR DISCUSSION

1. What type of demographics and GIS information would be relevant for deciding on a store location?
2. It has been mentioned that Starbucks encourages its customers to use its mobile app. What type of information might the company gather from the app to help it better plan operations?
3. Will the availability of free Wi-Fi at Starbucks stores provide any information to Starbucks for better analytics?

Sources: Digit.HBS.org. (2015). “Starbucks: Brewing up a Data Storm!” <https://digit.bbs.org/submission/starbucks-brewing-up-a-data-storm/> (accessed August 2022); Wheeler, C. (2014). Blogs.ESRI.com. “From Customers to CxOs, Starbucks Delivers World-Class Service.” (2014). <https://blogs.esri.com/esri/ucinsider/2014/07/29/starbucks/> (accessed August 2022).

In addition to the retail transaction analysis applications highlighted here, there are many other applications of combining geographic information with other data being generated by an organization. For example, network operations and communication companies often generate massive amounts of data every day. The ability to analyze the data quickly with an elevated level of location-specific granularity can better identify the customer churn and help in formulating strategies specific to locations for increasing operational efficiency, quality of service, and revenue.

Geospatial analysis can enable communication companies to capture daily transactions from a network to identify the geographic areas experiencing many failed connection attempts of voice, data, text, or Internet. Analytics can help determine the exact causes based on location and drill down to an individual customer to provide better customer service. You can see this in action by completing the following multimedia exercise.

A Multimedia Exercise in Analytics Employing Geospatial Analytics

Teradata University for Academics includes a BSI video on the case of dropped mobile calls. Please watch the video that appears on YouTube at the following link: https://www.youtube.com/watch?v=4WJR_Z3exw4 (accessed August 2022).

A telecommunication company launches a new line of smartphones and faces problems with dropped calls. The new rollout is in trouble, and the northeast region is the worst hit region as they compare effects of dropped calls on the profits for the geographic region. The company hires BSI to analyze the problems arising due to defects in smartphone handsets, tower coverage, and software glitches. The entire northeast region data is divided into geographic clusters, and the company solves the problem by identifying the individual customer data. The BSI team employs geospatial analytics to identify the locations where network coverage was leading to dropped calls and suggests installing a few additional towers where unhappy customers are located.

After the video is complete, you can see how the analysis was prepared at: slideshare.net/teradata/bsi-teradata-the-case-of-the-dropped-mobile-calls (accessed August 2022).

This multimedia excursion provides an example of a combination of geospatial analytics along with Big Data analytics that assist in better decision making.

Real-Time Location Intelligence

Many devices in use by consumers and professionals are constantly sending out their location information. Cars, buses, taxis, mobile phones, cameras, and personal navigation devices all transmit their locations thanks to network-connected positioning technologies such as GPS, Wi-Fi, and cell tower triangulation. Millions of consumers and businesses use location-enabled devices for finding nearby services, locating friends and family, navigating, tracking assets and pets, dispatching, and engaging in sports, games, and hobbies. This surge in location-enabled services has resulted in a massive database of historical and real-time streaming location information. It is, of course, scattered and not especially useful by itself. The automated data collection enabled through capture of cell phones and Wi-Fi hotspot access points presents an interesting new dimension in nonintrusive market research, data collection, and, of course, microanalysis of such massive data sets.

By analyzing and learning from these large-scale patterns of movement, it is possible to identify distinct classes of behaviors in specific contexts. This approach allows a business to better understand its customer patterns and make more informed decisions about promotions, pricing, and so on. By applying algorithms that reduce the dimensionality of location data, one can characterize places according to the activity and movement between them. From massive amounts of high-dimensional location data, these algorithms uncover trends, meaning, and relationships to eventually produce human-understandable representations. It then becomes possible to use such data to automatically make intelligent predictions and find important matches and similarities between places and people.

Location-based analytics finds its application in consumer-oriented marketing applications. Many companies are now offering platforms to analyze location trails of mobile users based on geospatial data obtained from the GPS and target tech-savvy customers with coupons on their smartphones as they pass by a retailer. This illustrates the emerging trend in the retail space where companies are looking to improve efficiency of marketing campaigns—not just by targeting every customer based on real-time location, but by employing more sophisticated predictive analytics in real time on consumer behavioral profiles to find the right set of consumers for advertising campaigns.

Yet another extension of location-based analytics is to use augmented reality. In 2016, Pokémon GO became a market sensation. It is a location-sensing augmented reality-based game that encourages users to claim virtual items from select geographic locations. The user can start anywhere in a city and follow markers on the app to reach a specific item. Virtual items are visible through the app when the user points a phone's camera toward the virtual item. The user can then claim this item. Business applications of such technologies are also emerging. For example, an app called Candybar allows businesses to place these virtual items on a map using Google Maps. The placement of this item can be fine-tuned using Google's Street View. Once all virtual items have been configured with the information and location, the business can submit items, which are then visible to the user in real time. Candybar also provides usage analytics to the business to enable better targeting of virtual items. The **virtual reality** aspect of this app improves the experience of users, providing them with a “gaming” environment in real life. At the same time, it provides a powerful marketing platform for businesses to reach their customers.

Augmented Reality is reportedly a big push at both Microsoft and Apple. These technologies depend upon the location information, images through camera, as well as other **sensors** (when available) to provide the user an immersive experience. We will see an example of this in a brief discussion of **digital twins** in a later section.

As is evident from this section, location-based analytics and ensuing applications are the most important front soon for organizations. A common theme in this section was the use of operational or marketing data by organizations. We will next explore analytics applications that are directly targeted at users and sometimes take advantage of location information.

Analytics Applications for Consumers

The explosive growth of the apps industry for smartphone platforms (iOS, Android, Windows, and so forth) and the use of analytics are creating tremendous opportunities for developing apps where the consumers use analytics without ever realizing it. These apps differ from the previous category in that these are meant for direct use by a consumer, as opposed to an organization that is trying to mine a consumer's usage/purchase data to create a profile for marketing specific products or services. Predictably, these apps are meant for enabling consumers to make better decisions by employing specific analytics. We highlight two of these in the following examples.

- Waze, a social Web app that assists users in identifying a navigation path and alerts users about potential issues such as accidents, police checkpoints, speed traps, and construction, based on other users' inputs, has become an exceedingly popular navigation app. Google acquired this app a few years ago and has enhanced it further. This app is an example of aggregating user-generated information and making it available for customers.
- Another transportation-related app that uses predictive analytics, ParkPGH, has been deployed since about 2010 in Pittsburgh, Pennsylvania. Developed in collaboration with Carnegie Mellon University, this app includes predictive capabilities to estimate parking availability. ParkPGH directs drivers to parking lots in areas where parking is available. It calculates the number of parking spaces available in several garages in the cultural arts district of Pittsburgh. Available spaces are updated every 30 seconds, keeping the driver as close to the current availability as possible. Depending on historical demand and current events, the app can predict parking availability and provide information on which lots will have free space by the time the driver reaches the destination. The app's underlying algorithm uses data on current events around the area—for example, a basketball game—to predict an increase in demand for parking spaces later that day, thus saving the commuters valuable time searching for parking spaces in the busy city. Success of this app has led to a proliferation of parking apps that work in many major cities and allow a user to book a parking space in advance, recharge the meter, even bid for a parking space, etc. Both iPhone app store and Google Play store include many such apps.

A recent blog describes how a data science professional built their own parking prediction app by collecting some local data. This is a very useful blog with links to the code on GitHub: <https://towardsdatascience.com/how-i-built-my-own-real-time-parking-availability-predictive-model-31332e1b7747> (accessed August 2022).

Analytics-based applications are emerging not just for fun and health but also to enhance one's productivity. For example, Google's Gmail app analyzes billions of e-mail transactions and develops automated responses for e-mails. When a user receives an e-mail and reads it in her Gmail app, the app also recommends short responses for the e-mail at hand that a user can select and send to the original sender.

As is evident from these examples of consumer-centric apps, predictive analytics is beginning to enable development of software that is directly used by a consumer. We believe that the growth of consumer-oriented analytic applications will continue and create many entrepreneurial opportunities for the readers of this book.

One key concern in employing these technologies is the loss of privacy. If someone can track the movement of a cell phone, the privacy of that customer is a big issue. Some of the app developers claim that they only need to gather aggregate flow information, not individually identifiable information. But many stories appear in the media that highlight violations of this general principle. Both users and developers of such apps must be very aware of the deleterious effect of giving out confidential information as well as collecting such information. We discuss this issue a bit further in Chapter 11.

► SECTION 10.3 REVIEW QUESTIONS

1. How does traditional analytics make use of location-based data?
2. How can geocoded locations assist in better decision making?
3. What is the value provided by geospatial analytics?
4. Explore the use of geospatial analytics further by investigating its use across various sectors like government census tracking, consumer marketing, and so forth.
5. Search online for other applications of consumer-oriented analytical applications.
6. How can location-based analytics help individual consumers?
7. Explore more transportation applications that may employ location-based analytics.
8. What other applications can you imagine if you were able to access cell phone location data?

10.4 IMAGE ANALYTICS/ALTERNATIVE DATA

As seen in this section, analytics techniques are being applied to many diverse industries and data. An area of growth has been analysis of visual images. Advances in image capturing through high-resolution cameras, storage capabilities, and deep learning algorithms have enabled remarkably interesting analyses. Satellite data have often proven their utility in many different fields. The benefits of satellite data at high resolution and in different forms of imagery including multispectral are significant to scientists who need to regularly check global change, land usage, and weather. In fact, by combining satellite imagery and other data including information on social media, government filings, and so on, one can surmise business planning activities, traffic patterns, changes in parking lots or open spaces. Companies, government agencies, and nongovernmental organizations (NGOs) have invested in satellites to try to image the whole globe every day so that daily changes can be tracked at any location and the information can be used for forecasting. In the last few months, many interesting examples of more reliable and advanced forecasts have been reported. Indeed, this activity is being led by different industries across the globe and has added a term to Big Data called **Alternative Data**. Here are a few.

- World Bank researchers used satellite data to propose strategic recommendations for urban planners and officials from developing nations. This analysis arose due to the recent natural disaster where at least 400 people died in Freetown, Sierra Leone. Researchers clearly proved that Freetown and some other developing cities lacked systematic planning of their infrastructure and that resulted in the loss of life. The bank researchers are using satellite imagery now to make critical decisions about risk-prone urban areas.
- **EarthCast** provides correct weather updates for a large commercial U.S. airline based on the data it pulls from a constellation of 60 government-operated satellites combined with ground and aircraft-based sensors, tracking anything from lightning to turbulence. It has even developed the capability to map out conditions along a flight path and supplies customized forecasts for everything from hot air balloons to drones.
- Imazon started using satellite data to develop a picture of close real-time information on Amazon **deforestation**. It uses advanced optical and infrared imagery that has led to the identifying of illegal sawmills. Imazon is now focused more on getting data to local governments through its “green municipalities” program that trains officials to find and curb deforestation.
- International nonprofit organization Global Fishing Watch (**globalfishingwatch.org**) processes satellite extracted information on ship movement to spot where and when vessels are fishing illegally. They have now partnered with many countries

to detect and report unreported and unregulated fishing using various data sources including the satellite data.

These examples illustrate just a sample of ways that satellite data can be combined with analytics to generate new insights. In anticipation of the coming era of abundant earth observations from satellites, scientists and communities must put some thought into recognizing key applications and key scientific issues for the betterment of society. Although such concerns will eventually be resolved by policymakers, what is clear is that new and interesting ways of combining satellite data and many other data sources are spawning a new crop of analytics companies.

Such image analysis is not limited to satellite images. Cameras mounted on drones and traffic lights on every conveyable pole in buildings and streets supply the ability to capture images from just a few feet high. Analysis of these images coupled with facial recognition technologies is enabling all kinds of new applications from customer recognition to governments' ability to track all subjects of interest. Applications of this type are leading to much discussion on privacy issues. In Analytics in Action 10.5, we learn about a more benevolent application of image analytics where the images are captured by a phone and a mobile application supplies immediate value to the user of the app.

ANALYTICS IN ACTION 10.5 Image Analysis Helps Estimate Plant Cover

Estimating how much ground is covered by green vegetation is important in analysis of a forest or even a farm. In case of a forest, such analysis helps users understand how the forest is evolving, its impact on surrounding areas, and even climate. For a farm, similar analysis can help understand plant growth and help estimate future crop yields. It is obviously impossible to measure all forest cover manually and is challenging for a farm. The common method is to record images of a forest/farm and then analyze these images to estimate the ground cover. Such analysis is expensive to perform visually and is also error prone. Different experts looking at the ground cover might estimate the percentage of ground covering differently. Thus, automated methods to analyze these images and estimate the percentage of ground covered by vegetation are being developed. One such method and an app to make it practical through a mobile phone has been developed at Oklahoma State University by researchers in the Department of Plant and Soil Sciences in partnership with the university's App Center and the Information Technology group within the Division of Agricultural Sciences and Natural Resources.

Canopeo is a free desktop or mobile app that estimates green canopy cover in near real time from images taken with a smartphone or digital camera. In experiments in corn, wheat, canola, and other crops, Canopeo calculated the percentage of canopy covering dozens to thousands of times faster than existing software without sacrificing accuracy. And unlike other programs, the app can get and analyze video images, says Oklahoma State University (OSU) soil physicist, Tyson Ochsner—a feature that should reduce the sampling error associated with canopy cover estimates. “We know that plant cover, plant canopies, can be quite variable in space,” says Ochsner, who led the app's development with former doctoral student Andres Patrignani, now a faculty member at Kansas State University. “With Canopeo, you can just turn on your [video] device, start walking across a portion of a field, and get results for every frame of video that you're recording.” By using a smartphone or tablet's digital camera, Canopeo users in the field can take photos or videos of green plants, including crops, forages, and turf, and import them to the app, which analyses each image pixel, classifying them based on its red-green-blue (RGB) color values. Canopeo analyses pixels based on a ratio of red to green and blue to green pixels as well as an excess green index. The result is an image where color pixels are converted into black and white with white pixels corresponding to green canopy and black pixels representing the background. Comparison tests showed that Canopeo analyses images more quickly and just as accurately as two other available software packages.

Developers of Canopeo expect the app to help producers judge when to remove grazing cattle from winter wheat in “dual-purpose” systems where wheat is also harvested for grain. Research

by others at OSU found that taking cattle off fields when at least 60% green canopy cover remained ensured a good grain yield. “So, Canopeo would be useful for that decision,” Patrignani says. He and Ochsner also think the app could find use in turfgrass management; in assessments of crop damage from weather or herbicide drift; as a surrogate for the Normalized Difference Vegetation Index (NDVI) in fertilizer recommendations; and even in UAV-based photos of forests or aquatic systems.

Analysis of images is a growing application area for deep learning as well as many other AI (Artificial Intelligence) techniques.

QUESTIONS FOR DISCUSSION

1. What is the purpose of knowing how much ground is covered by green foliage on a farm? In a forest?
2. Why would image analysis of foliage through an app be better than a visual check?
3. Explore research papers to understand the underlying algorithmic logic of image analysis. What did you learn?
4. What other applications of image analysis can you think of?

Source: Compiled from Patrignani, A., & Ochsner, T. E. (2015). “Canopeo: A Powerful New Tool for Measuring Fractional Green Canopy Cover.” *Agronomy Journal*, 107(6), pp. 2312–2320; Lollato, R., Patrignani, A., Ochsner, T. E., Rocatelli, A., Tomlinson, P. & Edwards, J. T. (2015). Improving Grazing Management Using a Smartphone App. www.bookstore.ksre.ksu.edu/pubs/MF3304.pdf (accessed August 2022); <http://canopeoapp.com/> (accessed August 2022); Oklahoma State University press releases.

ANALYTICS IN ACTION 10.6 How Unilever Used Image Analytics in Its Supply Chain to Examine Deforestation

Unilever is using both satellite imagery and geospatial location data to understand and get information about the source location of its raw materials used in manufacturing of its products, for example, Axe deodorant.

Marc Engel, chief supply chain officer, states that supply chains are traditionally long and are not transparent. For instance, palm oil, an important raw material in many daily products from biscuits to shampoos, is in limelight due to risk of deforestation as more trees in the forest are being cleared to plant more palm plantations. Unilever said that it is planning to achieve a deforestation-free supply chain by 2023.

Unilever is working with **Orbital Insight**, a software company to trace the crop’s journey from its origin to the manufacturing company, which is known as the “first mile” in supply chain. Orbital Insight is using cell phone data to connect the tracks from one location to another to form a map of the supply chain. Combining this data with advanced satellite imagery is being used to arrive at an estimate of deforestation. This is helping Unilever in alerting its suppliers about the source locations or farms.

Orbital Insight developed algorithms and trained them to get more information about the differences between managed forest and the actual deforestation. This technology is being used in various places around the world like Sumatra to check the palm oil plantation and in Brazil to look after soya bean production.

The story linked below also highlights another technology platform—Provenance. Its CEO and founder Jessi Baker points out that many brands now want to share the origin and the supply chain journey of a product to the customers. This can help customers/stakeholders be aware of information such as carbon footprint, and animal welfare.

QUESTIONS FOR DISCUSSION

1. How can the satellite imagery be used to understand deforestation threat?
2. Go to [orbitalinsights.com](https://www.orbitalinsights.com) website and identify other similar applications of imagery.

Source: <https://www.cnbc.com/2020/09/24/unilever-in-data-pilot-to-check-for-deforestation-in-its-supply-chain.html> (accessed August 2022).

► SECTION 10.4 REVIEW QUESTIONS

1. How can images supplement other data sources for providing information to decision makers?
2. Define alternative data.
3. Find other recent examples of applications of image data in decision making.
4. What would be some of the concerns in using image data?

10.5 IoT ESSENTIALS

The **Internet of Things (IoT)** is an evolving term with several definitions. In general, IoT refers to a computerized network that connects many objects (people, animals, devices, sensors, buildings, items) each with an embedded microprocessor. The objects are connected, mostly wirelessly, to the Internet forming the IoT. The IoT can exchange data and allow communication among the objects and with their environments. That is, the IoT allows people and things to be interconnected anytime and anyplace. Embedded sensors that collect and exchange data make up a major portion of the objects and the IoT. That is, IoT uses **ubiquitous computing**. Analysts predict that by the year 2025, more than 50 billion devices (objects) will be connected to the Internet, creating the backbone of IoT applications.

Embedding computers and other devices that can be switched on and off into active items anywhere and connecting all devices to the Internet (and/or to each other) permit extensive communication and collaboration between users and items. By connecting many devices that can talk to each other, one can create applications with new functionalities, increase the productivity of existing systems, and drive the benefits discussed later. This kind of interaction opens the door to many applications. Check the “Internet of Things Consortium” (iofthings.org) and its annual conferences. For an infographic and a guide, see intel.com/content/www/us/en/internet-of-things/infographics/guide-to-iot.html (accessed August 2022).

Definitions and Characteristics

There are several definitions of IoT. Kevin Ashton, who is credited with the term the “Internet of Things,” provided the following definition: “The Internet of Things means sensors connected to the Internet and behaving in an Internet-like way by making open, ad hoc connections, sharing data freely, and allowing unexpected applications, so computers can understand the world around them and become humanity’s nervous system” (term delivered first in a 1999 oral presentation. See Ashton, 2009).

Our working definition is:

The IoT is a network of connected computing devices including different types of objects (e.g., digital machines). Each object in the network has a **unique identifier (UID)**, and it is capable of collecting and transferring data automatically across the network. The collected data has no value until it is analyzed, as illustrated in the opening vignette. Note that the IoT allows people and things to interact and communicate at any time, any place, regarding any business topic or service.

According to Miller (2015), the IoT is a connected network in which:

- Large numbers of objects (things) can be connected.
- Each thing has a unique definition (IP address).
- Each thing has the ability to receive, send, and store data automatically.

- Each thing is delivered mostly over the wireless Internet.
- Each thing is built upon machine-to-machine (M2M) communication.

Note that, in contrast with the regular Internet that connects people to each other using computing technology, the IoT connects “things” (physical devices and people) to each other and to sensors that collect data. In a later subsection, we explain the process of IoT.

SIMPLE EXAMPLES A common example of the IoT is the autonomous vehicle. To drive on its own, a vehicle needs to have enough sensors that automatically monitor the situation around the car and take appropriate actions whenever necessary to adjust any setting, including the car’s speed, direction, and so on. Another example that illustrates the IoT phenomenon is the company **Smartbin**. It has developed trash containers that include sensors to detect their fill levels. The trash collection company is automatically notified to empty a trash container when the sensor detects that the bin has reached the fill level.

A common example people give to illustrate IoT is the idea that a refrigerator could automatically order food (e.g., milk) when it detects that the food has run out! Clorox introduced a new Brita filter so that a Wi-Fi-enabled mechanism can order water filters by itself when it detects that it is time to change them. In these examples, a human does not have to communicate with another human or even with a machine.

IoT IS CHANGING EVERYTHING According to McCafferty (2015), the IoT is changing everything. For how manufacturing is revolutionized by IoT, see Greengard (2016). Here are a few examples that were provided:

- “Real-time systems make it possible to know where anyone is at any moment, which is helpful to secured locations as military bases and seeking to push promotions to consumers.”
- “Fleet tracking systems allow logistics and transport firms to optimize routing, track vehicle speeds and locations, and analyze driver and route efficiencies.”
- “Owners and operators of jet engines, trains, factory equipment, bridges, tunnels, etc., can stay ahead of repairs through machines that monitor for preventive maintenance.” (opening case)
- “Manufacturers of foods, pharmaceuticals and other products monitor temperature, humidity and other variables to manage quality control, receiving instant alerts when something goes wrong.”

These changes are facilitated by AI systems, which enhance analytics and automate or support decision making.

The IoT Ecosystem

When billions of things are connected to the Internet with all the supporting services and connected IT infrastructure, we can see a giant complex, which can be viewed as a huge ecosystem. The **Internet of Things ecosystem** refers to all components that enable users to create IoT applications. These components include gateways, analytics, AI algorithms, servers, data storage, security, and connectivity devices. A pictorial view is provided in Figure 10.5. The figure is dated 2016, but it still presents a good overview of the group of technologies and applications in IoT. An example of an IoT application is provided in the opening vignette. It illustrates a collection of sensors that collects information, which is transmitted to a central place for processing and eventually for decision support. Thus, the IoT applications are subsets of the IoT ecosystem.

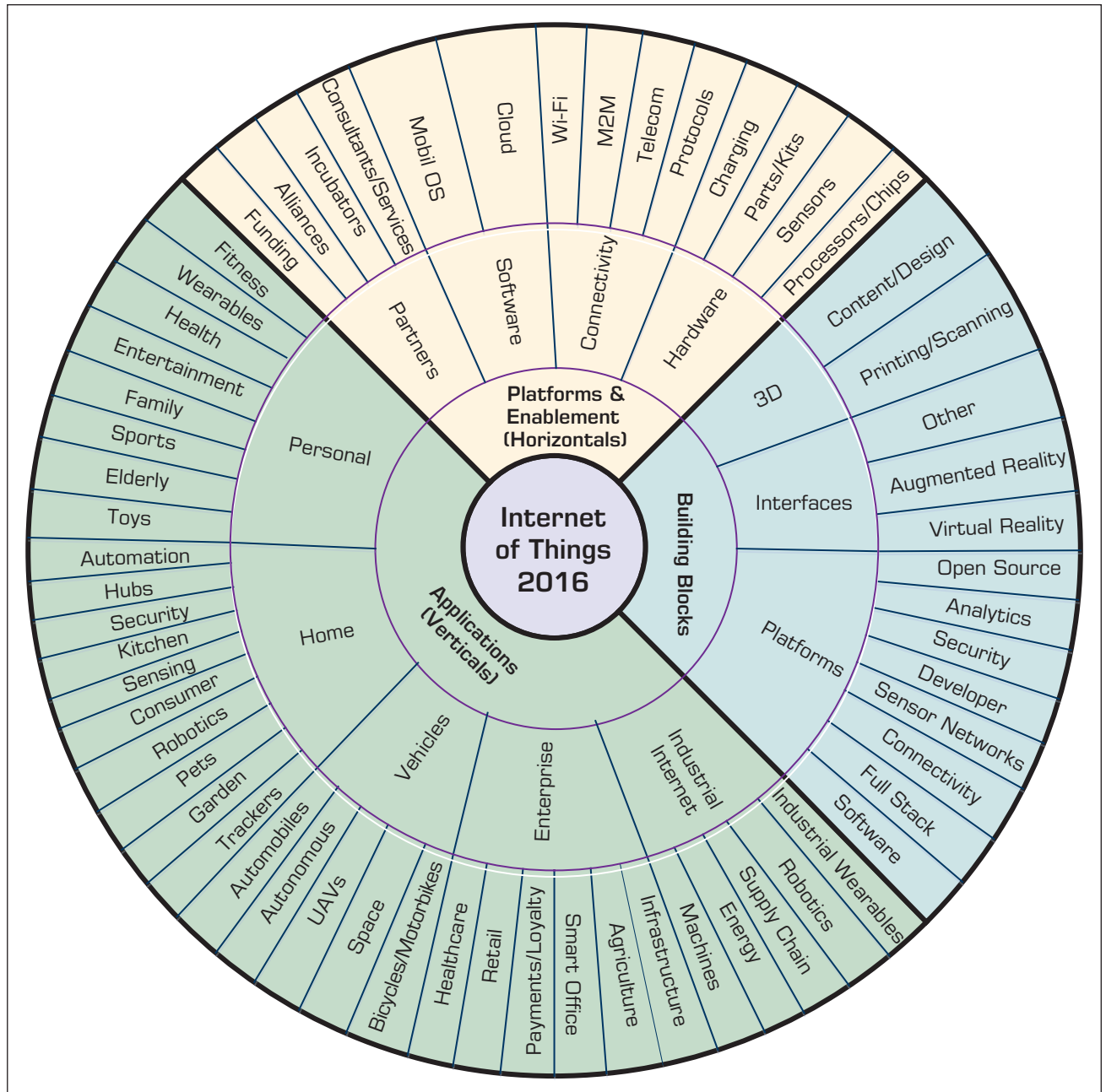


FIGURE 10.5 An Illustrative IoT Ecosystem.

Structure of IoT Systems

Things in IoT refers to a variety of objects and devices ranging from cars and home appliances to medical devices, computers, fitness tracers, hardware, software, data, sensors, and much more. Connecting things and allowing them to communicate is a necessary capability of an IoT application; but for more sophisticated applications, we need additional components: a control system and a business model. The IoT enables the things to sense or be sensed wirelessly across the network. A non-Internet example is a temperature control system in a room. Another non-Internet example is a traffic signal at intersections of

roads where camera sensors recognize the cars coming from each direction and a control system adjusts the time for changing the lights according to programmed rules. Later, we will introduce the reader to many Internet-based applications.

IoT TECHNOLOGY INFRASTRUCTURE From a bird’s-eye view, IoT technology can be divided into four major blocks. Figure 10.6 illustrates these.

- 1. **Hardware:** This includes the physical devices, sensors, and actuators where data are produced and recorded. The devices are the equipment that needs to be controlled, monitored, or tracked. IoT sensor devices could contain a processor or any computing device that parses incoming data.
- 2. **Connectivity:** There should be a base station or hub that collects data from the sensor-laden objects and sends those data to the “cloud” to be analyzed. Devices are connected to a network to communicate with other networks or other applications. These may be directly connected to the Internet. A gateway enables devices that are not directly connected to the Internet to reach the cloud platform.

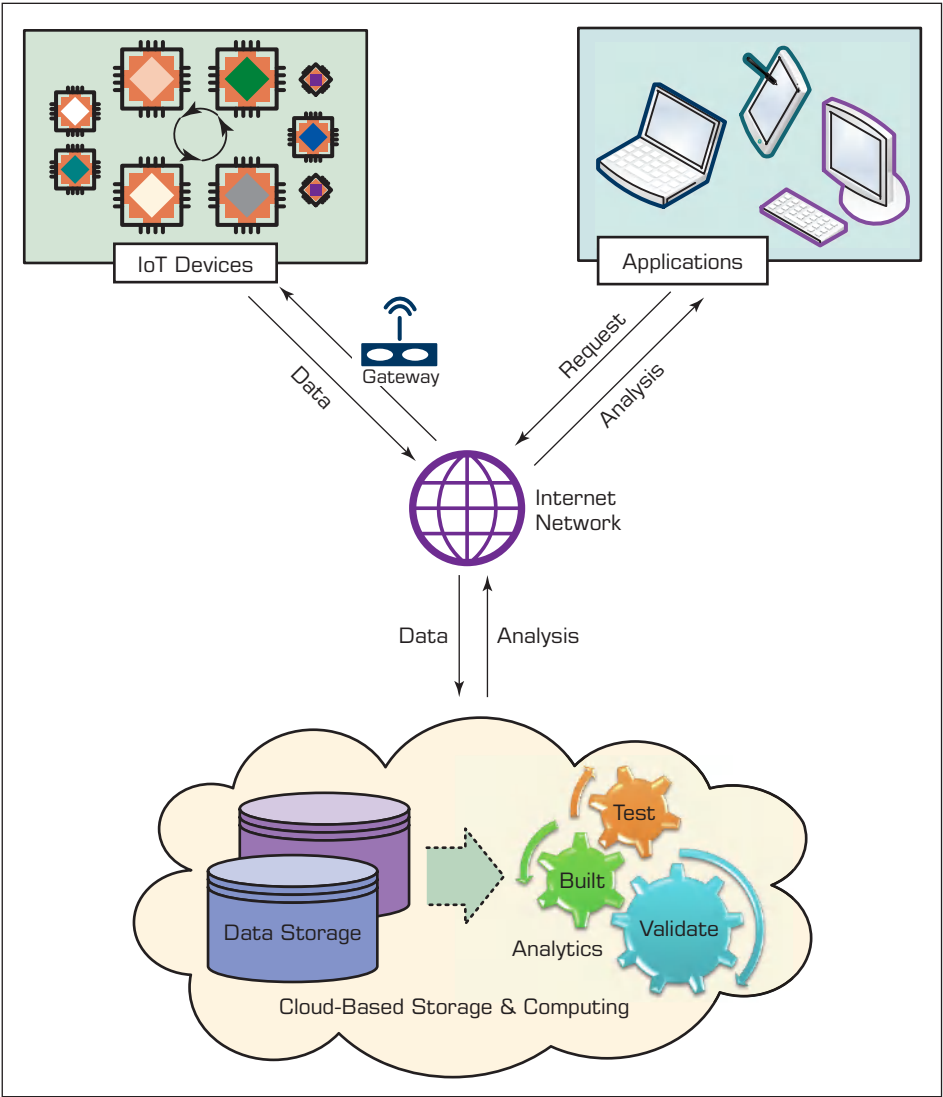


FIGURE 10.6 The Building Blocks of IoT.

3. **Software backend:** In this layer, the data collected are managed. Software backend manages connected networks and devices and provides data integration. This may very well be in the cloud.
4. **Applications:** In this part of IoT, data are turned into meaningful information. Many of the applications can run on smartphones, tablets, and PCs and do something useful with the data. Other applications can run on the server and provide results or alerts through dashboards or messages to the stakeholders.

To assist with the construction of IoT systems, one may use IoT platforms. For information, see Meola (2018).

IoT PLATFORMS Because IoT is still evolving, many domain-specific and application-specific technology platforms are also evolving. Not surprisingly, many of the major vendors of IoT platforms are the same ones who provide analytics and data storage services for other application domains. These include Amazon AWS IoT, Microsoft Azure IoT suite, Predix IoT Platform by General Electric (GE), and IBM Watson IoT platform (ibm.com/us-en/marketplace/internet-of-things-cloud) (accessed August 2022). Teradata Unified Data Architecture has similarly been applied by many customers in the IoT domain.

MAJOR BENEFITS AND DRIVERS OF IoT

The major objective of IoT systems is to improve productivity, quality, speed, and the quality of life. There are potentially several major benefits from IoT, especially when combined with AI.

Major Benefits of IoT

The following are the major benefits of IoT:

- Reduces cost by automating processes.
- Improves workers' productivity.
- Creates new revenue streams.
- Optimizes asset utilization (e.g., see the opening vignette).
- Improves sustainability.
- Changes and improves everything.
- May anticipate our needs (predictions).
- Enables insights into broad environments (sensors collect data).
- Enables smarter decisions/purchases.
- Provides increased accuracy of predictions.
- Identifies problems quickly (even before they occur).
- Provides instant information generation and dissemination.
- Offers quick and inexpensive tracking of activities.
- Makes business processes more efficient.
- Enables communication between consumers and financial institutions.
- Facilitates growth strategy.
- Fundamentally improves the use of analytics (see the opening vignette).
- Enables better decision making based on real-time information.
- Expedites problem resolution and malfunction recovery.
- Supports facility integration.
- Provides better knowledge about customers for personalized services and marketing.

Major Drivers of IoT

The following are the major drivers of IoT:

- The number of “things”—20 to 50 billion—may be connected to the Internet by 2020–2025.

- Connected autonomous “things”/systems (e.g., robots, cars) create new IoT applications.
- Broadband Internet is more widely available, increasing with time.
- The cost of devices and sensors is continuously declining.
- The cost of connecting the devices is decreasing.
- Additional devices are created (via innovations) and are interconnected easily.
- More sensors are built into devices.
- Smartphones’ penetration is skyrocketing.
- The availability of wearable devices is increasing.
- The speed of moving data is increasing to 60 THz.
- Protocols are developing for IoT (e.g., WiGig).
- Customer expectations are rising; innovative customer services are becoming a necessity.
- The availability of IoT tools and platforms is increasing.
- The availability of powerful analytics that are used with IoT is increasing.

How IoT Works

IoT is not an application. It is an infrastructure, platform, or framework that is used to support applications. The following is a comprehensive process for IoT applications. In many cases, IoT follows only portions of this process.

The process is explained in Figure 10.7. The Internet ecosystem (top of the figure) includes a large number of things. Sensors and other devices collect information from the ecosystem. The collected information can be displayed, stored, and processed analytically (e.g., by data mining). This analysis converts the information into knowledge and/or intelligence. Expert systems or machine learning may help in turning the knowledge into decision support (made by people and/or machines), which is evidenced by improved actions and results.

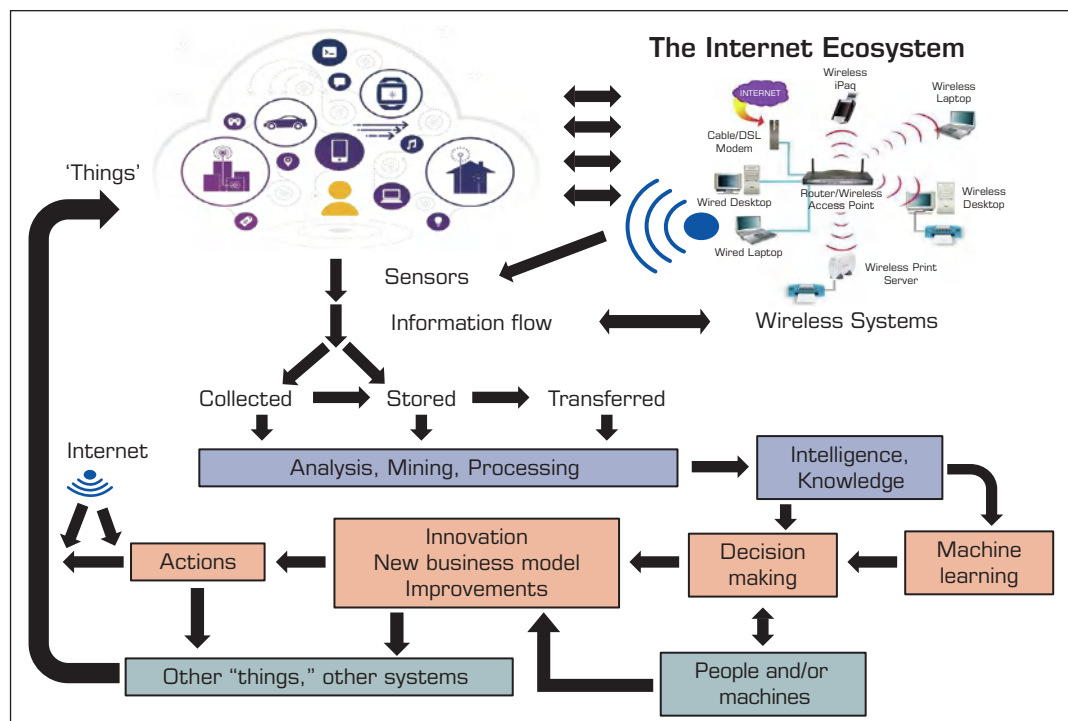


FIGURE 10.7 The Process of IoT.

The generated decisions can help in creating innovative applications, new business models, and improvements in business processes. These result in “actions,” which may impact the original scenario or other things. The opening vignette illustrates this process.

Note that most of the existing applications are in the upper part of the figure, which is called sensor to insight, meaning up to the creation of knowledge or to the delivery of new information. However, now, the focus is moving to the entire cycle (i.e., sensor to action).

The IoT may generate a huge amount of data (Big Data) that needs to be analyzed by various business intelligence methods, including deep learning, or advanced AI methods.

IoT and Decision Support

As stated earlier, the IoT creates knowledge and/or intelligence, which is submitted as support to decision makers or is inputted to automated decision support entities. The transition from data collection to decision support may not be simple due to the large amount of data, some of which are irrelevant. Large-scale IoT usually needs to filter the collected data and “clean” them before they can be used for decision support, particularly if they are used as a base for automated decision making.

Sensors and Their Role in IoT

As illustrated in the opening vignette to this chapter, sensors play a major role in IoT by collecting data about the performance of the things that are connected to the Internet and monitoring the surrounding environment, collecting data there too if necessary. Sensors can transmit data and sometimes even process it prior to transmission.

Brief Introduction to Sensor Technology

A **sensor** is an electronic device that automatically collects data about events or changes in its environment. Many IoT applications include sensors (see the opening vignette). The collected data are sent to other electronic devices for processing. There are several types of sensors and several methods for collecting data. Sensors often generate signals that are converted to human-readable displays. In addition to their use in IoT, sensors are essential components in robotics and autonomous vehicles. Each sensor usually has a limit on the maximum distance that it can detect (nominal range). Sensors of a very short range known as *proximity sensors* are more reliable than those that operate in larger ranges. Each IoT network may have millions of sensors.

How Sensors Work with IoT

In large-scale applications, sensors collect data that are transferred to processing in the “cloud.” Several platforms are used for this process as discussed in Analytics in Action 10.7.

ANALYTICS IN ACTION 10.7 Rockwell Automation Monitors Expensive Oil and Gas Exploration Assets to Predict Failures

Rockwell Automation is one of the world’s largest providers of industrial automation and information solutions. It has customers in more than 80 countries worldwide and around 22,500 employees. One of its business areas of focus is assisting oil and gas companies in exploration. An example is Hilcorp Energy, a customer company that drills oil in Alaska. The equipment used in drilling, extracting, and refining oil is very expensive. A single fault in the equipment can cost the company around \$100,000–\$300,000 per day in lost production. To deal with this problem, it needed technology to monitor the status of such piece of equipment remotely and to predict failures that are likely to happen in the future.

Rockwell Automation considered the opportunity to expand its business in oil and gas industries by gathering data from the exploration sites and analyzing them to improve preventive maintenance decision making regarding the critical equipment, thus minimizing downtime and drive better performance. The company utilizes its vision of Connected Enterprise with Microsoft's software to monitor and support oil and gas equipment placed in remote areas. Rockwell is now providing solutions to predict failure of equipment along the entire petroleum supply chain, monitoring its health and performance in real time, and to prevent failures in the future. Solutions are provided in the following areas.

- **Drilling:** Hilcorp Energy has its pumping equipment stationed in Alaska where it drills for oil 24 hours a day. A single failure in equipment can cost Hilcorp a large amount of money. Rockwell connected electrical variable drives of pumping equipment to be processed in the “cloud,” to control its machines thousands of miles away from the control room in Ohio. Sensors capture data, and through Rockwell's control gateway, these data are passed to Microsoft Azure Cloud. The solutions derived reach Hilcorp engineers through digital dashboards that provide real-time information about pressure, temperature, flow rate, and dozens of other parameters that help engineers monitor the equipment's health and performance. These dashboards also display alerts about any possible issues. When one of Hilcorp's pieces of pumping equipment failed, it was identified, tracked, and repaired in less than an hour, saving six hours of tracing the failure and the large cost of lost production.
- **Building smarter gas pumps:** Today, some delivery trucks use liquid natural gas (LNG) as fuel. Oil companies are updating their filling stations to incorporate LNG pumps. Rockwell Automation installed sensors and variable frequency drives at these pumps to collect real-time data about equipment operations, fuel inventory, and consumption rate. This data are transmitted to Rockwell's cloud platform for processing. Rockwell then generates interactive dashboards and reports using Microsoft Azure (an IoT platform). Results are forwarded to the appropriate stakeholders, giving them a good idea about the health of their capital assets.

The Connected Enterprise solution by Rockwell has accelerated growth for many oil and gas companies like Hilcorp Energy by bringing their operations data to the cloud platform and helping them reduce costly downtime and maintenance. It has resulted in a new business opportunity for industrial age stalwarts like Rockwell Automation.

QUESTIONS FOR DISCUSSION

1. What type of information would likely be collected by an oil and gas drilling platform?
2. Does this application fit the three Vs (volume, variety, velocity) of Big Data? Why or why not?
3. Which other industries (list five) could use similar operational measurements and dashboards?

Sources: customers.microsoft.com (2015); Rockwell Automation: Fueling the Oil and Gas Industry with IoT; <https://customers.microsoft.com/Pages/CustomerStory.aspx?recid=19922>; Microsoft.com. (n.d.). “Customer Stories | Rockwell Automation,” <https://www.microsoft.com/en-us/cloud-platform/customer-stories-rockwell-automation> (accessed August 2022).

Sensor Applications and Radio-Frequency Identification (RFID) Sensors

There are many types of sensors. Some measure temperature; others measure humidity. Many sensors collect information and transmit it as well. For a list of 50 sensor applications with a large number of related articles, see https://www.libelium.com/libeliumworld/top_50_iot_sensor_applications_ranking/ (accessed August 2022).

A well-known type of sensor that plays an important role in IoT is radio-frequency identification.

RFID SENSORS Radio-frequency identification (RFID) is part of a broader ecosystem of data capture technologies. Several forms of RFID in conjunction with other sensors play a major role in IoT applications. Let us see first what RFID is, as discussed in Technology Insights 10.1.

TECHNOLOGY INSIGHTS 10.1 RFID Sensors

RFID is a generic technology that refers to the use of radio-frequency waves to identify objects. Fundamentally, RFID is one example of a family of automatic identification technologies that also includes ubiquitous barcodes and magnetic strips. Since the mid-1970s, the retail supply chain (among many other areas) has used barcodes as the primary form of automatic identification. RFIDs can store a much larger amount of data than barcodes. Also, they can be accessed from a longer distance wirelessly. These potential advantages of RFID have prompted many companies (led by large retailers such as Walmart and Target) to aggressively pursue it as a way to improve their *supply chains* and thus reduce costs and increase sales.

How does an RFID work? In its simplest form, an RFID system consists of a tag (attached to the product to be identified), an interrogator (i.e., RFID reader), one or several antennae attached to the reader, and a computer program (to control the reader and capture the data). At present, the retail supply chain has primarily been interested in using passive RFID tags. Passive tags receive energy from the electromagnetic field created by the interrogator (e.g., a reader) and backscatter information only when it is requested. The passive tag remains energized only while it is within the interrogator's magnetic field.

In contrast, active tags have a battery to energize themselves. Because active tags have their own power source, they do not need a reader to energize them; instead, they can initiate the data transmission process on their own. As compared to passive tags, active tags have a longer read range, better accuracy, more complex rewritable information storage, and richer processing capabilities. On the negative side, their batteries cause active tags to have a limited life span, be larger in size than passive tags, and be more expensive. Currently, most retail applications are designed and operated with passive tags, each of which costs only a few cents. Active tags are most frequently found in defense and military systems, yet they also appear in technologies such as EZ Pass whose tags (called transponders) are linked to a prepaid account that, for example, enables drivers to pay tolls later, by driving past a reader rather than stopping to pay at a tollbooth. There are also semipassive tags with limited active tag capabilities.

The most commonly used data representation for RFID technology is the Electronic Product Code (EPC), which is viewed by many in the industry as the next generation of the Universal Product Code (UPC), most often represented by a barcode. Like the UPC, the EPC consists of a series of numbers that identifies product types and manufacturers across the supply chain. The EPC also includes an extra set of digits to uniquely identify items.

Use of RFID and Smart Sensors in IoT

Basic RFID tags, either active or passive, are not sensors. The purpose of the tags is to identify objects and determine their location (e.g., for the purpose of counting objects). To make them useful for most IoT applications, the tags need to be upgraded (e.g., by adding on-board sensors). These RFIDs called *RFID sensors* have more capabilities than RFID tags, or basic sensors. For a detailed discussion about the role of RFID in the IoT, see Donaldson (2017).

RFID sensors are wireless sensors that communicate, via mesh networks or conventional RFID readers, and they include identifiable ID. The RFID reader sends token information into gateways, such as AWS IoT service. This confirmation can be processed, resulting in some action.

SMART SENSORS AND IoT There are several types of smart sensors with different levels of capabilities when integrated into IoT. A **smart sensor** is one that senses the environment and processes the input it collects by using its built-in computing capabilities (e.g., a microprocessing). The processing is preprogrammed. Results are passed

on. Depending on the internal computing quality, smart sensors can be more automated and accurate than other sensors and can filter out unwanted noise and compensate for errors before sending the data.

Smart sensors are crucial and an integral element in the IoT. They can include special components, such as amplifiers, analog filters, and transducers, to support IoT. In addition, smart sensors for IoT can include special software for data conversion, digital processing, and communication capability to external devices.

SECTION 10.5 REVIEW QUESTIONS

1. What is IoT?
2. List the major characteristics of IoT.
3. What are the major components of an IoT technology?
4. List the major drivers of IoT.
5. Explain how the IoT works following the process illustrated in Figure 10.7.
6. How does IoT support decision making?
7. Describe the role of sensors in IoT.
8. What is RFID? What is a RFID sensor?
9. What role does the RFID perform in IoT?
10. Define *smart sensor* and describe its role in IoT.

10.6 IoT APPLICATIONS

We start with a well-known example: Imagine that your refrigerator can sense the amount of food in it and send you a text message when inventory is low (sensor-to-insight in Figure 10.8). One day refrigerators will also be able to place an order for items that need replenishment, pay for them, and arrange delivery (sensor-to-action). Let us look at some other, less futuristic enterprise applications.



FIGURE 10.8 The Components of a Smart Home.

Example French National Railway System's Use of IoT

SNCF, the French national railway system, uses IoT to provide quality, availability, and safety for its nearly 14 million passengers. The company sncf.com improved its operations using IoT (Estopace, 2017). To manage 15,000 trains and 30,000 kilometers of tracks is not simple, but IBM Watson, using IoT and analytics, helped to do just that. Thousands of sensors that are installed on the trains, tracks, and train stations gather data that Watson processes. In addition, all business process operations were digitized to fit into the system. Information concerning possible cyberattacks was also programmed into the system. All collected Big Data were prepared for decision support. IBM Watson's platform is scalable and can handle future expansions.

To understand the magnitude of this IoT network, consider that the mass transit lines in Paris alone required 2,000 sensors forwarding information from more than 7,000 data points each month. The systems enable engineers to remotely monitor 200 trains at a time for any mechanical and electrical operations and malfunctions while trains are moving. In addition, by using a predictive analytic model, the company can schedule preventive maintenance to minimize failures. Therefore, if you are one of the train travelers, you can relax and enjoy your trip.

Of all the consumer related IoT initiatives, three types are most well-known: smart homes and appliances, smart cities, and autonomous vehicles. We briefly introduce each of these below.

Smart Homes and Appliances

The concept of the smart home has been in the limelight for several years, even before the concept of the IoT took a front stage. A **smart home** is a home with automated components that are interconnected (frequently wirelessly), such as appliances, security, lights, and entertainment, and are centrally controlled and able to communicate with each other. For a description, see techterms.com/definition/smart_home (accessed August 2022).

Smart homes are designed to provide their dwellers with comfort, security, low energy cost, and convenience. They can communicate via smartphones or the Internet. The control can be in real time or at any desired intervals. Most existing homes are not yet smart, but they can easily and inexpensively be equipped to for at least partial smartness. Several protocols enable connections; well-known ones are XIO, UPB, Z-Wave, and EnOcean. These products offer scalability, so more devices can be connected to the smart home over time. In the United States followed by other countries, thousands of homes are already equipped with such systems.

Typical Components of Smart Homes

The following are typical components in smart homes:

- **Lighting.** Users can manage their home lighting from wherever they are.
- **TV.** This is the most popular component.
- **Energy management.** Home heating and cooling systems can be fully automated and controlled via a smart thermostat (e.g., see Nestnest.com/works-with-nest (accessed August 2022) about its product Nest Learning Thermostat).
- **Water control.** WaterCop (watercop.com) is a system that reduces water damage by monitoring water leaks via a sensor. The system sends a signal to a valve, causing it to close.
- **Smart speaker and chatbots.** Most popular are Echo and Alexa, and Google Assistant.
- **Home entertainment.** Audio and video equipment can be programmed to respond to a remote-control device. For instance, a Wi-Fi-based remote control

for a stereo system located in a family room can command the system to play on speakers installed anywhere else in the house. All home automation devices perform from one remote site and one button.

- **Alarm clock.** This tells everyone to go back to sleep or to wake up.
- **Vacuum cleaner.** Examples are iRobot Roomba, and LG Roboking vacuum
- **Camera.** This allows residents to see what is going on in their homes anytime from anywhere. Nest Cam Indoor is a popular product. Some smart cameras can even know how residents feel.
- **Refrigerator.** An example of this is Instaview from LG, which is powered by Alexa.
- **Home security and safety.** Such systems can be programmed to alert owners to security-related events on their property. As noted, some security can be supported by cameras for remote viewing of property in real time. Sensors can be used at home to detect intruders, keep an eye on working appliances, and perform several additional activities.

The major components of smart homes are illustrated in Figure 10.8. Note that only a few homes have all of these components. Most common are home security, entertainment, and energy management.

Smart Appliances

A **smart appliance** includes features that can remotely control the appliance operations, based on the user preferences. A *smart appliance* may utilize a *Home Network* or the Internet to communicate with other devices in the smart home.

Some common products that help make a home a smart home include:

- **Learning thermostat.** This device learns what temperature and humidity level that people like and controls the air conditioner/heating system accordingly. Google claims that its products provide an average energy savings of 13%, which could pay for the device in two years.
- **Smoke detector and alarm.** This device, which is controlled from a smartphone, tests itself automatically and lasts for about a decade.
- Webcam-based systems allow users to see what is going on in their homes from any location via smartphone or any desktop computer. The system can turn itself on automatically when nobody is at home. It can monitor pets, babies, and so on. A photo recorder allows users to go back in time.
- **Samsung Smart fridge.** Cameras check content; sensors check temperature and humidity.
- **10 in 1 device for the kitchen.** This stirs food such as scrambled eggs and has 10 cooking styles (e.g., baking, sauce making).
- **Instaview Refrigerator from LG.** Powered by Alexa (enabled by voice), this includes a 29-inch LCD touch screen display. It provides functions such as determining the expiration dates of food and notifying the user.
- **Whirlpool's smart top load washer.** This fully automated machine has smart controls. It saves energy and even encourages philanthropy by sending a small amount of money to "Habitat for Humanity" each time washer is loaded.
- TVs that can be used as a smart Hub for home appliances
- DIY home smart security cameras make sure there is an intruder, not just the cat, before alerting the police.
- Water controls for faucets, sprinklers, and flood detectors are available. In addition, sensors can detect water leakage and alert the homeowner.

Smart components for the home are available at home improvement stores (e.g., Lowes) and can be purchased directly from manufacturers (e.g., Nest).

A Smart Home Is Where the Bot Is

The virtual personal assistant that we introduced in Chapter 2 enables people to converse by voice with chatbots such as Alexa/Echo and Google Assistant. Such assistants can be used to manage appliances in smart homes.

In a comprehensive smart home, devices not only meet household needs but also are able to anticipate them. It is predicted that in the near future, an AI-based smart home will feature an intelligent and coordinated ecosystem of bots that will manage and perform household tasks and may even be emotionally connected with people. Smart homes will also have smart robots that can serve people snacks, provide support for people with disabilities, and even teach children different skills.

Barriers to Smart Home Adoption

The potential of smart homes is attractive, but it will take some time before there will be many of them. The following are some limiting barriers, per Venkatakrishnan (2017).

- **Compatibility.** There are too many products and vendors to choose from, making potential buyers confused. Many of these products do not “speak” to each other, so more industry standards are needed. In addition, it is difficult to match the products with consumers’ needs.
- **Communication.** Different consumers have different ideas on what the smart home should be. Therefore, the capabilities and benefits of a smart home need to be clearly communicated to users.
- **Concentration.** Brands need to concentrate on population segments that are most interested in smart homes

In addition are the issues of cost justification, invasion of privacy, security, and ease of use.

Smart homes, appliances, and buildings can be featured in smart cities, the subject of our next section.

Smart Components in Smart Cities and Smart Factories

The major objective of smart cities is to automate as many as possible public services such as transportation, utilities, social services, security, medical care, education, and economy. So, in the smart city overall project one may find several subprojects, some of which are independent of the master project.

Hong Kong has a project called a *smart mobility* for the improvement of road safety. A consortium of private and public organizations has introduced Intelligent Transport Services, including a warning mechanism for collision, and control assistance for finding parking. The system also manages speed and lane violations and traffic congestion. All of these increase safety and efficiency.

Transportation is a major area in which analytics and AI can make cities smarter. Other areas include economic development, crime fighting, and healthcare.

Other examples of smart city components can be found in a smart university, smart medical centers, smart power grid, and in airports, factories, ports, sport arenas, and smart factories. Each of these components can be treated as an independent IoT project, and/or as a part of the smart city overall project.

SMART (DIGITAL) FACTORIES Automation of manufacturing has been with us for generations. Robots are making thousands of products from cars to cellphones. Tens of thousands of robots can be found in Amazon’s distribution centers. Therefore, it is not surprising that factories are getting smarter with AI technologies and IoT applications. As such they may be considered a component of smart cities and may be interrelated with other components, such as clean air and transportation.

A **smart factory**, according to Deloitte University Press, is “a flexible system that can self-optimize performance across a broader network, self-adapt to and learn from new conditions in real or near real time, and autonomously run entire production processes.” For details, see the free Deloitte e-book at https://www2.deloitte.com/content/dam/insights/us/articles/4051_The-smart-factory/DUP_The-smart-factory.pdf (accessed August 2022). For a primer, see <https://www2.deloitte.com/insights/us/en/focus/internet-of-things/technical-primer.html>.

Tomás (2016) provides a vision of what industrial production will look like in the future. It will be essentially fully digitized and connected, fast, and flexible. The major idea is that there will be a command center in a factory equipped with AI technologies. The AI, combined with IoT sensors and information flow, will enable optimal organization and sequencing of business processes. The entire production chain, from raw material suppliers, logistics, and manufacturing to sales, will be connected to IoT systems for planning, coordination, and control. Planning will be based on analytic predictions of demand.

Production processes will be automated as much as possible and wirelessly controlled. Logistics will be provided on demand quickly, and quality control will be automated. IoT combined with sensors will be used for both predictive and preventive maintenance. Some of these elements exist in advanced factories, and more factories will be smarter in the future.

Improving Transportation in a Smart City

A major area of improvement in a smart city is transportation. A major problem in many cities is the increased number of vehicles and the inability to accommodate all of them effectively. Building more roads could add more pollution and lead to traffic jams. Public transportation can help alleviate the problem but may take years to complete. Quick solutions are needed. In some smart cities, innovators have already placed air quality sensors on bicycles and cars. Sensors also are taking data from cars on the roads to help generate data that can be analyzed, and results are transmitted to drivers. Examples of innovative projects are provided in the following examples.

Example 1

Valerann, an Israeli start-up, developed smart road studs to replace the reflective studs of today’s technology. Smart studs can transmit information of what they sense about what is occurring on the roads. Eventually, the studs will be incorporated with autonomous vehicles. The smart studs cost more than reflective studs but have a longer life.

Example 2

Smart Mobility Consortium (Hong Kong) works on mobility in the smart city of Hong Kong. More than 10 million people there use the public and private transportation systems every day. This transportation project includes several smart subsystems for parking, collision warning, and alerts for speeders and lane-changing violators.

Autonomous (Self-Driving) Vehicles

Autonomous vehicles, also known as **driverless cars**, robot-driven cars, self-driving cars, and autonomous cars, are already on the roads in several places. The first commercial autonomous car project was initiated by Google. Tesla is now almost synonymous with electric vehicles that also include the self-driving capability. France, Singapore, China, and several other countries already have these cars and buses on their roads. These cars are electric, and they can create a revolution by reducing emissions, accidents, fatalities, and traffic jams. Any statistics on these cars will be out of date as soon as we print them, so we exclude most such data.

Implementation Issues in Autonomous Vehicles

Autonomous vehicles such as cars, trucks, and buses are already on the roads in several cities worldwide. However, before we will see millions of them on the roads, it will be necessary to deal with several implementation issues. The following are reasons why full commercialization is going to take time:

- The cost of real-time 3D map technologies needs to be reduced and their quality needs to be increased.
- AI software must be nimble, and its capabilities increased. For example, AI needs to deal with many unexpected conditions, including that of the behavior of drivers of other cars.
- Bray (2016) posted an interesting question: “Are customers, automakers and insurers really ready for self-driving cars?” Customers seem to acknowledge that such cars are coming. But they resist boarding one. However, some daring people expect these cars to do a better job than humans in driving.
- The technology needs more research, which is very expensive. One reason is that the many sensors in the cars and on the road need to be improved and their cost reduced.
- The IoT is connecting many objects for autonomous vehicles, including those in clouds. The IoT systems themselves need to be improved. For example, data transmission delays must be eliminated. For more IT/AI generic implementation issues, see Chapter 14.

The Future of the IoT

With the passage of time, we see an increasing number of IoT applications, both external and internal to organizations and enterprises. Because all IoT networks are connected to the Internet, it will be possible to have some of the networks connected to each other, creating larger IoTs. This will create growth and expansion opportunities for many organizations.

AI ENHANCEMENT OF IoT There are several areas of potential development. One area where AI will enhance IoT is in its ecosystem. Many IoT applications are complex and could be improved with machine learning that can provide insights about data. In addition, AI can help in creating devices (“things”) that can self-diagnose problems and even repair them. Another future benefit of AI when combined with IoT is “shaping up to be a symbiotic pairing” (Hupfer, 2016). This pairing can create cognitive systems that are able to deal with and understand data that conventional analytics cannot handle. The AI and IoT combination can create an *embodied cognition* that injects AI capabilities into objects (such as robots and manufacturing machines) to enable the objects to understand their environments and then self-learn and improve their operation. Finally, AI can help the integration of IoT with other IT systems.

► SECTION 10.6 REVIEW QUESTIONS

1. Describe a smart home.
2. What are the benefits of a smart home?
3. List the major smart appliances.
4. List some benefits of a smart city to the residents.
5. What is the role of IoT in smart city initiatives?
6. What is a smart factory?
7. What are self-driving vehicles? How are they related to the IoT?
8. Why are Uber and similar companies interested in self-driving vehicles?
9. What AI technologies are needed to support autonomous vehicles?

10.7 5G TECHNOLOGIES AND IMPACT ON AI

Fifth-generation (5G) mobile technology is being touted as the catalyst for the **Fourth Industrial Revolution** (Davis, 2016). 5G's superior speeds and ultra-low latencies should enable diverse set of breakthrough use cases for consumers and businesses that are not possible with current networks. Key advantages and benefits of 5G technology overall are summarized as follows (i) it provides very high-speed access, even in densely populated areas, (ii) it connects everything, thereby supporting massive IoT and **machine-to-machine (M2M) communications**, (iii) it provides real-time latency, thereby minimizing delays in network response time and enabling entirely new services and applications for vertical industries, (iv) it delivers significant improvements to security and privacy, and (v) it provides a **service-based architecture (SBA)** that delivers increased flexibility and service diversity.

5G depends upon **edge computing**—performing as much of the computation/analytics at the edge rather than back in a central cloud. Without edge computing, 5G applications and services will rely upon connecting through the core network to centralized cloud resources for storage and computing, losing much of the positive impact of the latency reduction enabled by 5G. The basic idea behind edge computing is that by storing/accessing data and performing high-demand processing tasks closer to the network edge—that is, on the network gateway, customer premises or edge devices—network congestion and latency are significantly reduced. These reductions are crucial for latency-critical applications dependent on continuous corrective actions.

Edge computing is complementary to—rather than in competition with—cloud computing. In this way, benefits can be leveraged from both local (on the network gateway, customer premises, or edge devices) and cloud computing. Though edge computing can be used in a 4G environment, the combination with 5G optimizes throughput and latency—to the order of sub-10 milliseconds—thus laying the foundation for previously unimagined latency-dependent wireless solutions.

However, 5G is only one of a set of technologies coming together to kick off a new chapter in the history of telecommunications (Fox et al., 2020). The blending of 5G, edge computing, and artificial intelligence (AI) will impact industries and enable many new enterprise and industrial applications, such as smart factories, agricultural drones, robotic surgery, smart homes, AR/VR shopping, assistive robots, collaborative gaming, smart cities, and autonomous cars. Locating AI close to the edge is vital for applications where near-real-time feedback and optimization are a priority for applications—such as machine control, equipment monitoring, and remote surgery. Faster AI-based decision making and responses at the edge also result in better user experiences—for instance, in the areas of AI-assisted driving and advanced virtual reality applications (Snowdon et al., 2019). In addition, edge computing supports improved security because sensitive data can be

analyzed at the edge and doesn't need to be sent over the core network to a central cloud, in this way limiting the amount of data stored in any one location.

The combination of 5G, network slicing, edge computing, and AI will be a main driver of Industry 4.0 with billions of machines, devices, and sensors just waiting to be wirelessly connected. This will become the backbone of manufacturing and related services in the future. It will be powered by robotics, AI, IoT, 3-D printing, augmented reality (AR)/virtual reality (VR), and cloud technologies, all of which will use 5G technology to allow machine-to-machine communication.

Indeed, manufacturing is expected to be the largest beneficiary of 5G services (STL Partners, 2019). Industry 4.0 is completely changing the concept of manufacturing from massive production to massive customization. A flexible and programmable environment based on high-speed, low-latency wireless 5G connections, the use of edge computing and AI will link machines, processes, robots, and people. This environment sets up the opportunity for more flexible and dynamic production capabilities to meet rapidly evolving market needs and massive customization.

Somisetty (2018) grouped these applications into four categories: smart city, smart healthcare, connected automobile, and Industry 4.0. Since other sections have discussed examples of the first three categories of applications, we only introduce the Industry 4.0 category. By using the data from machines and sensors embedded throughout a supply chain from raw materials to factory systems to logistics supply chain, it is becoming more likely to extract every bit of efficiency in the supply chain. Many of the sensors used in this process may just be the RFID tags that include basic identification data, but by being able to monitor such tags and their movement throughout the supply chain, one can gain tremendous insights into the issues and inefficiencies in the supply process. This may require nothing more than just descriptive analytics and mapping of the data being generated. However, presenting and analyzing massive volumes of data being generated at a high rate creates research opportunities for developers of descriptive analytics to identify the best ways to aggregate and present the data visualizations to the decision makers.

Using such IoT data on machinery in use in a factory and by applying predictive analytics, we can develop efficient predictive maintenance models and procedures (Yan et al., 2017) to replace periodic maintenance of equipment with condition-based maintenance when a component is about to fail and should be serviced prior to any adverse event. Any equipment is taken offline for maintenance only when it is necessary rather than a scheduled maintenance regardless of the condition of the machinery. Developing such predictive applications for virtually every piece of equipment that requires periodic maintenance presents a huge opportunity that is significantly enabled by 5G to allow information on the equipment to be shared in near real time with decision makers or models to initiate timely predictive maintenance procedures before an item fails. Pasqua (2019) illustrated other potential applications of IoT and 5G in factory automation. These include remote control of robots particularly in hazardous situations, worker training for such situations, remote inspections, and management of emergency operations. As noted earlier, development of these use cases requires a combination of the three key technologies sensors, 5G, and appropriate analytics models.

Application of sensors into manufacturing processes enabled by 5G and advanced analytics is illustrated by a project titled Beer Fingerprinting Project, a collaboration of Microsoft, Carlsberg Brewery, Aarhus University, and the Technical University of Denmark. The project develops many beer samples each day, too many to be sampled and tested by human subjects. Instead, it uses sensors that capture the flavor characteristics of the samples, the yeasts used in brewing such beers, and transmits this voluminous data set using fast techniques such as 5G to an AI system where such data can be analyzed. Similar process-based sensor data collection is becoming common in many other industries. A recent application has been analyzing sewage samples through sensors to predict an early onset of COVID-19 infections.

One approach that predates this combination of the IoT, 5G, and AI for analyzing streams of data that have been realized, analyze the data, and propose a specific course of action/response is called Complex Event Processing (CEP). Many such applications were developed and proposed in various industries. See Luckham (2012), for example. The main difference between what was initiated in 1990s versus the current 5G/IoT/AI-enabled technologies is the size and scale of applications that is possible today, also leading to a new genre of research in developing predictive and prescriptive models. Furthermore, as noted earlier, distribution of data at the edge and in the cloud will also lead to new predictive (and possibly predictive) models.

There are two broad opportunities at the interface of 5G and analytics. The first group is development of all the innovative applications and techniques needed to build a new genre of applications to be powered by this combination of 5G, IoT, and AI. Several examples of interesting use cases already under development have been discussed above, but many more such uses cases will emerge.

AI opportunities in industry and consumer applications will also emerge in user acceptance of such rapid information generation. The privacy concerns will be mentioned in the next chapter. Opportunities exist for developing new ways of presenting fast-changing information in a form that helps a decision maker grasp the situation in context without increasing their cognitive load. For example, how often should such fast-changing information from a smart city or smart healthcare application be updated for a user? How should alerts be conveyed? The information also must be customized for delivery on different devices: industry monitors, portable computers, tablets, and/or mobile devices. AI can help with answers to such questions.

► SECTION 10.7 REVIEW QUESTIONS

1. Define 5G.
2. How does communication technology impact analytics?
3. What do we mean by edge computing?
4. Identify some recent applications of analytics deployed where 5G is a key enabler.

10.8 OTHER EMERGING AI TOPICS: ROBOTIC PROCESS AUTOMATION (RPA)

Robotic process automation is one of the major emerging trends in this digital transformation age. According to Nintex (www.nintex.com), **robotic process automation (RPA)** is defined as “a technology which reduces the workloads of the employees by automating the repetitive, high-volume steps in processes.” **Software robots** or “bots” replicate the actions of human workers. By handling the manual work of humans, RPA gives more time for humans to spend on other important activities. RPA can help achieve an increase in productivity, elimination of human errors, lower operational cost, automation of labor-intensive tasks, and faster completion of work, which saves time and money.

According to a Gartner report, the global RPA revenue grew 31% to \$2.4B in 2021, which is 16% above the software market growth rate. The research from Automation Now & Next (<https://www.automationanywhere.com/now-and-next>) (accessed August 2022) shows that the automation is growing significantly. It is reported that 77% of the business leaders are increasing budgets compared to last year and 90% of the business leaders said that automation has helped them overcome supply chain issues and address key employee shortages.

Though RPA is still in its infancy, the market is showing an increasing demand due to the deployment of applications in cloud platforms. The growth of RPA is likely to affect the future jobs of repeated tasks as RPA can handle them. As noted by Field and Murphy

(2020), the **coronavirus pandemic** has accelerated the growth of automation in almost all major industries such as retail, logistics and manufacturing, healthcare and delivery. Piotrowicz (2020) says that automation growth would not have been possible without the availability of wide network coverage and growth of cloud computing platforms that has allowed for rapid scaling, flexible storage availability and on demand computing power. Below are a few Analytics in Action where automation was used during the pandemic.

ANALYTICS IN ACTION 10.8 Monitoring quarantined persons in the state district

A district authority implemented strenuous wellness surveillance of quarantined individuals within the district. District authorities make checks on the quarantined individuals as well as gathering their medical details. For all those quarantined persons who show COVID-19 symptoms, a quick alert is delivered to the dedicated healthcare group. The dedicated healthcare group would reach out to the individuals and perform the required tests, as well as take up the other measures.

The district authority wants to build fast process automation so that the alerts from individuals reporting problems to the COVID-19 call center can be elevated to the responsible units using an automatic method.

Solution

In response to the COVID-19 pandemic, UiPath, a leading enterprise robotic process automation (RPA) software company, is supplying healthcare groups free of charge RPA software program that will accelerate the crucial procedures as well as free up employees so they can respond to issues more quickly.

The district authority produced a Web-based form, which was completed for each quarantined individual by the COVID-19 call center. Healthcare variables are put together in an online form. For individuals reporting signs and symptoms, the UiPath Software automatic robot forwards the data of symptomatic individuals as well as directs them to the correct medical group rather than needing to get it done by hand.

The group at Auxiliobits Technologies Pvt. Ltd. rapidly implemented the automation, by which the UiPath application robot filters and picks up the data of symptomatic individuals as well as directs the data to the respective medical groups and alerts them. This enables the District Authority to follow up with agility on the whole process.

Source: <https://www.uipath.com/resources/covid-automations/covid19-health-tracking-alerting-automation-for-state-government> (accessed August 2022).

ANALYTICS IN ACTION 10.9 Walgreens' HR shared service efficiency increased by 73% by Blue Prism's Digital Workforce

Business Challenges

Walgreens has been implementing changes in its primary HR and payroll methods with cloud-based substitutes, but supporting this large rollout was stretching the current employees thin. The team considered Robotic process automation (RPA) since they could not employ any more individuals to deal with the added labor. The RPA would take over the current labor intensive, transactional function of staff as well as adjustments to the duties of digital workers therefore, freeing bandwidth, and time for the staff.

Curt Burghardt, vice president of HR Shared Systems, and services at Walgreens, led the RPA initiative, removed any expectation he will bring down staff headcount, and also made certain everyone was fully aware of the effort. The interaction with the participants in the beginning as well as involving them to find qualified tasks for RPA was a crucial element for the satisfactory results on the first step, together with dealing with the IT (Information Technology) division for gathering data requirements and security.

Solution through RPA

Early steps showed that RPA was a lot more than simply an efficiency enabler for the HR group at Walgreen. For instance, Walgreens has around 2,000 workers on leave of absence on any working day. The Blue Prism loads the required information into all the essential applications instantly, which includes whether it is compensated or maybe unpaid leave. The Digital Worker transmits information between Walgreens systems as well as its program provider within the worker's compensation procedure.

Blue Prism's potential to work with the various methods was valuable in proving RPA's capability to Walgreen by showing precious time cost savings through automating the transactions.

The HR Lead at Walgreens was satisfied with Blue Prism's OCR (**Optical Character Recognition**) integration abilities as well as its potential to push scanned documents between the various systems. He supplies an example.

"It can read the employee ID and the store number on the document—it knows the difference. You can program the difference into whether it's a store, a distribution number or distribution center, or whether it is somebody in the support office. And if there is something specific that we want to do differently with a distribution center, it can kick it out into a separate queue. It then creates an e-mail to the proper address, sends it and saves it to the hard drive for case management. This is all done within a fraction of the time of a human performing the same tasks."

Business Impact

The Blue Prism RPA wedge supplies Walgreens with the bandwidth to offer extra effort and allows staff members to improve the customer experience as well as profits, while boosting effectiveness by 73 % within the HR shared providers team.

Source: How Blue Prism's Digital Workforce is increasing Walgreens' HR shared . . . , <https://cdn2.assets-servd.host/lively-jackal/production/uploads/resources/case-studies/blue-prism-Walgreens-case-study.pdf> (accessed August 2022).

► SECTION 10.8 REVIEW QUESTIONS

1. What is RPA?
2. Give examples of RPA in action.
3. Research and find some examples of RPA software providers.

10.9 BIOINFORMATICS AND HEALTH NETWORK SCIENCE

Bioinformatics in general and health analytics in particular have emerged as major areas of growth in AI. In this section, we begin with an application story that highlights the application of analytics to bioinformatics area. Then we introduce network science, an area that employs advanced analytical techniques.

ANALYTICS IN ACTION 10.10 Analyzing the genomics data to optimize animals' gut health and performance

Background

Both the human and animal gut contain a complex ecosystem of trillions of different microorganisms and their collective genomes, known as the gut microbiome. The microbiome is composed of all the microflora (bacteria, viruses, protozoa, and fungi) that live on or within a host organism. These microbes form a very complex ecological entity in body that plays an important role in many aspects of both nutrition and health. The food that animals eat influences the gut microbiome. Both good and bad bacteria can be found in the gut and are a

natural part of a balanced microbiome. However, when the bad bacteria outnumber the good bacteria, health and nutrition issues can arise. Thus, understanding which foods will improve gut microbiome is an important problem. Some of the main questions that we need to study are the following:

- a. How can we describe/identify if a group of bacteria are good or bad for the animals?
- b. How can we predict the performance and health attributes of animals using the microbiome/genomic data?
- c. How can we prescribe the solutions to the producers based on what you learned from predictions?

The Foundation—Purina MQ (Microbiome Quotient) Bioinformatics Platform

Purina uses gene sequencing technology in the Purina Emerging Technologies Lab to analyze the humongous sequencing data in petabytes generated from the samples. Samples get sequenced on an Illumina MiSeq (Illumina, San Diego, CA) machine. Gene sequencing technology allows us to identify members of the microbial population to the lowest possible species/genus level. However, a lot of complex processing has to be completed to get to where we can compare the microbial populations across different samples. The Data Intelligence team at Purina developed a cloud-based Purina MQ (**Microbiome Quotient**) Bioinformatics Platform enabled with parallel computing to automate the analysis of the sequencing data using the open-source package called QIIME2. QIIME2 helps us identify the bacteria at species level and translate raw sequence data into useful statistical results.

The team's idea here is to utilize the microbiome sequencing pipeline and evolve into an Integrated Animal Systems Biology Platform. It helps the team to process, analyze, store, and extract knowledge from sequence data and provides a cost effective, flexible, and scalable solution across all the animal species.

The main goals of developing this platform are the following:

- High performance and high throughput bioinformatics
- A central data repository to access and compile data sets
- Standardization of microbiome data management and analysis
- AI/ML and Advanced Statistical analysis

AI/ML and Biostatistical Methods for Data Analysis

Statistical Analysis

Relative abundance within and among groups is assessed using multiple different methods, for each metadata category, for example, one-way ANOVA, PERMANOVA, ANCOM, Kruskal–Wallis, and Wilcoxon Rank. (Several of these were introduced in Chapter 3.)

Supervised Learning

We employed multiple different AI/ML models like logistic regression (LR), linear discriminant analysis (LDA), effect size (LEfSe), random forest (RF), a wrapper around random forest called Boruta, **long short-term memory (LSTM)**, **convolutional neural network (CNN)**, etc., to compare the estimated phylotypes of animals/birds with disease and healthy microbiotas. Intergroup differences at the phylum, class, order, family, genus, and species level in each cluster were analyzed through all the above ML methods in R using the HMP, microbial, microbiome Bioconductor, LEfSe and mikroml packages.

LEfSe uses the two-tailed nonparametric Kruskal–Wallis test to evaluate the significance of differences in OTUs in two groups. A set of pairwise tests among two groups was performed using the unpaired Wilcoxon test. Finally, LDA was performed to estimate the effect size of each differentially abundant OTU. The results are expressed as the mean $\pm$ SEM. A strength of the LEfSe method compared with standard statistical approaches is that in addition to providing p values, it estimates the magnitude of the association between each OTU and the grouping categories, such as Control and FUDS. For stringency, the gut microbiotas were considered significantly different if their differences had a p value < 0.05 and an LDA score ($\log_{10}$) > 2.5 . The Kruskal–Wallis and Cramer V tests were used to determine associations between the groups of subjects and characteristics

Unsupervised Learning, Alpha and Beta Diversities

We use both alpha diversity (α -diversity) and beta diversity for understanding the different communities of animal samples. Alpha diversity is defined as the mean diversity of species in different sites or habitats within a local scale. Beta diversity measures the change in diversity of species from one environment to another. While alpha diversity is a measure of microbiome diversity applicable to a single sample, beta diversity is a measure of the similarity or dissimilarity of two communities. As for alpha diversity, many indices exist, each reflecting different aspects of community heterogeneity.

In simpler terms, it calculates the number of species that are not the same in two different environments. We use principal coordinate analysis for the Bray Curtis dissimilarity distances and UniFrac for each sample using Phyloseq for beta diversity. The significances of grouping in the PCoA plots is tested by analysis of dissimilarity (ADONIS) with 999 permutations from the vegan package in R Oksanen et al., 2013).

Outcomes and the Future

The Purina Research team has launched multiple microbiome quotient (MQ) research projects across all the species aimed at answering the question of what a normal, healthy animal's microbiome looks like. Also, we are looking the microbial community differences between healthy and diseased animals. We are also checking the difference in the microbiota for the animals who are on antibiotics vs. who are not on antibiotics. We are on a mission to create the largest database across both livestock and lifestyle species and learn as many things as we can to innovate personalized nutrition, probiotics products, and specialized diets.

Contributed By: Manohar Mohanlal Lahoti, Data Science and Digital Product Manager, Land O'Lakes Inc., Purina Animal Nutrition subsidiary.

NETWORK ANALYTICS

Network Analysis (NA) is a popular method for analyzing complex problems involving interactions among the features or the observations (Chen et al., 2012). While NA is not a new technique, it has recently gained momentum due to its suitability for analyzing large data sets and ability to understand the underlying relationships or connectedness. One major area that generates network-type data is social media, where relationships are explicitly embedded. However, there are other types of networks with implicit relationships that are defined using some underlying exchanges. Examples include product copurchase network (Dhar et al., 2014), ingredient network (Teng et al., 2012), comorbidity network (Hidalgo et al., 2009; (Kalgotra, Sharda and Croff, 2017), text network (Celardo and Everett, 2020), and others. More recently, we are seeing **network analytics** techniques used in conjunction with other approaches such as deep learning or natural language processing. Such innovative and hybrid use of methodologies in newer application domains such as COVID-19 spread, and infodemic can push the frontiers of scientific discovery.

Network science has been used as a *theory* to understand an emergent phenomenon and as a *methodology* to model the relationships. It has been used to create descriptive, predictive, and prescriptive analytics examples. The examples of descriptive analytics applications developed using network analysis include a study by Kalgotra, Sharda, and Croff (2020) published in *Nature's Scientific Reports* journal, in which the authors identified the health disparities among the different racial groups by comparing the **comorbidity networks** created using the electronic health records of millions of patients. Another example is the ingredients network created by Teng et al. (2012), in which the authors used the co-occurrence of multiple ingredients in the same recipe to genre the network and identify the clusters of ingredients to suggest new recipes. Similar other descriptive analytics applications are available that utilize network analysis as a core methodology to model interactions in the large data sets in any business and other scientific domains.

The traditional network analytics research creating predictive analytics applications focuses on predicting the attributes of the nodes in the network. For example, Dhar et al. (2014) predicted the sales of books in a book copurchase network created using the recommendation system on the Amazon Web site. Another typical network applications include the prediction of links in the network. Lü and Zhou (2011) highlight different methods and applications of link prediction in their survey. Recently, network analysis has been used as a tool to create new. The features created using network analysis are further used to develop machine learning models to predict outcomes exogenous to the network. For instance, Kalgotra and Sharda (2021) used the properties of the comorbidity network to predict the hospital length of stay for a patient. We summarize this application in the next section.

Finally, network analysis is a popular research area for prescriptive analytics. Examples include creating prescriptive models to identify cliques in the network, as discussed by Miao and Balasundaram (2017), to traverse the network and identify the shortest path between nodes (Selim and Zhan, 2016), etc.

ANALYTICS IN ACTION 10.11 Network analytics for predictive modeling: Predicting hospital length of stay using comorbidity networks

Here, we summarize an application by Kalgotra and Sharda (2021) in which network analysis has been used to predict an outcome which is exogenous to the network. This paper has been published in the *Journal of Management Information Systems*. Specifically, a comorbidity network was used to predict hospital length of stay (LOS). An early prediction of the hospital length of stay is an important problem for the patients, hospitals, and insurance companies. Accurate estimation can help hospitals manage and schedule resources such as beds and staff. An average cost of a stay in 2020 was about \$14,900, according to the Agency for Healthcare Research and Quality (AHRQ at <https://datatools.ahrq.gov/hcupnet>, accessed February 2023). Therefore, the insurance companies need to know the LOS for patients early for budgeting and revenue management.

In this paper, the authors used the Cerner electronic medical record (EMR) data warehouse containing information on more than 24.7 million patients across 662 U.S. hospitals over 16 years (2000–2015). The EMR includes the information about the multiple hospital visits by each patient. The variables available for this research included age, gender, race, date of admission, date of discharge, and diseases diagnosed. In addition, the primary reason for the hospital visit was also available. The authors used a two-step approach to create the machine learning model—creating comorbidity networks in the first step and then creating machine learning models for predicting LOS in the second step.

First, the medical records of about three million patients were extracted and used to create separate comorbidity networks for men and women because these groups manifest different types of diseases. In the comorbidity network, the diseases were the nodes, and two diseases were connected if they appeared in a patient during the same hospital visit. Each disease was recorded using the ICD-9-CM coding system. Across all the patients, 15,584 unique diseases were there. In the men network, there were 11,898 diseases connected through 249,902 connections, whereas in the women network, 12,719 diseases were connected through 267,170 links. The comorbidity networks were used to create new variables for the remaining patients who were not part of the network analysis. To understand the new features generated using comorbidity networks, consider a male patient who visits the hospital with a complaint of hypothetical disease A. Therefore, the only disease-related information available at the time of admission is disease A. In this case, the men network was searched for disease A, and the top five connected diseases were identified. These diseases were labeled as probable diseases as these were likely to be diagnosed during the hospital stay. A patient may have a history of diseases in the system, termed as historical diseases. Together, the probable and historical diseases were called as latent comorbidities. The new construct of latent comorbidities was further used in predicting LOS at the time of admission.

The predictive models for LOS were created with patient demographics, the known diseases at the time of admission, and latent comorbidities as the independent variables. The latent comorbidities were inputted as a two-dimensional array with historical diseases as the first and probable diseases as the second dimensions. Since the input variables included a two-dimensional construct, the Long-Short-Term Memory (LSTM) was adopted because it can handle the multidimensional variables. The models were created without and with the latent comorbidities to compute the explanatory and predictive power added by the proposed variables. In terms of variance explained, the new construct added 3.6%, and in terms of mean absolute percent error, the latent variable improved the number by 1.9%. Although the numbers seem low, these are equivalent to the improvement in the forecast by \$882.8 million. Therefore, the gain is practically significant.

Recalling the predictive modeling process using network analytics, the authors inferred the possibility of other probable diseases during the hospital visit using the comorbidity network. This information can help physicians adopt better treatment plans. Then, the comprehensive information about the patient, including the known diseases at the point of admission, historical diseases and probable diseases, and demographics, were used to create deep learning models.

Compiled from: Kalgotra, P. and R. Sharda. (2021). “When will I get out of the hospital? Modeling Length of Stay using Comorbidity Networks.” *Journal of Management Information Systems*, Vol. 38, No. 4, pp. 1150–1184.

► SECTION 10.9 REVIEW QUESTIONS

1. What is genomics data?
2. How did Purina analyze the genomics data for making decisions?
3. What is network science?
4. How can the network information be useful in predicting other variables of interest?

10.10 OTHER RECENT DEVELOPMENTS

Web 3.0

Another topic that is gaining much hype is **Web 3.0**. A good primer for Web 3.0 is in Murray et al. (2022), and many online articles in various magazines. First version of the Internet where we all logged into a Web site to find content or exchange e-mails was dubbed Web 1.0. In Web 1.0, a few entities created the content, but millions were able to consume it. In Web 2.0, consumers of the content also became producers. First, social media companies such as YouTube, TikTok, Instagram, virtually any Web site providing a user to share comments/content, have enabled “consumers” to also upload their own content and be seen by millions of other users. Second, every user generates data by about their usage patterns, interests, etc. that is collected by companies enabling these applications so as to mine this information and send targeted advertisements as well as suggest content that will keep you hooked to that platform. Thus, Web 2.0 became a platform for everyone to generate and share data/content, but enabled by companies such as Google, Meta, Twitter, and hundreds of other players. In this process, these companies became the central anchors in making Web 2.0 functional and useful. The idea of Web 3.0 is to decentralize the Web again. How it will all shape out in ten years is anybody’s guess, but as students of data science and AI, you should be watching the latest developments in this space.

Web 3.0 is considered decentralized because its fundamental building block is the open **blockchain** technology. Blockchain is largely a digital ledger that records all activity taking place on/with an entity. It is all public so that its authenticity can be verified. However, the access is also protected through a unique digital wallet, which

may be considered the next generation of authentication mechanisms as well as a collection of such mechanisms. Because all activity is public, it is not controlled by any one organization unlike the current content, which may be on any specific social media platform.

Web 3.0 also aims to employ cryptocurrencies, which are blockchain-driven mechanisms of financial transactions. This is a crucial component of Web 3.0. Another related idea is **non-fungible tokens (NFT)**, which allows someone to create/own/transfer a particular digital object (image, video, etc.) so that its authenticity can be assured. Another group of players that enable Web 3.0 are **decentralized autonomous organizations (DAOs)**, which allow the ability to vote on future protocols and proposals.

Another component of Web 3.0 is a **metaverse**. As noted in the earlier introduction of digital twins, metaverses are just virtual representations of real or imaginary physical environments. This implied a lot of 3D graphics. Facebook has embraced the concept of metaverses most directly, by even changing its name to Meta. Many initial applications of metaverses are through virtual games. When a user pays to play a game or is paid a winning, and the mechanism is a cryptocurrency, you can see some of the components of Web 3.0 coming together.

An early use of AI/data science in Web 3.0 will come through developing better user interfaces for accessing the blockchain-enabled Web 3.0. Anyone who has played with any blockchain/cryptocurrency technologies can vouch for the fact that currently these are not very use friendly. For these technologies to reach scale, much work will have to be done in making these tools user-friendly, intuitive, and using advances in AI (user interfaces, chatbots, deep learning, etc.). This is certainly an open frontier at this point.

Metaverse and Digital Twins

Recently, there has been much talk and buzz about the next major wave of technology metaverse. The major idea is to create a virtual replica of a physical object or system—a person, factory city, etc. where other virtual objects interact with each other. The objects can be other human beings who participate in this metaverse. Facebook has bet big on this vision and changed the company name to Meta. Most other major technology companies are investing in this space as well. Its previous incarnations have been labeled virtual reality, augmented reality, etc. Our main interest here is to recognize that emergence of this technology will have a significant impact on further development of AI/data science as well as many potential applications of current AI models in this domain. The behavior of the objects can be studied to identify patterns of behavior. Even more exciting frontier is the real-time digital twins. By incorporating inputs from IoT sensors, camera, etc. in real time, a digital twin can be an almost real-time replica of a physical object. This level of detail can provide significant decision support capability to the human decision makers without necessarily being in the physical environment at that specific moment. An interesting video explaining this capability and some recent real applications in factories, cities, etc. was broadcast in Japan. It can be viewed at <https://www3.nhk.or.jp/nhkworld/en/news/liveblog/4/> (accessed August 2022). The reader should also consult this Web site: <https://www.rtinsights.com/what-differentiates-real-time-digital-twins/> (accessed August 2022). This is a major emerging opportunity for the next phase of virtualization of reality, collection of real-time data, and being able to make automated as well as manual decisions on what is happening in the real world by observing the situation in its digital twin.

Some other recent advances in AI relate to conversational AI or chat bots (Chapter 2). We highlight three major examples that you should continue to follow for developments and new, interesting applications.

GPT-3/ChatGPT

Many of the new generations of natural language processing systems use machine learning at various levels to predict what the next element of a conversation can/should be. By training a model with billions of samples, many researchers are trying to build various prediction models in this genre. One example of such a system is GPT-3, a product of OpenAI (OpenAI.com). (OpenAI was started as a nonprofit company by some of the leading investors such as Elon Musk. However, it has since become a part of a for-profit company and Microsoft is the leading investor in the commercial applications of this technology.) GPT stands for Generative Pre-trained Transformer. Transformer is a type of neural network architecture that is best for sequential data—autoregressive. Most text, for example, would be considered to be somewhat sequential. This is also true of vision. No. 3 just denotes the current version of this technology. Generative Pre-training refers to predicting what is the next token. By training the model on billions of samples that do not require human labeling, the algorithm can scale easily. For example, GPT-3 has been used to build chatbots for some applications, create poetry, translate languages, develop images from text or vice versa, even write full prose, develop software code, explain the purpose of generated code, etc. OpenAI's Web site (<https://beta.openai.com/examples/>) lists several of the application prototypes that have been generated thus far using this technology. We introduced ChatGPT in Chapters 2 and 7 earlier. Manjoo (2020) and many other authors have discussed its applications as well as concerns about such automated technologies. ChatGPT was made available to general public recently and has become a widely tried application. It is reputed to create content for almost any topic, including code for use in applications through Microsoft's GitHub CoPilot program. A research project by Terwiesch (2023) pointed out that the answers provided by ChatGPT would have passed an MBA exam. Many academics are now discussing the impact of use of such tools in classrooms and ensuring students develop their own content. Microsoft is embedding ChatGPT functionality in its search engine Bing to enhance search results and provide a coherent answer to a search query rather than just a list of links that Google search typically provides (except for some queries where an answer may be provided, e.g., weather, sports scores, flight information). This development has put Google on the defensive. Exciting new advances are likely to emerge for consumers' benefits through this new rivalry among all major technology companies for employing AI in their offerings. Blenderbot (introduced below) is Meta's offering. See next for Google's developments.

LaMDA

LaMDA is an acronym for Language Model for Dialogue Applications, Google's adaption of technologies similar to GPT-3 in developing their chatbot. LaMDA is supposed to be able to manage conversations that begin on one topic but divert to other topics, a very common occurrence in human conversations. Similar to ChatGPT, LaMDA is also available to the public. Indeed, a chatbot based on LaMDA called Bard is now available as of February 2023. LaMDA was also in the news recently because a software engineer at Google argued that this chatbot was so advanced it could be considered to be sentient—humanlike. This engineer's claims were refuted by other leaders at Google, and a public feud erupted. We will discuss this issue further in the next chapter in our section on Responsible AI.

Blenderbot

Not surprisingly, Meta (the owner of Facebook) has also been investing heavily in developing AI applications and in August 2022 released their chatbot called Blenderbot 3. You can chat with it at <https://blenderbot.ai/> (accessed Aug. 2022). According to some preliminary results published by Meta AI researchers, based upon some feedback from the early users of this chatbot (about 65K conversations), less than 3% of the conversations were marked as inappropriate, nonsensical, or off the topic. Thus, almost 97% relevance of a conversation is a high mark of achievement for such technologies.

Here, we are just introducing you to the latest advances in conversational technologies that appear to make a human-computer conversation almost look like a human-to-human conversation. In addition to the conversational technologies, these same broad class of tools can also create content such as poetry, images, and text on the basis of a user's request. This branch of AI is called Generative AI. Besides the large technology companies identified here, Stability Diffusion is a major player in this space, among about 500 additional startups. This is a major exciting new frontier of AI. The next chapter will mention these again in the context of managerial, legal, and ethical considerations.

► SECTION 10.10 REVIEW QUESTIONS

1. What is Web 3.0?
2. Define metaverse. What technologies are likely to power metaverse?
3. Research further into Blenderbot and other chatbots, and identify recent developments.

Chapter Highlights

- Cloud computing is one of the major emerging fields in 21st century, which provides great opportunities for analytics.
- Cloud Computing offers a shared pool of resources like storage, computing power and network, etc.
- Major cloud service providers are Amazon Web Services (AWS), Microsoft Azure, Google Cloud, IBM Cloud.
- Cloud offers pay-as-you-pricing which means we only have to pay for the resources we used for the given amount of time.
- The three major cloud models are software as a service (SaaS), platform as a service (PaaS) and infrastructure as a service (IaaS).
- The other available cloud models are data as a service (DaaS) and desktop as a service (DaaS).
- Cloud services are deployed in three different ways as private cloud, public cloud, and hybrid cloud.
- Geospatial analytics helps organizations to get more informative location-based insights.
- Taking advantage of the location-based data will help organizations find potential growth opportunities at the granular level.
- Location-based data is being embraced by organizations to target the right customers.
- Loan approval decisions can be made with the use of geographic information system (GIS) and business intelligence (BI) integrated systems.
- Starbucks used AccuWeather's data to customize the menu according to the weather.
- Location information can be collected from many devices used by customers through technologies like GPS, Wi-Fi, and cell tower triangulation.
- Consumer-based applications make great use of location analytics for their marketing needs.
- Analytics-based applications are being used by organizations to get consumer data like their purchase history and services usage.
- Gmail analyses billions of e-mail conversations and suggests automated responses for e-mails.
- Urban Planners use satellite data for infrastructure planning to prepare for natural disasters.
- EarthCast provides weather updates for U.S. airline with the data collected from 60 satellites.
- Imazon uses real-time satellite data to track deforestation and illegal activities in Amazon.
- Indonesian government monitors illegal ship movement with the help of satellite data.
- Image analysis helps in estimation of green cover in a forest or farm area.
- Mobile app Canopeo estimates real-time green canopy cover from images from smartphone or digital camera.
- Unilever is working with Orbital Insight to use image analytics in its supply chain to examine deforestation with satellite imagery and geospatial data.
- Internet of Things helps people and the objects to be interconnected anytime and anyplace.
- Sensors are integral part of Internet of Things, and they are connected to Internet.
- Sensor is an electronic device which automatically collects data from its environment.
- In IoT, each object receives, sends, and stores data automatically over the wireless internet.
- An example of IoT is the Brita filter introduced by Clorox, which automatically places orders for filters by detecting the time to change the filters.
- French National Railway System uses IoT to improve safety of its passengers by installing thousands of sensors on the trains and tracks to collect information.

- Smart home is a home in which the components are interconnected wirelessly and centrally controlled to be able to communicate with each other.
- The components of smart homes are lighting, TV, energy management, water control, smart speakers, refrigerator and home security.
- Fifth-generation (5G) mobile technology offers ultra-low latency and provides high-speed access even in densely populated areas.
- Industry Revolution 4.0 will be driven by 5G, AI, network slicing, and edge computing.
- Robotic process automation (RPA) is a major emerging field which reduces the workloads of humans by automating time-consuming repetitive work.
- The wide network coverage and cloud computing platforms paved way for the growth of robotic process automation.
- RPA helped Walgreens HR shared service to increase its efficiency by 73%.
- Network analytics can be used to analyze large data sets and social media is an example for network type data.
- Network analytics helps us to understand the underlying relationships which are explicitly embedded.
- The idea of Web 3.0 is to decentralize the Web again and its fundamental building block is the open blockchain technology.
- The crucial component of Web 3.0 is the metaverse, which is virtual representation of real or imaginary physical environments.
- Web 3.0 will employ cryptocurrencies which are blockchain-driven mechanisms of financial transactions.
- Blenderbot, GPT-3, and LaMDA are examples of emerging new conversational and generative AI technologies.

Key Terms

Artificial Intelligence (AI)	tors (KPI)	Smart Sensor	Microbiome Quotient (MQ)
Internet of Things (IoT)	Geographic Information Systems (GIS)	Driverless Cars	Long-Short-Term Memory (LSTM)
Sensors	Radio Frequency Identification (RFID)	Fifth Generation (5G)	Convolutional Neural Network (CNN)
Machine Learning	ArcGIS	Service-Based Architecture (SBA)	Network Analytics
Cloud Computing	Alternative Data	Fourth Industrial Revolution	Comorbidity Networks
Business Intelligence (BI)	EarthCast	Edge Computing	Web 3.0
Virtualization	Canopeo	Virtual Reality (VR)	Metaverse
Mobile Extension Framework (MXFTM)	Deforestation	Robotic Process Automation (RPA)	Blockchain
Analytics as a Service (AaaS)	Orbital Insight	Software robots	Non-Fungible Tokens (NFT)
GO-JEK	Ubiquitous Computing	Coronavirus Pandemic	Decentralized Autonomous Organizations (DAOs)
Zion China	Unique Identifier (UID)	Digital Workforce	Digital Twins
Stream Analytics	Machine to Machine (M2M) Communication	Optical Character Recognition (OCR)	Deep learning
Cloud Security Alliance	Smartbin	Genomics	
Geospatial Analytics	IBM Watson		
Key Performance Indicators (KPI)			

Questions for Discussion

1. Compare the IoT with regular Internet.
2. Discuss the potential impact of autonomous vehicles on our lives.
3. Why is the IoT considered a disruptive technology?
4. Alexa is typically connected to smart home devices such as thermostats, TVs, and microwaves. Find examples of other appliances that are connected to Alexa and write a report.
5. Discuss the objective of smart cities to conserve the earth's limited resources.
6. What are the major uses of IoT?
7. Accidents involving driverless cars slow down the implementation of the technology. Yet, the technology can save hundreds of thousands of lives. Is the slowdown (usually driven by politicians) justifiable? Discuss.
8. What are the potential benefits of using geospatial data in analytics? Give examples.
9. What types of new applications can emerge from knowing locations of users in real time? What if you also knew what they have in their shopping cart, for example?
10. How can consumers benefit from using analytics, especially based on location information?

11. “Location-tracking–based profiling is powerful but also poses privacy threats.” Comment.
12. Is cloud computing “just an old wine in a new bottle?” How is it similar to other initiatives? How is it different?
13. Research and discuss the impacts of technologies such as chatGPT in a domain/industry of interest to you: sports, entertainment, education, retail, gaming, agriculture, manufacturing, finance, etc.
14. Research and discuss the impacts of technologies such as blockchains in a domain/industry of interest to you: sports, entertainment, education, retail, gaming, agriculture, manufacturing, finance, etc.
15. Research and discuss the impacts of technologies such as digital twins in a domain/industry of interest to you: sports, entertainment, education, retail, gaming, agriculture, manufacturing, finance, etc.

Exercises

1. Go to Discover Food site and study how they are using ML models and algorithms in revolutionizing the food safety.
2. Research the importance of temperature check in food industry and write a report.
3. Watch the Introduction to Google Cloud (22:24 min.) on Google Cloud Tech at <https://www.youtube.com/watch?v=leMYQ-qJeK4> and learn about various features of GCP. Write a report.
4. Go to the Web site of the National Institute of Standards and Technology <https://www.nist.gov/> and learn about their research programs.
5. Go to AWS Web site <https://aws.amazon.com/> and try using their “Try Analytics Services for Free.”
6. Go to <https://www.salesforce.com/> and <https://www.ibm.com/cloud> and find their capabilities. Summarize the various cloud offerings.
7. Find recent information on latest cloud developments. Write a report.
8. Study and write a report on various virtualizations like Network, Storage and Server.
9. Find the latest information on real-time incident reporting using cloud-mobile technology. Write a report.
10. Find information on the use of cloud technology in mobile apps. Write a report.
11. Explore the new cloud-related products and services of Google Cloud <https://cloud.google.com/>. Explore the various cloud offerings.
12. Find material on the impact of cloud on marketing and advertising. Write a report.
13. Identify applications other than those discussed in this chapter where cloud technology is being used industries.
14. Go to MineMyText.com site and study about their text mining capabilities in understanding the sentiments from customer reviews.
15. Go to <https://www.ptc.com/en/resource-center> or other sources, and select three IoT implemented cases. Write a summary of each.
16. It is said that the IoT will enable new customer service and B2B interactions. Explain how.
17. Enter smartcitiescouncil.com. Write a summary of the major concept found there; list the major enablers and the type of available resources.
18. Find the status of Bill Gates's futuristic smart city. What are some of its specific plans?
19. *City Brain* is the name of Alibaba's platform for smart cities. One project has been adopted in China and Malaysia. Find information and write a report.
20. The competition for creating and using autonomous cars is intensifying globally. Find 12 companies that are competing in this field.
21. Enter the McKinsey Global Institute Web site, mckinsey.com/mgi/overview, and find recent studies on IoT. Prepare a summary.
22. Alexa is now connected to smart home devices such as thermostats and microwaves. Find examples of other appliances that are connected to Alexa and write a report.

References

- Murray, A., Kim, D., & Combs, J. “The promise of a decentralized Internet: What is Web 3.0 and HOW can firms prepare?” *Business Horizons*, 2022. <https://doi.org/10.1016/j.bushor.2022.06.002> (accessed August 2022).
- Ashton, Kevin. “That ‘internet of things’ thing.” *RFID journal* 22.7 (2009): 97–114.
- Bray, E. “Are consumers, automakers and insurers really for self-driving cars?” *Tech Crunch*, August 10, 2016.
- Celardo, L., & Everett, M. G. (2020). Network text analysis: A two-way classification approach. *International Journal of Information Management*, 51, 102009., at <https://doi.org/10.1016/j.ijinfomgt.2019.09.005>
- Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165–1188.
- COVID-19 health tracking and alerting automation for State Government. (2022) UiPath. <https://www.uipath.com/resources/covid-automations/covid19-health-tracking-alerting-automation-for-state-government> (accessed August 2022).
- Demirkan, H., & Delen, D. (2013, April). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and Big Data in cloud. *Decision Support Systems*, 55(1), 412–421
- Donaldson, J. “Is the role of RFID in the internet of things being underestimated?” *Mojix*, May 2, 2017.

- Dhar, V., Geva, T., Oestreicher-Singer, G., & Sundararajan, A. (2014). Prediction in economic networks. *Information Systems Research*, 25(2), 264–284.
- Davis, N. 2016. “What is the fourth industrial revolution?” World Economic Forum Geneva, January 19, 2016, Retrieved from: <https://www.weforum.org/agenda/2016/01/what-is-the-fourth-industrial-revolution/> (accessed August 2022).
- Estopace, E. “French national railway operator taps IoT for rail safety.” *eGov Innovation*, February 21, 2017a.
- Fox, B., Viveros, M., and Dam, R. v. d. “Telecom’s 5G future.” IBM Institute for Business Value. February 2020. Retrieved from <https://www.ibm.com/thought-leadership/institute-business-value/report/5g-telecom#> (accessed August 2022).
- Field, H. & Murphy, M. (2020) COVID-19 will herald an automation boom, <https://www.protocol.com/automation-boom-caused-by-coronavirus> (accessed August 2022).
- Greengard, S. “How AI will impact the global economy.” *CIO Insight*, October 7, 2016.
- Gartner Says Worldwide Robotic Process Automation Software Revenue to Reach Nearly \$2 Billion in 2021. (2020). Gartner. <https://www.gartner.com/en/newsroom/press-releases/2020-09-21-gartner-says-worldwide-robotic-process-automation-software-revenue-to-reach-nearly-2-billion-in-2021> (Accessed August 2022).
- Hupfer, S. “AI is the future of IoT.” *IBM Blog*, December 15, 2016. [ibm.com/blogs/internet-of-things/ai-future-iot/](https://www.ibm.com/blogs/internet-of-things/ai-future-iot/) (accessed August 2022).
- Hidalgo, C. A., Blumm, N., Barabási, A. L., & Christakis, N. A. (2009). A dynamic network approach for the study of human phenotypes. *PLoS Computational Biology*, 5(4), e1000353.
- “How Blue Prism’s Digital Workforce is increasing Walgreens’ HR shared service efficiency by 73%.” *Blueprism*, 2022, <https://cdn2.assets-servd.host/lively-jackal/production/uploads/resources/case-studies/blue-prism-Walgreens-case-study.pdf> (accessed August 2022).
- Kalgotra, P., Sharda, R., & Croff, J. M. (2017). Examining health disparities by gender: A multimorbidity network analysis of electronic medical record. *International Journal of Medical Informatics*, 108, 22–28.
- Kalgotra, P., Sharda, R., & Croff, J. M. (2020). Examining multimorbidity differences across racial groups: a network analysis of electronic medical records. *Scientific reports*, 10(1), 1–9.
- Kalgotra, P. & Sharda, R. (2021). “When will I get out of the hospital? Modeling length of stay using comorbidity networks.” *Journal of Management Information Systems*. 38(4), pp. 1150–1184.
- Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical mechanics and its applications*, 390(6), 1150–1170.
- Luckham, D. C. 2012. *Event Processing for Business: Organizing the Real-Time Enterprise*. Hoboken, New Jersey: John Wiley & Sons, Inc. (accessed August 2022).
- Manjoo, F. “How do you know a human wrote this?” *New York Times*, July 29, 2020. ISSN 0362–4331 (accessed August 2022).
- Miao, Z., & Balasundaram, B. (2017). Approaches for finding cohesive subgroups in large-scale social networks via maximum k-plex detection. *Networks*, 69(4), 388–407.
- McCafferty, D. “How the Internet of Things is changing everything.” *Baseline*, June 16, 2015.
- Meola, A. “What is the Internet of Things (IoT)? Meaning & definition.” *Business Insider*, May 10, 2018.
- Miller, M. *The Internet of Things: How Smart TVs, Smart Cars, Smart Homes, and Smart Cities Are Changing the World*. Indianapolis, IN: Que Publishing, 2015.
- Oksanen, J., Blanchet, F. G., Kindt, K., Legendre, P., Minchin, P. R., O’Hara, R. B., Simpson, G. L., et al. “Package ‘vegan’.” Community ecology package, version 2, no. 9 (2013): 1–295
- Pasqua, E. “How 5G, AI and IoT enable “Intelligent Connectivity,” *IOT Analytics*. February 27, 2019. Retrieved from <https://iot-analytics.com/how-5g-ai-and-iot-enable-intelligent-connectivity/> (accessed August 2022).
- Piotrowicz, W. (2020) Digitization and technology in supply chain, during pandemics and beyond, <https://www.linkedin.com/pulse/digitization-technology-supply-chain-during-pandemics-piotrowicz/?fbclid=IwAR0Kp04DbrFuDbPZN3bHoP0kq4PWYfvr5vVmEaVn0TBogxZ5ffAeeg3KxeM> (accessed August 2022).
- Selim, H., & Zhan, J. (2016). Towards shortest path identification on large networks. *Journal of Big data*, 3(1), 1–18.
- Snowdon, J., Canepa, S., Fox, B., & Dam, R. v. d. “What consumers expect from 5G entertainment.” IBM Institute for Business Value, September 2019. Retrieved from <https://www.ibm.com/thought-leadership/institute-business-value/report/immersive-5g-entertainment> (accessed August 2022).
- Somisetty, M. 2018. “Big Data Analytics in 5G,” IEEE. Retrieved from <https://futurenetworks.ieee.org/images/files/pdf/applications/Data-Analytics-in-5G-Applications030518.pdf> (accessed August 2022).
- STL Partners. 5G’s impact on manufacturing: \$740Bn of benefits in 2030, October 2019. Retrieved from <https://stlpartners.com/research/5gs-impact-on-manufacturing-740bn-of-benefits-in-2030/> (accessed August 2022).
- Teng, C. Y., Lin, Y. R., & Adamic, L. A. (2012, June). Recipe recommendation using ingredient networks. In *Proceedings of the 4th Annual ACM Web Science Conference* (pp. 298–307).
- Terwiesch, C. “Would Chat GPT Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course”, Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania, 2023 (Available at <https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Christian-Terwiesch-Chat-GTP-1.24.pdf>, accessed Feb. 2023)
- Tomás, J. “Smart factory tech defining the future of production processes.” *RCR Wireless News*, March 28, 2016.
- Tulchinsky TH, “John Snow, Cholera, the Broad Street Pump; Waterborne Diseases Then and Now.” *Case Studies in Public Health*. 2018:77–99. doi: 10.1016/B978-0-12-804571-8.00017-2
- Venkatakrishnan, K. “Are connected consumers driving smart homes?” *Enterprise Innovation*, May 31, 2017.
- Yan, J., Meng, Y., Lu, L., & Li, L. 2017. “Industrial big data in an Industry 4.0 environment: Challenges, schemes, and applications for predictive maintenance.” *IEEE Access*, 5, pp. 23484–23491.

Ethical, Privacy, and Managerial Considerations in Analytics

LEARNING OBJECTIVES

- Describe the major implementation issues of intelligent technologies
- Discuss legal, privacy, and ethical issues
- Discuss the importance of ethical/responsible use of intelligent technologies
- Understand the deployment issues of intelligent systems
- Describe the major impacts on organizations and society
- Discuss and debate the impacts on jobs and work
- Discuss the arguments of utopia and dystopia in a debate of the future of robots and artificial intelligence (AI)
- Understand the data science profession, specifically the role of a citizen data scientist in analytics

In this concluding chapter, we cover a variety of issues related to the implementation and future of intelligent systems. We begin our coverage with technological issues such as security and connectivity. Then, we move to managerial issues that cover legality, privacy, and ethics. We next explore the impacts on organizations, society, work and jobs. Then, we conclude with a brief discussion of data science profession, including the new term *citizen data scientist*.

Note. In this chapter, we refer to all technologies covered in this book as *intelligent technologies or intelligent systems*.

This chapter has the following sections:

- 11.1** Opening Vignette: Lessons Learned from Analytics Journey in an Organization 630
- 11.2** Implementing Intelligent Systems: An Overview 633
- 11.3** Successful Deployment of Intelligent Systems 634
- 11.4** Implementing IoT and Managerial Considerations 637
- 11.5** Legal, Privacy, and Ethical Issues 640

- 11.6 Ethical/Responsible/Trustworthy AI 646
- 11.7 Impacts of Intelligent Systems on Organizations 651
- 11.8 Impacts on Jobs and Work 656
- 11.9 Potential Dangers of AI 662
- 11.10 Citizen Science and Citizen Data Scientist 664

11.1 OPENING VIGNETTE: Lessons Learned from Analytics Journey in an Organization

Chapters 1 and 10 highlight various analytics projects that were undertaken by Land O'Lakes. As noted in Chapter 1, Land O'Lakes is a cooperative of hundreds of dairy farmers. It started in 1921 and is a major producer of dairy products. A subsidiary of Land O'Lakes, Purina Animal Nutrition LLC (www.purinamills.com) is a national organization serving producers, animal owners and their families through more than 4,700 local cooperatives, independent dealers, and other large retailers throughout the United States. Over the years, they have developed over 235 nutritional solutions for animals across 24 animal species. By using a talented group of PhD nutritionists and veterinarians, they have conducted more than 24,000 research studies in fields across the country and registered over 125 patents. To date, their research teams have analyzed over 1,500 nutrient combinations to drive product development and rigorous testing of new diet formulations, innovative milk replacers, and nutritional programs. Chapters 1 and 10 highlight some data science projects undertaken by this organization. Mr. Manohar M. Lahoti, Manager of Data Science and Digital Products at Land O'Lakes, reflected on the lessons learned from the various projects. Those ideas are summarized in his contributions below.

ASKING QUESTIONS

Asking all the clarifying questions to the business users/researchers is a must before project planning. Questions play a very important role. Many times, the business users especially in the agriculture industry (likely true in many other industries as well) do not have full clarity when they reach out to the data science team. Going through Q&A sessions is helpful for both the business and data science teams.

Based on the discussion with the business stakeholders, data science professionals will have some clarity about the kind of problem you are solving with data science. Problem Statements can be categorized into two buckets:

- **Business Problem**

Business problems are the problems where you are adding value to the business, farmer and ranchers directly by solving them. These are the problems that are relatively short-term, and the data science team is asked and is responsible to optimize the key north star metrics, which will have a direct impact on revenue or other measures of interest.

- **Research Problem**

Research problems are the problems where the business or research teams are willing to learn some new things from the data and are often required to innovate a new method to solve these kinds of problems. These are long-term projects and often need help from academia. Eventually, the solutions are published, perhaps after filing for protecting the intellectual property.

DEFINING THE GOAL, VALUE, AND RIGHT BUSINESS/RESEARCH NORTH STAR METRICS

This step may seem very simple, but this is the most critical step and sometimes a tough one. This step paves the way for the data science team with the approach they are going to choose to solve the problem. The data science team often fails to see the bigger picture and jumps to solve the problem by modeling right away. Many times, an entirely new project begins at the end of the current project because often it does not meet the business expectations. Many of the readers are working for businesses so it is important to understand the scope of the data science problem in terms of the business metrics. A business/research metric is a quantifiable measure a business uses to track, monitor, and assess the success or failure of various business/research processes. The main point of using business/research metrics is to communicate an organization's progress toward their long- and short-term objectives.

IDENTIFY THE DATA SOURCES AND CHECK THE DATA QUALITY

Identifying the right data sources and later understanding the data quality are two very important steps before applying the data science techniques. We generally ask

- a. Can we access the data that exist within our organization to meet our project goals? This question helps us with the following issues:
 1. Need for external data sources
 2. Formats of the data sources—unstructured and structured data sources and the effort needed to pull all the data together in one place in a uniform structure
 3. Need for any manual intervention, such as labeling the data and coming up with different categories for the categorical data
 4. Calculated variables, lookups, and routines/functions to autopopulate some of the columns
- b. Have we considered data governance issues for data quality? This question will determine:
 1. Missing data issues
 2. Granularity of the data and related issues
 3. Messy data
 4. Incorrect data types
 5. Bad values

PREDICTIONS VS CAUSATION

This is a very common problem in the world of data science. Technically, it is known as a bias vs variance trade-off. Simple predictive models can be better explained, and you can define the relationship in terms of some mathematical formula or decision tree. However, most of the time these models underfit and cannot be used for predictions.

More complex AI/ML models are the black box and may be built with a heuristic approach. You will not understand the relationship between the data with these complex models because it iteratively fits and creates multiple layers before giving the final predictions. So, if the goal is to get accurate predictions, these are better. However, the con here is that these models are black box and clear relationships cannot be always defined.

CREATING THE RIGHT FRAMEWORK FOR DATA SCIENCE PROBLEMS

The solutions to the data science problem can be combinations of the following things:

1. Reports and dashboards
2. Exploratory data analysis, data visualization

3. Statistical analysis, hypothesis testing
4. Simple predictive models
5. Deep learning models
6. Causal models

So, depending upon the problem that needs to be solved, the approach will change. More often than not, you have to use multiple different techniques before arriving at an answer.

MAINTENANCE OF AI/ML MODELS

Depending on the scope of the project, the data science team must retrain the models deployed in the production systems and evaluate the results. The predictive models assume that the new data (test data) are very similar to the training data, and these are coming from a similar world. However, that is not always true, and thus many times predictions are way off in such situations.

EXPLAINING THE MODEL RESULTS TO BUSINESS AND PROVIDING THE RECOMMENDATION

Business stakeholders want things to be explained in terms of business metrics because ultimately that is what matters. However, you do not get the outputs from data science models in terms of business metrics. You get a bunch of statistical results. Explaining the appropriate and relevant statistical results with respect to business is an important step and often very hard.

TIMELINE EXPECTATIONS

Timelines depend on the quality of the data and what kind of problem you are solving. So, it is important to set realistic timeline expectations with the business team and give them a clear idea about all the intermediate steps you will be following.

INVOLVEMENT OF BUSINESS/PRODUCT TEAMS

At almost every checkpoint, the data science team needs to present their findings to the business clients and get their buy-in if they are moving on the right path. Staying on the right track is very important. Otherwise, you can get lost easily in the journey of solving the data science problems.

TALENT GAPS

Data science is needed in every single domain/vertical today. Data science is not just a mathematics or science discipline but it equally needs business and arts expertise. So, finding the right industry-specific talent is really tough in the current market. Employee retention rate is also poor as the data analytics job market is still hot.

LESSONS LEARNED FROM THIS VIGNETTE

This vignette reinforces many of the lessons the earlier chapters have mentioned. Too often, data scientists have a desire to jump to data analysis and modeling. However, it is far more important to understand the business problem and learn how to work with the business team to address the real issues. Furthermore, being able to tell the story of the results in the language understood by the client is of utmost importance. As we will see in this chapter, data science professionals need to be aware of many “soft” issues. Those are the topics we will cover in the following sections.

DISCUSSION QUESTIONS FOR THIS VIGNETTE

1. What is the difference between a business problem and a research problem?
2. Explain the difference between causation and predictions.
3. What are some of the data governance issues?

Adapted from a contribution by Manobar Mohanlal Lahoti Data Science and Digital Product Manager, Land O'Lakes Inc., Purina Animal Nutrition subsidiary.

11.2 IMPLEMENTING INTELLIGENT SYSTEMS: AN OVERVIEW

Now that you have learned the essentials of analytics, data science, artificial intelligence, and decision support activities, you may be tempted to ask: What can I do with all this in my organization? You learned about the great benefits, and you read about numerous companies that use intelligent systems. So, what you should do next? First, read some of the resources recommended in this book so you will have a better understanding about the technologies. Next, read this chapter that deals with the major issues that are involved in **implementing intelligent systems** in organizations.

Implementing business analytic/AI systems can be a complex undertaking. In addition to specific issues found in intelligent systems, there are issues that are common to many other computer-based information systems. In this section, we describe the major types of issues, some of which are discussed in this chapter. For several success AI implementation factors revealed in a survey of 3,000 executives, see Bughin, McCarthy, and Chui (2017).

The Intelligent Systems Implementation Process

This chapter is divided into three parts. In the first part, we describe some managerial-related implementation issues. In the second part, we describe the impacts of intelligent technologies on organizations, management, work, and jobs. The last part extends the issues related to implementation of these technologies, particularly regarding ethics and bias.

The implementation process of intelligent systems is similar to the generic process of other information systems. Therefore, we will present it only briefly. The process is illustrated in Figure 11.1.

THE MAJOR STEPS OF IMPLEMENTATION The major steps are:

Step 1 Need assessment. Need assessment needs to provide the *business case* for the intelligent systems, including their major parts. (This is a generic IT step and will not be discussed here.)

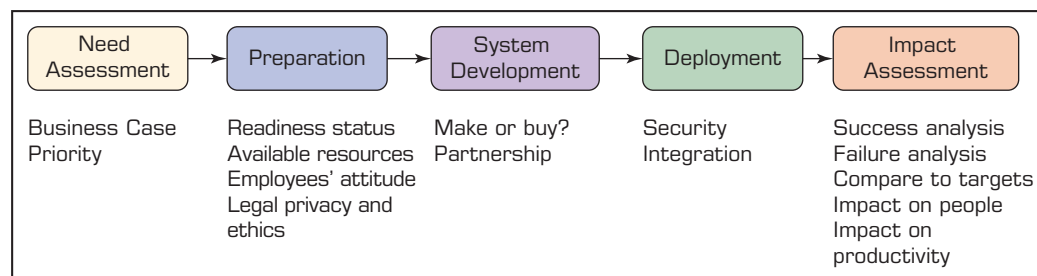


FIGURE 11.1 Implementation Process. Drawn by E. Turban.

- Step 2 Preparations.** In this step, it is necessary to examine the organization readiness for analytics and AI. It is necessary to check available resources, employees' attitudes for the change, projects' priorities, and so on. This generic IT activity will not be discussed here. However, it is useful to think about legal, privacy, and ethical issues as they are related to intelligent technologies as described in Section 11.5.
- Step 3 System acquisition.** Organizations need to decide on in-house or outsourcing approach (make or buy) or on a combination of the two and possibly with partnership with a vendor or another company. A consultant may help at this step. It is a generic IT step that will not be discussed here.
- Step 4 System development.** Regardless of who will develop the system, certain activities need to be done. These include security, integration with other systems, project management preparation, and other activities. Again, many of those are generic and will not be described here. Only selected ones are described in Section 11.3.
- Step 5 Impact assessment.** It is necessary to check the performance of the systems against plans. Again, this is a generic issue that will not be covered here.

► SECTION 11.2 REVIEW QUESTIONS

1. List the major steps in the implementation process.
2. Why is implementation an important subject?
3. Describe the major impact areas of intelligent systems.

11.3 SUCCESSFUL DEPLOYMENT OF INTELLIGENT SYSTEMS

Many experts, consultants, and researchers provide suggestions regarding intelligent systems' successful deployment. Given the importance of the topic, it is clear that companies need to get ready for the mass arrival of AI and other intelligent technologies. Here are some topics related to deployment strategy:

- When to embark on intelligent projects and how to prioritize them.
- How to decide whether to do it yourself or use partners, or to outsource.
- How to justify investments in intelligent projects.
- How to overcome employees' resistance (e.g., fear of job loss).
- How to arrange appropriate people-robot teams.
- How to determine which decisions to fully automate by AI.
- How to protect intelligent systems (security) and how to protect privacy.
- How to handle possible loss of jobs and retraining of employees
- How to determine whether you have the necessary up-to-date technology.
- How to decide what support top management should provide.
- How to integrate the system with business processes.
- How to find qualified personnel for building and using intelligent systems.

For more strategy issues, see Kiron (2017). We cover only several topics in this section and provide references to more. Most of the implementation topics are generic in nature and will not be covered here.

Top Management and Implementation

According to Chui et al. (2017), from McKinsey & Company, "Senior executives need to understand the tactical as well as the strategic opportunities (of AI), redesign their organizations, and commit to helping shape and debate about the future of work."

Specifically, the executives need to plan for integrating intelligent systems into their workplace, making a commitment to conduct a participating environment for the changes and provide sufficient resources. Many executives may be aware that intelligent systems will change their business, but they do not do much about it or know how to. Sometimes, if things appear to be going well, the senior management is reluctant to invest in long-term technologies because it is more expedient to return that money to shareholders. A major airline in the United States had to face a lot of criticism for having had to cancel thousands of flights (from December 2022 to January 2023) in light of some weather issues that made its older technology systems unable to provide scheduling and operational decisions. Much of the blame was laid to its older technology systems. Senior management was blamed by employees, customers, and analysts for not investing in technologies in favor of stock buybacks and other ways to reward shareholders.

KPMG, a large management service consultant, provides the following steps regarding digital labor. “KPMG’s holistic approach—from strategy through execution will assist companies on each step of implementation. The steps are:

- Establishing priority areas for technological innovation.
- Developing a strategy and a plan for the employees.
- Identify providers and partners for plans’ execution.
- Establishing a strategy and plans to realize benefits from the digital labor initiatives.”

A complete guide for KPMG is provided by Kiron (2017). It includes robotic process automation, enhanced process automation, and cognitive automation.

System Development Implementation Issues

Since *AI* and *business analytics* are broad terms, describing several technologies whose maturity levels vary, implementation issues may vary considerably. Shchutskaya and Zharovskikh (2020) identify at least the following three major problems:

1. *Development approach.* Business analytic and AI systems require an approach different from that of other IT/computer systems. Specifically, it is necessary to identify and deal with different and frequently large data sources. It is necessary to cleanse and curate these data. Also, if learning is involved, one needs to use machine training. Thus, special methodologies are needed.
2. *Learning from data.* Many AI and business analytics involve learning. The quality of the input data determines the quality of the applications. Also, the learning mechanism is important. Therefore, data accuracy is critical. In learning, systems must be able to deal with changing environmental conditions. Data should be organized in databases, not in files.
3. *No clear view is available of how insights are generated.* AI, IoT, and business analytic systems generate insights, conclusions, and recommendations based on the analysis of the data collected. Given that data are frequently collected by sensors and there are different types of them, we may not have a clear view of the insights that are generated.

Related important areas include problems with Big Data, ineffective information access, and limited integration capabilities (discussed next).

Connectivity and Integration

As part of the development process, it is necessary to connect the AI and analytic applications to existing IT systems, including the Internet, and other intelligent systems.

Example

The Australian government commissioned Microsoft in August 2017 to build hyperscale cloud regions to unlock the power of intelligent technologies. The system is expected to dramatically modernize how the government processes data and delivers services to its citizens. The system can handle both unclassified and protected data. The infrastructure is built inside, or near, the government data centers. The system will enable the government to use innovative applications based on machine learning, bots, and language translation, and it will improve healthcare, education, social services, and other government operations. Finally, the system will increase both security and privacy protection.

Integration needs to be done with almost every system that is being impacted by AI or business analytic. For example, it is necessary to integrate intelligent applications both to a digital marketing strategy and to marketing implementation. For a discussion, see searchenginejournal.com/artificial-intelligence-marketing/200852/ (accessed January 2023).

To overcome the integration difficulty, Huawei of China (a cellphone producer) is installing an AI system with its knowledge base inside the chips of its products. Other phones' manufacturers rely on connecting to the "cloud" to interact there with AI knowledge. For the implications on IoT connectivity, see Rainie and Anderson (2017). For considerations regarding IoT connectivity providers, see Baroudy et al. (2018).

Security Protection

Many intelligent applications are managed and updated in the "cloud" and/or connected to the regular Internet. Unfortunately, by adding Internet connection, new vulnerabilities may be created. Hackers use intelligent technologies to identify these vulnerabilities. For how criminals use AI and related issues, see Crosman (2017). In Section 11.9, we discuss the potential dangers of robotics. The safety of passengers in self-driving cars and others who may be involved in collisions with the self-driving cars is an important safety issue as well. Also, the safety of people working near robots has been researched for many decades. In addition, hacking robots, chatbots, and other intelligent systems are areas that require attention. Finally, the safety of robots themselves when they work on the streets is an issue. Some people may attack them, for instance, because people may feel uncomfortable or fearful.

Leveraging Intelligent Systems in Business

There are many ways to leverage intelligent systems, depending on the nature of the applications. Catliff (2017) suggests the following ways to do this, leveraging the intelligent technology capabilities to increase efficiency and provide more customer care. Specifically, he suggested:

1. Customize the customer experience (e.g., for interactions with customers).
2. Increase customer engagement (e.g., via chatbots).
3. Use intelligent technologies to detect problems and anomalies in data.

Singh (2017a) recommends the following as critical success factors: discover, predict, justify, and learn from experience. Ross (2018) raised the issue of the need to upgrade employees' skills and build an empowered AI-savvy workforce. One of the most important issues is how to handle the fear of job loss of employees. This is discussed in Section 11.8.

Intelligent System Adoption

Most of the issues related to intelligent systems' adoption are the same as or similar to that of any information systems. For example, employees may resist change, management may not provide sufficient resources, there could be a lack of planning and coordination, and so on. To deal with such issues, Morgan Stanley drew ideas from hundreds of conversations with experts. One important issue is to have an appropriate deployment and adoption strategy that should work in harmony with the implemented technologies and the people involved. In general, the generic adoption approach to information systems should work here, too.

SECTION 11.3 REVIEW QUESTIONS

1. Describe the systems deployment process.
2. Discuss the role of top management in deploying intelligent systems.
3. Why is connectivity such an important issue?
4. Describe system development issues.
5. Discuss the importance of security and safety, and how to protect them.
6. Describe some issues in intelligent systems adoption.

11.4 IMPLEMENTING IoT AND MANAGERIAL CONSIDERATIONS

In Chapter 10, we presented several successful IoT-based applications. The results so far are more than encouraging, especially in areas such as monitoring equipment performance to improve its operation and maintenance. However, this is only the tip of the iceberg. As we indicated earlier, the IoT can change everything. In this section, we present some of the major issues that are related to successful IoT implementation. Although there is considerable excitement about the growth and the potential of the IoT, there are issues that managers should be aware of.

Major Implementation Issues

McKinsey's Global Institute (Bughin et al., 2015) put together a comprehensive *Executive's Guide to the Internet of Things*. This guide identifies the following issues:

- **Organizational alignment.** Although it is true of several other technology initiatives, with IoT, the opportunities for operational improvements and creating new business opportunities mean that IT and operational personnel have to work as one team rather than separate functions. As noted by the guide's authors, "IoT will challenge other notions of organizational responsibilities. Chief financial, marketing, and operating officers, as well as leaders of business units, will have to be receptive to linking up their systems."
- **Interoperability challenges.** Interoperability is a huge detriment thus far in the growth of IoT applications. Few IoT devices connect seamlessly with each another. Second, there are many technological issues regarding connectivity. Many remote areas do not yet have proper Wi-Fi connection. Issues related to Big Data processing are also responsible for slow progress in IoT adoption. Companies are trying to reduce data at the sensor level so that only a minimal amount goes into clouds. Current infrastructure hardly supports the huge amount of data collected by IoT. A related problem is retrofitting sensors on devices to be able to gather and transmit data for analysis. In addition, it will take time for consumers to replace their analog objects with new IoT digital smart products. As an example, it is easier for people to replace mobile phones than a car, kitchen appliances, and other things that can benefit from having a sensor and being connected to IoT, assuming they can afford these.

- **Security.** Security of data is an issue in general, but it is an even bigger one in the context of IoT. Each device that is connected to IoT becomes another entry point for malicious hackers to get into a large system or at the least operate or corrupt a specific device. There are stories of hackers being able to breach and control automated functions of a car or to control a garage door opener remotely. Such issues require that any large-scale adoption of IoT involves security considerations from the very beginning.

Given that the Internet is not well secured, applying IoT networks requires special security measures, especially in the wireless sections of the networks. Perkins (2016) summarizes the situation as follows: “IoT creates a pervasive digital presence connecting organizations and society as a whole. New actors include data scientists; external integrators; and exposed endpoints. Security decision makers must embrace fundamental principles of risk and resilience to drive change.” For a free e-book about IoT, see McLellan (2017b).

Additional issues follow.

- **Privacy.** To ensure privacy, one needs a good security system plus a privacy protection system and policy. Both may be difficult to construct in IoT networks due to the large size of the networks and the use of the less protected Internet. For advice from top security experts, see Hu (2016).
- **Connection of the silos of data.** There are millions of silos of data on the Internet and many of them need to be interconnected in specific IoT applications. This issue is known as the need for a “fabric” and connectivity. This can be a complex issue for applications that involve many different silos belonging to different organizations. Connectivity is needed in machine to machine, people to people, people to machines, and people to services and sensors. For a discussion, see Rainie and Anderson (2017).
- **Preparation of existing IT architectures and operating models for IoT can be a complex issue in many organizations.** Integrating IoT into IT is critical for the data flow needed by the IoT and IoT-processed data to flow back to actions.
- **Management.** As in the introduction of any new technology, the support of top management is necessary. Bui (2016) recommends hiring a *chief data officer* in order to succeed in IoT due to the need to deal with silos of data described earlier. Using such a top manager can facilitate information sharing across all business functions, roles, and levels. Finally, it solves departmental struggles to own and control the IoT.
- **Connected customers.** There is evidence of an increased use of IoT in marketing and customer relationships. In addition, the IoT drives increased customer engagement. According to Park (2017), a successful deployment of IoT for customers requires “connected customers.” The connection needs to be for data, decisions, outcomes, and staff related to any contacts relevant to the IoT and marketing. The Blue Hill research organization provides a free report on this issue (see Park). IoT enables a better connection with key clients and improves customer service. Of special considerations are hospitality, healthcare, and transportation organizations.

Strategy for Turning Industrial IoT into Competitive Advantage

With so many implementation issues, an implementation strategy is necessary.

IoT collects large amounts of data that can be used to improve external business activities (e.g., marketing) as well as internal operations. SAS (2017) proposed a strategy cycle that includes the following steps:

1. *Specify the business goals.* They should be set with perceived benefits and costs so the initiatives can be justified. This step involves a high level of planning and examination of resources. Initial return on investment (ROI) analysis is advisable.
2. *Express an analytic strategy.* To support ROI and prepare a business case, it will be necessary to plan how Big Data will be analyzed. This involves the selection of

an analytic platform, which is a critical success factor. An examination of emerging AI technologies, such as deep learning, may be conducted. An appropriate selection will ensure a powerful IoT solution.

3. *Evaluate the needs for edge analytics.* Edge analytics is a technology that is needed for some, but not all, applications. It is designed to introduce real-time capabilities to the applications. It also filters data to enable automated decision-making, frequently in real time because only relevant data results from the filtering.
4. *Select appropriate analytics solutions.* There are numerous analytic solutions on the market offered by many vendors. In using one or several for IoT, it is necessary to **consider** several criteria such as fitness for IoT, ease of deployment, ability to minimize project risks, sophistication of the tools, and connection to existing IT systems (e.g., the quality of IoT gateways). Sometimes it is a good idea to look at a group of vendors that offer combined products (e.g., SAS and Intel). Finally, appropriate infrastructures, such as high-performance cloud servers and storage systems, need to be examined. These must work together as a scalable, effective, and efficient platform.
5. *Continues improvement closes the loop.* Like in any strategy cycle, performance should be monitored, and improvements in various steps of the process need to be considered, especially since IoT is evolving and changing rapidly. The extent of goal achievement is an important criterion and upgrading the goals should be considered.

A summary of the process is provided in Figure 11.2.

Weldon (2015) suggests the following steps for successful IoT implementation. These are not much different from general implementation steps identified earlier in this section:

- Develop a business case to justify the IoT project including a cost-benefit analysis and a comparison with other projects.
- Develop a working prototype. Experiment with it. Learn and improve it.
- Install the IoT in one organizational unit. Experiment with it. Learn lessons.
- Plan an organization-wide deployment if the pilot is a success. Give special attention to data processing and dissemination.

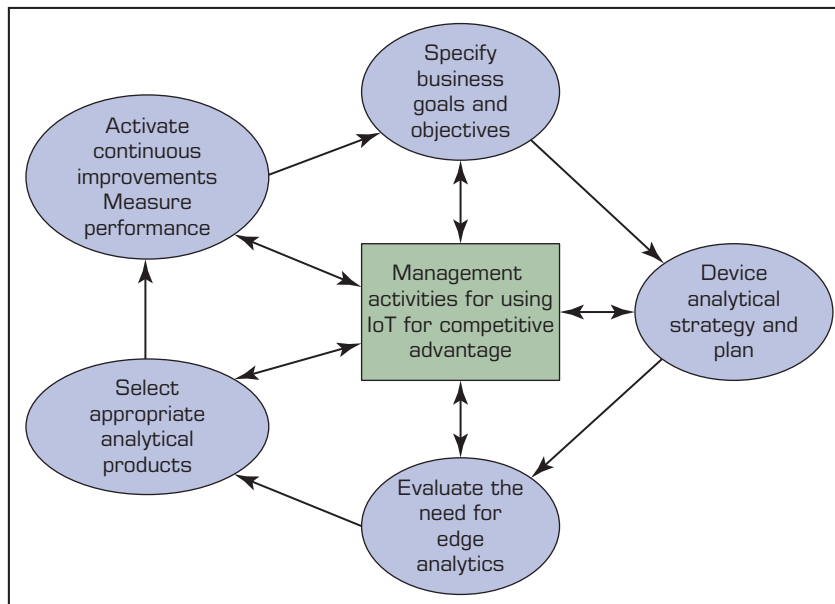


FIGURE 11.2 The IoT Strategy Cycle.

► SECTION 11.4 REVIEW QUESTIONS

1. What are some of the unique IoT implementation considerations as compared to those discussed in Section 11.3?
2. What are edge analytics?
3. What are some of the unique data issues in implementing IoT applications?

11.5 LEGAL, PRIVACY, AND ETHICAL ISSUES

As data science, analytics, cognitive computing, and AI grow in reach and pervasiveness, everyone may be affected by these applications. Just because something is doable through technology does not make it appropriate, legal, or ethical. Data science and AI professionals and managers have to be very aware of these concerns. Several important legal, privacy, and ethical issues are related to intelligent technologies, and they are interrelated. For example, several privacy issues are parts of ethics or have legal aspects. Here we provide only representative examples and sources as pointed out in Chapter 1. Our goal here is only to give the reader an exposure to these issues. For why we should care about the legal, ethical, and privacy of AI, see Krigsman (2017) and the next section.

Legal Issues

The introduction of intelligent technologies may compound a host of legal issues already relevant to computer systems. For example, questions concerning liability for the actions of advice provided by intelligent machines are beginning to be considered. In this section, we provide a sample of representative issues. Many other issues exist.

In addition to resolving disputes about the unexpected and possibly damaging results of some intelligent systems, other complex issues may surface. For example, who is liable if an enterprise finds itself bankrupt as a result of using the advice of an AI-based application? Will the enterprise itself be held responsible for not testing the system adequately before entrusting it with sensitive or volatile issues? Will auditing and accounting firms share the liability for failing to apply adequate auditing tests? Will the software developers of intelligent systems be jointly liable? As self-driving cars become more common, who is liable for any damage or injury when a car's sensors, network, or AI system fail to function as planned? A few recent cases involving a Tesla car accident where the driver died in a crash while the car was allegedly on "autopilot" mode have brought this issue to the front pages of newspapers and the legal profession.

A Sample of AI Potential Legal Issues

- What is the value of an expert opinion in court when the expertise is encoded in a computer?
- Who is liable for the wrong advice (or information) provided by an intelligent application? For example, what happens if a physician accepts an incorrect diagnosis made by a computer and performs a procedure that results in the death of a patient?
- What happens if a manager enters an incorrect judgment value into an intelligent application and the result is damage or a disaster?
- Can management force experts to contribute their expertise to an intelligent system? How will they be compensated?
- Who owns the knowledge that the new AI systems such as ChatGPT and Google's Bard gather and is then employed by an end user in some application?
- Is it okay for self-driving cars with in-vehicle back-up drivers to drive on public roads? (Already yes, in a few states, but still not a national policy.)

- Who should regulate driverless car: cities, states, or the federal government?
- Who should be held liable for any accidents caused by self-driving cars?
- Should delivery robots be allowed on sidewalks?
- Should robots have human rights? (What if they are citizens like Sophia in Saudi Arabia?) If they get rights, should they have legal responsibilities as well?
- Can companies be sued for the content recommended by their algorithms? A major case is under review by the US Supreme Court (see: <https://cacm.acm.org/magazines/2023/3/270203-a-legal-challenge-to-algorithmic-recommendations/>)

Example: Intellectual Property Protection

Ownership and protection of intellectual property is very important for technology-related companies. Indeed, it is a key enabler of technology innovation. In the case of intelligent technologies, these issues present some unique dimensions.

LEGAL ISSUES OF INTELLIGENT TECHNOLOGIES Several of the ethical issues described later need to be combined with legal issues. For example, take robots' legal rights. Do we need these rights? What for (an ethical issue)? Then, it will be necessary to develop the legal rights. Facebook, for example, has had legal issues regarding face recognition. Safety rules for robots were developed a long time ago. At the moment, there are very few laws regarding intelligent technologies. Most of the laws relate to safety.

AI AND LAW In addition to laws related to robotics and AI, there is a subfield of AI that is concerned with AI applications to the legal profession and the solution of some legal problems. According to Donahue (2018), the following are some major topics:

- Analyzing legal-related data (e.g., regulatory conflicts) to detect pattern
- Providing legal advice to consumers
- Document review
- Analyzing contracts
- Supporting legal research
- Predicting results (e.g., likelihood to win)
- AI impact on the legal profession

AI can execute routine legal-related tasks such as managing documents and drafting contracts. Legal issues may be strongly connected to our next topic, privacy.

OWNERSHIP OF CONTENT USED FOR TRAINING AI MODELS As we noted in multiple earlier chapters and the reader is likely already aware, there is much current enthusiasm about the potential applications of Generative AI to create new content based on a user's request. Companies such as ChatGPT, Stability AI, Midjourney Inc., and DeviantArt are touting their systems which let a user quickly create an image or text based upon a user's specified parameters. These systems are clearly trained using thousands if not millions of images and other knowledge and parameters to be able to respond to user's specifications for images. These companies build such corpus by scraping and compiling these images from various Web sites and online archives. There are multiple lawsuits already filed by various artists and image collection companies such as Getty Images to stop using the content of various artists without their permission, and to compensate them for their creative work. For example, read or listen to this story (Setty 2023) at <https://news.bloomberglaw.com/ip-law/first-ai-art-generator-lawsuits-threaten-future-of-emerging-tech>

(accessed February 2023). It also notes related lawsuits already underway against ChatGPT and GitHub's CoPilot offering, which has been trained using many programmers' submission of codes to GitHub.

Privacy Issues

Privacy means different things to different people. In general, **privacy** is the right to be left alone and the right to be free from unreasonable personal intrusions. Privacy has long been related to legal, ethical, and social issues in many countries. The right to privacy is recognized today in every state of the United States and by the federal government either by statute or by common law. The definition of *privacy* can be interpreted quite broadly. However, the following two rules have been followed fairly closely in past court decisions: (1) The right of privacy is not absolute. Privacy must be balanced against the needs of society. (2) The public's right to know is superior to the individual's right to privacy. These two rules show why it is difficult, in some cases, to determine and enforce privacy regulations. Privacy issues online have specific characteristics and policies. One area where privacy may be jeopardized is discussed next. Privacy issues are getting more and more important as the amount of data generated on the Internet is increasing exponentially, and in many cases, it is lightly secured. For an overview of privacy as it relates to AI, see Provazza (2017).

COLLECTING INFORMATION ABOUT INDIVIDUALS Intelligent technologies aim to provide targeted services and marketing to consumers; they do so by collecting information about these customers. In the past, the complexity of collecting, sorting, filing, and accessing information manually from numerous government agencies and other public databases was, in many cases, a built-in protection against the misuse of private information. The Internet in combination with large-scale databases has created an entirely new dimension of accessing and using data. The inherent power in intelligent systems that can access vast amounts of data and interpret them can be used for the good of society. For example, by analyzing records with the aid of business analysis, it is possible to eliminate or reduce fraud, crime, government mismanagement, tax evasion, welfare cheating, family-support filching, employment of illegal workers, and so on. However, what price must the individual pay in terms of loss of privacy so that the government can better apprehend criminals? The same is true on the corporate level. Private information about employees may aid in better corporate decision-making, but the employees' privacy may be compromised.

The use of AI technologies in the administration and enforcement of laws and regulations may increase public concern regarding privacy of information. These fears, generated by the perceived abilities of AI, will have to be addressed at the outset of almost any AI development effort.

VIRTUAL PERSONAL ASSISTANTS Amazon's Echo/Alexa, Google Assistant, iPhone Siri, and other similar devices listen to what is going on. They also may take photos. In other words, your voice assistant is spying on you.

Most advanced is the Echo/Alexa pair. You can ask Alexa and any of these devices to buy products today. Privacy advocates are not happy, but customers may be. For example Alexa can act as a fashion consultant using *style check*. The system combines the knowledge of a fashion specialist and AI knowledge. A recommendation provides you with two photos at a time, telling you which one to buy (based on color, current trends, etc.). To make it useful, Amazon is improving the privacy. This may not be easy since your record is stored in Amazon's cloud. Huff (2017) provides arguments about the risks of the assistant and the protection provided by Amazon.

MOBILE USER PRIVACY Many users are unaware of the private information being tracked through their smartphone usage. Many apps collect user data that track each phone call as it moves from one cell tower to another, from GPS-enabled devices that transmit users' locations, and from phones transmitting information at Wi-Fi hotspots. Major app developers claim that they are extremely careful and protective of users' privacy, but it is interesting to note how much information is available through the use of a single device, especially when smartphones contain more and more AI components. Apple claims to try and limit information sharing by the apps, but mobile users' privacy is a shaky concept at best.

PRIVACY IN IoT NETWORKS For privacy and security of the Internet of Things (IoT), see Hu (2016). More data are flowing with IoT networks. Note that AI data privacy issues are on the rise, especially when AI deals with consumers' data. There is a growing amount of data collected, for example, by machine learning and chatbots. Also, in the enterprise, employers collect and analyze more data on employees. How do we protect the data and guard against their misuse?

RECENT TECHNOLOGY ISSUES IN PRIVACY AND ANALYTICS With the growth of Internet users in general and mobile device users in particular, many companies have started to employ intelligent technologies to develop profiles of users on the basis of their device usage, surfing, and contacts. *The Wall Street Journal* has an excellent collection of articles titled "What They Know" (WallStreetJournal.com, 2016). These articles were compiled several years ago but continue to highlight the latest technology and privacy/ethical issues. One of the companies mentioned in this series is Rapleaf (now part of Towerdata). Rapleaf's technology claims to be able to provide a profile of a user just by knowing his or her e-mail address. Clearly, Rapleaf's technology enables it to gather significant related information. Another company that aims to identify devices on the basis of their usage is BlueCava. BlueCava technology attaches a personal profile to be able to recognize a user as one individual or a household, even though the user may be working with multiple mobile devices and laptops. All these companies employ analytics such as clustering and association mining to develop profiles of users. Of course, many of the analytics start-ups in this space claim to honor user privacy, but violations are often reported. For example, Rapleaf was collecting unauthorized information from Facebook users and was subsequently banned from Facebook. One user reported that an hour after he gave his e-mail address to a company that specializes in user information monitoring (reputation.com), the company was able to discover his Social Security number. So, violations of privacy create fears of criminal conduct regarding information. This area is a big concern overall and needs careful study. These examples not only illustrate the power of analytics in being able to learn more about target customers but also serve as a warning to AI and analytics professionals about being sensitive to privacy and ethical issues.

Another related application area of **privacy concerns** is analyzing employee behaviors based on the data collected from sensors that employees wear in a badge. One company, Humanyze, has reported several such applications of its sensor-embedded badges. These sensors track all movements of an employee. Indeed, such application have given rise to another term: workplace analytics.

Example: Using Sensors and IoT to Observe Bankers at Barclays Bank

Using heat and motion sensors, Barclays tracks how long its bankers are at their desks. The system was installed in the London, United Kingdom, branches. The formal explanation was to find out the occupancy of the cubes in the bank to optimally allocate and possibly reduce office space. The IoT network provided dashboards showing which

workstations (cubes) were underutilized, and what the usage trend was. The bank informed the employees and the union that this project did not measure productivity, only space utilization. The results can be used to better manage energy consumption in the cubes and to schedule a flexible work environment. As a result, Barclays was able to save office space and rent it out for \$45 million a year.

The bank uses a similar tracking system to find out how much time that different types of employees spend with customers. The union is watching this IoT application carefully to ensure that it is not used to spy on employees. Other banks in England use similar systems.

Of course, situations like those described create major privacy concerns. Should companies be able to monitor their employees this intrusively? What are the privacy rights of employees in such situations?

Finally, there is a possibility of ransomware, or hackers' attacks on robots, which could be used against businesses whose employees use such robots. Smith (2018) reported on research that identified 50 vulnerabilities in robots. Ransomware attacks may interrupt operations, forcing organizations to pay substantial ransoms.

OTHER ISSUES OF POTENTIAL PRIVACY VIOLATION The following are some more examples of potential privacy violations in the intelligent technology world:

- Delaware police are using AI dashcams to look for fugitives in passing cars. Photos and videos taken are sent to the clouds and analyzed there by AI algorithms.
- Facebook's face recognition systems create concerns regarding privacy protection.
- Epicenter offers its employees a microchip implant. It acts like a swipe card, opens doors, buys you food in the company store, and much more. But management can track you too. It is given only to volunteers.

Who Owns Our Private Data?

With the recent growth of data from our use of technology and the companies' ability to access and mine it, the privacy debate also leads to the obvious issue of whose property any user's data is. Take an example of a relatively new car. The car is equipped with many sensors starting with tire pressure sensors to GPS trackers that can keep track of where you have gone, how fast you were driving, when you changed lanes, and so on. The car may even know the passenger's weight added to the front seat. As Welch notes, a car connected to the Internet (most new cars are!) can be a *privacy nightmare* for the owner or a data "gold mine" for whoever can possess or analyze these data. A major battle is brewing between automobile manufacturers and technology providers such as Apple (CarPlay) and Google (Android Auto) on who owns these data and who can access them. This is becoming more crucial because as cars become smarter and eventually self-driving, the driver/passenger in the car could be a highly targeted prospect for marketers' services. For example, Google's Waze app collects GPS data from millions of users to track traffic information and help users find the best routes; but it also displays pop-up ads on the users' screens. Yelp, Spotify, and other apps popularly used in cars have similar approaches.

The bottom line is that intelligent systems professionals and users must be aware of the legal and ethical issues involved in collecting information that may be privileged or protected. Privacy issues are considered in many cases as important components of ethics.

Ethics Issues

Several ethical issues are related to intelligent systems. Personal values constitute a major factor in the issue of ethical decision-making. The study of ethical issues is complex because of their multidimensional nature. One story that upset many users (although it was

not illegal) some time ago was Facebook's experiment to present different News Feeds to the users and monitor their emotional reactions as measured by replies, likes, sentiment analysis, and so on. Most companies, including technology companies, run user testing to identify the features most liked or disliked and fine-tune their product offerings accordingly. Because Facebook is so large, running this experiment without the users' informed consent was viewed as unethical. Indeed, Facebook acknowledged its error and instituted a more formal review through Internal Review Boards and other compliance mechanisms for future testing.

Morgan (2017) said that it is necessary to be at the foundations of what AI does for both vendors and customers in order to stay ethical and have transparency of each situation. This way people can stay honest and adhere to the goals of AI, so it can play a significant role in our life and work.

Ethical Issues of Intelligent Systems

Many people have raised questions regarding ethical issues in AI, robotics, and other intelligent systems. For example, Bossmann (2016) raised the following issues:

- What is their impact on jobs (see Section 11.8)?
- How do machines (i.e., robots) affect our behavior and interactions?
- How can wealth created by intelligent machines be distributed?
- How can intelligent applications mistakes be guarded against? For example, how long should training programs in machine learning be?
- Can intelligent systems be fair and unbiased? How can bias in creation and operation of AI systems be eliminated?
- How can intelligent applications be kept safe from adversaries?
- How can systems be protected against unintended consequences (e.g., accidents in robot operations)? For example, Facebook researchers had to shut down an AI system that created its own poor language.
- How can we stay in control of a complex intelligent system?
- Should we develop robots' legal rights? How can we define and plan human treatment of intelligent machines?
- Should we allow a self-governing robot society to exist with ours?
- To what extent should we influence unintended robots' behavior (or even be able to)?
- How would we get around the question of smart machine ownership?

Additional issues are:

- Electronic surveillance
- Ethics in business intelligence (BI) and AI systems design
- Software piracy
- Invasion of individuals' privacy
- Use of proprietary databases and knowledge bases
- Use of personal intellectual property such as knowledge and expertise for the benefits of companies and the payment to the contributors
- Accuracy of data, information, and knowledge
- Protection of the rights of users
- Accessibility to information by AI users
- The amount of decision-making to delegate to intelligent machines.
- How AI can fail due to inappropriate ethics
- The ethics of legal analytics (Goldman, 2018)

Other Topics in Intelligent Systems Ethics

- Machine ethics is a part of the ethics of AI that is concerned with the moral behavior of artificially intelligent beings.
- Robotics is concerned with the moral behavior of designers, builders, and users of robots.
- Microsoft's Tay chatbot was closed due to its inability to understand many irrelevant and offending comments.
- Some are afraid that algorithm-based technologies, including AI, may become racists. We discuss algorithms and the bias issues in the next section.
- According to Spangler (2017), self-driving cars may one day face a decision of whom to save and whom to kill.
- Voice technologies enable the identification of callers to AI machines. This may be great on one hand, but it creates privacy concerns on the other.
- One area in which there are considerable ethical concerns (frequently combined with legal concerns) is the healthcare/medical field. Given the large efforts by Alphabet and many other health analytics organizations, this is not surprising.

COMPUTER ETHICS IN GENERAL Computer ethics focuses on the behavior of people toward information systems and computers in general. The study of ethics in intelligent systems is strongly related to the ethics of computers and information systems in general. The following are some resources.

THE TEN COMMANDMENTS OF COMPUTER ETHICS This well-known document is published by cybercitizenship (cybercitizenship.org/ethics/commandments.html).

1. Thou shalt not use a computer to harm other people.
2. Thou shalt not interfere with other people's computer work.
3. Thou shalt not snoop around in other people's files.
4. Thou shalt not use a computer to steal.
5. Thou shalt not use a computer to bear false witness.
6. Thou shalt not use or copy software for which you have not paid.
7. Thou shalt not use other people's computer resources without authorization.
8. Thou shalt not appropriate other people's intellectual output.
9. Thou shalt not think about the social consequences of the program you write.
10. Thou shalt not use a computer in ways that show consideration and respect.

MIT Media Lab and the Center for Internet & Society at Harvard University manage an initiative to research ethical and governance topics in AI. SAS, a major analytical and AI vendor, proposed three essential steps for AI ethics as described in sas.com/en_us/insights/articles/analytics/artificial-intelligence-ethics.html.

SECTION 11.5 REVIEW QUESTIONS

1. List some legal issues of intelligent systems.
2. Describe privacy concerns in intelligent systems.
3. In your view, who should own the data about your use of a car? Why?
4. List ethical issues in intelligent systems.
5. What are the 10 commandments of computer ethics?

11.6 ETHICAL/RESPONSIBLE/TRUSTWORTHY AI

We began discussion of ethics in AI in the last section but devote another full section to explore this issue in more detail. We first begin with a few stories that have been reported in the media.

The Curious Case of BlenderBot 3.0

We discussed chatbots in Chapter 2 and again in Chapter 10. One chatbot introduced by Meta is called BlenderBot 3. BlenderBot 3 can browse the Internet to chat about virtually any topic. Meta designed BlenderBot such that it learns how to improve its skills and safety through natural conversations and feedback from people “in the wild.” (Source: <https://ai.facebook.com/blog/blenderbot-3-a-175b-parameter-publicly-available-chatbot-that-improves-its-skills-and-safety-over-time/>.)

Meta released BlenderBot for testing by users in the United States in early August 2022. While the ability of BlenderBot to autonomously do all of the browsing and conversing looks like an advancement in the world of conversational AI, the early reports from several test users looked concerning as well. Many users reported that BlenderBot made anti-Semitic and biased remarks while chatting with them. Though it might seem beyond our comprehension that a machine could replicate the human way of thinking, it is not so surprising when one realizes that the BlenderBot might have been trained on the vast data originating from public activity on the Internet. In Meta’s defense, they implemented several measures to remove such bias from the model they were trying to build, but still, a few found their way into the final product (Meyer 2022)

A chatbot making offensive remarks in the early stages of its development might not ring any alarm bells. But imagine the situation if we fail to eradicate such human biases from the intelligent systems that automate crucial tasks and decision-making in critical fields like healthcare and the judicial system. Time and time again, there were several reports of intelligent systems turning rogue due to human biases, and adding to those reports is the fact that many AI systems are black-box models that are difficult for the public to interpret. It led to a growing suspicion in people toward the increasing usage of intelligent machines these days.

Before trying to generalize what ethical AI means, it helps to look at a few unethical use cases of AI.

Amazon’s Scrapped AI-Powered Recruiting Tool

A report published on Reuters.com in 2018 shed light on Amazon’s failed attempt to introduce automation in its hiring process by creating an AI tool that screens the applicant resumes and shortlists the top matches. But Amazon scrapped this tool after the company insiders realized that it showed discrimination against female applicants. This tool was trained on a large set of resumes submitted to the company over some time by the applicants, most of whom were male. Hence, the AI model behind this tool learned to penalize any resume with the word “woman’s” appearing on it (see Dustin 2018).

COMPAS

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is an AI-powered tool used for decision-making in several US courts. When defendants are jailed, they are made to answer a 137-questionnaire survey. These responses are fed into the COMPAS software to predict the risk of recidivism (the probability that a person will commit a crime again). A judge could use this prediction along with the facts of the case on hand in determining a prisoner’s sentence. In a perfect world, this software could lead to a reduction in the crime rate. But a report published by ProPublica, a nonprofit organization based in New York, shows that this software is not flawless. One of the findings published in this report points to racial bias in predicting recidivism (Larson et al 2016).

<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

In the cases of Amazon and COMPAS, it is unlikely that the engineers and developers designed these tools to discriminate based on gender and race. It is just a likely consequence of using incorrectly sampled data to train these models. But this shows how critical it is to screen unconscious and inherent biases from inputs to AI models. An **ethical AI** system should be free of all such biases and designed based on clear ethical and moral guidelines.

A few terms have popped up to describe this focus on making AI more ethical, fair, and relatively harmless to the public. These terms include *ethical AI*, *responsible AI*, and *trustworthy AI*.

There may be subtle differences among these three labels, but the focus is largely the same. A separate branch called Ethical AI has gained some attention to address the concerns of the common public regarding intelligent systems and to ensure fairness in the AI models, which can impact human life. Some people argue that Ethical AI is broader in scope than responsible AI. As per Wikipedia, Ethical AI is a division of ethics of technology that focuses on AI. Ethical AI goes hand in hand with other branches like Robot ethics and Machine ethics to address several ethical issues with intelligent machines.

On the other hand, Responsible AI can be viewed as practical guidelines adopted by a company to ensure that their intelligent systems meet the ethical AI standards and requirements.

Principles of Responsible AI

Every company has its version of responsible AI guidelines and framework that matches its core values. But the key principles embedded in all these guidelines are as below:

1. Fairness
2. Reliability and safety
3. Privacy and security
4. Inclusiveness
5. Transparency
6. Accountability

<https://learn.microsoft.com/en-us/training/modules/responsible-ai-principles/>

Cisco's version of these principles can be found in the below document:

https://www.cisco.com/c/dam/en_us/about/doing_business/trust-center/docs/cisco-responsible-artificial-intelligence-principles.pdf

IBM's version, which is not very different from Microsoft's or Cisco's, can be found below:

<https://www.ibm.com/artificial-intelligence/ethics>

Two of the major consulting companies Deloitte and PwC have developed their own frameworks for fairness in AI. While Deloitte uses the term *trustworthy AI*, PwC uses the term Responsible AI for their framework. Figure 11.3 describes PwC's framework as articulated by Dr. Anand Rao, a PwC senior lead in AI. They propose nine ethical dimensions:

1. Interpretability (Explainability, transparency, provability)—the model decision-making should be explainable, and justification for a specific prediction/decision
2. Reliability, robustness, security—should be able to operate over a period of time and with varying data sets
3. Accountability—some entity should be clearly accountable for the moral and legal implications of its use

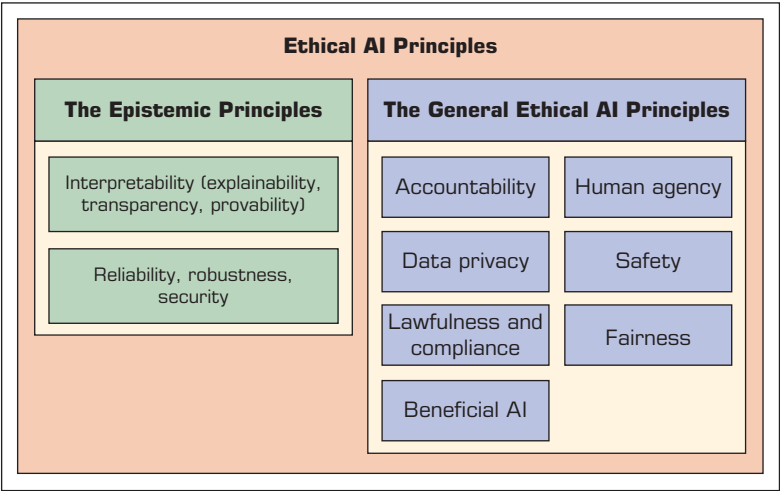


FIGURE 11.3 Ethical Principles of AI (adapted from a presentation by A. Rao, PwC).

4. Data privacy—individuals should be able to decide to let their data be used for modeling as well as and production systems, especially how the data can be shared/reused by other parties
5. Lawfulness and compliance—obviously, everyone is expected to operate within the legal framework of their environment
6. Beneficial AI—the systems should be aimed at creating common good
7. Human agency—humans must be involved in making the ethical decisions in proportion to the risk level
8. Safety—must be safe for all stakeholders' physical or mental health
9. Fairness—must be making similar decisions based on entity characteristics regardless of gender, race, religion, etc. That is, there should be no bias in the data sets being used for such models.

Why Ethical AI Should Matter to the Managers?

For a successful implementation of AI technologies, let us consider the case of Netflix. Upon its inception in 1997, Netflix was primarily a DVD sales and rental business. Later, Netflix scrapped the sales portion of its business model to focus solely on the rental side of it. In 1999, Netflix embraced the concept of monthly subscription for the first time, which turned out to be a game-changer. In 2007, Netflix turned its attention toward launching a streaming media service, which became a precursor to the company's stocks soaring. In efforts to keep expanding, Netflix entered the business of producing movies and TV shows on its own, which has also been a great success.

There is no doubt that Netflix showed business savviness by regularly exploring new ventures. But a large portion of the credit for its success should also go to how Netflix integrated innovative technology seamlessly into its business model. In 2016, Netflix's then VP of Product Innovation Carlos Gomez-Urbe (currently working as a senior director at Apple) and ex-Chief Product Officer Neil Hunt (current Chief Product Officer at Vibrant Planet) coauthored a technical paper (Gomez-Urbe and Hunt 2015) regarding Netflix's famous recommender system. They claimed in this paper that Netflix's recommender system was crucial in reducing user churn, which has an estimated impact of almost 1 billion USD saved each year.

Recommender systems are built using an AI algorithm that needs a large amount of user viewing habits data to predict the title of the next show these users would find interesting, which could be recommended next to these users. Like any other AI model, this system is not immune to biases, but this use case of Netflix demonstrates the true impact of a well-implemented AI.

It is rarely sufficient for large businesses to achieve success just with new product ideas without adapting the latest technology the way Netflix did. A report published in *Fortune Business Insights* states that the global AI market is expected to grow from \$387.45 billion in 2022 to \$1,394.30 billion by 2029. A leading professional services company, Accenture, researched 12 economies and forecasted that AI has the potential to increase profitability rates by an average of 38% by 2035. All such reports point to one fact: AI is a dominant player in the technology industry, and it is going to become an indispensable tool for even more organizations.

(Sources: <https://www.fortunebusinessinsights.com/industry-reports/artificial-intelligence-market-100114> (accessed January 2023); https://www.accenture.com/fr-fr/_acnmedia/36dc7f76eab444cab6a7f44017cc3997.pdf (accessed January 2023))

Hence, it is not a completely unforeseeable scenario that the managers at any level within these organizations would have to manage a project that either deals with AI implementation or requires the usage of AI tools.

Companies have a lot at stake while such projects are in the developmental phase. Consider the case of Amazon's scrapped recruiting tool. There must have been a large amount of capital and effort invested into building this tool, which ultimately failed to see the light of the day. Failure to adhere to proper ethical guidelines in all the phases of such projects could lead to wasted investments with an improper result. Also, any such news regarding unethical AI causes a stir among the public, endangering the company's reputation.

Also, when AI-powered products built without heeding ethical concerns are released to the public, they could easily attract tons of lawsuits. It could force the companies into paying billions of dollars as compensation, which only aggravates the irreparable damage already done to their reputation and goodwill. The parent company of chess-playing Deep Blue, IBM, had to endure one such lawsuit from the office of the Los Angeles city attorney in 2019 because of one of its subsidiaries. It was alleged in this lawsuit that IBM's weather channel mobile application, which had the permission to access the geolocation data of its users, sold it to third-party advertising companies behind the back of its users.

Hence, in addition to ensuring that the AI systems being built are fair and free from inherent biases, managers also have the responsibility to oversee that the guidelines for using data in these projects are in line with the appropriate data privacy and protection laws.

The O'Neil Claim of Potential Analytics' Dangers

Managers and data science professionals should be aware of the social and long-term effects of mathematical models and algorithms. Cathy O'Neil, a Harvard PhD in mathematics who worked in finance and the data science industry, expressed her experiences and observations in the popular book *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. We suggest you read the book, or at least visit the author's blog site at mathbabe.org/. The blog site highlights social issues related to analytics. A good summary/review of the book is available at knowledge.wharton.upenn.edu/article/rogue-algorithms-dark-side-big-data/.

In her book, O'Neil (2016) argued that models must satisfy three conditions. First, they must be transparent. That is, if the model is not understandable, its application can lead to unintended consequences.

Second, the model must have clear quantifiable objectives. For example, the celebrated application of analytics in the book and movie *Moneyball* includes a model that was aimed at increasing the number of financial wins. And the proposed input measures were well understandable. Rather than using the more commonly reported measure "run base in" (RBI), the analyst in *Moneyball* proposed and used on-base percentage and other measures (which were also easily calculated and understood by anyone with basic math skills). On the other hand, models built to assess the risk of mortgage-backed securities when no one fully understood the underlying assumptions of collateralized securities, but financial traders were trading, have been blamed for leading the financial crisis of 2008.

The third requirement is that the models must have a self-correcting mechanism and a process in place so that they are audited regularly, and new inputs and outputs are constantly being considered. This third issue is particularly critical in applying models in social settings. Otherwise, the models perpetuate the faulty assumptions inherent in the initial modeling stage. O'Neil discusses several situations where such is the case. For example, she describes the models built in the United States to identify underperforming teachers and reward better teachers. Some of these models utilized the test scores of the pupils to assess the teachers. O'Neil cited several examples where the models were used to fire "underperforming" teachers even though those teachers were loved by the students and parents. Similarly, models are used to optimize the scheduling of workers in many organizations. These schedules may have been developed to meet seasonal and daily demand variations, but the models do not take into account the deleterious impacts of such variability in schedules on the families of these usually lower-income workers. Other such examples include credit score assessment models that are based on historical profiles and thus may negatively impact underrepresented groups in training data. Without mechanisms to audit such models and their unintended effects, they can do more harm than good in the long term. So, model builders need to consider such concerns.

A comment: There is evidence that in some cases, O'Neil's claims are valid, and therefore model builders and implementers must pay attention to the issues. However, in general,

analytics are properly designed and bring considerable benefits to society. Furthermore, analytical models increase the competitiveness of companies and countries, creating many highly paid jobs. In many cases, companies have social responsibility policies that minimize biases and inequality. Finally, some people argue that algorithms and AI can be seen as great equalizers in bringing services that were traditionally reserved for a privileged few, to everyone. For example, AI could provide access to healthcare diagnostic services of experts more broadly around the world.

► SECTION 11.6 REVIEW QUESTIONS

1. What do the stories from Amazon, COMPAS, and BlenderBot have in common?
2. What are the nine principles of Responsible AI as proposed by PwC?

11.7 IMPACTS OF INTELLIGENT SYSTEMS ON ORGANIZATIONS

Intelligent systems are important components in the information and knowledge revolution. Unlike the slower revolutions of the past, such as the Industrial Revolution, this revolution is taking place very rapidly and affecting every facet of our work and lives. Inherent in this transformation is the impact on organizations, industries, and managers, some of which are described in this section.

Separating the impact of intelligent systems from that of other computerized systems is a difficult task, especially because of the trend toward integrating, or even embedding, intelligent systems with other computer-based information systems. Intelligent systems can have both micro- and macroimplications. Such systems can affect particular individuals and jobs as well as the work and structures of departments and units within an organization. They can also have significant long-term effects on total organizational structures, entire industries, communities, and society as a whole.

Explosive growth in analytics, AI, and cognitive computing is going to have a major impact on the future of organizations. The impact of computers and intelligent systems can be divided into three general categories: organizational, individual, and societal. In each of these, computers may have many possible impacts. We cannot possibly consider all of them in this book, so in the next paragraphs we cover topics we feel are most relevant to intelligent systems and organizations.

New Organizational Units and Their Management

One change in organizational structure is the possibility of creating an analytics department, a BI department, a data science department, and/or an AI department in which analytics plays a major role. Such special units (of any type) can be combined with or replace a quantitative analysis unit, or it can be a completely new entity. Some large corporations have separate decision support units or departments. For example, many major banks have such departments in their financial services divisions. Many companies have small data science or BI/data warehouse units. These types of departments are usually involved in training in addition to consulting and application development activities. Others have empowered a chief technology officer over BI, intelligent systems, and e-commerce applications. Companies such as Target and Walmart have major investments in such units, which are constantly analyzing their data to determine the efficiency of marketing and supply chain management by understanding their customer and supplier interactions. On the other hand, many companies are embedding analytics/data science specialties within functional areas such as marketing, finance, and operations. In general, this is one area where considerable job opportunities currently exist. Many organizations have appointed Chief Data Officers. Some are also considering adding other C-suite titles such as a Chief AI officer.

Growth of the BI and analytics has resulted in the formation of new units within IT companies as well. For example, a few years ago, IBM formed a new business unit focused on analytics. This group includes units in BI, optimization models, data mining, and business performance. More importantly, the group is focused not just on software but also significantly more on services/consulting.

Transforming Businesses and Increasing Competitive Advantage

One of the major impacts of intelligent systems is the transformation of businesses to digital ones. While such transformation has been going on with other information technologies for years, it has accelerated with intelligent technologies, mostly with AI.

In many cases, AI is only a supportive tool for humans. However, as AI has become more capable, machines have been able to perform more tasks by themselves or with people. The fact is that AI already is transforming some businesses. As seen in Chapter 2, AI already is changing all business functional areas, especially marketing and finance. The impact ranges from full automation of many tasks, including managerial ones, to an increase in human-machine collaboration. A comprehensive description of how AI is driving digital transformation is provided by Daugherty and Wilson (2018), who concluded that businesses that will miss the AI-driven transformation would be in a competitive disadvantage. Batra et al. (2018) point to a similar phenomenon and urge companies to use AI and utilize it for a wave of innovations.

USING INTELLIGENT SYSTEMS TO GAIN COMPETITIVE ADVANTAGE Use of intelligent technologies, and especially AI, is evidenced in many cases. For example, using robots, Amazon.com enabled the company to reduce cost and control online commerce. In general, by cutting costs, increasing customer experiences, improving quality, and speeding deliveries, companies will gain competitive advantage. Andronic (2017) points to competitive advantages of AI. The benefits include generating more demand (see Chapter 2), automating sales (Chapter 2), and identifying sales opportunities.

An important recent factor is the fact that new companies and blurring sector borders are influencing the competitive picture of many industries. For example, autonomous vehicles are impacting the competition in the automotive industry.

As we have seen throughout this book, a smart use of analytics offers top competitive advantage. The author provides advice on how organizations can get the full benefits from analytics. An example of how 1-800-Flowers.com is using analytics, AI, and other intelligent technologies to gain a competitive advantage is provided in Analytics in Action Case 11.1. You already saw the story about Netflix in Section 11.6.

ANALYTICS IN ACTION 11.1 How 1-800-Flowers.com Uses Intelligent Systems for Competitive Advantage

1-800-Flowers.com is a leading online retailer of flowers and gifts. The company moved from telephone to online ordering in the mid-1990s. Since then, it has grown to over \$1 billion in revenue and over 4,000 employees, despite fierce competition. In a world dominated by online giants such as Amazon.com and Walmart.com, and hundreds of other companies that sell online flowers and gifts, survival is not easy.

The company is using the following three key strategies:

- Enhancing the customer experience
- Driving demand more efficiently
- Building a workforce that supports the products and technology innovation (culture of innovation)

The company has been using intelligent technologies extensively to build a superb supply chain and to facilitate collaboration. Lately, it started to use intelligent systems to enhance its competitive strategies. Here are several technologies covered in this book that the company uses.

1. *Optimal customers experience.* Using SAS Marketing Automation and Data Management products, the retailer collects information regarding customers' needs and analyzes it. This information enables senders of flowers and gifts to find perfect gifts for any occasion. Senders want to make recipients happy, so appropriate recommendations are critical. The company uses advanced analytics and data mining from SAS to anticipate customers' needs. 1-800-Flowers.com marketers can then communicate with customers more effectively. Using the newest tools, company data scientists and marketing analysts mine data more efficiently. Today customer expectations are higher than ever because it is much easier for customers to compare vendors' offerings online. Analytics and AI enable the company to understand its customers' sentiments. Now the company is able to understand the emotional reasoning behavior for purchasing decisions and customer loyalty. This change results in product recommendations described later.
2. *Chatbots.* 1-800-Flowers.com has a bot on Facebook Messenger. As described in Chapter 12, such a bot can be useful as a source of information and as a vehicle for conversation. The company also offers chat on its Web site online, and chat using voice. In addition, mobile shoppers can use Google Assistant for voice ordering. The company also offers voice-enabled Alexa with its "one-shot intent" to expedite ordering.
3. *Customer service.* The company offers a portal and one-stop shopping similar to what Amazon.com offers, and self-service payment is available. The same capability is available when shopping with the company's bot on Facebook Messenger. Customers do not have to leave Facebook to complete an order.
4. *AI-based recommendation.* As you may recall from Chapter 12, e-commerce retailers excel by providing product recommendation (e.g., Amazon, Netflix). 1-800-Flowers.com is doing the same thing, offering recommendation and advice on gifts from their brand's Web sites (e.g., Harry and David). The recommendations are generated by IBM's Watson and are offered as a "cognitive concierge," making online shopping feel as having an in-store experience. This AI-based service is known as GWYN (Gifts When You Need) at 1-800-Flowers. Watson's natural language processing (NLP) enables easy shopper-machine conversations.
5. *Personalization.* SAS advanced analytics enables the company's marketing department to segment customers into groups with similar characteristics. Then the company can send promotions targeted to the profile of each segment. In addition to e-mails, special campaigns are arranged. Based on the feedback, the company can plan and revise marketing strategy. SAS also helps the company to analyze the "likes" and "dislikes" of the customers. All in all, the intelligent systems help the company and its customers to make informed decisions.

QUESTIONS FOR DISCUSSION

1. Why it is necessary to provide better customer experience today?
2. Why do data need sophisticated analytical tools?
3. Read the "Key benefit of SAS Marketing Automation." Which benefits do you think are used by 1-800-Flowers.com and why?
4. Relate IBM Watson to "personalization."
5. Relate "SAS Advanced Analytics" capabilities to their use in this case.
6. "SAS Enterprise Miner" is used to do data mining. Explain what is done and how.
7. SAS has a product called "Enterprise Guide" that 1-800-Flowers.com uses. Find how it is used based on the tools' capabilities.

Sources: Compiled from J. Keenan. (2018, February 13). "1-800-Flowers.com Using Technology to Win Customers' Hearts This Valentine's Day." *Total Retail*; S. Gaudin. (2016, October 26). "1-800-Flowers Wants to Transform Its Business with A.I." *Computer World*, SAS. (n.d.). "Customer Engagement Enhanced with Cloud-based Analytics and AI." *SAS Publication*, https://www.sas.com/en_us/customers/1800flowers.html (accessed January 2023).

Redesign of an Organization through the Use of Analytics

An emerging area of research and practice is employing data science technologies for studying organizational dynamics, personnel behavior, and redesigning the organization to better achieve its goals. Indeed, such analytics applications are known as *People Analytics*. For example, analytics are used by HR departments to identify ideal candidates from the pool that submits resumes to the organization or even from broader pools such as LinkedIn. Note that with AI and analytics, managers will be able to have a larger span of control due, for example, to the advice managers and employees can get from virtual assistants. The increased span of control could result in flatter organizational structures. Also, managers' job descriptions may have to change.

A more interesting and recent application area relates to understanding employee behavior by monitoring their movements within the organization and using that information to redesign the layout or teams to achieve better performance. A company called Humanyze (previously known as Sociometric Solutions) has badges that include a GPS and a sensor. When employees wear these badges, all of their movement is recorded. Humanyze has reportedly been able to assist companies in predicting which types of employees are likely to stay with the company or leave on the basis of their interactions with other employees. For example, it is stereotypical that those employees who stay in their own cubicles are less likely to progress up the corporate ladder than those who move about and interact with other employees extensively. Similar data collection and analysis have helped other companies determine the size of conference rooms needed or even the office layout to maximize efficiency. According to Humanyze's Web site, one company wanted to better understand characteristics of its leaders. By analyzing the data from these badges, the company was able to recognize that the successful leaders indeed have larger networks with which they interact, spend more time interacting with others, and are also physically active. The information gathered across team leaders was used to redesign the workspace and help improve other leaders' performance. Clearly, this may raise *privacy issues*, but within an organization, such studies may be acceptable. Humanyze's Web site has several other interesting case studies that offer examples of how Big Data technologies can be used to develop more efficient *team structures* and *organizational design*.

Intelligent Systems' Impact on Managers' Activities, Performance, and Job Satisfaction

Although many jobs may be substantially enriched by intelligent technologies, other jobs may become more routine and less satisfying. Some claim that computer-based information systems in general may reduce managerial discretion in decision-making and lead managers to be dissatisfied. However, studies of automated decision systems found that employees using such systems, especially those who are empowered by the systems, were more satisfied with their jobs. If using an AI system can do routine and mundane work, then it should free managers and knowledge workers to do more challenging tasks.

The most important task of managers is making decisions. Intelligent technologies can change the manner in which many decisions are made and can consequently change managers' job responsibilities. For example, some researchers found that a decision support system improved the performance of both existing and new managers as well as other employees. It helped managers gain more knowledge, experience, and expertise and consequently enhanced the quality of their decision-making. Many managers report that intelligent systems have finally given them time to get out of the office and into the field. They have also found that they can spend more time planning activities instead of putting out fires because they can be alerted to potential problems well in advance thanks to intelligent system technologies.

Another aspect of the managerial challenge lies in the ability of intelligent technologies to support the decision-making process in general and strategic planning and control decisions in particular. Intelligent systems could change the decision-making process and even decision-making styles. For example, information gathering for decision-making is completed much more quickly when algorithms are in use. Research indicates that most managers tend to work on a large number of problems simultaneously, moving from one to another as they wait for more information on their current problem. Intelligent technologies tend to reduce the time required to complete tasks in the decision-making process and eliminate some of the nonproductive waiting time by providing knowledge and information.

The following are some potential impacts of intelligent system on managers' jobs:

- Less expertise (experience) is required for making many decisions.
- Faster decision-making is possible because of the availability of information and the automation of some phases in the decision-making process.
- Less reliance on experts and analysts is required to provide support to top managers and executives. Today, they can decide by themselves with the help of intelligent systems.
- Power is being redistributed among managers. (The more information and analysis capability they possess, the more power they have.)
- Support for complex decisions makes solutions faster to develop and of better quality.
- Information needed for high-level decision-making is expedited or even self-generated.
- Automation of routine decisions or phases in the decision-making process (e.g., for frontline decision-making and using automated decision-making) may eliminate some managers.

In general, it has been found that the job of middle managers is the most likely job to be automated. Midlevel managers make mostly routine decisions, which can be fully automated. Managers at lower levels do not spend much time on decision-making. Instead, they supervise, train, and motivate nonmanagers. Some of their routine decisions, such as scheduling, can be automated; other decisions that involve cognitive aspects may not be automated. However, even if managers' decisional role is completely automated, many of their other activities could not be automated or could only be partially automated.

Impact on Decision-Making

Throughout the book, we have illustrated how intelligent technologies improve or automate decision-making. These technologies, of course, will impact managers' job. One aspect is the impact of intelligent technologies supported by the "cloud." It illustrates the flow of data from data sources and services via an information service to analytical services for different types of decision-making supported by analytics.

Uzialko (2017) describes how humans can use AI to predict and analyze the consequences of different potential solutions, streamlining the decision-making process. Also, by using machine learning and deep learning, more decisions can be automated.

Industrial Restructuring

A few authors have speculated on the impact of AI, analytics, and cognitive computing on the future of industry. A few interesting resources to consult are Autor (2016), and a special report by *The Economist* (Standage, 2016), and a book by Brynjolfsson and McAfee (2016). The report by *The Economist* is quite comprehensive and considers

many dimensions of the impact of the current developments on industry and society. The main arguments are that technology is now enabling more and more tasks that were done by humans using computers. Automating work, of course, has happened before, since the time of the Industrial Revolution. What makes the change this time around significantly more far reaching is that the technology is enabling many cognitive tasks to be done by machines. And the speed of change is so radical that the likely impact on organizations and society will be very significant and at times unpredictable. These authors do not agree in their predictions, of course. Let us focus first on the organizational impacts. Ransbotham (2016) argues that cognitive computing will convert many jobs done by humans to be done by computers, thus reducing costs for organizations. The quality of output may increase as well in cognitive work, which has been shown in several studies that compare a human's performance with a machine. Everyone is aware of IBM Watson having won in *Jeopardy!* or Google's system winning in the game of GO against human champions. But many other studies in specific domains such as speech recognition and medical image interpretation have also shown similar superiority of automated systems when the task is highly specialized yet routine or repetitive. Also, because machines tend to be available at all hours and at all locations, an organization's reach may increase, resulting in easier scaling and thus greater *competition* among organizations. These organizational impacts mean that yesterday's top organizations may not remain at the top forever because cognitive computing and automation can challenge established players. This is the case in the automotive industry. Although traditional car companies are trying quickly to catch up, Google, Tesla, and other technology companies are disrupting industry structure by challenging the leaders of the automotive age. Analytics and AI are empowering many of these changes.

► SECTION 11.7 REVIEW QUESTIONS

1. List the impacts of intelligent systems on managerial tasks.
2. Describe new organizational units that are created because of intelligent systems.
3. Identify examples of analytics and AI applications used to redesign workspace or team behavior.
4. How is cognitive computing affecting industry structure and competition?
5. Describe the impacts of intelligent systems on competition.
6. Discuss the impact of intelligent systems on decision-making.

11.8 IMPACTS ON JOBS AND WORK

One of the most discussed and debated topics in considering the impacts of intelligent systems is on jobs and work. There is a general agreement that:

- Intelligent systems will create many new jobs as automation always has.
- There will be a need to retrain many people.
- The nature of work will be changed.

The discussions, debates, and disagreements relating to the issues of when, how much, and how to deal with these phenomena occupy many researchers and are the topics of this section.

An Overview

According to Ransbotham (2016), financial advising is typically considered a knowledge-intensive task. As robot (robo) advisors provide personalized support for individuals, the costs of such services go down. This leads more people to demand such

services, eventually freeing more humans to address advanced financial issues. Robo advisors may also cause some people to lose their jobs.

Some authors argue that the automation segment, which is related to cognitive computing and AI, will accelerate what is called *polarization* of the labor market in the future. This entails significant job growth in the top *and* bottom tiers of the labor market but losses in the middle. Jobs requiring low but specialized skills, such as personal care, are continuing to grow. Similarly, jobs that require very high skill, such as graphics design work, and so on, are also growing. But jobs that require “middle skills” such as specialized knowledge that is applied over and over with some adaptation, are at the greatest risk of disappearing. Sometimes technology disintermediates itself! For example, IBM Watson Analytics now includes querying capabilities to begin asking questions that an intelligent system professional previously asked and, obviously, providing answers. Other analytics-as-a service offerings with similar services may result in a need for fewer people to be proficient at using analytics software.

A report by *The Economist* notes that even if AI does not replace workers directly, it will certainly require employees to acquire new skills to keep their jobs. Market disruption is always uncomfortable. The next few years will provide excellent opportunities for intelligent technology professionals to shape the future.

Are Intelligent Systems Going to Take Jobs—My Job?

Tesla’s Elon Musk envisions AI-based autonomous driving trucks all over the world within 10 years. There will be convoys of such trucks, each of which will follow a lead truck. Trucks will be electrical, economical, and pollution free. In addition, there will be fewer accidents—sounds great! But what about thousands of drivers who will lose their jobs? What about many thousands of employees in truck stops who will lose their jobs as well? The same scenario could happen in many other industries. Amazon has opened its first Go, a cashierless physical store. They plan 3,000 more in a few years. The post office in some countries already distributes mail using autonomous vehicles. In short, there is a chance for massive unemployment.

Example: Pilots at FedEx

FedEx has a fleet of close to 1,000 airplanes flying globally. According to Frank Tode, editor and publisher of *The Robot Report*, FedEx had hoped that around 2020 the company would have one global pilot center with three or four pilots who will operate the entire fleet. Clearly, that has not happened yet. Foxconn, an iPhone manufacturer in Taiwan, had planned to replace almost all of its employees (60,000) in Taiwan with robots (Botton, 2016). The company already produced 10,000 robots for this purpose. However, they still employ thousands of workers.

INTELLIGENT SYSTEMS MAY CREATE MASSIVE JOB LOSSES The debate regarding technology taking jobs has been going on since the beginning of the industrial revolution. The issue regarding intelligent systems is strongly debated now due to the following:

- They are moving very fast.
- They may take a large variety of jobs, including many office-focused, administrative, and nonphysical jobs.
- Their comparative advantage over manual labor is very large and growing rapidly.
- They are already taking some professional jobs from financial advisors, paralegals, and medical specialists.
- The capabilities of AI are growing rapidly.
- In Russia, robots are already teaching mathematics in schools (some do a better job than humans). Just think about what could happen to the teaching profession.

AI Puts Many Jobs at Risk

For the potential impact of AI on jobs, see Dormehl (2017), who explores the possibility of creative intelligent machines. For example, McKinsey's study estimates that AI is poised to take over 30% of all bank jobs in the near future. The study also predicts that robots will take 800 million jobs worldwide by 2030.

To research the potential danger of job loss, McKinsey & Company divided jobs into 2,000 distinct work activities, such as greeting customers and answering questions about products, which retail salespeople do. Its researchers (see Chui et al., 2015) found that 45% of all 2,000 activities could be economically and physically automated. The activities include physical, cognitive, and social types.

While autonomous vehicles are not taking jobs, yet they will take jobs from taxi drivers, Uber, and similar companies' drivers. Also, bus drivers may lose their jobs. Other jobs that have already been replaced by intelligent systems are listed in Analytics in Action Case 11.2.

ANALYTICS IN ACTION 11.2 Administrative/Office Jobs That Robots Have Already Taken

While it may be sometime before FedEx will have pilotless airplanes and schools will have no human teachers, some jobs, according to Sherman (2015), have already been taken by robots. They include:

- **Online marketers.** Using NLP, companies are automatically developing marketing ads and e-mails that influence people to buy (robo marketers). These are based on a dialog with potential buyers and on an automatic database search of historical cases. "Who needs an online marketer that may have inferior, biased, or incomplete knowledge?"
- **Financial analysts and advisors.** As were described in Chapter 2, robo advisors are all over the scene. Equipped with the ability to deal with Big Data in real time and conduct predictive analysis in seconds, these programs are liked by investors who pay about one-tenth of what human advisors charge. Furthermore, robo advisors can personalize recommendations.
- **Anesthesiologists, diagnosticians, and surgeons.** The medical field seems to be immune from AI. This is not the case. Expert systems for diagnosis have been in place for about 40 years. The FDA has already approved the J&J Sedasys system for delivery of low-level anesthesia in surgeries, such as colonoscopies. Finally, surgeons already use automated machines in some invasive procedures.
- **Financial and sports reporters.** These jobs involve gathering information, interviewing people, answering questions, analyzing the material, and writing reports. The Associated Press (AP) has experimented with AI machines since 2014. Results so far are virtually error- and bias-free (and no fake news!). Tools such as ChatGPT threaten more of such jobs.

Palmer (2017) reported an additional five jobs in danger, including middle management, commodity salespeople, report writers, accountants and bookkeepers, and some types of doctors.

McFarland (2017b) lists as high-risk jobs cashiers, toll booth operators, fast-food employees, and drivers. Low-risk jobs include nurses, doctors, dentists, youth sport coaches, and social workers.

QUESTIONS FOR DISCUSSION

1. Watch or listen to the 4:22-minute video about an interview with Palmer at [linkedin.com/pulse/5-jobs-robots-take-first-shelly-palmer/](https://www.linkedin.com/pulse/5-jobs-robots-take-first-shelly-palmer/). Discuss some of the assertions made regarding doctors.
2. Discuss the possibility of your checkup by a robot diagnostician. How would you feel?

3. With the bombardment of fake news and their biased creators, it may be wise to replace all of them by intelligent machines. Discuss such a possibility.
4. You are a defendant in a crime you did not commit. Would you prefer a traditional lawyer or one equipped with an AI e-discovery machine? Why?

Sources: Compiled from E. Sherman. (2015, February 25). "5 White-Collar Jobs Robots Already Have Taken." *Fortune.com*. fortune.com/2015/02/25/5-jobs-that-robots-already-are-taking (accessed January 2023), S. Palmer. (2017, February 26). "The 5 Jobs Robots Will Take First." Shelly Palmer.

Let us look at some other studies. A 2016 study done in the United Kingdom predicted that robots would take 50% of all jobs by 2026. Egan (2015) reports that robots already threaten the following jobs: marketers, toll booth operators and cashiers, customer service, financial brokers, journalists, lawyers, and phone workers. Note that automation may affect portions of almost all jobs to a greater or lesser degree. Experts estimate that about 80% of IT jobs may be eliminated by AI.

Given all this, you may wonder whether your job is at risk.

Which Jobs Are Most in Danger? Which Ones Are Safe?

If want to know about your job, it obviously depends on the type of job you are holding. Oxford University in the United Kingdom looked at 700 jobs and ranked them from zero (no risk of automation) to 1 (very high risk of automation). Straus (2014) provided a list of the top 100 most at-risk jobs (all above 0.95) and the 100 jobs with the lowest risk (with 0.02 or less). The top 10 "safe" and the 10 at risk are listed in Table 11.1.

A 2017 study conducted by the Bank of England found that almost half of the UK jobs (15 million out of 33.7 million) are at risk of loss within 20 years. Creative robots are the greatest threat because they can learn and increase their capabilities. While in the past, automation may not have decreased the total number of jobs, this time the situation may be different.

A side effect of this situation may be that workers will have less income while the owners of robots will have a larger income. (This is why Bill Gates suggested taxing the robots and their owners.)

Some More Job Losses Observations

- Kelly (2018) predicts that robots could eliminate many Las Vegas jobs. And indeed, in many casinos worldwide, you can play several traditional games on machines.
- People with doctoral degrees have a 13% chance of being replaced by robots and AI versus 74% for those with only a high school education (Kelly, 2018). Women may lose more jobs to automation than men (Krauth, 2018).

The Web site <https://willrobotstakeyourjob.com/> (accessed March 2023) can be used to get the latest information on which jobs are considered at risk of being automated.

Intelligent Systems May Actually Add Jobs

Despite the fear, uncertainty, and panic related to job losses, many reports contradict this. Here are some examples: de Vos (2018) reported that AI will create 2.3 million jobs while eliminating 1.8 million. Also, one needs to consider the great benefits of AI and the fact that human and machine intelligence will complement each other in many jobs. Also, AI will increase international trade, adding more jobs. de Vos also cites studies that show the

TABLE 11.1 Ten Top Safe and At-Risk Occupations

Probability of Job Loss
Low-Risk Jobs
0.0036 First-Line supervisors of firefighting and prevention workers
0.0036 Oral and maxillofacial surgeons
0.0035 Healthcare social workers
0.0035 Orthotists and prosthetists
0.0033 Audiologists
0.0031 Mental health and substance abuse social workers
0.0030 Emergency management directors
0.0030 First-line supervisors of mechanics, installers, and repairers
0.0028 Recreational therapists
High-Risk Jobs
0.99 Telemarketers
0.99 Title examiners, abstractors, and searchers
0.99 Sewers, hand
0.99 Mathematical technicians
0.99 Insurance underwriters
0.99 Watch repairer
0.99 Cargo and freight agents
0.99 Tax preparers
0.99 Photographic process workers and processing machine operators
0.99 New account clerks

Source: Based on Straus (2014) Straus, R.R. "Will You Be Replaced by a Robot? We Reveal the 100 Occupations Judged Most and Least at Risk of Automation." *ThisisMoney.com*, May 31, 2014. thisismoney.co.uk/money/news/article-2642880/Table-700-jobs-reveals-professions-likely-replaced-robots.html (accessed January 2023)

creation of jobs due to equipment maintenance and service that cannot be automated. The following are predictions on both sides of the issue:

- A PricewaterhouseCoopers (PwC) study forecast that robots will bolster UK economic growth. So, even though robots could destroy about 7 million jobs in the United Kingdom, they will create at least 7 million new jobs and probably more over 20 years (Burden, 2018).
- IBM's new deep learning service may help save IT jobs.
- There is a shortage of millions of skilled workers (e.g., about 50,000 truck drivers in the United States), so automation will reduce millions of unfilled positions.
- On the other hand, many new categories of human jobs have been created by AI.
- Some believe that there will be a total of increase in jobs due to AI-induced innovations.
- Violino (2018) contradicts those who claim that there is a huge fear among employees regarding job loss, saying that most workers see robots as an aid to their jobs. See also Leggatt (2017).

Jobs and the Nature of Work Will Change

While you may not lose your job, intelligent applications may change it. One aspect of this change is that low-skill jobs will be taken by machines, but high-skill jobs may not. Therefore, jobs may be redesigned either to be low skilled in order to be automated, or

to be high skilled so that they will be executed exclusively by humans. In addition, there will be many jobs where people and machines will work together as a team.

Changes in jobs and business processes will impact training, innovation, wages, and the nature of work itself. Manyika (2017) of McKinsey & Company analyzed the shifts that can be fundamental, and arrived at the following conclusions:

- Many activities done by humans will have the potential to be automated.
- Productivity growth from robotics, AI, and machine learning will be tripled compared to pre-2015.
- AI will create many new jobs paying high salaries.
- Since more than half the world is still offline, the changes will not be too rapid.

Example: Skills of Data Scientists Will Change

There will be a need to retrain or train scientists so they can deal with intelligent technologies and the changes in data science and in solving related real-world problems. Thus, proper education must evolve. The job requirements of data scientists are already changing. The scientists will need to know how to apply machine learning and intelligent technologies to build IoT and other useful systems. New algorithms improve operations and security, and data platforms are changing to fit new jobs. We address this topic further in the last section of this chapter.

Snyder (2017) found that 85% of executives knew that intelligent technologies would impact their workforce within five years, and 79% expected the current skill sets to be restructured. They also expected 79% productivity improvement. Employees fear that intelligent systems will take over some of their activities, but they hope that intelligent systems will also help with their work.

TIPS FOR SUCCESS A McKinsey study of 3,000 executives (Bughin, McCarthy, and Chui, 2017) reports the following success tips for implementing AI provided by the executives:

- Digital capabilities need to come before AI.
- Machine learning is powerful, but it is not the solution to all problems.
- Do not put technology teams solely in charge of intelligent technologies.
- Adding a business partner may help with AI-based projects.
- Prioritize a portfolio approach to AI initiatives.
- The biggest challenges will be people and business processes.
- Not every business is using intelligent systems, but almost all those that use them increase income and profit.
- Top leadership support is necessary for a transformation to AI.

DEALING WITH THE CHANGES IN JOBS AND THE NATURE OF WORK Manyika (2017) made the following suggestions for policymakers:

1. Use learning and education to facilitate the change.
2. Involve the private sector in enhancing training and retraining.
3. Have governments provide incentives to the private sector so employees can invest in improved human capital.
4. Encourage private and public sectors to create appropriate digital infrastructure.
5. Innovative income and wage schemes need to be developed.

6. Carefully plan the transition to the new work. Deal properly with displaced employees.
7. Properly handle new technology-enabled technologies.
8. Focus on new job creation, particularly digital jobs.
9. Properly capture the productivity increase opportunities.

Baird et al. (2017) of McKinsey & Company provide a video interview with industry experts discussing how to deal with the changing nature of work. Chui et al. (2015) researched the impact of automation on redefining jobs and business processes, including the impact on wages, and the future of creativity. Finally, West (2018) provides a comprehensive study on the future of work as it is influenced by robotics and AI-driven automation.

Conclusion: Let Us Be Optimistic!

Assuming that the disasters will not occur, then, as in the past, concerns about technology replacing many human jobs and reducing wages are hopefully exaggerated. Instead, intelligent technologies will clearly contribute to shorter work time for humans. Today, most people work long hours just for survival.

► SECTION 11.8 REVIEW QUESTIONS

1. Summarize the arguments of why intelligent systems will take away many jobs.
2. Discuss why job losses may not be catastrophic.
3. How safe is your job? Be specific.
4. How may intelligent systems change jobs?
5. In what ways may work be changed?
6. Discuss some measures to deal with the changes brought by intelligent systems.
7. One of the areas of potential job loss is due to autonomous vehicles. Discuss the logic of this.

11.9 POTENTIAL DANGERS OF AI

During the period 2016–2018, we witnessed a heated debate regarding the future of AI and particularly robots. Dickson (2017) called the optimistic approach *Utopia* and the pessimistic one *Dystopia*. The debate began with the industrial revolution regarding automation, and it has accelerated because of the rapid technological innovations of AI. We have already presented one aspect of this debate, the impact on jobs. In the center of the debate is the prediction of when AI's capabilities to reason and make decisions will become similar or even superior to that of people. Furthermore, will such a development be beneficial or dangerous to society?

Position of AI Dystopia

The camp that supports this prediction includes well-known tech executives. Here are three of them:

- **Elon Musk:** “We need to be super careful with AI. Potentially more dangerous than nukes.” (See or listen to the 10-minute video at [youtube.com/watch?v=SYqCbJ0AqR4](https://www.youtube.com/watch?v=SYqCbJ0AqR4).) Musk predicts that World War III will start because of AI. “Robots will kill us all, one day,” he said in his several presentations.

- **Bill Gates:** “I am in the camp that is concerned about super intelligence. Musk and some others are on this, and I don’t understand why some people are not concerned.” (Comments made on TV and interviews, several times.) He also suggested taxing the manufacturers and users of robots and other AI machines.
- **Stephen Hawking:** The late scientist stated, “The development of full artificial intelligence could spell the end of the human race.”

Many people are afraid of AI because they believe that computers will become smarter than we are. See or listen to Bostrom’s video of his famous TED presentation at [youtube.com/watch?v=MnT1xgZgkpk](https://www.youtube.com/watch?v=MnT1xgZgkpk). For how robots can learn motor skills through trial and error, see the video at [youtube.com/watch?v=JeVppkoloXs/](https://www.youtube.com/watch?v=JeVppkoloXs/). For more, see Pham (2018).

The AI Utopia’s Position

A good place to begin for information on this position is to watch or listen to the 26-minute documentary video on the future of AI at [youtube.com/watch?v=UzT3Tkwx17A](https://www.youtube.com/watch?v=UzT3Tkwx17A). This video concentrates on the contribution of AI to the quality of life. One example is crime fighting in Santa Cruz, California, where AI was able to predict where and when crimes will occur. Following the predictions, the police department has been planning its work strategies. The result is a 20% reduction in crime.

A second example is the prediction of the probability that a certain song will be a hit. The prediction helps both artists and managers to plan their activities. Great success has been made. In the future, AI is predicted to compose top songs.

Finally, there is a story about dating. The capabilities of AI enabled a scientist to find a perfect match in a population of 30,000 potential candidates.

A basic argument of the Utopianists expressed in interviews, TV lectures, and more is that AI will support humans and enable innovations. AI also will partner with humans. The Utopians believe that as AI expands, humans will become more productive and will have time to do more innovative tasks. At the same time, more tasks will be fully automated. Prices of products and services will drop, and the quality of life will increase.

A leading proponent of AI benefits is Mark Zuckerberg of Meta (Facebook). He was in a heated debate with Elon Musk (CEO of Tesla Corp), the unofficial leader of the Dystopia camp of believers. Zuckerberg criticized those that believe that AI will cause “doomsday scenarios” (see the next section). Musk claimed that Zuckerberg has a “limited understanding” of AI, and Zuckerberg answered by referring to his paper on AI that won an award at the “top computer vision conference.”

SOME ISSUES RELATED TO THE UTOPIA Several issues are related to the Utopianists’ position. Here are three examples:

1. AI will be so great that people will have a problem of what to do with their free time. If you have not yet seen Disney’s *Wall-E* movie, go and see it. It shows how humans are served by robots. Dennis Hassabis, a strong proponent of Utopia (from Deep Mind, an AI company), believes that AI will one day help people have a better life by understanding what makes humans unique, what the mysteries of the mind are, and how to enjoy creativity.
2. The road to AI Utopia could be rocky, for example, there will be impacts on jobs and work. It will take time to stabilize and adjust work and life of living with robots, chatbots, and other AI applications.
3. One day we will not drive anymore and there may not be human financial advisors; everything will be different, and the changes may be rapid and turbulent, and we may even face disasters, as projected by the Dystopia camp.

THE FRIENDLY AI Eliezer Yudkowsky, a cofounder of the Machine Intelligence Research Institute, developed the idea of *friendly AI*, according to which AI machines should be designed so that they will benefit humans rather than harm them (i.e., use a system of checks and balances in designing the AI capabilities).

CONCLUSION It is difficult to know what will happen in the future. But some actions are already being taken to prevent a disaster. For example, several major companies have declared that they will not produce or support killer robots.

► SECTION 11.9 REVIEW QUESTIONS

1. Summarize the major arguments of the Utopia camp.
2. Summarize the major arguments of the Dystopia camp.
3. What is the friendly AI?
4. What is Open AI? Relate it to the dystopia vision.
5. What are the potential risks in using modeling and analytics?

11.10 CITIZEN SCIENCE AND CITIZEN DATA SCIENTISTS

Citizen Science

A concept that has recently gained some popularity recently is the citizen data scientist. This idea is rooted in the concept of citizen science. Citizen science refers to members of general public voluntarily help conduct scientific research. This participation in scientific research may be at any stage of the scientific process—design experiments, collect data, help analyze data, develop reports, etc. For example, the US National Park Service invites citizen scientists to participate in its scientific research. According to their Web site (<https://www.nps.gov/subjects/citizenscience/citizen-science.htm>), “In national parks, most citizen scientists collect data with tools provided by project directors. These data help professional scientists and resource managers answer scientific questions and solve important problems. And the activity helps participants build meaningful connections to science.” A related similar project has been run by Audubon Society to get help in counting winter birds in various locations (<https://www.audubon.org/conservation/science/christmas-bird-count>). The data collected through such crowdsourced efforts have been used to better understand the impact of climate change, etc. on bird populations.

The US government maintains a Web site, <https://www.citizenscience.gov/>, to help citizens learn about many citizen science projects being undertaken by various branches of the US government. In early 2023, the list of such projects on the Web site is extensive, ranging from monitoring water quality in Kenai River in Alaska to spotting clouds in Martian atmosphere.

Citizen Data Scientist

The term Citizen Data Scientist is being used in two ways. First, as an extension of the citizen science introduced above, any projects where the citizens are involved in data science to help either collect data or analyze the data using data science methods can be called **citizen data scientists**. For example, many volunteers worked on COVID data to provide various metrics, visualizations, and predictions during the early days of the pandemic.

A more commonly used meaning of the term *citizen data scientist* was introduced by Gartner and adopted by the data science software industry in particular. Essentially, a citizen data scientist is a business user in an organization who has not been fully

trained in mathematics and statistics concepts of data science but has learned to use the software tools to build models for their own business use. Thus, the term refers to a domain expert in any specific domain who uses data science for a part of their work, in contrast to a data scientists who are hired to build such models and applications for their clients who would indeed be business (for profit or nonprofit) users. The citizen data scientists have learned to use the tools that make model exploration easy through a point-and-click interface. Indeed, some of the tools introduced in this book (Chapter 9) fall in this category of easy-to-use software that can allow a casual user of data science techniques to at least experiment with the models to see if there is any value in employing such techniques.

The idea of citizen data scientist was promoted by the data science industry to help alleviate the shortage of data science professionals. Now that we have many data science programs around the world, it may be less of an issue. Indeed, some experts wonder if the citizen data scientists run the risk of misapplication of a model due to their lack of specialized training. Perhaps an early prototype development by a citizen data scientist followed by involvement of a professional data scientists before a model is employed provides a good compromise between scalability and quality.

In any case, the first use of the term is still very relevant, needing more contributions by citizens in broader societal applications to help label data that may be used for machine learning models, provide visuals, run models, etc.

Involvement of citizens in scientific research could also impact acceptance of recommendation coming from such research. In many cases, future behavior changes require social acceptance and adaptation of changes suggested by researchers and policymakers. For example, many countries around the world face water shortage, particularly groundwater shortage. Any proposed reduction in groundwater use, for example, will be ignored by the populace unless the users better understand the problem. This is where citizen science in general and citizen data science in particular comes in. By involving the citizens in data collection, it makes them more aware of the problem. By employing analytics/data science to share information about what they are doing, what others are doing, and how the overall situation is evolving, the citizens have the potential to become a part of the decision-making process and thus accept the suggested interventions more easily. Engagement with citizen science projects results in participants being more interested in the natural world and their communities, by learning more about it and being more exposed to it, and therefore contributing more actively to its protection (Santori et al., 2021). For example, a project initiated by Dr. Basant Maheshwari of the University of Western Sydney in cooperation with several teams in India has initiated **MARVI** project, an international project (Australia-India), that has developed a participatory model to train villagers as groundwater citizen scientists and educate rural Indian communities about their groundwater resources, especially for sustainable water management (Maheshwari et al, 2022). The university is also leading citizen science related hydrological and wetland monitoring in Australia. The Hawkesbury River Keepers Streamwatch program is a citizen science water monitoring program that enables community groups to monitor the quality and health of local waterways. The 1 million Turtles Community Conservation program, which is a partnership between 10 universities, non-government conservation organizations, and government agencies (as well as more than 30 community groups), includes an Australia-wide wetland wildlife monitoring tool that has been used by more than 5,000 Citizen Scientists (see TurtleSAT.org.au). These data feed into Australia's premier conservation database, Atlas of Living Australia. Overall, the role of a citizen scientist and making the general public more aware of data-based decision-making are exciting frontiers for scaling up the applications and then enabling significant behavioral changes to achieve goals for societal good.

Concluding Remarks

New topics are constantly coming up in the field of analytics/data science/AI. Some are likely to be just fads whereas the others will make transformative difference in life. For example, as we complete this manuscript, the topic of Generative AI has gotten a lot of attention. Companies such as OpenAI are causing a lot of excitement about the potential of AI being able to create new content. For example, Open AI's ChatGPT is reputed to be able to answer questions about many topics with a clarity and efficiency expected of human experts. An earlier product released by Open AI called DALL-E can generate images that look very authentic similarly on the basis of a short description provided by the user. OpenAI's CoPilot product can be used for generating code in conjunction with GitHub. This ability to generate new text, images, sounds, etc. is termed Generative or Regenerative AI. Now, Microsoft and Google are in a race to integrate such tools in their search engines as well as other products. The field is exploding with over 450 firms currently calling themselves to be startups in this space. Similarly, conversational AI is a subset of these technologies that understand and produce conversations. These developments offer exciting opportunities and new frontiers for students and professionals in data science/AI, while also presenting new topics of issues to watch in the future.

This book has just scratched the surface of the analytics/data science/artificial intelligence technologies' applications and managerial issues. Since these topics are constantly evolving, we hope that our readers have not only learned some concepts, but even more importantly, the ways, means, and outlets to consult for other future topics.

Chapter Highlights

- Intelligent systems can affect organizations in many ways as standalone systems or integrated among themselves or with other computer-based information systems.
- The impact of analytics on individuals varies—it can be positive, neutral, or negative.
- Serious legal issues may develop with the introduction of intelligent systems; liability and privacy are the dominant problem areas.
- Many positive social implications can be expected from intelligent systems. These range from providing opportunities to people to lead the fight against terrorism. Quality of life, both at work and at home, is likely to improve as a result of the use of these technologies. Of course, there are potentially negative issues to be concerned about.
- Growth of intelligent systems is going to lead to major changes in industry structure and future employment.
- A major battle is brewing about who owns the user data that are being generated from the use of smartphones, cars, and so on.
- In deploying intelligent systems, it is necessary to consider legal, privacy, and ethical issues.
- Placing robots as coworkers in the work force raises legal and ethical issues.
- Intelligent technologies may impact business processes, organizational structure, and management practices.
- It may be necessary to create independent organizational units that deploy and manage intelligent systems.
- Intelligent systems may provide a considerable competitive advantage to their users.
- Intelligent systems may create massive unemployment mainly in routine and mid-management jobs.
- Eventually, intelligent system may cause unemployment even in skilled jobs. So, retraining may be needed.
- Intelligent systems may result in restructuring many jobs notably through human-machine collaboration.
- Intelligent systems will create many new jobs that require specialized training.
- The use of intelligent systems automation may result in a shorter workweek and a need to help provide some income to those people who will lose their jobs.
- Rapid advances in these technologies are enabling citizens to participate in codeveloping AI technologies.

Key Terms

implementing intelligent systems
ethical AI

privacy concerns
citizen data scientists

Questions for Discussion

1. Some say that analytics in general dehumanize managerial activities, and others say they do not. Discuss arguments for both points of view.
2. Diagnosing infections and prescribing pharmaceuticals are the weak points of many practicing physicians. It seems, therefore, that society would be better served if analytics-based diagnostic systems were used by more physicians. Answer the following questions:
 - a. Why do you think such systems are used minimally by physicians?
 - b. Assume that you are a hospital administrator whose physicians are salaried and report to you. What would you do to persuade them to use an intelligent system?
 - c. If the potential benefits to society are so great, can society do something that will increase doctors' use of such intelligent systems?
3. What are some of the major privacy concerns in employing intelligent systems on mobile data?
4. Identify some cases of violations of user privacy from current literature and their impact on data science as a profession.
5. Some fear that robots and AI will kill all of us. Others disagree. Debate the issue.
6. Some claim that AI is overhyped. Debate the issue. Place a question on Quora and analyze five responses.
7. Some claim that AI may become a human rights issue (search for Safiya Noble). Discuss and debate.
8. Discuss the potential impact of the GDPR on privacy, security, and discrimination.
9. Discuss ethics and fairness in machine learning.
10. Should robots be taxed like workers? Read Morris (2017) and write about the pros and cons of the issue.

Exercises

1. Identify ethical issues related to managerial decision-making. Search the Internet, join discussion groups/blogs, and read articles from the Internet. Prepare a report on your findings.
2. Search the Internet to find examples of how intelligent systems can facilitate activities such as empowerment, mass customization, and teamwork.
3. Investigate the American Bar Association's Technology Resource Center (americanbar.org/groups/departments_offices/legal_technology_resources.html) and nolo.com. What are the major legal and societal concerns regarding intelligent systems? How are they being dealt with?
4. Explore several sites related to healthcare (e.g., WebMD.com, who.int). Find issues related to AI and privacy. Write a report on how these sites suggest improving privacy.
5. Go to Humanyze.com. Review various case studies and summarize one interesting application of sensors in understanding social exchanges in organizations.
6. Research the issue of voice assistants and privacy protection.
7. Is granting advanced robots rights a good or bad idea?
8. Face and voice recognition applications are mushrooming. Research the state of their regulation in a country of your choice. Use the United States if your country is not regulating.
9. Research the ethical issues of self-driving cars.
10. Is your organization ready for AI? Research this issue and find all major activities that it includes.
11. Research the role of IoT as a tool for providing connectivity between sensors and analytics. Write a report.
12. Some people say that robots and chatbots may increase insurance risk and fees. Research this and write a report.
13. Listen to the video at youtube.com/watch?v=ww-uovuCfDU/ and comment about the robot's potential impacts.
14. Research the issue stated in quotation marks: "When will robots' rebel?" and "Will AI take control of the plant?"
15. Research the areas in which machines can replace humans and where they cannot (yet). Write a report.
16. Listen to the 3:38-minute video at youtube.com/watch?v=78-1Mlkxyql/. Relate it to Musk's predictions about robots reigning in this world.
17. Read the SAS report on AI ethics at sas.com/en_us/insights/articles/analytics/artificial-intelligence-ethics.html. Comment on each of the three proposed steps. Also comment on the human-machine collaboration in problem solving.

References

- Andronic, S. (2017, September 18). "5 Ways to Use Artificial Intelligence as a Competitive Advantage." Moonioa.com.
- Autor, D. H. (2016, August 15). "The Shifts—Great and Small—in Workplace Automation." *MIT Sloan Review*. sloanreview.mit.edu/article/the-shifts-great-and-small-in-workplace-automation/ (accessed January 2023).
- Baird, Z. et al. (2017, August). "The Evolution of Employment and Skills in the Age of AI." McKinsey Global Institute. Video at <https://www.mckinsey.com/featured-insights/future-of-work/the-evolution-of-employment-and-skills-in-the-age-of-ai#/> (accessed March 2023)
- Baroudy, K., et al. (2018, March). "Unlocking Value from IoT Connectivity: Six Considerations for Choosing a Provider." *McKinsey & Company*.
- Batra, G., A. Queirolo, & N. Santhanam. (2018, January). "Artificial Intelligence: The Time to Act Is Now." *McKinsey & Company*.
- Bossmann, J. (2016). "Top 9 Ethical Issues in Artificial Intelligence." *World Economic Forum*.
- Botton, J. (2016, May 28). "Apple Supplier Foxconn Replaces 60,000 Humans with Robots in China." *Market Watch*.
- Brynjolfsson, E., & A. McAfee. (2016). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. Boston, MA: W.W. Norton.
- Bughin, J., M. Chui, and J. Manyika. "An Executive's Guide to the Internet of Things." *McKinsey Quarterly*, August 2015.
- Bughin, J., B. McCarthy, & M. Chui. (2017, August 28). "A Survey of 3,000 Executives Reveals How Businesses Succeed with AI." *Harvard Business Review*.
- Bui, T. "To Succeed in IoT, Hire a Chief Data Officer." *Tech Crunch*, July 11, 2016.
- Burden, E. (2018, July 16). "Robots Will Bolster U.K. Growth and Create New Jobs, PwC says." *Bloomberg News*.
- Catlift, C. (2017, August 15). "Three Ways Your Business Can Leverage Artificial Intelligence." *The Globe and Mail*.
- Chui, M., J. Manyika, & M. Miremadi. (2015, November). "Four Fundamentals of Workplace Automation." *McKinsey Quarterly*. At <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/four-fundamentals-of-workplace-automation> (accessed March 2023)
- Chui, M., J. Manyika, & M. Miremadi. (2016, July). "Where Machines Could Replace Humans—and Where They Can't (Yet)." *McKinsey Quarterly*. At <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/where-machines-could-replace-humans-and-where-they-cant-yet> (accessed March 2023)
- Chui, M., K. George, & M. Miremadi. (2017, July). "ACEO Action Plan for Workplace Automation." *McKinsey Quarterly*. At <https://www.mckinsey.com/featured-insights/digital-disruption/a-ceo-action-plan-for-workplace-automation> (accessed March 2023).
- Crosman, P. (2017, August 17). "Why Cybercriminals Like AI As Much as Cyberdefenders Do." *American Banker*.
- Dastin, J. (2018), "Amazon scraps secret AI recruiting tool that showed bias against women," at <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G> (accessed January 2023)
- Daugherty, P. R., & J. Wilson. (2018). *Human+Machine: Reimagining Work in the Age of AI*. Boston, MA: Business Review Press.
- de Vos, B. (2018, July 11). "Opinion: These 3 Business Functions Will Be the First to Benefit from Artificial Intelligence." *Information Management*.
- Dickson, B. (2017, July 28). "What Is the Future of Artificial Intelligence?" *Tech Talk*.
- Donahue, L. "A Primer on Using Artificial Intelligence in the Legal Profession." *Jolt Digest*, January 3, 2018.
- Dormehl, L. (2017). *Thinking Machines: The Quest for Artificial Intelligence—and Where It's Taking Us Next*. New York, NY: TarcherPerigee.
- Egan, M. (2015, May 13). "Robots Threaten These 8 Jobs." *CNNMoney.com*.
- Gaudin, S. (2016, October 26). "1-800-Flowers Wants to Transform Its Business with A.I." *Computer World*.
- Goldman, S. (2018, March 22). "The Ethics of Legal Analytics." *Law.com*.
- Carlos A. Gomez-Urbe and Neil Hunt. 2015. The Netflix recommender system: Algorithms, business value, and innovation. *ACM Trans. Manage. Inf. Syst.* 6, 4, Article 13 (December 2015), 19 pages. DOI: <http://dx.doi.org/10.1145/2843948>
- Hu, F. *Security and Privacy in Internet of Things (IoT): Models, Algorithms, and Implementations*. Boca Raton, FL: CRC Press, 2016.
- Huff, E. (2017, January 17). "Proof That Amazon Devices Are Spies in Your Own Home: Alexa Automatically Orders Product after 'Hearing' Audio in Private Homes." *Natural News*.
- Keenan, J. (2018, February 13). "1-800-Flowers.com Using Technology to Win Customers' Hearts This Valentine's Day." *Total Retail*.
- Kelly, H. (2018, January 29). "Robots Could Kill Many Las Vegas Jobs." *Money.CNN.com*.
- Kiron, D. (2017, January 25). "What Managers Need to Know about Artificial Intelligence." *MIT Sloan Management Review*.
- Krauth, O. (2018, January 23). "Robot Gender Gap: Women Will Lose More Jobs Due to Automation Than Men, WEF Finds." *Tech Republic*.
- Krigsman, M. (2017, January 30). "Artificial Intelligence: Legal, Ethical, and Policy Issues." *ZDNet*.
- Larson J., Surya Mattu, Lauren Kirchner and Julia Angwin (2016). "How We Analyzed the COMPAS Recidivism Algorithm," at <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm> (accessed January 2023).

- Leggatt, H. (2017, June 7). "Biggest Stressor in U.S. Workplace Is Fear of Losing Jobs to AI, New Tech." *Biz Report*.
- Manyika, J. (2017, May). "Technology, Jobs, and the Future of Work." *McKinsey Global Institute*.
- Maheshwari, B.; Varua, M.; Ward, J.; Packham, R.; Chinnassamy, P.; Dashora, Y.; Dave, S.; Soni, P.; Dillon, P.; Purohit, R.; et al. (2022). The role of transdisciplinary approach and community participation in village scale groundwater management: Insights from Gujarat and Rajasthan, India. *Int. Open Access J. Water*, 6, 3386–3408.
- McFarland, M. (2017b, September 15). "Robots: Is Your Job at Risk?" CNN News.
- McLellan, C. "Cybersecurity in an IoT and Mobile World." Special Report. *ZDNet*, June 1, 2017b.
- Meyer, D. "It is painful to see some of these offensive responses"—Meta defends its occasionally anti-Semitic BlenderBot 3 chatbot," at <https://fortune.com/2022/08/09/painful-offensive-responses-meta-blenderbot-3-chatbot-tay/> (accessed January 2023)
- Morgan, B. (2017, June 13). "Ethics and Artificial Intelligence with IBM Watson's Rob High." *Forbes*.
- Morris, D. (2017, February 18). "Bill Gates Says Robots Should Be Taxed Like Workers." *Fortune.com*.
- O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown Publishing).
- Park, H. "The Connected Customer: The Why behind the Internet of Things." *Blue Hill Research*. White Paper. January 2017.
- Palmer, S. (2017, February 26). "The 5 Jobs Robots Will Take First." *Shelly Palmer*.
- Perkins, E. "Securing the Internet of Things." Report G00300281. *Gartner Inc.*, May 12, 2016.
- Pham, S. (2018, February 21). "Control AI Now or Brace for Nightmare Future, Experts Warn." *Money.cnn.com (News)*.
- Provazza, A. (2017, May 26). "Artificial Intelligence Data Privacy Issues on the Rise." *Tech Target (News)*.
- Rainie, L., & J. Anderson. (2017, June 6). "The Internet of Things Connectivity Binge: What Are the Implications?" *Pew Research Center*.
- Ransbotham, S. (2016). "How Will Cognitive Technologies Affect Your Organization?" <https://sloanreview.mit.edu/article/how-will-cognitive-technologies-affect-your-organization/> (accessed January 2023).
- Ross, J. (2018, Winter). "The Fundamental Flaw in AI Implementation." *MIT Sloan Management Review*. <https://sloanreview.mit.edu/article/the-fundamental-flaw-in-ai-implementation/> (accessed January 2023).
- Santori, C. Keith, R.J., Whittington, C.M., Thompson, M.B., Van Dyke, J.U., and Spencer, R.J. (2021). Changes in participant behaviour and attitudes are associated with knowledge and skills gained by using a turtle conservation citizen science app. *People and Nature* 3 (1), 66–76.
- SAS. "5 Steps for Turning Industrial IoT Data into a Competitive Advantage." *SAS White Paper 108670_G456z 0117.pdf*, January 2017.
- Setty, R. "First AI Art Generator Lawsuits Threaten Future of Emerging Tech," at <https://news.bloomberglaw.com/ip-law/first-ai-art-generator-lawsuits-threaten-future-of-emerging-tech> (accessed February 2023)
- Shchutskaya, V. and A. Zharovskikh (2020). (2020). "Major Problems of Artificial Intelligence Implementation" *InData Labs*. <https://indatalabs.com/blog/problems-of-artificial-intelligence-implementation> (accessed January 2023).
- Sherman, E. (2015, February 25). "5 White-Collar Jobs Robots Already Have Taken." *Fortune.com* <https://fortune.com/2015/02/25/5-jobs-that-robots-already-are-taking/> (accessed January 2023).
- Singh, G. (2017a, September 20). "Opinion: 5 Components That Artificial Intelligence Must Have to Succeed." *Health Data Management*.
- Smith, Ms. (2018, March 12). "Ransomware: Coming to a Robot near You Soon?" *CSO, News*.
- Snyder, A. (2017, September 6). "Executives Say AI Will Change Business, But Aren't Doing Much About It." *Axios.com*.
- Spangler, T. (2017, November 24). "Self-Driving Cars Programmed to Decide Who Dies in a Crash." *USA Today*.
- Standage, T. (2016) "The Return of the Machinery Question." Special Report. *The Economist*. economist.com/sites/default/files/ai_mailout.pdf (accessed January 2023).
- Straus, R. (2014, May 31). "Will You Be Replaced by a Robot? We Reveal the 100 Occupations Judged Most and Least at Risk of Automation." *ThisIsMoney.com*. <https://www.thisismoney.co.uk/money/news/article-2828370/The-robots-coming-One-three-UK-jobs-high-risk-replaced-machines-20-years.html> (accessed January 2023).
- Uzialko, A. (2017, October 13). "AI Comes to Work: How Artificial Intelligence Will Transform Business." *Business News Daily*.
- Violino, B. (2018, February 21). "Most Workers See Smart Robots as Aid to Their Jobs, Not Threat." *Information Management*. *WallStreetJournal.com*. (2016). "What They Know." wsj.com/public/page/what-they-know-digital-privacy.html (accessed Jan 2023).
- Weldon, D. "Steps for Getting an IoT Implementation Right." *Information Management*, October 30, 2015.
- West, D. (2018). *The Future of Work: Robots, AI, and Automation*. Washington, DC: Brookings Institute Press.

a priori algorithm The most commonly used algorithm to discover association rules by recursively identifying frequent itemsets.

activation function a mathematical derivations to determine the output of a PE, usually composed of a summation followed by a transfer function, producing a numerical output between 0 and 1 or -1 and 1.

Alexa A cloud-based chat bot developed by Amazon.

Alternative Data New generation of big data extensions that imagery and satellite data.

Anaconda Anaconda is a popular distribution option for Python and R programming languages for scientific computing that aims to simplify package management and deployment.

Analytics as a Service Analytics serviced provided by a third party, usually on a cloud platform.

analytics ready A state of preparedness for analytics projects, especially as it relates to data acquisition and preparedness.

ArcGIS A cloud-based mapping and spatial analytics software platform offered by Esri.

area under the ROC curve A graphical assessment technique for binary classification models where the true positive rate is plotted on the y -axis and false positive rate is plotted on the x -axis.

arithmetic mean A descriptive statistics measure—a simple average of a given sample of numeric data points.

artificial intelligence The subfield of computer science concerned with symbolic reasoning and problem solving.

artificial neural network (ANN) Computer technology that attempts to build computers that operate like a human brain. The machines possess simultaneous memory storage and work with ambiguous information. Sometimes called, simply, a *neural network*.

association A category of data mining algorithm that establishes relationships about items that occur together in a given record.

augmented intelligence The AI technology that augments computer tasks to extend human cognitive abilities, resulting in high performance.

authoritative pages Web pages that are identified as particularly popular based on links by other Web pages and directories.

automation Automating some or all the steps in the analytics solution development process.

autonomous car A sensor and AI enhanced car that can drive itself without a driver.

Autonomous vehicles A sensor and AI enhanced vehicles that can operate itself without the help of a human.

Autonomy Being self-sufficient to carry out tasks, not needing the help of a human.

AWS AWS (Amazon Web Services, Inc.) is a subsidiary of Amazon that provides on-demand cloud computing platforms and APIs to individuals, companies, and governments, on a metered, pay-as-you-go basis.

Backpropagation A learning method used to train MLP-type neural networks where the error is systematically fed backward into the neural network to gradually adjust its weights.

Big Data analytics Application of analytics methods and tools to Big Data.

Big Data Data that is characterized by its volume, variety, and velocity that exceeds the reach of commonly used hardware environments and/or capabilities of software tools to process.

black-box syndrome The concept of complex machine learning algorithms, e.g., neural networks, not being transparent to explain their prediction process.

blockchain A distributed database or ledger shared among a computer network's nodes.

bootstrapping A sampling technique where a fixed number of instances from the original data are sampled (with replacement) for training, and the rest of the data set is used for testing.

box plot A descriptive statistics tool that shows both central tendency and dispersion (quartiles) of a given sample of numeric data points in a graphical illustration.

box-and-whiskers plot A graphical method of displaying descriptive statistics (central tendency and dispersion) of a numerical variables.

Bubble charts A type of scatter plots where the size of the bubbles can be used to show a extra dimension of information.

business intelligence (BI) A conceptual framework for managerial decision support. It combines architecture, databases (or data warehouses), analytical tools, and applications.

business report Any communication artifact prepared with the specific intention of conveying information in a digestible form to whoever needs it, whenever and wherever they may need it.

Caffe A well-known, community developed, deep learning framework.

Canopeo A free desktop or mobile app that estimates green canopy cover in near real time from images taken with a smartphone or digital camera.

categorical data Data that represent the labels of multiple classes used to divide a variable into specific groups.

Categorical A nominal data type that contains two or more categories as its values.

centrality A group of metrics that aims to quantify the importance or influence (in a variety of senses) of a particular node (or group) within a network.

certainty The business situation where complete knowledge is available so that the decision maker knows exactly what the outcome of each course of action will be.

Chatbots A AI-based computer program that simulates and processes human conversation.

citizen data scientists A citizen data scientist is an individual who does some data science work for an organization but doesn't hold the title of data scientist or have a formal background in advanced analytics, statistics or related disciplines.

classification Supervised induction used to analyze the historical data stored in a database and to automatically generate a model that can predict future behavior.

Clementine Clementine is one of the earliest software tools that offered a visual intuitive workspace approach for data mining. Later, it was acquired by SPSS, and then SPSS acquired by IBM, changing its current name to IBM SPSS Modeler.

clickstream analysis The analysis of data that occur in the Web environment.

Closed-source software This is the opposite of open-source software, where the source code is not accessible or reusable to the end users.

cloud computing Information technology infrastructure (hardware, software, applications, and platform) that is available as a service, usually as virtualized resources.

Cloud Security Alliance (CSE) The world's leading organization dedicated to defining and raising awareness of best practices to help ensure a secure cloud computing environment.

Cloud-based software Server based software that runs on the cloud, often via a web browser, not needing a local installation.

clustering Partitioning a given data set into segments (natural groupings) in which the members of a segment share similar qualities.

Cognitive analytics A term that refers to cognitive computing-branded technology platforms, such as IBM Watson.

cognitive computing The use of computerized models to simulate the human thought process in complex situations where the answers may be ambiguous and uncertain.

Cognitive search A new generation search method that uses AI (advanced indexing, NLP, and machine learning) to return results that are much more relevant to users.

comorbidity networks A network diagrams that shows links/affinities between different diagnoses/comorbidities.

Computer vision An interdisciplinary field that deals with how computers can create high-level understanding of digital images or videos.

confidence In association rules, the conditional probability of finding the RHS of the rule present in a list of transactions where the LHS of the rule exists.

Connection weights They are the relative strength of connections between processing elements in neural networks.

Constant Error Carousel (CEC) Also known as the state unit, in LSTM deep learning architecture, is the integration mechanism of the gates their interactions with the other layers.

convolution function In CNN, it extracts simple patterns like the existence of horizontal or vertical lines or edges in different parts of a picture.

convolutional neural network (CNN) Is a variation of the deep MLP architecture, initially designed for computer vision applications but are also applicable to nonimage data sets.

coronavirus pandemic A worldwide pandemic caused by the corona virus in 2019.

corpus In linguistics, a large and structured set of texts (usually stored and processed electronically) prepared for the purpose of conducting knowledge discovery.

correlation A statistical measure that indicates the extent to which two or more variables change/fluctuate together.

CRISP-DM A cross-industry standardized process of conducting data mining projects, which is a sequence of six steps that starts with a good understanding of the business and the need for the data mining project (i.e., the application domain) and ends with the deployment of the solution that satisfied the specific business need.

cube In data warehousing, it is multi-dimensional data representation for conducting efficient data analysis.

dashboard A visual presentation of critical data for executives to view. It allows executives to see hot spots in seconds and explore the situation.

data integration Integration that comprises three major processes: data access, data federation, and change capture. When these three processes are correctly implemented, data can be accessed and made accessible to an array of ETL, analysis tools, and data warehousing environments.

data mart (DM) A departmental data warehouse that stores only relevant data.

data mining A process that uses statistical, mathematical, artificial intelligence, and machine-learning techniques to extract and identify useful information and subsequent knowledge from large databases.

data preprocessing A tedious process of converting raw data into an analytic ready state.

data quality The holistic quality of data, including their accuracy, precision, completeness, and relevance.

data scientist A new role/job/title commonly associated with Big Data or data science.

data taxonomy A structured representation of the subgroups/subtypes of data.

data visualization A graphical, animation, or video presentation of data and the results of data analysis.

data warehouse (DW) A physical repository where relational data are specially organized to provide enterprise-wide, cleansed data in a standardized format.

data warehouse administrator (DWA) A person responsible for the administration and management of a data warehouse.

datum A piece of information/fact; singular version of data.

decentralized autonomous organizations (DAOs) A group of players that work on Web 3.0 to permit the ability to vote on future protocols and proposals.

deception detection A way of identifying deception (intentionally propagating beliefs that are not true) in voice, text, and/or body language of humans.

decision analysis A modeling approach that deals with decision situations that involve a finite and usually not too large number of alternatives.

decision or normative analytics Also called prescriptive analytics, is a type of analytics modeling that aims at identifying the best possible decision from a large set of alternatives.

decision table A tabular representation of possible condition combinations and outcomes.

decision tree A graphical presentation of a sequence of interrelated decisions to be made under assumed risk. This technique classifies specific entities into particular classes based on the features of the entities; a root followed by internal nodes, each node (including root) is labeled with a question, and arcs associated with each node cover all possible responses.

decision variable The variable of interest.

deep belief networks (DBNs) A type of a large class of unsupervised deep neural networks called generative models.

deep learning A type of machine learning based on artificial neural networks in which multiple layers of processing are used to extract progressively higher-level features from data.

deep neural networks A neural network with arbitrary number of layers, usually with more than two hidden layers.

DeepQA A software architecture for deep content analysis and evidence-based reasoning (e.g., IBM Watson)

Deforestation Loss of forests due to overcutting of trees.

dependent data mart A data mart that depends on the existence of a data warehouse.

descriptive (or reporting) analytics An earlier phase in analytics continuum that deals with describing the data—answering the question of what happened and why did it happen.

descriptive statistics A branch of statistical modeling that aims to describe a given sample of data (*also see* inferential statistics).

digital twins A virtual model designed to accurately reflect a physical object.

dimension tables In a data warehouse, surrounding the central fact tables (and linked via foreign keys) are called dimension tables.

dimensional modeling A retrieval-based data querying system that supports high-volume, high-speed access to subsets of data.

dimensional reduction An iterative/heuristic process of reducing the number of input variables to a manageable number—identifying the most prevalent/important/contributing variables to include in the modeling activities.

discrete event simulation A type of simulation modeling where a system is studied based on the occurrence of events/interaction between different parts (entities/resources) of the system.

dispersion A descriptive statistics measure where the spread of a given sample of numeric data points is assessed.

distance measure A method used to calculate the closeness between pairs of items in most cluster analysis methods. Popular distance measures include Euclidian distance (the ordinary distance between two points that one would measure with a ruler) and Manhattan distance (also called the rectilinear distance, or taxicab distance, between two points).

drill down The investigation of information in detail (e.g., finding not only total sales but also sales by region, by product, or by salesperson). Finding the detailed sources.

driverless cars A car that can drive itself without a driver.

EarthCast An application that provides correct weather updates for a large commercial U.S. airline based on variety of data it pulls from disparate sources.

Echo A smart speakers system developed by Amazon to be used with Alexa.

edge computing An emerging computing paradigm where computations are performed at the devices near the user (as opposed to at the server).

enterprise application integration (EAI) A technology that provides a vehicle for pushing data from source systems into a data warehouse.

enterprise data warehouse (EDW) An organizational-level data warehouse developed for analytical purposes.

enterprise information integration (EII) An evolving tool space that promises real-time data integration from a variety of sources, such as relational databases, Web services, and multidimensional databases.

entropy A metric that measures the extent of uncertainty or randomness in a data set. If all the data in a subset belong to just one class, then there is no uncertainty or randomness in that data set, and therefore the entropy is zero.

environmental scanning and analysis A continuous process of intelligence building—identification of problems and/or opportunities via acquisition and analysis of data/information.

ethical AI Ethical AI is artificial intelligence that adheres to well-defined ethical guidelines regarding fundamental values, including such things as individual rights, privacy, non-discrimination, and non-manipulation.

extraction, transformation, and load (ETL) A data warehousing process that consists of extraction (i.e., reading data from a database), transformation (i.e., converting the extracted data from its previous form into the form in which it needs to be so that it can be placed into a data warehouse or simply another database), and load (i.e., putting the data into the data warehouse).

Fifth generation (5G) The latest and fastest cellular data transmission technology (mobile network).

forecasting Using the data from the past to foresee the future values of a variable of interest.

Forrester Forrester is one of the popular consultancy and advisory companies in IT and business management that offers a variety of services including research, consulting, and events.

Forrester Wave Forrester Wave is a graphical representation of Forrester's assessment of a specific market segment and is graphically presented using a detailed spreadsheet with exposed scores, weightings, and comments.

Fourth Industrial Revolution A fusion of technologies that is synergizing the physical, digital, and biological spheres.

Gartner Gartner, Inc. is a technological research and consulting firm that conducts research on technology and shares this research both through private consulting as well as executive programs and conferences.

Genetic algorithms A method for solving (via a heuristic search) optimization problems that is based on natural selection.

Genomics the study of human genes (the genome).

geographical information system (GIS) An information system capable of integrating, editing, analyzing, sharing, and displaying geographically referenced information.

geospatial analytics Use of locational data (in addition to traditional types of data) while performing data analytics.

Gini index A metric that is used in economics to measure the diversity of the population. The same concept can be used to determine the purity of a specific class as a result of a decision to branch along a particular attribute/variable.

goal seeking A prescriptive analytics method where first a goal (a target/desired value) is set, and then the satisfying set of input variable values is identified.

GO-JEK A forecasting and pricing platform that employs Google cloud.

Google Colab Collaboratory ("Colab" for short) is a data analysis and machine learning tool that allows end users to combine executable Python code and variety of data into a single document stored in Google Drive.

Google Lens A search engine for the real-world visual images.

Google's Assistant A virtual assistant software application developed by Google that is primarily available on mobile and home automation devices.

GoogLeNet A type of convolutional neural network based on the Inception architecture.

graphics processing unit (GPU) A specialized processor originally designed to accelerate graphics rendering but now it is also used for AI and deep learning applications.

Hadoop Distributed File System (HDFS) A distributed file management system that lends itself well to processing large volumes of unstructured data (i.e., Big Data).

Hadoop An open source framework for processing, storing, and analyzing massive amounts of distributed, unstructured data.

hidden layer The layer that exist between the input and output layers in an MLP neural network architecture.

high-performance computing A large-scale computing infrastructure to deal with Big Data.

histogram A statistical chart that shows frequency of bins for a given sample of numeric data.

Hive A Hadoop-based data warehousing-like framework originally developed by Facebook.

hub One or more Web pages that provide a collection of links to authoritative pages.

hyperlink-induced topic search (HITS) The most popular, publicly known, and referenced algorithm in Web mining, which is used to discover hubs and authorities.

IBM SPSS Modeler A popular, commercially available, comprehensive data, text, and Web mining software suite developed by SPSS (formerly Clementine).

IBM Watson A cognitive analytics based computer architecture popularized by competing and winning the TV gameshow, Jeopardy!.

ImageNet An image database organized according to the WordNet hierarchy.

independent data mart A small data warehouse designed for a strategic business unit or a department.

inferential statistics A branch of statistical modeling that aims to draw inferences or conclusions about the characteristics of the population based on a given sample of data (*also see* descriptive statistics).

influence diagram A graphical representation of a given mathematical model.

information gain The splitting mechanism used in ID3 (a popular decision tree algorithm).

intelligent agent An expert or knowledge-based system embedded in computer-based information systems (or their components) to make them smarter.

intermediate results Intermediate outcomes in mathematical models.

Internet of Things (IoT) The technological phenomenon of connecting a variety of devices in the physical world to each other and to the computing systems via the Internet.

interval data Variables that can be measured on interval scales.

JMP Pro JMP Pro is a menu-based, intuitive, statistics and predictive analytics software tool offered by SAS Institute.

Jupyter Notebook Jupyter Notebook is a server-client application that allows editing and running notebook documents, allowing computer coding in most popular languages such as Python and R, via a web browser.

KNuggets KNuggets is an online platform that disseminates information and resources on business analytics, big data, data mining, and data science.

Keras An API designed to consume existing deep learning libraries to perform deep learning tasks.

key performance indicator (KPI) Measure of performance against a strategic objective and goal.

k-fold cross-validation A popular accuracy assessment technique for prediction models where the complete data set is randomly split into k mutually exclusive subsets of approximately equal size. The classification model is trained and tested k times. Each time it is trained on all but one fold and then tested on the remaining single fold. The cross-validation estimate of the overall accuracy of a model is calculated by simply averaging the k individual accuracy measures.

KNIME A free, open source analytics software platform (can be accessed at knime.org).

KNIME Analytics Platform KNIME Analytics Platform is an intuitive, visual, free, and open-source software product that allows users to access, blend, analyze, and visualize data, without any coding.

KNIME Hub KNIME Hub is the place where the members of the KNIME community can search for nodes, extensions, components and workflows made available by the other member of the community.

knowledge discovery in databases (KDD) A machine-learning process that performs rule induction or a related procedure to establish knowledge from large databases.

knowledge Understanding, awareness, or familiarity acquired through education or experience; anything that has been learned, perceived, discovered, inferred, or understood; the ability to use information. In a knowledge management system, knowledge is information in action.

kurtosis A statistical measure to characterize the shape of a unimodal distribution—characterizing the peak/tall/skinny nature of the distribution (*also see* skewness).

KXEN KXEN is an analytics software company which existed as an independent entity until 2013 when it was acquired by SAP AG.

LDA Latent Dirichlet Allocation (LDA) is one of the most popular methods for statistical topic modeling.

learning A process of self-improvement where the new knowledge is obtained through a process by using what is already known.

lift A goodness-of-fit measure for classification as well as association rule mining models.

linear programming (LP) A mathematical modeling technique used to represent and solve constraint optimization problems.

linear regression A relatively simple statistical technique to model the linear relationship between a response variable and one or more explanatory/input variables.

link analysis The linkage among many objects of interest is discovered automatically, such as the link between Web pages and referential relationships among groups of academic publication authors.

logistic regression A very popular, statistically sound, probability-based classification algorithm that employs supervised learning.

long short-term memory (LSTM) A type of recurrent neural network capable of learning order dependence in sequence prediction.

machine learning A branch of artificial intelligence which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

machine vision The ability of a computer to see by employing one or more video cameras.

machine-to-machine (M2M) communications CH10

Magic Quadrant Gartner's Magic Quadrant is a visual snapshot of in-depth analyses and actionable guidance about an IT and business related market segment.

MapReduce A technique to distribute the processing of very large multistructured data files across a large cluster of machines.

mathematical programming A family of analytic tools designed to help solve managerial problems in which the decision maker must allocate scarce resources among competing activities to optimize a measurable goal.

mean absolute deviation An accuracy metric commonly used for regression-type prediction problems (e.g., time series forecasting) where the error is calculated as the average squared distance between the actuals and the predicted values.

median A central tendency measure in statistics that identifies the simple average of a given numeric data sample.

metadata Data about data. In a data warehouse, metadata describes the contents of a data warehouse and the manner of its use.

metaverse The emerging 3-D-enabled digital space that uses virtual reality, augmented reality, and other advanced internet technologies.

Microbiome Quotient A complex group of microorganisms most of which are bacteria.

Microsoft Enterprise Consortium Worldwide source for access to Microsoft's SQL Server software suite for academic purposes—teaching and research.

Microsoft SQL Server A popular RDBM system developed by Microsoft.

Miniconda Miniconda is a small, bootstrap version of Anaconda that includes only Conda, Python, the packages they depend on, and a small number of other useful packages.

Mobile Extension Framework (MXFTM) A new addition to the Mobile iOS SDK that allows apps to save events in app extensions.

mode A central tendency measure in statistics that identifies the value in the fiftieth percentile.

Monte Carlo simulation A simulation technique that relies on change/probability distribution to represent the uncertainty in the modeling of the decision problem.

multidimensional (modeling) analysis A modeling method that involves data analysis in several dimensions.

multilayer perceptron (MLP) A neural network architecture that is composed of two or more consecutive layers.

multiple goals Having more than just one goal to consider in an optimization problem.

Mutation A genetic algorithms operator for random switching of binary digits.

MYCIN An expert system developed to diagnose and treat bacterial infections.

natural language processing (NLP) Using a natural language processor to interface with a computer-based system.

Natural language processing A branch of artificial intelligence that is concerned with granting computers the ability to understand and create human language.

network analytics The analysis of network data to identify trends and patterns.

network structure The manner in which the neurons and layers are organized in different neural network architectures.

neural network See artificial neural network (ANN).

neuron A biological nerve cell, or the equivalent for a processing element in artificial neural networks.

nominal data A type of data that contains measurements of simple codes assigned to objects as labels, which are not measurements. For example, the variable *marital status* can be generally categorized as (1) single, (2) married, and (3) divorced.

non-fungible tokens (NFT) Unique cryptographic tokens that exist on a blockchain and cannot be replicated

NoSQL (Not Only SQL) A new paradigm to store and process large volumes of unstructured, semistructured, and multistructured data.

numeric data A type of data that represents the numeric values of specific variables. Examples of numerically valued variables include age, number of children, total household income (in U.S. dollars), travel distance (in miles), and temperature (in Fahrenheit degrees).

online analytical processing (OLAP) An information system that enables the user, while at a PC, to query the system, conduct an analysis, and so on. The result is generated in seconds.

online transaction processing (OLTP) Transaction system that is primarily responsible for capturing and storing data related to day-to-day business functions.

open-source software The software that is distributed with its source code, making it available for use, modification, and distribution with its original rights.

oper mart An operational data mart. An oper mart is a small-scale data mart typically used by a single department or functional area in an organization.

operational data store (ODS) A type of database often used as an interim area for a data warehouse, especially for customer information files.

Optical Character Recognition The process of converting an image of text into a machine-readable text format.

optimal solution The best possible solution to a problem.

Orange An open source, free, workflow-type analytics software tool.

Orbital Insight An AI-powered Geospatial Analytics platform designed to simplify the use of location data.

ordinal data Data that contain codes assigned to objects or events as labels that also represent the rank order among them. For example, the variable *credit score* can be generally categorized as (1) low, (2) medium, and (3) high.

ordinary least squares (OLS) A method that relies on the square of the distance measure to identify the best fitting line/plane/hyperplane in regression modeling.

Overfitting Over training of a machine learning model where not only signals but also noise in the training data is captured.

parameter Numeric constants used in mathematical modeling.

part-of-speech (POS) tagging The process of marking up the words in a text as corresponding to a particular part of speech (such as nouns, verbs, adjectives, adverbs) based on a word's definition and context of its use.

Perceptron An artificial neuron using a step function as the activation function.

performance function A function that is used to measure the network's performance.

pie chart A graphical illustration of proportions.

Pig Pig is an open-source high level data flow architecture that provides a simple language for queries and data manipulation.

polarity identification The process of identifying negative or positive connotations in text (in sentiment analysis).

polysemes Words also called *homonyms*; they are syntactically identical words (i.e., spelled exactly the same) with different meanings (e.g., *bow* can mean “to bend forward,” “the front of the ship,” “the weapon that shoots arrows,” or “a kind of tied ribbon”).

pooling It is a layer used in deep learning to consolidate elements in the input matrix in order to produce a smaller output matrix while maintaining the important features.

PowerBI A popular data visualization and visual analytics software tools and platform offered by Microsoft.

prediction The act of telling about the future.

predictive analytics A business analytical approach toward forecasting (e.g., demand, problems, opportunities) that is used instead of simply reporting data as they occur.

prescriptive analytics A branch of business analytics that deals with finding the best possible solution alternative for a given problem.

privacy concerns Privacy concerns refer to the worry that the personal data of individuals may be used by others in an unpredictable way.

processing elements [PEs] A simulation of biological neuron, where it receives a number of input signals and, based on those inputs, either generate a single output signal (fire) or do not.

Proprietary software Proprietary software is primarily commercial software, rights of which are owned by a vendor/developer, which can be bought, leased, or licensed from its owner.

PyCharm PyCharm is a hybrid platform developed by JetBrains as an IDE for Python coding and application development.

Python Python is an interpreted, object-oriented, high-level programming language with dynamic semantics, which has recently become the most popular tool to develop analytics and data science applications.

quantitative model Mathematical models that rely on numeric/quantifiable measures.

quartile One-fourth (i.e., quarter) of a sorted numeric/ordinal data sample.

R R is a free, community-supported software environment for statistical and analytics computing as well as graphics development.

radio-frequency identification (RFID) A generic technology that refers to the use of radio-frequency waves to identify objects.

range A statistics measure for dispersion—the distance between the smallest and largest values within a given sample of numeric data points.

RapidMiner A popular, open source, free-of-charge data mining software suite that employs a graphically enhanced user interface, a rather large number of algorithms, and a variety of data visualization features.

ratio data Continuous data where both differences and ratios are interpretable. The distinguishing feature of a ratio scale is the possession of a nonarbitrary zero value.

Rattle A data mining user interface developed for R.

real-time data warehousing (RDW) The process of loading and providing data via a data warehouse as they become available.

recurrent neural network (RNN) A type of artificial neural network which uses sequential data or time series data, and can have bi-directional connection among neurons and layers.

regression A data mining method for real-world prediction problems where the predicted values (i.e., the output variable or dependent variable) are numeric (e.g., predicting the temperature for tomorrow as 68°F).

report Any communication artifact prepared with the specific intention of conveying information in a presentable form.

representation learning A class of machine learning approaches that allow a system to discover the representations required for feature space.

reproduction A genetic algorithms operator to create new potential solutions.

result (outcome) variable A variable that expresses the result of a decision (e.g., one concerning profit), usually one of the goals of a decision-making problem.

risk analysis Use of mathematical modeling to assess the nature of risk (variability) for a decision situation.

risk A probabilistic or stochastic decision situation.

Robo advisors An online application that provides automated guidance and services.

Robot A machine or a physical device or software that with the cooperation of AI can accomplish a responsibility autonomously.

robotic process automation (RPA) A technology intended to reduce the workloads of the employees by automating the repetitive, high-volume steps in processes.

RStudio RStudio is an integrated development environment for R programming language.

SAS Enterprise Miner A comprehensive, commercial data mining software tool developed by the SAS Institute.

SAS Viya A cloud-based analytics platform developed by SAS Institute.

scatter plot A graph in which the values of two variables are plotted along two axes to illustrate the relationship between them.

scene recognition A form of artificial intelligence that uses deep neural networks to identify objects, people, and scenes in images and videos.

search engine A program that finds and lists Web sites or pages (designated by URLs) that match some user-selected criteria.

SEMMA An alternative process for data mining projects proposed by the SAS Institute. The acronym “SEMMA” stands for “sample, explore, modify, model, and assess.”

sensitivity analysis A study of the effect of a change in one or more input variables on a proposed solution.

sensors A variety of devices that detect and respond to some type of input from the physical environment.

sentiment analysis The technique used to detect favorable and unfavorable opinions toward specific products and services using a large numbers of textual data sources (customer feedback in the form of Web postings).

SentiWordNet An extension of WordNet to be used for sentiment identification. *See* WordNet.

sequence mining A pattern discovery method where relationships among things are examined in terms of their order of occurrence to identify associations over time.

service-based architecture (SBA) A modular computational framework that provides cloud-based services on an as needed basis.

simple split Data is partitioned into two mutually exclusive subsets called a *training set* and a *test set* (or *holdout set*). It is common to designate two-thirds of the data as the training set and the remaining one-third as the test set.

Simulation Simulation is the process of mathematical modelling, performed on a computer, which is designed to predict the behavior of, or the outcome of, a real-world or physical system.

singular value decomposition (SVD) Closely related to principal components analysis, it reduces the overall dimensionality of the input matrix (number of input documents by number of extracted terms) to a lower dimensional space, where each consecutive dimension represents the largest degree of variability (between words and documents).

Siri Apple’s AI-based virtual assistant for iOS and macOS devices that uses voice.

skewness A statistical measure to characterize the shape of a unimodal distribution—characterizing the asymmetry (sway) of the distribution (*also see* kurtosis).

Skype Translator A speech-to-speech translation of different languages application developed by Skype, which has operated as a division of Microsoft since 2018.

Smartbin A new high technology that integrated waste containers with smart sensors, which allows you to track through the waste management processes.

smart sensor A smart sensor is a device that takes input from the physical environment and uses built-in intelligence to perform predefined functions.

snowflake schema A logical arrangement of tables in a multidimensional database in such a way that the entity relationship diagram resembles a snowflake in shape.

social media analytics Application of analytics tools to social media and social network data.

social network A network of social interactions and personal relationships.

social robot A robot capable of interacting with humans and other robots.

software as a service (SaaS) Software that is rented instead of sold.

Software robots Simple or complex computer programs designed to do specific actions, such as automating repetitive tasks or simulating human users.

Speech (voice) understanding Computer systems that can understand human speech and voice.

spider *See* Web crawler.

Spyder Spyder is a free and open source scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts.

standard deviation A descriptive statistics measure for dispersion. It is the square root of the variance.

star schema Most commonly used and simplest style of dimensional modeling.

Statistica Statistica is an analytics software product that provide data analysis, data management, statistics, data mining, machine learning, text analytics and data visualization procedures (now owned by TIBCO).

statistics A collection of mathematical techniques to characterize and interpret data.

stemming A process of reducing words to their respective root forms to better represent them in a text mining project.

stochastic gradient descent (SGD) An iterative method for optimizing an objective function with suitable smoothness properties.

stop words Words that are filtered out prior to or after processing of natural language data (i.e., text).

storytelling A case with rich information and episodes. Lessons may be derived from this kind of case in a case base.

stream analytics A term commonly used for extracting actionable information from continuously flowing/streaming data sources.

structured data Data that is formatted (often into tables with rows and columns) for computers to easily understand and process.

summation function A function that combines the various input activations into a single output activation.

supervised learning A subcategory of machine learning and artificial intelligence that learns from labelled data.

support The measure of how often products and/or services appear together in the same transaction; that is, the

proportion of transactions in the data set that contain all the products and/or services mentioned in a specific rule.

Tableau Tableau is a popular visual analytics platform for data analytics and visualization.

Tensorflow An end-to-end open-source machine learning platform for deep learning.

Teradata Vantage Teradata Vantage is the connected multi-cloud data platform for enterprise analytics that unifies data lakes, data warehouses, machine learning, and new data sources and types.

term-document matrix (TDM) A frequency matrix created from digitized and organized documents (the corpus) where the columns represent the terms and rows represent the individual documents.

text mining The application of data mining to nonstructured or less structured text files. It entails the generation of meaningful numeric indices from the unstructured text and then processing those indices using various data mining algorithms.

Theano A Python library for manipulating and evaluating mathematical expressions, especially matrix-valued ones, in deep learning applications.

threshold value In neural network, it is a value that determines the dichotomous outcome of an activation function.

TIBCO TIBCO Software Inc. is an American business intelligence and data analytics software company founded in 1997 in Palo Alto, California.

time series forecasting A prediction model that relies solely on the past occurrences/values of the variable of interest to estimate/calculate the expected future values.

tokenizing Categorizing a block of text (token) according to the function it performs.

Torch An open-source machine learning library, a scientific computing framework, and a scripting language based on Lua.

transfer function It is a mathematical transformation of the input (that comes from the summation functions) into output of the processing element.

trend analysis The collecting of information and attempting to spot a pattern, or *trend*, in the information.

Turing Test A test for intelligence in a computer, requiring that a human being should be unable to distinguish the machine from another human being by using the replies to questions put to both.

ubiquitous computing The concept of using small internet connected and inexpensive computers to help with everyday functions in an automated fashion.

uncertainty A decision situation where there is a complete lack of information about what the parameter values are or what the future state of nature will be.

uncontrollable variable A mathematical modeling variable that has to be taken as given—not allowing changes/modifications.

unique identifier (UID) An identifier that is guaranteed to be unique among all identifiers used for those objects and for a specific purpose.

unstructured data Data that do not have a predetermined format and are stored in the form of textual documents.

variable selection *See* dimensional reduction.

variance A descriptive statistics measure for dispersion. It is the square of standard deviation.

virtual personal assistant (VPA) An application program that understands natural language voice commands and completes tasks on behalf of its human user.

virtual reality The use of computer technology to create simulated environments.

Virtualization A process that allows for more efficient utilization of physical computer hardware and is the foundation of cloud computing.

visual analytics An extension of data/information visualization that includes not only descriptive but also predictive analytics.

visual interactive modeling (VIM) A visual model representation technique that allows for user and other system interactions.

visual interactive simulation (VIS) A visual/animated simulation environment that allows for the end user to interact with the model parameters while the mode is running.

Visual programming Visual programming is a programming language that lets humans describe processes and workflows using graphical illustration (e.g., nodes) as opposed to syntax-driven textual codes/descriptions.

Visual Studio Code Visual Studio Code, also commonly referred to as VS Code, is a source-code editor made by Microsoft with the Electron Framework, for Windows, Linux and macOS.

voice of the customer (VOC) Applications that focus on “who and how” questions by gathering and reporting direct feedback from site visitors, by benchmarking against other sites and offline channels, and by supporting predictive modeling of future visitor behavior.

Watson Analytics Watson Analytics, developed by IBM, is a cloud application that tries to bring sophisticated data discovery and predictive analytics to a business user.

Web 3.0 The next evolution (the third generations) of the World Wide Web where the user can own and control the use of his/her content on the internet.

Web analytics The application of business analytics activities to Web-based processes, including e-commerce.

Web content mining The extraction of useful information from Web pages.

Web crawler Also known as *spider* is an application used to swift/crawl/read through the content of a Web sites automatically.

Web mining The discovery and analysis of interesting and useful information from the Web, about the Web, and usually through Web-based tools.

Web service An architecture that enables assembly of distributed applications from software services and ties them together.

Web structure mining The development of useful information from the links included in Web documents.

Web usage mining The extraction of useful information from the data being generated through Web page visits, transactions, and so on.

Weka A popular, free-of-charge, open source suite of machine-learning software written in Java, developed at the University of Waikato.

what-if analysis It is an experimental process that helps determine what will happen to the solution/output if an input variable, an assumption, or a parameter value is changed.

word embeddings A type of word representation that allows words with similar meaning to have a similar representation.

word2vec A two-layer neural net that processes text by “vectorizing” words.

WordNet A popular general-purpose lexicon created at Princeton University.

Zion China A startup company that provides life management solutions for diabetics.

Note: 'f' and 't' refer to figures and tables respectively.

A

- AaaS. *See* Analytics-as-a-service (AaaS)
- AARP, Inc., 212–213
- Abandonment rates, 354
- Absolutist approach, 198
- Academic applications, 317
- Accessibility, 123, 358
- Accidental falls, 43
- Accounting, 451t
- Accuracy, 123, 273, 274t, 362
- Accuracy rate, 272
- Acquisitions, 211
- Acronyms, 28
- Activation function, 376, 382
- Active tags, 607
- Axiom, 292
- AdaBoosting, 277
- Adidas, robotics, 93
- Advanced analytical methods
 - and evaluation, 39
- Advanced analytics, 222–223, 361
- Affective computing, 326
- Affinity analysis, 282
- Agent-based models, 477
- Agglomerative classes, 281
- Aggregated information, 190
- Aggregation, 129, 334
- Agility, 219–220
- Agricultural applications, 590
- Agrobot, 98–99
- AI. *See* Artificial intelligence (AI)
- Airlines, 262–263
- AISERA, 108
- Alexa (Amazon), 109–110
 - defined, 110
 - Echo, 27, 109
 - skills, 110
- AlexNet, CNN, 405, 405f, 406
- Algorithms
 - AI, 80
 - Apriori algorithm, 266, 284–285, 285f
 - association rule mining, 260, 283, 284
 - backpropagation, 386–387, 412
 - classification modeling, 276–277
 - clustering algorithms, 280–282
 - data mining, 266, 294–295
 - decision trees, 278–280, 472
 - genetic, 277
 - in-database processing technology, 222
 - k-means, 282
 - machine learning algorithms, 126
 - optimization via, 449t
 - predictive, 122, 126
 - search engines, 349
 - sentiment classifiers, 336
 - SGD, 386
 - See also Specific algorithms*
- Alhea, 348
- All England Lawn Tennis Club (AELTC), 327–330
- Alpha diversity, 620
- Alternative data, 596–598
- Alteryx, 553–555
- Amazon (Amazon.com), 36, 45, 48, 154, 578, 580, 652
 - AI, 79, 86
 - Alexa. *See* Alexa (Amazon)
 - cloud computing, 578
 - Elastic Beanstalk, 584–585
 - IaaS, 581
 - QuickSight, 586
 - Smart Assistant Shopping Bots, 111
 - WorkSpaces, 580
- Ambari, 147
- American Cancer Society, 266
- Analysis, 31, 42
- Analysis ToolPak tool, 164
- Analytical decision modeling
 - with decision tables/trees, 471–473
 - goals/goal seeking, 467–468, 470
 - mathematical models, 450–452
 - mathematical programming optimization, 458–466
 - model-based, 444–450
 - sensitivity analysis, 68
 - with spreadsheets, 453–456
 - what-if analysis, 469
- Analytical layer, PANDA platform, 38, 38f
- Analytical processing, 31, 190
- Analytical support, 19
- Analytics, 19, 25, 31, 34
 - advanced, 361
 - applications, selected domains, 43–57
 - Big Data, 27, 140–143
 - (*See also* Big Data analytics)
 - business. *See* Business analytics, statistical modeling for
 - cognitive, 428
 - data science, machine learning and AI, 42
 - decision/normative, 41
 - department, 651
 - descriptive (or reporting), 35, 158
 - (*See also* Descriptive analytics)
 - developments, 18–19
 - evolution of computerized decision support, 25–28, 25f
 - evolving needs for, 17–19
 - healthcare domain, 43–45
 - image, 596–598
 - in-memory, 142
 - levels of, 34, 35f
 - location-based. *See* Location-based analytics
 - management, impact on, 654–655
 - organizational redesign, 654
 - overview, 34–42
 - predictive, 35–36, 126
 - (*See also* Predictive analytics)
 - prescriptive, 36, 41.
 - (*See also* Prescriptive analytics)
 - ready, 122
 - retail value chain, 45–48, 46f, 47t
 - sports, 174
 - stream. *See* Stream analytics
 - and text mining, 305–308, 305f
 - traffic congestions, 398–400
 - types, 35f
 - video, 77
 - visual. *See* Visual analytics
 - web technologies, 350–351
 - See also* Business analytics
- Analytics-as-a-service (AaaS), 585–586
- Analytics in action
 - AARP, Inc., 212–213
 - administrative/office jobs, robots, 658–659
 - animals' eating behavior and weight, 37–41
 - animals' gut health and performance, 618–620
 - cancer research, 266–267
 - Canopeo, 597–598
 - Cary, North Carolina, 167–168
 - cloud infrastructure applications, 587–589
 - cloud-mobile technology, 582–584
 - customer experience, financial services, 192–193
 - deliver innovation, customer sentiments, 310–312
 - disease patterns, 566–567
 - elevators, 22
 - emergency departments, 423–425
 - emotional support robots, 89–91
 - EY, Deloitte, and PwC, 83
 - face recognition, 407–408
 - gaming companies, 379–380
 - Georgia DOT variable speed limit, 398–400
 - GO-JEK, 587
 - Gulfstream Park Casino, 49–51
 - Hollywood movies, 287–290
 - home loan, 591–592
 - KNIME and R, 534–535
 - lies, 314–316
 - monitoring quarantined persons, COVID-19, 617
 - NCAA Bowl Game outcomes, 174–179
 - network analytics, predictive modeling, 621–622
 - 1-800-Flowers.com, 652–653
 - research literature survey, 323–325
 - Rockwell Automation, 605–606
 - Silvaris Corporation, 36
 - social media campaigns, 240–241, 359–360
 - specialty steel bar company
 - and available-to-promise dates, 41
 - Starbucks, 592–593
 - student attrition, 130–136
 - Target, 293
 - traffic accidents, 393–395
 - Unilever, 598

- Walgreens, 617–618
 - WeChat, 107
 - Wimbledon, 327–330
 - Analytics Leadership Award, 44
 - Analytics* magazine, 444
 - Analytics tools
 - commercial, 552–570
 - importance of, 497–504
 - multidimensional categorization, 498–500
 - programming languages, 505–527
 - visual tools, 527–552
 - Android, 89, 90
 - ANN. *See* Artificial neural network (ANN)
 - Antecedent, 284
 - AOL Search, 348
 - Apache Hadoop, 148–149. *See also* Hadoop
 - Apache Software Foundation (ASF), 148
 - Apple, 643
 - CarPlay, 644
 - Siri, 110, 415, 427
 - Appliances, 142
 - Application cases
 - Canadian Football League (CFL), 445–446
 - Metro Meals on Wheels Treasure Valley, 455–456
 - Pennsylvania Adoption Exchange (PAE), 454–455
 - refinery modeling, 447–448
 - steel tubing manufacturer, 474
 - University of Tennessee Medical Center, 458–459
 - Application programming interface (API), 422–423
 - Applications of AI
 - in accounting, 83–84, 83
 - in financial services, 84–86
 - in HRM, 85–86
 - in marketing, advertising, and CRM, 86–87
 - Applied mathematics, 42
 - Appraisal extraction, 326
 - Apriori algorithm, 266, 284–285, 285f
 - AR. *See* Augmented reality (AR)
 - ArcGIS, 590, 592
 - Architecture file, 421
 - Archival public records, 256
 - Area under the ROC curve, 275–276, 276f
 - ARIMA, 180
 - Arithmetic mean, 158–159
 - Artificial intelligence (AI), 27, 29, 368, 575, 642, 657
 - applications. *See* Applications of AI
 - benefits, 68–69
 - business analytics and, 635
 - capabilities, 70, 73
 - characteristics, 66
 - vs.* cognitive computing, 427–429, 427t
 - dangers of, 662–664
 - decision-making process, 79–82
 - definitions, 66
 - drivers, 68–70
 - Dystopia, 662–663
 - elements, 66
 - ethical/responsible/trustworthy, 646–651
 - examples, 66–67, 69
 - fifth-generation (5G) technologies, 614–616
 - functionalities and applications, 66–67, 67f
 - goals, 67–68
 - human intelligence, 72, 73t
 - innovation and, 19
 - limitations, 69–70
 - measuring, 72–73
 - ML and biostatistical methods, 619
 - Schrage's models, 82
 - technologies. *See* Technologies of AI
 - Turing Test, 72, 74f
 - Utopia, 663–664
 - vignette, 64–65
 - Artificial neural network (ANN), 368
 - backpropagation, 386–388, 387f
 - black box of, 389–395
 - elements of, 381
 - overfitting, 388, 388f
 - supervised learning, 385, 386f
 - transfer function, 382, 383f
 - See also* Neural networks
 - Artificial neuron
 - multiple-input, 378f
 - single-input, 376f
 - Ask, 348
 - Assisted intelligence, 70
 - Association/rule learning, 257, 260–261, 322
 - Association rule mining, 266, 282–285
 - Assumed risk, 453
 - Astronomy, 255
 - Atlas, 592
 - Atos, 240, 359–360
 - @RISK, 478
 - Attributes, 278
 - Augmented intelligence, 70–71
 - Augmented reality (AR), 79, 594
 - Authoritative pages, 342
 - Authority, 342
 - Automated data collection systems, 121
 - Automated decision-making, 81
 - Automatic sensitivity analysis, 468
 - Automatic summarization, 312
 - Automation, 657
 - defined, 88
 - See also* Robotics
 - Autonomous AI, 70
 - Autonomous car, 101
 - Autonomous robots, 89
 - Autonomous vehicles
 - computer centers in cars, 100
 - deep learning, 100
 - defined, 613
 - development, 100–101
 - implementation issues in, 613
 - maps, 100
 - mobile phones, 100
 - self-driving cars, 101–102
 - wireless Internet, 100
 - Autonomy, 88
 - Available-to-promise (ATP) decisions, 41
 - Average, 158–159
 - Average page views per visitor, 351
 - Average pooling function, 403
 - Averaging methods, 180
 - Avro system, 147
- B**
- Back-office business analytics, 4
 - Backpropagation (back-error propagation), 386–388, 387f
 - Bagging-type decision tree ensembles, 277
 - Bag-of-words model, 308
 - Balanced scorecard-type reports, 225
 - BAM. *See* Business activity management (BAM)
 - Banking industry, 262
 - Banking services, 283
 - AI in, 84–85
 - association rule mining, 283
 - data mining, 262
 - Bar charts, 228
 - Barcode, 607
 - Bayes/Bayesian classifiers, 276
 - Beane, Billy, 3
 - Bernoulli trial, 173
 - Bertin, Jacques, 226
 - Best-of-breed components, 34
 - Beta diversity, 620
 - Beyond the brand, 362
 - Bias, 128
 - BI Competency Center, 32
 - Big Data analytics, 27, 117, 121, 136, 141f, 191, 569–570
 - business problems addressed by, 143
 - challenges, 142–143
 - conceptual architecture for, 140f
 - critical success factors, 141–142, 141f
 - definition of, 136–139
 - fundamentals of, 140–143
 - Hadoop. *See* Hadoop
 - high-level conceptual architecture, 140f
 - high-performance computing, 142
 - management of, 19
 - MapReduce, 145–146, 145f
 - need for, 140–141
 - NoSQL, 151
 - and stream analytics, 153–156
 - technologies, 143–152
 - value proposition, 139, 140
 - variability, 139
 - variety, 138
 - velocity, 138–139
 - veracity, 139
 - volume, 137
 - Bing, 348
 - Bioinformatics, 618–620
 - Biomedical applications, 316–317
 - BI systems. *See* Business intelligence (BI) systems
 - Black-box syndrome, 390
 - Black-hat SEO, 348, 349
 - Black holes, 31
 - Blenderbot, 624–625
 - BlenderBot 3.0, 647–649
 - BlueCava technology, 643
 - BM SPSS Modeler, 288, 289f
 - Bolivian chatbot (BO.T), 105
 - Boosting ensemble method, 277
 - Boosting-type decision tree ensembles, 277
 - Bootstrapping, 275
 - Bot. *See* Chatbots (Chat robot)
 - Bottom-up approach, 284–285
 - Box-and-whiskers plot, 161–162, 162f, 167f
 - Box-office prediction system, 288, 289f
 - Box plot, 161–162
 - BPM. *See* Business performance management (BPM)
 - Branch, 278
 - Brand management, 331
 - Break-even point, goal seeking, 470
 - Bridge, 357
 - Brokerages, 262
 - Brokers and traders, data mining, 262

- Brontobytes (BB), 138
 - Broussard, Bruce, 45
 - Browser-native technologies, 227
 - BSI Videos (Business Scenario Investigations), 30
 - Bubble charts, 229
 - Building the Data Warehouse* (Inmon), 191
 - Bullet graphs, 230
 - Burt, Ronald, 357
 - Bus architecture. *See* Data mart bus architecture
 - Business activity management (BAM), 33
 - Business analytics, 29
 - back-office business analytics, 4
 - and business intelligence, 188, 189, 189f
 - cloud computing. *See* Cloud computing/cloud-based analytics
 - front-office business analytics, 4
 - statistical modeling, 157–168
 - See also* Analytics
 - Business analytics, statistical modeling for, 157–158
 - analytics in action, 167–168
 - arithmetic mean, 158–159
 - box-and-whiskers plot/box plot, 161–162, 162f, 167f
 - charts and graphs, 228–231, 231f, 232f
 - descriptive statistics, 158
 - kurtosis, 164
 - mean absolute deviation, 161
 - measures of centrality, 158
 - measures of dispersion, 160
 - median, 159
 - mode, 159–160
 - quartiles, 161
 - range, 160
 - shape of distribution, 162–164, 163f
 - skewness, 163–164
 - standard deviation, 161
 - variance, 160
 - Business data warehouse, 191
 - Business intelligence (BI) systems, 25–28, 157, 188
 - AARP, Inc., 212–213
 - acquisition of, 33
 - architecture of, 29, 30f
 - and business analytics, 188, 189, 189f
 - business strategy, alignment with, 31–32
 - cost-benefit analysis, 33
 - and data visualization tools, 569
 - definition, 28
 - department, 651
 - development of, 33
 - drivers of, 29–30
 - evolution of, 25f, 29f
 - framework for, 28–34
 - high-level architecture, 30f
 - history of, 28–29
 - integration of systems and applications, 33–34
 - justification, 33
 - multimedia exercise in, 30–31
 - origins and drivers of, 29–30
 - planning and alignment, 31–32
 - privacy, 33
 - real-time, on-demand BI, 32–33
 - real-time BI applications, 32
 - security, 33
 - tax fraud, targeting, 186–188
 - transaction processing *vs.* analytic processing, 31
 - Business need, 141
 - Business objective, 265
 - Business Objects, 191, 197, 205t
 - Business performance management (BPM), 28, 224
 - Business process management, 224
 - Business reporting, 223–225
 - balanced scorecard, 225
 - dashboard, 224
 - functions, 223
 - in managerial decision making, 224f
 - metric management reports, 224
 - Business reports, 223
 - Business rules, 215
 - Business Scenario Investigations (BSI) videos, 30
 - Business strategy, 31–32
 - Business-user comments, 243
 - Buzzwords, 28
- C**
- C4.5, 279
 - C5, 279
 - Caffe/Caffe2 (learning framework), 421–422
 - Calculated risk, 453
 - Calculation rules, 215
 - California Institute of Technology (Caltech), 190–191
 - Call detail records (CDR), 155
 - Cancer research, 266–267
 - Candidate generation method, 284–285
 - Candybar, 594
 - Canopeo, 597–598
 - Capacities, LP model, 459
 - Cary, North Carolina, 167–168
 - Cascaded decision tree model, 6, 7f
 - Case-based reasoning, 276
 - Catastrophic loss, 472
 - Categorical data, 125, 260, 265
 - Categorical representation, 126
 - Categorization, 307, 321
 - CDR. *See* Call detail records (CDR)
 - Center for Health Systems Innovation, 566
 - Centers for Disease Control and Prevention (CDC), 43
 - Central fact tables, 208
 - Central Intelligence Agency (CIA), 313–314
 - Centrality, 162, 357
 - Centralized data warehouse, 200f, 201, 203
 - Central processing unit (CPU), 395, 421–422
 - Centroid, 282
 - CEP. *See* Complex event processing (CEP)
 - Cerner Corporation, 566
 - Certainty, decision making, 452–453
 - The Championships, 327–330
 - Change capture, 211
 - Changing business environments, 17–19
 - Channel analysis, 47t
 - Charts
 - bar charts, 228
 - basic charts, 228–229
 - bubble charts, 229
 - choice of, 230–231
 - Gantt charts, 229
 - line charts, 228
 - PERT charts, 229
 - pie charts, 228
 - specialized charts, 229–230
 - taxonomy, 230, 231f
 - See also* Graphs
 - Chat-based Generative Pretrained Transformer (ChatGPT), 416, 624
 - applications, 418–419
 - biases, 419
 - computer programming, 419
 - customer service, 418
 - false answers, 419–420, 420f
 - four-step process, 416f, 417–418
 - healthcare, 419
 - limitations, 419–420
 - marketing, 419
 - obsolete data, 420
 - Chatbots (Chat robot), 64, 86, 102
 - benefits, 105
 - chatting with, 104f
 - components and use, 104
 - defined, 103
 - drivers of, 103–104
 - enterprise. *See* Enterprise chatbots
 - evolution, 103–104
 - person-machine interaction process, 104, 104f
 - representative, 105
 - revolution, 102
 - smart assistant shopping bots, 111
 - types of, 103
 - ChatGPT. *See* Chat-based Generative Pretrained Transformer (ChatGPT)
 - Chime, 588
 - Chi-squared automatic interaction detector (CHAID), 279
 - Choice phase of decision-making, 20, 24
 - Chromosomes, 482
 - CIO Insight*, 59
 - Citibank, 191
 - Citizen data scientists, 664–665
 - Citizen science, 664
 - Classification, 321, 448
 - accuracy of, 273–280, 273f
 - area under the ROC curve, 275
 - Bayesian classifiers, 276
 - bootstrapping, 275
 - case-based reasoning, 276
 - common accuracy metrics, 274t
 - in data mining, 258–260, 272–273
 - decision tree analysis, 276
 - genetic algorithms, 258
 - jackknifing, 275
 - k*-fold cross-validation, 275
 - leave-one-out, 275
 - matrix, 273
 - neural networks, 276
 - N–P polarity, 332–333
 - problem, 23, 273
 - rough sets, 258
 - simple split, 274–275, 274f
 - statistical analysis, 276
 - techniques, 276–277
 - text mining, 321
 - Classification and regression trees (CART), 279, 287t
 - Classification tools, 258–259
 - Class label, 278
 - Class/response variable, 173
 - Click map, 352
 - Click paths, 352
 - Clickstream analysis, 349–350

- Client/server architecture, 197
 - Cliques and social circles, 357
 - Cloaking technique, 349
 - Closed-circuit television (CCTV), 46
 - Cloud-based systems, 18
 - Cloud-based technologies, 18
 - Cloud computing/cloud-based analytics, 218, 575, 576, 578
 - analytics-as-a-service (AaaS), 585–586
 - analytics in action, 582–583
 - cloud deployment models, 584
 - cloud infrastructure application, 587–589
 - cloud oriented support system, 579f
 - cloud platform providers, 584–585
 - data-as-a-service (DaaS), 579–580
 - defined, 576
 - essential technologies, 581–584
 - illustrative analytics applications, 587–589
 - infrastructure-as-a-service (IaaS), 578, 580–581
 - major cloud providers, 584–585
 - platform-as-a-service (PaaS), 578, 580
 - representative analytics, 586
 - software-as-a-service (SaaS), 580
 - support system, 579f
 - technologies, 581–582, 581f
 - virtualization, 581–582
 - Cloud deployment models, 584
 - Cloudera, 148
 - Cloud Security Alliance, 588
 - Clusters/clustering, 258, 260, 280, 321–322, 324, 448
 - algorithms, 35, 281
 - analysis, 280–282
 - coefficient, 357
 - data mining, 260, 280–282
 - k*-means algorithm, 282, 282f
 - optimal number, 281
 - query-specific, 322
 - scatter/gather, 322
 - text mining, 321–322
 - CNN. *See* Convolutional neural networks (CNN)
 - Cobots. *See* Collaborative robots (Cobots)
 - Cognitive analytics, 428–429
 - Cognitive computing system, 79, 368–369, 425–434
 - vs.* AI, 427–429, 427f
 - attributes, 426–427
 - cognitive search, 429, 430f
 - framework, 426f
 - IBM Watson, 429–434
 - work process, 426–427
 - Cognitive limits, 19
 - Cognitive search method, 429, 430f
 - Cognos, 191
 - Cohesion, 357
 - Collaboration process, 18
 - Collaborative filtering, 380
 - Collaborative robots (Cobots), 94, 99
 - Collection, 334
 - Columnar database, 220–221
 - Column-oriented database management system, 220–221
 - Communication networks, 355
 - Community networks, 356
 - Comorbidity networks, 568f, 620
 - Comparative measures, 241
 - COMPAS. *See* Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)
 - Competitive advantage, 261
 - Complex event processing (CEP), 616
 - Comprehensive database, 197
 - Comprehensiveness, 123
 - Computer Associates, 205t
 - Computer-based information system (CBIS), 383, 633, 651, 654
 - Computer ethics, 350
 - Computer hardware and software, 262. *See also* Hardware; Software
 - Computerized decision support framework, 20–24
 - analytics/data science, 25–28, 25f
 - Computer science, 42
 - Computer vision, 76–77
 - Compute Unified Device Architecture (CUDA), 421
 - Concept hierarchies, 129
 - Concept linking, 307
 - Concepts, 307
 - Condition-based maintenance, 262
 - Confidence, 284
 - Confidence gap, 478
 - Confidence metric, 284
 - Confusion matrix, 273, 273f
 - Congressional Floor-Debate Transcripts, 336
 - Connections, 356
 - Connection weights, 382
 - Consequent, 284
 - Consistent data, 123
 - Constant Error Carousel (CEC), 414
 - Constant variance (of errors), 173
 - Constraints, 318, 451, 458, 459
 - Consumer-centric apps, 595
 - Contentenka, 348
 - Content groupings, 353
 - Contextual metadata, 242–243
 - Contingency table, 273
 - Continuous data, 126
 - Continuous distributions, 477, 477t
 - Conversion statistics, 353–354
 - Convolutional neural networks (CNN), 400–411, 619
 - Caffe/Caffe2, 421–422
 - for extracting features, 402f
 - face recognition technology, 407–408
 - function, 401–403
 - image processing, 404–408
 - input matrix, 402, 402f
 - pooling layer, 401, 403–404
 - for relation extraction, 410f
 - text processing, 408–411
 - unit, 401f
 - Convolution function, 401–403, 402f
 - Convolution kernel, 401, 402f
 - Convolution layer, 401
 - Cornell Movie-Review Data Sets, 336
 - Coronavirus Worldwide Data, 52
 - Corporate cloud, 584
 - Corporate information factory. *See* Hub-and-spoke architecture
 - Corpus, 307, 319
 - Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), 647–648
 - Correlation *vs.* regression, 169
 - Cost-benefit analysis, 33, 41
 - Cost reduction, 143
 - COVID-19 pandemic, 18, 28, 617
 - analytics applications, 52–53
 - mask mandates in Oklahoma, 53–57, 57t
 - COVID Data Tracker, 52
 - Credibility assessment, 314. *See also* Deception detection
 - Credit card transactions, 283
 - Crime analysis, 590
 - Criminal networks, 356
 - CRISP-DM (Cross-Industry Standard Process for Data Mining), 264, 264f, 318
 - CRM. *See* Customer relationship management (CRM)
 - Cross-Industry Standard Process for Data Mining (CRISP-DM), 264, 264f
 - business understanding, 265
 - data preparation, 266
 - data understanding, 265–266
 - deployment, 269
 - model building, 266
 - standardized methodologies, 269–270, 270f, 271f
 - testing and evaluation, 269
 - Cross-linking, 347
 - Cube, 209
 - Currency, 124
 - Customer buying patterns, 293
 - Customer churn analysis, 47t
 - Customer experience, 143
 - Customer relationship management (CRM), 4, 31, 261–262, 310, 313
 - AI in, 86
 - customer experiences and, 86
 - data mining, 261–262
 - Cutting, Doug, 144
 - Cybersecurity, 155
- ## D
- DaaS. *See* Data-as-a-service (DaaS)
 - Dashboards, 18, 29, 238–243, 239f
 - analytics in action, 240–241
 - benchmarks, 242
 - best practices, 242
 - business-user comments, 243
 - characteristics, 242
 - contextual metadata, 242–243
 - design, 241–243
 - guided analytics, 243
 - information level, 243
 - KPI, 242
 - management, 241
 - metrics, 242–253
 - monitoring, 241
 - presentation of information, 243
 - prioritization of alerts/exceptions, 243
 - ranking of alerts/exceptions, 243
 - usability specialist, 243
 - user comments, 243
 - validation methods, 243
 - visual construct, 243
 - Web analytics, 354f
 - Dashboard Spy Web, 241
 - Dashboard-type reports, 224
 - Data, 25, 125
 - analytics in action, 130–136
 - Big Data. *See* Big Data analytics
 - data-related tasks, 123
 - data size, 138
 - data to knowledge continuum, 121, 122f
 - dirty data, 127

Data (*Continued*)

- identification and selection, 265
- management, 19
- nature of, 121–124
- preprocessing, 127–129, 128f, 130t, 266
- quality, 121
- readiness level of, 123–124
- reduction, 129
- representation schemas, 127
- science, 25–28, 42
- scientists, 42, 145
- scrubbing, 127
- security, 123
- shape of distribution, 162–168
- storytelling with, 235–236
- taxonomy of, 125–127, 125f
- transportation of, 215
- types of, 125–126
- value proposition, 122
- variable types, 127
- See also Specific types of data*
- Data access, 211
- Data accessibility, 123
- Data acquisition, PANDA platform, 38, 38f
- Data acquisition (back-end) software, 197
- Data Advantage Group, 205t
- Data analysis, 199
- Data analyst, 42
- Data archaeology, 256. *See also* Data mining
- Data-as-a-service (DaaS), 579–580
- Database management system (DBMS), 199, 222
- Data blending, 127
- Data cleansing, 127–128, 130t
- Data concentration, 158
- Data consistency, 123
- Data consolidation, 130t
- Data content accuracy, 123
- Data currency/data timeliness, 124
- Data dredging, 256.
 - See also* Data mining
- Data-driven customer experience, financial services, 192–193
- Data-driven decision making, 196f
- Data-driven marketing, 118–121
- Data engineering challenges, 37–38
- Data extraction, 196
- Data federation, 211
- Data for good projects, 151–152, 152t
- Data governance, 142
- Data granularity, 124
- Data infrastructure, 142
- Data-in-motion analytics.
 - See* Stream analytics
- Data integration, 142, 211–214
- Data lakes, 218–220
- Data loading, 197
- Data management, 456
 - improved data management, and decisions, 19
 - technologies and practices, 221–222
- Data mart (DM) approach, 194, 204, 206t, 207t
- Data mart bus architecture, 200–201, 200f, 202
- Data migration, 199, 215
- Data mining, 18, 27, 35, 251, 256, 258
 - accuracy metrics, 273–280, 274t
 - applications, 261–263
 - association, 260–261
 - benefits, 256–257
 - blend of multiple disciplines, 257f
 - in cancer research, 266–268
 - categories, 258
 - characteristics and objectives, 256–257
 - classification, 258–260, 272–273
 - clustering, 260, 280–282
 - concepts, 254–261
 - CRISP-DM. *See* Cross-Industry Standard Process for Data Mining (CRISP-DM)
 - defined, 256, 306
 - ideas behind, 255
 - methods, 272–285
 - of multiple disciplines, 257f
 - myths and blunders, 293–295, 294t
 - objectives, 256–257
 - other names associated with, 256
 - patterns, 257–258
 - predictions, 258, 287–289, 293
 - privacy issues, 292
 - process, 264–271, 264f, 270f, 271f
 - software tools, 286–291, 287t, 291f
 - vs.* statistics, 261
 - tasks, categories of, 258–260
 - taxonomy for tasks, methods, and algorithms, 259f
 - in text analytics, 305f
 - vs.* text mining, 305–307
 - time-series forecasting, 261
 - use of term, 255
 - value proposition, 294
 - vignette, 251–254
 - visualization, 261
- Data mining methods, 272–285
 - association rule mining, 282–285
 - classification, 258–260
 - cluster analysis, 280–282
 - decision trees. *See* Decision trees
 - ensemble models, 277–278
- Data mining process, 264–271
 - business understanding, 265
 - CRISP-DM, 264, 264f
 - data preparation, 266
 - data understanding, 265–266
 - deployment, 269
 - evaluation, 269
 - model building, 266–268
 - other standardized processes and methodologies, 269–271
 - ranking of processes and methodologies, 271f
 - testing, 269
- DataMirror, 205t
- Data-oriented cloud systems, 588
- Data-oriented DSS, 30–31
- Data preparation, 266, 294. *See also* Data preprocessing
- Data preprocessing, 127, 266, 423, 423f
 - art and science of, 127–136
 - essence of, 129, 130t
 - purpose of, 266
 - steps, 128f
 - value proposition of, 129
- Data privacy, 123
- Data quality, 121–122, 215
- Data reduction, 129, 130t
- Data relevancy, 124
- Data retrieval, 199
- Data richness, 123
- Data science, 31, 446
 - vs.* analytics, 42
 - challenges, 39
 - department, 651
- Data Science Central, 59
- Data scientists, 145
- Data scrubbing, 127
- Data security, 123
- Data sources, 27, 196, 215
 - for business applications, 265
 - reliability, 123
- Data stream analytics, 139
- Data taxonomy, 125–127, 125f
- Data transformation, 128–129, 130t, 196
 - tools, 215
- Data validity, 124
- Data visualization, 225
 - in BI and analytics, 232, 233f, 569
- dashboards. *See* Dashboards
- data mining and, 261
- future of, 227
- history of, 225–227, 226f, 227f
- storytelling, 235–236
- visual tools, need for, 227
- Data volume, 142
- Data warehouse administrator (DWA), 217
- The Data Warehouse Toolkit* (Kimball), 191
- Data warehouse/warehousing (DW), 18, 19, 29, 31, 185, 189
 - administration, 217
 - all-in-one solutions, 221
 - appliance, 221
 - architectures, 197–203, 197f, 198f, 200f
 - business value of, 206
 - characteristics of, 193–194
 - centralized data warehouse, 200f, 201, 203
 - client/server architecture, 194
 - as component of BI system, 29
 - data analysis, 209
 - data-driven decision making, 196f
 - data from, 26
 - vs.* data lake, 219t
 - data mart (DM) approach, 194, 204
 - data migration tools, 199
 - development, 203–216
 - direct benefits, 203
 - DW-driven DSSs, 26–27
 - DW model of traditional BI systems, 33
 - enterprise data warehouse (EDW), 195, 199, 204, 206t, 207t
 - federated data warehouse, 201
 - future of, 217–223
 - giant data warehouses, 19,
 - high-level depiction, 196f
 - historical perspective, 190–192, 190f
 - hosted data warehouse, 207–208
 - hub-and-spoke architecture, 199, 200f, 201
 - indirect benefits, 203–204
 - infrastructure, 220–222
 - integration, 194
 - management of giant data warehouses, 19
 - metadata, 194
 - migration of data, 215
 - modern approaches, origins of, 29–30
 - multidimensional structure, 194
 - nonvolatile, 194
 - privacy, 217
 - process, 195–197
 - real-time data warehousing (RDW), 27, 194, 221
 - relational structure, 194

- representation of data, 208–209
 - right-time, 27
 - security issues, 217
 - sourcing, 217–218
 - subject orientation, 193
 - tax fraud, targeting, 186–188
 - time variant (time series), 194
 - vendors, 205t
 - Web-based applications, 194
 - wide variety of data, 31
 - The Data Warehousing Institute, 33, 59, 204
 - Datum, 125. *See also* Data
 - Davenport, Thomas H., 255
 - Da Vinci Surgical System, robotics, 98
 - DBN. *See* Deep belief network (DBN)
 - DB2, 199
 - Decentralized autonomous organizations (DAOs), 623
 - Deception detection, 314–316, 315f, 315t
 - Decision analysis, 41
 - decision tables, 471–472
 - decision trees, 472–473
 - defined, 471
 - Decision-making process, 18
 - AI support for, 79–81
 - automated, 81
 - certainty, 452–453
 - decision modeling with spreadsheets, 453–457
 - example, 22
 - forecasting, 447
 - intelligent technologies, support of, 655
 - model-based. *See* Model-based decision-making
 - phases of, 20
 - probabilistic decision-making situation, 453
 - risk, 453
 - stochastic decision-making situation, 453
 - uncertainty, 453, 472
 - zones, 452f
 - Decision modeling with spreadsheets, 453–457
 - Decision/normative analytics, 41
 - Decision support
 - developments, 18–19
 - evolution of computerized, 25–28, 25f
 - evolving needs for, 17–19
 - Decision support mathematical models
 - components of, 450–451, 451t
 - decision variables, 451
 - intermediate result variables, 451
 - profit model, 451
 - result (outcome) variables, 450
 - structure of, 450–452
 - uncontrollable variables, 451
 - Decision support systems (DSSs), 19, 25, 26
 - add-ins, 454
 - DW-driven DSSs, 26–27
 - Keen and Scott-Morton's definition, 25
 - knowledge-based modeling, 448
 - mathematical models for, 450–452
 - mathematical programming optimization, 458–466
 - resources and links, 59
 - with spreadsheets. *See* Spreadsheets
 - technologies for, 18–19
 - and visual interactive models, 479
 - Decision Support Systems* (journal), 59
 - Decision tables, 471–472, 471t
 - Decision tree analysis, 276
 - Decision tree models, 35
 - Decision trees, 258, 260, 266, 278–280, 472–473
 - algorithms, 278–279
 - bagging-type ensembles for, 277
 - boosting-type ensembles for, 277
 - cascaded decision tree model, 6, 7f
 - Decision variables, 450f, 451, 451t, 459, 471
 - Dedicated url, 352
 - Deep belief network (DBN), 396
 - Deep Blue (chess program), 430
 - Deep feedforward networks, 394f, 396
 - Deep knowledge, 303
 - Deep learning (DL) technology, 76
 - AI-based learning methods, 374f
 - computer frameworks. *See* Libraries (software)
 - overview, 373–375
 - vignette, 369–372
 - Deep Learning tools, 290
 - Deep neural networks, 375, 395–400
 - classification-type, 297f
 - deep feedforward networks, 394f, 396
 - hidden layers *vs.* neurons, 397
 - random weights in MLP, 396–397
 - DeepQA architecture, 303, 304f, 430
 - Defense, 262
 - Defense coordinator visual tool, 10, 10f
 - Deforestation (Amazon), 596
 - Delivery lengths, bowling, 13–14
 - Dell, 205t
 - Dell Statistica, 287t
 - Delta, 385–386
 - Demirkan, Haluk, 578
 - Demographic details, 590
 - Demos, 59
 - Denial of service (DDoS) attacks, 330
 - Density, 357
 - Department for Homeland Security, 313
 - Dependent data mart, 194
 - Dependent variables, 450
 - Deployment, 269
 - Deployment of intelligent systems, 634–637
 - adoption, 637
 - connectivity and integration, 635
 - implementation issues, 635
 - leveraging in business, 636
 - management and implementation, 634–635
 - security protection, 636
 - Descriptive analytics, 35, 42, 223, 189, 361
 - branches of, 157
 - business intelligence. *See* Business intelligence (BI) systems
 - data warehousing. *See* Data warehouse/warehousing (DW)
 - descriptive statistics, 158
 - OLAP. *See* Online analytical processing (OLAP)
 - statistical methods, 157
 - statistics, 157–158
 - Descriptive statistics, 158, 164
 - arithmetic mean, 158–159
 - box-and-whiskers plot, 161–162, 162f
 - for descriptive analytics, 158
 - interquartile range, 161
 - mean absolute deviation, 161
 - measures of centrality tendency, 158
 - measures of dispersion, 160
 - median, 159
 - Microsoft Excel, 164–167
 - mode, 159–160
 - quartiles, 161
 - ranges, 160
 - role in business analytics, 158
 - shape of a distribution, 162–168
 - standard deviation, 161, 163
 - variance, 160
 - Desktop as a Service (DaaS), 580
 - Design phase of decision-making, 20, 23–24
 - Devlin, Barry, 191
 - Dice, 210
 - Dictionary, 320
 - Difference-in-differences (DiD) model, 54, 56, 57, 57t
 - Digital twins, 594, 623
 - Dimensional modeling, 208
 - Dimensional reduction, 129
 - Dimension tables, 208
 - Direct searches, 352
 - Dirty data, 127
 - Discover Foods, 575–576
 - Discrete data, 125
 - Discrete distributions, 477, 477t
 - Discrete event simulation, 477, 478
 - Discretization, 129, 260
 - Discriminant analysis, 157, 258
 - Disease patterns, 566–567
 - Disease spread prediction, 590–591
 - Dispersion method, 160, 163f
 - Distance, 357
 - Distance measure, 281
 - Distant supervision method, 411
 - Distributed database management system, 191
 - Distributions, 357
 - Divisive classes, 281
 - DL. *See* Deep learning (DL) technology
 - DM approach. *See* Data mart (DM) approach
 - DNA microarray analysis, 316
 - Document hierarchy, 342
 - Document indexer, 345, 345f
 - Document matcher/ranker, 345f, 346–347
 - Dogpile, 348
 - Domain experts, 26
 - Domain of interest, 319
 - Downloads, 351–352
 - Drill down/up, 210
 - Driverless cars. *See* Autonomous vehicles
 - DSS. *See* Decision support systems (DSSs)
 - DuckDuckGo, 348
 - DW. *See* Data warehouse/warehousing (DW)
 - DWA. *See* Data warehouse administrator (DWA)
 - Dynamic advertisements, 332
 - Dynamic data, 126
 - Dynamic models, 448, 479
 - Dynamic networks, 412
 - Dynamic pricing, 5, 6f
 - Dystopia (pessimistic approach), 662
- E**
- EAI. *See* Enterprise application integration (EAI)
 - EarthCast, 596
 - eBay, 154

- ECHELON surveillance system, 313
 Echo, 109
 e-commerce, 154–155, 332, 349, 350
 Economining, 336
The Economist, 655–657
 ED. *See* Emergency department (ED)
 Edge computing, 614
 EDW. *See* Enterprise data warehouse (EDW)
 EEE (exposure, experience, and exploration) approach, 2
 EII. *See* Enterprise information integration (EII)
 EIS. *See* Executive information systems (EISs)
 Electronic Product Code (EPC), 607
 Elitism, 482
 Embarcadero Technologies, 205t
 EMC Greenplum, 237
 Emergency department (ED), 423–425
 End-user modeling tool, 454
 Energy industry, 154, 154f
 Ensemble models, 134–135, 277–278
 Enterprise application integration (EAI), 214
 Enterprise chatbots
 AISERA, 108
 Amazon's Alexa and Echo, 109–110
 analytics in action, 107
 Apple's Siri, 110
 Facebook's chatbots, 106
 financial services, 107
 Google Assistant, 110
 improving customer experience, 106
 industry-specific bots, 108
 inside enterprises, 108
 interest of, 106
 knowledge for, 109
 messaging services, 106
 platforms, 108
 as professional advisors (robo advisors), 111
 service industries, 107–108
 virtual personal assistants, 109
 See also Chatbots (Chat robot)
 Enterprise data warehouse (EDW), 195, 199, 204, 206t, 207t
 Enterprise information integration (EII), 214
 Enterprise resource planning (ERP), 26, 31
 Entertainment industry, data mining, 263
 Entropy, 280
 Environmental effects, 591
 Environmental scanning and analysis, 446
 Equivariance, 403
 EPC. *See* Electronic Product Code (EPC)
 ERP. *See* Enterprise resource planning (ERP)
 Errors, 180
 ES. *See* Expert system (ES)
 ESRI (esri.com), 590
 Ethical/responsible/trustworthy AI, 646–651
 Ethics, 644–646
 ETL. *See* Extraction, transformation, and load (ETL)
 Euclidian distance, 281
 EUROPOL, 313
 Evidence-based medicine, 308
 Evolutionary algorithms, 480. *See also* Genetic algorithms (GA)
 Exabytes (EB), 137, 138
 Excel. *See* Microsoft
 Executive champion, 141
 Executive dashboard, 238–239, 239f. *See also* Dashboards
 Executive information systems (EISs), 26, 28–29
 Executive's Guide to the Internet of Things, 637–638
 Exit rates, 354
 Experience, 2, 3
 Experimentation, 42
 Expert system (ES), 26, 103
 Explanatory variables, 170, 173
 Explicit sentiment, 327
 Exploratory data analysis, 39
 Explore, 2, 3
 Exponential smoothing, 180
 Exposure, 2, 3
 “The Extended ASP Model”, 218
 eXtensible Markup Language (XML)-based tools, 201, 214
 Extraction, transformation, and load (ETL), 223, 214–215, 215f
 Extract/transfer/load batch update, 33
 Extranet, 198–199
- F**
 Fabio (robot), 97
 Facebook, 147, 373, 643, 645
 Caffe2, 421
 chatbots, 106
 ethical issues, 645
 proponent, 663
 Rapleaf, 643
 weakly supervised training, 406
 Facial recognition technology, 77
 Fact-based decision-making culture, 142
 Falls, 43–44
 Feature engineering, 39
 Federated architecture, 200f, 203
 Federated data warehouse, 201
 Feedforward-multilayered perceptron (MLP), 381, 385
 random weights, 396–397
 -type deep networks, 394f, 396
 FICO Decision Management, 287t
 Fifth-generation (5G) technologies, 614–616
 Filter, 401
 Finance sectors, 255
 Financial investment, 451t
 Financial markets, sentiment analysis, 331
 Financial model, 451–452
 Financial robo advisors, 111
 Financial services, Big Data, 156
 Flume framework, 147
 Fool's gold, 295
 Forecasting, 169, 258, 446–448
 Foreign language reading, 312
 Foreign language writing, 312
 Forget/feedback gate, 414
 Forrester, 502
 Fourth Industrial Revolution, 88, 614
 Fraud
 fraud detection, 280
 tax fraud, 186–188
 Frequency, 358
 Frequency plot, 163
 Friendly AI, 664
 Friendship network, 567
 Frontline Systems Inc. (solver.com), 454
 Front-office business analytics, 4
 Functionality, 32
 Fuzzy logic, 281
- G**
 GA. *See* Genetic algorithms (GA)
 GA-EAIP. *See* Gaming Analytics Enterprise Intelligence Platform (GA-EAIP)
 Galton, Francis, 169
 Gaming Analytics Enterprise Intelligence Platform (GA-EAIP), 49–50, 50f
 Gaming Analytics platform, 49–51, 50f
 Gaming industry applications, 48
 Gantt charts, 229
 Gapminder, 231, 232f
 “Garbage in garbage out–GIGO” concept/principle, 117
 Gartner, Inc., 32, 138, 232, 355
 analytics tools, 501
 business intelligence platform, 233–234
 social analytics, 355
 Gartner Group, 28
 GDOT. *See* Georgia Department of Transportation (GDOT)
 Gegobytes (GeB), 138
 Gene/protein interaction identification, 316–317, 317f
 Generative models, 396
 Genetic algorithms (GA), 258, 277, 281, 480
 applications, 485
 chromosomes, 482
 crossover, 482
 elitism, 482
 knapsack problem, 484–485
 mutation, 482
 process, 483–484, 483f
 reproduction, 482
 Vector game, 480–482
 Genomic data, 255
 Geocoding, 591
 Geographic information systems (GIS), 230, 590, 592–593
 Geographic maps, 229–230
 Geography, 353
 Georgia Department of Transportation (GDOT)
 bottlenecks and traffic turbulence, 399–400, 399f
 highway speed variability, 398–399
 new insight on highway congestion, 400
 VSL pilot project, 398
 Geospatial analytics, 589–593
 applications for consumers, 595
 concept, 589–591, 593
 real-time location intelligence, 594
 See also Location-based analytics
 Geospatial data, 589
 Gephi, 567
 GhostMiner, 287t
 Giant data warehouses, 19
 Gigabytes (GB), 137, 138
 Gini index, 279
 GIS. *See* Geographic information systems (GIS)
 Global Fishing Watch (globalfishingwatch.org), 596
 Goal seeking, 456, 470, 470f
 GO-JEK, 587
 Google, 343, 348, 373, 595, 656
 Android Auto, 644
 cloud-based speech-to-text service, 415
 Google App engine, 585

- Google Assistant, 110, 642
- Google Cloud Platform, 587
- Google Home, 27, 427
- Google Lens, 406, 407f
- Google Maps, 227
- GoogLeNet, 405–406, 406f
- Google Now, 415, 429
- TPU, 422
- word2vec project, 408–410, 409t
- Google's Neural Machine Translation (GNMT) platform, 415, 416f
- Gopal, Vipin, 43
- Government, 156, 262
 - and defense, data mining, 262
 - intelligence, sentiment analysis, 331–332
- GPT-3, 624
- Gramm-Leach-Bliley privacy and safeguards rules, 217
- Grand View Research, 4
- Grant Thornton, 64–65
- Granularity, 124
- Graphical user interface (GUI), 198, 370, 395
- Graphs, 49
 - basic graphs, 228–229
 - bullet graphs, 230
 - choice of, 230–231
 - highlight tables, 230
 - histogram, 163, 229
 - scatter plots, 170f, 228–229
 - specialized graphs, 229–230
 - taxonomy, 230–231, 231f
- See also* Charts; Maps
- Green Card, 105
- Greenplum, 205t
- Grid computing, 142
- Grimes, Seth, 232
- Grindgis.com, 590
- Group collaboration, 18
- Group communication, 18
- Guessing, 258
- Guided analytics, 243
- H**
- Hadoop, 144–151
 - benefits, 149
 - clusters, 147–148
 - components, 146–147
 - data scientists, 145
 - defined, 144
 - HDFS, 144
 - misconceptions, 148–149
 - pros and cons, 147–148
 - Spark *vs.*, 149–151
 - technical components, 146–147
 - working principle, 144–145
- Hadoop/Big Data tools, 290, 291
- Hadoop Distributed File System (HDFS), 144, 146, 150
- Hardware
 - data mining used in, 262
 - IoT technology, 602
- Harte-Hanks, 205t
- Harvard Business Review*, 255
- Harvard Business Review Analytic Services, 360
- Hazie, chatbot, 105
- HBase database, 147
- HCatalog storage management, 147
- HDFS. *See* Hadoop Distributed File System (HDFS)
- Healthcare, 43–47, 263
- Health Insurance Portability and Accountability Act (HIPAA), 123, 217
- Health insurers, 44
- Health Sciences, Big Data, 156
- Health services, 156
- Heat maps, 6, 7f, 230
- Heterogeneous ensemble method, 277–278, 278f
- Heuristics, 26, 449t
- Hidden layer, 381–382, 387
 - vs.* neurons, 397
- Highlight tables, 230
- High-performance computing, 142, 236
- High-volume query access, 208
- Hilcorp Energy, 605–606
- HIPAA. *See* Health Insurance Portability and Accountability Act (HIPAA)
- Histogram, 163, 229
- Historical data, 194
- HITS. *See* Hyperlink-induced topic search (HITS)
- Hive framework, 147
- Holdout set, 274
- Hollywood movies, 287–289
- Homeland security, data mining, 263
- Homogeneous-type ensembles, 277
- Homonyms, 307
- Homophily, 356
- Homoscedasticity, 173
- Hortonworks, 147, 148
- Hospital systems, 156
- Hosted data warehouse, 207–208
- Hotels/resorts, 262–263
- HP, 205t
- Hub-and-spoke architecture, 199, 200f, 201, 203
- Hubs, 343
- Humana, Inc., 43–45
- Human-computer interaction (HCI), 427
- Human-mediated machine-learning approach, 373
- Human resource management (HRM), 85–86
- Humanyze Company, 643, 654
- Hummingbird Ltd., 205t
- Hybrid BAM-middleware providers, 33
- Hybrid cloud, 584
- Hyperion Solutions, 191, 205t
- Hyperlink-induced topic search (HITS), 343
- Hyperlinks, 342–343
- Hypothesis development, 39
- Hypothesis testing, 169
- I**
- IaaS. *See* Infrastructure-as-a-service (IaaS)
- IBM, 139, 191, 199, 327–330, 428, 555
 - on analytics, 652
 - cloud, 586
 - cognitive computing, 368–369
 - Deep Blue (chess program), 430
 - IBM® Bluemix®, 328
 - IBM InfoSphere BigInsights, 205t
 - IBM Ireland, 191
 - IBM SPSS Modeler, 287t, 288
 - IBM Watson, 287t, 302–304, 329, 429–431, 585, 656
 - education, 433
 - finance, 433
 - government, 433
 - healthcare and medicine, 432
 - MYCIN, 432
 - research, 434
 - retail, 433
 - security, 433
 - Watson Tone Analyzer, 431
- ID3, 126–127, 279
- if-then-else rules, 26
- Image analytics, 596–597
 - application, 597–598
 - satellite data, 596–597
- ImageNet data set, 404–405
- ImageNet Large Scale Visual Recognition Challenge (ILSVRC), 405
- Image processing technology, 76, 404–408
- Imagery data, 126
- Immediacy, 358
- Imperfect input, 309
- Implementation
 - defined, 24
 - phase of decision-making, 20, 24
- Implicit sentiment, 327
- Improved search precision, 322
- Improved search recall, 322
- Inception, 405–406, 406f
- Include terms, 320
- In-database analytics, 142, 222
- In-database processing technology, 222
- Independence, 259
- Independence (of errors), 172
- Independent data mart, 194, 199–200, 200f, 203
- Indiana University Kelly School of Business, 44
- Indices, 320
- Individual impacts, 202
- Industrial restructuring, 655–656
- Industrial Revolution, 651, 656
- Industry standards, 242
- Inferences, 158
- Inferential statistics, 158, 164, 169–180
- Influence diagram, 450
- Influencers, 362
- Info, 348
- Informatica, 205t
- Information, 223
 - to decision makers, 223
 - dashboards. *See* Dashboards
 - extraction, 305f, 306
 - fusion, 277
 - gain, 279–280
 - harvesting, 256 (*See also* Data mining)
 - quality, 202
 - reporting, 223–225
 - visualization, 225, 227, 234 (*See also* Data visualization)
 - warfare, 263
- Information-as-a-service (IaaS), 580–581
- Information Builders, 204
- Information systems (IS), 19
- Information Systems Research* (ISR), 323–325
- INFORMS. *See* Institute for Operations Research and the Management Sciences (INFORMS)
- Infospace, 348
- Infrastructure, 32, 220–223
- Infrastructure-as-a-service (IaaS), 27, 578, 580–581

- In-memory analytics, 142
 - In-memory storage technology, 222
 - Inmon, Bill, 191, 202, 204, 207t
 - In-motion analytics, 139
 - Innovation and AI, 19
 - Innovation networks, 356
 - Input gate, 414
 - Input/output (technology) coefficients, 459
 - Input/output of network, 382
 - Insightful Miner, 287t
 - Instagram, 406
 - Institute for Operations Research and the Management Sciences (INFORMS), 34, 444, 449
 - Insurance industry
 - AI in, 85
 - association rule mining, 283
 - data mining, 262
 - service products, 283
 - Integration
 - business intelligence (BI), 33–34
 - data warehousing, 194
 - technologies, 214
 - Intelligence, 87
 - assisted, 70
 - augmented/augmentation, 70–71
 - and automated decision support, 81
 - capabilities, 72
 - content, 72
 - government, 331–332
 - human intelligence *vs.* AI, 72, 73t
 - types, 71
 - Intelligence phase of decision-making, 20
 - classification of problems, 23
 - data collection, 22
 - decomposition of problems, 23
 - identification of problems, 21–22
 - problem ownership, 23
 - Intelligent agent (IA), 33, 75
 - Intelligent bots, 103
 - Intelligent systems
 - adoption, 637
 - in business, 636
 - ethical issues, 644–646
 - impacts on organizations. *See* Organizations, intelligent systems
 - implementation process, 633–634, 633f
 - on jobs and work, 656–662, 660t
 - legal issues, 640–642
 - privacy issues. *See* Privacy in intelligent technology
 - private data, 644
 - successful deployment. *See* Deployment of intelligent systems
 - vignette, 630–633
 - Intercept, 171
 - Interfaces, 444
 - Intermediate result variables, 451
 - Internal cloud, 584
 - International Classification of Diseases, 567
 - Internet, 198–199, 433, 642. *See also* Web analytics; Web mining
 - data visualization, 227
 - marketing strategy, 347
 - search engine. *See* Search engines
 - Internet of Things (IoT), 59, 301
 - applications, 608–613
 - benefits of, 603
 - building blocks of, 602f
 - changing everything, 600
 - characteristics, 599–600
 - and decision support, 605
 - defined, 599
 - drivers of, 603–604
 - ecosystem, 600, 601f
 - essentials, 599–608
 - French national railway system's use, 609
 - hardware, 602
 - and managerial considerations, 637–640
 - platforms, 603
 - privacy in, 643
 - process of, 604f
 - RFID and smart sensors in, 606–608
 - sensors and, 605–608
 - strategy cycle, 639f
 - structure of, 601–603
 - technology infrastructure, 602–603, 602f
 - work process, 604–605
 - Interoperability, 637
 - Interpersonal communication skills, 18
 - Interpretability, 273
 - Interquartile range, 161
 - Interval data, 126, 260
 - Intranet, 198–199
 - Inventory optimization, 47t
 - IoT. *See* Internet of Things (IoT)
 - Irregular input, 309
 - Islands of data, 191
 - IT strategy, 142
 - ixQuick, 348
- J**
- Jackknifing, 275
 - Java, 42
 - JavaScript Object Notation (JSON), 421, 588
 - Jeopardy, 302–304
 - JetBlue Airlines, 292
 - JMP Pro, 558–564
 - Job tracker, 144
 - Joint distribution, 392
 - Journal of Management Information Systems (JMIS), 323–325
- K**
- kaggle.org, 277
 - KDD. *See* Knowledge discovery in databases (KDD)
 - KNuggets.com, 290, 502–504
 - Keras (learning framework), 422–423
 - Key performance indicators (KPIs), 224, 242, 590
 - business reports, 224
 - dashboards, 238, 242
 - Keywords, 353
 - k*-fold cross-validation, 275, 275f
 - Kilobytes (KB), 138
 - Kimball, Ralph, 191, 202, 204, 207t
 - Kip chatbot, 105
 - k*-means clustering algorithm, 157, 260, 266, 281, 282, 282f
 - k*-modes, 281
 - KNIME tool (data mining tool), 240–241, 286, 359–360, 527–528
 - correctness and flexibility, 530–531
 - data science in enterprise, 531–533
 - extensive coverage, 531
 - KNIME Hub, 529–530
 - learning curve, 528–529
 - predictive analytics, 534–535
 - time and money, 528
 - workflow, 535–541
 - Knowledge
 - acquisition, 78, 78f
 - of context, 411
 - data, 121, 121f
 - graph, 49, 50
 - management, 19
 - patterns, 269
 - representation, 78
 - Knowledge-based modeling, 448
 - Knowledge discovery in databases (KDD), 270, 271f
 - Knowledge discovery in textual databases. *See* Text mining
 - Knowledge extraction, 256. *See also* Data mining
 - Knowledge management systems (KMS), 19
 - Knowledge Miner, 287t
 - KPIs. *See* Key performance indicators (KPIs)
 - Kurtosis, 164
- L**
- Labor market, 657
 - LaMDA. *See* Language Model for Dialogue Applications (LaMDA)
 - Landing page profiles, 353
 - Language Model for Dialogue Applications (LaMDA), 624
 - Large Language Models (LLMs), 417, 418
 - Large data, 261
 - Latent Dirichlet allocation (LDA), 338–339, 338f
 - Latent semantic analysis (LSA), 337
 - Latent semantic indexing (LSI), 308, 337
 - Law enforcement, 155, 263
 - agencies, 254
 - and Big Data, 155
 - data mining, 263
 - LDA. *See* Latent Dirichlet allocation (LDA)
 - Leads, 353
 - Leaf node, 278
 - Learning, 173
 - Learning chatbots, 103
 - Learning process in ANN, 385–386
 - backpropagation, 386–388, 387f
 - Leave-one-out method, 275, 390
 - LeClaire, Brian, 43
 - Left-hand side (LHS), 284
 - Legal issues in intelligent systems, 640–642
 - Legislation, 30
 - Lewis, Michael, 3
 - Lexicon, 332, 333f, 334–335
 - LHS. *See* Left-hand side (LHS)
 - Libraries (software), 421
 - Caffe, 421
 - Keras, 422–423
 - TensorFlow, 422
 - Theano, 422
 - Torch, 421
 - Lies, 314–316
 - Lift metric, 284
 - Lindo Systems, Inc. (lindo.com), 466
 - Linear discriminant analysis, 173
 - Linearity, 172
 - Linear programming (LP)
 - defined, 458
 - modeling, 459–465

- Linear regression model, 170f
 - assumptions in, 172–173
 - development of, 170–171
 - effectiveness of, 171
 - See also* Regression modeling
- Line charts, 228
- Link analysis, 260
- LinkedIn, 42, 654
- Link function, 173
- Links, 59
- Livestock producer, 39
- LLMs. *See* Large Language Models (LLMs)
- Local interpretability, 391–393
- Local Interpretable Model-Agnostic Explanations (LIME), 392
- Localization, 100
- Location-based analytics
 - applications for consumers, 595
 - classification, 589f
 - geospatial analytics, 589–593
 - multimedia exercise in analytics, 593
 - real-time location intelligence, 594
- Location intelligence, 590
- Logistic function, 174, 174f
- Logistic regression, 35, 157, 173–179, 174f, 258
- Logistic regression coefficients, 174
- Logistics, data mining, 262
- Long short-term memory (LSTM) network, 396, 411–416, 619
 - applications, 415–416
 - architecture, 414f
 - Caffe, 421
- Long-term memory, 413
- LP. *See* Linear programming (LP)
- Lua programming language, 421
- LSTM network. *See* Long short-term memory (LSTM) network
- M**
- MA. *See* Medicare Advantage (MA)
- MAARS (Modular Advanced Armed Robotic System), 96
- Machine-learning, 374, 575
 - algorithms, 126, 373, 421, 336
 - explainability and transparency in, 389–393
 - techniques, 75–76, 373, 374, 374f, 385
- Machine-to-machine (M2M) communications, 614
- Machine translation of languages, 77–78, 312, 415
- Machine vision, 76
- Macros, 456
- MAE. *See* Mean absolute error (MAE)
- Magic bullet syndrome, 294
- Magic Quadrant for Business Intelligence and Analytics Platforms, 232, 233–234
- Mahindra & Mahindra Ltd., 95–96
- Mahout, 147
- Male comorbidity networks, 568f
- Mainframes, 190
- Major League Baseball, 263
- Management, 654–655
- Management information systems (MIS), 25
- Manhattan distance, 281
- Manual methods, 190
- Manufacturers, 262
- Manufacturing, 262, 451t
- MAPE. *See* Mean absolute percent error (MAPE)
- Mapping and localization, 100
- MapReduce technique, 144–146, 149
 - defined, 145
 - graphical depiction of, 145f
 - use, 146
- Maps
 - geographic maps, 229–230
 - heat maps, 6, 7f, 230
 - tree maps, 230
 - See also* Graphs
- Market-basket analysis, 46, 47t, 260, 282–283
- Marketing, 451t
- Marketing applications, 313
- Market segmentation, 280
- MARVI, 665
- Maryland, 186–188
- Mask mandates, COVID-19, 53–57, 57t
- Massive parallelism, 303
- Mass spectrometry proteomics, 316
- Master data management, 221
- Mathematical models, 461f
 - decision support mathematical models. *See* Decision support mathematical models
 - implementation, 465–466
 - linear programming model, 459–465
 - mathematical programming optimization, 458–466
- Mathematical programming tools
 - application, 458–459
 - components of, 450–451
 - defined, 458
 - implementation, 465–466
 - LP model, 460–465
 - optimization, 458–466
 - structure of, 451–452
- Mathematical representation, 158
- Matrix size, 321
- Max pooling function, 403, 404f
- McKinsey & Company management consultants, 17
- McKinsey's Global Institute, 637–638
- Meals on Wheels America, 455–456
- Mean, 158–159, 162
- Mean absolute deviation, 161
- Mean absolute error (MAE), 180
- Mean absolute percent error (MAPE), 180
- Mean squared error (MSE), 180
- Measures of centrality tendency, 158
- Measures of dispersion, 160
- Measures of location or centrality, 158
- Measures of spread decentrality, 160
- Median, 158, 159, 162
- Medical devices, 156
- Medical records, 283
- Medicare Advantage (MA), 43–45
- Medicine, data mining, 263, 283
- Megabytes (MB), 138
- Mergers, 211
- Message feature mining, 314
- Metadata, 194, 195, 197
- Metaverse, 623
- Metric management reports, 224
- Metro Meals on Wheels Treasure Valley, 455–456
- Microsoft, 191, 205t, 606
 - Azure, 37, 585
 - Cortana, 111, 415
- Enterprise Consortium, 59, 290
- Excel, 164–167, 165f, 166f, 167f, 455–456, 457f, 461–462, 478
- Skype Translator service, 415–416, 416f
- SQL Server, 290
- TrueText, 415
- MicroStrategy, 569
- Middleware tools, 197
- Minard, Charles Joseph, 225–226
- MineMyText.com, 586
- MIS. *See* Management information systems (MIS)
- MIS Quarterly (MISQ), 323–325
- Missing values, 128
- Mixed-integer programming model, 458–459
- Mobile user privacy, 643
- Mode, 158, 159–160
- Model-based decision making, 444–450
 - application, 445–446
 - current trends in modeling, 449–450
 - environmental scanning and analysis, 446
 - knowledge-based modeling, 448
 - model categories, 448–450, 449t
 - model management, 448
 - prescriptive analytics, 446
 - of problem and environmental analysis, 446–448
 - vignette, 442–444
- Model building, 266–268
- Model categories, 448–450, 449t
- Model ensembles. *See* Ensemble models
- Modeling and analysis
 - certainty, uncertainty, and risk, 452–453
 - decision analysis, 471–473
 - goals, 467–468, 472, 473t
 - goal seeking analysis, 470
 - mathematical models for decision support, 450–451
 - mathematical programming optimization, 458–466 (*See also* Linear programming (LP))
 - sensitivity analysis, 468–469
 - with spreadsheets, 453–456 (*See also* Spreadsheets)
 - what-if analysis, 469, 469f
- Modeling process, 39–40, 40f
- Model management, 448
- Models, 25, 257
- Moneyball* (Lewis), 3, 650
- Monte Carlo simulation, 477, 478
- Morphology, 308
- Motorola, Inc., 216
- Moving average, 180
- MPQA Corpus, 336
- MSE. *See* Mean squared error (MSE)
- Multicollinearity, 173
- Multicriteria decision analysis, 471
- Multidimensional analysis (modeling), 449
- Multilayer perceptron (MLP), 396
- Multilevel text analysis, 316–317, 317f
- Multiple-Aspect Restaurant Reviews, 336
- Multiple databases, 195
- Multiple goals, 467–468, 472, 473t
- Multiple-input neuron, 377, 378f
- Multiple linear regression, 171
- Multiple models, 277
- Multiple regression analysis, 157, 170
- Multiplexity, 356
- Murphy, Paul, 191

Mutuality/reciprocity, 356
 MYCIN expert system, 432
 MyWebSearch, 348

N

Name node, 146
 Napoleon's Army, 227f
 National Basketball Association (NBA), 263
 National Centre for Text Mining, 317
 National Collegiate Athletic Association (NCAA), 263
 National Highway Traffic Safety Administration (NHTSA), 393
 National Institute of Standards and Technology (NIST), 576
 National Institutes of Health, 317
 Natural language generation, 312
 Natural language processing (NLP), 77–79, 305, 308–312, 329, 410
 concept, 308–312
 defined, 309
 as text analytics, 305f
 Natural language understanding, 312
Nature, 317
 NCAA Bowl Game outcomes, 174–179
 Netezza, 205t
 Net input function, 376
 Network
 architectures, 381
 closure, 356
 diagrams, 229
 gradients, 386
 science, 42
 structure, 381–382
 virtualization, 581
 Network analytics, 620–622
 Neural computing, 386
 Neural network models, 394
 Neural networks, 35, 258–260, 276, 281, 375
 with backpropagation, 386–388, 387f
 convolutional. *See* Convolutional neural networks (CNN)
 deep. *See* Deep neural networks
 development process, 383–388, 384f
 with layers and neurons, 378f, 381f
 shallow, 375–383
 transfer functions in, 376, 377f
 See also Artificial neural network (ANN)
 Neuron, 374, 381
 artificial. *See* Artificial neuron
 backpropagation of error, 387f
 vs. hidden layers, 397
 summation function for, 382, 383f
 New Member Predictive Model (NMPM), 44–45
 Newness, 220
 New organizational units, 651–652
 New store analysis, 47t
 New visitors, 353
 NLP. *See* Natural language processing (NLP)
 Nodes, 567
 Noise words, 307
 Nominal data, 125, 265
 Non-fungible tokens (NFT), 623
 Normal distribution, 163, 163f
 Normality, 259
 Normality (of errors), 173
 Normative analytics, 41
 NoSQL, 148, 151

Novelty, 220
 N–P (negative/positive) polarity classification, 332–333, 335f
 n-tiered architectures, 197
 Nuclear physics, 255
 Numeric data, 126, 159, 265
 Numeric representation, 126

O

Objective function, 459
 Objective function coefficients, 459
 Obsolete data, 194
 Occurrence matrix, 308
 OCR. *See* Optical character recognition (OCR)
 ODS. *See* Operational data store (ODS)
 Offline campaigns, 352
 Oil and gas exploration assets, 605–606
 OLAP. *See* Online analytical processing (OLAP)
 OLTP. *See* Online transaction processing (OLTP)
 1-800-Flowers.com, 652, 652–653
 O'Neil, Cathy, 650–651
 Online analytical processing (OLAP), 18, 31, 157, 223, 209–210, 210t, 456
 Online campaigns, 352
 Online transaction processing (OLTP), 31, 223, 215, 209, 210t
 On-site Web analytics, 350–351
 Oozie system, 147
 Open AI, 670
 Opening vignettes
 balancing delivery routes, production schedules, and inventory, 442–444
 fighting fraud with DL and AI, 369–372
 Grant Thornton, 64–65
 IBM Watson, 302–304
 predictive policing, 251–252
 Seagate, 495–497
 SiriusXM Radio, 118–121
 sports analytics, 3–17
 tax fraud, 186–188
 Openshift, 585
 Open source software, 218
 Operational databases, 31
 Operational data store (ODS), 194–195
 Operations research (OR), 26
 Oper marts, 195
 Opinion mining, 326
 Opinion-oriented search engines, 332
 Optical character recognition (OCR), 312, 618
 Optimal solution, 459
 Optimistic approach (Utopia), 472, 662, 663–664
 Optimization
 algorithms, 449t
 analytic formula, 449t
 deep MLPs, 396
 mathematical programming, 458–466
 SEO, 347–349
 OR. *See* Operations research (OR)
 Oracle Corporation, 18, 191, 199, 205t
 Oracle Data Mining (ODM), 287t
 Orange Data Mining Tool, 287t, 541–547
 Orbital Insights, 598
 Ordinal data, 126, 265–266
 Ordinal multiple logistic regression, 126

Ordinary least squares (OLS), 56, 171
 Organization, 32
 Organizational alignment, 637
 Organizational impacts, 202
 Organizational redesign, 654
 Organizational structure, 651
 Organizations, intelligent systems, 651–656
 business transformation, 652
 competitive advantage, 652–653
 industrial restructuring, 655–656
 new units and management, 651–652
 organizational design, 654
ORMS Today, 444, 479
 O–S Polarity (Objectivity–Subjectivity Polarity), 332
 Outliers, 159, 161
 Output gate, 414
 Output variable, 169
 Overall Analysis System for Intelligence Support (OASIS), 313
 Overall classifier accuracy, 273
 Overall F-test, 171
 Overfitting in ANN, 388, 388f

P

PaaS. *See* Platform-as-a-service (PaaS)
 Page-loading speed, 199
 Page views, 351
 Palisade Corp. (palisade.com), 478, 484
 PANDA. *See* Purina Animal Nutrition Data Analytics (PANDA)
 Parallel processing, 199
 Parameters, 451
 Parameter sharing, 401
 Partitioning, 199
 ParkPGH, 595
 Part-of-speech tagging, 307, 309, 317
 Passive tags, 607
 Patil, D. J., 42
 Pattern analysis, 256. *See also* Data mining
 Pattern searching, 256. *See also* Data mining
 Pearson, Karl, 169
 Pennsylvania Adoption Exchange, 454–455
 Penzias, Arno, 254
 People Analytics, 654
 Pepper robot, 97
 Perceptron, 375
 Per class accuracy rates, 273
 Performance dashboards. *See* Dashboards
 Performance function, 385
 Periodicals, 59
 Personal values, 644
 PERT charts, 229
 Pervasive confidence estimation, 303
 Pessimistic approach (Dystopia), 472, 662
 “Petabyte age”, 588
 Petabytes (PB), 137, 138
 Physical data integration, 214
 Pie charts, 228
 Pig query language, 147
 Pivot, 210
 Planning, 31–32, 362
 Platform-as-a-service (PaaS), 578, 580
 Playfair, William, 225, 226f
 Pokémon Go, 594
 Polarity identification, 334
 “Polarization” of the labor market, 657
 Police departments, predictive analytics, 251–254

- Politics, 331
 - PolyAnalyst, 287t
 - Polysems, 307
 - Pooling layer, 401, 403–404
 - Popular Data Analytics Software Tools, 291f
 - Power industry, 155–156
 - Power users, 197
 - Precision, 274t
 - Prediction model performance, 424, 424f
 - Predictions, 169, 258
 - Predictive accuracy, 272–273
 - Predictive analytics, 35–36, 126, 446, 594, 595
 - algorithms, 126
 - data mining. *See* Data mining
 - data types, 126–127
 - ensemble models, 277–278
 - forecasting, 446–447
 - logistic regression, 173
 - social analytics. *See* Social analytics
 - social media analytics. *See* Social media analytics
 - statistical methods, 157
 - text analytics. *See* Text analytics
 - Web analytics. *See* Web analytics
 - See also* Data mining
 - Predictive models, 44–45, 449t, 512–515
 - Predictive policing, 251–254
 - Predictor variables, 173
 - Prescriptive analytics, 36, 41, 442–443
 - application, 447–448
 - decision analysis. *See* Decision analysis
 - goal seeking, 456, 470
 - model-based decision making, 444–450
 - model examples, 446
 - multiple goals, 467–468
 - optimization. *See* Optimization
 - predictive analytics, 446–447
 - sensitivity analysis, 468–469
 - simulation. *See* Simulation
 - what-if analysis, 456, 469, 469f
 - Preset robots, 99
 - Pressure points, 90
 - Price elasticity, 47t
 - PricewaterhouseCoopers (PwC), 660
 - Primary data, 261
 - Privacy in intelligent technology, 640, 642
 - business intelligence (BI), 33
 - collection of information, 642
 - data mining, 292
 - data privacy, 123
 - data warehousing, 217
 - example, 643–644
 - in IoT, 643
 - mobile user, 643
 - ownership of private data, 644
 - technology issues, 643–644
 - violations, 644
 - Privacy lawsuits, 292
 - Private cloud, 584
 - Probabilistic decision-making situation, 453
 - Probabilistic simulation, 477, 477t
 - Probability distribution, 266
 - Problem identification, 446–448
 - Problem ownership, 23
 - Process efficiency, 19, 143
 - Processing capabilities, 142
 - Processing element (PE), 376, 381
 - ProClarity, 191
 - Production, data mining, 262
 - Productivity, 595
 - Products, 59
 - Profit model, 451
 - Program, 458
 - Programmability, 456
 - Programming languages, 215, 290
 - Project management, 24
 - Proximity, 356
 - Proximal Policy Optimization (PPO), 418
 - Proximity sensors, 605
 - Public cloud, 584
 - Pure-play BAM, 33
 - Purina Animal Nutrition Data Analytics (PANDA), 37–38, 40
 - platform architecture, 38f
 - Purina MQ (Microbiome Quotient) Bioinformatics Platform, 619
 - Python, 42, 97, 290, 421, 422, 584–585
 - classification, imbalanced data set, 525–527
 - importing data, 521–525
 - installation, 520–521
- ## Q
- Qualitative data, 265, 266
 - Quality, 358
 - Quantitative data, 265, 266
 - Quantitative models, 467
 - decision variables, 451
 - defined, 450
 - intermediate result variables, 451
 - result (outcome) variables, 450
 - structure of, 450f
 - uncontrollable variables, or parameters, 451
 - Quartiles, 161
 - Query analyzer, 345f, 346
 - Query-specific clustering, 322
 - Question answering, 307, 312
 - Queuing, 479
- ## R
- R (open source platforms), 290
 - application tutorial, 507–510
 - data visualization, 510–512
 - explainable AI, 515–516
 - installation, 505–507
 - machine learning, predictive modeling, 512–515
 - Rattle, 516–517
 - R^2 (R -squared), 171
 - Radio-frequency identification (RFID), 32, 262, 606–608
 - Random Forest, 277, 287t
 - Ranges, 160
 - RapidMiner software, 286, 290, 549–552
 - Rapleaf Software Company, 643
 - Rathi, Abhishek, 45
 - Ratio data, 126
 - RDBM. *See* Relational database management (RDBM)
 - Reach, 358
 - Real-time BI applications, 32
 - Real-time data analytics, 153. *See also* Stream analytics
 - Real-time data warehousing (RDW), 27, 221
 - Real-time location intelligence, 594
 - Real-world data, 127
 - Recall, 274t
 - Rectilinear distance, 281
 - Recession, 218
 - Reciprocity, 356
 - Recurrent neural network (RNN), 396, 412–414, 413f
 - Red Brick Systems, 191
 - Referral Web sites, 352
 - Regional Neonatal Associates, 458
 - Regression, 169, 272
 - Regression modeling, 169–180
 - application, 174–179
 - correlation *vs.* regression, 169
 - effectiveness of model, 171
 - evaluation of fit, 171
 - linear regression model, development of, 170–171
 - linear regression model, important assumptions in, 172–173
 - logistic regression, 173–179, 174f
 - recognizing good model, 171
 - simple *vs.* multiple regression, 170
 - time series forecasting, 179–180
 - Regular bots, 103
 - Regularization strategy, 388
 - Regulation, 30
 - Regulatory requirements, 211
 - Reinforcement Learning from Human Feedback (RLHF), 417, 418
 - Relational database management (RDBM), 26, 191, 199, 220, 222
 - Relational triads, 356
 - Relation extraction, 410, 410f
 - Relevancy, 124
 - Reliability, 123
 - Remote-controlled robots, 99
 - Rental car companies, 262–263
 - Report, 223
 - Reporting analytics, 35
 - Representation learning technique, 374, 374f
 - Representative analytics as service offerings, 586
 - Reproduction, 482
 - Research literature survey, 323–325
 - Resources, 59
 - Responding Cycle, 345f, 346–347
 - Response variable, 169, 170
 - Result (outcome) variables, 450, 450f, 451t
 - Retailing industry, data mining, 262
 - Retail sector, 45–48, 255
 - Retail value chain, 45–48, 46f, 47t
 - Retrieval speed, 219
 - Returning visitors, 353
 - Return on investment (ROI), 204, 210, 351
 - Review-oriented search engines, 332
 - Reward Model (RM), 418
 - RFID. *See* Radio-frequency identification (RFID)
 - RHS. *See* Right-hand side (RHS)
 - Richness, 123
 - Ride sharing by Taxi Bot, 105
 - Right-hand side (RHS), 284
 - Right-time data warehousing, 27
 - Ripple effect, 362
 - Risk, 453
 - calculated risk, 453
 - catastrophic loss, 472
 - management, 143
 - treating risk, 472

- RLHF. *See* Reinforcement Learning from Human Feedback (RLHF)
- RMSE. *See* Root Mean Square Error (RMSE)
- RNN. *See* Recurrent neural network (RNN)
- Robo advisors
 defined, 111
 financial advisors, 111
- RoboCoke, 105
- Robotic process automation (RPA), 616–618
- Robotics
 Adidas, 93
 Agrobot, 98–99
 BMW, collaborative robots, 94
 changing precision technology, 93
 Da Vinci Surgical System, 98
 in defense industry, 96
 history, 91–93
 illustrative applications, 93–102
 Mahindra & Mahindra Ltd., 95–96
 overview, 88–89
 Pepper, 97
 San Francisco Burger Eatery, 94–95
 Spyce, 95
 Tega, 94
 The Robotics Institute of America, 88
 Robots, 88–89
 (robo) advisors. *See* Robo advisors
 autonomous, 89
 categories of, 99–100
 collaborative, 94, 99
 defined, 88
 dangers of, 662–664
 in defense industry, 96
 emotional support, patients and children, 89–91
 Huggable Robot, 90–91, 90f
 in motion, 100–102 (*See also* Autonomous vehicles)
 preset, 99
 remote-controlled, 99
 stand-alone, 99
 supplementary, 100
 See also Robotics
 Robustness, 273
 ROC curve, 275–276, 276f
 Rockwell Automation, 605–606
 ROI. *See* Return on investment (ROI)
 Roll-up, 210
 Root Mean Square Error (RMSE), 171
 Rotation estimation, 275
 Rough sets method, 258, 277
 Rule-based expert systems (ESs), 26
 Rule induction, 260
 Russian campaign (1812), 227f
- S**
- SaaS. *See* Software-as-a-service (SaaS)
- Sales/conversions, 354
- Salesforce's AI Einstein, 86–87
- Sales transactions, 283
- Sam M. Walton College of Business, 290
- San Francisco Burger Eatery, robotics, 94–95
- Sankey chart/diagram, 8, 8f
- SAP InfiniteInsight (KXEN), 287t
- Sarbanes-Oxley Act, 30
- SAS Enterprise Miner, 287t
- SAS Institute, Inc., 34, 168, 234, 205t, 269, 586
- SAS Visual Analytics, 236, 237, 237f, 586
- SAS Visual Statistics, 586
- SAS Viya, 555–558
- Scalability, 273
- Scatter/gather clustering, 322
- Scatter plots, 170f, 228–229
- Scene recognition, 77
- Schmidt, Eric, 349
- Schrage's models for AI, 82
- SCM. *See* Supply chain management (SCM)
- Search, 342
- Search engine optimization (SEO), 347–349
- Search engine poisoning, 348
- Search engines, 344–349, 352
 algorithms, 349
 anatomy of, 344–346
 defined, 344
 development cycle, 344–346
 document indexer, 345
 document matcher/ranker, 346–347
 effectiveness, 344
 efficiency, 344
 evaluation metrics, 344
 most popular, 348
 optimization, 347–349
 and organic search traffic, 349
 poisoning, 343
 query analyzer, 346
 response cycle, 346–347
 taxonomy, 341f
 Web crawlers, 345
- Search precision, 322
- Search recall, 322
- Search spam, 348
- Secondary data, 261
- Secondary node, 146
- Securities and Exchange Commission, 224
- Securities trading, 262
- Security, 637
 applications, 313–316
 business intelligence (BI), 33
 cybersecurity, 155
 data security, 123
 data warehousing, 217, 220
 threats, 330
- Seeds, 345
- Segmentation, 357
- Self-driving vehicles. *See* Autonomous vehicles
- Self-organizing maps, 260
- Semantic orientation, 336
- Semistructured data, 125
- Semistructured problems, 23
- SEMMA (sample, explore, modify, model, and assess) process, 269–270, 270f
- Seniors, 43–44
- Sensitivity analysis, 24, 289, 390, 391f, 468–469
 on ANN model, 393
 application, 393–395
 graphical depiction, 391f
 injury severity factors, traffic accidents, 393–395
 leave-one-out methodology, 390–391
 top 15 risk factors, 424–425, 424f
- Sensors, 75, 88, 594
 applications and RFID, 606–608
 camera-based, 98–99
 defined, 605, 607
 and IoT, 605–608
 smart, 607–608
 technology, 605
- Sensor to insight, 605
- Sensory data, 262
- Sentiment analysis, 155, 310–312, 326–336
 aggregation, 334
 applications, 330–332
 brand management, 331
 collection, 334
 concept, 326–327
 defined, 310
 financial markets, 331
 government intelligence, 331–332
 lexicon, 334–335
 N–P polarity classification, 332–333
 polarity identification, 334
 politics, 331
 process, 332–334, 333f
 semantic orientation, 336
 sentiment detection, 332
 target identification, 334
 training documents, 335–336
 voice of the customer (VOC), 330–331
 voice of the employee (VOE), 331
 voice of the market (VOM), 331
- Sentiment detection, 332
- SentiWordNet, 335
- Sequence mining, 260
- Sequential relationships, 258
- Serial analysis of gene expression (SAGE), 316
- Server capacity, 199
- Server virtualization, 582
- Service-based architecture (SBA), 614
- Service-oriented architecture (SOA), 27, 214
- Services, 451t
- SGD. *See* Stochastic gradient descent (SGD)
- Shallow knowledge, 303
- Shallow neural networks, 375–383
- Shallow-parsing, 317
- Shapley Additive Explanations (SHAP), 392
- Shells, 33
- ShopiiBot, 105
- Shopper analytics, 46
- Shopper insight, 47t
- Short-term memory, 413
- Siemens, 205t
- Sigmoid transfer functions, 376, 387
- Silvaris Corporation, 36
- Simon's process of decision-making, 20, 21f
- Simple average, 180
- Simple linear regression, 173
- Simple logistic regression, 173
- Simple present-value cash flow model, 452
- Simple regression, 170
- Simple regression analysis, 170–171
- Simple split, 274–275, 274f
- Simulation, 449t, 473
 advantages, 475
 application, 474
 characteristics, 473–474
 defined, 473
 disadvantages, 475
 discrete event, 477, 478
 experiments, conduct of, 473
 inadequacies, 478
 methodology, 475–476
 Monte Carlo simulation, 477
 packages, 475

- probabilistic, 477
- process, 476f
- software, 479–480
- test and validation, 476
- time-dependent, 477
- time-independent, 477
- types, 477–479
- visual interactive simulation (VIS), 479
- Simultaneous goals, 467
- Single-input neuron, 376f
- Singular value decomposition (SVD), 308, 321, 337
- Siri (Speech Interpretation and Recognition Interface), 110, 415, 427
- SiriusXM Radio, 118–121
- Skewed data, 159
- Skewness, 163–164
- Skills availability, 143
- Skype Translator service (Microsoft), 415–416, 416f
- Slice, 210, 211f
- “Slice and dice”, 210
- Slope, 171
- Smart assistant shopping bots, 111
- SmartBin, 600
- Smart cities
 - improving transportation in, 612
 - smart components and smart factories in, 611–612
 - smart (digital) factories in, 612
- Smart homes and appliances
 - barriers to adoption, 611
 - Bot, 611
 - components, 608f, 609–610
 - defined, 609
 - smart appliances, 610–611
- Smartphone platforms, 595
- Smart sensor, 607–608
- Smart speaker, 27
- sncf.com, 609
- Snowflake, 586, 588
- Snowflakes schema, 208–209, 209f
- SOA. *See* Service-oriented architecture (SOA)
- SOA coarse-grained services, 214
- Social analytics, 355–362
 - connections, 356
 - distributions, 357
 - segmentation, 357
 - social media analytics. *See* Social media analytics
 - social network analysis, 355–356
- Social capital, 357
- Social circles, 357
- Social media, 217–218
- Social media analytics, 357–358
 - accessibility, 358
 - accuracy of text analysis, 362
 - best practices, 361–362
 - beyond the brand, 362
 - concept, 357–358
 - connections, 356
 - defined, 357–359
 - distributions, 357
 - elusive sentiment, tracking, 361–362
 - frequency, 358
 - impact, 361
 - influencers, 362
 - intelligence, 362
 - measurement, 361
 - not rating system, 361
 - planning, 362
 - quality, 358
 - reach, 358
 - ripple effect, 362
 - tools, 362
 - updatability, 358
 - usability, 358–360
 - user engagement, 358–359, 359f
- Social media campaigns, 240–241, 359–360
- Social network analysis, 355–356, 361
- Social network analysis metrics, 356–357
- Social networking web sites, 578
- Social networks, 355
- Social robot, 90
- Soft data, 256
- Softlayer®, 329
- Softmax transfer function, 410
- Software
 - AI, 613
 - backend, 603
 - data mining, 262, 287t
 - libraries. *See* Libraries (software)
 - robots, 616
 - simulation, 475, 479–480
 - Tableau, 586
 - tools, 286–291, 291f
 - Weka, 286
- Software-as-a-service (SaaS), 218, 569, 580
- Software monitors, 33
- Solution, 458
- Solution cost, 143
- Solver file, 421
- Sonatica, 238
- Spamdexing, 348
- Spam filtering, 308
- Spark *vs.* Hadoop, 149–151
- Spatial data, 126
- Special weapons observation reconnaissance detection system (SWORDS), 96
- Speech acts, 309
- Speech analytics, 309
- Speech recognition, 312
- Speech synthesis, 312
- Speech (voice) understanding
 - technology, 77
- Speed, 273
- Spiders (web crawlers), 342, 345
- Split point, 278
- Sponsorship, 141
- Sports, data mining, 263
- Sports analytics, 3–17, 4f, 174
 - business office, 5–6
 - the Commissioner, 15–16
 - Cricket, 12–14
 - football, 6–8
 - high school coach, 8–12
- Spreadsheets, 453–457
 - application, 454–456
 - decision modeling and, 453–456
 - excel dynamic model, 457f
 - static model, 457f
- SPRINT (Scalable PaRallelizable INduction of Decision Trees), 279
- Spyce, robotics, 95
- SQL queries, 197
- SQL Server, 290
- SQL Server Data Mining, 287t
- Sqoop tool, 147
- Stand-alone robots, 99
- Standard deviation, 161, 164
- Standardization of encoded attributes, 215
- Stanford—Large Movie Review Data Set, 336
- Stanford University, 309
- Starbucks, 592–593
- Star schema, 208, 209f
- State of nature, 471
- State unit, 414
- Static data, 32, 126
- Static models, 448, 479
- Static network, 412
- Statista.com, 3–4
- Statistical analysis, 276
- Statistical modeling, 42, 157–168
 - for business analytics. *See* Business analytics, statistical modeling for
- Statistics, 42, 157–158, 166f, 276
 - conversion, 353–354
 - vs.* data mining, 261
 - and descriptive analytics, 157f, 158, 164
 - inferential, 158, 164, 169–180
 - statistical software packages, 164
 - text analytics, 305f
- Statistics-based classification techniques, 258–259
- Stemming process, 307, 320
- ST_GEOMETRY, 591
- Stochastic decision-making situation, 453
- Stochastic gradient descent (SGD), 386
- Stock Keeping Units (SKUs), 46
- Stop terms, 319
- Stop words, 307, 319
- Storage
 - and cognitive limits, 19
 - data warehouses, 219
- Storage virtualization, 581
- Store layout, 47t
- Stories, 235–236
- Story structure, 235–236
- Storytelling, 235–236
- Strategy, 31–32, 141
- Stream analytics, 142, 153–154, 154f, 588
 - applications, 154–156
 - and Big Data, 153–156
 - cybersecurity, 155
 - defined, 142, 153–154
 - e-commerce, 154–155
 - financial services, 156
 - government, 156
 - health services, 156
 - law enforcement and cybersecurity, 155
 - power industry, 155–156
 - telecommunications, 155
 - use case of, 154f
- Structural holes, 357
- Structured data, 125, 307
- Student attrition, 130–136
- Subject matter experts, 295
- Subject orientation, 193
- Subjectivity analysis, 326
- Summarization, 307
- Summarization rules, 215
- Summation function, 382, 383f
- Supervised Fine-Tuning (SFT), 417–418
- Supervised induction, 258–260
- Supervised learning process, 258, 385, 386f, 619
- Supplementary robots, 100

Supply chain management (SCM), 31
 Support, 19, 284
 Support metric, 283
 Support vector machines (SVMs), 258
 SVD. *See* Singular value decomposition (SVD)
 SVMs. *See* Support vector machines (SVMs)
 Sybase, 205t
 Synonyms, 307, 320
 Syntactic ambiguity, 309
 System dynamics, 477
 System quality, 202

T

Tableau, 36, 586
 Target, 293
 Target identification, 334
 Tax fraud, 186–188
 Taxicab distance, 281
 TDM. *See* Term–document matrix (TDM)
 TDWI.org, 33
 Technologies of AI, 74, 74f
 autonomous business decisions, 82
 computer vision, 76
 DL, 76
 examples, 75–77
 IA, 75
 machine learning, 75–76
 machine translation of languages, 77–78
 machine vision, 76
 NLP, 77–79
 speech (voice) understanding, 77
 Technology insight
 augmented intelligence, 71
 calculating descriptive statistics in Excel, 164–167
 data lakes, 218–220
 data size, 138
 elements of ANN, 381–383
 explainability and transparency, machine learning, 389–393
 Gartner, Inc.'s business intelligence platform, 233–234
 Hadoop, demystifying facts, 148–149
 hosted data warehouse, 207–208
 LP, 460
 popular search engines, 348
 predictive text mining and sentiment analysis, 336
 RFID sensors, 607
 Schrage's models for AI, 82
 search engines, 348
 storytelling, 235–236
 text mining, 307–308
 Tega, robotics, 94
 Telecommunications, 155, 283
 TensorBoard (visualization module), 422
 TensorFlow (learning framework), 422
 Tensor Processing Unit (TPU), 422
 Terabytes (TB), 137, 138
 Teradata, 119–120, 190–191, 199, 205t, 221, 237, 564–567
 Teradata Aster, 567
 Teradata University for Academics (TUA), 59
 Teradata University Network (TUN), 564
 Teradata Vantage
 application, 566–567
 architecture, 565f
 Teradata Warehouse Miner, 287t

Term, 307
 Term-by-document matrix, 308, 346
 Term dictionary, 307
 Term–document matrix (TDM), 319–321, 320f
 Term Frequency-Inverse Document Frequency (TF-IDF), 311
 Test sample estimation, 274
 Test set, 274
 Text analytics, 19, 305–306, 305f. *See also* Text mining
 Text mining
 Text-based deception-detection process, 315f
 Text categorization, 321
 Text clustering, 311
 Text data mining. *See* Text mining
 Text mining, 27, 306
 academic applications, 317
 application areas, 306–307
 applications, 313–317
 bag-of-words model, 308
 benefits of, 306
 biomedical applications, 316–317
 CNN for relation extraction in, 410
 combined data set, 324t
 context diagram for, 318f
 Corpus, 319
 and customer relationship management (CRM), 313
 defined, 306
 knowledge extraction, 321–325
 marketing applications, 313
 natural language processing (NLP), 308–312
 overview, 305–308
 process, 318–325, 318f
 research literature survey with, 323–325
 security applications, 313–316
 term-document matrix creation, 319–321
 terminology, 307–308
 text analytics and, 305f
 textual data, 305f
 three-step/task, 319f
 Text mining process, 318–325
 association, 322
 classification, 321
 clustering, 321–322
 context diagram, 318f
 corpus, establishment of, 319
 extraction of knowledge, 321–323
 reduction of dimensionality of matrix, 321
 representation of indices, 320
 term–document matrix (TDM), 319–321, 320f
 three-step/task, 319f
 trend analysis, 323
 Text processing using CNN, 408–411
 Text proofing, 312
 Text segmentation, 309
 Text-to-speech, 312
 Textual data, 126, 335–336
 Theano (software), 422
 Three-tier data warehouse, 197–198, 197f
 Threshold value, 382
 TIBCO, 567
 Tie strength, 357
 Time compression, 475
 Time-dependent *vs.* time-independent simulation, 477
 Time of day, 353
 Timelines, 124
 Time on site, 351
 Time-sensitive environments, 139
 Time series analytics, 6, 8f
 Time series data, 180f
 Time series forecasting, 179–180, 180f, 261
 Time series line chart, 226f
 Time-series forecasting, 261
 Tokenizing, 307
 Tone Analyzer, Watson, 431
 Topic tracking, 306
 Topic modeling, 336–338
 Topologies, 381
 Torch (computing framework), 421
 Torch Concepts, 292
 TPU. *See* Tensor Processing Unit (TPU)
 Traders, 262
 Traffic sources, 352
 Training, 358
 Training documents, 335–336
 Training process, 378, 427
 Training set, 274
 Transaction layer, PANDA platform, 38, 38f
 Transaction processing *vs.* analytic processing, 31
 Transfer function, 376
 Transformation-tool approach, 216
 Transitivity, 356
 Transportation, 451t
 Travel industry, data mining, 262–263
 TreeAge Pro (TreeAge Software Inc., treeage.com), 472
 Tree maps, 230
 Trend analysis, text mining, 323
 Trial-and-error sensitivity analysis, 469
 True negative rate, 274t
 True positive rate, 274t
 TrueText (Microsoft), 415
 Tufte, Edward, 225
 Tukey, John W., 161
 TUN. *See* Teradata University Network (TUN)
 Tuples, 153
 TurboRouter, 444
 Turing Test of AI, 72, 74f
 Twitter, 240, 360, 361
 Two-person game, 471
 Two-tier data warehouse, 197, 198f

U

Ubiquitous computing, 599
 Uncertainty, decision making, 452f, 453, 472
 Uncontrollable parameters, 451
 Uncontrollable variables, 450f, 451, 451t, 471
 “Understanding the customer”, 255
 Unique identifier (UID), 599
 Unimodal distribution, 164
 Universal Product Code (UPC), 607
 University of Arkansas, 290
 University of California, Berkeley, 317
 University of Liverpool, 317
 University of Manchester, 317
 University of Tennessee Medical Center, 458–459
 Unstructured data, 125, 307
 Unsupervised learning process, 258, 396, 620
 UPC. *See* Universal Product Code (UPC)

Updatability, 358
 Urban sociology, 356
 U.S. Census, 196
 U.S. Department of Agriculture, 77
 U.S. Department of Education, Center for Educational Statistics, 130
 U.S. Federal Bureau of Investigation (FBI), 313–314
 U.S. government, 217
 Usability, 351–352, 358
 Usability specialist, 243
 User interface, 29
 Users, 220
 Utopia (optimistic approach), 662, 663–664

V

Validity, 124
 Vanishing gradient problem, 405
 Vantage Customer Experience (Vantage CX), 192–193
 Variability, 139
 Variable identification, 446
 Variables, 451
 class/response variable, 173
 decision variables, 451, 459, 471
 dependent variables, 450
 explanatory variable, 170
 explanatory/predictor variables, 173
 intermediate result variables, 451
 output variable, 169
 predictor variables, 173
 response variable, 169, 170
 result (outcome) variables, 450
 uncontrollable variables, 451, 471
 Variable selection process, 129
 Variable speed limit (VSL), 398–400
 Variance, 160
 Variety, 138
 vCreaTek.com LLC, 45
 Velocity, 138–139
 Vendors, 59
 Veracity, 139
 Video analytics, 46, 47t, 77
 Video data, 126
 Vignettes. *See* Opening vignettes
 VIM. *See* Visual interactive modeling (VIM)
 Virtual digital assistants, 429
 Virtual personal assistant (VPA), 642
 Amazon's Alexa and Echo, 110–111
 Apple's Siri, 110
 defined, 109
 Google Assistant, 110
 other personal assistants, 111
 Virtualization, 581–582
 Virtual reality, 594
 VIS. *See* Visual interactive simulation (VIS)
 Vision, 141
 Visitor profiles, 353
 Visual analytics, 232, 234–236, 238f, 261
 emergence of, 232–237
 high-powered visual analytics environments, 236–237
 story structure, 234–236
 Visual interactive modeling (VIM)
 defined, 479
 and DSS, 479
 See also Visual interactive simulation (VIS)
 Visual interactive problem solving, 479
 Visual interactive simulation (VIS)
 concept, 479
 conventional simulation inadequacies, 479
 defined, 479
 models and DSS, 479
 simulation software, 479–480
 Visualization, 50, 261
 PANDA platform, 38, 38f
 tools, 35, 256
 See also Data visualization
 Voice data, 126
 Voice of the customer (VOC), 330–331
 Voice of the employee (VOE), 331
 Voice of the market (VOM), 331
 Volume, 137

W

Waiting-line management, 479
Wall Street Journal, 643
 Walmart, 651
 Walnut chatbot, 105
 Warehousing strategy, 204
 WaterCop (watercop.com) system, 609
 Watson, IBM, 91
 analytics, 657
 Deep Blue (chess program), 430
 DeepQA architecture, 303, 304f, 430
 Personality Insight, 431
 Tone Analyzer (IBM), 431
 See also IBM Watson
 Watson Tone Analyzer, 431
 Waze, 595, 644
Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (O'Neil), 650
 Weather effects, 591
 Web analytics, 341, 349–354
 conversion statistics, 353–354
 dashboard, 354f
 defined, 341, 341f, 349
 metrics, 351–354
 on-site, 350–351
 technologies, 350–351
 traffic sources, 352
 visitor profiles, 353
 Web site usability, 351–352
 See also Web mining
 Web-based data management tools, 210
 Web-based data warehouses, 198, 198f
 Web-based e-mail, 578
 Web browser, 198
 Web content mining, 305f, 341, 342–343
 Web crawlers, 342, 345, 345f, 348
 Web mining, 339–343
 challenges, 340
 defined, 305f, 341, 341f
 overview, 339–343
 taxonomy of, 341f
 Web content mining, 341, 342–343
 Web structure mining, 341, 342–343
 Web usage mining, 341, 349–354
 Web-oriented languages, 42
 Web services, 33
 Web site usability, 351–352
 Web spiders, 345
 Web structure mining, 341, 342–343
 defined, 343
 taxonomy, 341f
 text analytics, 305f
 Web 3.0, 622–623
 Web usage mining, 305f, 341, 349–354, 350f.
 See also Web analytics
 WeChat's super chatbot, 107
 Weighted averages, 281–282
 Weighted Lease Squares (WLS) model, 56
 Weighted moving average, 180
 Weight function, 376
 Weka software, 286, 547–549
 Wells Fargo Bank, 191
 What-if analysis, 456, 469, 469f
 “What They Know”, 643
 White-hat SEO, 348
 Wi-Fi hotspots, 643
 Wikipedia, 578
 Wimbledon, 327–330
 Word disambiguation, 305f
 Word frequency, 307
 WordNet, 310, 334–335
 defined, 310
 web site (wordnet.princeton.edu), 334
 WordNet-Affect, 335
 Word sense disambiguation, 309
 word2vec project, Google, 408–410, 409t
 Word vectors/embeddings, 408, 409f, 410
 Worker nodes, 146
 World Health Organization (WHO), 52
 World Wide Web, 339–340, 345, 345f
 Wow, 348

X

XLMiner, 287t

Y

Yahoo! Search, 147, 343, 347, 348
 Yield management, 263
 Yottabytes (YB), 138
 YouTube, 593

Z

Zementis Predictive Analytics, 287t
 Zettabytes (ZB), 137, 138
 Zion China, 587–588

This page intentionally left blank