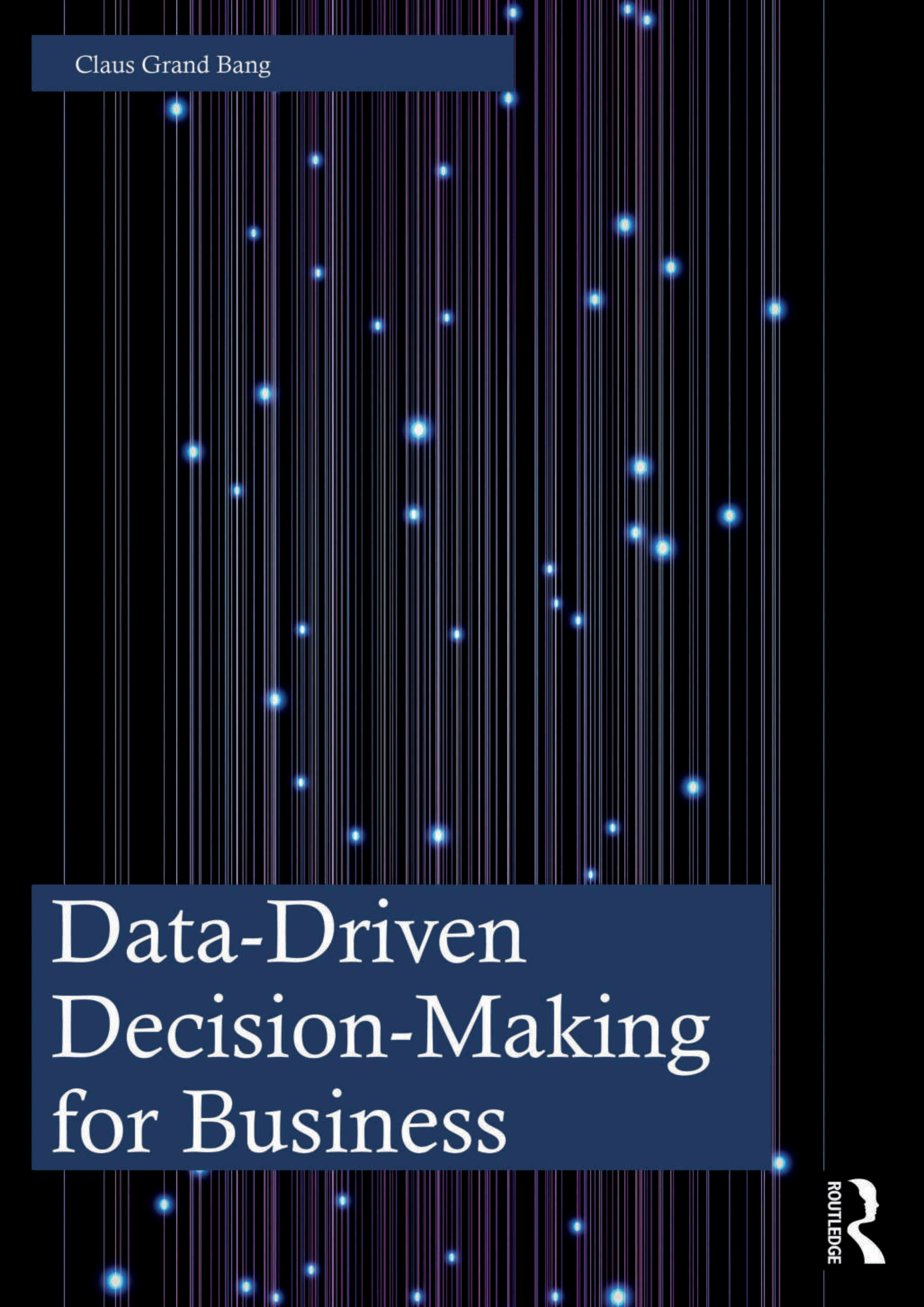


Claus Grand Bang

Data-Driven Decision-Making for Business

Data-Driven Decision-Making for Business

Research shows that companies that employ data-driven decision-making are more productive, have a higher market value, and deliver higher returns for their shareholders. In this book, the reader will discover the history, theory, and practice of data-driven decision-making, learning how organizations and individual managers alike can utilize its methods to avoid cognitive biases and improve confidence in their decisions. It argues that value does not come from data, but from acting on data.

Throughout the book, the reader will examine how to convert data to value through data-driven decision-making, as well as how to create a strong foundation for such decision-making within organizations. Covering topics such as strategy, culture, analysis, and ethics, the text uses a collection of diverse and up-to-date case studies to convey insights which can be developed into future action. Simultaneously, the text works to bridge the gap between data specialists and businesspeople. Clear learning outcomes and chapter summaries ensure that key points are highlighted, enabling lecturers to easily align the text to their curriculums.

Data-Driven Decision-Making for Business provides important reading for undergraduate and postgraduate students of business and data analytics programs, as well as wider MBA classes. Chapters can also be used on a standalone basis, turning the book into a key reference work for students graduating into practitioners. The book is supported by online resources, including PowerPoint slides for each chapter.

Claus Grand Bang is Associate Professor at Dania Academy, Denmark. He has more than ten years of business experience developing companies based on data and another ten years in academia teaching students from all over the world. As a lecturer, he has specialized in the fields of applied data analysis, supply chain management, and project management. He created one of the first applied data analytics degrees in Europe. Now, as Head of Data and IT, at a global biotech company he applies what he has taught in academia.

Data-Driven Decision-Making for Business

Claus Grand Bang

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

Designed cover image: © Flavio Coelho/Getty Images

First published 2025
by Routledge
4 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
605 Third Avenue, New York, NY 10158

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book has been requested

ISBN: 978-1-032-60153-3 (hbk)
ISBN: 978-1-032-60149-6 (pbk)
ISBN: 978-1-003-45778-7 (ebk)

DOI: 10.4324/9781003457787

Typeset in Bembo
by Apex CoVantage, LLC

Access the Support Material: www.routledge.com/9781032601496

Contents

Figures

Illustrations

Tables

CHAPTER 1

Introduction

What Is Data-Driven Decision-Making and Why Does It Matter?

This chapter explains the concept and benefits of data-driven decision-making, as well as some challenges and best practices for implementing it in organizations. We will also cover some common cognitive biases that hinder our data-driven decision-making:

- The history of data-driven decision-making
- The benefits and challenges of data-driven decision-making
- The data-driven decision-making process
- The decision-making styles and links to data.

Chapter case: Saxo Bank, finance, Singapore

Imagine sitting in the corner office. You are looking out across the skyline. In a few minutes, you will be going down the hallway to the big meeting room where the decision regarding the allocations of resources for the next 6 months will be made. How does your stomach feel? If you are like most people, you will be a bit nervous about settling on a final decision.

Now imagine that you knew what the probability of your decision was. Imagine that you could say that, of the potential choices I have, these are the two with the highest probability of success. This is what data-driven decision-making tries to achieve. We are not entirely there yet, but we are getting closer and during this book, you will hopefully have come close to the point where you are the one sitting in the chair with the confidence of data to back decisions.

LEARNING GOALS:

L1.1 Understand how data-driven decision-making (DDDM) is defined and its application

L1.2 Have an understanding of the links between classical decision theory and DDDM

L1.3 Be able to link major cognitive biases to DDDM

L1. Be able to link decision-making styles to DDDM

DOI: 10.4324/9781003457787-1

Let us start with a definition and then unfold it a bit.

The definition we will be using in this book is:

> Data-driven decision-making refers to the systematic collection, analysis, examination, and interpretation of data, usually through the application of analytics or machine learning methods and techniques, to reach informed decisions.
>
> (Elgendy, Elragal, & Päivärintaa, 2022)

It also links very well with the definition from one of the primary software companies in the field where DDDM is defined as using facts, metrics, and data to guide strategic business decisions that align with your goals, objectives, and initiatives (Tableau, 2023).

If you were to look up the definition on Wikipedia (Wikipedia, 2023) you would also get a good indication of its origin in the latest iterations. Here it specifically refers to the use case in the educational sector and how a combination of data points can inform an educator in his choices regarding school and school district management.

In the business world, this translates to a variety of activities (Stobierski, 2019):

- Gather survey responses to ascertain customer preferences for products, services, and desired features
- Perform user testing to observe customer behavior and usage patterns, as well as to identify and address potential issues before a full-scale launch
- Introduce a new product or service in a trial market to gauge its performance and gain insights into its potential market reception
- Evaluate changes in demographic data to identify business prospects or potential risks to the organization
- …

Common for all is that the leader will enhance the decision-making process using data. That does not mean that intuition is of no value, but it will not be standing alone, and you are less likely to fall into the cognitive bias traps, which are explained later in this chapter.

THE HISTORY OF DATA-DRIVEN DECISION-MAKING

Data-driven decision-making is not a new concept. It is just that the amounts of data available and the tools to analyze them are new. That also means that to understand DDDM we should go back a bit and explore how decisions originally are seen to have been made and how the data availability and tools have had an impact on it.

Three different perspectives have historically been the focus when describing decision-making:

- Decision-making process
- Decision-maker
- Decision (type).

This also means that any decision made is dependent on those factors. That is still true when we add data and analytics to the decision, which you will see in the DECAS theory that we will focus on later in this chapter.

The **decision-making process** is concerned with the structure of the process. It can be formalized with the description of the process and the checking that it should go through. There are in general two elements that define how structured a process is. It is the maturity of the organization and the time available to make the decision. That could lead you to think that the more mature the organization is the better the decisions, but it is unfortunately not that simple. Structure takes time and time is not always available. Drucker (1967) defines a linear approach, which we will get back to a little later. More recently, this structured way has been challenged by, amongst others, Mintzberg and Westley (2001) as decisions cannot always be recognized and will therefore follow a more unstructured process.

The **decision-maker**, or the one making the final decision will be bringing his or her personal preference and experiences to the decision-making process. We will be looking into cognitive biases later in this chapter, as they have a potentially unwanted impact on the decision-making process. Historically, decision-making has been thought of as a rational process (Mintzberg, 1990) along with the economics field. That has over the last 30 years been loosened with the understanding of bias and the acceptance that not all can be known about the alternatives and the consequences of each of the alternatives.

The **decision** quality includes timeliness, accuracy, and correctness of the decision (Jannsen et al., 2017). That also means that the consequences of the decision and the time available must be considered as elements. Jeff Bezos from Amazon is thought to have said that all his decisions are evaluated on "Is it possible to reverse the decision if we find it to be erroneous?". He calls them Type 1 and Type 2 decisions. Type 1 decisions are almost impossible to reverse and require careful analysis and planning. Type 2 decisions are relatively easy to reverse and can be made quickly and changed if needed. He advises avoiding using too much energy on Type 2 decisions and using intuition and experimentation instead. Experimentation is one type of decision analysis that we will cover later, in Chapter 7 where different types of analysis are covered.

When talking about the origin of DDDM we also need to touch upon the term "**bounded rationality**". That means that our rational decisions are bounded, or limited, by what we know. We will be using data and analytics to expand those borders and therefore be able to make rational decisions in more areas than without the data to help and guide us.

Classical decision theories

Classical decision theory assumes that decision-makers have complete and consistent preferences for overall outcomes and that they can assign precise probabilities to all uncertain events. Some of the main contributors to classical decision theory are Daniel Bernoulli, Adam Smith, John von Neumann, and Leonard Savage. Herbert Simon, Daniel Kahneman, Henry Mintzberg and more later concluded that some assumptions needed to be reconsidered. Table 1.1 provides an overview of some of the major milestones within decision theory.

Until Herbert Simon's groundbreaking introduction of the "Administrative Behavior (1947)" and concepts of "satisfying" and "bounded reality" decision theory was thought of as being a normative/mathematical exercise that could/should be optimized to find a single

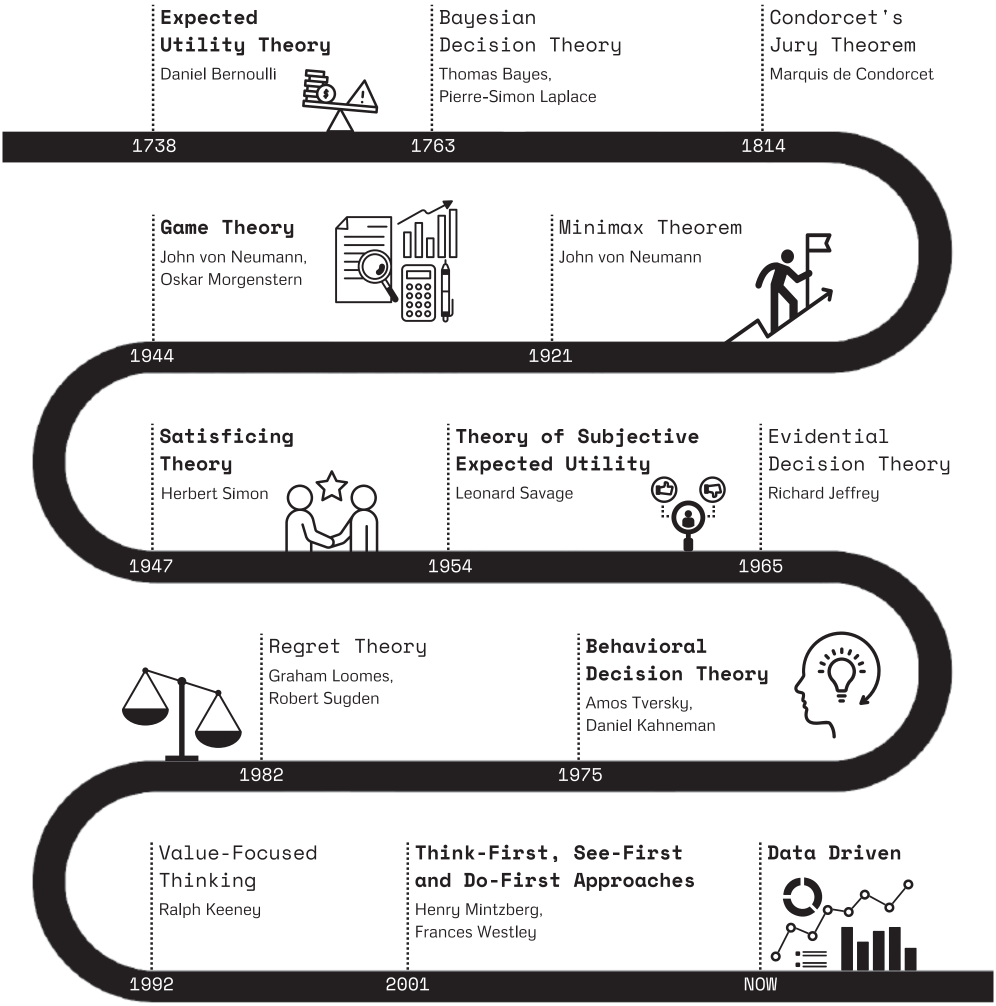


FIGURE 1.1 Timeline on decision theory

correct decision, this contrasts with the "economic man" concepts and rationality of Adam Smith in his works "The Wealth of Nations (1776)" and "The Theory of Moral Sentiment (1759)" as well as the thoughts of Daniel Bernoulli.

Now it is more of a descriptive field that relies heavily upon psychology and neuroscience. In the following sections, some of the theories from the table will be expanded upon.

Expected utility theory was a response to what Bernoulli called the St. Petersburg lottery problem. The idea is a coin is tossed and every time heads are the result the payout will be doubled. The player will have to pay an amount to enter the game and will get the amount accumulated according to the number of "heads" tossed in a streak. How much would you pay to enter the game if we start with 2$ and double for each head shown? From a purely

TABLE 1.1 History of decision-making theory

Year	Theory	Author(s)	Category	Description
1738	**Expected utility theory**	Daniel Bernoulli	Decision-maker	A theory that explains how rational agents should make choices under risk, by maximizing the expected value of their utility function.
1763	Bayesian decision theory	Thomas Bayes and Pierre-Simon Laplace	Decision-making process	A theory that applies the principles of probability and statistics to decision-making under uncertainty, by updating beliefs based on new evidence and choosing the action with the highest expected utility.
1814	Condorcet's jury theorem	Marquis de Condorcet	Decision (type)	A theorem that shows that under certain conditions, the probability of a correct decision by a group of individuals is greater than the probability of a correct decision by any single individual.
1921	Minimax theorem	John von Neumann	Decision (type)	A theorem that proves the existence of optimal strategies for zero-sum games, where one player's gain is another player's loss. The theorem states that the optimal strategy for each player is to minimize their maximum possible loss.
1944	**Game theory**	John von Neumann, Oskar Morgenstern	Decision (type)	A theory that studies strategic interactions among rational agents, using mathematical models to analyze situations of conflict and cooperation.
1947	**Satisficing theory**	Herbert Simon	Decision-maker	A theory that proposes that agents often settle for satisfactory rather than optimal solutions, due to bounded rationality and cognitive limitations.
1954	**Theory of subjective expected utility**	Leonard Savage	Decision-making process	A theory that assumes that a rational decision-maker has a subjective probability distribution over the possible states of the world and a utility function over the possible consequences of her actions. The theory proposes a set of axioms that characterize such preferences and imply the existence and uniqueness of a subjective expected utility representation.
1965	Evidential decision theory	Richard Jeffrey	Decision-making process	A theory that generalizes Savage's theory by allowing preferences to depend not only on the outcomes but also on the evidence or information that is revealed by the choice. The theory proposes a different set of axioms that characterize such preferences and imply the existence and uniqueness of a Jeffrey expected utility representation. The theory also allows for updating preferences by any new evidence.
1967	**The effective decision**	Drucker	Decision-making process	The effective executive will have to manage his employees and delegate decisions based on how his organization is set up and the type of tasks that they need to do.

(*Continued*)

TABLE 1.1 (Continued)

Year	Theory	Author(s)	Category	Description
1975	**Behavioral decision theory**	Amos Tversky, Daniel Kahneman	Decision-maker	A theory that incorporates psychological insights into decision-making, such as heuristics, biases, framing effects, and prospect theory.
1982	Regret theory	Graham Loomes, Robert Sugden	Decision-maker	A theory that accounts for the emotional impact of outcomes on decision-making, by comparing the actual outcome with the best possible outcome in each choice situation.
1992	Value-focused thinking	Ralph Keeney	Decision-making process	A theory that advocates for identifying and structuring values before generating and evaluating alternatives, to make more creative and effective decisions.
2001	**Think-first, see-first, and do-first approaches**	Henry Mintzberg, Frances Westley	Decision-making process	A framework that distinguishes between three ways of making decisions based on the dominant mode of cognition: thinking, seeing, or doing. Think-first is the rational approach that follows a logical sequence of steps. See-first is the creative approach that relies on visual imagery and intuition. Do-first is the experimental approach that involves learning by doing.

mathematical point, any amount would be worth it as there is the possibility of an infinite payout, but most people would stop way before.

> The determination of the value of an item must not be based on the price, but rather on the utility it yields … There is no doubt that a gain of one thousand ducats (red. currency) is more significant to the pauper than to a rich man though both gain the same amount.
>
> (Daniel Bernoulli, 1738)

In other words that means you will only pay in as much as you can afford to lose even though the payout could be much higher.

One of the most important assumptions in "expected utility theory" is that you don't consider other actors' actions in your decision, so the "all things equal" assumption is very strong and must be considered when working from this perspective.

When thinking about this from a data and analytics point of view, you would only invest as much as you are certain to get and the only thing that changes is your decision. That also means that more data and better analysis will allow you to invest more wisely and be better prepared to make the "right" decision.

Von Neuman introduced **game theory** in the first half of the 20th century to expand the expected utility theory to multiple actors and their interactions. That led to the introduction of the concepts of cooperative and non-cooperative games, equilibrium points, dominant and dominated strategies, mixed and pure strategies, and various solution concepts for different types of games.

Game theory also extends expected utility theory to situations where the agents have incomplete or asymmetric information about each other's preferences, beliefs, or actions. For example, game theory studies signaling games, screening games, Bayesian games, and mechanism design. Expected utility theory assumes that the agent has complete and perfect information about the state of the world and the consequences of their actions.

Game theory challenges some of the assumptions and implications of expected utility theory, such as risk neutrality, transitivity, independence, and Bayesian updating. For example, game theory shows that some paradoxes and anomalies that violate expected utility theory can be explained by strategic behavior or psychological factors. Game theory is also linked to alternative models of decision-making under uncertainty, such as regret theory (Loomes & Sugden), prospect theory (Kahneman), and bounded rationality (Simon).

The **theory of subjective expected utility** is an extension of the expected utility theory that Savage proposes in his book *Foundation of Statistics* (1954). He recognizes that not all decisions can be made under the full knowledge assumption as defined by the classic expected utility theory. He goes on to define choice-based subjective probabilities as the foundation for making the choices. He thereby bridges the gap to the Bayesian ideas of updating your decisions as more information becomes available. These ideas were then expanded and generalized by Jeffries in his evidential decision theory.

These theories, that are focused on rational decisions, are the foundation for the data-driven decision paradigm as we with data will try to approach the "full knowledge" and provide an objective and correct decision base.

As seen in the list, the theories started with a mathematical foundation and it was mathematicians that tried to come to grips with the optimization of decision-making. We also therefore have a natural extension into the data-driven decision domain, where mathematics and statistics are employed to improve decision-making. The underlying assumption is, however, still that *there is a correct decision to be made*.

With the publishing of *Administrative Behavior* and the **satisficing theory** in 1947, Herbert Simon laid the foundation for a new direction in decision theory. Where the focus until then had been on reaching the optimal solution, he focused on reaching a decision/solution that would reach a minimum satisfiable option.

One of the classic books on decision-making is *The Effective Decision* by Peter Drucker, published in 1967. In this book, he presents a systematic approach to making effective decisions in various contexts and situations.

Drucker defines a decision as a judgment and a choice between alternatives. He argues that effective decisions are not based on speed, technique, or cleverness, but on conceptual understanding, impact, and soundness. He also distinguishes between decisions based on principle and decisions based on pragmatism, and between right and wrong compromises.

Drucker outlines six steps for making effective decisions:

1 Classify the problem: Determine whether the problem is generic, unique, or new. Generic problems require a rule or a policy to solve them. Unique problems require a one-time solution. New problems require a new rule or policy to be developed.
2 Define the problem: Identify what the problem is and what it is not. Use facts and data to clarify the situation and avoid assumptions and opinions. Focus on the root cause and not the symptoms of the problem.
3 Specify the answer: Define the criteria and objectives for a satisfactory solution. Specify the boundary conditions, such as time, budget, quality, and resources. State what the solution must do and what it must not do.
4 Decide what is right: Choose the best alternative that meets the criteria and objectives, without compromising on principles or values. Consider the consequences and risks of each alternative. Use logic and intuition to make the final judgment.
5 Build action into the decision: Plan how to implement the decision, who will be involved, and what resources are needed. Assign responsibilities and accountabilities to the people who will carry out the decision. Communicate the decision clearly and convincingly to all stakeholders.
6 Test the decision: Monitor the results and feedback of the decision and adjust or revise as necessary. Compare the actual outcomes with the expected outcomes. Learn from experience and feedback and improve future decisions.

Drucker emphasizes the importance of learning from experience and feedback, and of developing a clear vision and purpose for decision-making. He also warns against common pitfalls and challenges that can hinder effective decisions, such as complexity, uncertainty, ambiguity, and resistance.

Tversky and Kahneman came along with their **behavioral decision theory**, which challenged the traditional economic assumption that people make rational choices based on their self-interest. They showed that people often fail to fully analyze situations where they have to make complex judgments, and instead rely on heuristics and biases that can lead to suboptimal decisions. One of their most influential contributions was the **prospect theory**, which describes how people asymmetrically evaluate their losses and gains. According to this theory, people are more sensitive to losses than to gains of the same magnitude, a phenomenon known as loss aversion. They also tend to overweight low probabilities and underweight high probabilities, which can explain why they buy insurance and gamble. Prospect theory also suggests that people's choices are influenced by how a given situation is framed, such as whether it involves a sure gain or a sure loss. Tversky and Kahneman's work has had a profound impact on various fields, such as economics, finance, psychology, and management. Their findings have implications for understanding consumer behavior, risk perception, negotiation, and decision-making under uncertainty.

The last "classic" model we will cover before moving onto the specific data and analytics-focused models was proposed by Mintzberg and Westley in 2001. The **think-first, see-first, and do-first approaches** are not strictly a decision theory, but they propose three ways of deciding that each has to be supported in different ways by data and analytics. Think-first is classic full-scale analysis before making a decision, which could lead one to think that this is similar to DDDM. It doesn't however take the creative/explorative parts of analytics into account. Do-first is the approach where you act first and then adjust. It borrows from the Bayesian thoughts by Thomas Bayes and its need to be adaptable to new information and the need to act quickly. This is similar to running experiments in data analysis, or often just A/B testing where two variations of a campaign are launched and the best performing is the one that "wins" after an initial trial period. This could then lead to thinking that the see-first approach might not be relevant for a data-driven approach. However, in the data analysis, we have two major categories, explorative and explanative (further described in Chapter 7 Data Analysis). The explorative is an inherently creative process where insights are not given a priori, and the decisions come through an explorative process. This means that all the "decision modes" that Mintzberg and Westley propose can be data driven, even though they are not thought of as such from the beginning, as the data and analytics tools were yet to be developed when they developed their model.

The DECAS model

Elgendy et al. (2022) acknowledged that the world had changed with the exponential growth of stored and accessible data as well as tools for analyzing them. The **DECAS model** is based on three core claims.

First, (big) data and analytics (machine) should be considered separate elements in decision-making, rather than being conflated or ignored. Data refers to the raw or processed information that can be used for analysis, while analytics refers to the methods and tools that can extract insights and patterns from data. Both data and analytics have their characteristics, quality, limitations, and implications that affect decision-making.

Second, collaboration between the (human) decision-maker and the analytics (machine) can result in collaborative rationality, extending beyond the classically defined bounded rationality. Bounded rationality is the idea that human decision-makers are limited by their cognitive abilities, information availability, and time constraints, and therefore they cannot always make optimal decisions. Collaborative rationality is the idea that human decision-makers can leverage the complementary strengths of analytics (machine), such as speed, accuracy, scalability, and objectivity, to overcome some of their limitations and biases, and to make more informed and possibly better decisions.

Third, meaningful integration of the classical decision-making elements with data and analytics can lead to more informed, and better, decisions. The classical decision-making elements are the decision-making process, the decision-maker, and the decision. The decision-making process is the sequence of steps or activities that lead to a decision. The decision-maker is the person or entity who makes or influences the decision. The decision is the outcome or choice that is made among alternatives. The DECAS model suggests that data and analytics can play different roles and have different impacts on each of these elements, depending on the context and the type of decision.

The **decision-making process** in organizations comes in two flavors – a structured and unstructured approach to making decisions. Simon's and Drucker's involve sequential steps of gathering data, analyzing alternatives, choosing the best option, and implementing it. However, some decisions are not linear or predefined, and require an iterative process of defining, diagnosing, designing, and deciding as described by Mintzberg and others.

The concept of bounded rationality means that the **decision-makers** are not fully rational, because they face limitations in their information, computation, and environment. It means that decision-makers choose the first satisfactory solution rather than the optimal one. The decision-maker part of the DECAS does not assume that decision-makers know all the alternatives and their consequences. They are bound by human perception and cognition.

Due to their bounded rationality, decision-makers use a simplified model of rationality to make satisficing decisions, rather than optimal ones. It also distinguishes between small worlds and large worlds, where the classical model of rationality is not applicable and different approaches are needed. The factors that affect the decision quality are illustrated in Figure 1.3.

That leads to what is termed collaborative rationality. It was originally termed as the collaboration between multiple people to reach a better decision (e.g. Condorcet's Jury Theorem (1814)) but is not extended to be the collaboration between humans and computers (data and analytics).

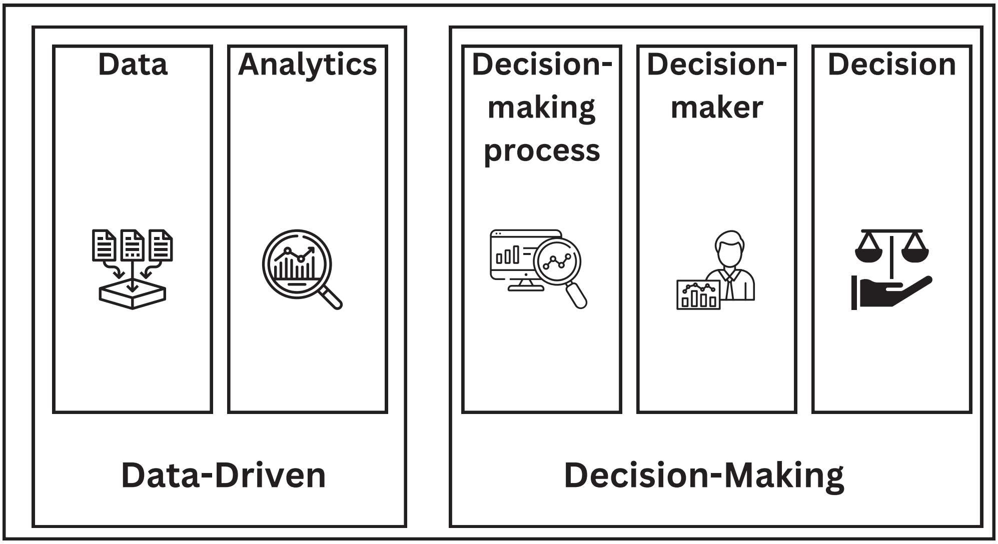


FIGURE 1.2 Elements of data-driven decision-making

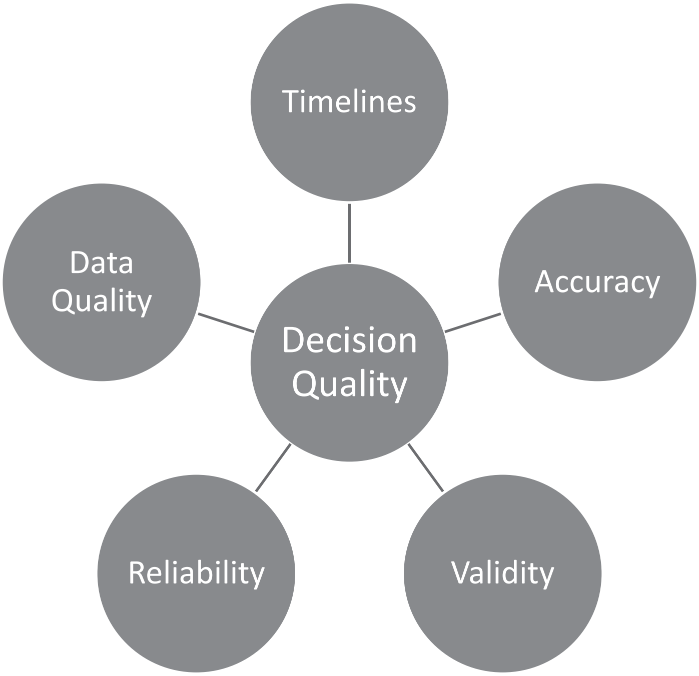


FIGURE 1.3 Decision quality elements

The notion of collaboration in the decision-making between humans and data and analytics has also been explored in the HBR article "What will working with AI really require" (Mohammad Hossein Jarrahi, 2023) where they explore how to work with AI effectively and collaboratively, and how to balance the competitive and cooperative skills of humans and AI. It argues that humans and AI systems need to develop data-centric skills, such as understanding, validating, and visualizing data results; AI literacy, such as explaining how algorithms work and how to communicate with them; emotional intelligence, such as recognizing and reflecting on one's own and others' emotions; holistic and strategic thinking, such as considering the big picture and the ethical implications of AI decisions; creativity and outside-the-box thinking, such as using algorithms in novel and innovative ways; and critical and ethical thinking, such as assessing and addressing the risks and responsibilities of AI systems. It also argues that AI systems need to improve their explainability**,** adaptability, personalization, and context awareness skills, such as providing clear and understandable explanations of their processes and results, learning from previous interactions and personalizing responses, understanding the context in which an interaction is taking place, and offering relevant and personalized content. The article concludes that organizations should democratize data to foster the continuous development of competitive human and machine skills, look outside their own organization's walls for cooperative human skills, and not let geography limit the skills they are hiring for. By doing so, organizations can reap the benefits of an infinity loop between AI and human competitive skills, where humans can leverage both the partnership with machines and their own competitive edge against the machine.

Alternative DDDM framework

The DECAS model extends on classical decision theory, but others have tried to provide entirely new frameworks.

Elisabeth Teal (2011) combines the creative, psychological, and managerial contexts. That shows where data and analytics can fit and enhance the decision-making process. Alternative solutions in the creative context, bias handling in the psychological context, and analytics or decisions support systems in the managerial context.

Barbara Wixom (2023) takes an approach that focuses on data monetization and thereby suggests a framework of value-creation that bases decisions on data and skills. Those can be used as an extension to the DECAS model's data and analytics elements.

COGNITIVE BIAS AND DECISION-MAKING STYLES

Cognitive biases are systematic patterns of deviation from rationality or normative judgment. People use mental shortcuts or heuristics to process information and make decisions, often unconsciously and automatically. Cognitive biases can affect various aspects of human cognition, such as perception, memory, reasoning, and behavior.

The term cognitive bias was first coined in the 1970s by Israeli psychologists Amos Tversky and Daniel Kahneman, who used this phrase to describe people's flawed patterns of thinking in response to judgment and decision problems.[1] They were inspired by Herbert Simon's principle of bounded rationality, which addressed the specific constraints faced by agents in their environments, such as limited time, information, and cognitive capacity.[2] They also drew on research on perceptual biases, which showed that the human perceptual system was prone to errors and illusions due to the imperfect cues from the external world.[3]

Cognitive biases can be classified into three broad categories, based on the cognitive processes they affect:

- **Social biases:** These are biases that affect how people interact with others and interpret social situations. They are influenced by factors such as expectations, stereotypes, emotions, and motivations. For example, managers may exhibit confirmation bias when they seek out or interpret information that supports their existing beliefs or hypotheses about their employees or customers, and ignore or discount information that contradicts them.[4]
- **Memory biases:** These are biases that affect how people encode, store, retrieve, and recall information. They are influenced by factors such as attention, salience, familiarity, and consistency. For example, managers may exhibit availability bias when they judge the frequency or probability of events based on how easily they can recall or generate examples from memory, rather than on the actual evidence or statistics.
- **Decision-making and behavioral biases:** These are biases that affect how people form beliefs, evaluate evidence, make choices, and act on them. They are influenced by factors such as anchors, frames, availability, and risk. For example, managers may exhibit a framing effect when they perceive and evaluate information differently depending on the way it is presented, such as positive or negative, emphasized or de-emphasized, contextualized, or isolated.

Cognitive biases can have significant implications for data-driven decision-making for business, as they can lead to errors and inefficiencies in various domains, such as marketing, sales, negotiation, forecasting, valuation, and risk management. To identify and overcome cognitive biases in DDDM for business, decision-makers should:

- Be aware of their own cognitive biases and how they might influence their judgments and decisions
- Seek out diverse and credible sources of data and information that can provide a more comprehensive and accurate picture of the situation
- Consider multiple factors and criteria that can affect the frequency, probability, or importance of events or outcomes, rather than relying on a single factor or criterion
- Evaluate the quality and relevance of the data and information objectively and critically, without being influenced by prior preferences or expectations
- Update their judgments and decisions based on new data and feedback, and admit when they are wrong.

Some examples of how cognitive biases can be used positively in different domains:

- **HR:** HR managers can use the halo effect to enhance the attractiveness and reputation of their organization by highlighting its positive attributes and achievements, and attracting more qualified and motivated candidates. They can also use the similarity bias to foster a sense of belonging and cohesion among employees by emphasizing their common values and goals and reducing conflicts and turnover.
- **Marketing:** Marketing managers can use the scarcity bias to increase the demand and value of their products or services by creating a sense of urgency and exclusivity, and inducing customers to act quickly before they miss out. They can also use social proof bias to influence customers' preferences and choices by showing them how their products or services are popular and approved by others, such as experts, celebrities, or peers.
- **Negotiation:** Negotiation managers can use the anchoring bias to gain an advantage in bargaining by setting a high initial offer or a low initial demand, and making the other party adjust from that point. They can also use the contrast effect to make their offer or demand seem more reasonable and acceptable by presenting it after a more extreme or unfavorable one.
- **General people management:** People managers can use the optimism bias to motivate and inspire their employees by setting high but realistic expectations and goals, and providing positive feedback and recognition. They can also use confirmation bias to reinforce their employees' confidence and self-efficacy by providing them with information and evidence that supports their abilities and achievements.
- **Project management:** Project managers can use the planning failure probability analysis to avoid overconfidence and underestimation in their project planning by considering the best-case, worst-case, and most likely scenarios, and using historical data and expert opinions. They can also use the hindsight bias to learn from their project outcomes by analyzing what went well and what went wrong and applying the lessons learned to future projects.
- **Risk management:** Risk managers can use the availability bias to assess and prioritize the risks that are most likely and impactful by using relevant data and statistics, rather than

relying on intuition or memory. They can also use the loss aversion bias to encourage risk mitigation and prevention by emphasizing the potential losses and costs of inaction, rather than the potential gains and benefits of action.

Confirmation bias

We notice what confirms our prior belief.

Confirmation bias is a cognitive bias that affects how people search for, interpret, and remember information. It causes people to favor information that confirms their existing beliefs or desires, and to ignore or discount information that contradicts them. Confirmation bias can influence many aspects of decision-making, such as problem identification, hypothesis testing, evidence evaluation, and judgment formation.

Confirmation bias is also a broad category that encompasses several related biases, such as:

- Selective exposure: The tendency to seek out or pay attention to information that supports one's beliefs and avoid or disregard information that challenges them.
- Confirmation search: The tendency to look for or interpret evidence in ways that confirm one's hypotheses and expectations.
- Biased assimilation: The tendency to accept evidence that supports one's beliefs more readily than evidence that opposes them.
- Belief perseverance: The tendency to maintain one's beliefs even after they have been discredited by new evidence or logical arguments.
- Polarization: The tendency for people with opposing views to become more extreme in their opinions after being exposed to conflicting information.

Confirmation bias can have significant implications for data-driven decision-making, as it can lead to errors in reasoning, faulty assumptions, inaccurate predictions, and poor outcomes. To reduce the effects of confirmation bias, decision-makers should:

- Be aware of their own beliefs and assumptions and how they might influence their information processing
- Seek out diverse and credible sources of information that can provide different perspectives and challenge their views
- Consider alternative hypotheses and explanations that can account for the available evidence.
- Evaluate the quality and relevance of the evidence objectively and critically, without being influenced by prior beliefs or expectations
- Update their beliefs and decisions based on new evidence and feedback, and admit when they are wrong.

Confirmation bias can also be used strategically in marketing and sales to influence customers' preferences and choices. For example, marketers can:

- Segment customers based on their existing interests, needs, values, and behaviors, and target them with products or services that match their profiles

- Frame products or services in ways that appeal to customers' beliefs and goals, and highlight the benefits and features that confirm their expectations
- Encourage customers to write positive reviews or testimonials after purchasing a product or service, as this can reinforce their satisfaction and loyalty
- Use the trade-off extension bias to present customers with options that make their preferred choice more attractive and their alternative choice less attractive.

Anchoring bias

Anchoring bias is a cognitive bias that affects how people estimate values, probabilities, or outcomes. It causes people to rely too much on the first piece of information they receive (the anchor) and to adjust their judgments insufficiently from that point. Anchoring bias can influence many aspects of decision-making, such as numerical estimation, forecasting, valuation, and negotiation.

Anchoring bias can occur for various reasons, such as:

- **Insufficient adjustment:** People tend to make small adjustments from the anchor, rather than large ones, due to cognitive laziness or overconfidence.
- **Selective accessibility:** People tend to retrieve or generate information that is consistent with the anchor, rather than information that is inconsistent or contradictory.
- **Focalism:** People tend to focus on the anchor and neglect other relevant factors or information that could affect their judgments.
- **Priming:** People tend to be influenced by the anchor because it activates certain associations or schemas in their memory.

Anchoring bias can have significant implications for data-driven decision-making, as it can lead to errors in calculation, prediction, evaluation, and negotiation. To reduce the effects of anchoring bias, decision-makers should:

- Be aware of their tendencies to be influenced by anchors and how they might affect their judgments
- Seek out diverse and credible sources of information that can provide different reference points and perspectives
- Consider multiple alternatives and scenarios that can account for the uncertainty and variability of the situation
- Evaluate the quality and relevance of the information objectively and critically, without being influenced by prior expectations or anchors
- Update their judgments and decisions based on new information and feedback and admit when they are wrong.

Anchoring bias can also be used strategically in marketing and sales to influence customers' preferences and choices. For example, marketers can:

- Present customers with the most important or attractive information first, as this can create a positive impression and set the standard for the rest of the information

- Highlight the best features or benefits of their products or services and compare them with less favorable alternatives or competitors
- Use price anchoring to make their products or services seem more valuable or affordable by showing a higher initial price or a lower discounted price
- Use the trade-off contrast effect to present customers with options that make their preferred choice more attractive and their alternative choice less attractive.

Availability bias

If we think we have the right or enough information we are unlikely to search for additional information.

Availability bias is a cognitive bias that affects how people judge the frequency, probability, or importance of events. It causes people to base their judgments on the ease with which they can recall or generate examples from memory, rather than on the actual evidence or statistics. Availability bias can influence many aspects of decision-making, such as risk assessment, estimation, prediction, and evaluation.

Availability bias can occur for various reasons, such as:

- **Recency:** People tend to remember and weigh more heavily the information that they have encountered recently, rather than the information that they have encountered earlier or less frequently.
- **Salience:** People tend to remember and weigh more heavily the information that is more vivid, emotional, or distinctive, rather than the information that is more abstract, neutral, or common.
- **Retrievability:** People tend to remember and weigh more heavily the information that is more easily accessible or familiar, rather than the information that is more difficult to access or unfamiliar.
- **Search set:** People tend to remember and weigh more heavily the information that is more compatible or consistent with their existing beliefs or expectations, rather than the information that is more incompatible or inconsistent.

Availability bias can have significant implications for data-driven decision-making, as it can lead to errors in judgment, perception, and reasoning. To reduce the effects of availability bias, decision-makers should:

- Be aware of their tendencies to be influenced by the availability of information and how it might affect their judgments
- Seek out diverse and credible sources of information that can provide a more comprehensive and accurate picture of the situation
- Consider multiple factors and criteria that can affect the frequency, probability, or importance of events, rather than relying on a single factor or criterion
- Evaluate the quality and relevance of the information objectively and critically, without being influenced by prior memories or experiences
- Update their judgments and decisions based on new information and feedback and admit when they are wrong.

Availability bias can also be used strategically in marketing and sales to influence customers' preferences and choices. For example, marketers can:

- Present customers with frequent and consistent information about their products or services, as this can increase their awareness and recall
- Use vivid and emotional appeals to capture customers' attention and interest and highlight the benefits and features that make their products or services stand out
- Use familiarity and repetition to enhance customers' recognition and preference for their products or services and associate them with positive attributes or outcomes
- Use priming and framing to activate certain associations or schemas in customers' memory that can make their products or services more appealing or persuasive.

Framing bias

The context of the information provided is important.

Framing effect is a cognitive bias that affects how people perceive and evaluate information. It causes people to be influenced by the way information is presented, rather than by the content of the information itself. The framing effect can influence many aspects of decision-making, such as preference formation, choice selection, risk assessment, and judgment formation.

Framing effect can occur for various reasons, such as:

- **Valence:** People tend to react differently to information that is presented positively or negatively, such as gains or losses, benefits or costs, advantages or disadvantages.
- **Emphasis:** People tend to pay more attention to information that is highlighted or emphasized, such as bolded or italicized text, numbers or percentages, images, or graphs.
- **Context:** People tend to interpret information differently depending on the surrounding information or situation, such as comparison or contrast, reference point or anchor, goal or expectation.
- **Perspective:** People tend to view information differently depending on their point of view or the source of the information, such as self or others, expert or novice, friend or foe.

The framing effect can have significant implications for data-driven decision-making, as it can lead to errors in perception, interpretation, and evaluation. To reduce the effects of framing, decision-makers should:

- Be aware of their own tendencies to be influenced by the framing of information and how it might affect their decisions
- Seek out diverse and credible sources of information that can provide different ways of presenting and analyzing the same information
- Consider multiple aspects and dimensions of the information that can affect its meaning and value, rather than relying on a single aspect or dimension
- Evaluate the quality and relevance of the information objectively and critically, without being influenced by prior preferences or expectations
- Update their decisions and judgments based on new information and feedback, and admit when they are wrong.

The framing effect can also be used strategically in marketing and sales to influence customers' preferences and choices. For example, marketers can:

- Present customers with information that is framed positively and appealingly, and highlight the benefits and features that match their needs and goals.
- Use emphasis and salience to capture customers' attention and interest, and show them how their products or services can solve their problems or satisfy their desires.
- Use context and comparison to influence customers' perceptions and expectations, and show them how their products or services are superior or unique compared to others.
- Use perspective and source to enhance customers' trust and credibility, and show them how their products or services are endorsed or recommended by experts or peers.

We will look more into how we can positively use cognitive biases in the chapters about creating a data-driven culture (Chapter 4) and data visualization (Chapter 6).

Decision-making styles

There are two widely recognized decision-making style measures: The Decision Style Inventory (DSI) and the General Decision-making Style (GDMS). We will start by going through them and then tie them into the archetypes.

According to Scott and Bruce (1995), the five (GDMS) decision-making styles are: rational, intuitive, dependent, avoidant, and spontaneous.

- A **rational decision-maker** follows a structured process of analysis and evaluation, using logic and facts to support their choice.
- An **intuitive decision-maker** relies on their intuition and gut feelings, trusting their inner voice and personal experience.
- A **dependent decision-maker** seeks advice and input from others, often deferring to someone else's opinion or preference.
- An **avoidant decision-maker** tries to avoid or delay making a decision, often feeling overwhelmed or anxious by the alternatives.
- A **spontaneous decision-maker** acts quickly and impulsively, without much deliberation or planning.

These styles are not mutually exclusive, and people may use different styles depending on the situation or context.

The DSI was developed by Rowe and Boulgarides (1992) and is based on Adizes four decision-making styles: directive, analytical, conceptual, and behavioral.

- A **directive decision-maker** has a task focus, low structure, low ambiguity, and an orientation toward action. They make quick and decisive decisions based on their knowledge and judgment.
- An **analytical decision-maker** has a task focus, high structure, high ambiguity, and an orientation toward analysis. They make logical and rational decisions based on facts and data.

- A **conceptual decision-maker** has a social focus, low structure, high ambiguity, and an orientation toward creativity. They make intuitive and innovative decisions based on their vision and values.
- A **behavioral decision-maker** has a social focus, high structure, low ambiguity, and an orientation toward collaboration. They make participative and empathetic decisions based on the input and feelings for others.

The DSI can help people understand their strengths and weaknesses as decision-makers, as well as appreciate the diversity of decision-making styles in others. The DSI can also help people adapt their style to different situations or contexts that may require different approaches to decision-making.

The DSI and the GDMS are both measures of decision-making styles, but they have some differences in their theoretical background, dimensions, and terminology.

Two of the most widely used instruments for measuring decision-making styles are the Decision Style Inventory (DSI) and the General Decision-making Style (GDMS). However, these instruments have some notable differences in their theoretical foundations, dimensions, and terminologies.

The DSI styles reflect two underlying dimensions: the focus of the decision-maker (either task-oriented or people-oriented) and the tolerance for ambiguity (either low or high). The DSI uses a 10-point scale to assess the degree to which a person uses each style and also identifies the dominant pattern of styles for each person.

The GDMS also reflects two underlying dimensions: The cognitive complexity of the decision-maker (either low or high) and the value orientation of the decision-maker (either task-oriented or people-oriented). The GDMS uses a 5-point scale to assess the degree to which a person uses each style and does not identify any dominant pattern of styles for each person.

The DSI and the GDMS have some similarities in their dimensions, such as the task orientation vs people orientation dimension, but they also have some differences, such as the structure vs cognitive complexity dimension and the orientation vs value orientation dimension. Moreover, the DSI and the GDMS use different terminologies for their styles, with none of the styles having a direct correspondence between the two instruments. For example, the directive style in the DSI is not equivalent to the rational style in the GDMS, although they may share some characteristics. Therefore, it is important to be aware of these differences when using or comparing these instruments for measuring decision-making styles.

There are several available tools online that you can use to assess your DSI or GDMS.

Knowing your decision-making style can help you adapt your style to different situations or contexts that may require different approaches to decision-making. For example, you may need to use a more analytical style when dealing with complex or uncertain problems, or a more behavioral style when dealing with people-related issues.

DSI and GDMS can be linked with DDDM by understanding how each style relates to the use of data and analytics in decision-making. For example:

- A directive decision-maker (DSI) or a rational decision-maker (GDMS) may use data and analytics to support their quick and decisive decisions based on logic and facts
- An analytical decision-maker (DSI) may use data and analytics to conduct a thorough and systematic analysis of complex and uncertain problems

- A conceptual decision-maker (DSI) or an intuitive decision-maker (GDMS) may use data and analytics to generate creative and innovative solutions based on their vision and values
- A behavioral decision-maker (DSI) or a dependent decision-maker (GDMS) may use data and analytics to facilitate participative and empathetic decisions based on the input and feelings of others
- An avoidant decision-maker (GDMS) may use data and analytics to overcome their tendency to postpone or delay decisions due to anxiety or overwhelm
- A spontaneous decision-maker (GDMS) may use data and analytics to balance their impulsiveness and lack of planning with some evidence and rationality.

Many companies are going through a digital transformation these years as that is a prerequisite for becoming data-driven in their decision-making. Philipp Korherr and colleagues came up with four archetypes of managers along with their preferences for management and hence decision style (Korherr et al., 2022). The following table summarizes their findings.

The analytical thinker is an archetype of a manager who has a strong technical focus and expertise in implementing analytical models for decision-making. They are rational, analytical, and detail-oriented. They are early adopters of new technologies and methods, and they often become subject matter experts and advisors in their organization. They lead their employees and projects closely by tracking technical KPIs.

The coach is an archetype of a manager who focuses on the social and human aspects of digital transformation. They are communicative, open, and diplomatic. They care about the well-being of their employees and project members. They facilitate collaboration and cultural understanding among different stakeholders.

TABLE 1.2 Decision archetypes

Archetypes	Characteristics	Ontological categories capabilities	Contribution
Analytical thinker	• Technology interest • Analytic mindset • Early adopter	• KPI-driven leadership • Technical know-how • Sustainable operation	• Exploiting technical limits • Working agile (incremental) • Seeking perfection (solution)
Coach	• Diplomatic interest • Social competence • Situational awareness	• Laissez-faire leadership • Collaboration facilitation • Cultural understanding	• Collaboration with specialists • Actively driving change • Involving external specialists
Guide	• Market/Industry expertise • Extensive network • Trusted advisor	• Authoritarian leadership • Lean management • Setting boundaries	• Challenging alternatives • Active involvement • Project management
Strategist	• Open-minded • Trendsetter & innovator • Thought leader	• Visionary leadership • Fostering personal growth • Enabling teamwork	• Team custody • Thinking out of the box • Incorporating corporate strategy

Source: Adapted from Korherr et al., 2022, Table 3.

The guide is an archetype of a manager who has a lot of experience and expertise in the company and the industry. They are a trusted advisor and a change enabler for digital transformation. They have a broad network and a strong leadership style. They set the strategic direction and guide younger colleagues. They can handle uncertainty and change with confidence.

The strategist is an archetype of a manager who has a strategic, visionary mindset and a passion for innovation. They are creative, open-minded, and resilient. They keep the corporate vision for digital transformation in mind and respond to unexpected events with unconventional methods and tools. They are thought leaders in the organization and visionary leaders for their employees. They enable their employees to be creative and grow personally.

PROCESS FOR DATA-DRIVEN DECISION-MAKING

Decisions are prone to bias as mentioned earlier in this chapter. We should therefore scope a decision-making process that will counter as many of the biases as possible.

The process for making a data-driven decision will of course vary heavily based on the risks involved in making the decision and the time available, as noted in the section about decision-making styles.

Data-driven decision-making types can be categorized in the **Eisenhower matrix** (Covey, 1989) or the **urgent–important matrix**, which prioritizes decisions/tasks based on their urgency and importance.

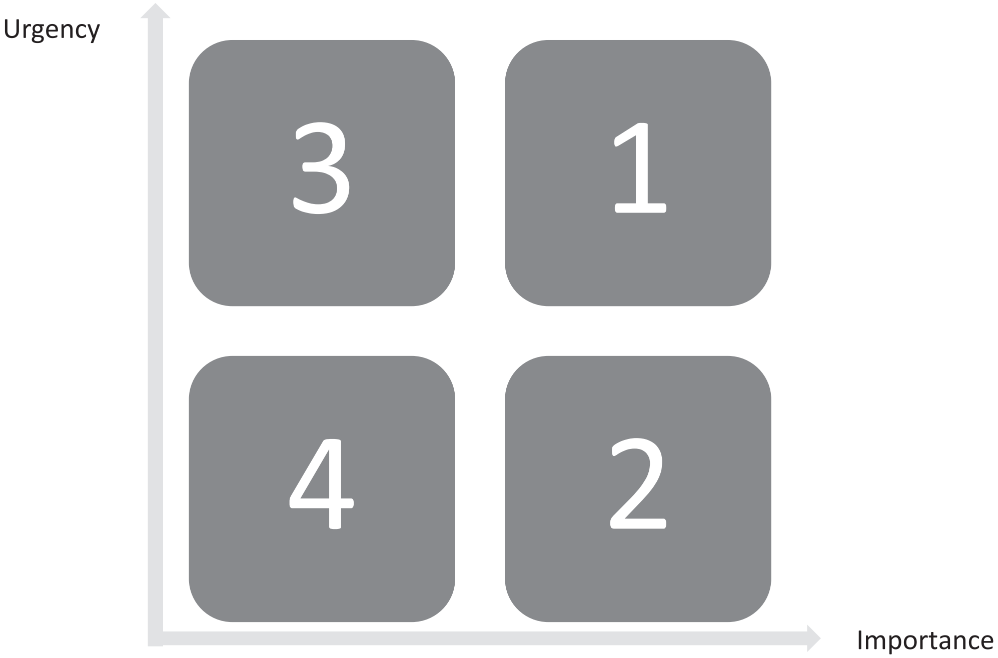


FIGURE 1.4 Eisenhower matrix

The matrix has four quadrants:

- **Quadrant 1**: Urgent and important tasks. These are tasks that need to be done as soon as possible and have a high impact on your goals or outcomes. Examples: crises, deadlines, emergencies.
- **Quadrant 2**: Not urgent but important tasks. These are tasks that do not have a pressing deadline but are still valuable and contribute to your long-term goals or vision. Examples: planning, learning, and relationship building.
- **Quadrant 3**: Urgent but not important tasks. These are tasks that demand your immediate attention but do not have a significant impact on your goals or outcomes. They are often distractions or interruptions. Examples: phone calls, emails, meetings.
- **Quadrant 4**: Not urgent and not important tasks. These are tasks that have little or no value and do not contribute to your goals or outcomes. They are often timewasters or low-priority activities. Examples: entertainment, gossip, browsing the internet.

The idea is to focus on the tasks in Quadrant 1 and Quadrant 2 while minimizing or eliminating the tasks in Quadrant 3 and Quadrant 4. This way, you can make better decisions that align with your priorities and goals.

The mapping of the Eisenhower matrix to data-driven decision-making is shown in Figure 1.5.

The usage of Bayesian experiments in quadrant 1 might sound complicated but it doesn't have to be. At the core, you will try a little bit that you believe is ok, but then make sure that you have ample room for adjusting course/reversing your decision. An example could be that a competitor changes the product's price, which is a direct substitute for your key product. You will have to respond quickly, but how? You do not have data on how the market responds yet, and you know:

1 Great decisions are shaped by consideration of many different viewpoints.
2 Great decisions are made as close as possible to the action.
3 Great decisions address the root cause, not just the symptoms.
4 Great decisions are made by a clearly accountable person.
5 Great decisions consider the holistic impacts of a problem.
6 Great decisions balance short-term and long-term value.
7 Great decisions are communicated well to stakeholders.
8 Great decisions are timely.

(Moore, 2022)

Best practices for implementation

We'll be taking our starting point in the DECAS model in this section, but I'd also recommend going to Chapter 4, which deals with implementing a data-driven culture, if you want further details.

There is no single or definitive way to implement DECAS, as it depends on the specific context and type of decision that is made. However, here are five steps that can help to apply DECAS in a general way:

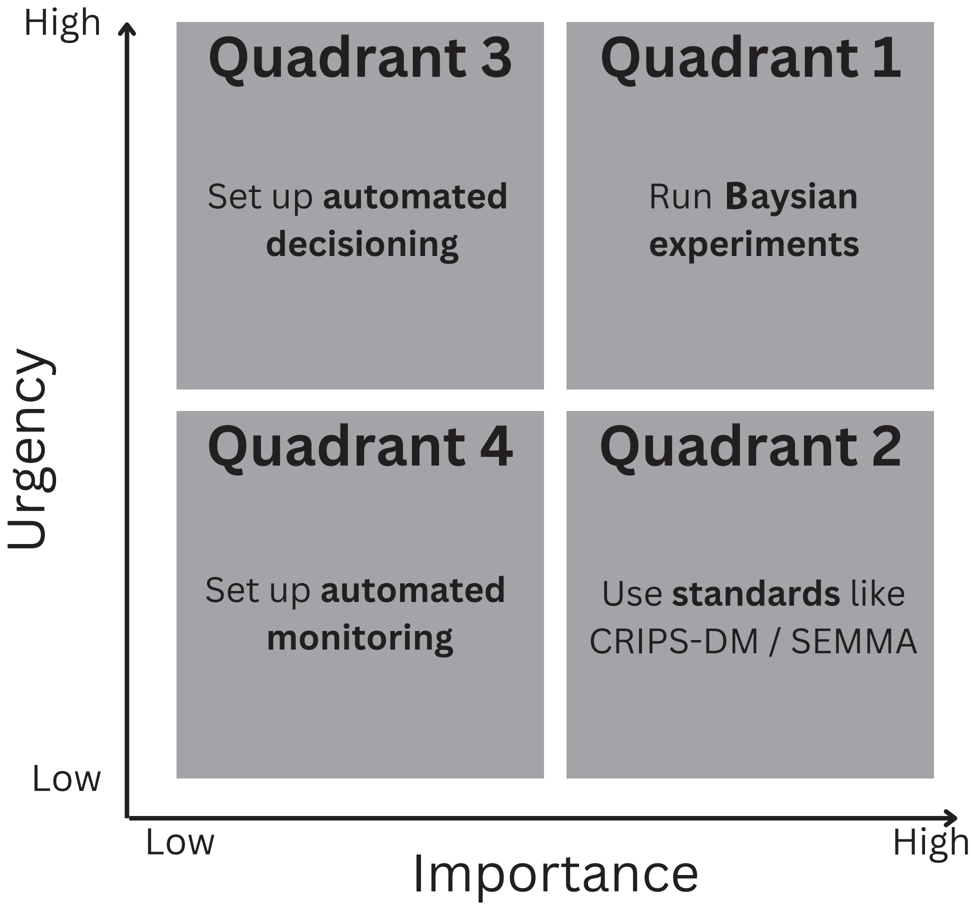


FIGURE 1.5 Data-driven support for important/urgent matrix decisions

Identify and frame the decision problem: What is the goal or objective of the decision? What are the alternatives or options? What are the criteria or factors that matter for the decision? What are the constraints or limitations that affect the decision?

Collect and analyze data and analytics: What are the sources and types of data and analytics that are relevant and available for the decision? How can they be accessed and processed? What are the methods and tools that can be used to analyze them? What are the insights and patterns that can be extracted from them?

Collaborate with data and analytics systems: How can human decision-makers interact with data and analytics systems to obtain information and guidance for the decision? How can human decision-makers evaluate and challenge the data and analytics systems' outputs and recommendations? How can human decision-makers leverage their expertise and judgment to complement the data and analytics systems' strengths and overcome their weaknesses?

Integrate data and analytics with classical decision-making elements: How can data and analytics inform and influence the decision-making process, the decision-maker, and the decision? How can data and analytics be aligned with the goals and values of the decision-makers and the stakeholders? How can data and analytics be subject to ethical standards and regulations? How can data and analytics be transparent and explainable?

Monitor and evaluate data-driven decisions: How can data-driven decisions be measured and assessed for their outcomes and impacts? How can data-driven decisions be revised or corrected if necessary? How can data-driven decisions be responsive to feedback and learning? How can data-driven decisions be open to new data and analytics that may improve them?

Some of the companies that are known to have successfully integrated data-driven decision-making are:

Google: Google is known for its data-driven culture and its use of people analytics to make decisions about leadership development, hiring, promotion, compensation, and team formation. Google also uses data and analytics to improve its products and services, such as search, advertising, maps, and cloud computing.

Amazon: Amazon is another example of a data-driven company that uses data and analytics to drive sales, optimize operations, enhance customer experience, and innovate new offerings. Amazon uses data and analytics to provide personalized recommendations, dynamic pricing, fast delivery, voice assistant, and cloud services.

Southwest Airlines: Southwest Airlines is an example of a data-driven company in the airline industry that uses data and analytics to improve its performance, efficiency, and customer satisfaction. Southwest Airlines uses data and analytics to optimize its routes, schedules, fares, fuel consumption, maintenance, and loyalty program.

CASE

Saxo Bank: Making decisions based on data

Saxo Bank, a pioneering financial institution, has been at the forefront of democratizing the global financial markets since its inception as a bank in 2001. Even before it transitioned into a bank, it operated as a brokerage firm with a unique vision – to provide data and access to information to as many people as possible. This vision was driven by the belief that knowledge is power and that by equipping individuals with the right information, they could make informed investment decisions.

This approach was revolutionary at the time, as it challenged the traditional model where such data and insights were reserved for in-house usage in regular trading companies. By breaking down these barriers, Saxo Bank empowered its customers, giving them the tools and resources they needed to navigate the complex world of financial markets.

Today, Saxo Bank is considered a global leader in the financial industry, renowned for the “Saxo experience”. This experience is characterized by its commitment to transparency, accessibility, and the democratization of financial information. Saxo Bank provides its customers with comprehensive data, sophisticated analysis tools, and real-time market insights, empowering them to make informed investment decisions.

But Saxo Bank's mission doesn't stop there. To continue playing this pivotal role, they are always on the lookout for new ways of supporting the decision-making of their more than 800,000 clients, as well as internally within the organization. This involves staying ahead of industry trends, understanding the evolving needs of their customers, and continuously innovating their products and services.

Saxo Bank invests heavily in technology and human capital to ensure they remain at the forefront of the industry. They foster a culture of learning and growth, encouraging their employees to continually improve their skills and knowledge. This commitment to continuous improvement is a key part of their strategy to provide the best possible service to their customers.

In 2018, Saxo Bank welcomed Graham Sterling to their team. At the time, the bank was starting to feel the pressure of getting the siloed data into the hands of the right people. The data was getting there, but every data set was a unique case that required manual extraction, transformation, and loading into the data warehouse.

Recognizing the inefficiencies and limitations of this approach, Graham Sterling spearheaded the implementation of a new data infrastructure. This involved the adoption of a data mesh architecture, which decentralized data ownership and made it easier for different teams within the organization to access and utilize the data they needed.

This new approach not only improved the speed and efficiency of data delivery but also fostered a more collaborative and data-driven culture within the organization. By breaking down data silos and promoting cross-functional collaboration, Saxo Bank was able to leverage its data more effectively, leading to better decision-making and improved business outcomes.

This case illustrates Saxo Bank's commitment to leveraging data to drive decision-making, both for its customers and within its organization. It highlights the importance of having the right data infrastructure in place and the role of leadership in driving data-driven transformation. It also underscores the value of continuous innovation and adaptation in staying ahead in the rapidly evolving financial industry.

Case questions:

1) Using the DECAS model you are to place the customer decisions in that different section.
2) What kind of cognitive biases should the Saxo Bank experience be countering and how?
3) Using the DECAS model you are to place the Saxo Bank decisions about the data mesh project in that different section.
4) What kind of decision-making style would you estimate Graham Sterling to exhibit and why?

SUMMARY

This chapter explains the concept and benefits of data-driven decision-making, as well as some challenges and best practices for implementing it in organizations. It also covers some common cognitive biases that hinder our data-driven decision-making:

- The history of DDDM
- The benefits and challenges of DDDM
- The DDDM process
- The decision-making styles and links to data.

Data-driven decision-making refers to the systematic collection, analysis, examination, and interpretation of data, usually through the application of analytics or machine learning methods and techniques, to reach informed decisions (Elgendy et al., 2022).

It also links very well with the definition from one of the Tableau (IT company, 2023) where DDDM is defined as using facts, metrics, and data to guide strategic business decisions that align with your goals, objectives, and initiatives. DDDM has been especially popular in education management historically with numerous papers referencing that "industry".

Common for all definitions is that the leader will enhance the decision-making process using data. That does not mean that intuition is of no value, but it will not be standing alone, and you are less likely to fall into the cognitive bias traps.

Data-driven decision-making is not a new concept. It is just that the amounts of data available and the tools to analyze them are recent. That also means that to understand DDDM we should go back a bit and explore how decisions originally are seen to have been made and how the data availability and tools have had an impact on it.

Three different perspectives have historically been the focus when describing decision-making:

- Decision-making process
- Decision-maker
- Decision (type).

This also means that any decision made is dependent on those factors. That is still true when we add data and analytics to the decision, as with the DECAS theory.

The decision-making process is concerned with the structure of the process. There are in general two elements that define how structured a process is. It is the maturity of the organization and the time available to make the decision. That could lead you to think that the more mature the organization is the better the decisions, but it is unfortunately not that simple. Structure takes time and time is not always available. More recently, this structured way has been challenged by, amongst others, Mintzberg and Westley (2001) as decisions cannot always be recognized and will therefore follow a more unstructured process.

The decision-maker, or the one making the final decision, will be bringing his or her personal preference and experiences to the decision-making process.

The cognitive biases have a potentially unwanted impact on the decision-making process. Confirmation is when you notice information that supports your original position. Anchoring bias is when you don't value new information as highly as the first piece. Availability bias is when you stop looking for information when you think you have enough but don't.

Managers can have a preferred decision-making style that either enhances or hinders data-driven decision-making. A **directive decision-maker** has a task focus, low structure, low ambiguity, and an orientation toward action. An **analytical decision-maker** has a task focus, high structure, high ambiguity, and an orientation toward analysis. They make logical and rational decisions based on facts and data. A **conceptual decision-maker** has a social focus, low structure, high ambiguity, and an orientation toward creativity. They make intuitive and innovative decisions based on their vision and values. A **behavioral decision-maker** has a social focus, high structure, low ambiguity, and an orientation toward collaboration. They make participative and empathetic decisions based on the input and feelings of others.

Implementing data-driven decision-making in an organization is not a straightforward process just like any other organizational change process. Taking an analytical approach that identifies decision types and their potential for data analytics support is recommended, as it will show how the process should be in the future. This can be supported by classical change management models like Kotter's 8-step model.

KEY TERMS

Analytics systems: Technological systems that analyze data to provide insights, guidance, and recommendations.

Behavioral decision theory: The theory that states that people are not always making rational choices, but are influenced by numerous biases.

Bounded rationality: The concept that human decision-makers face limitations in information, time, and mental capacity, preventing fully rational decisions.

Classical decision theory: The study of how rational decisions can be made, assuming the decision-maker has full knowledge of alternatives, consequences, and preferences.

Data-driven decision-making: The systematic collection, analysis, examination, and interpretation of data, usually through the application of analytics or machine learning methods and techniques, to reach informed decisions.

DECAS model: Modern data-driven decision theory highlighting the collaboration between humans and analytics.

Decision-maker: Person or entity responsible for making the decision.

Decision-making process: Sequence of activities and steps taken to reach a decision.

Decision: The choice made between alternative options based on criteria.

Expected utility theory: Theory stating that rational agents should choose alternatives that maximize their expected utility.

Game theory: Theory analyzing strategic decision-making when multiple rational parties with potentially conflicting interests interact.

Prospect theory: Theory describing how people evaluate decision outcomes asymmetrically for losses versus gains.

Satisficing theory: Theory stating that decision-makers often settle for satisfactory rather than optimal solutions due to limited rationality.

Theory of subjective expected utility: Expansion of expected utility theory by allowing the agent to make estimations due to lack of full knowledge.

REVIEW QUESTIONS

1 What is the definition of data-driven decision-making covered in the text?
2 What are the three perspectives that have historically described decision-making?
3 Who first coined the term "cognitive bias" and when?
4 What are the three types of cognitive biases?
5 What are the five decision-making styles in the GDMS model?
6 What are the four decision-making styles in the DSI model?
7 Who published the book *The Effective Decision* and when?
8 What are the six steps outlined by Peter Drucker for making effective decisions?
9 What are the three core claims of the DECAS model?
10 What are examples of companies successfully implementing data-driven decision-making?
11 What quadrants are in the Eisenhower matrix linked to data-driven decision-making?
12 What steps are involved in applying the DECAS model?
13 What are some common cognitive biases that can impact data-driven decision-making?
14 How can understanding decision-making styles help with data-driven decision-making?
15 What are some positives of the confirmation bias?
16 What is an example application of the anchoring bias?
17 How can the availability bias be applied positively?
18 What can lead to the framing bias?
19 What are the characteristics of the "strategist" manager archetype?
20 Why is monitoring outcomes important for data-driven decision-making?

Answers to review questions

1. Data-driven decision-making refers to the systematic collection, analysis, examination, and interpretation of data, usually through the application of analytics or machine learning methods and techniques, to reach informed decisions.
2. The decision-making process, the decision-maker, and the decision (type).
3. Israeli psychologists Amos Tversky and Daniel Kahneman in the 1970s.
4. Social biases, memory biases, decision-making, and behavioral biases.
5. Rational, intuitive, dependent, avoidant, and spontaneous.
6. Directive, analytical, conceptual, and behavioral.
7. Peter Drucker in 1967.
8. Classify the problem, define the problem, specify the answer, decide what is right, build action into the decision, test the decision.
9. Data and analytics should not be conflated, collaboration between humans and analytics enables better decisions, and integrating classical elements with data and analytics improves decisions.
10. Google, Amazon, Southwest Airlines.

11. Urgent/Important, Not urgent/Important, Urgent/Not important, Not urgent/Not important.
12. Frame the decision problem, collect/analyze data, collaborate with analytics systems, integrate analytics insights, monitor/evaluate/update decisions.
13. Confirmation bias, anchoring bias, availability bias, framing bias.
14. Links styles to data/analytics use and suggests adapting style by situation.
15. Reinforce employee confidence, enhance organizational attractiveness.
16. Setting a high initial price to increase perceived product value.
17. Present frequent/consistent product info to aid customer recall and preference.
18. Differences in information presentation, emphasis, context, or perspective.
19. Visionary leadership, fostering innovation/growth, incorporating strategy.
20. Enables evaluation, improvements, and responses to feedback.

NOTES

[1] *Encyclopedia of Behavioral Neuroscience*, 2nd edition. https://doi.org/10.1016/B978-0-12-809324-5.24105-9

[2] Cognitive bias. Wikipedia.https://en.wikipedia.org/wiki/Cognitive_bias

[3] Charlotte Ruhl (2023). Cognitive bias: How we are wired to misjudge. simplypsychology.org

[4] Cognitive bias 101: What it is and how to overcome it. health.clevelandclinic.org

BIBLIOGRAPHY

Confluent. (2021). Placing Apache Kafka at the heart of a data revolution at Saxo Bank. https://developer.confluent.io/podcast/placing-apache-kafka-at-the-heart-of-a-data-revolution-at-saxo-bank/ (Accessed January 10, 2024).

Confluent. (2023). Saxo Bank: A data mesh journey with Apache Kafka and Confluent. www.confluent.io/blog/saxo-bank-data-mesh-journey-apache-kafka-confluent/ (Accessed January 10, 2024).

Covey, S. R. (1989). *The 7 habits of highly effective people: Powerful lessons in personal change*. New York: Simon & Schuster.

DECAS: a modern data-driven decision theory for big data. www.tandfonline.com/doi/full/10.1080/12460125.2021.1894674

Drucker, P. F. (1967). *The effective decision*. New York, NY: Harper & Row.

Elgendy, N., Elragal, A., & Päivärintaa, T. (2022). DECAS: A modern data-driven decision theory for big data and analytics. *Journal of Decision Systems*, *31*(4), 337–373.

Graham, C. R. (2014). Data-driven decision-making in the k12 classroom. Academia.edu. www.academia.edu/4146226/Data_driven_decision_making_in_the_k12_classroom

Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality, *Journal of Business Research*, *70*, 338–345, https://doi.org/10.1016/j.jbusres.2016.08.007

Korherr, P., Kanbach, D. K., Kraus, S., & Mikalef, P. D. K. (2022). From intuitive to data-driven decision-making in digital transformation: A framework of prevalent managerial archetypes. *Digital Business*, *2*(2). https://doi.org/10.1016/j.digbus.2022.100045

Mckinsey. (n.d.). How six companies are using technology and data to transform themselves. www.mckinsey.com/capabilities/mckinsey-digital/our-insights/how-six-companies-are-using-technology-and-data-to-transform-themselves

Mintzberg, H. (1990). The manager's job: Folklore and fact. *Harvard business Review*. hbr.org

Mintzberg, H., & Westley, F. (2001, April 15). Decision making: It's not what you think. *MITSloan Management Review*. https://sloanreview.mit.edu/article/decision-making-its-not-what-you-think/

Mohammad Hossein Jarrahi, K. M. (2023, June 8). What will working with AI really require. *Harvard Business Review*. https://hbr.org/2023/06/what-will-working-with-ai-really-require

Moore, M. G. (2022, March 22). How to make great decisions, quickly. *Harvard Business Review*. Retrieved from https://hbr.org/2022/03/how-to-make-great-decisions-quickly

Rowe, A. J., & Boulgarides, J. D. (1992). *Managerial decision making*. New York: Macmillan Publishing.

Savage, L. J. (1954). *The foundations of statistics*. New York: John Wiley.

Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational Research*, *61*(3), 257–273. https://doi.org/10.1080/00131881.2019.1625716

Scott, S. G., & Bruce, R. A. (1995). General decision-making scale (GDMS*). APA PsycTests*. https://doi.org/10.1037/t08399-000

Simon, H. A., & Barnard, C. I. (1947) Administrative behavior: A study of decision-making processes in administrative organization. New York: Macmillan Publishing.

Stobierski, T. (2019). The advantages of data-driven decision-making. *Business Insights*. Harvard Business School Online. https://online.hbs.edu/blog/post/data-driven-decision-making

Tableau. (2023, May 25). data-driven-decision-making. www.tableau.com/learn/articles/data-driven-decision-making

Teal, E. J. (011). Strategic decision making under uncertainty from the foundations of creativity, psychology, and management research: An examination and synthesis. *Journal of Business Administration Online*, *10*(1), www.atu.edu/business/jbao/spring2011/Teal-Creative%20Decision%20Making.pdf

Thoughtworks. (2023). Saxo Bank: Building a self-service data catalog and quality platform. www.thoughtworks.com/case-studies/saxo-bank-data-catalog-quality-platform (Accessed January 10, 2024).

Unscramble. (2021). 6 inspiring examples of data-driven companies (Key . . . – UИSCЯAMBL. https://unscrambl.com/blog/data-driven-companies-examples/

Utika. (2019). Three examples of how companies make data-driven decisions. https://programs.online.utica.edu/resources/article/data-driven-decisions

Wikipedia. (2023, May 25). Data-based decision-making. https://en.wikipedia.org/wiki/Data_based_decision_making

Wixom, B. H., Beath, C. M., & Owens, L. (2023). *Data is everybody's business: The fundamentals of data monetization*. Cambridge, MA: The MIT Press.

CHAPTER 2

Data Strategy

How to Align Data Initiatives with Business Goals and Objectives

This chapter provides a framework for developing a data strategy that supports the business vision, mission, values, and goals. It covers how to identify data needs, data sources, stakeholders, and governance needs.

- Overview of classic and modern strategy frameworks
- Developing a data vision, mission, and values
- Identifying the data needs and potential sources internally and externally
- Defining data stakeholders and governance needed.

LEARNING GOALS:

L2.1 Explain the difference between classic and modern strategy frameworks and how they apply to data initiatives

L2.2 Develop a data vision, mission, and values statement that aligns with the business vision, mission, and values

L2.3 Identify the data needs and potential sources of the business internally and externally using various methods and tools

L2.4 Define the data stakeholders and their roles and responsibilities in the data strategy

L2.5 Establish a data governance framework that covers data quality, security, privacy, ethics, and compliance

L2.6 Evaluate the effectiveness and impact of the data strategy on the business goals and objectives.

Imagine that you are sitting at your kitchen table, and you receive a text message about a party happening later that night. Excitement fills your heart as you realize it's an event you've been looking forward to attending. Without hesitation, you accept the invitation and begin to prepare for the evening.

DOI: 10.4324/9781003457787-2

You hurry into your bedroom to find your "party clothes" at the bottom of the laundry basket. Your "backup" set is clean but wrinkled in the bag of clean clothes that was washed three days ago. You decide to iron the wrinkles and hope for the best. While doing the ironing your friend Emma calls. She offers to pick you up on the way to the party. You are happy to have that problem off your list, but while you are chatting you forget the iron on your shirt, and a burn mark is made on the front.

You are running out of both options and time, so you decide to run to a nearby store to buy a new shirt. They have some good options, but when you check the time, you realize that Emma will be on the way to pick you up. You text her, to ask if she can pick you up at the store, instead thinking that it is only a "small detour". She responds with only an "Ok". A few minutes later another message "Here now, coming out???". You decide to keep the shirt you have on in the dressing room and go to pay at the counter. It all goes quickly, but when you jump into the car you feel the tension building. You are less than 30 minutes late for the party, which is within the time frame you usually arrive, but Emma quickly moves on to talk to some of the other guests.

The party is ok, but as you take a taxi home, the realization sinks in – strategies are not only for business or major life decisions. They are also essential for the little things that shape our everyday experiences. Whether it's organizing our time, prioritizing tasks, or anticipating potential obstacles, having a strategy can save us from unnecessary stress and disappointment.

Also, in the business of data-driven decision-making, everything starts with the strategy. If not, you will be bound to make costly mistakes that in the worst case could bring down the entire company.

INTRODUCTION

Data initiatives are projects or activities that involve collecting, processing, analyzing, or using data to achieve a specific purpose or outcome. Data initiatives can range from simple tasks such as creating a report or dashboard, to complex endeavors such as building a data warehouse or a machine learning model. Data initiatives can also vary in scope, scale, duration, and complexity depending on the data needs and capabilities of the business.

However, not all data initiatives are equally beneficial or meaningful for the business. Some data initiatives may be irrelevant, redundant, or ineffective in addressing the business problems or opportunities. Some data initiatives may be misaligned with the business vision, mission, and values, or may conflict with the interests or expectations of the business stakeholders. Some data initiatives may be inefficient, costly, or risky in terms of data quality, security, privacy, ethics, or compliance.

Therefore, businesses need to align their data initiatives with their business goals and objectives. Business goals and objectives are the desired outcomes or results that the business wants to achieve in each period. Business goals and objectives can be derived from the business vision, mission, and values, which are the overarching statements that define the purpose, direction, and principles of the business. Business goals and objectives can also be influenced by external and internal factors that affect the performance and competitiveness of the business.

Aligning data initiatives with business goals and objectives has several benefits for the business and its stakeholders. First, it ensures that the data initiatives are relevant, valuable, and impactful for the business and its stakeholders. By aligning data initiatives with business goals and objectives, the data team can focus on delivering data solutions that address the most important and urgent business needs and opportunities. By aligning data initiatives with business goals and objectives, the data team can also demonstrate the value and impact of their work and communicate it effectively to the business stakeholders.

Second, it avoids wasting time, money, and resources on data initiatives that are not aligned with the business vision, mission, and values. By aligning data initiatives with business goals and objectives, the data team can prioritize and select the most appropriate and feasible data projects or activities that support the business vision, mission, and values. By aligning data initiatives with business goals and objectives, the data team can also avoid duplication or contradiction of efforts or outcomes with other teams or departments in the business.

Third, it creates a clear and consistent direction and purpose for the data initiatives and the data team. By aligning data initiatives with business goals and objectives, the data team can establish a clear and shared vision, mission, and values for their work and role in the business. By aligning data initiatives with business goals and objectives, the data team can also set clear and measurable targets and milestones for their data projects or activities that guide their planning and execution.

Fourth, it fosters a data-driven culture and mindset in the business that supports evidence-based decision-making and innovation. By aligning data initiatives with business goals and objectives, the data team can encourage and enable the use of data as a strategic asset and a source of insight and value for the business. By aligning data initiatives with business goals and objectives, the data team can also promote and facilitate a culture of learning and experimentation that leverages data to test hypotheses, validate assumptions, optimize processes, improve products or services, and discover new opportunities.

Fifth, it measures and monitors the progress and outcomes of the data initiatives and their contribution to the business goals and objectives. By aligning data initiatives with business goals and objectives, the data team can define and collect relevant and reliable indicators and metrics that track and evaluate the performance and impact of their data solutions. By aligning data initiatives with business goals and objectives, the data team can also review and update their data strategy based on feedback and changing business needs.

CLASSIC AND MODERN STRATEGY FRAMEWORKS

This section reviews some of the classic and modern strategy frameworks that can be used to guide data strategy development, such as Porter's generic strategies, blue/red ocean, and strategic scenarios. It explains the strengths and limitations of each framework and how they can be adapted to data initiatives.

Introduction to strategy (history)

Strategy is a term that has been used for thousands of years in various contexts, such as politics, warfare, business, and sports. Strategy can be defined as a plan of action or policy designed to achieve a major or overall aim. The history of strategy can provide valuable insights and lessons for modern strategic management.

One of the earliest-known discussions of strategy is offered in the Old Testament of the Bible,[1] where Moses delegated authority to other leaders to help him implement his strategies for leading the Hebrews out of Egypt. This shows the importance of creating a hierarchical command structure that allows the top leader to focus on the most critical decisions and delegate the rest.

Another ancient source of wisdom on strategy is Sun Tzu's *The Art of War*,[2] written in 400 BC in China. Sun Tzu emphasized the creative and deceptive aspects of strategy, such as avoiding confrontation with a stronger enemy, exploiting the enemy's weaknesses, and creating surprise and confusion. Sun Tzu also advocated for winning a battle without fighting, which implies finding a unique competitive advantage that makes the enemy surrender or retreat.

The history of strategy also includes many examples of military strategies that have been applied to business contexts, such as Clausewitz's *On War*, written in 1832, which introduced the concept of friction (the difference between plans and reality) and the need to adapt to changing circumstances. Another influential military strategist was Alfred Thayer Mahan, who wrote *The Influence of Sea Power upon History* in 1890, which argued that naval power was the key to national prosperity and security. Mahan's ideas influenced many business leaders who sought to create dominant positions in their industries by controlling key resources and markets.

The history of strategy as a field of study can be traced back to the early 20th century when scholars such as Henri Fayol and Frederick Taylor developed theories of management and organization that emphasized rationality, efficiency, and control. However, it was not until the 1960s and 1970s that strategy emerged as a distinct discipline within management, thanks to the contributions of academics such as Igor Ansoff, Kenneth Andrews, Michael Porter, and others. These scholars developed frameworks and tools for analyzing the external and internal environments of firms, identifying their strengths and weaknesses, formulating their goals and objectives, and choosing alternative courses of action.

Porter's generic strategies

One of the most influential frameworks for strategic analysis is Porter's generic strategies (Porter, 1980), which proposes that firms can achieve competitive advantage by pursuing one of three generic strategies: cost leadership, differentiation, or focus. Cost leadership involves offering products or services at lower prices than competitors while maintaining acceptable levels of quality and features. Differentiation involves offering products or services that are perceived as unique or superior by customers, who are willing to pay a premium price for them. Focus involves targeting a narrow segment of customers with products or services that meet their specific needs or preferences better than those of competitors.

Porter's generic strategies provide a useful starting point for understanding how firms can position themselves in their industries and markets. However, they are not sufficient to explain all the nuances and complexities of strategic management. Therefore, it is important to complement them with other perspectives and tools that can help managers cope with uncertainty, ambiguity, dynamism, and complexity in their strategic environments.

Data strategy support for Porter's generic strategies

A data strategy is a plan that outlines how a firm can use data to achieve its strategic goals and objectives. A data strategy can support each of the three generic strategies proposed by Porter in different ways:

A cost leadership strategy involves offering products or services at lower prices than competitors while maintaining acceptable levels of quality and features. A data strategy can support this strategy by helping the firm reduce its operational costs, improve its efficiency and productivity, optimize its processes and resources, and identify and eliminate waste and errors. This means automated processes and waste identification should be in focus. A simple route optimization in the distribution channel would be an example here. Another could be working on preventive maintenance of production equipment. However, there is always a system cost consideration to be taken into account here.

A differentiation strategy involves offering products or services that are perceived as unique or superior by customers (but still at large scale), who are willing to pay a premium price for them. A data strategy to support this strategy could be focused on enhancing its innovation and creativity, developing new products or services, customizing its offerings to different segments or preferences, and creating a strong brand image and reputation. This means that the focus would be the data initiatives that provide insights into market trends and changes, such as changes in click patterns on company-owned websites and SoMe (social media) channels.

A focus strategy involves targeting a narrow segment of customers with products or services that meet their specific needs or preferences better than those of competitors. A data strategy can support this strategy by helping the firm understand its target market and customers, segment and personalize its marketing and communication, tailor its value proposition and delivery, and build loyalty and trust. An example could be the use of generative AI to tailor messages on an individual level and 3D printing/micro production for custom products.

Data variation of the strategies

In a cost leadership data strategy, you would be making data choices focused on low operating costs, but still collecting vast amounts of data.

The differentiation data strategy will lead you down a specialization path, doing specialized solutions for large parts of the company. That could be a self-service dashboard solution with a training program that enables employees to make clustering analysis

If we choose the focus data strategy there will be initiatives that are either technology-, process-, or department-specific. It could be an anomaly detection solution in our finance department

In many cases, a company would start with a focus strategy and then move on to either a cost leadership or differentiation strategy based on which the company follows in general. One of the key points that Porter makes about his strategies is that you should not mix them, as everything the company does from procurement to development to production and sales should work to enhance the strategy.

Blue ocean strategy

Blue ocean strategy is a strategic framework that aims to create and capture new market spaces and make the competition irrelevant. It is based on the idea that market boundaries and industry structure are not fixed but can be reshaped by the actions and beliefs of industry players. The term "blue ocean" refers to the uncontested market space that offers ample opportunities for growth and profitability, while the term "red ocean" refers to the crowded and competitive market space that is shrinking and commoditized (Kim, 2005).

Concept

The main principles of blue ocean strategy are:

- Pursue both differentiation and low cost. This is called "value innovation", which means creating value for customers and reducing costs for the company. Value innovation breaks the trade-off between quality and price and creates a leap in value for both buyers and sellers.
- Focus on noncustomers rather than customers. This means identifying and reaching out to the people who have not yet been served or are underserved by the existing industry offerings. By understanding their unmet needs and preferences, a company can create new demand and expand the market.
- Reconstruct market boundaries rather than compete within them. This means challenging the conventional wisdom and assumptions that define how an industry operates and competes. By applying a set of analytical tools, such as the strategy canvas and the four actions framework, a company can discover new ways to differentiate itself from the competition and create a blue ocean.
- Align the whole system of activities with the strategic choice. This means ensuring that all the activities of the company, from operations to marketing to human resources, are consistent with and support the value innovation proposition. This creates a coherent and compelling strategy that is hard to imitate by competitors.

Some examples of companies that have successfully applied the blue ocean strategy are Netflix, Cirque du Soleil, Southwest Airlines, and Yellow Tail (ClearPoint Strategy, 2023).

Data strategy support

Data analytics can be a powerful tool to support the blue ocean strategy in various ways (Davenport, 2017). Data analytics can help a company to identify new market spaces and customer segments. By using data sources such as social media, web analytics, surveys, and customer feedback, a company can gain a deeper understanding of the current and potential customers' needs, wants, behaviors, and preferences. This can help a company to discover new opportunities for value innovation and noncustomer attraction.

These new ideas and offerings can then be tested and validated by using data methods such as experiments, simulations, predictive modeling, and optimization, a company can evaluate the feasibility and desirability of new products, services, or business models. This can help a company to reduce risks and uncertainties and enhance customer satisfaction and loyalty.

Finally, monitoring for performance and impact should be set up using data metrics such as key performance indicators (KPIs), balanced scorecards, dashboards, and benchmarks. The company can then track and measure the results and outcomes of its strategy execution. This can help a company to identify strengths and weaknesses, learn from feedback, and make the adjustments needed.

Data variation

The data variation of a blue ocean strategy creates a new and unique data platform and culture. This would also be a source of great competitive advantage. Several companies have come into this category and gone on to monetize the tools that they build for internal purposes. Airbnb (airflow), Lyft (Amundsen), Netflix (data, 2023) and Amazon (AWS) are notable examples of this.

Strategic scenarios

The two previous strategies have a very important assumption that any business leader should be aware of. They are based on the classic assumption from the economic theory of "all other things being equal". The implication in strategy is that you expect the world to continue (mostly) "as is". If the later years have taught us anything, then it is that the world can quickly take an unexpected turn (e.g. the COVID-19 pandemic).

This is the situation that the strategic scenario framework tries to solve. With a foundation in vision and mission, Paul de Ruijter's (2014) multiple potential scenarios and dynamic strategies are built.

Concept

The strategic scenario model by Paul de Ruijter (2014) consists of eight elements that guide the strategic conversation from the present situation to the desired future.[3] The eight elements are:

- Mission: The reason for the organization's existence and its core values.
- Trends: The external factors that influence the organization's environment and shape its future possibilities.
- Scenarios: The plausible and relevant stories of how the future might unfold, based on different combinations of key uncertainties.
- Options: The strategic choices that the organization can make to respond to the scenarios, including both proactive and reactive actions.
- Vision: The preferred direction and destination for the organization, based on its mission and options.
- Roadmap: The plan of action that outlines the steps and milestones to achieve the vision, as well as the indicators to monitor progress and performance.
- Action: The implementation of the roadmap, involving the allocation of resources, the communication of the strategy, and the engagement of stakeholders.
- Monitoring: The evaluation of the strategy, based on feedback from the environment and the organization, and the adjustment of the roadmap and action as needed.

The strategic scenario model is a cyclical and iterative process that requires constant dialogue and learning among the participants. It helps organizations to navigate the future by anticipating change, exploring alternatives, creating alignment, and fostering agility.[4]

One of the benefits of using the strategic scenario model is that it helps organizations to anticipate and prepare for future changes and uncertainties, rather than reacting to them after they happen. By exploring multiple possible futures and challenging existing assumptions and mental models, scenario planning stimulates creativity and innovation. Another benefit is that it enhances strategic alignment and communication by involving diverse stakeholders and perspectives in the strategic conversation and creating a shared vision and roadmap. Furthermore, scenario planning fosters strategic agility and resilience by enabling organizations to identify and seize opportunities, mitigate risks, and adapt to changing conditions.

However, using the strategic scenario model also poses some challenges. One of them is that it requires a significant amount of time, resources, and commitment from the organization and its leaders to conduct a thorough and meaningful scenario-planning process. Another challenge is that it can be difficult to balance the trade-off between complexity and simplicity, as scenarios need to be both plausible and relevant, but also clear and concise. A third challenge is that it can be hard to integrate scenario planning with other strategic tools and processes, such as forecasting, budgeting, and performance management. A fourth challenge is that it can be hard to measure the impact and value of scenario planning, as it is not a predictive tool but a learning tool that aims to improve strategic thinking and decision-making.

Data strategy support

A data strategy is a plan that defines how an organization collects, stores, analyzes, and uses its data to achieve its business objectives (Luther & Ali, 2022). If a company were to be run by strategic scenarios, it would need a data strategy that supports the following aspects (AWS, n.d.; IBM, n.d.):

Data collection: The company would need to collect data from various sources, both internal and external, that are relevant to its scenario-planning process. The data sources should provide information about the current situation, the driving forces, the key uncertainties, and the potential outcomes of different scenarios.

Data storage: The company would need to store the data in a secure, scalable, and accessible way that enables data integration and sharing across different departments and stakeholders. The data storage should also support data quality and governance standards to ensure data accuracy and consistency.

Data analysis: The company would need to analyze the data using various methods and tools to generate insights and scenarios. The data analysis should involve both quantitative and qualitative techniques, such as statistical modeling, machine learning, text mining, visualization, and storytelling. The data analysis should also be iterative and collaborative, allowing for feedback and refinement of the scenarios.

Data usage: The company would need to use the data and the scenarios to inform its strategic decision-making and action planning. The data usage should involve communicating the scenarios and their implications to the relevant audiences, evaluating the strategic options and trade-offs, selecting the preferred vision and roadmap, implementing the action plan, and monitoring the results and feedback.

Data variation

If scenario planning is applied to data-driven decision-making it is focusing on building options into the plans and having a set of end goals that we are targeting against.

The company strategy is for a large part an implementation of the company mission and vision, in the same way the data strategy is an implementation

DATA VISION, MISSION, STRATEGY, AND VALUES

This section explains how to develop a data vision and mission statement that aligns with the business vision and mission. It provides examples and best practices for crafting a clear and compelling data vision, mission, and values statement that communicates the purpose, direction, and principles of the data strategy. Implementation is the strategy which is done through a governance program.

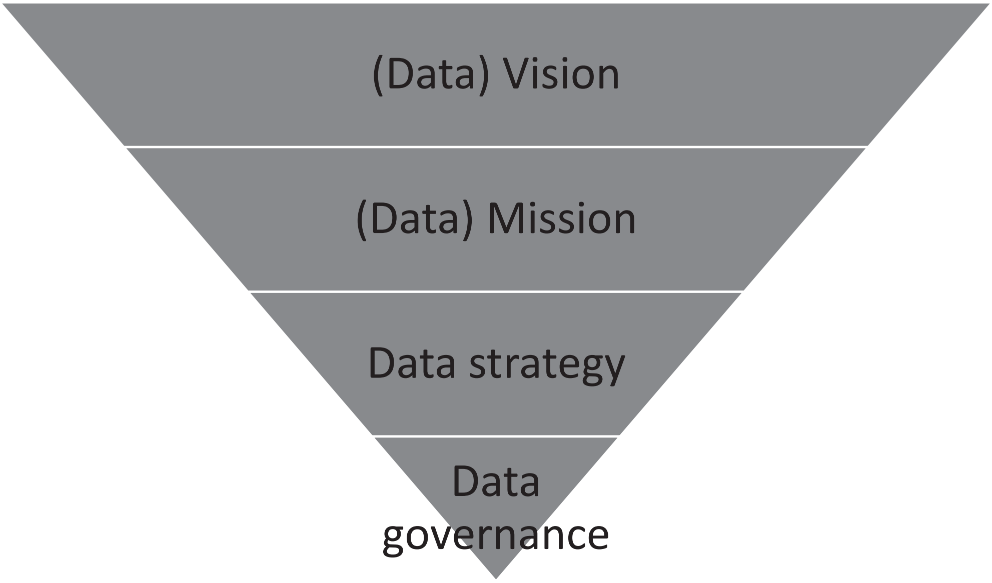


FIGURE 2.1 Hierarchy of vision, mission, strategy, and governance

The data vision and mission must be closely aligned with the company vision. In most cases, it is beneficial to have a dedicated implementation as described below, but it does depend on how easy it is to operationalize the general company vision and mission.

A **data vision** statement describes *why* an organization cares about data and what is the ultimate aspiration for data. It should support and be consistent with the organization's overall vision statement. It should answer questions like "Why do we care about data?" and "Why do the data organization and processes exist?"

A **data mission** statement describes *what* an organization is going to accomplish with data and *how* it is going to do it. It should support and be consistent with the organization's overall mission statement. It should also be the starting point for the data strategy. It should answer questions like "What are we going to achieve with data?" and "How are we going to achieve it?" (Treder, 2020).

Strategy is the implementation of the mission and vision which will be described in the coming sections and the data governance after that.

According to Gartner (n.d.), there are four (4) pillars to an AI strategy. This can be compared to a data strategy in this context. These are vision, risks, value, and adaptation. Vision is already covered in the previous section. The risks, value, and adaptation pillars are dealt with in the coming sections.

IMPLEMENTATION, WHERE TO START

A natural place to start is making a data audit. That means identification of data quality issues, governance gaps, data security issues, integrity of the data that is available, and compatibility between the different sources of data.

It can be done through a traditional project management process that starts with Interviews of both key stakeholders and data users. This should then be supplemented with surveys of surface usability, trust, system incompatibilities, and accessibility of the existing setup. Even if there isn't any formalized setup this should be done to ensure that the processes that act as "workarounds" for employees are captured in the following initiatives. A couple of key questions to use here are:

- What can't you do, that you need to, because of lack of data?
- What would you like to do if you had data available?

The interview and survey are part of the data issue surfacing process. In addition to reviewing documentation, a data council should be established that meets regularly; here escalation of issues and technological reviews are considered. The data council can be compared to a steering committee for a strategic project but is more permanent. Over time, as the organization matures in its use of data to drive decision-making, the topics can be rolled into the normal organizational hierarchy.

When the as-is situation is sufficiently established a strategy can be developed to move the organization towards the company (data) vision and (data) mission. This should fall within one of the previously described paradigms which align across the entire organization.

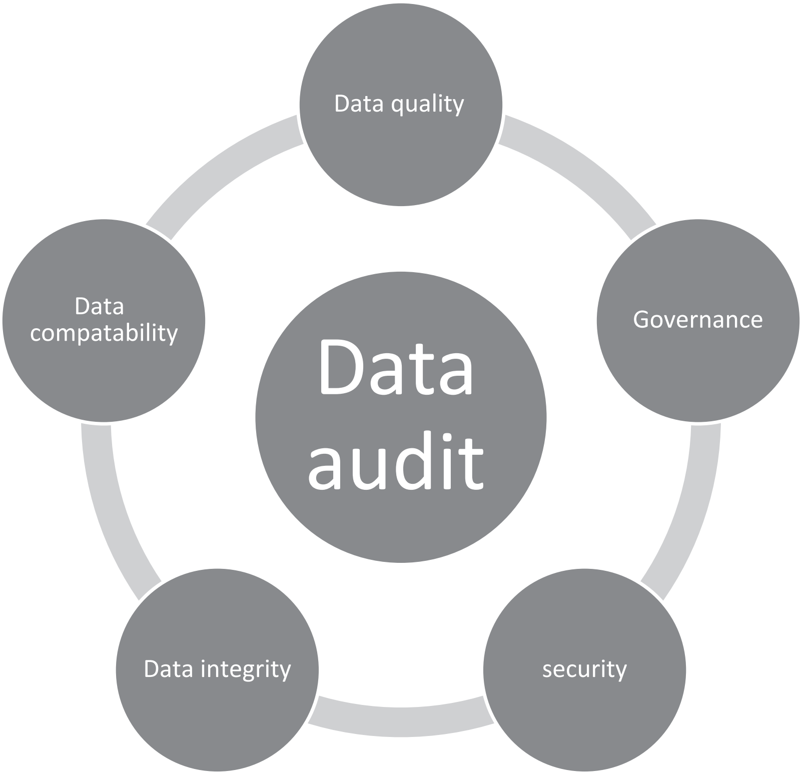


FIGURE 2.2 Data audit components

DATA NEEDS AND POTENTIAL SOURCES

This section covers how to identify the data needs and potential sources of the business internally and externally. It discusses various methods and tools for data discovery, such as interviews, surveys, workshops, data catalogs, etc. It also covers how to assess the availability, quality, and relevance of the data sources and how to prioritize them based on the business goals and objectives.

In rough terms, there are two places where you can get the data for your analysis that leads to data-driven decision-making. Internal data sources are the systems within the company that generate data. This could be financial systems, enterprise resource planning systems, customer relation systems, etc. Sometimes internal data also needs to be generated through surveys and interviews.

External data can also come from surveys and interviews with customers, partners, or other stakeholders. Often, however, it is not that bespoke external data that the company builds decisions on but rather the combination of generic information like statistics from public sources and databases that are procured, but that anyone can buy access to. Chapter 5 is dedicated to this topic alone.

DATA GOVERNANCE

It has, correctly, been said about data governance that "It's a practice, not a project" – and therefore a way of working with data continuously.

According to data from the Data Governance Institute (DGI), governance is:

> a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods.
>
> (Data Governance Institute, 2004)

This section defines the data stakeholders and their roles and responsibilities in the data strategy. It covers how to map the data stakeholders, using tools such as stakeholder analysis matrix, RACI matrix, etc. It also establishes a data governance framework that covers data quality, security, privacy, ethics, and compliance. It discusses how to define data policies, standards, roles, processes, and metrics for data governance. Finally, we will link data governance to DataOps.

Data governance is a critical component of any data-driven strategy and enables data-driven decision-making. Robert Seiner (2023) defines it as the "The execution and enforcement of authority over data".

It provides a systematic approach to:

- Managing data assets
- Ensuring data quality, and
- Enabling appropriate access and
- Usage.

Without proper data governance, organizations run the risk of making decisions based on inaccurate or incomplete data, which can have detrimental consequences.

Implementation of data governance involves defining the roles, responsibilities, and processes for managing data assets. This encompasses various aspects, including data stewardship, data quality, data privacy, and data security. By implementing a well-defined data governance framework, organizations can establish trust in their data and thereby promote a data-driven culture with confidence.

Successful data governance, however, starts with ensuring executive commitment and support. If proper alignment with the business vision, mission, and strategies has been done, this step should be rather easy to achieve. You can check if the different strategy models align in the first part of this chapter.

However, in a recent study by KPMG (KPMG, 2021), 48% responded that their data strategy was out of sync with the business.

Then consider the digital employee experience (DEX) as a key factor for ensuring the adaptation of the governance program. Employees will optimize their working conditions and if the governance hinders that optimization the adaptation rate will go down. This topic will be covered further in Chapter 4 about culture and data literacy.

Data governance (tactic) vs data management (operations)

"Data governance" and "data management" are regularly used indiscriminately which can cause problems when defining and communicating the scope of responsibility.

Data governance is the tactical aspect of the data strategy, which defines the policies for data usage, handling, and storage. It involves the participation of executives, stakeholders, and data trustees who set a framework for data governance in the organization.

Data management is the operational aspect of enterprise data management, which implements the policies for data handling throughout its lifecycle. It covers various domains such as master data management, data quality and security, and database operations. Data custodians and stewards oversee the application and maintenance of the technology and enforce the procedures and policies specified in the data governance framework. (Andrieiev, 2023)

The data management will be covered in further detail in Chapter 4.

The following is a reasonable short stakeholder analysis as the creation of a full stakeholder analysis is also covered in Chapter 6 about visualization and data communication.

Stakeholders in the data governance

In principle, everyone with the company and the ones who get in touch with the company are stakeholders in the data governance. Their roles and their impact can vary significantly over time.

Senior executive leadership: They are the sponsors and champions of data governance, who set the vision, strategy, and priorities for data-driven initiatives (Ladley, 2019). They also allocate resources and resolve conflicts across different data domains (DAMA International, 2017). Often a Chief Data Officer (CDO) is part of this group and has the main responsibility for the strategy, and reporting on the governance status

Data trustees: They are the business or operational leaders who represent their units in the data governance committee. They are accountable for the data definitions, data quality, and compliance with data management policies and standards for their respective data domains (DAMA International, 2017). They will often be the signatories of data contracts that are discussed later in this chapter. Note that these people can also be external to the company if they are providing data that the company depends on.

Data stewards: They are the subject matter experts in their data domains, who support the business unit staff in data management and access (Ladley, 2019). They are often the (data) managers for their areas, who define, document, and maintain the metadata, business rules, and data quality metrics (DAMA International, 2017). The day-to-day responsible employees. Like the data trustees, these people can also be external partners.

Data custodians: They are the IT or system managers who review and grant access rights to the data they oversee. They are responsible for data security, quality, and integrity (Ladley, 2019). They also implement the technical solutions and infrastructure for data storage, processing, and integration (DAMA International, 2017). They are mostly IT infrastructure people as their primary role.

Data consumers: They are the end users of the data, who need it for various purposes such as reporting, analysis, decision-making, or innovation. They include internal staff, external partners, or customers. They rely on the data to be accurate, complete, consistent, and timely (Gartner IT Glossary, n.d.). If governance is not in place they will not be able to trust data that they get, and might get data that put the organization at risk.

Data analysts: They are the specialists who perform data analysis, modeling, visualization, or storytelling. They use various tools and techniques to transform raw data into meaningful insights and recommendations (Gartner IT Glossary, n.d.). They communicate their findings and suggestions to the data consumers or other stakeholders (Ladley, 2019).

Data architects: They are the experts who design and maintain the data architecture of the organization. They define how the data is structured, stored, integrated, and accessed across different systems and platforms (Gartner IT Glossary, n.d.). They ensure that the data architecture is aligned with the business needs and objectives (Ladley, 2019). In the chapter on the modern data stack, this role is elaborated on significantly.

Data engineers: They are the developers who build and optimize the data pipelines and workflows. They use various programming languages and frameworks to extract, transform, load (ETL), or stream data from various sources to various destinations (Gartner IT Glossary, n.d.). They ensure that the data is available, reliable, and scalable (Ladley, 2019).

Data scientists: They apply advanced analytics, machine learning, or artificial intelligence to discover new patterns, trends, or opportunities from the data. They use various statistical methods and algorithms to create predictive models or experiments (Gartner IT Glossary, n.d.). They generate new value or solutions from the data (Ladley, 2019). They need significant volumes of trustworthy data.

Compliance and security experts: They are the professionals who monitor and enforce the compliance and security standards for data governance. They ensure that the data is protected from unauthorized access, use, or disclosure (Gartner IT Glossary, n.d.). They also ensure that the data governance adheres to the relevant laws, regulations, or policies (Ladley, 2019). This regulation on storage and access to personal information is regulated in e.g., the European GDPR legislation.

For each of them, it can make sense to create a RACI matrix for the following common tasks in data governance. A company will rarely have all the positions in their organization as separate job roles but rather have several combined in a few positions.

RACI stands for Responsible, Accountable, Consulted, and Informed, which are the four types of participation that can be assigned to each stakeholder for each data governance activity. A RACI matrix is usually presented as a table or a chart, where the rows represent the data governance activities, and the columns represent the stakeholders. For each cell in the matrix, one of the following symbols or letters is used to indicate the level of participation:

- R: Responsible. This means that the stakeholder is directly involved in performing the activity or delivering the output.
- A: Accountable. This means that the stakeholder has the authority to approve or reject the activity or output and is ultimately answerable for its quality and outcome.

- C: Consulted. This means that the stakeholder provides input or feedback on the activity or output and has some influence on its design or execution.
- I: Informed. This means that the stakeholder is kept updated on the progress or results of the activity or output but has no direct involvement or influence on it.

The benefits of using a RACI matrix for data governance are:

- It provides clarity and transparency on who is responsible for what in data management processes, reducing ambiguity and confusion.
- It enhances collaboration and communication among stakeholders, ensuring that everyone is aware of their roles and expectations and that they are aligned on the goals and objectives of data governance.
- It streamlines decision-making and accountability, ensuring that decisions are made by the appropriate stakeholders, and that they are supported by relevant information and feedback.
- It improves efficiency and effectiveness, ensuring that tasks are assigned to the most suitable stakeholders and that duplication of efforts and conflicts are avoided.
- It supports compliance and quality, ensuring that data governance activities are performed per data policies, standards, and regulations and that data quality issues are identified and resolved.

The challenges of using a RACI matrix for data governance are:

- It requires a clear understanding of the data governance framework, including the data governance activities, stakeholders, and their relationships.
- It requires a high level of commitment and involvement from all stakeholders, especially those who are accountable or responsible for data governance activities.
- It requires regular review and update to reflect changes in data governance requirements, processes, or stakeholders.
- It may not capture all the nuances and complexities of data governance situations, such as multiple levels of responsibility or accountability, or overlapping or conflicting roles.

The following list is an example of tasks that could be listed in the data governance RACI matrix:

- Define data governance vision and strategy (McKinsey, 2023)
- Establish data governance committee and roles (Data.org, 2023)
- Develop data governance policies and standards (McKinsey, 2023)
- Define and plan the scope of the program (AIMultiple, 2023)
- Develop ways to improve data quality and security (AIMultiple, 2023)
- Create and manage metadata (AIMultiple, 2023)
- Evaluate the suitability of new data sources (AIMultiple, 2023)
- Monitor and enforce compliance with data policies (AIMultiple, 2023)
- Prevent unauthorized access to data through access management (DataCamp, 2023)

- Prevent personal or commercially sensitive data being leaked publicly through data masking and encryption (DataCamp, 2023)
- Prevent data being deleted or corrupted through disaster recovery (DataCamp, 2023)
- Implement the architecture and processes to achieve the goals of data governance (NetApp, 2023)
- Encourage improvement and ensure data policies are enforced (NetApp, 2023)
- Support data analysis, modeling, visualization, or storytelling (McKinsey, 2023)
- Communicate the value and benefits of data governance to stakeholders (Data.org, 2023).

Later in the chapter, their daily responsibilities will be expanded upon. However, the first step would be to define the data governance framework.

Data governance framework

The data governance framework outlines the structure and processes for governing data within an organization. It defines the roles and responsibilities of different stakeholders involved in managing data. The framework also establishes policies and procedures for data management, including data quality standards, data classification, and data retention policies.

There are several standards out there. Two of the most common ones are COBIT 5 (Control Objective for Information and Related Technology – 2012) and the Data Governance Institute (DGI) framework (2004). These frameworks provide a comprehensive and holistic approach to data governance and management, covering different aspects and dimensions of data as an asset. They also use a **maturity model** to measure the progress and effectiveness of data governance and management practices.

COBIT 5 is a framework developed by ISACA in 2012 to guide the governance and management of enterprise IT. It is based on five principles: meeting stakeholder needs, covering the enterprise end-to-end, applying a single integrated framework, enabling a holistic approach, and separating governance from management. It defines 37 processes that support the achievement of 17 governance and management objectives. It also provides a capability maturity model to assess the performance of each process (ISACA, 2012).

The **DGI framework** is a framework developed by the Data Governance Institute in 2004 to guide the governance and management of data as an asset. It is based on eight components: data governance organization, data governance strategy, data governance policy, data quality, metadata, data architecture, data security, and data compliance. It defines ten best practices that support the implementation of each component. It also provides a maturity model to assess the level of data governance maturity in an organization (Data Governance Institute, 2004).

The main similarities between COBIT 5 and the DGI framework are that they both aim to provide a comprehensive and holistic approach to data governance and management; they both align with the business goals and stakeholder needs of the organization; they both use a maturity model to measure the progress and effectiveness of data governance and management practices; and they both reference each other and other frameworks to ensure compatibility and integration.

The main differences between COBIT 5 and the the DGI framework are that COBIT 5 covers the entire spectrum of enterprise IT, while the DGI framework focuses specifically on data as an asset; COBIT 5 uses principles, objectives, and processes as the main elements of its framework, while the DGI framework uses components, strategies, and policies as the main elements of its framework; COBIT 5 has more processes (37) than the DGI framework has best practices (10), which may imply that COBIT 5 is more detailed and prescriptive than the DGI framework; COBIT 5 uses the CMMI Institute performance management scheme to measure capability and maturity levels, while the DGI framework uses its own maturity model based on six levels.

In the following sections, DGI will be the focal framework as it is specifically focused on data governance.

The DGI framework provides a comprehensive and holistic approach to data governance, covering eight components that address the WHY, WHAT, WHO, and HOW of data governance. The eight components are:

Data governance organization: This component defines the roles and responsibilities of the data governance participants, such as the data governance office, the data governance council, the data stewards, and the data owners. It also establishes the reporting and escalation mechanisms for data governance issues and decisions.

Data governance strategy: This component defines the vision, mission, goals, objectives, and value proposition of the data governance program. It also aligns the data governance program with the business strategy and stakeholder needs.

Data governance policy: This component defines the high-level, top-down, data-related policies that guide the data governance program. It also translates the policies into operational rules and standards that can be implemented and enforced by the data governance participants.

Data quality: This component defines the criteria and metrics for measuring and improving the quality of data. It also implements processes and tools for data quality assessment, monitoring, reporting, and remediation.

Metadata: This component defines the processes and tools for creating, managing, and sharing metadata. Metadata is the data about data, such as definitions, descriptions, classifications, lineage, relationships, and usage.

Data architecture: This component defines the structure, design, and integration of data across the enterprise. It also ensures that the data architecture supports the business requirements and complies with the data governance policies and standards.

Data security: This component defines the processes and tools for protecting data from unauthorized access, use, disclosure, modification, or destruction. It also ensures that the data security complies with the legal, regulatory, and ethical obligations of the organization.

Data compliance: This component defines the processes and tools for ensuring that data complies with the external and internal rules and regulations that govern its collection, storage, processing, and dissemination. It also monitors and audits the compliance status of data and reports any violations or issues.

The DGI framework also provides ten best practices that support the implementation of each component. These best practices are:

- Define a clear scope for your data governance program
- Establish a cross-functional team to lead your data governance program

- Develop a business case to justify your data governance program
- Assess your current state of data governance maturity
- Define your desired state of data governance maturity
- Identify gaps between your current state and the desired state of data governance maturity
- Prioritize your data governance initiatives based on business value and feasibility.
- Implement your data governance initiatives using a phased approach
- Monitor your data governance performance using KPIs
- Continuously improve your data governance practices based on feedback.

The DGI framework has a maturity model to assess the level of data governance maturity in an organization. The maturity model has six levels:

Each level has specific characteristics that describe how well an organization performs in each component of data governance (Data Governance Institute, 2004). This model is also similar to several other maturity frameworks like the Gartner for data maturity (LightsOnData, n.d.) and the CMMI for (IT) processes (CMMI Institute, n.d.).

The first step to starting a data governance program is to establish a single executive owner. It's best to have one ultimate decision-maker rather than two or three, so decisions can be made quickly, and the project can keep moving forward.

This individual should be accountable to the overall program, with budget authority and the ability to make decisions. The owner doesn't need to be involved in the day-to-day workings of data management and governance, but they must have a vested interest in the pursuit of better data.

They should also be involved with the company's highest level strategic planning efforts and have a broad view of the organization, with a commitment to supporting the program and helping to overcome major roadblocks.

Data management maturity assessment

Where to start can be defined after a data management maturity assessment which shows the areas where the organization's current practices support success, and where they could use improvement. "You need to understand your current state before you can create a roadmap for data maturity advancement." There are quite a few available assessment tools, but Doupe recommends the CMMI Institute's Data Management Maturity Model.

The CMMI model looks at six major categories: data governance, data management strategy, data quality, platform and architecture, data operations, and supporting processes. Each of these six categories is evaluated using a one-to-five-point score. Ideally, a third-party data management consulting company should assess them to provide an unbiased external viewpoint.

Even if you are using the DGI framework, in general the comprehensive framework from COBIT 5 can be beneficial as it will also highlight associated areas, especially within IT.

TABLE 2.1 Data maturity levels

Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
Unaware	Aware	Reactive	Proactive	Managed	Optimized

People in the governance project

The first person to identify is a lead for the data governance program, someone to drive governance and be willing to go down a rabbit hole if it's necessary. This should be the same person who defines and drives the company's data strategy: ideally, the CDO, CIO, or CTO.

It's possible to have a data governance program manager in this position but the most important takeaway is that the program needs a person who is 100% dedicated to data governance.

A three-level data governance program structure is recommended. The top strategic/executive level should consist of two to four C-level executives with wide influence, one of whom serves as the executive sponsor. The second level, tactical/data owners, is composed of three to ten senior managers or directors who are data owners accountable for the data within their domains. Operational/data stewards and subject matter experts with significant organizational knowledge about the company's data make up the third level. These are people who often do data management activities, even if those duties are not explicitly listed in their job description.

Creating a formal data governance committee charter clarifies roles and responsibilities and engenders a sense of responsibility in participants. It provides a forum to discuss progress and make decisions on some of the more challenging conversations regarding data management.

Formalizing the committee via a charter also sets up the purpose and mission of the team, the objectives, how results will be measured, the scope of the committee and what decisions they will be making, how the committee should operate, and how they will make decisions.

Gathering feedback upfront from all members of the data governance committee and incorporating that feedback into the team's charter can generate a tremendous amount of goodwill. Having a formal vote on the charter early in the process allows committee members to put their charter-defined voting process into action from the beginning.

Determine an objective

The next step is to explore why data is important to the company. This process should involve consulting with different people across the organization, and their input will form the data governance objectives for the program.

Answers might range from improving analytics or making more data-driven decisions to creating a sustainable long-term advantage in the marketplace or changing the company's business model. It helps to craft a "future state" vision statement to guide decisions.

The data steward community

Data stewardship involves the responsibilities and activities related to the oversight and management of data assets. Data stewards are individuals or teams responsible for ensuring the quality, integrity, and security of data. They collaborate with data owners, data custodians, and other stakeholders to define and enforce data governance policies and procedures. Data stewards play a crucial role in maintaining data accuracy and consistency throughout its lifecycle

Data governance cannot be successful without data stewards with significant first-hand knowledge of the data, and those employees need a clear understanding of expectations before they are engaged in the process.

Training is critical for data stewards because most of them have day jobs and being a data steward is not their full-time role. Most data stewards will enjoy opportunities to enhance their skills and become more effective, and the investment will show them that their work is valued.

A data summit event builds community among stewards and provides information that can help them in their roles. Summit topics don't necessarily have to be focused on data – guest speakers and executives can talk about how steward's activities contribute to the success of the broader company strategy.

Data stewards are those responsible for part of the data (implementation) and the operation model on implementation (incremental implementation).

Data governance tools

Finding the funds to procure third-party tools rather than creating homegrown options can be recommended. "Tools from the vendors have been built with feedback from dozens of clients that they have. As a result, there's a lot of functionality" (Doupe, 2021), which can immediately improve data management practices.

The minimum needed is a business glossary, which documents the program's standardized terminology and definitions, as well as a data dictionary, which contains various metadata attributes concerning data assets across the organization. This is often also called metadata management and ensures that data is discoverable.

A second must-have tool is a data quality tool for data profiling. Data profiling documents the descriptive statistics of a data set, such as the format of the data, the number of rows or values, and how often they occur. *A data quality tool will also allow you to create and run data rules to produce data quality metrics*. **Data lineage**, how the data has changed over time is also becoming a compliance requirement in many industries and geographies. This tool will be used to live up to that requirement.

Not essential but nice to have is a dashboard that makes data quality metrics highly visible to everyone in the organization. Most organizations have PowerBI or similar tools available. Those can be used to showcase the metrics.

Data quality management and data privacy

Data quality management focuses on ensuring the accuracy, completeness, consistency, and timeliness of data. It involves establishing data quality standards, conducting data profiling and cleansing activities, and implementing data quality controls. By monitoring and improving data quality, organizations can enhance decision-making processes and reduce the risk of relying on erroneous or outdated information.

Privacy is a big area in the governance model and is covered in compliance, security, and general policies. An example of maturity assessment regarding privacy could be:

- Level 0 (Unaware), an organization has no awareness of data privacy issues or requirements
- Level 3 (Proactive), an organization has established data privacy policies and standards and implemented data security controls
- Level 5 (Optimized), an organization has achieved a culture of data privacy excellence and continuous improvement.

Defining the goals and creating a roadmap

An objective data management maturity assessment will provide both beginning and ending points on a roadmap to success. Your roadmap should answer two questions: One, what are you trying to accomplish? And two, how are you going to get there?

Fill in the steps in an implementation plan by mapping out milestones along the way. Ensure that objectives are S.M.A.R.T.: specific, measurable, attainable, relevant, and time-bound. It should be very clear whether a milestone or deliverable was achieved; this should be able to be answered with a "yes" or a "no."

A data management roadmap should be built and developed with two key inputs: Undeveloped areas in need of improvement as defined by the maturity assessment, and issues or areas the data governance committee has identified that need to be addressed.

Figure 2.3 shows the recommended way through the implementation of the DGI data governance framework.

Data contracts

In recent years policies and standard operating procedures (SOP) have been recognized as being "dead" documents that were made and then placed in the drawers of management. The response to that has been the implementation of data contracts to counter it.

Data contracts are a way to *prevent issues and enforce data quality* across data teams and upstream engineering teams. They are similar to API agreements that define the shape, meaning, and constraints of data. Data contracts can help data teams *save time, resources, and frustration* by avoiding issues introduced by upstream teams that change or rename data fields. Data contracts

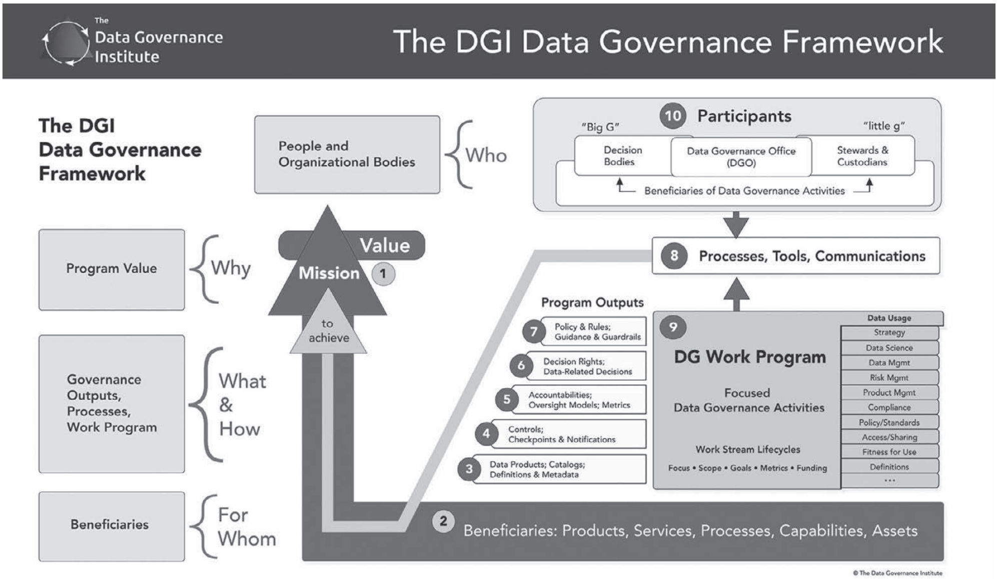

The Data Governance Institute
The DGI Data Governance Framework
The DGI Data Governance Framework
People and Organizational Bodies
Who
10 Participants
"Big G"
"little g"
Decision Bodies
Data Governance Office (DGO)
Stewards & Custodians
Beneficiaries of Data Governance Activities
Program Value
Why
Value
Mission 1
to achieve
8 Processes, Tools, Communications
Program Outputs
7 Policy & Rules; Guidance & Guardrails
6 Decision Rights; Data-Related Decisions
5 Accountabilities; Oversight Models; Metrics
4 Controls; Checkpoints & Notifications
3 Data Products; Catalogs; Definitions & Metadata
9 DG Work Program
Focused Data Governance Activities
Work Stream Lifecycles
Focus • Scope • Goals • Metrics • Funding
Data Usage
Strategy
Data Science
Data Mgmt
Risk Mgmt
Product Mgmt
Compliance
Policy/Standards
Access/Sharing
Fitness for Use
Definitions
...
Governance Outputs, Processes, Work Program
What & How
Beneficiaries
For Whom
2 Beneficiaries: Products, Services, Processes, Capabilities, Assets
© The Data Governance Institute

FIGURE 2.3 DGI data governance framework

can also *improve data reliability, accountability, and culture* by ensuring that data conforms to standards and expectations. However, data contracts also face challenges and limitations, especially enforcement (Dengsøe, 2023).

ORGANIZING FOR ANALYTICS

In the previous sections, we looked at what data strategy comes from the business strategy and what we are looking for when having to convert it into real life. This means we have also come across many tasks that need to be handled by people within the organization.

If we want our employees to make decisions based on data, we need to bring the analysis/ data to where them or at least make it clear where they will be able to get it.

In general, you can organize your analytics function either as a centralized center of excellence or in a hub and spoke structure. What works best is dependent on two things:

- How complex technically is the analysis required and the volume of the analysis
- How deep is the analytical skill of employees, or what is the culture like (See Chapter 4 for more on culture and working with data literacy).

It is of course not an either–or, but more a continuum where few companies have fully centralized or fully decentralized analytics capability.

The data team has to perform three roles where the data engineer primarily extracts data from various sources and loads it into the analytics platform, the analytics engineer/data scientist transforms the data to make it useful for the data analyst who should extract business insights

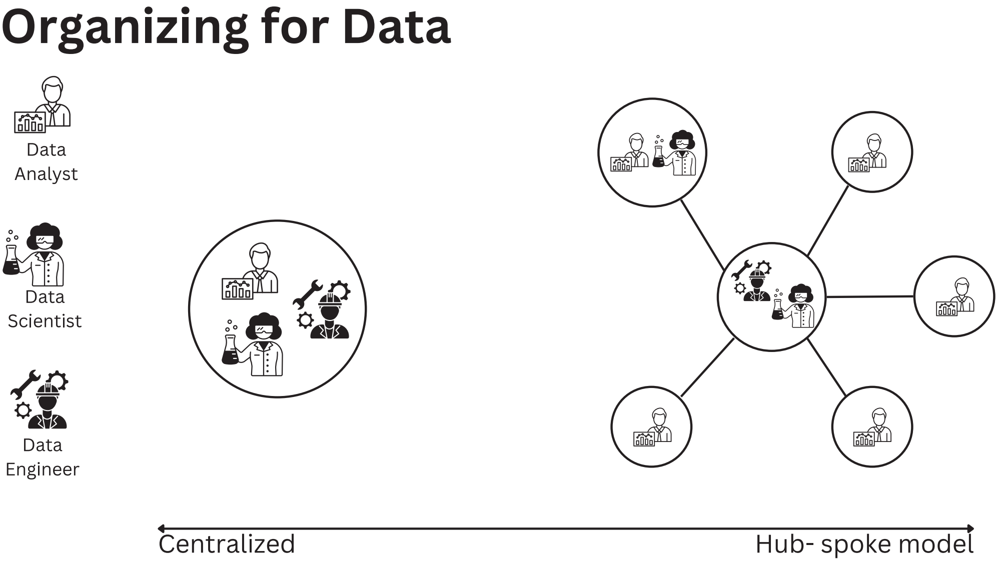


FIGURE 2.4 Organizing the data function within the organization

TABLE 2.2 Functions and roles in the data organization

Data engineer	Data scientist	Data analyst
• Build custom data ingestion • Manage pipeline orchestration • Manage endpoint (system) connections • Build and maintain the data platform • Performance optimization and monitoring	• Build machine learning models • Create data products • Provide data models (semantic/cubes) for analyst • Make non-visual data analysis	• Maintain data • Analyze "simple" data • Train users on data platforms and usage • Maintain documentation • Build dashboards • Link with the data function and business

from it (dbt Learn, n.d.) The roles are somewhat overlapping and the functions can be adjusted based on the skills available in the organization.

Table 2.2 is a small overview over what each of the three roles covers.

Depending on the size and the maturity of the organization the roles can shift and functions can either converge or diverge.

THE DATA GOVERNANCE PROGRAM

A RACI matrix is a tool that helps to clarify the roles and responsibilities of different stakeholders in a data governance project. It assigns four levels of involvement for each stakeholder: Responsible, Accountable, Consulted, and Informed (CIO Wiki, n.d.). Here is an example of a RACI matrix for the implementation of a data governance project, based on data stakeholder definitions:

Communicate

The last piece of the puzzle is to continually broadcast the importance of data. If you're leading a data governance program, your side role needs to be chief communicator. This is critical.

Communication about the data governance program generates two things: It generates awareness, and it generates desire.

Awareness is generated by answering these questions:

- "Why do we have a data governance program?"
- "Why is data important to the organization?"
- "Why should we be focused on improving our data management practices?"

Desire is generated by answering the question: "What's in it for me?"

A data steward, for example, might want to devote themselves to data governance efforts to learn new skills, to advance along a career path, or to provide opportunities to be recognized. "You've got to put yourself in their shoes and figure out what incentivizes them to be part of this effort," he said.

TABLE 2.3 RACI at the different steps of data governance implementation

Task/Milestone	Senior executive leadership	Data trustees	Data stewards	Data custodians	Data consumers	Data analysts	Data architects	Data engineers	Data scientists	Compliance and security experts
Define the data governance vision and strategy	A	C	I	I	I	I	I	I	I	C
Establish data governance committee and roles	A/R	R/C	C/I	C/I	C/I	C/I	C/I	C/I	C/I	C
Develop data governance policies and standards	A/C/I	R/C/I	R			I	C	C	I	R/I
…										

Key messaging should stress:

- Why is data important?
- What are we doing to get better?
- What do we need from you?
- Who are the "unsung heroes" among the data stewards?

Use existing communication channels as well as add new ways to promote the program, such as town halls or putting out governance program newsletters.

Quality and accuracy

A McKinsey survey found that about 30% of all time in an organization is spent on non-value-adding tasks due to data's poor quality and availability.

DAMA, which is an international organization for "data management professionals" is trying to do something about that issue. This is primarily through the DAMA-DMBOK (Data Management Book Of Knowledge). It can be found on their website in updated versions: www.dama.org/cpages/dmbok-2-image-download. This mirrors the widely recognized PMBOK for project management, even though the DMBOK is not in as wide a circulation (yet). The book covers how to ensure data quality through the use of metadata and processes.

Compliance and risk management

Compliance and risk management are covered extensively in Chapter 9, but the implementation of a data governance program is a key cornerstone to ensuring that the organization doesn't open up for legal issues. These could come in the form of privacy violations covered by e.g., the European GDPR legislation, algorithmic violations covered by the US presidential act, or various other legislations covering regulated industries like the financial or pharmaceutical.

Accessibility and availability

The basis for data-driven decision-making is that the data is available for analysis that can guide the decision. Therefore, storing and making data available in the right formats at the right times is crucial. This is covered extensively in Chapter 8 where the data stack is described. The tools for finding data, especially unstructured data, are evolving rapidly at the moment, so it's recommended to draw upon industry specialists to get the latest information available. It is, however, important to keep compliance and risk management in mind when making data available to the organization.

Implementation (policy creation)

Implementation of data governance starts with the establishment of objectives and scope. Based on that, stakeholders can be identified and buy-ins obtained. Without the right level of involvement from the entire organization, a data governance program is likely to fail. See the previous section for the potential stakeholders in the process.

Then the policy creation can begin by going through the parts of the organization using the four-step process:

1. Conduct a data assessment
2. Define roles and responsibilities
3. Develop policy components
4. Check against business objectives.

Some parts will need extensive policies based on the nature of the business and the maturity of the processes, while others can be managed with loose descriptions.

DATA STRATEGY EVALUATION

This section evaluates the effectiveness and impact of the data strategy on the business goals and objectives. It covers how to measure and monitor the progress and outcomes of the data initiatives using tools such as KPIs, dashboards, reports, etc. It also covers how to review and update the data strategy based on feedback and changing business needs.

Once a data strategy has been implemented, it's essential to evaluate its effectiveness and impact on the business goals and objectives. This involves measuring and monitoring the progress and outcomes of the data initiatives using tools such as KPIs, dashboards, reports, and others. The purpose of this evaluation is to determine whether the data strategy is delivering the expected benefits and to identify areas for improvement.

Measuring and monitoring progress and outcomes

To measure and monitor the progress and outcomes of a data strategy, organizations should establish clear KPIs that align with their business goals and objectives. These KPIs should be specific, measurable, achievable, relevant, and time-bound (SMART). Examples of KPIs for a data strategy may include:

- Data quality metrics, such as data accuracy, completeness, and consistency
- Data adoption metrics, such as the percentage of employees using data tools and the frequency of use
- Data-driven decision-making metrics, such as the percentage of decisions supported by data
- Business outcome metrics, such as revenue growth, customer satisfaction, and operational efficiency.

Dashboards and reports can also be used to monitor progress and outcomes. Dashboards provide a visual representation of data, making it easier to identify trends and patterns. Reports offer a more detailed analysis of data and can be used to track progress over time.

Reviewing and updating the data strategy

Regular review and update of the data strategy are critical to ensuring it remains relevant and effective. Feedback from stakeholders, changes in business needs, and emerging technologies may require adjustments to the data strategy.

Senior leadership support and understanding are essential for the success of a data strategy. They should be involved in the review and update process to ensure that the data strategy remains aligned with the organization's overall goals and objectives.

Resources for administration are also necessary to ensure that the data strategy is successfully implemented and maintained. This includes investing in technology, training employees, and hiring data experts.

Clear approval and enforcement consistently across the organization are critical for ensuring that the data strategy is followed throughout the organization. Policies and procedures should be developed and communicated to ensure that everyone understands their roles and responsibilities in implementing the data strategy.

In conclusion, evaluating the effectiveness and impact of a data strategy is crucial for organizations that want to make data-driven decisions. By establishing clear KPIs, using dashboards and reports, and regularly reviewing and updating the data strategy, organizations can ensure that their data strategy is delivering the expected benefits and making a positive impact on their business goals and objectives.

Best practices:

- Senior leadership support and understanding
- Resources for administration
- Clear approval and enforcement consistently across the organization.

SUMMARY

This chapter covers developing a data strategy that aligns with and supports business goals and objectives. It provides frameworks for crafting a data vision, mission, and values, identifying data needs and sources, defining data governance, and evaluating the effectiveness of the data strategy.

Classic strategy frameworks like Porter's generic strategies of cost leadership, differentiation, and focus can be adapted for data initiatives. Data can help drive efficiency, innovation, or customization. Blue ocean's strategy aims to create new demand by understanding unmet customer needs. Scenario planning builds strategic options against different futures.

A data vision describes the aspiration for data while the data mission defines what will be accomplished and how. These should align with the business vision and mission. Data principles state the values that guide data efforts. Strategy implements the mission/vision via governance.

Key steps include auditing current data issues, defining needs via interviews and surveys, and prioritizing initiatives based on business value. Potential internal and external data sources are assessed for relevance, quality, and availability.

Data governance provides systematic data quality, security, privacy, and ethics. It involves policies, standards, roles, and metrics. Key roles include senior leaders as sponsors, trustees accountable for data domains, stewards who define and document data, and IT teams who implement infrastructure. Data contracts help codify expectations between data producers and consumers.

Maturity models like DGI and COBIT 5 assess current governance and management capability levels to identify gaps and improvement priorities. An executive owner and cross-functional data governance committee help drive the program. Formalizing via a committee charter clarifies expectations.

Priorities come from assessing gaps versus desired states and issues raised by the governance committee. Data quality and metadata tools provide visibility. Dashboards make metrics visible across the organization. Collaboration, community building, and communication drive adoption.

Evaluating data strategy impact involves defining SMART KPIs on dimensions like data quality, decision-making, and business outcomes. Dashboards visualize progress. Regular review ensures the data strategy stays relevant to evolving business needs. Leadership commitment, investment in skilled resources, and consistent reinforcement sustain success.

In summary, an effective data strategy requires aligning initiatives with business goals, crafting a compelling data vision/mission, systematically governing data via policies and roles, and regularly evaluating progress toward desired data and organizational outcomes.

KEY TERMS

Business goals: The desired outcomes or results that the business wants to achieve in a given period.

COBIT framework: A framework that guides governance and management of enterprise IT, with a maturity model to measure capability.

Dashboard: A data visualization interface that tracks KPIs and metrics to provide insights into current status or progress toward goals.

Data accuracy: The degree to which data correctly describes the real-world object or event being represented.

Data adoption metric: A metric that measures the usage and adoption of data and analytics tools.

Data analyst: A specialist who analyzes data and transforms it into insights, recommendations, visualizations, or stories.

Data architect: An expert who designs and maintains the overall data architecture of an organization.

Data architecture: The master blueprint that standardizes how data is treated across systems, applications, and processes to serve business needs.

Data catalog: A centralized repository of information to help discover, understand, and manage data.

Data completeness: The degree to which data expected to be captured is present.

Data consistency: The absence of difference, when comparing two or more representations of a thing against a definition.

Data consumer: An end user of data, who relies on its accuracy, completeness, consistency, and timeliness.

Data contract: Agreements that define expectations around data shape, meaning, constraints, and handling between data teams.

Data custodian: An IT or system manager responsible for reviewing access rights to data, and overseeing security, quality, and integrity.

Data discovery: The process of identifying, locating, and accessing data from various internal or external sources across an organization.

Data engineer: A developer who builds and optimizes the technical data pipelines and workflows of an organization.

Data governance: A systematic approach to managing data assets that ensures data quality, security, and privacy and enables appropriate access and usage.

Data lineage: Documentation of the origins and movement of data from its point of entry through integration, transformation, and use.

Data mission: A statement that describes what an organization is going to accomplish with data and how it is going to do it.

Data pipeline: An automated workflow that moves data from one system to another on a scheduled basis for integration, processing, analytics, or reporting.

Data profiling: The process of examining, categorizing, and organizing data to understand its meaning, relationships, and quality.

Data quality: The accuracy, completeness, consistency, and timeliness of data.

Data scientist: A specialist, often with statistical or mathematical background focused on the advanced predictive ML models or algorithms.

Data steward: An individual responsible for ensuring the quality, integrity, and security of data assets by supporting data governance policies and procedures.

Data strategy: A plan that outlines how a firm can use data to achieve its strategic goals and objectives.

Data trustee: A business or operational leader who represents their unit in data governance and is accountable for data definitions, quality, and compliance in their domain.

Data vision: A statement that describes why an organization cares about data and what is the ultimate aspiration for data.

DGI framework: A framework focused specifically on data governance, with maturity model levels based on awareness, proactivity, management, and optimization.

ETL (extract, transform, load): The process of extracting data from source systems, transforming it into an analysis-friendly format, and loading it into a data warehouse or other target system.

KPI (key performance indicator): A quantifiable metric used to define, track, and communicate performance against strategic goals and objectives.

Maturity model: A methodology used to assess the current capabilities or proficiency level of an organization across different domains or practices.

Metadata: Data that describes and gives information about other data, such as definitions, descriptions, classifications, and usage.

RACI matrix: A matrix that clarifies roles and responsibilities by defining levels of involvement in tasks as responsible, accountable, consulted, and informed.

REVIEW QUESTIONS

1 What is a data strategy and why is it important to align it with business goals?
2 What are Porter's three generic business strategies and how can data support each one?
3 What is the blue ocean strategy and how can data analytics help implement it?
4 What are the key elements of the strategic scenario planning model?
5 What should an effective data vision and mission statement communicate?
6 What are some methods for identifying data needs and potential sources?
7 What are the main components of the DGI data governance framework?
8 What are the different maturity levels in a data governance maturity model?
9 What are the key roles and responsibilities in data governance?
10 What is a RACI matrix and what are its benefits for data governance?
11 What tools can help monitor and enforce data governance?
12 What is the difference between data governance and data management?
13 Why are data contracts important for data quality and accountability?
14 What is the recommended path for implementing a data governance initiative?
15 How can you evaluate the effectiveness and impact of a data strategy?
16 What are some key performance indicators (KPIs) to measure a data strategy?
17 Why is it important to regularly review and update the data strategy?
18 What best practices help sustain the success of a data strategy over time?
19 What are the benefits of centralizing vs decentralizing data analytics capabilities?
20 What factors determine the best analytics organizational structure?

Answers to review questions:

1 What is a data strategy and why is it important to align it with business goals? (Introduction section)
2 What are Porter's three generic business strategies and how can data support each one? (Porter's Generic Strategies section)
3 What is the blue ocean strategy and how can data analytics help implement it? (Blue Ocean Strategy section)
4 What are the key elements of the strategic scenario planning model? (Strategic Scenarios section)
5 What should an effective data vision and mission statement communicate? (Data Vision, Mission, Strategy, and Values section)
6 What methods can be used for identifying data needs and potential sources? (Data Needs and Potential Sources section)
7 What are the main components of the DGI data governance framework? (Data Governance Framework section)
8 What are the different maturity levels in a data governance maturity model? (Data Governance Framework section)
9 What are the key roles and responsibilities in data governance? (Stakeholders in Data Governance section)

10 What is a RACI matrix and what are its benefits for data governance? (Data Governance Program section)
11 What tools can help monitor and enforce data governance? (Data Governance Tools section)
12 What is the difference between data governance and data management? (Data Governance vs Data Management section)
13 Why are data contracts important for data quality and accountability? (Data Contracts section)
14 What is the recommended path for implementing a data governance initiative? (Figure in Data Governance section)
15 How can you evaluate the effectiveness and impact of a data strategy? (Data Strategy Evaluation section)
16 What are some key performance indicators to measure a data strategy? (Data Strategy Evaluation section)
17 Why is it important to regularly review and update the data strategy? (Reviewing and Updating the Data Strategy section)
18 What best practices help sustain the success of a data strategy over time? (Data Strategy Evaluation section)
19 What are the benefits of centralizing vs decentralizing data analytics capabilities? (Organizing for Analytics section)
20 What factors determine the best analytics organizational structure? (Organizing for Analytics section)

NOTES

[1] Bruce D. Henderson (1989), The origin of strategy. *Harvard Business Review*.
[2] Richard Horwath (2020, July 1), The origin of strategy (blog). Strategic Thinking Institute.
[3] Paul de Ruijter (2014), *Scenario based strategy: Navigate the future* (1st ed.) Routledge.
[4] Ibid.

BIBLIOGRAPHY

AIMultiple. (2023). *What is data governance? Use cases, best practices & tools*. https://research.aimultiple.com/data-governance/ (Accessed October 7, 2023).

AWS (no date) *What is data strategy? Data analytics strategy explained*. https://aws.amazon.com/what-is/data-strategy/ (Accessed January 10, 2024).

CIO Wiki. (n.d.). *RACI matrix*. https://cio-wiki.org/wiki/RACI_Matrix (Accessed September, 2023).

ClearPoint Strategy. (2023, June 21). *The Blue Ocean Strategy Summary (With 4 Examples)*. www.clearpointstrategy.com/blog/blue-ocean-strategy

CMMI Institute (n.d.) Appraisal levels. https://cmmiinstitute.com/learning/appraisals/levels (Accessed January 10, 2024).

DAMA International. (2017). *DAMA-DMBOK: Data management body of knowledge* (2nd ed.). Technics Publications.

DAMA. Earley, S., Henderson, D., & Sebastian-Coleman, L. (Eds.) (2017). *The DAMA guide to the data management body of knowledge (DAMA-DM BOK)*. Bradley Beach, NJ: Technics Publications.

Data Governance Institute. (2004). *The DGI data governance framework*. https://datagovernance.com/the-dgi-data-governance-framework/ (Accessed October 7, 2023).

Data.org. (2023). *The complete guide to data governance roles and responsibilities*. https://data.org/resources/the-complete-guide-to-data-governance-roles-and-responsibilities/ (Accessed October 7, 2023).

DataCamp. (2023). *Data governance fundamentals cheatsheet*. www.datacamp.com/cheat-sheet/data-governance-fundamentals-cheatsheet (Accessed October 7, 2023).

Davenport, T. H. (2017). *Competing on analytics: The new science of winning*. Boston, MA: Harvard Business Review Press.

dbt Learn (n.d.) *Analytics Engineer*. https://courses.getdbt.com/courses/take/fundamentals/lessons/27962291-analytics-engineer (Accessed January 10, 2024).

Dengsøe, M. (2023, June 6). *Activating ownership with data contracts in dbt*. https://medium.com/@mikldd/activating-ownership-with-data-contracts-in-dbt-4f2de41c4657

de Ruijter, P. D. (2014). *Scenario based strategy: Navigate the future* (1st ed.). London: Routledge. https://doi.org/10.4324/9781315607689

Doupe, R. (2021, April). Getting started with Data Governance. Webinar organized by DAMA International.

Gartner. (n.d.). Learn to build an ai strategy for your business. www.gartner.com/en/information-technology/topics/ai-strategy-for-business (Accessed April 13, 2024).

Gartner IT Glossary. (n.d.). *Data governance*. www.gartner.com/en/information-technology/glossary/data-governance#:~:text=Data%20governance%20is%20the%20specification,executing%20effective%20data%20governance%20initiatives

IBM. (n.d.) *Design your data strategy in six steps*. www.ibm.com/resources/the-data-differentiator/data-strategy (Accessed January 10, 2024).

ISACA. (2012). *A business framework for the governance and management of enterprise IT*. COBIT®5, Information Systems Audit and Control Association – ITPro Collection.

Kim, W. C. (2005). *Blue ocean strategy*. Boston, MA: Harvard Business School Press.

KMPG. (2021, November). the-data-imperative.pdf. https://assets.kpmg.com/content/dam/kpmg/xx/pdf/2021/11/the-data-imperative.pdf

Ladley J. (2019). *Data governance: How to design deploy and sustain an effective data governance program* (2nd ed.). Burlington, MA: Morgan Kaufmann.

LightsOnData (n.d.). *Data governance maturity models – Gartner*. www.lightsondata.com/data-governance-maturity-models-gartner/ (Accessed January 10, 2024).

Luther, D., & Ali, R. (2022). *Scenario planning: Strategy, steps and practical examples*. www.netsuite.com/portal/resource/articles/financial-management/scenario-planning.shtml (Accessed January 10, 2024).

McKinsey. (2023). *Designing data governance that delivers value*. www.mckinsey.com/capabilities/mckinsey-digital/our-insights/designing-data-governance-that-delivers-value (Accessed October 7, 2023).

NetApp. (2023). *Data governance: Roles, policies, and challenges*. www.netapp.com/media/102374-policy-driven-data.pdf (Accessed October 7, 2023).

Porter, M. E. (1980). *Competitive strategy*. New York: The Free Press (MacMillan).

Seiner, R. (2023). *Non-invasive data governance strikes again: Gaining experience and perspective*. Sedona, AZ: Technics Publications.

Treder, M. (2020). *The chief data officer management handbook*. Berkeley, CA: Apress.

CHAPTER 3

Data Products

You need to make a decision. You want it to be data-driven, but where to start? You need information and a way to manipulate the data within the tool to ensure that it is not just data, but information that you can build decisions upon. That could easily be an example of what in the chapter is defined as a data product.

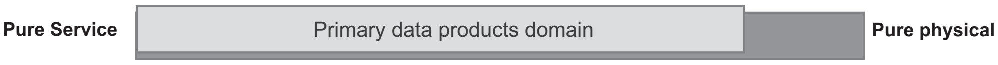
This chapter explores how to design and build data products that solve problems or create opportunities for customers and internal users. It also covers how to measure and communicate the value of data products using metrics and feedback loops.

LEARNING GOALS:

L3.1 Understand what a data product is
L3.2 Have insights into product portfolio management for data products
L3.3 Be able to describe a data product using a value proposition canvas
L3.4 Understand the use of the Data Product Canvas

Data products come in several different shapes and forms and there is no single agreed-upon definition. For this book, we will start by defining them as either internally or externally focused *solutions that have data as the core component*. The primary focus is on data products for internal consumption, but most of the principles applied for these internal solutions come from regular product management and can therefore be applied across both domains.

We use the term *solution* as it indicates a task that needs to be solved and that solution might be a physical product, a service, or anything in between. For the most part, we are however looking at data services or data that is enhanced or carried by a physical product.



You can argue that books and printed reports are purely physical data products, but for the sake of clarity, we are addressing data in the digital format only.

DOI: 10.4324/9781003457787-3

VALUE CREATION FOR DATA PRODUCT DEFINITION

What does it then mean that data is at the core? It comes down to the value that the data drives in a business context. Is the primary value based on data or is it just a supplement to what is being done? A way to assess it is by listing the pain relievers and gain creators (Alexander Osterwalder, 2014) in the solution provided, but let's take a look at how to define a data product based on value creation.

The value proposition canvas is often used for describing how value is created for a customer by a product or a service. By utilizing that model, we can assess whether we are talking about a data product/service or a regular product/service. Figure 3.1 shows the concept graphically, where the value map that describes the three product elements is listed on one side and the three elements that describe the customers are listed on the other side. If the value map primarily describes data-based features, we are looking at a data product.

Customer profile

The customer of a data product is primarily defined as someone who requires knowledge. Knowledge is, according to Webster's Dictionary, "the fact or condition of knowing something with familiarity gained through experience or association". That means we require information to be presented or experienced in the solution offered. As we are primarily focusing on data-driven decision-making in this book we will define the customer as someone who needs information or tools to enhance decisions. This can then be elaborated on by defining the pains and wants of that decision-maker.[1]

Pains are often associated with the risks of making an erroneous decision or not getting the information to make a decision/action in a timely manner.

Some examples could be:

- Losing market share or reputation by making wrong strategic decisions due to inaccurate or outdated data
- Wasting time and money by using a data product that is complex, slow, or incompatible with existing systems
- Facing legal penalties or customer complaints by violating data protection regulations or ethical standards.

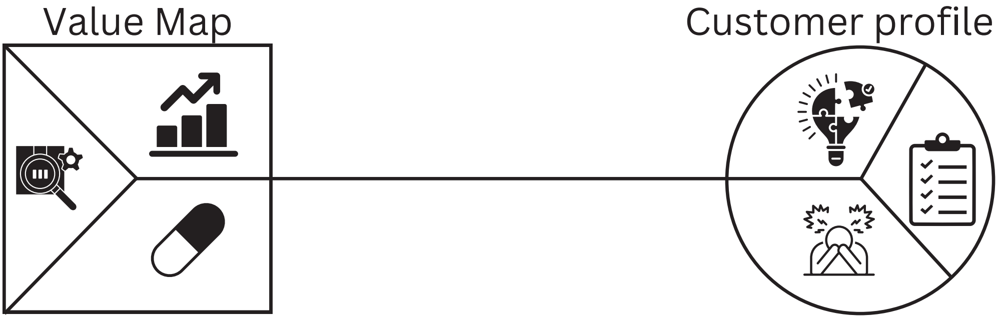


FIGURE 3.1 Value proposition model for data products

Gains are most often related to the improvement of a decision that would be taken anyway or with the possibility of making new decisions that the customer was not aware of before he/she was presented with the information.

Some examples could be:

- Increasing sales revenue by 10% by using the data product to identify the most profitable customer segments and products
- Reducing operational costs by 15% by using the data product to optimize inventory and supply chain management
- Enhancing customer satisfaction and loyalty by using the data product to personalize offers and recommendations.

Value map

The value map in the value proposition canvas is a description of how the data product delivers value to the customer. It consists of three elements: products and services, pain relievers, and gain creators.

Products and services are the tangible or intangible offerings that the data product provides to the customer. They can be data, insights, visualizations, dashboards, reports, algorithms, models, etc. All are described as features.

Pain relievers are the ways that the data product alleviates or eliminates the customer's pains. They can be features, functions, benefits, or outcomes that address the customer's problems, frustrations, or challenges.

Gain creators are the ways that the data product enhances or creates the customer's gains. They can be features, functions, benefits, or outcomes that enable the customer to achieve their goals, aspirations, or desires.

The value map should be aligned with the customer profile, meaning that *every pain reliever and gain creator should correspond to a specific pain or gain of the customer*. The value map should also be differentiated from the competing solutions, meaning that it should offer a unique value proposition that is superior or distinctive from other data products in the market.

When used internally a value proposition map should be presented to the "customer" or stakeholder to get confirmation of the pains and gains before embarking on the development of the data product/solution.

But how do we come up with these pain relievers and gain creators? A good approach is the design thinking methodology as it will channel and enhance the creative process along the lines of value creation while keeping the focus on the customer. The following section will step you through the process before we go deeper into the data product structures, implementation, and management.

DESIGN THINKING FOR CREATING DATA PRODUCTS

Why is it relevant to work with design thinking? Design thinking can help you or your team uncover the unmet needs of the people you are creating solutions for. It can reduce the risk associated with launching new ideas and generate solutions that are revolutionary and not just incremental. Design thinking generally helps organizations learn faster and is an inherently data-driven approach to innovation. Design thinking is a process for designing or a process for innovation (5 Examples of Design Thinking in Business, 2023) and a method for solving complex problems and creating innovative solutions (8 Great Design Thinking Examples, 2023). Design thinking can be used in many different contexts and has proven to be effective in both large and small organizations (The 5 Stages of the Design Thinking Process, 2023). Some examples of companies that have used design thinking to create innovative solutions are:

- GE Healthcare focused on user-centered design to improve a product that seemingly had no problems (5 Examples of Design Thinking in Business, 2023)
- Oral B used design thinking to test initiatives before they were implemented (5 Examples of Design Thinking in Business, 2023)
- Netflix used design thinking to create a personalized experience for viewers (5 Examples of Design Thinking in Business, 2023)
- Airbnb used design thinking to transform its business from a failing start-up to a billion-dollar company (Pengertian, 2023)
- UberEats used design thinking to improve customer experience and increase sales (5 Examples of Design Thinking in Business, 2023).

The design thinking story has a history that goes back to the 1950s and '60s and has roots in the study of design cognition and design methods. Design thinking can be considered as a process for designing or a process for innovation (Design thinking, Wikipedia, 2023). It was the cognitive scientist and Nobel laureate Herbert A. Simon who first mentioned design as a way of thinking in his 1969 book *The Sciences of the Artificial* (Simon, 1969; The History of Design Thinking, 2023). He then contributed many ideas in the 1970s that are now regarded as principles of design thinking (The History of Design Thinking, 2023). IDEO, a non-profit organization, has supported the approach since 1991 and hosted the first Design-Thinking Research Symposia (The History of Design Thinking, 2023). In 2005, IDEO helped found Stanford University D School (officially Hasso Plattner Institute of Innovation and Design) (The History of Design Thinking, 2023). The model, regardless of which variation of design thinking (sometimes called HCD – human-centered design), is based on three pillars:

- Empathy (inspiration) – Understanding the needs of those you are designing for
- Ideation (ideation) – Generating many ideas. Brainstorming is a technique, but there are many others
- Experimentation (implementation) – Testing these ideas with prototypes. This is illustrated in this way by Stanford D School in Paris.

The inspiration phase or empathy phase in design thinking is the first phase in the process (Interaction Design Foundation, 2020). The purpose of the empathy phase is to gain a deep understanding of the problems and realities that the people you are designing for face (empathize-dalam-design-thinking, 2023). Empathy requires that you learn about the difficulties that people face and reveal their latent needs and desires to explain their behavior (Interaction Design Foundation, 2020).

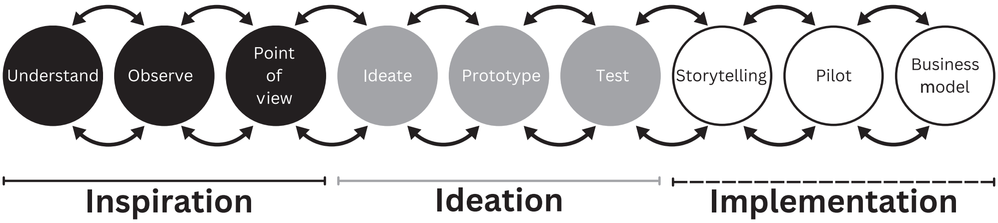


FIGURE 3.2 The design thinking process

The inspiration phase, or empathy phase in design thinking, contains several techniques that can be used to understand the problem (understand), validate the problem (observe), and ensure that you solve the problem user's face (POV, point of view/user perspective). Define your target groups (understand). Who should we design/develop for? If you have not fully understood their world, you can be sure that you will not reach your goal with the use of your solution (Linke, 2017). If there are several target groups, you should also make sure that there are no conflicts between them when the solutions are tested.

Interview types (understand/POV)

Understanding the customer/user of the solution is an important part of the design thinking process. Here are some questions you can ask in an interview to understand the customer better:

- What is your biggest problem with [product/service]?
- What is most important to you when using [product/service]?
- How do you use [product/service] in your daily life?
- What other products/services have you tried before?
- What can we do for usability?

These questions can help you understand the customer's needs and desires better and give you an idea of how you can improve your product or service. Often, an open or semi-structured interview will be the way forward to uncover the problems. However, you can also start with an expert interview. An expert who knows the target group can in some cases be better than them at explaining what they do and experience. However, there is a risk of falling into the traditions if you weigh an expert interview too high. Another form of interview is the focus group interview. Here is an example of how you can conduct a focus group interview in connection with your design thinking process:

- Define the purpose of the interview and the questions you want to ask. For example, you can ask about their needs and desires concerning your product or service.
- Choose a group of people who represent your target group. For example, you can choose people from different age groups or geographic locations.

- Plan the interview and choose a suitable place and time. For example, you can choose to hold the interview online or offline.
- Conduct the interview and make sure to take notes. You can also record the interview to ensure that you do not miss anything important.
- Analyze the results and use them to inform your design thinking process.

Analogous inspiration (observe)

Analogous inspiration is a way of finding solutions in different contexts that can be relevant to your problem or inspire an idea (Design Thinking for Brands, 2023). Analogous inspiration can help create new ways of thinking about a challenge and identify and observe experiences that are not directly related to the industry being designed for but have a relatable aspect (Poetz, 2023). An example of how such an exercise can be conducted (activity-explore-analogous-inspiration, 2023):

Step 1. Choose a part of a service, experience, or problem that you want to focus on. Example: We want to increase access to our school lunch programs

Step 2. Identify a feeling that you want to evoke in your focus audience. Example: We want the children to feel proud instead of ashamed of using our programs

Step 3. Brainstorm other services, experiences, or solutions that evoke that feeling. Example: Seeing their work displayed on a glass wall, being selected for a team, receiving a compliment on the project they are choosing to proceed with. Example: Being selected for a team

Step 4. Explore how the analogous service, experience, or solution evokes that feeling. Be specific. Example: Uniforms and matching equipment help them show their identity as part of the team, they get the opportunity to spend time with a new group of friends who share their passion, they have a coach who values them and wants to include them

Step 5. Fill out this Madlib statement: How can we make (our service, experience, or problem) more like (analogous service, experience, or solution)? Example: How can we make signing up for our school lunch programs more like being selected for a team?

Step 6. Use this Madlib as a framework for another brainstorming to generate new ideas for your context.

Extremes and mainstreams (POV)

"Extremes and mainstreams" is a method used in the design thinking process to identify and understand different perspectives and needs of the users (Extremes and Mainstreams: Design Toolkit, 2023). The method involves interviewing both extreme and mainstream users to get a broader understanding of the problem and possible solutions. To perform this exercise, you can follow these steps:

- Identify your extreme users – those who are far from the average in terms of age, gender, income, or other factors

- Identify your mainstream users – those who are close to the average in terms of age, gender, income, or other factors.

Interview both groups of users and ask questions about their needs and desires concerning your product or service. Analyze the results and use them to inform your design thinking process.

Nonverbal communication, also sometimes called body language, is defined as a series of nonverbal signs and signals that include facial expressions, hand, leg, and foot movements, and posture. From our body language, we can better interpret and understand what others want to tell us (kropssprog-hvad-behoever-vide-forstaa-signaler, 2023). There are many ways to study nonverbal communication in people (10 Tips to Improve Your Nonverbal Communication - Verywell Mind. , 2023). Some of them include paying attention to nonverbal signals, looking for incongruent behavior, focusing on tone of voice, using good eye contact, asking questions, using signals to add meaning, looking at signals as a whole, and considering the context. Observe and note the following:

- posture
- movements
- location in space
- touches
- facial expressions
- eye movements and gestures.

Body language is often interpreted unconsciously and can therefore be difficult to observe in a structured way, but studies have shown that up to 80% of information transmission occurs nonverbally (Mehrabian, 1969).

Posture

If a person has their shoulders back and their back straight, it is a sign that they are engaged, listening, and open to the ideas or information you present. If they show poor posture with their shoulders hanging or raised and their back bent, they may be nervous, anxious, or angry. If a person has their arms down by their sides, on the table, or arranged in another open way, it is a sign that they feel positive and ready to absorb information. If their arms are crossed or closed, they may experience some kind of negative emotion. If a person has both feet flat on the ground, it is a sign that they feel open to communication. If their legs are crossed or arranged in another closed formation, they may feel irritated or stressed. If you communicate with someone who frowns or has tight lips, you can pause to make sure they don't feel confused, angry, or any other negative emotions. If you communicate with someone who has a soft smile, relaxed facial muscles, or gently raised eyebrows, it is a sign that they feel comfortable with the information you present (Nonverbal-communication-skills, 2023).

5xWhy

“5 times Why” is a technique to find the root cause of a problem by repeatedly asking “Why?” until the symptoms are traced back to the source. It was developed by Toyota and is also known as “5 Whys”. The method is most effective when the answers come from people who have practical experience with the process or problem in question. The method is remarkably simple: When a problem occurs, you drill down to its root cause by asking “Why?” five times (Five whys – Wikipedia, 2023).

Photojournal

A photojournal is a collection of photographs that tell a story or document an event. The best practice for making a photojournal includes having a clear idea of the story or event you want to document (How to Easily Make a Cool Photo Journal, 2023). It is also important to take pictures in the right order to provide an easy understanding of any event. You can also consider including both routine and extraordinary events and taking pictures from different angles and perspectives.

“Framing the Design Challenge”

“Framing the design challenge” is a process of defining and organizing a problem to create a clear understanding of the problem and possible solutions (How to Properly Frame Your Design Challenge, 2023). It helps to turn problems into opportunities and provides a clear idea of where to focus the design process (How To Properly Frame Your Design Challenge, 2023). To do this, you can start by asking questions like “What is the problem?”, “Who are the users?”, “What are the goals?”, and “What are the current challenges?” (Daily Creations’s Framing your Design Challenge template, 2023).

Ideation

The ideation phase consists of three elements, which build on the understanding of the problem field and empathy with the target group. In the ideation phase of the design thinking process, the purpose is to generate as many ideas as possible and then select the best ideas to continue prototyping and testing. The ideation phase can include different techniques such as brainstorming, brainwriting, the worst possible idea, and SCAMPER (The 5 Stages of the Design Thinking Process, 2023).

Brainstorming

Brainstorming is for some just a listing of the things they have just come up with. However, it can be much better by following some simple principles.

1. Set a time limit and stick to the topic
2. Start with a problem statement or a goal and keep focusing on the topic
3. Postpone judgment or criticism, including non-verbal
4. Encourage wild and crazy ideas

5. Say yes to all ideas and say thank you for your contributions
6. Build on each other's ideas. These best practices can help create a positive and productive brainstorming session.

You can optionally combine numbers 5 and 6 in the "Idea Relay" exercise. Here each participant chooses their "best idea" and then the task for the other participants is to build on the idea. For this purpose, Figure 3.3 can be used.

Worst possible idea (Idea generation)

Worst possible idea (Idea generation) is a technique to generate ideas by thinking of the worst possible solutions to a problem and is, therefore, a mirror of the above brainstorming. The purpose of this technique is to create a positive and productive brainstorming session by removing fear of criticism and encouraging creative thinking. To use this technique, you ask the participants to think of the worst possible ideas for a problem or challenge. This can help to unleash creativity and lead to more innovative solutions. Once you have generated many bad ideas, you can then work on refining and improving them.

SCAMPER (Idea generation)

SCAMPER technique is a method of creative thinking and problem solving that consists of a set of questions and action verbs that help generate new ideas and solutions (SCAMPER – Teknik Untuk Pemecahan Masalah Kreatif, 2023). The process consists of:

- substitute
- combine

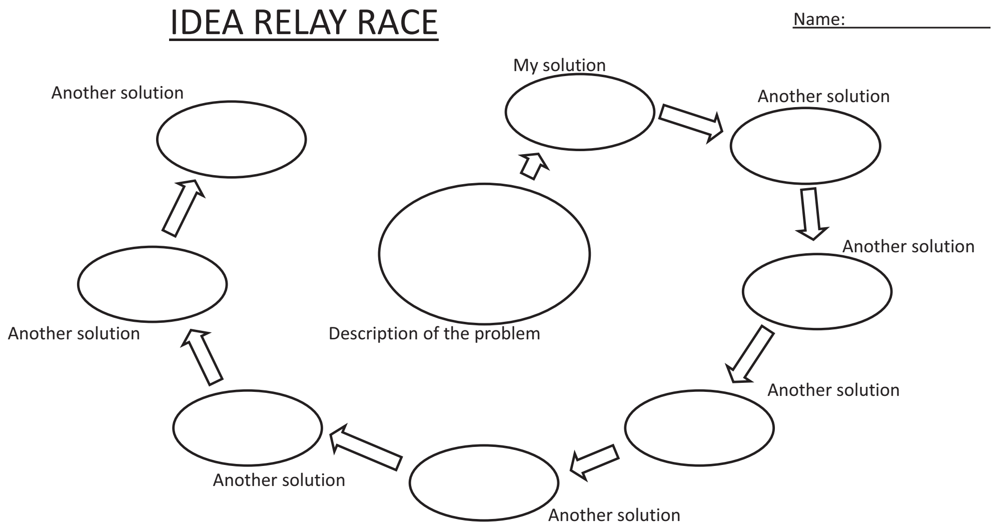


FIGURE 3.3 Model for collaborative idea relay race

- adjust
- modify
- put to other uses
- eliminate
- reverse. (SCAMPER – Improving Products and Services, 2023)

To use the SCAMPER technique, you ask questions about existing products using each of the seven prompts above. These questions help you come up with creative ideas for developing new products and improving current products.

Role-playing and picture cards

Sometimes you can get caught in what is called converging thinking. This means that you "just" keep coming up with ideas in the same direction. This often happens in connection with group brainstorms where the first idea "feeds" the next, but where it is just a a variation on a theme. To break the chain of thoughts of ideas, you need to get an external influence. Here, role-playing and picture cards can be used. You can buy role-playing cards dedicated to both purposes but also start with the list below for role-playing. In practice, you choose a role/a number and say:

"As a [role] I would solve [the problem] by [solution]:

Teacher
Nurse
Pedagogue/Educator
Social Advisor
Dentist
Physiotherapist
Psychologist
Lawyer
Journalist
Engineer
Architect
Graphic Designer
IT Consultant
Project Manager
Sales Manager
Marketing Manager
HR Manager
Finance Manager
Production Manager
Store Manage
Restaurant Manager
Hotel Manager
Auditor
Researcher
Consultant
Real Estate Agent
Financial Advisor
Insurance Advisor
Recruitment Consultant
Customer Service Representative

Instead of buying picture cards, you can find one of the online websites that find random pictures that are uploaded to the internet. It could be this page: Random Pictures – 1000+ Random Images (www.randomwordgenerator.com). Here you choose to display 1–3 pictures and use them to build ideas.

Prototyping

Prototyping is a process where design teams implement ideas in tangible forms from paper to digital. The teams build prototypes of varying degrees of accuracy to capture design concepts and test on users. With prototypes, you can refine and validate your designs, so your brand can release the right products (Prototyping, 2023). Prototyping has many benefits, including allowing you to test the product's functionality as well as work quickly, focus on quality, and use materials. Another benefit of prototyping is the speed at which it can be done. Rapid prototyping systems, such as 3D printers, can create prototypes in hours (What are the advantages and disadvantages of prototyping? 2023). But prototyping also has some disadvantages such as additional development costs for the process and some issues with the accuracy of the design (What are the advantages and disadvantages of prototyping?, 2023). If you are working with visual/digital products, tools like Figma can be used with advantage. More generally, tools like Mural and Canva can also be used to some extent to illustrate ideas.

Prototyping can take many forms and can be a useful way to visualize a product and test its functionality. Here are some of the most common types of prototypes:

Sketches and diagrams

The most basic form of prototyping simply involves sketching an idea on paper. These can vary in detail, but such paper prototypes are a useful starting point for conceptualizing an idea for a new product. These allow ideas to be shared, so a design can be formalized for later development.

Physical models

Physical prototypes can range from simple paper-based designs to more complex versions. These provide a rough idea of a design and show a scaled-down version of a concept before creating a larger-scale model. Used for a variety of different designs, these are particularly suited for smaller objects but can be used for larger projects such as architectural designs.

3D printing has revolutionized prototyping and allows engineers to create realistic production design models quickly. These 3D models mean that companies can identify any errors or areas for development and move quickly toward the production phase. With rapid modeling, these prototypes can be adjusted, and new versions created, allowing for rapid testing and simplifying and reducing large designs to a more manageable scale.

Wireframes

Wireframes are digital diagrams or layouts of a product, often used for software, websites, or other digital assets to present a visual information architecture blueprint. They allow designers and other project workers to navigate a digital structure and place content as well as assess user interface and user flow, allowing for later usability testing to find any usability issues. Such digital representations can be presented as low- or high-fidelity prototypes.

Virtual or augmented reality

Virtual or augmented reality tools can be used for some designs and allow users to "experience" a design as if it were in the physical world. This can be used for building design, amusement parks, and other real settings.

These prototypes can be either digital or physical models and are used to test features that can be added later in the design process. These allow prototyping designers to enhance or customize an existing prototype with additional features.

Working models

More complex than an initial prototype, a working model allows designers to test whether a product works as intended. These are typically used for mechanical inventions or designs that need to move or fit in a certain way. These allow designers to see if their designs work.

Video prototypes

Animated videos and simulations can be used as a graphical representation of a product, while video can also be used to show other prototypes to help designers, managers, and consumers visualize a product.

Horizontal prototypes

Primarily used in software design, horizontal prototypes show a design from a user's perspective, including menus, windows, and screens, so user interactions can be tested.

Vertical *prototypes*

Used for database design, vertical prototypes are digital "backend" models that allow testing of software features before the next design phase.

What are the challenges of prototyping?

Because prototypes are not fully functional final versions, there can be challenges. The higher the level of detail, the easier it is to test, but it still requires taking into account any potential problems and performing essential tests. You should also make sure to choose the right prototyping process and avoid wasting time on less important factors such as aesthetics. It's also easy to spend too much time working on new prototype iterations, which can unnecessarily delay the creation of a final product.

How do you choose the right prototype?

Your prototype should match the final real-world functionality of the product as closely as possible. This applies both to the design of an entirely new product or the refinement of an existing one. The closer you can get to the final product, the better. In some cases, a physical prototype may be preferred over a digital one, although advances in computer technology mean that modern prototyping can be achieved without the need for multiple physical iterations (What Is Prototyping, 2023).

Testing

While testing has its section in the model, it is integrated with prototyping. The process typically involves the following six steps:

- Define your goals and success criteria
- Choose test participants
- Choose a testing method
- Conduct the test
- Analyze the results
- Refine your prototype.

When testing a prototype, it's important to keep in mind the ultimate goals of your product. This includes understanding user needs and how your product will meet them. Testing is crucial, but you shouldn't become too bogged down in repeated testing. It's best to conduct your testing, assess the results, make changes, and repeat the process until you're satisfied with the outcome.

There are many different testing methods you can use when testing prototypes. Some of the most common include (Prototype Testing: Definition, Benefits, How-To, 2023):

1. User testing: Test your prototypes with real users to get feedback on usability and functionality.
2. A/B testing: Test different versions of your prototype to see which works best.
3. Usability testing: Test your prototypes to see if they're easy to use and understand.
4. Qualitative testing: Test your prototypes with a smaller group of users to get more detailed feedback.
5. Quantitative testing: Test your prototypes with a larger group of users to get more general feedback.

Implementation

Implementation in design thinking refers to the last phase of the design thinking process, where you implement the solution you have developed (The 5 Stages of the Design Thinking Process, 2023). Implementation is a critical part of the design thinking process because it is here that you take your prototype and bring it to life (Randhowa et al., 2021).

Implementation can be challenging because it requires collaboration and coordination across different departments and teams (Design Thinking, explained, 2023). But when done correctly, it can lead to significant advantages for the organization and its customers (4 Steps to Implementing Design Thinking, 2023).

Storytelling

Storytelling in design thinking is a way to understand and communicate with users (What is Storytelling?, 2023). It is a method for creating an emotional connection with users and understanding their needs and desires (How to Use Storytelling in Design Thinking, 2023).

Storytelling can be used as part of the implementation process in design thinking by helping to communicate the solution to stakeholders (5 Effective Ways of Using Storytelling in Design, 2023). It can also help create a sense of ownership and engagement among team members (How to use Storytelling in Design Thinking, 2023).

Aristotle (a Greek philosopher) defined good stories based on Figure 3.4. Good stories are the foundation of good storytelling in design thinking. This means that the story that is told around the solution should contain most or all of the following elements:

- Plot: What do the users want to achieve/overcome?
- Character: Who are the users, not just their demographics, but what insights do you need to understand about them and their needs?
- Theme: How can you establish a credible presence for them and still stand out from competitors? How can you reflect on the overall obstacles that users must overcome?
- Dialogue/Dictation: What will your design say to users and how? Does a formal/informal tone suit their expectations? How much text is appropriate?
- Melody: How will the overall design pattern appear pleasant and predictable to users and move them emotionally?
- Decoration: How will you present everything so that the graphics match the atmosphere that users can perceive? Would a classic design or a stylish, niche layout meet their expectations?
- Spectacle: How can you make your design remarkable so that users will remember it?

An example of using these elements could be the following story: A persona named "Rick", a 47-year-old manager who struggles with his work–family life balance. He even works on his commute. He feels drained and wishes he had better control over his life.

Create a plot with conflict to make the characters heroes and imagine how they can overcome specific problems using your design. Do this as a short journey or storyboard with each persona's goals clearly defined. For example:

- Rick discovers your (not yet designed) time management app online. He downloads it and fills out your questionnaire about work commitments, family, expenses, etc.
- He begins to use your app and allows it to collect data from his phone and fitness tracker about time spent on various tasks/activities, stress levels, sleep patterns, etc.

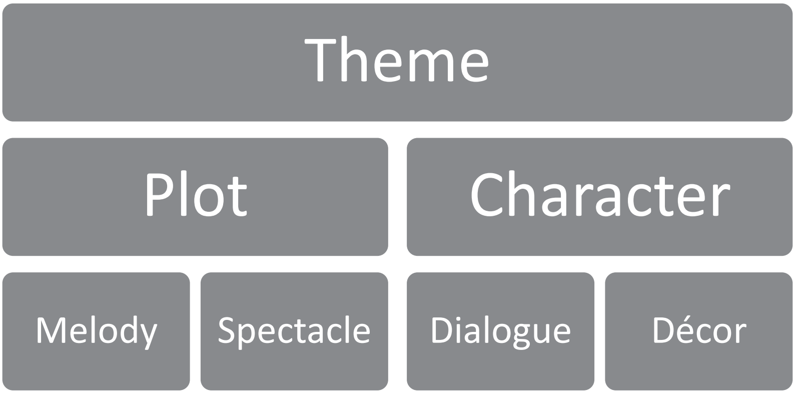


FIGURE 3.4 Aristotle's story components

- After a week, your app shows him his tasks and activities in diagrams, including sleep, heart rate data, etc.
- By pressing a button on his phone, Rick sees suggestions for time management, such as how he can become more productive, relaxed, etc.
- He has the option to continue or pause monitoring (e.g. if he's on vacation).

Give your solution a supporting role – show how it improves your persona's/user's life and how easy it is to use. Consider, for example, how many steps Rick needs to take to use your app and whether voice-controlled devices at home can affect its suggestions.

Work with settings – when and where users use your design is crucial for building empathy. For Rick, it's home, commuting, and the workplace. But what about traveling professionals who work from home?

Customize appearance/feeling – the appearance of your design is important regardless of its functional advantages, so design the most suitable (e.g.) layout, colors, and typography. For example, Rick prioritizes an intuitive design, but calming colors would complement larger font sizes, etc.

Pilot testing or pilot project – this means testing your solution on a limited target group. The solution should be fully functional but can be considered a "minimum viable product" (MVP). The purpose of a pilot test is to identify potential problems and flaws in a real-life situation before launching the final version.

In connection with design thinking, pilot tests are used to test a prototype of a product or service in a real-life situation to identify potential problems and flaws and improve the prototype before the final version is launched. Pilot testing is like the earlier testing of ideas, but with the significant difference that we only expect adjustments and not full rejection of our solution – although that is also possible.

A pilot test can be conducted in different ways depending on the product or service and the purpose of the test. Here are some steps that can be followed to conduct a pilot test:

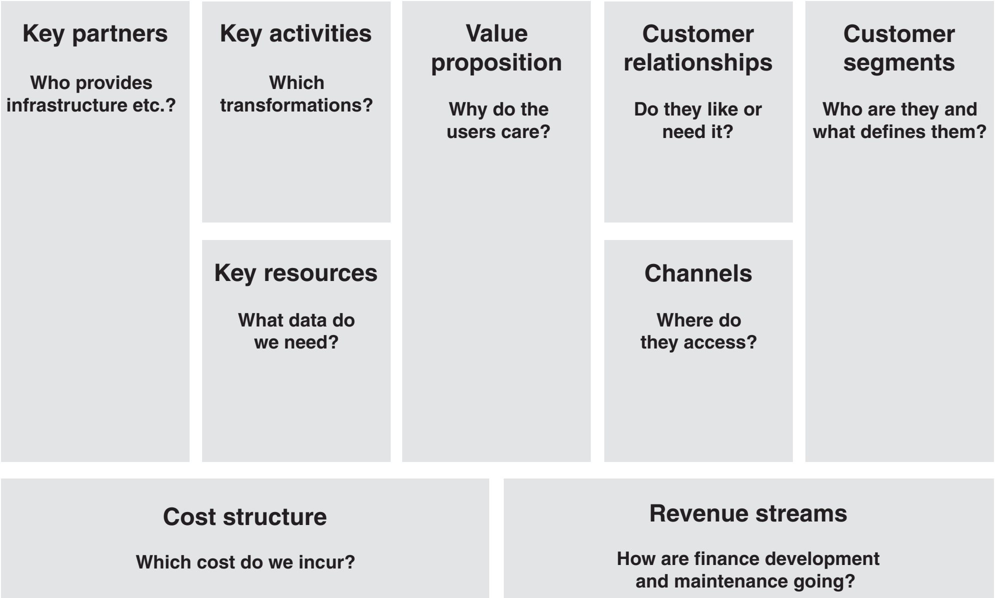
1. Identify the purpose of the test: What do you want to test? What is the purpose of the test?
2. Identify the test group: Who will you test the product or service on? How will you recruit the test group?
3. Identify the test environment: Where will you conduct the test? How will you create a realistic scenario?
4. Develop a test plan: How will you conduct the test? What are the steps involved in the test?
5. Conduct the test: Conduct the test according to the test plan.
6. Collect feedback: Collect feedback from the test group on their experience with the product or service.
7. Analyze feedback: Analyze the feedback to identify any issues or problems.
8. Make adjustments: Make adjustments to the product or service based on the feedback from the pilot test.

Business model

The last step in the design thinking process is to create a business model around the solution. This is where all the data comes together in a coherent solution. It is important to find resources to bring the solution to life. The Business Model Canvas is a useful tool for this purpose, as it provides a visual overview of the elements that describe a company's value proposition, infrastructure, customers, and financials. It helps businesses adjust their activities by illustrating their business model in an easily understandable way. It also ensures that all the data is available for deciding about an innovation. That can be a data product or any other solution.

The Business Model Canvas consists of nine building blocks, which cover the three main areas of a business: desirability, feasibility, and viability. The nine building blocks are:

1. Value Proposition: What is your product or service offering? What sets it apart from other products or services on the market?
2. Customer Segments: Who are your customers? What are their needs? How can you reach them?
3. Channels: How do you reach your customers? Which channels do you use to reach them?
4. Customer Relationships: What kind of relationship do you have with your customers? How do you build trust and loyalty with them?
5. Revenue Streams: How do you make money? What prices do you charge for your product or service?
6. Key Partners: Who do you work with to create value? Which partners are crucial for your business?
7. Key Activities: What are the most important activities for creating value? What do you need to do to deliver your product or service to your customers?



Key partners
Who provides infrastructure etc.?

Key activities
Which transformations?

Key resources
What data do we need?

Value proposition
Why do the users care?

Customer relationships
Do they like or need it?

Channels
Where do they access?

Customer segments
Who are they and what defines them?

Cost structure
Which cost do we incur?

Revenue streams
How are finance development and maintenance going?

FIGURE 3.5 Business model canvas illustration, with data product adaptation

8. Key Resources: What resources do you need to create value? Which resources are essential for your business?
9. Cost Structure: What costs do you need to cover to create value? What costs are associated with running your business?

If you are interested in digging deeper into the "Business Model Canvas", you can find a PDF version of the original book by Alexander Osterwalder on his website and register there (Osterwalder, 2010).

This entire process should lead you to the identification of some of the potential data products. These can then be detailed out using the Data Product Canvas and a Data Value Chain Model. Sometimes these descriptions can be used as your prototype.

DATA PRODUCT CANVAS

According to J. Majchrzak et al. (2022), a "*data product is an autonomous, read-optimized, standardized data unit containing at least one dataset (Domain Dataset), created for satisfying user needs*".

Professor Leandro Carvalho (Carvalho, 2022) took inspiration from the widely used Business Model Canvas (BMC) by Osterwalder (Osterwalder, 2010) when creating his Data Product Canvas.

The core idea is to create innovative products that get used by tying the solution to the business. This is done in three areas, product vision – backend (1–4), vision of strategy – front end (5–7), and business vision –management (8–10). These areas are then divided into ten blocks, which we will

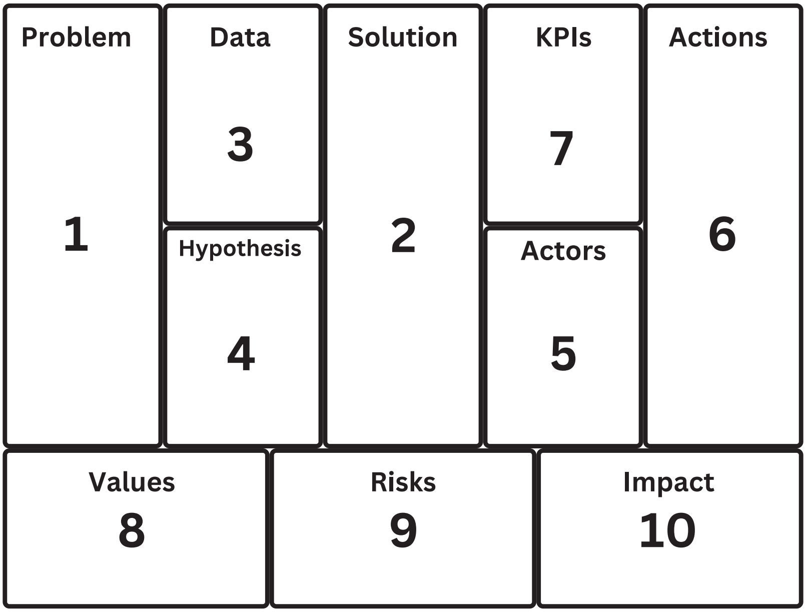


FIGURE 3.6 Data Product Canvas

Source: adapted from Carvalho 2022.

dive into in a moment, but first, it is important to realize that the work with the model starts with the problem. That means if you cannot identify a distinct problem to solve you should not even embark on filling the rest of the Canvas and build a data product. It is the same as with the BMC where you shouldn't start a business unless people are feeling an "important enough" problem.

Like the BMC, the Data Product Canvas is *not a static model* but a board that should be reverted to and updated continuously for the duration of the product life.

Problem definition

The problem definition is the most critical step in the data product discovery process. It ensures that you are solving the right problem with the right solution. Therefore, as a data product lead, you need to guide your team and stakeholders to focus on the problem first, before jumping to any solution ideas or hypotheses.

A well-defined problem is a problem that is clear, specific, and agreed upon by all the relevant parties. You could consider using SMART[2] goals. It also aligns with the expected outcomes and goals of the project. A well-defined problem can help you fill in the rest of the Canvas and even change or cancel the project if necessary, as mentioned before.

To define the problem, you should use the following three questions:

- What is the problem?
- Why is it a problem?
- Whose problem is it?

These questions will help you understand the nature, causes, and impacts of the problem, as well as the target audience and stakeholders. For each answer, you should ask at least three more follow-up questions to dig deeper and uncover the root cause of the problem. This technique is called the "5 Whys" and it can help you avoid superficial or inaccurate problem definitions and ensure that you are addressing the root cause.

The solution that will be adopted

A data product should solve a business problem, not a technical problem. Therefore, it is important to focus on the practical solution to the problem rather than falling in love with fashionable terms such as deep learning, neural networks, big data, and generative AI. For example, it is not always necessary to use deep learning when logistic regression works or to use big data if you don't even do the basic statistics of your data. Sometimes, a data product is as simple as searching for the data itself because if you don't have data, you should start there.

To identify the simplest and most objective solution possible for solving the problem, you should ask three questions:

- What kind of solution will be adopted? (Ex.: analytics, machine learning, AI, etc.)
- What will be the solution? For example:
    - If the adopted solution is machine learning, we must take into account that: for each problem, we have different approaches; for each approach, several algorithms; for each algorithm, several parameterizations. That is, there is not and never will be a "best

algorithm" for a given problem. But in any case, having the mapping of what you want will guide the development of the project.

- What is expected of the solution? What would the outputs of the product be?
    - E.g.: A report with the final product of an analysis? A specific prediction about a data type?

At the end of each of these questions, always ask "Why?" three times to dig deeper and uncover the root cause of the problem. Remember to circle back to the first block of the Canvas and establish clearly and objectively the problem you are solving. This is a crucial and non-negotiable point. The purpose of this (lean methodology) is to test fast, fail fast, and adjust fast, so you don't end up having to deal with too much legacy.

Data mapping

As a data product leader, you need to be aware of the challenges and opportunities of working with data projects. One of the key aspects of data projects is the data discovery and management process, which involves finding, accessing, transforming, and validating the data sources for your product. This is explored in detail in Chapter 5 on data sources.

The data discovery and management process is applicable at every point in the project life cycle. It affects the quality and reliability of your product outputs and outcomes. Therefore, you need to follow some steps that will help you ensure that your data sources are adequate and appropriate for your product goals. These steps are:

1. Identify the source of the data, i.e., where the data comes from. Example: Is it on a system? Is it a set of files? Does it have structured formatting?
2. Assessing the quality of the data, i.e., how accurate, complete, consistent, and relevant is the data for your analysis and models.
3. Verify the access and availability of the data, i.e., whether you have permission and the capability to access and use the data.
4. Define the process and transformation of the data, i.e., how you will read, clean, integrate, and manipulate the data for your product. The technical aspects of this are discussed in Chapter 8 where the data stack is described.
5. Specify the output formats of the data, i.e., how you will present and deliver the data to your stakeholders and customers. This is explained in further detail in Chapter 6 on data communication and visualization.
6. Establish the test, training, and validation strategies for the data, i.e., how you will split, sample, and evaluate the data for your product.

Remember, sometimes the data product is the data capture itself. After all, if you don't have the data or access to it, you won't be able to develop your product.

The hypotheses that will be tested

Hypothesis testing is a vital step in the data product discovery process. It helps you validate that the proposed solution will address the real problem identified and deliver value to the business.

Therefore, as a data product manager, you need to guide your team and stakeholders to formulate and test a set of hypotheses that will monitor the performance and impact of your product.

A hypothesis is a tentative statement that expresses a relationship between variables or outcomes. These should preferably be gathered in qualitative interviews with business stakeholders. They can be tested empirically using data and evidence. To formulate and test hypotheses, you should ask three questions:

- What are the hypotheses we want to test?
- What are the expected responses for each of them?
- What to do from each answer? In other words, what strategy should we follow?

For example, in the case of an e-commerce company, it is concluded that it has a churn[3] problem (bottom of the funnel) and not a customer attraction problem (top of the funnel). Therefore, we could ask: Will the proposed solution reduce the churn rate? With the proposed solution, is it possible to predict the customers who are thinking of leaving?

By answering these questions, we can complete the first domain area of the data product discovery process, which is planning the product vision.

This is the basic definition of what our product will be and how it will solve the problem.

All actors (customers and stakeholders) involved

The second domain area of the Data Product Canvas is the product strategy view, which is the forward-looking part of filling the Data Product Canvas. It guides us in the elaboration of the actions that must be implemented when the final solution is ready. This ensures that the solution continues to be used after its development, as we will already know in advance what to do with the created product.

To develop the product strategy view, it is important to identify all actors that will have some involvement with the product. For each identified actor, we must validate what was understood in the first domain area; that is, we must validate with them the problem definition, the identification of the solution, the mapping of the data, and the hypotheses that will be tested.

The mapping of actors must be done continuously during the product life cycle. Therefore, we must ask multiple times:

- Who is the sponsor/project/product owner?
- Who is the final user of the product?
- Who are the interested parties and stakeholders during development?
- Who will pay for the solution?
- Who will be impacted by the solution after launch?

It is important to assess whether the project should go ahead if you cannot find an organizational sponsor for the product that has proper internal weight. This is a warning signal that should not be overlooked. The bigger the business impact, the higher the organization the sponsor should be. However, if you cannot identify the product's client and their linked problem, you should immediately stop the project. After all, there is no product if there is no customer. Chapter 6 on data communication provides several frameworks that can be used to deepen this stakeholder analysis.

The strategic actions that will be developed

The action section is where you bring the data product to life and define the strategic actions that will be derived from the use of the product. That means you are looking at decisions/actions that are taken because of the product, not the product itself.

In this area, the stakeholders play a crucial role in guiding the product development, while the data product lead plays a supporting role in recording and facilitating. This ensures that the product will be adopted by the business unit and will not be abandoned or neglected shortly after its development.

An example: Suppose that the e-commerce company has developed a predictive model of churn prevention. Once the product has been delivered and is ready for use, what will the business do with it? What actions will be taken? Given the signal that a customer is likely to leave the company, what can be done to retain him or her? Will it be an automatic action (email marketing) or a human action (personal contact)?

To ensure that your product is not forgotten in a "drawer" or rather, on a cloud server costing the company money, you must ensure, even before the start of product implementation, that the business area will be able to map the strategic actions that will be executed.

Therefore, you should ask:

- What actions will be taken?
- Which communication campaigns should be created?
- How to generate value for the business from the use of the data product developed?

We will later take a look at data product portfolio management to help guide the process across multiple products that arise quickly in companies that embark on the data-driven decision-making journey.

The KPIs that should be monitored

The "Key Performance Indicators" (KPI) view is the area where you define the indicators that will help you monitor and evaluate the quality and impact of your product. These indicators are essential for data-driven decision-making, as they provide feedback and evidence on the performance and value of your product.

There are two types of indicators that you need to consider: technical and business.

Technical indicators measure the quality of the developed product, such as how accurate, reliable, and robust it is. Remember to set a procedure to assess model drift if you are using "black box" neural network models, as their performance will tend to change over time. For example, if your product is a machine learning model for classification, you can use metrics such as accuracy, precision, recall, or F1-score to measure its technical quality.

Business indicators measure the impact of the product on the business outcomes, such as how much revenue, customer satisfaction, or retention it generates. To measure its business impact, you can use metrics such as conversion rate, churn rate, customer lifetime value, or net promoter score.

Depending on the type and objective of your product, you will need to choose different metrics for each type of indicator.

To define the product metrics view, you could ask the following questions:

- How to evaluate the quality (technical/business) of the finished product?
- What metrics should be used/can we act upon it?
- How to measure the results of the actions that will be derived from the use of the product?
- If it includes A/B testing, how do we ascertain the "winning" side?
- How much uncertainty can we deal with?

Keep in mind that no data product is 100% effective/accurate. After all, the past (the data we build on) is not a mirror to the future. Therefore, every data product manager must identify, together with its main stakeholders, how much uncertainty the company can tolerate based on the results that will be produced by the data product in question. This will help you set realistic expectations, manage risks, and set the limits for retiring a data product.

The values (the size of the problem)

The value is where you estimate the magnitude and impact of the problem that you are trying to solve with your product. This is essential for prioritizing the development efforts and assessing the feasibility of the project, that is, whether the benefits or savings generated will justify the execution of what was planned. To estimate the problem size, you should ask three questions:

- How big is your problem?
- What is the baseline?
- What is the expected benefit or savings from using the product?

For each question, you should ask "Why?" at least three times to dig deeper and uncover the root cause and the potential value of the problem. For example, if your problem is customer churn, you can ask:

- How big is your problem? (Ex. 20% of customers leave every year.)
    1. Why? (Ex. Customers are dissatisfied with the service quality or price.)
    2. Why? (Ex. Customers have more options or better offers from competitors.)
    3. Why? (Ex. Customers are not loyal or engaged with the brand.)
- What is the baseline? (Ex. The industry average churn rate is 15%.)
    1. Why? (Ex. The industry has high competition and low switching costs.)
    2. Why? (Ex. The industry has low differentiation and commoditization.)
    3. Why? (Ex. The industry has low innovation and customer focus.)
- What is the expected benefit or savings from using the product? (Ex. Reducing the churn rate by 5% would increase revenue by $1 million per year.)
    1. Why? (Ex. Retaining customers is cheaper and more profitable than acquiring new ones.)
    2. Why? (Ex. Retaining customers increases customer lifetime value and referrals.)
    3. Why? (Ex. Retaining customers builds trust and loyalty with the brand.)

By asking such questions, you can quantify and qualify the problem that you are trying to solve with your product and evaluate its feasibility and priority.

It should also be used when assessing which solution to go with. If using an advanced non-transparent model only provides a 2% lift from the baseline you should probably not add the complexity and risk to your solution.

The risks

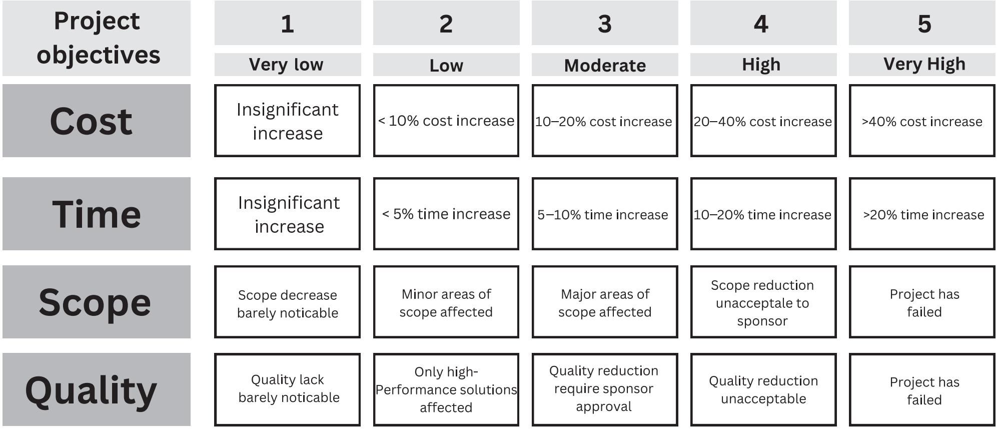
The risk management view is the area where you identify and map the potential risks that could affect the development, implementation, and usage[4] of your product. Risks are uncertain events or conditions that could hurt your product goals, such as quality, performance, or value. Risk management is essential for ensuring complete and healthy planning throughout the development journey and for mitigating or avoiding the adverse effects of risks. Therefore, you should always keep this in mind and maintain a list of the main identified risks. To identify and map the risks, you can start by asking two questions for the development:

- What are the risks?
- What could these risks block during product development?

For example, some common risks for data products are:

- Data quality issues, such as missing, inaccurate, inconsistent, or outdated data
- Data access issues, such as lack of permission, availability, or security of data sources
- Data privacy issues, such as compliance with regulations, ethical standards, or customer expectations
- Technical issues, such as bugs, errors, failures, or performance issues of the product
- Business issues, such as changes in requirements, expectations, or priorities of the stakeholders or customers.

These risks could then be scored according to Figure 3.7:



Project objectives	1	2	3	4	5
	Very low	Low	Moderate	High	Very High
Cost	Insignificant increase	< 10% cost increase	10–20% cost increase	20–40% cost increase	>40% cost increase
Time	Insignificant increase	< 5% time increase	5–10% time increase	10–20% time increase	>20% time increase
Scope	Scope decrease barely noticable	Minor areas of scope affected	Major areas of scope affected	Scope reduction unacceptale to sponsor	Project has failed
Quality	Quality lack barely noticable	Only high-Performance solutions affected	Quality reduction require sponsor approval	Quality reduction unacceptable	Project has failed

FIGURE 3.7 Project objective risk matrix

And then be structured in a risk matrix as shown in Table 3.1 by assessing its likelihood and impact.

The dark gray areas show the risks that need to be addressed before they happen, while the white can most likely be tolerated.

For each risk, you then create a mitigation or contingency plan. For example, if your risk is data quality issues, you can ask:

- How likely is it that the data quality will be poor? (Ex. High, Medium, Low)
- How much impact will it have on the product quality and value? (Ex. High, Medium, Low)
- How can we mitigate or avoid this risk? (Ex. Perform data cleaning, validation, and integration before using the data for the product.)
- How can we deal with this risk if it occurs? (Ex. Report and resolve the data quality issues as soon as possible.)

This risk response chart is then used and updated at each status meeting.

When the data product moves from development to production the risk management changes slightly. There is still a need to assess the potential impact of risk occurrence, but the focus moves from project risk to business risk.

The business risks of a data product failing can be significant and diverse. Some of the common risks are:

- Financial risks, such as loss of revenue, increased costs, or legal liabilities
- Reputation risks, such as loss of trust, credibility, or brand value
- Operational risks, such as disruption of services, processes, or systems
- Compliance risks, such as violation of regulations, standards, or policies
- Security risks, such as data breaches, cyber-attacks, or privacy breaches.

The impact and likelihood of these risks will depend on the type and objective of your data product, the quality and quantity of your data sources, the accuracy and reliability of your models and algorithms, the effectiveness and efficiency of your processes and systems, and the expectations and requirements of your stakeholders and customers. To manage these risks, you need to identify them early and proactively, assess their likelihood and impact, define

TABLE 3.1 Risk assessment matrix

Probability	5		▨	▨	■	■
	4		▨	▨	■	■
	3			▨	▨	■
	2			▨	▨	■
	1				▨	▨
		1	2	3	4	5
	Impact					

mitigation or contingency plans, monitor their status, and progress, and communicate them transparently to your stakeholders just like the project risks. You could also benefit by establishing a risk management culture that promotes awareness, accountability, and continuous improvement.

The performance and/or impacts of the product on the business (values generated or saved)

The impact area is where you infer the value and feasibility of your product to the company executives. It involves estimating the impact and performance of your product on the company's goals, such as revenue, profit, customer satisfaction, or market share. These are the areas that you must monitor in your continuous production-level risk assessment. The impact section is essential for securing the support and resources for your product development, implementation, and continued maintenance. It also helps you monitor and evaluate the value and impact of your product throughout its life cycle. To fill the impact section, you should start by asking the following three questions:

- What is the impact on the business?
- How to measure it?
- Where and how can we see this improvement or impact/performance?

For example, if your product is a recommender system, you may ask:

- What is the impact on the business?
    - Ex. Increasing sales, cross-selling, and up-selling by providing personalized recommendations to customers.
- How to measure it?
    - Ex. Using metrics such as conversion rate, average order value, customer lifetime value, or revenue per visitor.
- Where and how can we see this improvement or impact/performance?
    - Ex. Using dashboards, reports, or analytics tools that show the results of the recommender system compared to a baseline or a control group.

By answering these questions, you can create a compelling and data-driven business case for your product that demonstrates its value and feasibility to the company executives.

THE INNOQ DATA PRODUCT DESIGN BOARD

Another model was developed by INNOQ and takes a more technical approach to data products, which means that the two approaches can be combined, and this acts as the data and solution elements of the Data Product Canvas.

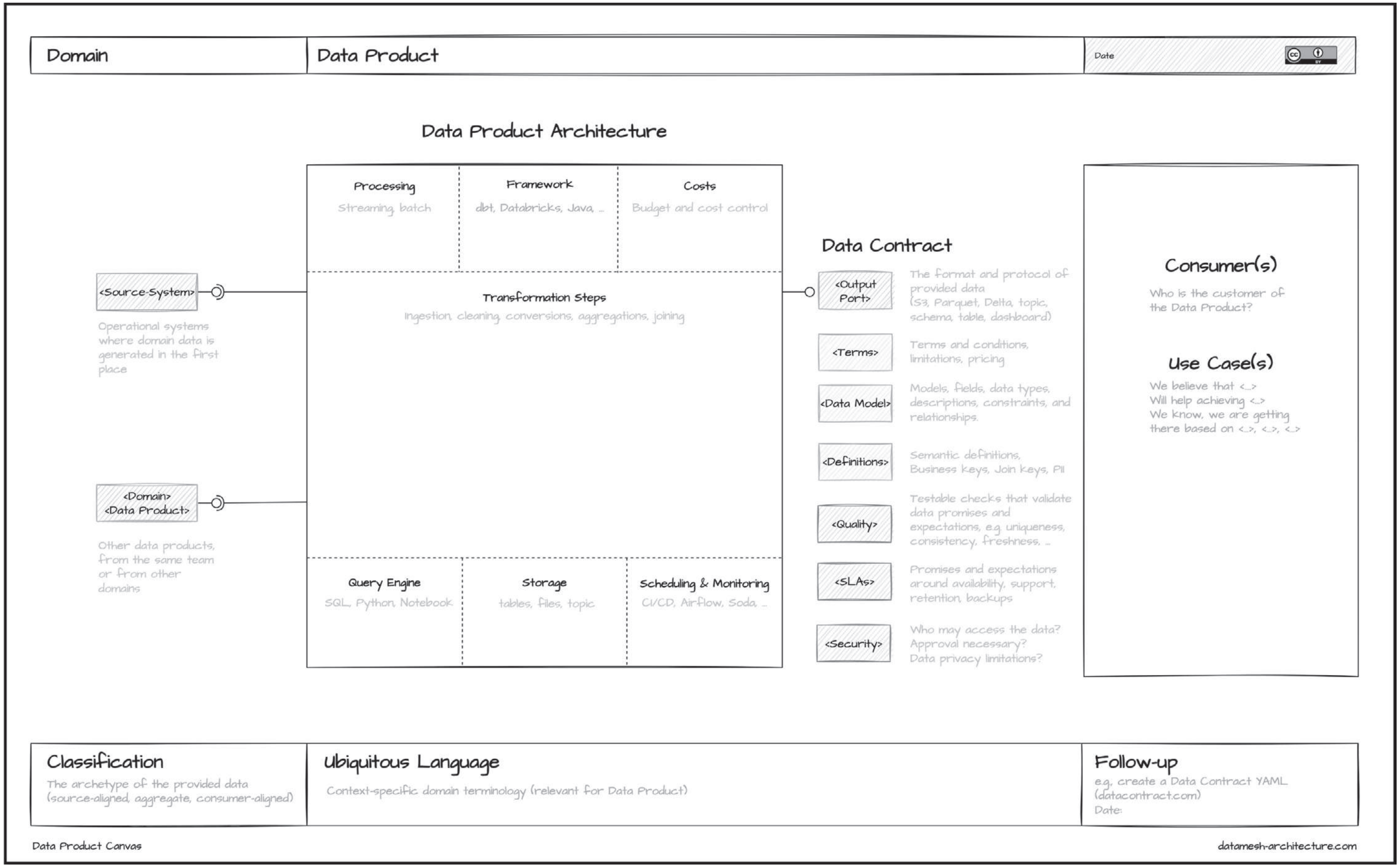

Domain
Data Product
Date
Data Product Architecture
Processing
Streaming, batch
Framework
dbt, Databricks, Java, ...
Costs
Budget and cost control
Transformation Steps
Ingestion, cleaning, conversions, aggregations, joining
<Source-System>
Operational systems where domain data is generated in the first place
<Domain> <Data Product>
Other data products, from the same team or from other domains
Query Engine
SQL, Python, Notebook
Storage
tables, files, topic
Scheduling & Monitoring
CI/CD, Airflow, Soda, ...
Data Contract
<Output Port>
The format and protocol of provided data (S3, Parquet, Delta, topic, schema, table, dashboard)
<Terms>
Terms and conditions, limitations, pricing
<Data Model>
Models, fields, data types, descriptions, constraints, and relationships.
<Definitions>
Semantic definitions, Business keys, Join keys, PII
<Quality>
Testable checks that validate data promises and expectations, e.g. uniqueness, consistency, freshness, ...
<SLAs>
Promises and expectations around availability, support, retention, backups
<Security>
Who may access the data? Approval necessary? Data privacy limitations?
Consumer(s)
Who is the customer of the Data Product?
Use Case(s)
We believe that <...>
Will help achieving <...>
We know, we are getting there based on <...>, <...>, <...>
Classification
The archetype of the provided data (source-aligned, aggregate, consumer-aligned)
Ubiquitous Language
Context-specific domain terminology (relevant for Data Product)
Follow-up
e.g. create a Data Contract YAML (datacontract.com)
Date:
Data Product Canvas
datamesh-architecture.com

FIGURE 3.8 INNOQ – Data Product Canvas (www.datamesh-architecture.com/data-product-canvas)

Data products are the core components of data-driven decision-making. They are how data is transformed into insights and actions that create value for an organization. To design and develop effective data products, we need a systematic approach that aligns the data product vision with the consumer needs and technical capabilities. In this chapter, we introduce the Data Product Canvas, a visual tool that helps you specify the key aspects of a data product in a concise and structured way.

The INNOQ Data Product Canvas consists of nine building blocks that cover the following dimensions of a data product:

- **Domain:** The domain defines the scope and ownership of the data product. It answers questions such as: Who is accountable for the data product? Who specifies its requirements? Who will answer questions about the data product? Who fixes it when it breaks?
- **Data product name:** The name is a unique identifier for the data product within an organization. It should follow a consistent naming convention that reflects the domain and the purpose of the data product.
- **Consumer and use case(s):** This building block describes the target audience and the value proposition of the data product. It answers questions such as: Who will use the data product and for what purpose? What are the analytical use cases and organizational objectives that the data product supports?
- **Output ports:** The output ports define how the data product delivers its value to the consumers. They specify the format and protocol of the data output, such as a database table, a file, an application programming interface (API), or a visualization.
- **Metadata:** The metadata provides additional information about the data product that helps to understand, access, and use it. It includes elements such as ownership, data schema, semantics, security, and versioning.
- **Input ports:** The input ports describe the sources and types of data that feed into the data product. They specify the format and protocol of the data input, such as operational source systems or other data products (internal or external).
- **Data product design:** This building block describes the logic and processes that transform the input data into the output data. It covers aspects such as data ingestion, storage, transport, wrangling, cleaning, transformations, enrichment, augmentation, analytics, sales qualified lead (SQL) statements, or data platform services.
- **Observability:** This building block describes how to monitor and measure the performance and quality of the data product. It covers aspects such as quality metrics, operational metrics, and service level objectives (SLOs).
- **Ubiquitous language:** This building block describes a common language that is shared between everyone involved in the project. It is usually a context-specific domain terminology that is relevant for operational systems and data products.
- **Classification:** This building block specifies the nature of the exposed data. We classify our data product as either source-aligned, aggregate, or consumer-aligned.

The Data Product Canvas is designed to be used collaboratively by a cross-functional team that includes domain experts, data engineers, data scientists, and product owners. The recommended way to use it is to fill out each building block in sequence during a workshop session. The following sections provide more details on each building block and how to complete them.

DATA VALUE CHAIN MODEL

With the maturing of data operations in companies, they also start to need an operating model that can focus and structure work processes in a way that can be generally recognized across the entire business. Resource allocations are also more readily available when "the business" understands the implications of the data (product) requests.

Fortunately, we have such a model to build upon. Back in 1985, Michael Porter spawned his value chain model to describe how a business operates, a model that today is still part of the core curriculum in most business schools. As a general model, it has many challenges, but most people in business will readily recognize its elements, base concept, and structure (for supply chain management (SCM) process analysis recommend the supply chain operations reference (SCOR) model instead).

The natural arena for Porter's model is the production company with primary activities focused on delivering (mass) products and supporting activities that enable that production.

The data solutions can be described using the same terminology and by doing it that way we ensure the full operating model is accounted for.

Primary activities

Let's start with the primary activities that move raw data to value-created products. Deployment is the last stage in the CRISP-DM model (**Cr**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining).[5] With the movement of the analytics products and models from the development phases the primary activities start.

Inbound logistics is called ingestion in the data world. How are we getting information, where are we getting it from, and who is responsible for the quality are questions that need to be addressed. Some things can be handled "contractually" with data contracts in "procurement", but there still needs to be continuous checks like with all other raw materials for production.

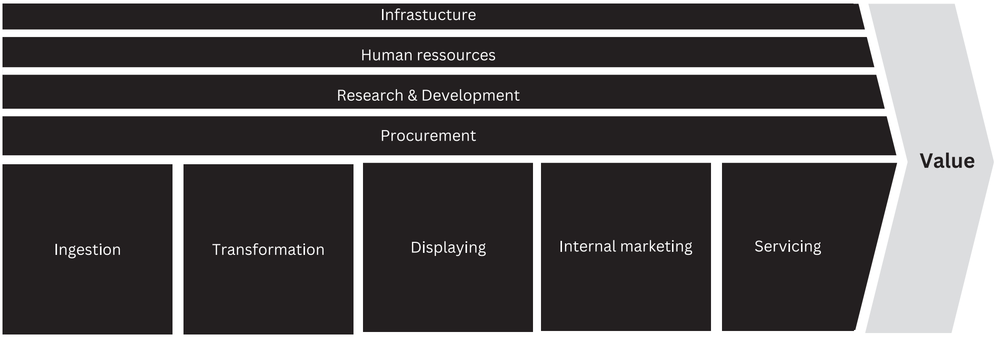


FIGURE 3.9 Data value stream

Production (transformation) is the transformation of data. Depending on your data stack (ETL/ELT+data warehouse or direct[6]) there might need to be a cleaning process here. The output is the model which can be AI/ML-based or just a simple summary column in an Excel sheet. Often these will be placed in dashboards or standardized reports.

Outbound logistics (displaying) is the medium for communication, that being internal websites, email, meetings, etc. Data visualizations are widely recognized as useful, but if the users cannot find the information they need and digest it you are missing out on the value creation of your data product.

Sales & Marketing (internal) are closely related to data literacy and data advocacy. Like with the communications channels above it's all about ensuring the value creation for the data product. You can, with management support, try to shove your data products down the throats of your users, but it will never become a long-term sustainable success that way. You can watch my YouTube videos on selling and decision-making if you want to dig deeper into this area.

Servicing is an often-forgotten element when providing data products, or if not forgotten then a place where the analyst ends up swamped by doing tech support. The use of service level agreements (SLAs) and service level objectives (SLOs) can be used to align expectations. Sometimes data contracts also become relevant here if you have stakeholders that are manipulating data themselves.

Supporting activities

These are the activities needed to develop and run your data products. Some areas might naturally be outsourced to the base organization.

Infrastructure is the "stack" and the domain of the data engineers and the IT department. Responsibility can therefore be outsourced to them but remember to set your servicing level requirements like you spoke about in the servicing section. A part of the infrastructure that cannot be outsourced is the data dictionary/meta-data description which is owned by infrastructure. Part of your governance that ensures the rights management should also be managed here.

HR is similar to the traditional model and often outsourced to the base organization but with the added complexity of covering roles not limited to the data team, but also "external"/ embedded analysts. Remember that the embedded analysts will often be pulled in multiple directions like the classic matrix organization and will therefore need to have quite a strong individual mandate to prioritize tasks.

R&D is where the solutions are built and tested before they are deployed to production. CRISP-DM which was mentioned earlier can be a method to focus on here or using DATAOPS as the framework. I'd make some separate materials related to this later.

Procurement is about data governance and data contracts. If you are in regulated industries like finance/health/medical/ or need to follow the US Sarbanes Oxley regulation, this is where your lineage documentation should reside. Data contracts with the data owners will need to be enforced. That doesn't necessarily mean that you are going to have written contracts with all the paragraphs and clauses, but you need to address critical things such as availability, updates, and how schema changes are handled.

DATA PRODUCT PORTFOLIO MANAGEMENT

Over time any organization will get more and more data products into the production environment. This can be a dashboard for management, data pipes for suppliers and customers, algorithms for production equipment maintenance, etc.

If this proliferation of data products is not managed the company will end up with an unmanageable swamp of solutions where the maintenance task will eclipse the value creation and support the development of the business.

Data product portfolio management (DPPM) is created to handle that. "DPPM does not exist in a vacuum—by its nature, DPPM is closely related to and interdependent with enterprise architecture practices like business capability management and project portfolio management. DPPM itself may therefore also be considered, in part, an enterprise architecture practice" (Mayrhofer, 2023).

To create a mapping of data products in the portfolio we will use an adaptation of the well-known BCG Growth Share Matrix.[7] This will categorize our products in two dimensions:

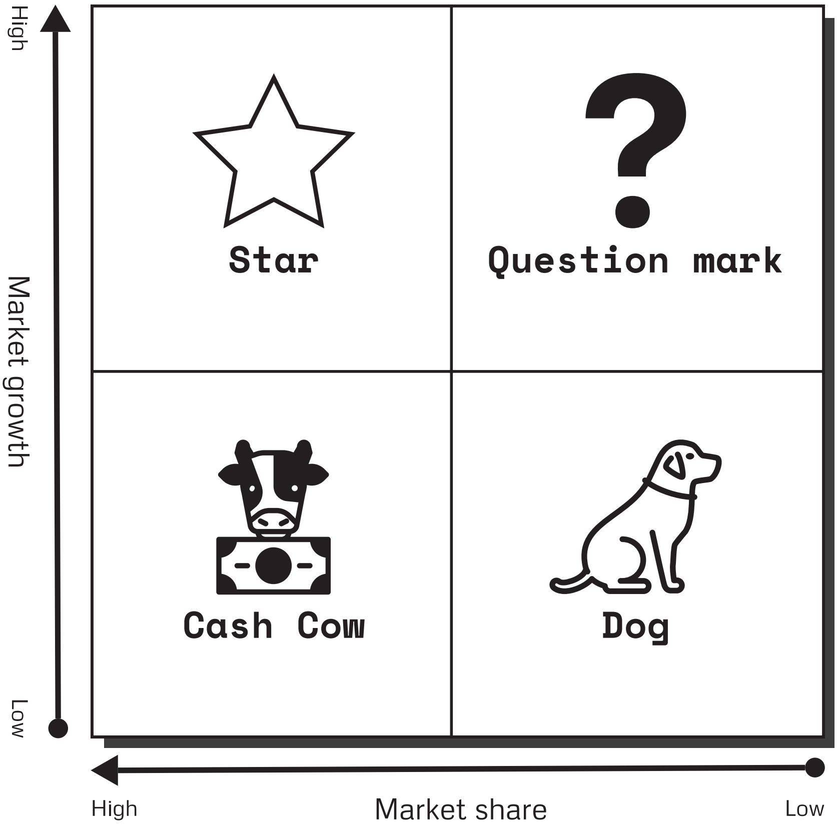


FIGURE 3.10 Product portfolio assessment

the market share, or usage penetration in the addressable market and the growth, and strategic impact potential. We need this update on the dimensions as some of our data products are entirely internal, but we still want to assess them in a framework that is easily explainable and widely recognized by all stakeholders.

Bruce Henderson wrote in "The Product Portfolio" (Henderson, 1970): "*A company should have a portfolio of products with different growth rates and different market shares. The portfolio composition is a function of the balance between cash flows*". Since then, the focus on cash flows has been supplemented with value creation and expanded from dealing with strategic business units (SBUs) in conglomerate companies. The speed of business has also changed, so business, and therefore products, spend less time in each quadrant. The profitability has also shifted towards stars that exhibit both high market share and high growth (Reeves, 2014).

To get it properly adapted to the world of internal data product portfolio management, we will start by adjusting the two scales to reflect a focus on value creation instead of direct profitability. The profitability of the solution should, however, be an effect of the focus on value creation. We are focusing on the internal solutions as customer-facing solutions should be handled as all other products with a focus on profitability.

The market we define for our internal solutions is the usage by the decision-makers that could base their work on the data solutions.

- What are the key data domains within the organization? What are the key data products within these domains needed to solve current business problems? How do we iterate on this discovery process to add value while we are mapping our domains?
- Who are the consumers in our organization, and what logical, regulatory, physical, or commercial boundaries might separate them from producers and their data products?
- How do we encourage the development and maintenance of key data products in a decentralized organization?
- How do we monitor data products against their SLAs, and ensure alerting and escalation on failure so that the organization is protected from bad data?
- How do we enable those we see as being autonomous producers and consumers with the right skills, the right tools, and the right mindset to want to (and be able to) take more ownership of independently publishing data as a product, and consuming it responsibly?
- What is the life cycle of a data product? When do new data products get created, and who is allowed to create them? When are data products deprecated, and who is accountable for the consequences to their consumers?
- How do we define "risk" and "value" in the context of data products, and how can we measure this? Whose responsibility is it to justify the existence of a given data product?

KRIFA CASE

Kristelig Fagforening, also known as Krifa, is a cross-professional union in Denmark that accepts both wage earners and self-employed individuals.

Krifa is known for its affordable prices and independence. It fights against exclusive agreements of trade unions and instead prefers a more individualized approach. Due to this stance, Krifa is often referred to as a "yellow union", which is considered an ideological counterpart to the more well-known "red" unions.

The mission of Krifa is to fight for the well-being of each member in their everyday life. As a member, you can also get unemployment insurance (a-kasse) and wage insurance (lønsikring) from Krifa, which gives you the right to unemployment benefits, legal coverage, advice, and subsidies for education and courses.

Krifa offers two types of memberships: Krifa Basis and Krifa Plus. Both memberships provide access to unemployment insurance and thus the right to unemployment benefits. They also offer job security under favorable conditions, as well as assistance with job search and career guidance.

Krifa has data-products for their members. Based on a modern data warehouse, which allows them to make better decisions based on a common ground, they have access to a wealth of information about the members and the labor market in Denmark. The system contains data such as statistics on members, time spent on cases, absences, financial data, and member data which they compile for various statistical purposes. This data is collected in an automated way, making it easy for Krifa's employees to create reports and dashboards in Power BI. These reports provide insight into the business easily and intuitively. They also have a digital human, Aida, powered by AI, installed as a kiosk in Krifa's reception. Members can here interact with this digital receptionist seamlessly.

While the AI receptionist works ok, it does require that the members show up at their physical location and only has limited functionality related to "reception work". They are therefore looking into creating a purely digital version. As a lot of the information is highly personal and sensitive at Krifa, it is also the core reason they can be useful to their members, so just exposing information "freely" is not a practical option.

Create a Data Product Canvas which can be used by the management to decide on a way forward for their AIDA project.

SUMMARY

Data products are solutions built around data that solve problems or create opportunities. They can be designed for internal or external use. This chapter explores how to design and build data products using design thinking and data-focused frameworks like the Data Product Canvas.

Design thinking is a human-centric approach to problem-solving that focuses on understanding user needs. Key steps include:

In the inspiration phase, the focus is to define target users and understand their needs through interviews and research. Tools like analogous inspiration and extremes and mainstreams reveal insights.

In the ideation phase, the brainstorming of solutions starts. Techniques like worst possible idea, SCAMPER, role-playing, and picture cards spark creativity.

The implementation phase builds prototypes to test ideas. Types of prototypes include sketches, 3D printing, wireframes, virtual reality, and suggested models. Then user testing is conducted to refine prototypes using goals and metrics. Steps include choosing test methods, running tests, analyzing feedback, and iterating.

Storytelling techniques are then used to communicate the solution to stakeholders and create engagement. Elements include plot, character, theme, dialogue, melody, decoration, spectacle. The next chapter dives deeper into storytelling.

The Data Product Canvas is a framework for designing data products. Key building blocks:

- Problem definition using techniques like 5 Whys to uncover root causes
- Proposed solution, starting simple
- Data mapping to track data pipelines
- Hypotheses to validate solution will solve the problem
- Actors impacted by the solution
- Strategic actions enabled by the solution
- KPIs to track performance
- Estimated value/impact of solving the problem
- Risks and mitigation plans.

A data product can also be seen from the perspective of the Data Value Chain Model. The model adapts Porter's value chain for data. The primary activities are:

- Ingestion that brings information into the system
- Transformation cleans and models data
- Communication channels display insights
- Internal sales and marketing spreads adoption
- Servicing provides user support and maintains the solution.

Support activities enable this flow, like infrastructure and R&D.

Data product portfolio management

Organizations need to manage multiple data products over time. Metrics to categorize data products in a portfolio can be scored on two dimensions. Market penetration, which is evaluated by usage rates, and growth potential, which is assessed by strategic value. This allows the organization to identify high-potential products to invest in while at the same time identifying products to phase out.

In summary, this chapter provides pragmatic frameworks to design, deliver, and manage data products that create business value. Design thinking puts the user front and center, while data-centric models focus on core data components. Together they enable data-driven decision-making.

KEY TERMS

Actors: The users and stakeholders impacted by the data product (Data Product Canvas, block 5).

Brainstorming: Technique to generate many ideas without judgment (Ideation Phase).

Communication: Medium to present insights from data product (Data Value Chain Model, Primary activities).

Data mapping: Details on data inputs, pipelines, and outputs for the product (Data Product Canvas, block 3).

Data Product Canvas: A framework to design and develop data products (Data Product Canvas).

Data products: Solutions that have data as the core component to solve problems or create opportunities (Introduction).

Data value chain: Model to describe data operations (Data Value Chain Model).

Design thinking: A human-centric approach to problem-solving that focuses on understanding user needs (Design Thinking for Data Products).

Empathy: Understanding the needs of the people you are designing for (Inspiration Phase).

Growth potential: Estimated future strategic value of a product (Data Product Portfolio Management).

Hypotheses: Assumptions about the solution that can be validated (Data Product Canvas, block 4).

Ideation phase: The second phase of design thinking focused on generating solution ideas (Design Thinking for Data Products).

Implementation phase: The third phase of design thinking focused on prototyping and testing solutions (Design Thinking for Data Products).

Infrastructure: Systems and technologies supporting data operations (Data Value Chain Model, Support activities).

Ingestion: Bringing information into the data system (Data Value Chain Model, Primary activities).

Inspiration phase: The first phase of design thinking focused on understanding the customer and problem (Design Thinking for Data Products).

Internal sales & marketing: Spreading adoption and usage internally (Data Value Chain Model, Primary activities).

KPIs: Metrics to measure performance and quality (Data Product Canvas, block 7).

Market penetration: Usage rates of a data product (Data Product Portfolio Management).

Portfolio management: Managing multiple data products across an organization over time (Data Product Portfolio Management).

Problem definition: Clearly and specifically articulating the root problem to be solved (Data Product Canvas, block 1).

Procurement: Data governance, policies, and contracts (Data Value Chain Model, Support activities).

Proposed solution: Mapping the approach to solve an identified problem (Data Product Canvas, block 2).

Prototyping: Building representations of a solution to test ideas (Implementation Phase).

R&D: Where data solutions are built and tested (Data Value Chain Model, Support activities).

Risks: Potential issues that could impact the data product's success (Data Product Canvas, block 9).

Servicing: Providing technical user support (Data Value Chain Model, Primary activities).

Strategic actions: Business decisions and changes enabled by the product (Data Product Canvas, block 6).

Transformation: Cleaning and manipulating source data (Data Value Chain Model, Primary activities).

User testing: Testing prototypes with target users to refine the solution (Implementation Phase).

Value/impact: Magnitude or size of the problem area (Data Product Canvas, block 8).

REVIEW QUESTIONS

1 What are data products?
2 What is the key focus of design thinking?
3 What is the first phase of design thinking?
4 What techniques can be used in the ideation phase?
5 What does the implementation phase involve?
6 What framework helps design data products?
7 What aspect of the Canvas focuses on the root problem?
8 Where are details on the data inputs and outputs captured?
9 What assumptions need validation?
10 Who uses the strategic actions from the data product?
11 How is data product performance measured?
12 What does the value/impact block estimate?
13 Where are risks and mitigations captured?
14 What model describes data operations?
15 What primary activity brings in source data?
16 What transformations occur to input data?
17 What enables and supports the data pipeline?
18 Why manage multiple data products over time?
19 What metrics categorize data products in a portfolio?
20 What should drive data product discovery?

Answers to review questions

1 Solutions that have data as the core component to solve problems or create opportunities (Introduction).
2 Understanding user needs through a human-centric approach (Design Thinking for Data Products).
3 The inspiration phase focused on empathy and understanding the customer (Design Thinking for Data Products, Inspiration Phase).
4 Brainstorming, worst possible idea, SCAMPER, role-playing, picture cards (Design Thinking for Data Products, Ideation Phase).
5 Building prototypes and testing them with users (Design Thinking for Data Products, Implementation Phase).
6 The Data Product Canvas (Data Product Canvas).
7 The problem definition block (Data Product Canvas, block 1).

8 The data mapping block (Data Product Canvas, block 3).
9 The hypotheses defined in the respective block (Data Product Canvas, block 4).
10 The actors outlined in block 5 (Data Product Canvas, block 5).
11 Using KPIs defined in block 7 (Data Product Canvas, block 7).
12 The magnitude or size of the problem area (Data Product Canvas, block 8).
13 In the risks block (Data Product Canvas, block 9).
14 The Data Value Chain Model (Data Value Chain Model).
15 Ingestion (Data Value Chain Model, Primary activities).
16 Cleaning, integration, manipulation (Data Value Chain Model, Primary activities).
17 The supporting activities like infrastructure and R&D (Data Value Chain Model, Support activities).
18 To maximize value creation across solutions (Data Product Portfolio Management).
19 Market penetration and growth potential (Data Product Portfolio Management).
20 Alignment to business problems that need solving (Key questions).

NOTES

[1] Remember that a decision-maker is not necessarily a leader.
[2] Specific, measurable, attainable, realizable, time-bound.
[3] Customers are not repeating their purchases.
[4] Usage is an extension to the original model by Carvalho
[5] Explained in further detail in Chapter 7 along with the analysis methods.
[6] E = Extract from source system, T = Transform the data to analysable formats, L = Loading into the system that stores it.
[7] Often referred to as the Boston matrix or BCG matrix.

BIBLIOGRAPHY

4 steps to implementing design thinking. (2023, April 16). Degreed Blog. https://blog.degreed.com/4-steps-to-implementing-design-thinking-at-your-organization/

5 examples of design thinking in business. (2023, April 16). HBS Online. https://online.hbs.edu/blog/post/design-thinking-examples

5 *effective ways of using storytelling in design*. (2023, April 16). Pepper Content: www.peppercontent.io/blog/5-effective-ways-of-using-storytelling-in-design/

8 great design thinking examples (2023, April 4). Voltage Control. https://voltagecontrol.com/blog/8-great-design-thinking-examples/

10 tips to improve your nonverbal communication. (2023, April 15). Verywell Mind. www.verywellmind.com/top-nonverbal-communication-tips2795400

Carvalho, L. (2022, August 15). *Data Product Canvas: A practical framework for building high-performance data products*. https://medium.com/@leandroscarvalho/data-product-canvas-a-practical-framework-for-building-high-performance-data-products-7a1717f79f0

Daily creation's framing your design challenge template. (2023, April 4). https://miro.com/miroverse/framing-your-design-challenge/https://doi.org/10.1111/jpim.12599

Design thinking, explained. (2023, April 16). https://mitsloan.mit.edu/ideas-made-to-matter/design-thinking-explained

Design_thinking. Wikipedia. (2023, April 14). https://en.wikipedia.org/wiki/Design_thinking

Design thinking for brands: Making a case for analogous inspiration. (2023, April 14). sureshdinakaran.com: www.sureshdinakaran.com/blog/2017/05/18/design-thinking-forbrands-making-a-case-for-analogous-inspiration/

empathize-dalam-design-thinking. (2023, April 14). https://medium.com/@lilyanastasia75/empathize-dalam-design-thinking-3c1c84fdcc21

Extremes and mainstreams: design toolkit. (2023, April 14). https://designthinking.ideo.com/resources/extremes-and-mainstreams-design-toolkit-by-ideo-org

Five whys. Wikipedia. (2023, April 15). https://en.wikipedia.org/wiki/Five_whys

Henderson, B. (1970). *The Product Portfolio*. The Boston Consulting Group, Boston, MA.

History | IDEO | Design Thinking: History. (2023, April 14). https://designthinking.ideo.com/

How to easily make a cool photo journal (Photo diary). (2023, a, April 15). https://expertphotography.com/photo-journal

How to properly frame your design challenge. (2023, April 15). https://uxdesign.cc/how-to-properly-frame-your-design-challenge-eccb4d89cb83

How to use storytelling in design thinking. (2023, April 16). www.d-thinking.com/blog/how-to-use-storytelling-in-design-thinking/

Hvad er fordele og ulemper ved prototyping? (2023, April 16). https://da.ebrdbusinesslens.com/39-info-8390108-advantages-disadvantages-prototypingl-71519

kropssprog-hvad-behoever-vide-forstaa-signaler. (2023, April 15). https://bedrelivsstil.dk/kropssprog-hvad-behoever-vide-forstaa-signaler/

Interaction Design Foundation (IxDF). (2020, June 16). *What is empathize?*. www.interaction-design.org/literature/topics/empathize

Linke, R. (2017). *Design thinking, explained. Ideas made to matter design*. MIT Sloan.

Mayrhofer, F. H. (2023, August 14). *The art and science of data product portfolio management*. AWS Big Data Blog: https://aws.amazon.com/blogs/big-data/the-art-and-science-of-data-product-portfolio-management

Majchrzak, J., Balnojan, S., & Siwiak, M., with Sieraczkiewicz, M. (2022). *Data mesh in action*. Manning.

Mehrabian, A. (1969). Some referents and measures of nonverbal behavior. *Behavior Research Methods & Instrumentation*, *1*, 203–207.

Nonverbal-communication-skills. (2023, April 15). www.indeed.com/career-advice/career-development/nonverbal-communication-skills

Osterwalder, A., & Pigneur, Y. (2010). *Business model generation*. Chichester: Wiley.

Osterwalder, A., & Pigneur, Y. (2014). *Value proposition design: How to create products and services customers want* (The Strategyzer Series). Chichester: Wiley.

Pengertian, T. D. (2023, April 15). *Design thinking: Pengertian, Tahapan, dan Contoh Penerapannya*. www.gramedia.com/literasi/design-thinking/

Poetz, M., Franke, N., & Schreier, M. (2023, April 14). *Sometimes the best ideas come from outside your industry*. https://hbr.org/2014/11/sometimes-the-best-ideas-come-from-outside-your-industry

Prototype testing: definition, benefits, how-to. (2023, April 16). https://maze.co/blog/prototype-testing/

Porter, M. E. (1985). *Competitive advantage: Creating and sustaining superior performance*. New York: Simon & Schuster.

Prototyping. (2023, April 16). The Interaction Design Foundation. www.interaction-design.org/literature/topics/prototyping

Randhawa, K, Nikolova, N., Ahuja, S., & Schweitzer, J. (2021). Design thinking implementation for innovation: An organization's journey to ambidexterity. *Journal of Product Information Management*, 38(6) 666–700.

Reeves, M. (2014, June 4). *BCG classics revisited: The growth share matrix*. www.bcg.com: www.bcg.com/publications/2014/growth-share-matrix-bcg-classics-revisited

SCAMPER – Improving products and services. (2023, April 15). www.mindtools.com/ao2rt8j/scamper

SCAMPER – Teknik Untuk Pemecahan Masalah Kreatif. (2023, April 15). www.affde.com/id/scamper.html

Simon, H. A. (1969). *The sciences of the artificial*. Cambridge, MA: MIT Press.

The 5 stages of the design thinking process [ELI5 Guide]. Springboard Blog. (2023, April 15). www.springboard.com/blog/design/design-thinking-process/

The easy guide to the Business Model Canvas. (2023, April 16). https://creately.com/guides/business-model-canvas-explained/

The history of design thinking: a short history – Clever UI. (2023, April 14). https:///the-history-of-design-thinking-a-short-history/

The history of design thinking. (2023, April 14). www.interaction-design.org/literature/article/design-thinking-get-a-quick-overview-of-the-history

What is prototyping. (2023, April 16). TWI Global: www.twi-global.com/technical-knowledge/faqs/what-is-prototyping#TypesofPrototyping

What is storytelling? (2023, April 16). The Interaction Design Foundation. www.interaction-design.org/literature/topics/storytelling

CHAPTER 4

Data Culture

How to Foster a Data-Driven Mindset (Data Literacy) and Behavior

"Culture eats strategy for breakfast." That quote was made famous by the management consultant and writer Peter Drucker. The same is true if you were to put "data" in front of both culture and strategy. That is why an entire chapter is devoted to the topic and linking it to data-driven decision-making, which is an element of both strategy and culture.

LEARNING GOALS:

L4.1 Analyze the benefits and challenges of creating a data culture in an organization using relevant theories

L4.2 Identify the key elements and characteristics of a data culture

L4.3 Compare and contrast different levels of data culture across various organizational contexts and domains

L4.4 Apply various assessment tools and frameworks to diagnose the current state of data culture in an organization and identify gaps and opportunities for improvement

L4.5 Design and implement evidence-based strategies and interventions to foster a data-driven mindset and behavior among employees and leaders, considering the organizational culture, structure, and processes

L4.6 Apply best practices and principles of data literacy to communicate, interpret, and use data effectively

L4.7 Evaluate the impact and outcomes of data culture initiatives using relevant metrics and indicators and identify areas for course correction

"Data-Driven Decision-Making is Bull★★★★ … and what to do about it" was the provoking headline of the event by LinkedIn learning teachers Gini von Courter and Bill Shander (Courter, 2022). The headline was of course created to ensure it grabbed people's attention and was to a large extent misleading. However, it brought across the point that even though we say we want to do data-driven decision-making we are inherently bad at practicing it in real life. This has also been shown in various papers where professionals working along with AI/recommender systems

DOI: 10.4324/9781003457787-4

make worse decisions than when they are just doing it without support. The AI also does better than when the human overrides. The conclusion was that we should look at our processes and culture. It unfortunately didn't go deeper because what is "data culture"?

There is no definitive answer to what constitutes a data culture in an organizational context, as different scholars have different perspectives and emphases. However, three possible definitions of data culture are widely recognized and cited:

- Data culture is "*the extent to which individuals within an organization utilize data in their decision-making*". This definition focuses on the individual level of data usage and the degree to which data informs and influences decisions (Popovic, 2018).
- Data culture is "*a set of organizational values, processes, and practices that support and encourage the use of data for decision making, learning, and improvement*". This definition emphasizes the organizational level of data adoption and the role of data in enhancing performance and innovation (Gummer, 2016).
- This definition highlights the collective level of data awareness and the importance of data literacy and data-driven action (Frank, 2016).

To conclude, we are looking at *data culture as the way an organization and its individuals gather, analyze, and use data to act*.

BENEFITS AND CHALLENGES OF DATA CULTURE

The potential benefits of data-driven decision-making are immense – better performance, innovation, competitive differentiation, and more. Yet instilling a culture that captures this potential requires surmounting common challenges around change resistance, data quality, governance, ethics, and other pitfalls.

The upside is, however, great as the company can secure a (sustainable) competitive advantage that is the basis of a company's existence.

A recent Forrester report on data culture declares "Leadership starts with a data culture" (Herrington et al., 2023). What underpins this claim?

Three areas are contributing to this conclusion.

- **Resource-based view:** Data analytics generate valuable strategic assets and capabilities that are hard to imitate, substituting sustainable differentiation for competitors in an information economy.
- **Dynamic capabilities:** The agility and customer insight enabled by data create sensing, seizing, and reconfiguring abilities to continuously reshape activities to emerging opportunities.
- **Knowledge-based view:** Data analysis unlocks intangible knowledge assets that power better decisions and services, which rivals cannot easily replicate.

A 2011 MIT study by Brynjolfsson of 179 large public firms concluded that those adopting data-driven decision-making increased productivity by 5–10% while outputting higher return

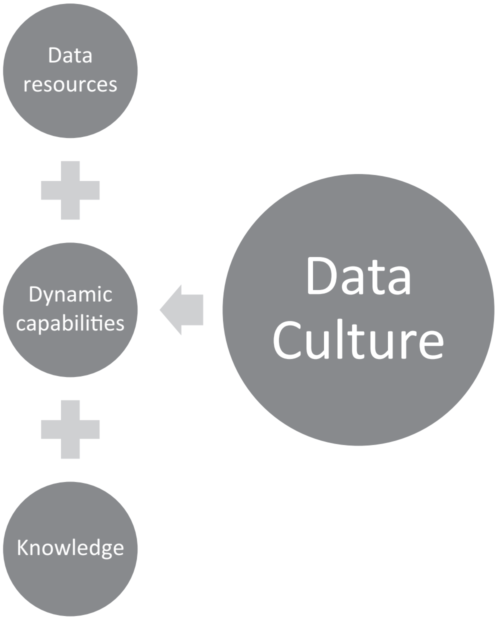


FIGURE 4.1 Data culture elements according to Forrester

on equity (ROE). Data informs strategies and optimizations that directly enhance organizational performance.

This could indicate that a data-driven culture should be the core focus of any management team.

The downside is that many will encounter transformation barriers when trying to implement a data culture.

Multiple challenges arise during implementation, as evidenced by 60% of digital transformation efforts stalling. Applying change management and culture theory helps diagnose common data culture pitfalls. A few of the classic change management models are relevant to bring in to help counter the challenges of this transformation process.

Lewin's Force Field Model: Data initiatives confront restraining forces like technology costs, job insecurity, lack of skills, and distrust of data validity. These must be alleviated while boosting driving forces around executive vision, training, communication, and participation.

Schein's Culture Levels: Analytics adoption can be hampered if alignment is lacking between operations, group norms, and underlying assumptions held by organizational members. Misalignment will trigger immune responses that data efforts must address across formal and informal culture layers.

Kotter's Change Model: Data culture founders when organizations skip needed steps like communicating urgency, training staff, and celebrating small wins.

Leaders in general and analytics leaders particularly should aim to develop a data-driven culture that institutionalizes data practices by following established change management best practices.

A data culture balancing act

Effectively fostering data-driven decisions demands handling both the upside potential and downside challenges. The following two examples highlight this balance:

*American Express (*AmEx) utilized analytics to transform itself from an outdated cheque processor to a top travel service company. This shift required balancing the benefits and challenges. Data mining spending history created high-value customer insights that powered new personalized products and value – increasing profit by 69% from 1994 to 2016.

It wasn't all easy though. Early data projects confronted unclear objectives, mismatches to business needs, and distrust of outputs. Adoption expanded by training employees to be "data translators", bridging technical findings with practical decisions.

Leave.EU and Eldon Insurance are examples of data culture tilting. An insurance company and the organization promoting Brexit don't seem to have much in common, and they shouldn't have from a data perspective. However, they were both run by Aaron Banks and were fined significantly for pushing information from the insurance company to the NGO for providing promotional messages (Hern, 2019).

Data was collected for one purpose but used for another. This case was clear as it was two different companies, but what if it was inside the same company?

As these cases illustrate, effectively transitioning towards data-driven decisions requires engaging both offensive and defensive considerations – making the strategic advantages palpable to employees while also smoothing out adoption barriers that manifest. Using frameworks like McKinsey's data culture model (McKinsey, 2021) can help maintain this balance across the multiple interdependent dimensions needed to turn potential into sustainable reality.

ELEMENTS AND CHARACTERISTICS OF DATA CULTURE

A data culture refers to the values, practices, and behaviors within an organization that support and encourage fact-based decision-making using data and analytics. Several key elements work together to create an effective data culture, and like any culture model, the right elements need to work together to ensure success.

Data vision and strategy

A clear data vision provides direction and justification for using data in decision-making. It explains why data matters and how it creates value for the organization. The vision is then translated into a data strategy with specific goals, priorities, roles, and metrics.

For example, the non-profit conservation organization Rare established a vision for "basing decisions on the best available science and data" to guide its programs protecting endangered species. This vision led to the creation of a data team responsible for impact measurement and analytics.

Data leadership

Committed leadership is vital for data culture change. Executives and managers must actively promote the adoption of data and analytics through policy, investment, and data-driven actions.

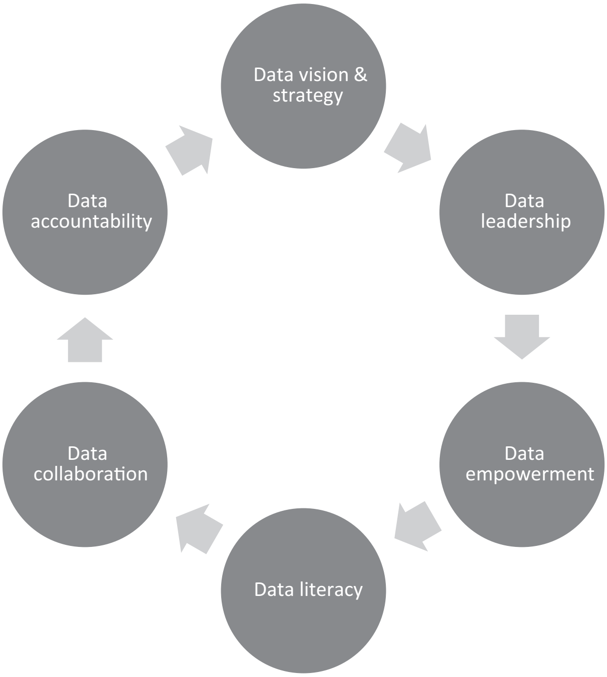


FIGURE 4.2 Interplay of data culture elements

At HP, the former CEO Carly Fiorina championed, "The goal is to turn data into information and information into insight", and made significant investments in analytics talent, tools, and training for employees. This consistent leadership has been instrumental to HP's success in those years (1999–2005).

Data empowerment

Enabling broad access and participation in using data is key. Employees at all levels should be able to retrieve, analyze, interpret, and act on data relevant to their roles. Data portals, self-service analytics tools, and data literacy programs can democratize data use.

For instance, Johnson & Johnson has over 50,000 employees using self-service dashboards via their enterprise data portal. This allows frontline staff like sales representatives to track key metrics and adapt their actions accordingly, without needing IT or analyst support.

Data literacy

Data literacy encompasses the ability to read, analyze, question, and communicate with data. Building employee data literacy through training helps ingrain data-driven thinking and removes barriers to data use. Data literacy programs can range from basic presentations and visualization skills to more advanced analytics and translation abilities, depending on the target audience and needs.

Greenpeace Brazil offers access to over 40 different data skills courses to their activists and staff, depending on their level and job function, empowering them to use data for planning and executing campaigns.

Data collaboration

Cross-functional collaboration breaks down data silos allowing different teams to combine insights. It enables data standardization across units for integrated analysis. Collaboration mechanisms like data working groups, knowledge-sharing sessions, and internal data customers can connect people working with data.

The Australia Tax Office has data collaboration forums that virtually bring together over 6,500 employees working with tax data to share knowledge, tools, and analytics. This has increased efficiency by reducing duplicated data efforts across teams and identifying new ways to leverage analytics.

Data accountability

Accountability procedures ensure high data quality standards are maintained and analytics adheres to organizational ethics. This includes instituting data governance frameworks outlining policies for security, transparency, privacy, and results validation. Establishing Chief Data Officer roles to oversee governance can further ingrain accountability.

Organizations like Walmart use a variety of data accountability methods spanning data monitoring, external audits focused on ethics and algorithmic testing for bias and overarching responsible AI principles reinforced by leadership commitments to fairness, transparency, and compliance.

All those elements need to work together to foster a culture centered around data usage. An often-used way of working with culture change is to use stories that people can associate with. The above stories could be about many companies and it's the job of the person in charge of the process to find and promote them.

MCKINSEY'S DATA CULTURE MODEL

McKinsey's data culture model, like most consulting companies, emphasizes the importance of a healthy data culture in modern organizations. The key principles that underpin a healthy data culture according to McKinsey are as follows:

Deep business engagement: Data culture should be integrated into the business, not segregated or left to specialists. It should support operations rather than the other way around.

Creating employee pull: Employees should be encouraged to actively seek data-driven solutions.

Cultivating a sense of purpose: The use of data should be purposeful and aimed at supporting the organization's goals.

Flexibility and common frameworks: While flexibility is necessary, there should also be an insistence on common frameworks and tools.

Competitive advantage: A culture that brings data talent, tools, and decision-making together can unleash competitive advantage.

These principles suggest that a data culture can't be imported or imposed but must be developed organically within the organization. It's also important to note that the journey towards a strong data culture is ongoing and doesn't have a definitive end (McKinsey, 2021). Many of the elements are linked to classic organizational development theory and inspiration can be sought there. In the following sections, the McKinsey model will be expanded to match the broader organizational context.

IMPLEMENTING A HOLISTIC DATA CULTURE

Leading companies foster data culture change by addressing all these key elements holistically. Piecemeal efforts may enable pockets of analytics adoption but fall short of transforming organizational decision-making overall.

For example, BBVA, the global bank, has systematically focused leadership priorities, talent development, easy data access, and governance practices to enable a widespread shift towards data-driven decisions across all business units globally. This multifaceted strategy resulted in analytics directly contributing over €150 million in quantified financial value by 2019. They did, however, start small and still work diligently on showing monetization from their data projects.

In summary, while specifics will differ across organizations, instilling an effective data culture fundamentally requires mobilizing around elements of vision, leadership, empowerment, literacy, collaboration, and accountability. Carefully coordinating interventions across these building blocks can ultimately realize the many benefits of becoming a truly data-driven enterprise.

LEVELS AND CONTEXTS OF DATA CULTURE

Data culture exists along a spectrum within organizations. The level and context of data culture can be characterized based on dimensions like analytical maturity, data management approach, and alignment to value creation. Different positions on these dimensions have implications for performance, innovation, and learning. Data maturity assessment has primarily been the purview of consulting companies and the models provided here are therefore primarily based on their work.

Maturity levels

One lens to assess data culture is maturity, which indicates how advanced and embedded data and analytics have become. Gartner works with a data maturity model that can be used in a data culture context as well (Gartner, 2015). In that, the maturity stages can be adapted to culture in the following ways:

- **Initial** – Little to no systematic data use beyond basic reports. Dashboards nonexistent.
- **Developing** – Basic analytics and dashboards established in silos. Data-driven decision-making is irregular.
- **Defined** – Standardized data access and tools. Some cross-functional sharing. Analytical skills are still limited.
- **Competent** – Widespread analytics adoption and skills. Collaborative data practices across units. Data-driven decisions are the norm.
- **Optimizing** – Self-service analytics and automation. Continual optimization of data impact and value through experimentation and feedback.

The maturity level affects the scale and sophistication of analysis possible and in turn, the business value achieved. For example, Apple maintains a competent level, with retail store managers empowered to access and act on customer engagement data, driving billions in sales annually. Meanwhile, nascent fintech Provizo is still developing basic data infrastructure and struggles to translate insights into growth. LaValle et al. (2011) report the results of a global survey of executives on how they use analytics to support decision-making. They identify four levels of analytical sophistication: aspirational, experienced, transformed, and analytical leaders. They also analyze how different industries and regions vary in their analytical capabilities and outcomes.

Readiness dimensions

Data culture readiness considers the foundational elements in place to progress maturity. MIT CISR (Wixom, 2023) research outlines five dimensions:

- Data - availability, quality, and infrastructure
- People – skills, leadership and participation
- Processes – data collection, sharing and usage
- Structure – roles, teams, priorities
- Values – strategic beliefs and incentives.

Organizations can assess readiness by capability and gaps within each dimension, such as identifying missing data sources, underdeveloped skills, or cultural inertia regarding data-driven decisions. Addressing limitations then enables maturation along the stages above.

For instance, UPS reached a readiness tipping point by establishing enterprise data lakes, upskilling managers through analytics training, and restructuring to have coordinated analytics teams advising all business functions. This foundation subsequently enabled sophisticated modeling and optimization benefiting routing, logistics, and customer experience.

Data culture alignment

Another view considers the alignment of data culture to business value and strategy. Four context archetypes exist:

1 Uncontrolled: Data analysis happens but is disconnected from core operations and priorities.
2 Local control: Data use and governance are siloed within individual business units.
3 Process control: Enterprise data management facilitates cross-functional efficiencies but analytical innovation is still decentralized.
4 Strategy control: Data and analytics are tightly integrated at all levels to competitive and organizational advantage.

High alignment contexts like strategy control are demanding but drive greater performance impact as data permeates decision DNA. For example, Amazon's ruthless customer-centricity is reinforced by deeply ingrained data feedback loops, testing, and algorithmic personalization that keeps user experience and innovation marching in step.

Transitioning culture levels

Elevating data culture depends on the current context. Going from local control to process control may require more data centralization, common standards, and collaboration forums before exploring advanced techniques like AI. Skipping steps can backfire. An example could be data lake initiatives floundering without people able to access and make sense of available data.

Tools like the Data Culture Framework help tailor initiatives – whether targeting skills, incentives, processes, or infrastructure – to nudge organizations positively along the dimensions above. The framework maps interventions and metrics to monitor progress specific to the stage and alignment challenges facing an enterprise.

With thoughtful stage-appropriate efforts, data culture can transition to higher levels marked by greater adoption, sophistication, and strategic integration. The resultant performance and competitive gains manifest the true potential of becoming a data-driven organization.

DIAGNOSING DATA CULTURE GAPS

Data culture exists along a spectrum. Organizations can gauge where they fall and identify areas for improvement using assessment tools evaluating dimensions like strategy, data quality, skill levels, and adoption barriers. Common diagnostic approaches include surveys, interviews, audits, benchmarks, and dashboard monitoring.

Several comprehensive guides and tools have been developed to assess the maturity and progress of these initiatives within organizations.

Eckerson (2020) provides a robust guide that evaluates the maturity of such initiatives based on 12 dimensions. These dimensions include vision and strategy, organization and culture, governance and standards, business alignment and value, analytics competency center, business requirements definition, data management, data quality, data integration, reporting, analysis, and delivery methods. Alongside this, Eckerson offers a scoring system and a roadmap for improvement, providing organizations with a clear path towards enhancing their business intelligence and analytics capabilities.

The same year, Gartner introduced a tool specifically designed to measure the progress of data and analytics initiatives within organizations. This tool has continuously been updated and now focuses on seven key dimensions: strategy and operating model, building organization, managing value creation, managing function, advanced AI maturity, creating and maintaining analytics content, integrating and managing data, and governing data and analytics assets. To aid organizations in their progress, Gartner also provides benchmarks and best practices for each dimension, offering a practical guide to improving their data and analytics initiatives (Gartner, 2024).

In a similar vein, TDWI (Halper, 2020) presents a tool aimed at evaluating an organization's readiness to adopt advanced analytics technologies and practices. This tool is based on five dimensions: organization, resources, data infrastructure, analytics, and governance to further assist organizations, TDWI provides recommendations and resources for each dimension, equipping them with the necessary tools to successfully adopt advanced analytics technologies and practices. The five stages are: nascent, early, established, mature, and advanced/visionary. It's important to note that there is a chasm between established and mature which can be challenging to cross.

These three resources offer comprehensive and practical guides for organizations to assess and improve their business intelligence, data, and analytics initiatives. They provide a multi-dimensional approach, considering various aspects of the organization and its operations, and offer valuable insights, benchmarks, and resources for enhancement.

They also explore how different organizational factors affect the progression along the stages.

Surveys

Surveys efficiently gather wide input through standardized data collection. Surveys can reach many respondents quickly, but careful design is needed so self-reported data draws an accurate picture vs just assumed strengths. Supplementing numeric ratings with interviews yields richer insights.

Example survey frameworks include:

- Data Culture Index: A research-validated instrument measuring alignment across 60+ factors like skills, leadership, infrastructure, and collaboration. Scores benchmark against peers.
- MIT Data Readiness Index: Examines data strategy, technology, skills, and connections to value. Highlights capability gaps limiting data use sophistication and business impact.

Interviews and focus groups

Interviews provide qualitative textures around barriers and needs. Focus groups add interactive dynamics eliciting insights a single interview may miss. Questions can probe topics like:

- Day-to-day data pain points
- Accessibility of data and analytics
- Enablers of/resistance to adoption
- Decision-making influences beyond data
- Perceived value and vision for data role.

Financial firm TP ICAP interviewed 150 leaders globally when seeking to upgrade analytics. The findings highlighted infrastructure limitations, data literacy needs, and trust issues that quantitative audits alone missed guiding technology investments.

Data culture audits

Audits take an inventory of existing assets and activities. They can cover:

- Data landscape – stores, models, flow monitoring
- Processes – collection, sharing, access protocols
- Skills – literacy proficiencies by role
- Governance – security, quality oversight
- Technologies – reporting, analysis, visualization.

Using a maturity framework like TDWI Analytics Capability Assessment, audits grade against best practices, illuminating capability and gap areas to address. Audits demand significant effort so are best periodically or during strategy shifts.

Benchmarks

It's one thing to have room for improvement but also vital to ascertain how far behind leaders an organization lags. Benchmarks contrast against peers:

- Data literacy rates
- Analytics adoption velocity
- Data-driven decision-making frequency
- Data talent ratios
- Data infrastructure spend.

Benchmarks contextualize gap magnitude – helping size investments appropriately. However, care should be taken when comparing across industries and processes given contextual differences.

Monitoring dashboards

Finally, dashboards track improvement initiatives and cultural shifts, displaying metrics like:

- Training participation
- Data portal usage
- Analysis of output volume
- Decision satisfaction score
- Data-informed optimizations.

Effective dashboards spotlight progress towards targets and roadmaps developed from aforementioned assessments. Data culture indicators should be monitored just as rigorously as other performance metrics.

Employing assessments holistically builds a comprehensive data culture picture – illuminating how to strategically improve data-driven decisions through tailored initiatives per gaps identified. Often a first step after creating the data vision is to assess the data culture gaps within literacy, usage, and platforms.

STRATEGIES AND INTERVENTIONS FOR DATA CULTURE

As data culture diagnostic assessments illuminate adoption gaps and needs, the next imperative is applying evidence-backed interventions tailored to the organizational context. Strategies span nurturing data skills, aligning incentives and processes, role modeling behaviors, and more.

Training programs

Equipping employees with literacy and analytics competencies removes adoption barriers. Curriculums should blend both hard technical proficiencies as well as "soft" skills for applying analysis. Targeted modules can serve specific audiences:

Leaders: Courses on framing problems data can solve, questioning assumptions, avoiding overvaluing HiPPOs (highest paid person's opinion), and monitoring analytics adoption key performance indicators through balanced scorecards or OKRs.

Managers: Sessions communicating the business case for data initiatives relevant to their function and tools for keeping teams accountable to data-informed decisions and actions.

Frontline: Basic data interpretation, visualization principles, and customer experience attribute analysis to ingrain a measurement mindset.

Data novices: Introductory analytics fluency pillars around handling, analyzing, interpreting, and discussing data.

Data intermediates: Statistical essentials, SQL basics, explanatory descriptive models, and storytelling for translating analysis into recommendations.

Data experts: Advanced modeling methods like machine learning and optimization algorithms to further impact.

Well-structured programs align competency building to genuine needs analysis while factoring in learner backgrounds. Education app Duolingo eased rollout through proper sequencing – delivering the right lessons at the right proficiency level via adaptive algorithms.

Sharma et al. (2014) propose some strategies and interventions for fostering a culture of innovation through data analytics in organizations, such as cultivating curiosity, encouraging diversity, promoting learning, rewarding failure, and facilitating communication.

Incentives and accountability

Humans naturally anchor to long-held intuitions and experiences. Counteracting this inclination towards HiPPO decision-making requires incentives nudging consideration of data.

Reward structures: Tie bonuses and advancement criteria to actual data-driven decisions vs vague aspirations. Spanish bank BBVA links executive variable pay to analytics usage and value metrics.

Data policies: Institute team data review requirements before finalizing recommendations or performance program changes, creating an accountability forcing function.

Data decision justification: Require leaders to explain quantitative and qualitative evidence backing significant judgments similar to structured evidence-based management protocols.

Applied through principles of behavioral design, incentives refocus what gets valued and where effort channels. As data initiatives mature, reinforcement gives way to intrinsic habitual usage.

Role modeling

Monkey see, monkey do – employees emulate visible behaviors modeled by influential members. Their passion and practices legitimize cultural change.

Hire/Develop data ambassadors: Ensure respected senior leaders or emerging standouts digest analytics and actively question traditional HiPPO notions in meetings.

Public support: Senior executives visibly using data insights in presentations signal its importance. Data champions can also detail its value connecting company strategy to frontline work in town halls.

Immersive experiences: Have skeptics spend time in analytics units to gain firsthand exposure to real business impacts, converting intuition to belief.

Applied thoughtfully, role modeling circumvents inertia around unfamiliar practices employees can't yet visualize delivering value.

Cultural touchpoints

Data should permeate, not silo. Subtle nudges through existing cultural channels drive that integration:

Data artwork: Adorn office walls with data-inspired murals, infographics, or sculptures communicating its creative flair. For comic relief, include Dilbert cartoons lampooning dysfunctional data culture sticking points.

Data storytelling: Spotlight case studies in company newsletters of teams who solved problems or found optimizations through data, humanizing its application. Also, highlight uses relevant to readers' roles.

Gamification: Frame data skill development as a game, having teams level up across literacy tiers and compete on leaderboards motivating progression. Duolingo's exponential user growth confirms gaming's stickiness.

Touchpoints piggyback off rooted norms perpetuating familiarity through different mediums over time as data transforms from novelty to norm.

Instilling an analytics culture demands strategic focus across numerous complementary fronts. But patient persistent efforts can ultimately unlock data's potential.

Brynjolfsson et al. (2011) identified some additional strategies and interventions for enhancing the adoption and impact of data-driven decision-making in organizations, such as creating a shared vision, empowering employees, experimenting continuously, embedding analytics into processes, and collaborating across boundaries.

PROMOTING DATA LITERACY FOR DECISION-MAKING

Data-driven decisions demand accurate interpretation and effective communication of analytics. Honing data literacy through training on underlying methods, proper visualizations, and real-world use cases ingrains critical knowledge exchange abilities.

Mastering key methods

Various analytical techniques inform robust insights. Training should advance proficiency using appropriate methods per context. Chapter 6 and Chapter 7 introduce many of the following concepts in further detail.

Descriptive statistics: Measures of central tendency, normalization, correlation analysis.
Data mining: Clustering, segmentation, regression, anomaly detection.
Predictive modeling: Forecasting via ARIMA, machine learning algorithms like random forest, neural networks.
Optimization: Linear programming, simulation, heuristics.
Experimentation: A/B testing, design of experiments.

Implementing discipline around selecting scientifically sound approaches prevents statistical pitfalls from undermining credibility. Literate professionals know their toolbox and applicability. Frank et al. (2016) describe some best practices and principles for developing and delivering data literacy training for professionals and practitioners, such as conducting a needs assessment, designing a training plan, selecting appropriate methods and materials, facilitating active learning, and evaluating the impact.

Designing compelling visuals

The best analysis holds little sway if not translated into digestible formats. Visualization principles enable data interface dexterity (Chapter 6 is focused on data visualization):

- Reduce clutter guided by the 5-second rule – simplify to key figures and flows
- Spot confusing elements like 3D effects obscuring rather than clarifying
- Use color judiciously to draw attention and encode meaning
- Align charts to natural reading gravity and layout (Gestalt principles)
- Align dashboards towards audience and decisions at hand.

Grounding in application

Standards cement theoretical understanding into workplace behaviors realizing data's latent potential.

- Document decisions requiring data justification to systematize consultation
- Log model monitoring procedures to maintain accuracy as conditions change
- Set traceability rules for modeling reproducibility across versions
- Implement version control and sandbox environments for experimental trials
- Flag biases identified through algorithm audits to address in future iterations.

Instilling foundational knowledge, communication tools, and grounded standards unlocks data literacy as a renewable competitive advantage through continual learning and refinement at scale.

MEASURING DATA CULTURE VALUE

Data-driven decision-making shows benefits on paper, but initiatives must demonstrate realized performance gains. Quantifying value requires selecting relevant KPIs reflecting strategic priorities – from efficiency to innovation to customer experience.

Lavalle et al. (2011) examine the impact and outcomes of data culture on the competitiveness and resilience of organizations in various industries, based on a survey of executives and case studies. They find that data culture enables organizations to anticipate and respond to market changes, optimize operations, enhance customer loyalty, and create new business models.

Productivity and financial returns

Data analytics, when viewed through the lens of financial metrics, provides compelling evidence of its tangible impact on an organization's bottom line. It manifests in the form of cost savings, achieved through the optimization of supply chains and staffing efficiencies, and the reduction of waste or redundancies. This optimization, driven by data, leads to a significant decrease in operational expenses.

Simultaneously, data analytics propels an upward trajectory in revenue. By enabling better targeting, pricing strategies, and conversion rates, it catalyzes a noticeable lift in sales. This revenue enhancement is a direct consequence of the strategic application of data analytics.

Furthermore, data analytics contributes to the expansion of profit margins. Whether it's through cost reduction or the ability to command premium pricing due to differentiation, data analytics plays a pivotal role in widening profit margins.

A testament to the power of data analytics is Netflix's elasticity modeling. By revealing optimized price points and content licensing costs, it ensures the maximization of subscriber growth and retention within targeted margin ranges. This strategic application of data analytics underscores its potential in driving business growth and profitability.

Beyond these traditional financial metrics, the value of data analytics can also be assessed through alternative metrics such as the return on data analytics investment. This metric provides a holistic view of the efficacy of data initiatives, evaluating whether the business gains yielded are proportional to their costs. Thus, it offers a comprehensive assessment of the payback from data analytics, further substantiating its value in the business landscape.

Ransbotham et al. (2017) explore the impact and outcomes of data culture on the transformation and differentiation of organizations in different domains, based on a survey of managers and interviews with experts. They find that data culture helps organizations to reinvent their products, services, processes, and strategies, as well as to gain a competitive edge in their markets.

Customer and market outcomes

Data potentially aids in reducing churn, a critical aspect of customer retention. By lowering customer turnover, organizations can avoid revenue loss, thereby improving their position in the competitive business landscape.

Moreover, data provides valuable insights into feature utilization. A higher usage rate not only signals a strong product-market fit but also guides the enhancement of product features, ensuring that the product continues to meet evolving customer needs.

A prime example of data-driven customer centricity is Uber's surge pricing algorithms. These algorithms balance supply and demand in real time, incentivizing an adequate number of drivers to meet ride demands. This strategic move keeps customers' waiting times low and Uber's sales high, creating a win–win situation for all parties involved.

Furthermore, advanced analytics improves market intelligence, enabling organizations to stay one step ahead. It aids in predicting emerging consumer needs and identifying potential competitive threats, thereby equipping organizations with the knowledge to navigate the dynamic market landscape effectively. In this way, data becomes an invaluable asset in the quest for customer centricity.

Transformation and innovation

In the realm of organizational transformation, data analytics emerges as a powerful catalyst, seamlessly mixing various aspects of business operations. It paves the way for the creation of new revenue streams by enabling organizations to design offerings that cater specifically to the nuanced needs of microsegments. This targeted approach not only enhances customer satisfaction but also uncovers previously untapped avenues for revenue generation.

Simultaneously, data analytics enables the reinvention of business models. It equips organizations with valuable insights, empowering them to pivot towards more lucrative value propositions and redefine their business landscapes.

The role of data analytics also extends to the acceleration of digitalization. By optimizing legacy processes and enriching customer experiences, it fast-tracks the digital transformation journey of organizations, making them more agile and responsive in an increasingly digital world.

The testament to the power of data-driven experiments lies in the success stories of industry giants like Amazon. Such initiatives now account for a significant 20% of Amazon's annual revenue, underscoring the immense potential of data analytics.

However, the true essence of data analytics lies in its ability to foster an experimental data culture within organizations. In this culture, every decision-making juncture echoes with the question, "What does the data suggest?" Leaders who adopt this approach ensure that data is at the core of strategic decisions, thereby driving continuous innovation and maintaining the momentum of transformation. This culture, steeped in data, is the cornerstone of sustained organizational transformation in the modern business era.

Bughin et al. (2018) analyze the impact and outcomes of data culture on the performance and innovation of organizations across different sectors and regions, based on a survey of executives and a database of indicators. They find that data culture is associated with higher revenue growth, profitability, productivity, customer satisfaction, and innovation.

Measurement implementation

KPI tracking should serve insight, not bureaucracy. Reports synthesize indicators, benchmarks, and qualitative observations painting a nuanced picture of progress toward strategic goals. Surfacing lessons then refine efforts for maximal lift. With data permeating decisions enterprise-wide, ultimately cultural change is its perpetual indicator.

CONCLUSION

This chapter explored multiple facets of instilling an effective organizational data culture that fully harnesses analytics' potential for enhanced decisions and performance.

We defined data culture as the degree to which data and analytics permeate values, skills, behaviors, processes, and leadership priorities throughout an enterprise. Six foundational data culture pillars were outlined encompassing vision, empowerment, literacy, collaboration, accountability, and leadership buy-in. Embracing these characteristics engenders competitive advantage through accelerated innovation and responsiveness from data-driven insights.

However, realizing this potential requires overcoming common barriers around skills gaps, data silos, poor data quality, change resistance, and ethical blind spots. Assessment tools like maturity models, capability audits, surveys, and monitoring dashboards can diagnose an organization's starting point challenges and opportunities. Tailored interventions spanning training, role modeling, governance policies, and behavioral incentives can then strategically address diagnosed gaps.

For example, needs-aligned analytics and interpretation education enhance individual data literacy unlocking wider participation. Meanwhile, incentives making data review mandatory for decisions instill rigor, countering bias towards gut feelings. Progress indicators tracking usage, proficiencies attained, and decision quality guide refinement.

With patient cultivation, data practices become normalized ingraining sustainable differentiation. Effectively designed dashboards spotlight leading return metrics reflecting data culture advances translating to efficiency gains, revenue growth, customer retention, and transformation velocity. Ultimately a thriving data culture manifests in an insatiable organizational appetite to interrogate assumptions with data and embrace insights through perpetual learning and experimentation.

In summary, data culture requires coordinated enablement across infrastructure, processes, skills, and leadership commitments. However, organizations investing to advance along data-driven transformation dimensions stand to amplify performance on all fronts by leveraging analytics as an essential enabler embedded into their decision DNA.

KEY TERMS

A/B testing: Randomized experiment methodology comparing outcomes between a current standard and an alternative to determine what performs best.

Analytical maturity: Metric gauging an organization's analytics progression across dimensions like access, skills, usage, impact, and governance.

Analytics-based KPIs: Performance indicators tracking data culture progress through usage, proficiencies gained, and decision quality metrics.

Business intelligence: Using data analytics and reporting to generate insights informing business decisions and strategy.

Chief Data Officer (CDO): Executive role instituted to provide leadership overseeing data strategy and governance implementation.

Data accountability: Policies and roles ensuring high data standards via governance oversight.

Data ambassadors: Influential analytics-proficient employees demonstrating and speaking to data use cases winning over skeptics by example.

Data collaboration: Breaking silos by fostering data sharing and coordinated analytics efforts across teams and units.

Data culture index: Measurement framework benchmarking data practices alignment to leadership priorities, empowerment, skills, behaviors, and attitudes.

Data culture: The values, practices, behaviors, and leadership support enabling data and analytics usage in decisions and processes organization-wide.

Data dashboard: Consolidated graphical data display highlighting key performance metrics and insights.

Data democratization: Enabling broad decentralized participation in data analysis activities through access, tools, and skills.

Data empowerment: Enabling broad employee participation in data analysis through access, tools, and skills.

Data governance: Policies and accountability models ensuring data practices adhere to quality, security, and ethics standards.

Data literacy: The ability to access, interpret, analyze, communicate, and act using data effectively.

Data storytelling: Communicating data narratives and visualizations in compelling ways clarifying meaning for audiences and influencing decisions.

Data vision: Strategic outlook articulating why and how data creates value, guiding data efforts.

Data-driven decision-making (DDD): Basing choices and strategies substantially on data and analytics versus instinct or observation alone.

DataOps: Collaborative agile project management approach coordinating data teams towards common objectives.

Descriptive analytics: Quantitative methods summarizing trends, patterns, and relationships in historical data to explain what happened.

HiPPO effect: Proclivity towards relying more on highest highest-paid person's opinion versus objective data analysis when making decisions.

Level of analytics: Segmenting analysis types into four tiers based on complexity spanning descriptive, diagnostic, predictive, and prescriptive methods.

Predictive analytics: Statistical and machine learning techniques analyzing current and historical data to forecast what could happen in the future.

Prescriptive analytics: Algorithms suggesting actions, optimizations, and decisions focused on how to leverage insights most advantageously towards goals.

Self-service analytics: Facilitating employees analyzing data independently without intermediaries via interfaces like BI tools or portals.

Small data: Complementing big data analytics with contextual insights from limited niche datasets requiring qualitative interpretation.

REVIEW QUESTIONS

1. What does a clear data vision provide for an organization seeking to build its data culture?
2. What MIT research dimension concerns the availability, quality, and flow of data assets?
3. True or false – data literacy only involves technical skills like data analysis and coding.
4. What change management theory suggests data culture efforts consider formal policies, ad hoc behaviors, and underlying assumptions?
5. What data literacy training methodology grounds concepts into workplace behaviors and standards?
6. Name 3 measures that help quantify the financial value created from improved data culture.
7. Which visualization principle aims to simplify charts to key figures based on cognitive load?
8. True or false – Surveys are the only valid way to assess organizational data culture maturity.
9. What McKinsey model balances data culture offensive and defensive considerations?
10. Name 3 metrics that help track customer outcomes conferred by enhanced data use.
11. True or false – Effective data culture transformations require broad isolated interventions across single dimensions.
12. What accountability policy requires quantitative evidence alongside decision records?
13. Name 3 types of analytics education curriculums that enhance data literacy.
14. What General Electric failure example highlights the perils of focusing overly on technical data tools?

15 What BBVA example showcases how financial incentives can help nudge the adoption of data practices?
16 True or false – Effective data culture measurement should serve insight not bureaucracy through streamlined reports.
17 What mechanism helps leaders digest analytics trends and influence colleagues by example?
18 Name 3 cultural nudges that can permeate data practices through existing norms.
19 What theory suggests data analysis confers hard-to-imitate competitive advantages?
20 What hospitality chain example showcases self-service analytics success?
21 Name 3 analysis methods useful for predictive modeling.
22 True or false – Aligning maturity-level interventions with current challenges increases the likelihood of positive transitions.
23 What conservation organization example showcases a linkage from data vision to roles?
24 Name 2 change management pitfalls data culture initiatives should sidestep.
25 Who famously flagged change resistance challenges by stating "Culture eats strategy for breakfast"?
26 True or false – Data culture ultimately permeates as an intrinsic normalized asset needing little active management.
27 What taxi firm showcases real-time data to improve customer experience?
28 Name 3 ethical considerations data efforts should address.
29 What bank example showcases holistic data culture advancement increasing quantified value eight-fold?
30 True or false – Data culture requires no new skills just updated technology infrastructure.

Answers to review questions

1 Direction and justification for using data in decision-making
2 Data
3 False (also involves communication and interpretation)
4 Schein's culture model
5 Application to real-world problems
6 Cost reductions, revenue increases, margin improvements
7 5 second rule
8 False (Interviews and audits also offer value)
9 Data culture balance model
10 Net promoter score, churn rates, feature utilization
11 False (Coordinated efforts across multiple dimensions work best)
12 Data justification documentation protocols
13 Leader, manager, frontline
14 Lacking change management and skills development resulted in distrust and non-usage
15 Executive variable pay tied to the usage of analytics
16 True
17 Data ambassadors
18 Data artwork, storytelling, gamification
19 Resource-based view

20 Johnson & Johnson
21 ARIMA, Machine learning, neural networks
22 True
23 Rare data team
24 A. Skipping needed steps like training or not celebrating small wins
25 Peter Drucker
26 True
27 Uber
28 Privacy, bias, transparency
29 BBVA generating over €150 million
30 False – Significant skill building in literacy and analytics is imperative

BIBLIOGRAPHY

Brynjolfsson, E., Hitt, L. M., Kim, H. H. (2011). Strength in numbers: How does data-driven decisionmaking affect firm performance? *S&P Global: Market* Intelligence. http://dx.doi.org/10.2139/ssrn.1819486.

Bughin, J., Seong, J., Manyika, J., Chui, M., & Joshi, R. (2018). Notes from the AI frontier: Modeling the impact of AI on the world economy. *McKinsey Global Institute*, *4*. www.McKinsey.com

Courter, G. V. (2022, November 22). *Data-driven decision-making is bull**** … and what to do about it*. Linkedin. www.linkedin.com/events/6998031457935900672

Eckerson, W. (2020). Creating an analytics culture part I+II: 12 characteristics. Eckerson Group. www.eckerson.com/articles/creating-an-analytics-culture-part-i-12-characteristics (Accessed April 14, 2024)

Frank, M. (2016). Developing data literacy: Theoretical frameworks to inform education and training initiatives. *International Journal of Data Science and Analytics*, 2(4), 231–244.

Gartner. (2015). Enterprise Information Management Maturity Model. www.gartner.com/en/documents/3136418 (Accessed January 16, 2024).

Gartner. (2024). *Gartner data and analytics maturity score for CDAOs*. Retrieved April 14, 2024, from www.gartner.com/en/data-analytics/research/data-analytics-maturity-score (Accessed April 14, 2024).

Gummer, M. A. (2016). *Data literacy for educators: Making it count in teacher preparation and practice*. New York: Teachers College Press.

Halper, F. (2020). TDWI analytics maturity model. *TDWI Research*, *22*. TDWI.org

Hern, A. (2019, February 1). Leave.EU and Arron Banks insurance firm fined £120,000 for data breaches. *The Guardian*. www.theguardian.com/uk-news/2019/feb/01/leave-eu-arron-banks-insurance-company-fined-data-breaches-information-commissioner-audit

Herrington, K., Sridharan, S., Sabri, F., & Barton, J. (2023). Maximize curiosity velocity to improve your data culture: Help employees seek, solicit, and speak in a safe to share environment. *Best Practice Report*. Forrester. www.forrester.com/report/maximize-curiosity-velocity-to-improve-your-data-culture/RES178997

Lavalle, S., Lesser, E., Shockley, R., Hopkins, M., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review*, *52*, 21–32.

McKinsey (2021). The importance of a healthy data culture in modern organizations. www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/Why%20data%20culture%20matters/Why-data-culture-matters.ashx (Accessed January 10, 2024).

Popovic. (2018). Creating a data culture: A dynamic capability perspective. *Journal of Strategic Information Systems*, *27*(4), 296–310.

Ransbotham, S., Kiron, D., Gerbert, P., & Reeves, M. (2017). Reshaping business with artificial intelligence: Closing the gap between ambition and action. *MIT Sloan Management Review*, *59*(1), 1–17.

Sharma, R., Mithas, S., & Kankanhalli, A. (2014). Transforming decision-making processes: A research agenda for understanding the impact of business analytics on organisations. *European Journal of Information Systems*, *23*. 433–441. https://doi.org/10.1057/ejis.2014.17

Wixom, B. H., Beath, C. M., & Owens, L. (2023*). Data is everybody's business: The fundamentals of data monetization*. Boston, MA: MIT University Press.

CHAPTER 5

Data Sources

How to Find, Collect, and Manage Data for Business Value

This chapter expands independently on the data collection topic introduced in Chapter 2: How to identify relevant internal and external sources of data. It also covers the emergence of data contracts and data contract management.

- Identifying relevant internal and external data sources
- Collecting quality, reliable, and timely data
- Managing your data assets using metadata, standards, contracts and policies

Chapter case: Happy Pops, Boutique Ice Cream, Canada

LEARNING GOALS:

L5.1 To evaluate the strengths and limitations of different data sources and methods for addressing specific business questions and challenges

L5.2 To identify the potential risks and challenges of data collection and management, such as data privacy, security, ownership, quality, and compliance issues

L5.3 To apply the concepts and techniques of data collection and management to real-world business cases and scenarios

L5.4 To explore emerging trends and opportunities in data collection and management, such as big data, open data, cloud computing, and artificial intelligence

L5.5 To appreciate the importance of data contracts and data contract management for establishing trust, accountability, and collaboration among data providers and consumers

L5.6 To develop skills in data management, such as creating and using metadata, standards, contracts, and policies to organize, document, and protect your data assets

L5.7 To understand the different types of data sources and how to select the most relevant ones for your business problem

L5.8 To learn how to collect data in a systematic, ethical, and legal way that ensures quality, reliability, and timeliness

DOI: 10.4324/9781003457787-5

DATA SOURCES

Data comes in many shapes, forms, and formats. In data analysis, we are often looking at either structured or unstructured data as that is the step just before analysis or transformation.[1]

Internal data

Definition and examples of internal data

Internal data is data that the organization controls. It might be proprietary, bought, or collected (discussed further in the data collection section of this chapter).[2]

Data can be gathered from "sales/marketing", "finance", "production", and "human resource".

"Sales/marketing" data and systems

Sales and marketing data is data that relates to the activities and performance of the sales and marketing functions of a business. Sales and marketing data can help the business understand its customers, competitors, and market trends, as well as measure and improve its sales and marketing strategies and outcomes.

Some examples of sales and marketing data are:

- Customer data: This includes information about the customers' demographics, preferences, behaviors, needs, satisfaction, loyalty, and feedback. Customer data can help the business segment and target its customers, personalize its offerings, and improve its customer service and retention.
- Sales data: This includes information about the sales pipeline, sales cycle, sales volume, sales revenue, sales costs, sales quotas, sales commissions, and sales forecasts. Sales data can help the business monitor and manage its sales performance, identify and prioritize sales opportunities, and optimize its sales processes and resources.
- Marketing data: This includes information about the marketing campaigns, marketing channels, marketing costs, marketing ROI, marketing metrics, and marketing analytics. Marketing data can help the business plan and execute its marketing activities, evaluate and improve its marketing effectiveness, and optimize its marketing budget and allocation.

Many systems can capture sales and marketing data for later analysis. Some of the common systems are:

- CRM systems: CRM stands for customer relationship management. CRM systems are software applications that help the business manage its interactions and relationships with its customers and prospects. CRM systems can capture and store customer data, as well as track and manage sales activities and processes. Some examples of CRM systems are Salesforce1, HubSpot2, Zoho, etc.

- Marketing automation systems: Marketing automation systems are software applications that help the business automate and streamline its marketing tasks and workflows. Marketing automation systems can capture and store marketing data, as well as execute and measure marketing campaigns across multiple channels. Some examples of marketing automation systems are Mailchimp, Marketo, HubSpot2, etc.
- Web analytics systems: Web analytics systems are software applications that help the business collect and analyze data about its website visitors and their behavior. Web analytics systems can capture and store web data, such as page views, bounce rate, conversions, etc., as well as provide insights into website performance, user experience, and optimization. Some examples of web analytics systems are Google Analytics, Adobe Analytics, Semrush, etc.

"Finance" data and systems

Finance data is data that relates to the financial activities and performance of a business, such as revenues, expenses, assets, liabilities, cash flows, budgets, forecasts, etc. Finance data can help the business measure and improve its financial health, efficiency, and profitability, as well as comply with accounting and tax regulations.

Various systems capture finance data that later can be used for analysis and provide data for decision-making. Some of the common systems are:

- ERP systems: ERP stands for enterprise resource planning. ERP systems are software applications that integrate and automate various business processes and functions, such as accounting, finance, procurement, inventory, sales, etc. ERP systems can capture and store finance data from different sources and provide a consolidated view of the financial performance and position of the business. Some examples of ERP systems are SAP1, Oracle2, Microsoft Dynamics3, etc.
- Business intelligence systems: Business intelligence systems are software applications that help the business collect, analyze, and visualize data to support decision-making and reporting. Business intelligence systems can capture and store finance data from various sources and provide insights into the financial trends, patterns, and anomalies of the business. Some examples of business intelligence systems are Tableau, Power BI, Qlik, etc.
- Financial planning and analysis systems: Financial planning and analysis systems are software applications that help the business plan, budget, forecast, and monitor its financial performance and goals. Financial planning and analysis systems can capture and store finance data from various sources and provide tools for scenario analysis, variance analysis, profitability analysis, etc. Some examples of financial planning and analysis systems are Adaptive Insights, Anaplan, Cognos, etc.

"Production" data and systems

Production data is data that relates to the manufacturing or creation of goods and services by a business, such as production volume, production costs, production quality, production efficiency, production capacity, etc. Production data can help the business measure and improve its operational performance, productivity, and profitability, as well as optimize its production processes and resources.

Production data that can be analyzed are many different systems. Some of the common examples are:

- MES systems: MES stands for manufacturing execution system. MES systems are software applications that monitor and control the production activities and processes on the shop floor. MES systems can capture and store production data from various sources, such as machines, sensors, operators, etc., and provide real-time feedback and guidance to improve production efficiency and quality. Some examples of MES systems are Siemens, Rockwell Automation, GE Digital, etc.
- SCADA systems: SCADA stands for supervisory control and data acquisition. SCADA systems are software applications that collect and analyze data from remote or distributed devices and systems, such as pumps, valves, motors, generators, etc., that are involved in the production process. SCADA systems can capture and store production data from various sources and provide visualization and control capabilities to optimize production performance and reliability. Some examples of SCADA systems are Wonderware, Inductive Automation, ABB, etc.
- ERP systems: ERP systems are also in this category of production systems and supplied by the same vendors. The extent of their relevance here depends on their configuration which varies significantly from company to company.

"Human resource" data and systems

Human resource data is data that relates to the people and activities involved in the human resource management function of a business, such as recruitment, retention, performance, compensation, training, development, etc. Human resource data can help the business measure and improve its human capital, employee engagement, and organizational culture, as well as comply with labor and employment laws and regulations.

Human resource data is often captured in specialized systems due to their confidential nature. Examples are:

- HRIS systems: HRIS stands for human resource information system. HRIS systems are software applications that store and manage employee data, such as personal information, job history, payroll, benefits, attendance, etc. HRIS systems can capture and store human resource data from various sources, such as employee records, timesheets, tax forms, etc., and provide basic reporting and analytics capabilities to support human resource administration and compliance. Some examples of HRIS systems are BambooHR, Workday, ADP, etc.
- HCM systems: HCM stands for human capital management. HCM systems are software applications that optimize and align human resource processes and practices with the strategic goals and objectives of the business. HCM systems can capture and store human resource data from various sources, as well as provide advanced features and functionalities for human resource planning, talent management, performance management, learning management, etc. Some examples of HCM systems are Oracle HCM Cloud, SAP SuccessFactors, Cornerstone OnDemand, etc.

- Employee engagement systems: Employee engagement systems are software applications that measure and improve the level of employee satisfaction, motivation, and commitment to the business. Employee engagement systems can capture and store human resource data from various sources, such as surveys, feedback, recognition, rewards, etc., and provide insights and recommendations to enhance employee engagement and retention. Some examples of employee engagement systems are Culture Amp, Glint, Peakon, etc.

Advantages and disadvantages of internal data

Internal data have several advantages when used in analysis. Some of the advantages include:

- Increased reliability: Internal data is private information that the company in question oversees directly. The company is responsible for collecting, storing, and maintaining internal data, which means it's more accurate and credible in comparison to information from third parties.
- Better security: Outside parties require express permission from a company before they can access or study internal data. Such high security ensures data remains free from manipulation and is only accessible to those the company authorizes.
- Easily accessible: Internal data allows for almost instant access to information, making it possible for businesses to make quick decisions.

However, there are also some disadvantages of using internal data in analysis. Some of the disadvantages include:

- Limited scope: Internal data only provides information on the company's transactions and current practices by pulling facts and statistics from internal databases. This means that it may not provide a complete picture of the market or industry trends.
- Biased perspective: Internal data is specific to the operations of the company in question, which means that it may not reflect the broader market or industry trends. This can lead to a biased perspective that may not be representative of the larger context.
- Costly and time-consuming: Collecting, storing, and maintaining internal data can be costly and time-consuming, especially if the company has a large amount of data or if the data is spread across multiple systems.

Overall, internal data can be a valuable source of information for businesses when used in analysis. However, it's important to consider both the advantages and disadvantages of using internal data and to balance it with external data sources to get a more complete and accurate picture of the market and industry trends.[2]

Criteria and methods for selecting and accessing internal data

Accessing internal data for analysis depends on the type of data and the systems used to store and manage it. Some common methods for accessing internal data include:

- Querying databases: Many companies store their internal data in databases, such as SQL Server, Oracle, MySQL, etc. To access this data, you need to know the structure and schema of the database, as well as the SQL or other query language used to extract the data. You can use tools like SQL Server Management Studio, Oracle SQL Developer, MySQL Workbench, etc., to query the databases and retrieve the data you need.
- Exporting reports: Many companies generate reports from their internal systems, such as ERP, CRM, HRIS, etc., that contain useful data for analysis. To access this data, you can export the reports in various formats, such as Excel, CSV, PDF, etc., and then import them into your analysis tools or platforms. You can use tools like Microsoft Excel, Google Sheets, Tableau, Power BI, etc., to import and analyze the reports.
- Using APIs: Many companies expose their internal data through APIs (application programming interfaces) that allow external applications to access and retrieve the data in a structured and secure way. To access this data, you need to know the API endpoints and parameters used to query the data. You can use programming languages like Python, Java, C#, etc., to write scripts or applications that interact with the APIs and retrieve the data you need.

It's important to note that accessing internal data for analysis requires proper authorization and security measures to ensure that only authorized users can access and use the data. It's also important to ensure that the data is accurate, complete, and relevant to your analysis needs (Davenport, 2021).

External data

Definition and examples of external data

External data sources are data that originate from outside of an organization and can be used to supplement or complement internal data sources. Examples of external data sources include job portals, career websites, business-focused social networks, patents, university data, learning offerings, online courses, public statistics, government sources, social media, satellite, consumer transactions, geo-location, and employment data. External data can provide valuable insights into market trends, customer behavior, competitive intelligence, industry benchmarks, and other external factors that affect business performance and strategy (Kilduff, 2023) However, external data also has some limitations and challenges, such as data quality issues, data privacy concerns, data integration complexity, and data interpretation ambiguity (Talend, 2023). Therefore, it's important to evaluate the relevance, reliability, and validity of external data sources before using them in analysis or decision-making.

As an example, external data sources such as job portals, career websites, and business-focused social networks are linked to internal human resource functions and their data. These sources provide a wealth of information for businesses to find new talent, research competitors, and stay up-to-date on industry trends. For example, job portals like Indeed and Monster allow businesses to post job openings and search résumés to find the right candidates for their open positions. Career websites like Glassdoor provide company reviews, salary information, and job listings to help job seekers find the right fit. Business-focused social networks like LinkedIn offer a platform for professionals to connect, share industry news, and build their personal brand. All of this data can be used by internal human resource teams to make informed decisions about hiring, compensation, and employee engagement.

External data sources can be used by internal human resource teams to make informed decisions about hiring, compensation, and employee engagement in various ways. For example:

- Hiring: External data sources can help human resource teams find and attract qualified candidates for their open positions, as well as assess their skills, experience, and fit. Job portals, career websites, and business-focused social networks can provide a large pool of potential applicants, as well as insights into their profiles, preferences, and feedback. Patents, university data, learning offerings, and online courses can provide information on the latest innovations, research, and education in the relevant fields. Public statistics and government sources can provide data on the labor market, employment trends, and regulations.
- Compensation: External data sources can help human resource teams design and implement fair and competitive compensation packages for their employees, as well as benchmark their performance and rewards. Job portals, career websites, and business-focused social networks can provide data on the salary ranges, benefits, and incentives offered by competitors and industry peers. Consumer transactions and geo-location data can provide data on the cost of living, inflation, and purchasing power in different regions. Public statistics and government sources can provide data on the tax rates, minimum wage, and labor laws.
- Employee engagement: External data sources can help human resource teams measure and improve the level of employee satisfaction, motivation, and commitment to the organization, as well as identify and address any issues or concerns. Social media, satellite, consumer transactions, and geo-location data can provide data on the sentiment, behavior, and feedback of employees and customers. Public statistics and government sources can provide data on the social, economic, and environmental factors that affect employee well-being and happiness.

These are just some of the ways that external data sources can be used by internal human resource teams to make informed decisions about hiring, compensation, and employee engagement. However, it's important to note that external data sources should be used with caution and care, as they may have limitations and challenges in terms of data quality, data privacy, data integration, and data interpretation. Therefore, it's advisable to use external data sources in combination with internal data sources to get a more complete and accurate picture of the human resource situation.

Some advantages of external data are:

- Increased scope: External data can provide a broader and more diverse range of information than internal data, as it comes from various sources and contexts outside the organization. This can help businesses gain new insights into market trends, customer behavior, competitive intelligence, industry benchmarks, and other external factors that affect their performance and strategy.

- Reduced costs: External data can be less expensive than internal data, as it's often available for free or at a lower cost than internal data. This can help businesses save money on data acquisition and management, as well as reduce the risks of data bias and manipulation (Brown, 2021).
- Improved accuracy: External data can provide more accurate and reliable information than internal data, as it's often collected and verified by independent third parties with specialized expertise and resources. This can help businesses avoid errors, biases, and conflicts of interest that may affect their internal data (Marr, 2022).

Some disadvantages of external data are:

- Limited relevance: External data may not be directly relevant or applicable to the business needs or goals of an organization, as it comes from various sources and contexts outside the organization. This can make it difficult to interpret, analyze, and use external data effectively in decision-making.
- Limited quality: External data may not be of high quality or reliability, as it's often collected and managed by independent third parties with different standards and practices. This can make it difficult to ensure the accuracy, completeness, and consistency of external data (Brown, 2021).
- Limited accessibility: External data may not be easily accessible or available to all businesses, as it's often subject to legal, ethical, or technical restrictions that limit its use or distribution. This can make it difficult to obtain or use external data for some businesses (Marr, 2022).

These are just some examples of the advantages and disadvantages of external data. The actual advantages and disadvantages may vary depending on the type, source, quality, relevance, accessibility, and context of the external data used.

There are several ways to access external data, depending on the type and source of the data. Some common methods include:

- Web scraping: Web scraping involves extracting data from websites and web pages using automated tools or scripts. Web scraping can be used to collect data on competitors, customers, products, prices, reviews, etc. Some examples of web scraping tools are BeautifulSoup, Scrapy, Selenium, etc.
- APIs: APIs (application programming interfaces) are software interfaces that allow different applications to communicate and exchange data with each other in a structured and secure way. APIs can be used to collect data from various sources, such as social media, weather, maps, news, etc. Some examples of APIs are Twitter API, OpenWeatherMap API, Google Maps API, NewsAPI, etc.
- Data brokers: Data brokers are companies that collect and sell data from various sources to other businesses or organizations. Data brokers can provide access to large and diverse datasets on consumers, businesses, industries, etc. Some examples of data brokers are Acxiom, Experian, Equifax, Dun & Bradstreet, etc.
- Public databases: Public databases are online repositories of data that are maintained by governments or other public organizations. Public databases can provide access to various

types of data on demographics, health, education, crime, environment, etc. Some examples of public databases are Data.gov, Eurostat, World Bank Open Data, etc.

The actual methods may vary depending on the type and source of the data and the tools or platforms used for analysis.

When accessing external data, there are several factors to consider to ensure that the data is relevant, reliable, and valid for your analysis needs. Some of the factors to consider are:

Data quality: External data may have quality issues, such as errors, biases, inconsistencies, or incompleteness. Therefore, it's important to evaluate the quality of the external data before using it in analysis. You can use various methods to assess the quality of external data, such as data profiling, data cleansing, data enrichment, or data validation.

Data relevance: External data may not be directly relevant or applicable to your business needs or goals. Therefore, it's important to identify the most relevant and useful external data sources for your analysis needs. You can use various methods to identify relevant external data sources, such as market research, competitor analysis, customer feedback, or industry reports.

Data privacy: External data may contain sensitive or confidential information that requires proper authorization and security measures to access and use. Therefore, it's important to comply with legal and ethical standards for data privacy when accessing and using external data. You can use various methods to ensure data privacy, such as data masking, encryption, anonymization, or access control.

Data integration: External data may come in different formats, structures, or systems that require proper integration and alignment with your internal data sources. Therefore, it's important to ensure that the external data is compatible with your internal systems and processes. You can use various methods to integrate external data with internal data sources, such as ETL (extract, transform, load), API (application programming interface), or various types of middleware.

Data interpretation: External data may have different meanings or interpretations depending on the context and purpose of your analysis. Therefore, it's important to interpret the external data correctly and accurately for your analysis needs. You can use various methods to interpret external data correctly, such as statistical analysis, machine learning algorithms, or expert judgment.

These are just some of the factors to consider when accessing external data for analysis. The actual factors may vary depending on the type and source of the external data and the specific needs and goals of your analysis.

Data collection

In the previous section, internal and external data was described and the ways to access it were briefly set out. In this section, we will focus on gathering or generating primary data. There are two data types that we can gather for our decision-making. Qualitative and quantitative data. The data type you need defines the method that you use. In some cases, both data types are relevant to gather. That would often start with qualitative data that will provide a hypothesis and then quantitative that confirms or disapproves the hypothesis.

Qualitative data methods

Interviews, focus groups, observations, and case studies are examples of qualitative data collection methods. Qualitative methods are used to collect non-numerical data that cannot be easily quantified. Qualitative data is descriptive and subjective, and it is analyzed by grouping it in terms of meaningful categories or themes. Qualitative methods are useful for exploring complex topics and gaining insights into the experiences and opinions of the participants.[3,4]

Quantitative data methods

Surveys, experiments, and observations with numerical data are examples of quantitative data collection methods. Quantitative methods are used to collect numerical data that can be analyzed using statistical analysis. Quantitative data is objective and measurable, and it is analyzed using mathematical calculations. Quantitative methods are useful for testing hypotheses and establishing causal relationships between variables.[5,6]

Mixed methods

Mixed methods research is a research approach that combines both quantitative and qualitative data collection and analysis methods (George, 2022). Some of the mixed methods for data collection are:

Convergent parallel design: In this design, both quantitative and qualitative data are collected at the same time and analyzed separately. The results are then compared to identify similarities and differences between the two types of data.

Embedded design: In this design, both quantitative and qualitative data are collected and analyzed within a larger quantitative or qualitative design. The qualitative data is used to provide a more in-depth understanding of the quantitative data.

Explanatory sequential design: In this design, quantitative data is collected first, followed by qualitative data. The qualitative data is used to explain or expand on the quantitative findings.

Exploratory sequential design: In this design, qualitative data is collected first, followed by quantitative data. The quantitative data is used to test or confirm the qualitative findings.

There are many methods for collecting data, depending on the type, purpose, and scope of the research. Some of the most widely recognized methods are:

Surveys: Surveys are questionnaires that gather both qualitative and quantitative data from subjects. Surveys can be conducted online, by phone, by mail, or in person. Surveys are useful for collecting large amounts of data from a diverse population (Harvard Business School Online, 2021).

Interviews: Interviews are face-to-face or remote conversations that elicit in-depth information from a small number of participants. Interviews can be structured, semi-structured, or unstructured, depending on the level of flexibility and standardization. Interviews are useful for exploring complex topics and gaining insights into the experiences and opinions of the participants (Bhandari, 2023).

Focus groups: Focus groups are group discussions that involve a moderator and a selected group of participants who share their views on a given topic. Focus groups can generate rich data through the interaction and dynamics of the group. Focus groups are useful for exploring the attitudes, preferences, and motivations of the participants.

Observations: Observations are systematic recordings of the behavior or phenomena of interest in their natural setting. Observations can be participant or non-participant, depending on the role and involvement of the researcher. Observations are useful for studying processes, contexts, and interactions that are not easily captured by other methods.

Experiments: Experiments are controlled manipulations of one or more variables to measure their effect on another variable. Experiments can be conducted in a laboratory or the field, depending on the realism and validity of the setting. Experiments are useful for testing hypotheses and establishing causal relationships between variables.

Data collection process

The general steps that must always be done when embarking on data collection are:

Identify the research question: The first step in data collection is to identify the research question that the data will help answer. The research question should be specific, measurable, and relevant to the research problem.

Determine the data needed: The second step is to determine what data is needed to answer the research question. The data should be relevant, reliable, and valid for the research purpose.

Choose the data collection method: The third step is to choose the most appropriate data collection method based on the research question and the type of data needed. The choice of method should consider factors such as feasibility, cost, time, and ethical considerations.

Design the data collection instrument: The fourth step is to design the instrument that will be used to collect the data. The instrument should be valid, reliable, and appropriate for the research question and the chosen method.

Pilot test the instrument: The fifth step is to pilot test the instrument with a small sample of participants to identify any problems or issues with the instrument and refine it accordingly.

Collect the data: The sixth step is to collect the data using the chosen method and instrument. The data collection process should be standardized, systematic, and ethical.

Clean and organize the data: The seventh step is to clean and organize the collected data to ensure its accuracy, completeness, and consistency. Data cleaning involves identifying and correcting errors, inconsistencies, or missing values in the dataset.

Analyze the data: The eighth step is to analyze the cleaned and organized data using appropriate statistical or qualitative methods depending on the research question and type of data.

Interpret and report the findings: The final step is to interpret and report the findings of the analysis in a clear, concise, and meaningful way that answers the research question and contributes to knowledge in the field.

These steps are not necessarily linear or sequential but rather iterative and interactive. Researchers may need to revisit some steps or modify them based on new information or insights gained during the process.

There are many types of interviews that employers use to evaluate candidates for a job. Each type has its advantages and disadvantages, depending on the purpose, scope, and context of the interview. Here are some of the common types of interviews and how they are conducted, along with their benefits and challenges:

Face-to-face interviews: These are interviews where you and the interviewer meet in person to discuss your credentials. The interviewer can be the employer, manager, someone from the HR department, or a third-party recruitment consultant hired by the company. The interview usually takes place on the company premises. Face-to-face interviews allow you to establish rapport and trust with the interviewer, demonstrate your communication skills and personality, and show your interest and enthusiasm for the job. However, face-to-face interviews can also be stressful, time-consuming, and costly, especially if you have to travel long distances or take time off from your current job (Indeed Editorial Team, 2023).

Panel interviews: These are interviews where you face questions from multiple interviewers at the same time. The interview panel may include people from different departments or disciplines. Panel interviews enable you to showcase your skills and knowledge to a diverse group of people, impress multiple decision-makers at once, and get a broader perspective of the company and the role. However, panel interviews can also be intimidating, challenging, and difficult to manage, especially if you have to balance your attention and eye contact among several people.

Phone interviews: These are interviews where you and the interviewer talk over the phone. Phone interviews are often used as a screening tool to narrow down the pool of candidates before inviting them for face-to-face or other types of interviews.[7] Phone interviews save time and money, reduce travel hassles, and allow you to focus on your verbal communication skills. However, phone interviews also limit your ability to convey your non-verbal cues, such as body language and facial expressions, create technical issues or distractions, and make it harder to build rapport and trust with the interviewer.

Video interviews: These are interviews where you and the interviewer communicate through a video conferencing platform, such as Skype or Zoom. Video interviews are similar to face-to-face interviews, except that they are conducted remotely. Video interviews are becoming more popular due to the COVID-19 pandemic and the rise of remote work. Video interviews offer convenience, flexibility, and safety, as you can conduct them from anywhere with a stable internet connection and a webcam. Video interviews also allow you to demonstrate your visual communication skills and personality. However, video interviews also pose technical challenges, such as poor audio or video quality, lagging or freezing screens, or background noise or interruptions. Video interviews also require you to prepare your environment, such as lighting, camera angle, and background.

Group interviews: These are interviews where you are interviewed along with other candidates for the same or similar positions. Group interviews are usually conducted by one or more interviewers who observe how you interact with other candidates and perform various tasks or activities.[8] Group interviews enable you to display your teamwork, leadership, problem-solving, and interpersonal skills. Group interviews also allow you to learn from other candidates and compare yourself with them. However, group interviews can also be competitive, stressful, and distracting. Group interviews can also make it harder for you to stand out from the crowd or express your individuality.

Behavioral interviews: These are interviews where you are asked to describe specific situations or scenarios from your past work experience that demonstrate how you handled various challenges or tasks. Behavioral interview questions usually start with phrases such as "Tell me about a time when …" or "Give me an example of how you …". Behavioral interview questions are based on the assumption that past behavior is the best predictor of future performance.[9] Behavioral interviews allow you to showcase your relevant skills and achievements in a concrete and detailed way. Behavioral interviews also help you to highlight your strengths and weaknesses in different situations. However, behavioral interviews can also be tricky, time-consuming, and demanding. Behavioral interviews require you to recall specific examples from your memory, use the STAR (Situation-Task-Action-Result) method to structure your responses (Indeed, 2023), and avoid vague or irrelevant answers.

Situational interviews: These are interviews where you are asked to imagine hypothetical situations or scenarios that you may encounter in the job role and explain how you would handle them. Situational interview questions usually start with phrases such as "What would you do if …" or "How would you deal with …". Situational interview questions are designed to test your problem-solving, decision-making, creativity, and adaptability skills.[10] Situational interviews allow you to demonstrate your potential and readiness for the job role. Situational interviews also help you to explore different possibilities and outcomes in various situations. However, situational interviews can also be challenging,

Sure, here are some of the common types of surveys, along with their benefits and challenges:

Cross-sectional surveys: Cross-sectional surveys collect data from a sample of individuals at a single point in time. Cross-sectional surveys are useful for describing the prevalence, distribution, and associations of variables in a population. The benefits of cross-sectional surveys are that they are easy to administer and analyze. However, they cannot establish causality or temporal relationships between variables.

Longitudinal surveys: Longitudinal surveys collect data from the same sample of individuals over time. Longitudinal surveys can be used to study changes, trends, and causal relationships in a population. The benefits of longitudinal surveys are that they can establish causality and temporal relationships. However, they are expensive and prone to attrition and bias.

Cohort surveys: Cohort surveys follow a specific group of individuals over time who share a common characteristic or experience. Cohort surveys can be used to study the incidence, risk factors, and outcomes of a disease or event. The benefits of cohort surveys are that they can provide detailed information on risk factors and outcomes. However, they require long-term follow-up and may suffer from selection bias.

Panel surveys: Panel surveys collect data from the same sample of individuals at multiple points in time. Panel surveys can be used to study stability, change, and causality in a population. The benefits of panel surveys are that they can provide rich information on stability and change. However, they require careful sampling and management to avoid attrition and panel conditioning.

Trend surveys: Trend surveys collect data from different samples of individuals at different points in time. Trend surveys can be used to study changes and patterns in a population over time. The benefits of trend surveys are that they can provide valuable insights into social change. However, they require careful measurement and interpretation to avoid spurious trends.

Diagnostic surveys: Diagnostic surveys are used to identify problems or needs in a population. Diagnostic surveys can be used to assess the quality, performance, or satisfaction of a product, service, or program. The benefits of diagnostic surveys are that they can provide actionable feedback on quality improvement. However, they require clear objectives and benchmarks for comparison.

Attitude and opinion surveys: Attitude and opinion surveys are used to measure people's beliefs, values, and attitudes towards a topic. Attitude and opinion surveys can be used to assess public opinion, political preferences, or social norms. The benefits of attitude and opinion surveys are that they can provide valuable information on public opinion. However, they require careful wording and sampling to avoid response bias.

Here are some tips and tricks to help you ensure you are developing a good survey questionnaire:

- Clearly state your intentions with the research
- Include instructions with your survey questionnaire
- Don't ask for personal information unless you need it
- Keep the questions concise
- Ask only one question at a time (avoid double-barreled questions)
- Make sure the questions are unbiased
- Ask questions that can be answered by your subjects
- Order/group questions according to subject.

The key to developing a good survey questionnaire is to keep it short while ensuring that you capture all of the information that you need. Before you even begin to design your survey questionnaire, you should develop a set of objectives for your research and list out the information that you are trying to capture.[11,12]

Different types of response scales can be used for questionnaires, such as Likert scales, numeric scales, semantic differential scales, and visual analog scales (Longe, 2024; Choudhury, 2019).

The choice of response scale depends on the purpose of the question and what is being measured. For example, a numeric scale might be more appropriate for measuring intensity or frequency, while a Likert scale can be used to measure agreement with a statement. Dichotomous questions, which only allow for a "yes" or "no" answer, are also an option, but they have the disadvantage of not allowing for a middle perspective.

An example of a semantic differential scale question:

How would you rate your experience with our customer service (QuestionPro, n.d.)?

Friendly – Unfriendly
Helpful – Unhelpful
Responsive – Unresponsive
Knowledgeable – Unknowledgeable

A Likert scale is a rating scale used to measure opinions, attitudes, or behaviors. It consists of a statement or a question, followed by a series of five or seven answer statements. Respondents choose the option that best corresponds with how they feel about the statement or question. Because respondents are presented with a range of possible answers, Likert scales are great for capturing the level of agreement or their feelings regarding the topic in a more nuanced way. However, Likert scales are prone to response bias, where respondents either agree or disagree with all the statements due to fatigue or social desirability or have a tendency toward extreme responding or other demand characteristics. Likert scales are common in survey research, as well as in fields like marketing, psychology, or other social sciences (Bhandari & Nikolopoulou, 2023).

DATA MANAGEMENT

Data management is a broad term that encompasses various aspects of collecting, processing, storing, and analyzing data. There are many methods and techniques for data management, depending on the type, source, quality, and purpose of the data (IBM, n.d.).

Some of the widely recognized methods for managing and storing data are:

Relational database management systems (RDBMS): These are software systems that store data in tables with rows and columns, and allow users to query and manipulate data using a structured query language (SQL). RDBMS are suitable for structured data that has predefined schemas and relationships. Examples of RDBMS include Oracle, MySQL, PostgreSQL, and Microsoft SQL Server.

NoSQL database management systems: These are software systems that store data in non-relational formats, such as key-value pairs, documents, graphs, or columns. NoSQL databases are suitable for unstructured or semi-structured data that has dynamic schemas and does not require strict consistency. Examples of NoSQL databases include MongoDB, Cassandra, Neo4j, and Redis (SAP, n.d.).

Data warehouse (DW) and data lake: These are large-scale repositories that store data from various sources for analytical purposes. A DW is a centralized database that integrates and transforms data into a consistent format, while a data lake is a distributed storage system that preserves data in its original or near-original format. A DW is suitable for structured or semi-structured data that has high quality and reliability, while a data lake is suitable for raw or unprocessed data that has high variety and volume (Davenport & Verma, 2018).

Hadoop and other open-source projects: These are software frameworks that enable distributed processing and storage of large-scale data using clusters of commodity hardware. Hadoop consists of several components, such as HDFS (a file system), MapReduce (a programming model), YARN (a resource manager), and Hive (a data warehouse). Other open-source projects that complement or extend Hadoop include Spark (a fast processing engine), Kafka (a streaming platform), and Flume (a data ingestion tool).

Master data

Master data is the critical business information that describes the core entities of an organization and provides context for business transactions (Gartner, n.d.). It includes fundamental entities such as customers, products, suppliers, employees, assets, and chart of accounts. Master data is typically persistent, with relatively low rates of change over time.

Some examples of master data entities are:

- **Customer data:** name, address, contact details, demographics, preferences.
- **Product data:** specifications, manufacturing/sourcing details, packaging info.
- **Supplier data:** name, location, contact info, payment terms.
- **Employee data**: compensation, skills, education, training.

The key characteristics that distinguish master data are:

- **Non-transactional:** provides context, not activity.
- **Relatively static**: slow to change.
- **Key business entities**: customers, products, etc.
- **Shared**: across systems, departments, applications.
- **Enterprise-wide impact**: critical to business operation.

High-quality master data enables informed business decisions, operational efficiency, regulatory compliance, and improved customer experience. Master Data Management focuses on the centralization, standardization, and governance of this critical business information.

Sometimes this master data is also referred to as golden records as they provide a "single source of truth" about the key business aspects.

Transactional data

Transactional data provides evidence that business activity has occurred, capturing the specifics of day-to-day operations, transactions, and interactions. This includes information such as orders, payments, shipments, service records, purchases, website clicks, and application usage data. In contrast to master data which changes slowly, transactional data is very dynamic, with high volumes of new activity data generated continuously.

Some examples of transaction data are:

- **Sales order transactions**: date, customer, product, quantity, price.
- **Payment records**: date, amount paid, payment method.
- **Shipping/fulfillment info**: delivery date, address, status.
- **Service ticket details**: date, issue summary, resolution steps.
- **Web clickstream data**: pages visited, date/time, user ID.
- **Video views**: video ID, viewer ID, minutes watched.

The key characteristics of transactional data include:

- **Activity-based**: represents an event/transaction.
- **Constantly changing:** high volume over time.
- **Detailed snapshots**: specifics of operations.
- **Time-related**: date/time stamps.
- **Analysis opportunities**: trends, performance.

Analyzing transactional data enables crucial business use cases like sales forecasting, promotion optimization, usage metrics, and more. The high volume requires specialized, scalable data infrastructure.

Metadata

Metadata is simply data that provides information about, or documentation of other data managed within an organization's systems and databases. It describes the structure, content, and context of the actual business data to help identify, manage, and understand that information.

Some examples of metadata are:

- **Data dictionaries**: outline the datasets, fields, definitions, and properties of captured data.
- **Data maps and models**: illustrate relationships between data elements.
- **Taxonomies**: classification systems, categories and tags.
- **Data lineage**: traces the origins and processing of data.
- **Data quality metrics**: accuracy, completeness, validity indicators.
- **Application and system logs**: audit trails tracking data transactions.

Key types of metadata

Various categories of metadata serve different purposes:

Descriptive metadata: Provides information to understand what the data represents and what it can be used for. This may include field names, definitions, data types, rule standards, encoding specifics, allowable values, and more.

Structural metadata: Describes the interrelationships, organization, and grouping of data within structures like relational database tables, data warehouse schemas, or JSON data objects. Defines entities and relationships.

Administrative metadata: Contains information to manage data like ownership, permissions, data sources and pipelines, refresh schedules, retention policies, privacy restrictions, etc. Supports data governance.

Statistical metadata: Provides statistical context about data like accuracy, completeness, quality metrics, algorithm/methodology details, variable distributions, reasons for missing values, etc. Important for analyzing and interpreting data.

Key purposes and usage

Metadata serves many invaluable functions across the data lifecycle including to:

- Enable data discovery, traceability, and understanding
- Streamline searchability with organization and context
- Provide transparency into origins, processing, and meaning
- Assess and monitor the quality, validity, and usability of data
- Manage security, access controls, and data policies
- Integrate, transform, map, and connect disparate data
- Support analytics usage, tools, and algorithms
- Maintain regulatory compliance on information storage
- Identify issues like bias, inaccuracies in data.

With comprehensive, well-structured metadata integrated into data management practices, organizations can fully leverage their information assets to drive key objectives while also governing those assets responsibly.

Metadata architectures and tools

There are dedicated metadata management platforms and repositories to store, organize, and share metadata elements. Metadata can also be embedded in, synchronized with, or linked to operational systems and databases.

Common components include:

- Metadata registries/repositories to systematize metadata
- Tagging, annotation, and classification tools
- Mappings to reconcile metadata across systems
- Capabilities to automatically capture metadata
- Metadata modeling, visualization
- Metadata standards and schemas
- Master data management integration
- Analytics to provide metadata insights
- Metadata life cycle management
- APIs and event-driven approaches.

With robust metadata management practices, companies can enable self-service access to reliable, understandable data while also providing critical transparency, governance, and oversight. The more technical aspects of data management are covered in Chapter 8 which deals with data infrastructure.

DATA CONTRACTS

For both internal and external data sources a "data contract" is a really good idea. There are of course distinct issues when we are talking about sourcing external data or providing access to data for external parties. In this section, we'll not go into the legal details but focus on the business scope of the contracts.

Internal data contract

An internal data contract is a formally documented agreement between departments or divisions within the same company that sets forth standardized rules, policies, constraints, and provisions concerning shared access and usage of proprietary data assets existing internally across the organization (Schmarzo, 2022).

It codifies the specifications, permissions, obligations, remedies, and restrictions that apply to the department, providing sensitive data resources and the department consuming those data, even though they operate under the broader corporate umbrella. This contract would be established between designated data stewards or executives representing their respective departments.

Aspects addressed typically include definitions of specific datasets covered, precise access permissions, notification and revocations protocols, query management, data masking/security precautions, allowed retention and destruction time frames, usage audit rights, infringement penalties, dispute resolution mechanisms, and overarching objectives of internal data sharing. Such governance facilitates equitable data circulation internally while upholding information security.

The Open Data Contract Standard that is being maintained by a 501 non-profit organization called "AIDA User Group (Artificial Intelligence, Data, and Analytics User Group)" (Group, 2023) is probably the most widely recognized framework.

Data-driven decision-making is the process of using data to inform and optimize business decisions. DDDM requires reliable, high-quality, and timely data that can be easily accessed and analyzed by data consumers. However, in a distributed data architecture, where data is produced and consumed by different services, teams, and applications, ensuring data quality and consistency can be challenging. This is where data contracts come in.

A data contract is an agreement between the producer and the consumer of a data product and is often facilitated by the "data team". A data product is any unit of data that can be exchanged, such as a file, a table, a stream, or an API. A data contract defines the structure, format, content, and meaning of the data product, as well as the expectations and obligations of both parties regarding the functionality, manageability, and reliability of the data product.

The purpose of a data contract is to ensure that both parties understand what they are sending and receiving, how to use it, how to handle errors or changes, and how to measure its

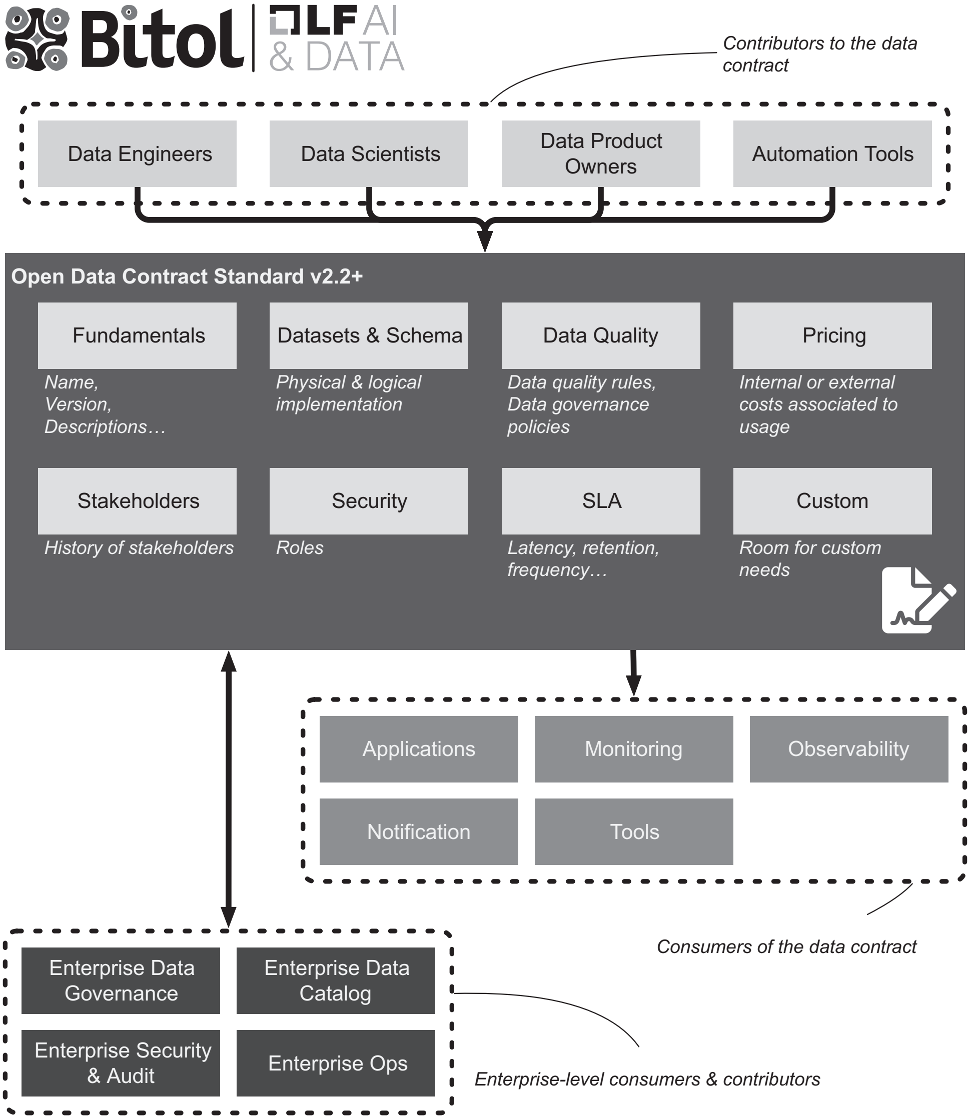

Bitol | LF AI & DATA
Contributors to the data contract
Data Engineers
Data Scientists
Data Product Owners
Automation Tools
Open Data Contract Standard v2.2+
Fundamentals
Name, Version, Descriptions…
Datasets & Schema
Physical & logical implementation
Data Quality
Data quality rules, Data governance policies
Pricing
Internal or external costs associated to usage
Stakeholders
History of stakeholders
Security
Roles
SLA
Latency, retention, frequency…
Custom
Room for custom needs
Applications
Monitoring
Observability
Notification
Tools
Consumers of the data contract
Enterprise Data Governance
Enterprise Data Catalog
Enterprise Security & Audit
Enterprise Ops
Enterprise-level consumers & contributors

FIGURE 5.1 Illustration of a data contract, its principal contributors, sections, and usage

quality and performance. A data contract also helps to establish trust and accountability between data producers and consumers, as well as to facilitate collaboration and communication.

A data contract can be expressed in different ways, depending on the type and complexity of the data product. For example, a data contract for a file or a table can be a schema that specifies the columns, types, constraints, and descriptions of the data. A data contract for a stream or an API can be a specification that defines the endpoints, parameters, responses, and documentation of the data. A data contract can also include metadata that provides additional information about the data product, such as its source, owner, purpose, lineage, quality metrics, SLAs, etc.

A data contract can be created and validated by different methods, depending on the tools and processes used by the data producer and consumer. For example, a data contract can be manually defined by using documentation or code annotations. A data contract can also be automatically generated or inferred by using schema registry or discovery tools. A data contract can be enforced by using validation or testing tools that check the compliance of the data product with the contract.

A data contract is a key concept for DDDM because it enables data consumers to access and analyze data with confidence and ease. By using data contracts, data consumers can:

- Discover and understand available data products and their characteristics
- Integrate and consume data products without errors or inconsistencies
- Monitor and evaluate the quality and performance of data products
- Provide feedback and request changes or improvements to data products.

Data contracts also benefit data producers by helping them to:

- Design and develop data products that meet the needs and expectations of data consumers
- Document and communicate the features and functionality of data products
- Manage and maintain data products with minimal disruptions or conflicts
- Improve and optimize data products based on feedback and metrics.

Data contracts can be seen as agreements between data producers and consumers that define and enforce the quality and consistency of data products in a distributed data architecture. Data contracts are often essential for enabling effective data collaboration and facilitating DDDM.

Contributors to the data contract

In addition to the regular people involved in data management (data engineers) and generation (data scientist and data product owners) we also have the systems that generate data. These have in the previous sections been described as internal and external data sources.

Consumers of the data contract

The consumers in a data contract can both be systems and people. In the case of systems, the system owner will in most cases stand in for the consumer as they are responsible for the quality of the actions taken based on the data. That could be a recommendation in a recommender-system that guides customers to the "right solution".

The people as consumers are those who analyze the data further or make use of it in their decision-making.

In both cases, there is a dependency on quality data by the consumer and they are therefore interested in a contract that can be enforced.

Components of the data contract

The contract itself has seven sections plus a custom area.

Fundamentals describe what the data is about, how it's stored/structured, who owns it, and the usage rights. This is defined in a table with 16 sections that are mandatory and 12 that are optional. This information should quickly allow a data-consumer to assess whether the dataset is relevant and useful for the context needed.

Dataset and schema dives deeper into the content of the dataset with naming and descriptions of individual tables and columns. For each of the columns both business name, data type, and definitions are provided. This will give a data consumer a good idea about the data modeling required in transforming the data.

Data quality is defined at both dataset and column levels. As opposed to the first two sections it doesn't have very many mandatory sections. Only the name of the tool used to complete the quality check and the quality template are required.

Pricing only has three variables to set and none are mandatory. It's the amount, currency, and unit. If data is to be fully valued within an organization, this should be carefully considered as with other transfer pricing settings.

Stakeholders are also devout of mandatory fields. Like with pricing it should be considered from a management perspective. Potentially using the RACI matrix to define who is responsible, owners, must be consulted and Informed of the data usage/transformation.

Security (roles) have a few mandatory sections. This is primarily to define who is capable of providing access to the data. There is some level of overlap with the stakeholder section here.

Service level agreement (SLA) defines the relevant lifespan of the dataset along with information like update frequency, availability, etc.

All these elements are provided in an XML machine-readable format that facilitates easy search and management. As of writing, this the most updated version, which is 2.2 and in rapid development, so it's recommended to check the project for updates.

Managing the data contract

Data contracts can be managed and enforced similarly to other contracts that are set up between departments and externally to the company.

Definition and purpose of external data contracts

A data contract can also be a legal agreement that specifies the rights and responsibilities of parties exchanging or collaborating on data assets. It then defines the allowed uses, liabilities, security controls, and other terms that the involved organizations agree to abide by in their treatment of the data.

Data contracts then explicitly spell out protections, permissions, consents, constraints, and processing details concerning the information being shared in a data supply chain. They can govern data exchanged between companies, provided to third-party vendors, accessed by research partners, or processed across borders.

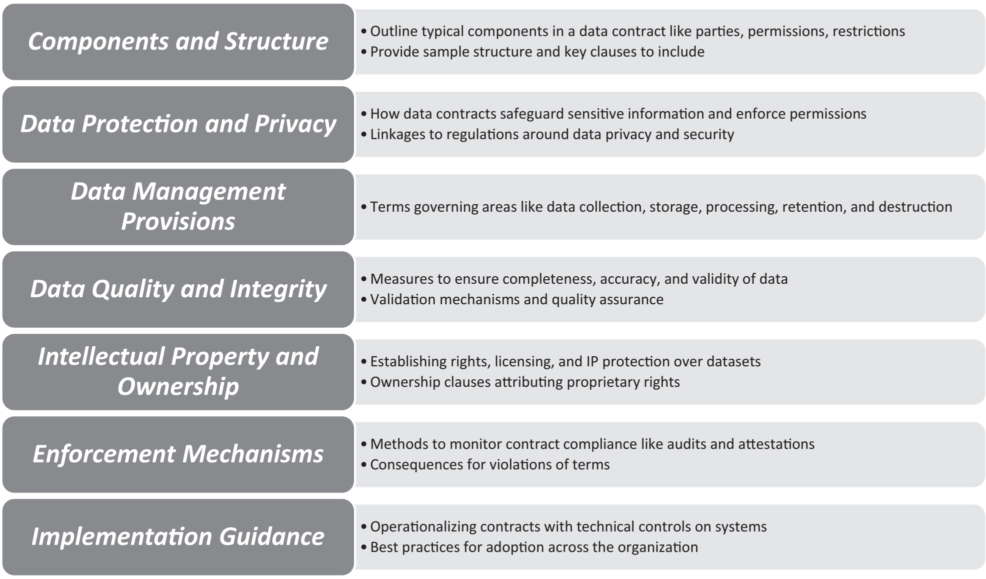
Purpose and use cases

Data contracts enable greater control, mutual understanding, and protection of valuable data assets. Key purposes include:

TABLE 5.1 Data contract purpose

Which data	Data access
Codifying what data is covered along with appropriate contracts allows organizations to specify datasets, permitted uses like analysis or marketing, prohibited uses, requisite data de-identification, or deletion requirements, etc. based on agreements or regulations.	Establishing security standards between data provider and data consumer. Contracts can demand encryption, access controls, breach notifications, third-party risk assessments, and other protections on systems handling the data.
Outlining quality criteria for completeness, accuracy, and delivery schedules of data flows. Documenting transparency and consent requirements especially around consumer data sharing to ensure privacy rights.	Clarifying legal responsibilities between collaborating organizations on data projects – defining liability, intellectual property rights, indemnification, regulatory applicability, etc.

Thus data contracts enable broader, mutually beneficial data use while providing guardrails to prevent misuse or overreach. They provide a mechanism to preemptively address key aspects of data partnerships.



Components and Structure
- Outline typical components in a data contract like parties, permissions, restrictions
- Provide sample structure and key clauses to include

Data Protection and Privacy
- How data contracts safeguard sensitive information and enforce permissions
- Linkages to regulations around data privacy and security

Data Management Provisions
- Terms governing areas like data collection, storage, processing, retention, and destruction

Data Quality and Integrity
- Measures to ensure completeness, accuracy, and validity of data
- Validation mechanisms and quality assurance

Intellectual Property and Ownership
- Establishing rights, licensing, and IP protection over datasets
- Ownership clauses attributing proprietary rights

Enforcement Mechanisms
- Methods to monitor contract compliance like audits and attestations
- Consequences for violations of terms

Implementation Guidance
- Operationalizing contracts with technical controls on systems
- Best practices for adoption across the organization

FIGURE 5.2 Data contract overview

EMERGING TRENDS

As data has become more and more in focus, the focus on data monetization has also increased. Companies that have gathered or are generating data that could be relevant for a wider audience have emerged as data sellers. With data sellers and buyers, brokers and data platform companies have emerged as well.

Data brokers

A data broker is a company that collects information from a variety of sources and then sells that data to other companies, typically for marketing or advertising purposes. Most often this has been personal information in anonymized form, which has become highly regulated by GDPR (see also Chapter 8).

Some key things to know about data brokers before engaging with them:

Data collection: Data brokers gather data from both online and offline sources, often without consumers' knowledge or direct consent. Sources include public records, retailer loyalty programs, social networks, surveys, website cookies, and more (Federal Trade Commission, 2014).

Types of data: Data brokers aggregate and trade in various types of consumer information including demographic, behavioral, socioeconomic, and interest data as well as contact information and unique identifiers.

Clients: The clients of data brokers include major corporations, specialized marketing firms, risk mitigation companies, people search services, and the government.

Industry size: There are hundreds of data brokers globally including large companies like Acxiom, Experian, and Oracle. The data brokerage industry generates billions in annual revenues.

Privacy concerns: Consumer advocates have raised objections regarding transparency, notice to consumers, consent requirements, potential errors or inaccuracies in collected data, and misuse or unauthorized access to data.

However, companies have emerged that act as brokers but are acting with people's consent and sharing the revenue with them. An example of that is Italian Togggle which stores personal information and thereby ensures it's only shared with people whom the user chooses.

Data marketplaces

A data marketplace is an online platform that enables companies or individuals to sell or purchase a variety of data assets.

Data assets: Structured data products such as datasets, data feeds, algorithms, data models, and analytics tools that have business value.

Examples of data marketplaces are:

- AWS Data Exchange: Enables data providers to package and sell datasets. Over 250 providers offer products.
- Dawex: Specialized in data exchange and data hub provider services across industries.
- Snowflake Data Marketplace: Allows buying and selling of data feeds, apps, algorithms, and learning models.
- Microsoft Azure Data Marketplace: Offers datasets from both Microsoft and third-party data providers.

Business models: Data providers can monetize data through subscriptions, usage-based pricing, or flat-fee downloads. Marketplaces typically take a percentage commission.

To summarize, data marketplaces enable companies to generate value from proprietary data while allowing other organizations to access datasets that can power business insights.

CASE

Happy Pops

Happy Pops is an upcoming Canadian manufacturer of high-quality popsicles (ice on stick based on fruit and water). They have over the last years experienced enormous success at specialist stores, events, cafés, and theme parks.

It's the summer of 2023 and Leila Keshavjee, the CEO, is standing in front of a major decision. Big retailers are showing an interest in adding Happy Pops to their product lineup. However, FMCG[13] is an area with high volumes and low margins. That means small errors in demand prediction and production can be very costly, and she would need to be sure that she was running the right campaigns at the right times as she would be carrying the cost and therefore the risk. Her family has been supplying retail for years and she follows industry trends and news through websites and newsletters regularly to keep up with what is happening.

According to Leila, the success of her company is to a large extent based on an unwavering focus on quality, service, and branding. That means they have been able to charge a higher price than competing popsicle manufacturers. This is, however, also needed as they have higher costs due to limited scale advantages. They are now procuring from the large-scale raw material manufacturers, but as they are still small they are not getting the big discounts, which can be a significant percentage.

The quality is controlled by the line staff and then Leila does the sampling of the boxes.

Another way that Happy Pops have been able to differentiate themselves is with alternative flavor combinations and collaborations with other brands.

Notably the organization behind the kids' show Sesame Street© recently reached out to Happy Pops and they have created two unique popsicles which are just now coming on sale.

Not all has been instant success, though. They tried to create a popsicle with pomegranate flavor, but the bitterness and the seeds meant that they now have a large number of wrappings that they cannot use. Another example is the tamarind flavor, which got to market but didn't have great sales volumes.

What sells and therefore how much to produce of each flavor is based on Leila's experience in running the business. They are using Quickbooks for their invoicing, but the rest of their numerical data is captured by using various Excel sheets stored in a local server. This includes the information that they are getting from their retailers by email when they run low on certain product variants.

A special thing that Happy Pops has noticed when looking at their sales over time is that they are not seeing a significant drop in sales over the winter period as would be expected for a product like Popsicles. They have an idea about it being used as a post-fitness snack substituting granola bars or gels. This information comes from comments on their social media channels. The product is both vegan and gluten free so it checks several of the boxes that "healthy living people" care about.

Several years ago Happy Pops made a list of cool places to be sold at. The list is comprised of companies in the primary target group that align with the values of Happy Pops. This is a focus on healthy, good-tasting products that are still fun. It doesn't include any of the large retailers as that was not really a realistic goal at the time, but now they are also re-evaluating the list and the criteria that shape it.

In addition to the retail business, Happy Pops also do corporate custom events. These are handheld activities where a special Popsicle wrapping is created and then handed out by Happy Pops employees. These events has so far worked as the sampling that is usually needed as instore sampling in the grocery chains; but it is also a very costly way of promoting the products, as the experience is that people don't buy the products, but only try them. They have so far not done post-event calculations of the profitability but are considering doing so.

Case question:

- Categorize the data sources that Happy Pops have available currently and do an assessment of their strengths (quality).
- What are the data gaps that you would recommend that Happy Pops fill before making a deal with a large retailer or would be required as part of a contract with them?
- Which available data would you primarily be basing your strategic decision on if you were Leila?
- What data sources would need to be monitored at production level now and in the future if the company were to make a deal for national retail distribution?

CONCLUSION

This chapter covers various aspects of identifying, collecting, and managing data to drive business value.

It starts by defining internal and external data sources. Internal data comes from systems within the organization like sales, marketing, finance, production, and HR. It provides a deeper look at the types of data within each system, such as customer data in CRM systems, financial data in ERP systems, manufacturing data in MES systems, and employee data in HRIS systems. Advantages of internal data include increased reliability, better security, and easier accessibility. Disadvantages include limited scope beyond the organization's transactions, potential bias, and high costs of collection and maintenance.

External data comes from outside the organization and includes sources like job portals, patent filings, public statistics, social media, and more. External data provides broader market insights, may reduce costs compared to collecting primary data, and offers independence from internal biases. However, relevance and quality issues can make applying external data challenging. Methods to access external data include web scraping, APIs, purchasing from data brokers, and public databases. Evaluating relevance, privacy, integration requirements, and interpretability is important when using external data sources.

The chapter then covers primary qualitative and quantitative data collection methods like interviews, surveys, focus groups, observations, and experiments. There is an overview of research project planning steps like stating goals, choosing methods, designing instruments, collecting data, analyzing, and reporting on findings. Different types of interviews and surveys are compared, highlighting their respective benefits and limitations. Guidelines on creating effective survey questionnaires are also provided.

Managing data assets is also discussed, touching on master data like customer data, transactional data like sales records, and metadata, which describes and gives context to underlying data. Metadata architectures, schemas, repositories, and tools are important for data governance, security, compliance, and enabling self-service access.

Emerging practices around data contracts are explored. Internal data contracts set permissions, restrictions, security protocols, and usage transparency between departments sharing datasets. External data contracts serve similar purposes between organizations and their partners, enforcing privacy, quality, intellectual property, liability, and regulatory clauses on shared data.

Finally, monetization models like data brokers and marketplaces are noted as trends in deriving business value from data. With increasing dependence on data and analytics, concepts around properly managing, exchanging, and commercializing data assets continue to gain priority.

KEY TERMS

Anonymization: Removing personally identifiable information from data to preserve anonymity.

APIs: Programming interfaces enabling software systems to exchange data.

Cohort surveys: Surveys tracking a group of people sharing a common characteristic over time.
CRM: Customer relationship management.
Cross-sectional surveys: Surveys collecting data at a single point in time.
Data brokers: Companies that aggregate consumer, business, and other data and then resell it.
Data contract: Agreement governing exchange, usage, and management of shared data assets.
Data lake: Highly scalable architecture to store large volumes of raw data efficiently.
Data marketplace: Platforms allowing the buying and selling of data products and assets.
Data masking: Encrypting or obscuring portions of data to ensure privacy.
Data warehouse: Central repository to store and analyze data for business intelligence.
Diagnostic surveys: Surveys conducted to evaluate the performance or quality of products, services, or programs.
ER: Enterprise resource planning system.
ETL (extract, transform, load): Process of pulling data out of source systems, modifying it, and loading it into a database or warehouse for analysis.
Experiments: Studies that manipulate variables to test hypotheses about causal relationships.
External data: Data that comes from outside an organization, e.g. social media, public records.
Focus groups: Moderated group discussions exploring perspectives on a topic.
GDPR: General Data Protection Regulation – European data privacy legislation.
Hadoop: Open source big data framework enabling distributed storage and processing on clusters.
Internal data: Data that comes from an organization's systems and processes.
Interviews: Direct conversations with selected participants to gather in-depth insights.
Longitudinal surveys: Surveys collecting data from the same group over an extended time frame.
Master data: Critical business entities like customers, products, and employees that provide context.
Metadata: Data providing information about other data assets.
Observations: Systematically recording behaviors, interactions, and processes.
Panel surveys: Surveys gathering data from the same group at multiple points in time.
Qualitative data: Descriptive, non-numerical data like interview transcripts.
Quantitative data: Numerical data that can be measured and analyzed statistically.
ROI: Return on investment.
SLA: Service level agreement that defines expected levels of service between providers and clients.
Surveys: Structured questionnaires to collect information from a sample population.
Transactional data: Activity and operational data like sales transactions and website clicks.
Trend surveys: Surveys gathering data from different samples to identify shifts over time.
Web scraping: Automated extraction of data from websites.

REVIEW QUESTIONS

Here are 25 review questions to test understanding of the key points in the chapter:

1 What are some examples of internal data sources that can provide valuable business insights?
2 What are some risks or limitations companies should be aware of when using internal data?

3 Name three sources organizations can leverage to collect relevant external data.

4 True or false: External data sources provide completely objective and accurate insights.

5 Qualitative data collection is useful for gathering _______ while quantitative data enables ________.

6 Interviews provide more breadth while surveys provide more depth. True or false?

7 Name two ways a longitudinal survey differs from a cross-sectional survey.

8 What are the two defining traits of master data versus transactional data?

9 Metadata serves which vital need for managing data as an asset?

10 What key purpose does an internal data contract serve between company divisions?

11 What are the two main parties involved in an external data contract?

12 Name two privacy or security-focused clauses that would be important to cover in an external data contract

13 Data brokers aggregate consumer data from various online and offline sources for _______.

14 True or False: Data marketplaces enable organizations to share data internally across divisions.

15 List two precautions businesses should take when launching surveys.

16 Name one difference between a data warehouse and a data lake.

17 What does the term API stand for and what is its relation to data?

18 Which data collection method could assess user satisfaction with a new software feature?

19 What does PII stand for and why must it be protected per data privacy regulations?

20 True or false: Data lakes and data warehouses serve the same functions for organizations.

21. What are two best practices for effectively managing master data?

22 External data relevance and quality should be evaluated carefully before ________________.

23 What technology is valuable for automating the capture and organization of metadata?

24 According to the chapter, what are some emerging technologies relevant to leveraging data as an asset?

25 What does GDPR legislation relate to?

Answers to review questions

1 Possible answers: Sales data, marketing data, financial data, production data, customer data, web analytics

2 Possible answers: Biased perspectives, limited scope, costly data management

3 Possible answers: Social media, public government databases, web scraping consumer review sites, data brokers

4 False (Quality and relevance issues can exist with external data)

5 Possible answers: Subjective viewpoints/experiences; Statistical analysis

6 False

7 Possible answers: Collects data over a longer timeframe; tracks same respondents over time

8 Possible answers: Master data changes slowly, provides core business context/transactional data captures daily operations and events

9 Possible answers: Provides descriptive information about datasets to enable discovery, tracking, interpreting, and governing data

10 Possible answer: Establishes standardized rules, policies, and permissions around internal data sharing

11 Possible answer: Data provider and data consumer organizations
12 Possible answers: Data encryption requirements, breach notification specifications, geographic restrictions onata hosting
13 Possible answer: Reselling the data to other companies, mainly for marketing and advertising
14 False
15 Possible answers: Ensure proper permissions, avoid collecting unnecessary personal data
16 Possible answer: Data warehouses store transformed data for analysis while data lakes store raw data
17 Possible answer: Application Programming Interface – Used to exchange data between software systems
18 Possible answer: Survey
19 Possible answer: Personally Identifiable Information. It is protected to safeguard individual privacy rights
20 False
21 Possible answers: Centralize into a master data management system, carefully govern changes to core business entities
22 Possible answer: using it for business analytics or decisions
23 Possible answer: Metadata management platforms
24 Possible answers: Data exchanges, marketplaces, AI/ML algorithms
25 Possible answer: Data privacy protection for individuals in the EU

NOTES

[1] Changing data so it can be analyzed e.g. image categorization or renaming according to standard terminology.
[2] https://uk.indeed.com/career-advice/career-development/advantages-of-internal-data
[3] www.scribbr.com/methodology/qualitative-quantitative-research/
[4] https://us.sagepub.com/sites/default/files/upm-assets/106363_book_item_106363.pdf
[5] https://careerfoundry.com/en/blog/data-analytics/difference-between-quantitative-and-qualitative-data/
[6] www.leadquizzes.com/blog/data-collection-methods/
[7] https://in.indeed.com/career-advice/interviewing/types-of-interviews
[8] Ibid.
[9] Ibid.
[10] Ibid.
[11] www.sciencebuddies.org/science-fair-projects/references/how-to-design-a-survey
[12] www.england.nhs.uk/wp-content/uploads/2018/01/bitesize-guide-writing-an-effective-questionnaire.pdf
[13] Fast Moving Consumer Goods.

BIBLIOGRAPHY

Bhandari, P. (2023, June 21). Data collection: Definition, methods & examples. Scribbr.

Bhandari, P., & Nikolopoulou, K. (2023, June 22). What Is a Likert scale? | Guide & Examples. Scribbr1.

Brown, S. (2021, February 18). *Why external data should be part of your data strategy*. MIT Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/why-external-data-should-be-part-your-data-strategy

Choudhury, S. B. R. (2019, January 31). *Types of measurement scales in survey questionnaires*. Project Guru. www.projectguru.in

Davenport, T. A. (2021, December 22). *AI can help companies tap new sources of data for analytics*. *Harvard Business Review*. https://hbr.org/2021/03/ai-can-help-companies-tap-new-sources-of-data-for-analytics

Davenport, T. H., & Verma, A. (2018, July 20). Data management techniques, approaches, and tools. *Deloitte Insights*. www2.deloitte.com/us/en/insights/topics/analytics/data-management-techniques-approaches-tools.html

Federal Trade Commission. (2014). *Data brokers: A call for transparency and accountability*. www.ftc.gov/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014

Gartner. (n.d.). Definition of Master Data Management (MDM): IT Glossary1 Definition of Master Data Management (MDM). Gartner.

George, T. (2022, October 22). *Mixed methods research: Definition, guide, & examples*. Scribbr. www.scribbr.co.uk/research-methods/mixed-methods/

Group, A. U. (2023, December 10). *Bitol*. Github. https://github.com/bitol-io/open-data-contract-standard

Harvard Business School Online. (2021, December 2). 7 data collection methods in business analytics. *Harvard Business School Online's Business Insights Blog*. https://online.hbs.edu/blog/post/data-collection-methods

IBM. (n.d.). *What is data management?* IBM. www.ibm.com/topics/data-management

Indeed Editorial Team. (2023). https://uk.indeed.com/career-advice/career-development/advantages-of-internal-data

Indeed Editorial Team. (2023, June 9). 7 types of job interviews (plus how to prepare and tips). https://in.indeed.com/career-advice/interviewing/types-of-interviews

Kilduff, E. (2023, 12 13). *Thoughtspot*. Internal Data + External Data = Augmented Decision-Making. www.thoughtspot.com/data-chief/internal-data-and-external-data-equals-augmented-decision-making

Longe, B. (2024, January 12). Survey scale: Definitions, types + [question examples]. Formplus1. www.formpl.us/blog/survey-scale

Marr, B. (2022, March 30). *Why external data is so important for every business*. Forbes. www.forbes.com/sites/bernardmarr/2022/03/30/why-external-data-is-so-important-for-every-business/?sh=1826ebdd9119

QuestionPro. (n.d.). Semantic differential scale, example, and question types. www.questionpro.com/semantic-differential-scale.html

SAP. (n.d.). *What is data management?* SAP1. www.sap.com/products/technology-platform/what-is-data-management.html

Schmarzo, B. (2022, March 23). *The Internal Data Contract*. LinkedIn. www.linkedin.com/pulse/internal-data-contract-bill-schmarzo/

Talend. (2023, December 23). *What is a data source? Definitions and examples*. Talend. www.talend.com/resources/data-source/

CHAPTER 6

Data Visualization and Presentation

How to Present and Communicate Data Effectively for Decision-Making

In today's world, data is everywhere. It can help us understand complex phenomena, identify trends and patterns, and make better decisions. However, data alone is not enough. To make sense of data, we need to present and communicate it in a way that is easy to understand, engaging, and actionable. This chapter will teach you how to do that. You will learn how to analyze your stakeholders and context for your data communication, how to choose and create data visualizations that suit your purpose and audience, and how to present and communicate your data findings using storytelling techniques. You will also learn how to apply data visualization in different scenarios, such as reports, graphics, and interactive dashboards.

Data is a powerful tool for making decisions, but only if it is presented and communicated effectively.

- How can you turn raw data into compelling stories that inform and persuade your audience?
- How can you choose the right data visualization techniques that highlight the key insights and messages in your data?
- How can you tailor your data presentation and communication to different stakeholders and contexts?

These are some of the questions that this chapter will address. You will learn how to use the CED (conclusion, evidence, data) framework to structure your data findings, how to select and create appropriate data visualizations that convey your message clearly and persuasively, and how to present and communicate your data findings to different audiences and contexts using storytelling techniques. You will also learn how to use data visualization in various formats, such as reports, graphics, and interactive dashboards.

DOI: 10.4324/9781003457787-6

LEARNING GOALS:

L6.1 Analyze your stakeholders and context for your data communication and identify their needs and expectations

L6.2 Select and create data visualizations that suit your purpose and audience, using principles of design, clarity, and persuasion

L6.3 Present and communicate your data findings using storytelling techniques, such as narrative, emotion, and action

L6.4 Apply data visualization in different formats, such as reports, graphics, and interactive dashboards, and evaluate their effectiveness and impact

PREPARING FOR DATA PRESENTATION

Before you present and communicate your data findings, you need to analyze your stakeholders and the context for your data communication. This will help you identify their needs and expectations and tailor your message accordingly. In this section, you will learn how to use various models and tools to conduct a stakeholder and context analysis and how to apply the results to your data communication plan.

Several models are relevant to cover for analyzing your stakeholders and the context for your data communication and identifying their needs and expectations.

Power–interest grid

This model categorizes stakeholders based on their power (their authority in the project's completion) and their interest (their level of concern over the project's success). It helps to prioritize the communication and attention given to different stakeholders according to their position on the grid (Indeed, n.d.).

The power–interest grid is a model that helps to categorize stakeholders based on their power and interest in the project. Power refers to the stakeholder's authority or influence over the project's completion, and interest refers to the stakeholder's level of concern or involvement in the project's success. The model uses a grid with four quadrants, where power increases along the vertical axis and interest increases along the horizontal axis. The stakeholders are placed on the grid according to their level of power and interest. The model helps to prioritize the communication and attention given to different stakeholders according to their position on the grid (Project Management, n.d.).

The four quadrants are:

- High power – High interest:
  These are the stakeholders who have the most influence and impact on the project and require the most attention and engagement. They should be closely managed and consulted throughout the project.

  These could be the project sponsor, the senior management, the key customers, the main suppliers, or the regulatory bodies. They are the ones who have a direct stake in the project's success and can make or break the project.
- High power – Low interest:
  These are the stakeholders who have the authority to affect the project but may not be very interested or involved in it. They should be kept satisfied and informed about the project's progress and benefits.

  These could be the project owner, the finance department, the legal department, the media, or the competitors. They are the ones who have the authority to influence the project but may not be very concerned or engaged with it.
- Low power – High interest:
  These are the stakeholders who have a high level of interest or involvement in the project but may not have much influence or authority over it. They should be kept informed and consulted about the project's issues and outcomes.

  These could be the project team, the end-users, the local community, the interest groups, or the advocates. They are the ones who have a high level of interest or involvement in the project but may not have much power or control over it.
- Low power – Low interest:
  These are the stakeholders who have little influence and impact on the project and may not be very interested or involved in it. They should be monitored and communicated with occasionally to ensure they do not become problematic or dissatisfied (KnowledgeHut, n.d.).

  These could be the other departments, the public, the minor suppliers, the subcontractors, or the consultants. They are the ones who have little influence and impact on the project and may not be very interested or involved in it.

Some possible advantages and disadvantages of using this model are:

Advantages:

- It is a simple and easy-to-use tool that can help to identify and categorize stakeholders quickly and visually
- It can help to prioritize
- the communication and attention given to different stakeholders according to their level of power and interest
- It can help to plan the stakeholder engagement strategies and activities based on their position on the grid
- It can help to avoid overlooking or neglecting any important stakeholders who may affect or be affected by the project.

Disadvantages:

- It is a subjective and qualitative tool that may not capture the complexity and dynamics of stakeholder relationships and interactions
- It may not account for the different types and levels of power and interest that stakeholders may have, such as formal or informal, direct or indirect, positive or negative, etc.
- It may not reflect the changes in stakeholder power and interest over time or across different stages of the project
- It may not consider the potential conflicts or synergies among different stakeholders who may have different or competing interests and expectations.

The same approach can also be used in impact assessment matrixes. Each comes with its own set of advantages and disadvantages but it's to a large extent the focus and available data that should guide in choosing which to go forward with.

Influence–impact grid

While quite like the power–interest grid, this model categorizes stakeholders based on their influence (their ability to control project decisions) and their impact (their ability to change the result of the project). It helps to identify the stakeholders who have the most influence and impact on the project and plan the engagement strategies accordingly. So, in combination, the two grids can provide a comprehensive overview of the stakeholder landscape. The usage is similar so sometimes it makes sense to combine them in the power–influence grid.

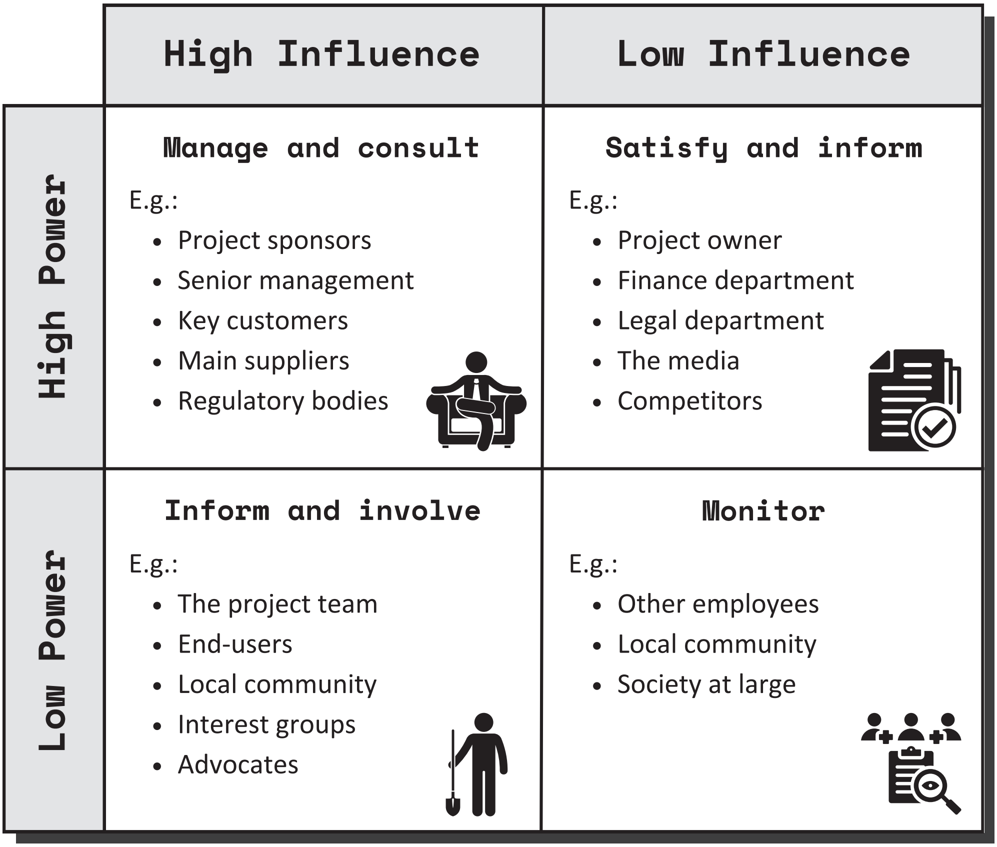
Power–influence grid

This model combines the previous two models and categorizes stakeholders based on their power and influence. It helps to assess the stakeholder's potential to support or oppose the project and manage the risks and opportunities accordingly (BrightHub PM, n.d.).

The power–influence grid is a model that helps to categorize stakeholders based on their power and influence in the project. Power refers to the stakeholder's authority or influence over the project's completion, and influence refers to the stakeholder's ability to affect the project's decisions. The model uses a grid with four quadrants, where power increases along the vertical axis and influence increases along the horizontal axis. The stakeholders are placed on the grid according to their level of power and influence. The model helps to assess the stakeholder's potential to support or oppose the project and manage the risks and opportunities accordingly (Project Management, n.d.).

The four quadrants are:

High power – High influence: These are the stakeholders who have the most authority and impact on the project and require the most attention and engagement. They should be closely managed and consulted throughout the project.



	High Influence	Low Influence
High Power	**Manage and consult** E.g.: • Project sponsors • Senior management • Key customers • Main suppliers • Regulatory bodies	**Satisfy and inform** E.g.: • Project owner • Finance department • Legal department • The media • Competitors
Low Power	**Inform and involve** E.g.: • The project team • End-users • Local community • Interest groups • Advocates	**Monitor** E.g.: • Other employees • Local community • Society at large

FIGURE 6.1 Power–influence grid illustrated

These could be the project sponsor, the senior management, the key customers, the main suppliers, or the regulatory bodies. They are the ones who have a direct stake in the project's success and can make or break the project.

High power – Low influence: These are the stakeholders who have the authority to affect the project but may not have much impact on the project's decisions. They should be kept satisfied and informed about the project's progress and benefits.

These could be the project owner, the finance department, the legal department, the media, or the competitors. They are the ones who have the authority to influence the project but may not have much impact on its decisions.

Low power – High influence: These are the stakeholders who have a high level of impact on the project's decisions but may not have much authority or control over them. They should be kept informed and involved in the project's issues and outcomes.

These could be the project team, the end-users, the local community, the interest groups, or the advocates. They are the ones who have a high level of impact on the project's decisions but may not have much power or control over it.

Low power – Low influence: These are the stakeholders who have little authority and impact on the project and may not be very concerned or engaged with it. They should be monitored and communicated with occasionally to ensure they do not become problematic or dissatisfied (JanBask Training, n.d.).

These could be the other departments, the general public, the minor suppliers, the subcontractors, or the consultants. They are the ones who have little authority and impact on the project and may not be very interested or involved in it.

Some possible advantages and disadvantages of using this model are:

Advantages:

- It is a simple and easy-to-use tool that can help to identify and categorize stakeholders quickly and visually
- It can help to assess the stakeholder's potential to support or oppose the project and manage the risks and opportunities accordingly
- It can help to plan the stakeholder engagement strategies and activities based on their position on the grid.

Disadvantages:

- It is a subjective and qualitative tool that may not capture the complexity and dynamics of stakeholder relationships and interactions
- It may not account for the different types and levels of power and influence that stakeholders may have, such as formal or informal, direct or indirect, positive or negative, etc.
- It may not reflect the changes in stakeholder power and influence over time or across different stages of the project
- It may not consider the potential conflicts or synergies among different stakeholders who may have different or competing interests and expectations.

To use any of the three grid models efficiently you should prioritize the use of multiple sources of information and data to identify and assess the stakeholders, such as existing documents, interviews, surveys, workshops, etc. It can also help to involve the project team and other relevant stakeholders in the process of creating and validating the grid, to ensure accuracy and consensus.

When populating the grid make sure you use clear and consistent criteria to define and measure the power, impact, influence, and interest of the stakeholders. That could be their authority, influence, resources, involvement, expectations, etc.

Then make sure that you review and update the grid regularly throughout the project life cycle, to capture any changes in stakeholder power, impact, and interest or any new or emerging stakeholders.

The grid should be used as a starting point for planning and implementing stakeholder engagement activities, such as communication, consultation, collaboration, etc. In the following sections, we will dive deeper into the management after the stakeholder analysis is done.

Stakeholder analysis

Stakeholder analysis is the general term for a range of techniques that map and understand the power, positions, and perspectives of the stakeholders who have an interest in or are affected by a particular policy reform. It helps to analyze the stakeholder's interests, expectations, alliances, conflicts, and influence on the analysis and decision process (Forbes Advisor, n.d.). The previous sections form the basis for your stakeholder management. Remember that the stakeholder landscape can change over time.

Stakeholder management data from a project manager perspective

Stakeholder management is the process of identifying, engaging, and communicating with the people who have an interest or influence in the project. Stakeholders can be internal or external to the organization, and they can have different levels of power, interest, and influence on the project outcomes. Effective stakeholder management is essential for project success, as it helps to build trust, alignment, and support among the stakeholders, and to avoid or resolve conflicts and issues that may arise during the project life cycle (Northeastern University, n.d.).

One of the key aspects of stakeholder management is to support their decision-making process using data. Data-driven decision-making (DDDM) is the process of making organizational decisions based on actual data rather than intuition or observation alone.

Data can help project managers and stakeholders to:

- Define clear and measurable project objectives and indicators of success
- Identify and prioritize the key project risks and opportunities
- Monitor and evaluate the project's progress and performance
- Identify and address any gaps, issues, or deviations from the project plan
- Learn from the project outcomes and improve future projects.

To support stakeholder decision-making using data, project managers need to:

- Know the mission, vision, and goals of the project and the organization
- Identify the key stakeholders and their roles, expectations, and information needs
- Collect relevant, reliable, and valid data from various sources, such as surveys, interviews, observations, documents, etc.
- Analyze and interpret the data using appropriate methods and tools, such as statistics, visualization, modeling, etc.
- Communicate the data insights and recommendations to the stakeholders in a clear, concise, and persuasive way, using reports, dashboards, presentations, etc.
- Involve the stakeholders in the data analysis and decision-making process, by soliciting their feedback, input, and opinions

- Follow up with the stakeholders on the implementation and impact of the decisions made based on data.

Stakeholder management using data

Stakeholder management is the process of identifying, engaging, and communicating with the people who have an interest or influence in the project. Stakeholders can be internal or external to the organization, and they can have different levels of power, interest, and influence on the project outcomes. Effective stakeholder management is essential for project success, as it helps to build trust, alignment, and support among the stakeholders, and to avoid or resolve conflicts and issues that may arise during the project life cycle.

One of the key aspects of stakeholder management is to support their decision-making process using data. DDDM is the process of making organizational decisions based on actual data rather than intuition or observation alone. Data can help project managers and stakeholders to define clear and measurable project objectives and indicators of success; identify and prioritize the key project risks and opportunities; monitor and evaluate the project progress and performance; identify and address any gaps, issues, or deviations from the project plan; learn from the project outcomes and improve future projects.

To support stakeholder decision-making using data, project managers need to know the mission, vision, and goals of the project and the organization; identify the key stakeholders and their roles, expectations, and information needs; collect relevant, reliable, and valid data from various sources; analyze and interpret the data using appropriate methods and tools; communicate the data insights and recommendations to the stakeholders in a clear, concise, and persuasive way; involve the stakeholders in the data analysis and decision-making process; follow up with the stakeholders on the implementation and impact of the decisions made based on data.

By using data to support stakeholder decision-making, project managers can enhance their stakeholder relationships and increase their chances of project success.

Context model

The context model (Medium: Business Architected, n.d.) is a visual tool that shows the relationship between a system (such as a project, a product, or a service) and its environment (such as its users, customers, suppliers, competitors, regulators, etc.). It helps to define the scope and boundaries of the system, identify the relevant stakeholders and their roles, and communicate the system's context clearly and efficiently.

To elaborate on the definition of a context model, I will provide some examples and explain how to create and use a context model.

A context model is a visual tool that shows the relationship between a system (such as a project, a product, or a service) and its environment (such as its users, customers, suppliers, competitors, regulators, etc.). It helps to define the scope and boundaries of the system, identify the relevant stakeholders and their roles, and communicate the system's context clearly and efficiently.

A system context diagram shows the physical scope of the system being designed, which could include the user as well as the environment and other actors. For example, a system context diagram for a catering company could show how it interacts with its customers, suppliers, delivery services, health authorities, etc.

A context-sensitive grammar defines the surrounding text of a lexical element. This enables a parser to disambiguate the meaning of the element based on its context. For example, a context-sensitive grammar for natural language processing could show how the word "bank" can have different meanings depending on whether it is followed by "of", "on", or "account".

A gene sequence analysis identifies the surrounding elements in a gene sequence. This helps to disambiguate the role of the gene based on its context. For example, a gene sequence analysis for bioinformatics could show how a gene is influenced by its neighboring genes or regulatory elements.

An ontology provides disambiguation of a subject via semantic analysis of information related to the subject. For example, an ontology for knowledge management could show how a concept is related to other concepts in terms of properties, categories, or hierarchies (Wikipedia, n.d.).

To create a context model, you need to:

1 Identify the system that you want to model and its purpose or goal
2 Identify the external entities that interact with the system or have an interest or influence in it. These could be organizations, departments, systems, processes, people, etc.
3 Identify the interactions between the system and the external entities. These could be exchanges of data, physical objects, funds, etc.
4 Label the interactions to show the flow or direction of the exchange
5 Draw a diagram that shows the system as a central element and the external entities as boxes around it Connect them with lines that represent the interactions and label them with flows
6 Review and validate the diagram with stakeholders and experts to ensure its accuracy and completeness.

To use a context model, you can:

- Use it as a communication tool to explain the scope and boundaries of the system to stakeholders and users
- Use it as an analysis tool to identify potential impacts of changes or issues on the system or its environment.
- Use it as a design tool to start functional decomposition or requirements discovery for the system or its components
- Use it as an estimation tool to understand the number and complexity of integrations or interfaces for the system
- Use it as a testing tool to create an adequate test strategy and coverage for the system.

A context model is a simple and powerful tool that can help you understand and communicate the context of a system in relation to its environment. It can also help you improve your decision-making and problem-solving skills by providing you with relevant information and insights.

Often describing the context model using a standard notation like UML (universal markup language) can be beneficial in the communication of the context.

DATA VISUALIZATION MODELS

Data visualization can be viewed from different perspectives. Which you choose depends on the purpose, function, and stakeholder you want to communicate to and which data and decision that is sought.

The data visualization process model describes the steps involved in creating effective data visualizations, such as defining the purpose and audience, selecting the data, choosing the visual form, applying design principles, and evaluating the results (Unzueta, n.d.). This model helps you plan and execute your data visualization project by following a systematic and iterative approach. It also helps you avoid common mistakes and pitfalls in data visualization by providing you with best practices and guidelines for each step. The data visualization process model can be represented as a cycle of six stages: define, collect, explore, analyze, design, and present.

The data visualization grammar model defines the basic components and rules of constructing data visualizations, such as data, marks, channels, scales, axes, legends, etc. (Tableau, n.d.). This model helps you understand and communicate the structure and logic of data visualizations by using a common vocabulary and syntax. It also helps you create and customize data visualizations by using a modular and flexible approach. The data visualization grammar model can be represented as a formula: data + marks + channels + scales + axes + legends = visualization.

The data visualization types model classifies different types of data visualizations based on their functions, such as comparison, correlation, distribution, composition, etc. (Substack, n.d.). This model helps you select and create data visualizations that suit your purpose and audience by providing you with a taxonomy and a framework. It also helps you understand and interpret data visualizations by providing you with a context and a rationale. The data visualization types model can be represented as a matrix of four categories: relationship (how variables relate to each other), composition (how parts make up a whole), distribution (how values are spread out), and change (how values change over time). We will take a look at the Abela chart choosing framework which is an example of this.

The data visualization perception theory explains how humans perceive and process visual information, such as color, shape, size, position, etc., and how to use them effectively to convey data insights (Hevo Data, n.d.). This theory helps you design and evaluate data visualizations by providing you with principles and guidelines based on cognitive psychology and neuroscience. It also helps you optimize and enhance data visualizations by providing you with techniques and tricks based on perceptual illusions and biases. The data visualization perception theory can be represented as a set of rules or heuristics for choosing and using visual encodings. In Chapter 1 cognitive biases were discussed that related to this.

The data visualization cognition theory explores how humans understand and interpret data visualizations, such as preattentive processing, gestalt principles, mental models, etc., and how to use them to enhance clarity and persuasion (IBM, n.d.). This theory helps you communicate and present data visualizations by providing you with strategies and methods based on cognitive science and rhetoric. It also helps you influence and persuade your audience by providing you with tools and tactics based on emotional appeal and storytelling. The data visualization cognition theory can be represented as a set of skills or competencies for creating and delivering visual stories.

These models and theories can help you select and create data visualizations that suit your purpose and audience by providing you with a framework, a vocabulary, a taxonomy, and a set of guidelines for data visualization. They can also help you avoid common pitfalls and errors in data visualization by informing you of the best practices and principles of design. By learning these models and theories, you can improve your data visualization skills and create more effective and engaging visual stories with data.

The following sections focus on the grammar model along with the perception and cognition theories.

The data visualization grammar model

The data visualization grammar model is a framework that enables us to emphasize the importance of separating data from its visual representation and proposes a structured framework for creating graphics. It is based on the idea that visualizations can be broken down into discrete components, such as data, aesthetics, and geometry. These components can be combined in various ways to create different types of visualizations (Sarkar, 2018).

The data visualization grammar model consists of the following core components:

Data: The actual variables to be plotted. These can be either continuous (numeric) or discrete (categorical) variables. For example, in a scatter plot, the data could be the height and weight of different individuals.

Aesthetics: Aesthetics deals with the axes based on the data dimensions, positions of various data points in the plot, and other visual properties such as color, shape, size, etc. These properties can be mapped onto different data variables to encode information. For example, in a scatter plot, the x-axis could be mapped to height, the y-axis could be mapped to weight, and the color could be mapped to gender.

Geometry: Geometry refers to the type of graphical element used to represent the data. These can be points, lines, bars, pies, etc. Different geometries are suitable for different types of data and purposes. For example, in a scatter plot, the geometry is point, which is good for showing relationships between two continuous variables.

Scales: Scales are functions that map data values to aesthetic values. They define how the data is transformed and displayed on the plot. For example, a linear scale maps data values proportionally to aesthetic values, while a logarithmic scale maps data values exponentially to aesthetic values. Scales can also affect the appearance of axes and legends.

Axes: Axes are graphical elements that show the range and scale of the data along each dimension. They usually have labels and ticks to indicate the values and units of the data. Axes help orient and guide the viewer in interpreting the plot.

Legends: Legends are graphical elements that explain the meaning of different aesthetics used in the plot. They usually have labels and symbols that correspond to the aesthetics. Legends help clarify and enhance the plot.

The data visualization grammar model can be represented as a formula: data + aesthetics + geometry + scales + axes + legends = visualization. This formula shows how different components can be combined to create a visualization.

The data visualization grammar model helps us understand and communicate the structure and logic of data visualizations by using a common vocabulary and syntax. It also helps us create and customize data visualizations by using a modular and flexible approach. By learning this model, we can improve our data visualization skills and create more effective and engaging visual stories with data.

Especially when collaborating on the creation of visualizations it's important to have a common vocabulary that can be agreed upon. A use-case could be where the analyst is working with the CFO to prepare presentations for the management team.

The data visualization perception theory

The data visualization perception theory is a theory that explains how humans perceive and process visual information, such as color, shape, size, position, etc., and how to use them effectively to convey data insights. This theory is based on the principles of cognitive psychology and neuroscience, and it helps us design and evaluate data visualizations by providing us with guidelines and best practices based on human visual perception (Donska, 2020).

The data visualization perception theory consists of the following core concepts:

Pre-attentive processing: The ability of our visual system to detect certain basic features of a scene in a very short time (less than 200 milliseconds), without conscious attention. These features include color, orientation, size, shape, motion, etc. Pre-attentive processing can be used to highlight important or unusual data points or patterns in visualization by using different visual encodings for them. For example, using a different color or shape for an outlier can make it stand out from the rest of the data.

Gestalt principles: A set of rules that describe how our visual system organizes and groups visual elements into meaningful wholes. These principles include proximity (elements that are close together are perceived as belonging together), similarity (elements that are similar in some way are perceived as belonging together), continuity (elements that form a smooth line or curve are perceived as belonging together), closure (elements that form a closed shape are perceived as belonging together), symmetry (symmetrical elements are perceived as belonging together), and common fate (elements that move in the same direction are perceived as belonging together). Gestalt principles can be used to create coherent and clear visualizations by grouping related data points or patterns using appropriate visual encodings. For example, using the same color or shape for data points that belong to the same category can make them appear as a group.

Mental models: Internal representations of how things work or behave in the real world. They help us understand and interact with complex systems or phenomena. Mental models can be influenced by our prior knowledge, experience, expectations, and assumptions. Mental models can affect how we perceive and interpret data visualizations by influencing what we pay attention to, what we ignore, what we infer, and what we question. Mental models can be used to create effective and persuasive visualizations by aligning them with the audience's existing mental models or by challenging them with new or surprising data insights. For example, using familiar metaphors or analogies for data visualizations can help the audience relate to them better, while using contrasting or conflicting data visualizations can help the audience question their assumptions or beliefs. The mental models should be revealed in the stakeholder analysis discussed previously in this chapter.

The data visualization perception theory helps us communicate and present data insights by providing us with strategies and methods based on how humans see and understand visual information. It also helps us optimize and enhance data visualizations by providing us with techniques and tricks based on perceptual illusions and biases.

In practical terms, the data visualization perception theory can help us with the following:

- Choose appropriate visual encodings for different types of data and purposes
- Highlight important or unusual data points or patterns using preattentive features
- Group related data points or patterns using Gestalt principles
- Align or challenge the audience's mental models using familiar or surprising data insights
- Avoid misleading or confusing data visualizations by following perceptual guidelines and best practices.

STORYTELLING

This section focuses on presenting and communicating the data findings using storytelling techniques, such as narrative, emotion, and action. Storytelling has evolved to transmit information that must be retained and acted upon. It's therefore an effective tool to bridge data and decision-making. In the following sections, the following topics will be covered:

- What is data storytelling?
- The CED framework (conclusion, evidence, data)
- How to use emotion and action to engage the audience to act
- Data visualizations that support the narrative
- Data storytelling best practices
- Evaluation of the effectiveness and impact.

As storytelling evolves into data storytelling, special considerations must be made.

Why is it important for data communication?

Data storytelling is the ability to effectively communicate insights from a dataset using narratives and visualizations. It is used to put data insights into a context and inspire action (decision) from your audience (Cote, 2021).

Data storytelling is important for data communication because it helps bring data to life and make it meaningful and accessible for other people than the highly data literate analyst. Data storytelling goes beyond the data itself and uses narratives and often data visualizations to help an audience understand the conclusions of data analysis. Simply put, many people who look at raw data won't understand its significance initially. Data storytelling helps them relate to the content and make critical decisions quickly and more confidently.

Data storytelling also adds value to your data and insights, provides a human touch to your data, offers value to your audience and industry, and builds credibility as an industry and topic thought leader. Data storytelling is a powerful tool that can move a person to act based on your data findings (ProjectPro 2023).

The CED framework (conclusion, evidence, data)

To build a narrative that connects your data insights with your audience's needs and expectations, using the CED framework (conclusion, evidence, data), use the following structure (O'Neill, 2019):

Start with the conclusion: This is the main message or takeaway that you want your audience to remember and act on. It should be clear, concise, and relevant to your audience's goals and challenges. For example, "*Our customer retention rate has increased by 15% in the last quarter thanks to our new loyalty program*." This is also sometimes referred to as the McKinsey style as it's the preferred style for presentations by consultants from the McKinsey consulting group. In an interactive dashboard that would often mean presenting the key KPIs at the top of the screen.

Provide the evidence: This is the data analysis that supports your conclusion and shows how you arrived at it. It should be accurate, complete, and trustworthy. You can use descriptive, diagnostic, predictive, and prescriptive analytics to explain what happened, why it happened, what will happen, and what should happen. For example, "*We analyzed customer behavior data from our CRM system and found that customers who enrolled in our loyalty program were more likely to make repeat purchases, refer new customers, and provide positive feedback*."

Show the data: This is the data visualization that illustrates your evidence and makes it easy for your audience to understand and remember. It should be simple, clear, and persuasive. You can use charts, graphs, diagrams, pictures, or videos to highlight the key points and trends in your data. For example, "*Here is a bar chart that compares the retention rate of customers who enrolled in our loyalty program versus those who did not*."

This is exemplified in Figure 6.2.

By following the CED framework, you can create a data story that connects your data insights with your audience's needs and expectations. You can also use storytelling techniques such as narrative, emotion, and action to make your data story more engaging and impactful. You can find more information and examples of data storytelling using the CED framework in the search results below.

Emotional appeal for driving action

To use emotion and action to engage your audience and inspire them to act based on your data findings you should use words, images, colors, and sounds that appeal to your audience's senses, feelings, and values. For example, you can use positive or negative emotions, such as

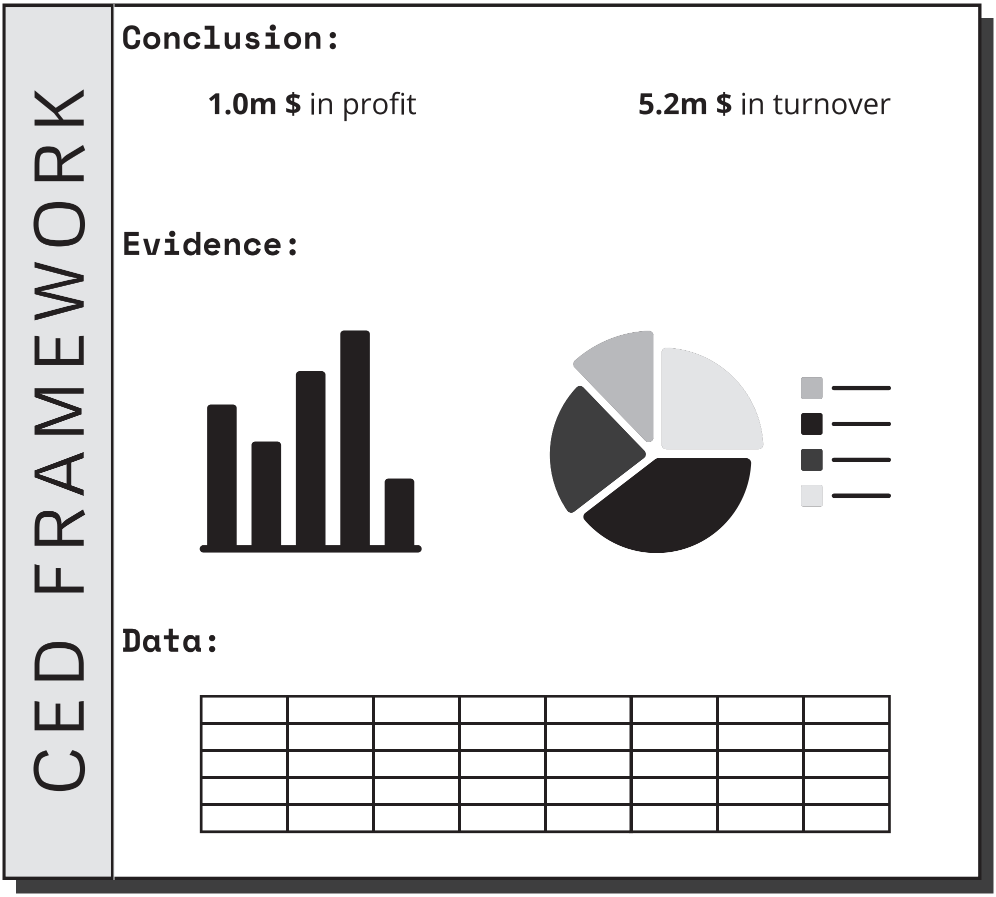


FIGURE 6.2 Conclusion-evidence-data (CED) framework illustrated

joy, surprise, anger, or fear, to capture attention and create a connection with your audience. This will enable what's called the "hook" that captures the attention and ensures that the rest is understood; which to use comes from your previous stakeholder analysis.

Then use stories, examples, analogies, or metaphors to make your data more relatable and memorable. For example, you can use a personal story or a customer testimonial to illustrate how your data findings have impacted someone's (work)life or solved a problem.

Data visualization should then be used to enhance the emotional appeal and lead to action. For example, you can use colors, shapes, sizes, or animations to highlight the most important or surprising data points and create contrast or tension. Be aware that most companies have brand guidelines for colors, fonts, and the like to guide visual presentations. Adhering to them lowers the cognitive load in the interpretation and allows the viewer to focus on the data. This also means that "breaking" the rules will draw attention.

Linking the emotional context and hard data in your data storytelling to influence others will enhance recall of the information later. For example, you can use data to back up your emotional claims or use emotions to emphasize your data insights.

Lastly, it is important to provide a clear call to action for your audience based on your data findings. For example, you can use words like "now", "today", or "immediately" to create a sense of urgency and motivate your audience to act. This is like classic marketing and sales, where nothing happens without the recipient of the communication acts.

By using emotion and action in your data storytelling, you can engage your audience and inspire them to act based on your data findings (Lobel, n.d.).

Data visualizations to support the story

To choose and create appropriate data visualizations that support your narrative and convey your message clearly and persuasively, you should consider the type of data you have and the purpose of your visualization.

The type of data you are presenting should guide your choice of visualization format. For example, if you are working with categorical data, you might use a bar chart, while continuous data might be best displayed with a line chart. Similarly, the purpose of your visualization should determine what kind of pattern or relationship you want to show. For example, if you want to compare values across categories, you might use a bar chart or a pie chart, while if you want to show trends over time, you might use a line chart or an area chart. In the last section of this chapter, the Abela chart chooser will be covered as a guide for choosing the right chart.

Knowing your goal and message is important to consider when adding visualizations to your data story. A goal can be to present information as objectively as possible, but information and messages are not the same things. *A message is the selected set of information to be communicated while information is the set of messages selected by the information source.* You should have a clear goal and message for your visualization and focus on the most relevant and important data points that support it. You should also avoid cluttering your visualization with too much information or unnecessary details that distract from your message (Knaflic, 2015).

Then, using predictable patterns for layouts as often described in company brand manuals will help lift the cognitive load from the viewers. The same is the case if you keep the Gestalt principles in mind. Humans are visual creatures by nature as seen by a significant portion of the brain being dedicated to visual processing. You should therefore use predictable patterns for your visualization layouts, such as placing the most important information at the top left corner, using consistent colors and fonts, aligning elements neatly, and using white space effectively along with the visualization grammar model.

Colors are a powerful tool for data visualization as they can draw attention, create contrast, show patterns, and evoke emotions. However, color can also be misleading, confusing, or distracting if used poorly. You should use color wisely by remembering palette use, color consistency, and accessibility:

Use a color palette that suits your data and message: For example, use sequential colors for showing gradients or ranges of values, diverging colors for showing positive and negative values or differences from a baseline, and categorical colors for showing distinct categories or groups.

Use color sparingly and consistently: For example, use one color for highlighting the most important data point or category, use different shades of the same color for showing subcategories or hierarchies, and use the same color for the same meaning across different visualizations.

Use color accessibility tools: This will ensure that your visualization is readable and understandable by people with color vision deficiencies or other visual impairments. For example, use tools like ColorBrewer or Color Oracle to test your color choices and avoid using colors that are too similar or hard to distinguish.

By following these tips, you can choose and create appropriate data visualizations that support your narrative and convey your message clearly and persuasively.

How to apply data storytelling best practices

To apply data storytelling best practices, such as turning metrics into actionable concepts, eliminating clutter, simplifying and making connections, and relying on the right data tools (Solutions Review 2023), you should:

- Turn metrics into actionable concepts
- Eliminating clutter
- Simplify and make connections
- Rely on the right data tools.

Turning metrics into actionable concepts: Metrics are quantitative measures that describe some aspect of your data, such as sales, revenue, or conversion rate. However, metrics alone are not enough to tell a compelling data story. You need to turn metrics into actionable concepts that explain what they mean, why they matter, and how they can be improved. For example, instead of saying "*Our conversion rate is 10%*", you can say "*We are converting one out of every ten visitors into customers, which is below the industry average of 15%. We need to optimize our landing page and offer more incentives to increase our conversions*." By turning metrics into actionable concepts, you can make your data story more relevant and make decision-making based on data much easier for your audience.

Eliminating clutter: Clutter is anything that distracts from your main message or makes your data story hard to understand. Clutter can include unnecessary details, redundant information, irrelevant data points, or excessive visual elements. In traditional graphics design, you say that you should remove elements until you lose the core meaning. You can eliminate clutter by following these guidelines:

- Focus on the most important and relevant data points and insights that support your message and goal
- Use clear and concise language and avoid jargon or technical terms that your audience may not understand
- Use simple and consistent visualizations that highlight the key points and trends in your data
- Use white space, labels, titles, legends, axes, and annotations to guide your audience through your visualization and explain what they are seeing.

Simplifying and making connections: Simplifying means reducing the complexity of your data story by presenting it logically and coherently. Making connections means linking your data insights with your audience's needs and expectations by providing context and recommendations. You can simplify and make connections by following these steps:

- Start with a clear goal and message for your data story and structure it using the CED framework (conclusion, evidence, data)
- Provide context for your data by explaining where it came from, how and why you gathered it, and what challenges or opportunities it reveals
- Provide recommendations for your audience based on your data findings and explain how they can solve a problem or achieve a goal
- End with a clear call to action for your audience that motivates them to take the next steps.

Relying on the right data tools: Data tools are software applications that help you collect, analyze, visualize, and communicate data. Relying on the right data tools means choosing the ones that suit your purpose, data type, audience, and skill level. For example, you can use tools like Excel or Google Sheets for basic data analysis and visualization; tools like Tableau or Power BI for more advanced data analysis and visualization; tools like Canva or Piktochart for creating infographics or presentations; or tools like Shiny, Datapine, or Toucan Toco for creating interactive dashboards or reports. By relying on the right data tools, you can make your data storytelling process more efficient and effective; especially if sharing data and insight outside the organization is an issue that must be carefully considered due to the potential cybersecurity and privacy issues.

Evaluating the effectiveness and impact

The first step in evaluation is defining your success criteria and key performance indicators (KPIs) for your data story. Depending on your goal and audience, you may want to measure different aspects of your data story, such as reach, engagement, comprehension, retention, persuasion, or action. For example, you can use metrics like views, shares, comments, likes, ratings, surveys, conversions, or revenue to track how well your data story is performing (Cote, 2021). This can be both internally on corporate intranets or externally on social media platforms.

Feedback is essential for improving your data storytelling skills and understanding how your data story is received and perceived by your audience and stakeholders. You should, and can collect feedback using various methods, such as surveys, interviews, focus groups, polls, or mining comments. You can also use tools like Google Analytics or Hotjar to analyze how your audience interacts with your data story online.

Once you have collected feedback and metrics for your data story, you should analyze them and look for patterns, trends, gaps, or outliers. You should also compare your results with your success criteria and KPIs and see if you have met them or not. Based on your analysis, you should identify areas where you can improve your data story or adjust your approach for the next time.

After identifying areas for improvement, you should implement them and test them with a new or existing audience. You should also collect feedback and the same metrics again to see if they have improved or not. You should repeat this process until you are satisfied with your data story's effectiveness and impact. Starting with an internal partner will almost always be a good idea if you feel they can provide trustworthy feedback.

By evaluating the effectiveness and impact of your data storytelling using feedback and metrics, you can ensure that your data story is achieving its purpose and meeting its audience's needs and expectations. You can also learn from your mistakes and successes and improve your data storytelling skills over time.

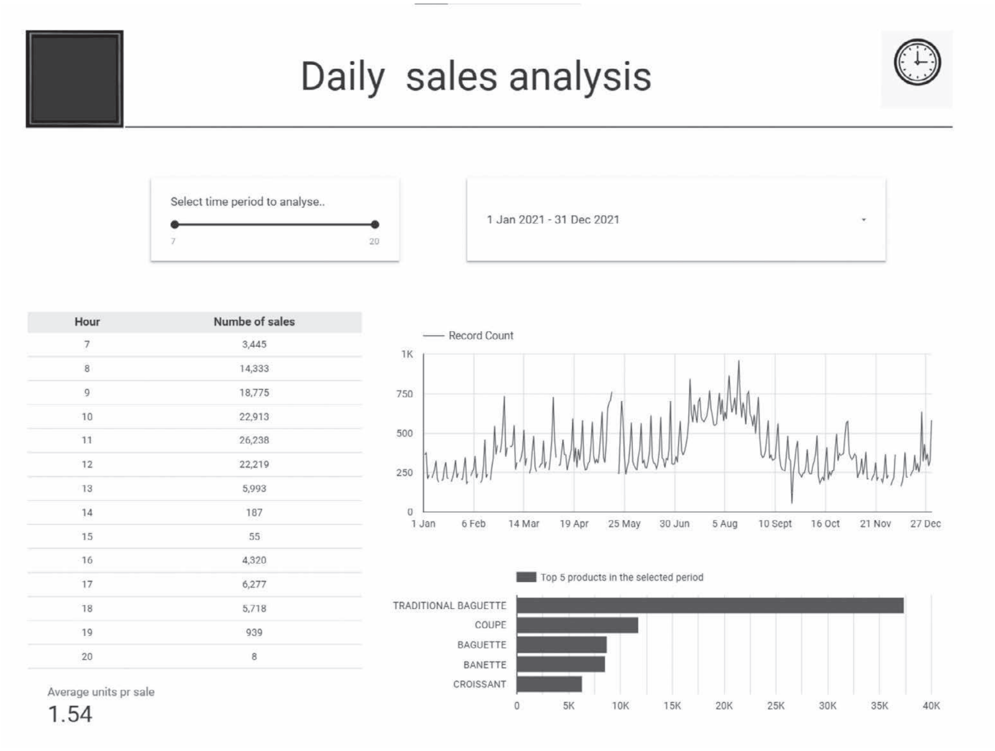
DATA VISUALIZATION FORMATS

We will cover three different formats for visualizations and then dive further into choosing the right charts as what you want to explain often defines what charts you can use.

Reports: Reports are defined as written documents that combine text, tables, charts, and other visual elements to communicate data findings and recommendations. Reports are useful for providing detailed and comprehensive information, analysis, and evidence to support decision-making. Reports can also be formatted and printed for easy distribution and reference. However, reports can also be lengthy, complex, and time-consuming to produce and read. Reports may not capture the attention and interest of the audience, especially if they are not well-structured, well-organized, and well-designed. Reports also rarely allow for interactivity and exploration of the data by the audience. They are, however, good at controlling the storytelling and ensuring that a message is thoroughly covered

Graphics: Graphics are visual representations of data, such as charts, graphs, maps, diagrams, and infographics. Graphics are useful for highlighting key patterns, trends, and insights from the data clearly and concisely. Graphics can also be attractive, engaging, and persuasive for the audience, especially if they use colors, shapes, icons, and images coherently. Graphics can also be easily shared and embedded in different media platforms, such as websites, social media, and presentations. However, graphics can also be misleading, inaccurate, or confusing if they are not well-designed, labeled, and scaled. Graphics may also not provide enough context or explanation for the data or the message behind it. Often the graphics cannot stand alone unless they are in the form of infographics.

Interactive dashboards: Interactive dashboards are web-based applications that display data visualizations dynamically and interactively. Interactive dashboards are useful for allowing the audience to explore and manipulate the data according to their needs and interests. Interactive dashboards can also provide real-time updates and feedback on the data and the performance indicators. Interactive dashboards can also be customized and personalized for different users and scenarios. However, interactive dashboards can also be challenging to create and maintain, requiring technical skills and resources. Interactive dashboards may also overwhelm or distract the audience with too many options or features. Interactive dashboards may also not work well on different devices or browsers or require an internet connection or software installation.



Daily sales analysis

Select time period to analyse..

1 Jan 2021 - 31 Dec 2021

Hour	Numbe of sales
7	3,445
8	14,333
9	18,775
10	22,913
11	26,238
12	22,219
13	5,993
14	187
15	55
16	4,320
17	6,277
18	5,718
19	939
20	8

Average units pr sale
1.54

ILLUSTRATION 6.1 Dashboard example made in Google Looker studio

Choosing the format

Different formats for data visualization can suit different purposes, audiences, and contexts depending on factors such as:

- The goal or objective of the data communication:
  - What is the main message or takeaway that you want to convey to your audience?
  - What action or decision do you want them to make based on your data?
- The type and complexity of the data:
  - What kind of data do you have?
  - How large or diverse is your data set?
  - How many variables or dimensions do you need to show?
- The characteristics and preferences of the audience:
  - Who is your audience?
  - What is their background knowledge and level of expertise on the topic?

- What are their needs and expectations?
- How do they prefer to consume and interact with data?

- The context and medium of the data communication:
  - Where and when will your data communication take place?
  - What format or platform will you use to deliver your data communication?
  - How much time or space do you have?

If your purpose is to provide a *comprehensive overview* of a large and complex data set to a technical audience who needs to make informed decisions based on your analysis, you may want to use a report format that contains detailed text, tables, charts, and other visual elements that explain your methodology, findings, and recommendations.

If your purpose is to highlight a *few key insights or trends* from a simple or small data set to a general audience who needs to be informed or persuaded by your message, you may want to use a graphic format that uses simple charts, graphs, maps, diagrams, or infographics that illustrate your message clearly and concisely.

If your purpose is to *enable a diverse or dynamic audience to explore* and manipulate a real-time or interactive data set according to their interests or needs, you may want to use an interactive dashboard format that displays data visualizations dynamically and interactively that allows the audience to filter, sort, zoom in/out, drill down/up, etc.

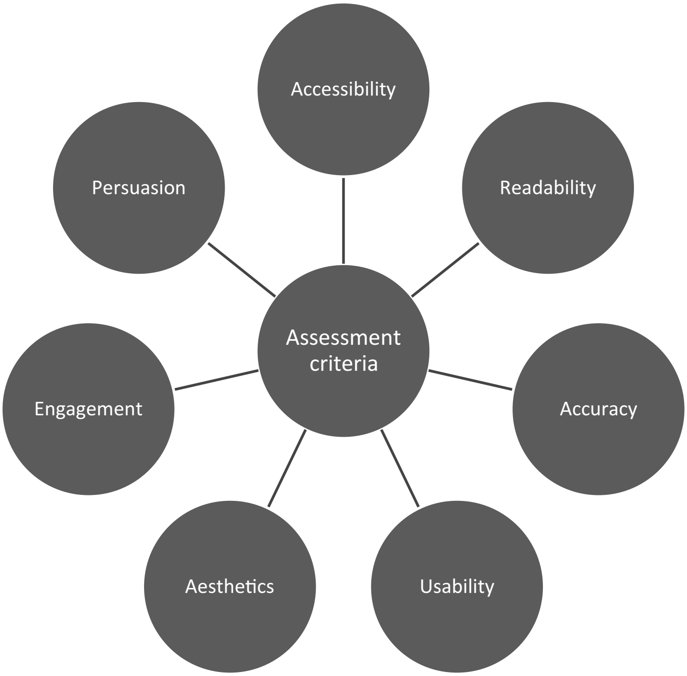


FIGURE 6.3 Data visualization assessment criteria

Criteria and methods for evaluating data visualizations

That's a good question. Here are some potential criteria and methods for evaluating the effectiveness and impact of data visualization in different formats, along with some explanations and comparisons:

Usability refers to how easy and intuitive it is for the audience to use and interact with the data visualization. Usability can be measured by testing the data visualization with real users and observing their behavior, feedback, and satisfaction. Usability can also be assessed by applying heuristics or guidelines for user interface design, such as consistency, visibility, feedback, and error prevention. Usability is important for all formats of data visualization, but especially for interactive dashboards that require user input and manipulation. Tableau (dashboard software developer) has a good library of articles that provides tips and best practices for designing user-friendly data visualizations (Tableau, 2023). It also shows some examples of how to use filters, tooltips, legends, and other features to enhance the usability of your data visualization.

Accessibility refers to how inclusive and adaptable the data visualization is for people with different abilities, needs, and preferences. Accessibility can be measured by testing the data visualization with diverse users and ensuring that it meets the standards and best practices for web accessibility, such as providing alternative text, captions, transcripts, color contrast, keyboard navigation, and screen reader compatibility. Accessibility is important for all formats of data visualization, but especially for graphics and interactive dashboards that rely on visual elements and features.

Readability refers to how clear and understandable the data visualization is for the audience. Readability can be measured by testing the data visualization with different levels of expertise and knowledge on the topic and evaluating their comprehension, recall, and interpretation of the data and the message. Readability can also be assessed by applying principles and techniques for effective communication, such as simplicity, structure, hierarchy, emphasis, and annotation. Readability is important for all formats of data visualization, but especially for reports and graphics that contain text and labels.

Accuracy refers to how correct and reliable the data visualization is in representing the data and the reality. Accuracy can be measured by verifying the data sources, methods, calculations, and assumptions behind the data visualization and ensuring that they are valid, consistent, and transparent. Accuracy can also be assessed by applying principles and techniques for ethical data visualization, such as honesty, integrity, fairness, and accountability. Accuracy is important for all formats of data visualization, as it affects the credibility and trustworthiness of the data communication.

Aesthetics refers to how appealing and attractive the data visualization is for the audience. Aesthetics can be measured by testing the data visualization with different preferences and tastes and evaluating their emotional response, engagement, and enjoyment. Aesthetics can also be assessed by applying principles and techniques for visual design, such as color theory, typography, layout, alignment, balance, etc. Aesthetics is important for all formats of data visualization as it affects the attention and interest of the audience.

Engagement refers to how interactive and immersive the data visualization is for the audience. Engagement can be measured by testing the data visualization with different scenarios and goals and evaluating their behavior, motivation, feedback, etc. Engagement can also be assessed by applying principles and techniques for gamification or storytelling such as challenges, feedback, rewards, narrative, emotion, action, etc. Engagement is important for all formats of data visualization but especially for interactive dashboards that allow user exploration and manipulation.

Persuasion refers to how influential and impactful the data visualization is for the audience. Persuasion can be measured by testing the data visualization with different outcomes or decisions and evaluating their attitude change, behavior change, or action taking. Persuasion can also be assessed by applying principles and techniques for rhetoric or psychology such as ethos, logos, pathos, framing, anchoring, priming, etc. Persuasion is important for all formats of data visualization but especially for graphics or reports that aim to inform or persuade the audience.

These are some of the potential criteria and methods for evaluating the effectiveness and impact of data visualization in different formats. They are not mutually exclusive or exhaustive but rather complementary or overlapping. Depending on the purpose, audience, context, and format of your data communication, you may want to prioritize or combine some of them over others.

Choosing the right chart

The chart chooser is a decision tree model developed by Andrew Abela to guide selection of the most appropriate visualization type based on the goal, data characteristics, and audience. The chart chooser is a decision tree model developed to help people select the most appropriate type of data visualization based on the data being presented and the goal of the visualization. It was first introduced in his book *Advanced Presentations by Design* in 2008 (Abela, 2008).

The model starts by asking what the purpose of the visualization is – to show composition, distribution, relationships, or to make comparisons. Then it branches out based on the data available for the visualization.

Comparison

Comparing static data is primarily done in the various forms of bar charts. Which to choose depends again on the variables, and more specifically on the number of variables to compare. The simplest is a single variable in a bar chart that's either horizontally or vertically placed.

Small multiples display subset time trends side-by-side for easy comparison but can quickly become confusing and difficult to interpret. In those cases, colors can also be a way to split variables.

For comparisons of data that change over time, line charts are great for highlighting overall patterns and trends. But they can obscure individual data points. In those cases, a column or bar chart can be used. If you are presenting data that is cyclical or seasonal in nature you should consider a circular line or area chart. Be aware that the historical data must have a certain volume or otherwise you are better served by adding a "last year data (LY)".

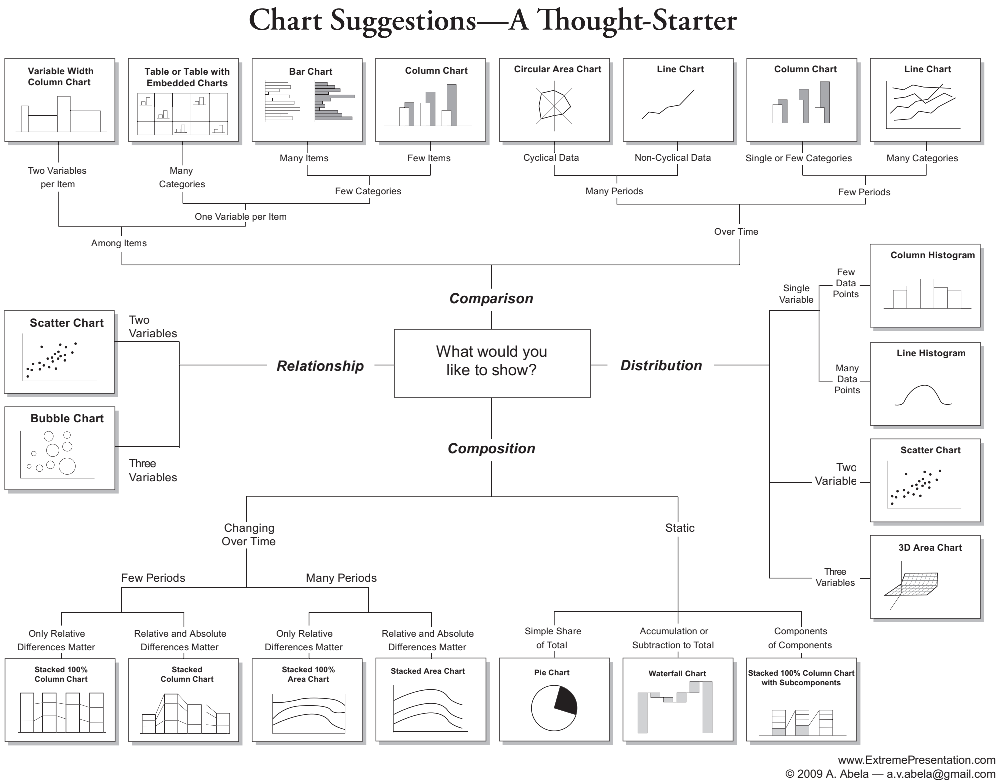


FIGURE 6.4 The Abela chart choosing framework © 2024 Dr. Andrew V. Abela, PhD. www.ExtremePresentation.com

Distribution

When the question of how the data is distributed or "is there links between groups and how many are there", the first question to be asked is how many variables that should be considered. Bar or bullet charts better emphasize individual values if you only have a few observations. With many observations, you will get a line histogram, but a word of warning should be made here. If the data is not continuous in nature the line might lead to false interpretations like "each car salesman must sell 2.6 cars per week".

The three variables displayed in a 3D space are difficult for many people to interpret, and you should carefully consider whether the visualization could be shown in 2D, but with an added color grading for the third variable.

Composition

Like with the comparisons, the first branch in composition visualization asks whether there is a time dimension in the data.

If the answer is yes, the stacked bar charts allow comparing multiple time series simultaneously. Again, the stacked charts come in two versions that are either scaled to 100% or shown in absolute values. The question posed here is whether there is a layer of information in the absolute values or the comparisons that you wish to draw out. What is most important? How many products do we sell in absolute terms or how are our sales composed of products A, B, C, etc.?

If we are looking at static data or a snapshot in time, we are looking at the classic pie chart primarily. The usage of the pie chart is, however, controversial as people in general have difficulty comparing angles. A good rule of thumb is therefore to limit the shares to a maximum of 4–5.

The waterfall chart is often used in financial reporting to show how you get from one value to another by adding and subtracting components. That could e.g., be to show how we go from last year's budget to this year's budget.

The components of components cart is often used when visualizing cost components. The system costs 100K USD, of which 20K USD is labor. The 20K labor is then split into salary, benefits, and pension. The pension is then again split into ….

With categorical data, heatmaps show the intensity of association between categories. Mosaic plots illustrate multi-way relationships. These are special versions that are not originally part of the Abela chart chooser and are not necessarily available in your visualization tool of choice.

Relationships

For revealing relationships, again start with figuring out how many variables you need to consider. The basic two-variable relationship is shown in the classic scatterplot. As humans are good at pattern recognition as described previously in this chapter, relationships are often readily visible. With more variables, you need to add graphical elements to the chart. The first variable to use is the size of the elements and then create a bubble chart. Then, adding color or shape can be used to illustrate additional dimensions in the data. An example could be combining the size of a bubble to illustrate turnover and the color to indicate whether it matches the expected budget.

Common sources of error and misinterpretation include:

- Axes scaling – inconsistent ranges can distort trends and relationships
- Using pie charts for comparison – hard to visually compare angles and area
- Lack of context or reference points – no baseline or goal lines to give meaning
- Correlation vs causation – scattering plots reveal correlation but not necessarily causality
- Sampling bias – non-random sampling can lead to misleading conclusions
- Audience bias – interpreting based on preconceptions vs evidence.

Thoughtfully choosing visual encodings like shape and color to match data types helps the audience interpret relationships accurately. Interactivity allows drilling down into details. Animation can clarify changes over time. The key is selecting visualizations matched to the specific data structure, relationships, and audience goals.

The chart chooser guides the analysts and designers through a structured decision process considering these key factors. Using this model as a tool can lead to more intentional, effective visualizations that intuitively communicate insights, tell a story, and avoid common pitfalls. Overall, it encourages visualization choices tailored to the data specifics and audience - leading to impactful data graphics that inform rather than mislead.

CONCLUSION

Effectively communicating data-driven insights is an indispensable capability in an increasingly competitive data economy. The way we collect, analyze, interpret, and present data can have profound implications for strategic decision-making across all facets of an organization. This chapter has provided an in-depth exploration of frameworks, models, and techniques to help convey data perspectives persuasively to diverse audiences.

The chapter emphasizes the importance of stakeholder and context analysis before crafting any data communication strategy. Models like the power–interest grid and influence–impact grid help categorize stakeholders based on their authority and involvement in the project outcomes. Conducting detailed stakeholder mapping illuminates priorities, alliances, conflicts, and expectations that must be considered when presenting data findings and recommendations. Beyond stakeholders, visual context models define the boundaries, flows of information, dependencies, and relationships between business systems and processes. These crucial first steps ensure data insights resonate with stakeholder needs and operational realities.

With audiences and settings clearly defined, data visualization emerges as a potent mechanism for storytelling. Whether highlighting key trends in performance dashboards, summarizing market dynamics in an investor presentation, or reporting weekly sales figures, impactful visualization transforms raw data into a compelling narrative. The frameworks covered help match visual design choices to data types and audience profiles. The data visualization grammar provides a modular toolkit to construct graphics systematically, while visualization classification schemes help select appropriate visual metaphors. Perceptual and cognitive theories shed light on how viewers process charts, gauge effectiveness, and extract meaning.

The CED structure (conclusion, evidence, data) lends logical flow in revealing data insights. Audience-centric narrative techniques, grounded in behavior science, boost engagement and motivation to act. Eliminating clutter, establishing a clear information hierarchy, and adhering to aesthetic design principles increase memorability and trust in the data artifacts presented. Guidelines help avoid misapplication and manipulation when portraying statistical findings.

With the exponential growth in data volume and analytical sophistication, fluency in "data storytelling" will serve aspiring leaders well. Evaluating communication outcomes using key performance indicators and user feedback provides crucial input for continuous refinement and improvement. Whether designing a real-time operational dashboard, an employee engagement survey report, a predictive sales forecast model, or even a machine learning recommendation system, conveying data narratives effectively to shape decisions will prove a key leadership skill for tomorrow's data-centric business landscape.

KEY TERMS

Abela chart chooser: A decision tree model developed by Andrew Abela to guide the selection of the most appropriate visualization type based on the goal, data characteristics, and audience.

CED framework: A structured approach to data storytelling that starts with the conclusion, provides evidence, and then shows the data.

Context model: A visual tool that shows the relationship between a system (such as a project or product) and its environment (users, customers, suppliers, etc.).

Data storytelling: The ability to effectively communicate insights from data using narratives and visualizations to inspire action or decision-making.

Data visualization grammar: A framework that breaks down visualizations into discrete components, such as data, aesthetics, and geometry, allowing for a structured approach to creating graphics.

Data visualization: The representation and presentation of data in a visual format, such as charts, graphs, or diagrams, to aid in understanding and communication.

Gestalt principles: A set of rules that describe how the human visual system organizes and groups visual elements into meaningful wholes. Can both distract and focus attention to certain data points.

Graphics: Visual representations of data, such as charts, graphs, maps, diagrams, and infographics.

Interactive dashboards: Web-based applications that display data visualizations dynamically and interactively, allowing users to explore and manipulate the data.

Mental models: Internal representations of how things work or behave in the real world, which can influence how individuals perceive and interpret data visualizations.

Power–interest grid: A model that categorizes stakeholders based on their power (authority over the project) and their interest (level of concern for the project's success).

Pre-attentive processing: The ability of the human visual system to detect certain basic features of a scene in a very short time, without conscious attention.

Reports: Written documents that combine text, tables, charts, and other visual elements to communicate data findings and recommendations.

Stakeholder analysis: The process of identifying and analyzing the individuals or groups who have an interest or influence in a data project or data-driven decision.

Usability: The ease and intuitiveness with which an audience can use and interact with a data visualization.

DISCUSSION QUESTIONS

- What are some of the benefits and challenges of using data visualization for decision-making?
- How would you approach stakeholder and context analysis for your data communication? What are some of the factors and criteria that you would consider?
- What are some of the principles and best practices of data visualization design? How can you apply them to your data visualizations?
- How can you use storytelling techniques to present and communicate your data findings? What are some of the elements and strategies that you would use to create a compelling story with data?
- How can you use different formats of data visualization, such as reports, graphics, and interactive dashboards, to suit different purposes and audiences? What are some of the advantages and disadvantages of each format?

Potential answers for the discussion questions

Some of the benefits of using data visualization for decision-making are that it:

- can help simplify complex data
- reveals patterns and trends
- highlights key insights and messages
- engages and persuades the audience, and
- facilitates action.

Some of the challenges of using data visualization for decision-making are that it:

- can be time-consuming and resource-intensive
- requires technical and design skills
- involves ethical and privacy issues, and
- is subject to bias and misinterpretation.

To approach stakeholder and context analysis for data communication, I would first identify who are the primary and secondary stakeholders of my data communication, and what are their roles, interests, expectations, and preferences. Then I would analyze the context of my data communication, such as the purpose, scope, format, medium, and timing. Based on these factors and criteria, I would tailor my data communication to suit the needs and expectations of my stakeholders and context.

Some of the principles and best practices of data visualization design are: choose the right type of chart or graph that matches your data and message, use appropriate colors, shapes, sizes, labels, and legends to enhance clarity and readability, avoid clutter and unnecessary elements that distract from the main message, use contrast, alignment, hierarchy, and balance to create visual harmony and emphasis, and use annotations, captions, titles, and narratives to provide context and explanation.

To use storytelling techniques to present and communicate my data findings, I would use the following elements and strategies: create a narrative structure that has a beginning (introduction), a middle (body), and an end (conclusion), use emotion to connect with the audience and elicit their interest and empathy, use an action to show how the data findings relate to the problem or goal that I am trying to solve or achieve, use characters to personalize the data findings and make them more relatable and memorable, use visuals to support and enhance the verbal message and create a more engaging experience.

To use different formats of data visualization to suit different purposes and audiences, I would consider the following advantages and disadvantages of each format: reports are good for providing detailed information and analysis in a structured and formal way, but they can be lengthy and boring to read; graphics are good for highlighting key insights and messages concisely and attractively, but they can be oversimplified or misleading; interactive dashboards are good for allowing the audience to explore the data in a dynamic and customized way, but they can be complex and confusing to use.

ADDITIONAL RESOURCES

How to Master Design Storytelling and Data Visualization: This article provides tips and tools for creating effective data visualizations and stories using software such as Sketch, Figma, Adobe XD, Tableau, Power BI, Google Data Studio, Canva, Piktochart, and Infogram. LinkedIn. www.linkedin.com/advice/0/how-do-you-develop-your-skills-knowledge-design

Ultimate Guide to Using Data Visualization in Your Presentation: This guide covers the basics of data visualization, such as what it is, why it matters, how to choose the right type of chart or graph, how to design and format your data visualizations, and how to use them in PowerPoint presentations. 24Slides. https://24slides.com/presentbetter/the-ultimate-guide-to-using-data-visualization-in-your-presentation/

Data Presentation Guide – Best Visuals, Charts, and Storytelling: This guide outlines the key objectives of data presentation, such as visual communication, audience, and context analysis, focus on important points, design principles, storytelling, persuasiveness, and dashboards. It also provides examples of good and bad charts and graphs and how to avoid common pitfalls. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/business-intelligence/data-presentation-guide/

Golden Rules for Creating a Data Visualization PowerPoint: This article provides some golden rules for creating a data visualization PowerPoint presentation, such as knowing your audience, choosing the right visuals, using colors wisely, keeping it simple and clear, and adding interactivity and animation. 24Slides. https://24slides.com/presentbetter/data-visualization-powerpoint/

What Is Data Visualization? | Microsoft Power BI: This webpage introduces the concept and benefits of data visualization and how it can help users develop powerful business insights quickly and effectively. It also showcases some examples of data visualization using Microsoft Power BI. Microsoft Power BI. https://powerbi.microsoft.com/en-us/data-visualization/

BIBLIOGRAPHY

24Slides. (n.d.). *Ultimate guide to using data visualization in your presentation*. https://24slides.com/presentbetter/the-ultimate-guide-to-using-data-visualization-in-your-presentation/ (Accessed March 30, 2023).

Abela, A. (2008). *Advanced presentations by design: Creating communication that drives action*. Pfeiffer.

Arcadis Gen. (n.d.). *Using data driven insights to better inform stakeholder communication*. www.arcadisgen.com/en/insights/data-and-stakeholder-management (Accessed March 31, 2023).

Bing. (n.d.). *Using data driven insights to better inform stakeholder communication*. https://bing.com/search?q=stakeholder+management+data-driven+decision+making (Accessed March 31, 2023).

BrightHub PM. (n.d.). *Using a power/influence grid (power/influence matrix) in stakeholder*. www.brighthubpm.com/resource-management/81140-what-is-the-power-influence-grid-or-matrix/ (Accessed March 30, 2023).

CIO. (n.d.). *What is data visualization? Presenting data for decision-making*. www.cio.com/article/191640/what-is-data-visualization-presenting-data-for-decision-making.html (Accessed March 31, 2023).

Cote, C. (2021, November 23). *Data storytelling: How to tell a story with data – Business Insights blog. Harvard Business School Online*. https://online.hbs.edu/blog/post/data-storytelling (Accessed March 31, 2023).

Donska, T. (2020). *Exploring data visualization psychology*. Toptal. www.toptal.com/designers/data-visualization/data-visualization-psychology (Accessed March 31, 2023).

Forbes Advisor. (n.d.). *What is a stakeholder analysis? Everything you need to know*. www.forbes.com/advisor/business/what-is-stakeholder-analysis/ (Accessed March 30, 2023).

Hevo Data. (n.d.). *Data modeling and visualization: A detailed comparison*. https://hevodata.com/learn/data-modeling-and-visualization/ (Accessed March 31, 2023).

IBM. (n.d.). *What is data visualization?*. www.ibm.com/topics/data-visualization (Accessed March 31, 2023).

Improvement Service. (n.d.). *Power/interest grid*. www.improvementservice.org.uk/business-analysis-framework/consider-perspectives/powerinterest-grid (Accessed March 30, 2023).

Indeed. (n.d.). *15 types of stakeholder analysis (and why it's important)*. www.indeed.com/career-advice/career-development/types-of-stakeholder-analysis (Accessed March 31, 2023).

Indeed. (n.d.). *What is a power interest grid and how do you use one?*. https://uk.indeed.com/career-advice/career-development/power-interest-grid (Accessed March 31, 2023).

JanBask Training. (n.d.). *How can we use Power/Influence Grid (Power/Influence Matrix) in stakeholder prioritization?*. www.janbasktraining.com/community/business-analysis/how-can-we-use-powerinfluence-grid-powerinfluence-matrix-in-stakeholder-prioritization (Accessed March 30, 2023).

Knaflic, C. N. (2015). *Storytelling with data: A data visualization guide for business professionals*. New York: Wiley.

KnowledgeBurrow. (n.d.). *What is a power influence grid?*. https://knowledgeburrow.com/what-is-a-power-influence-grid/ (Accessed March 30, 2023).

KnowledgeHut. (n.d.). Power interest grid: How to use, benefits, examples. www.knowledgehut.com/blog/project-management/power-interest-grid(Accessed March 30, 2023).

Lobel, G. (n.d.). *Data storytelling: linking emotions and rational decisions*. Toucan Toco. www.toucantoco.com/en/blog/data-storytelling-dataviz (Accessed March 31, 2023).

Medium: Business Architected. (n.d.). *How to create context models that stakeholders understand*. https://medium.com/business-architected/how-to-create-context-models-that-stakeholders-understand-bd337d3a35f2 (Accessed March 30, 2023).

More Than Digital. (n.d.). *Data-driven decision-making explained: Making smarter business*. https://morethandigital.info/en/data-driven-decision-making-explained/ (Accessed March 31, 2023).

Northeastern University. (n.d.). *Data-driven decision making: A primer for beginners*. www.northeastern.edu/graduate/blog/data-driven-decision-making/ (Accessed March 31, 2023).

ODI: Think change. (n.d.). *Mapping political context: Stakeholder analysis*. https://odi.org/en/publications/mapping-political-context-stakeholder-analysis/ (Accessed March 30, 2023).

O'Neill, B. T. (2019, April 17). *CED: A UX framework for designing analytics tools that drive decision making*. *Designing for Analytics*. https://designingforanalytics.com/resources/c-e-d-ux-framework-for-advanced-analytics/

Project Management. (n.d.). *Stakeholder analysis using the power interest grid*. www.projectmanagement.com/wikis/368897/Stakeholder-Analysis–using-the-Power-Interest-Grid (Accessed March 30, 2023).

Project Management. (n.d.). *Stakeholder analysis using the power interest grid*. www.projectmanagement.com/wikis/368897/Stakeholder-Analysis–using-the-Power-Interest-Grid

ProjectPro. (2023, October 12). *The Importance of data storytelling*. JCU Online. https://online.jcu.edu.au/blog/the-importance-of-data-storytelling (Accessed March 31, 2023).

Sarkar, D. (2018, September 12). A comprehensive guide to the grammar of graphics for effective visualization of multi-dimensional data. *Towards Data Science*. https://towardsdatascience.com/a-comprehensive-guide-to-the-grammar-of-graphics-for-effective-visualization-of-multi-dimensional-1f92b4ed4149

SAS. (n.d.). *Data visualization: What it is and why it matters*. www.sas.com/en_us/insights/big-data/data-visualization.html (Accessed March 30, 2023)

Solutions Review. (2023). *4 data visualization best practices through data storytelling*. https://solutionsreview.com/business-intelligence/data-visualization-best-practices-through-data-storytelling/ (Accessed March 31, 2023).

Substack. (n.d.). *What would a theory of data visualization look like?*. https://filwd.substack.com/p/theory-in-vis (Accessed March 31, 2023).

Tableau. (2023). *What is data visualization? Definition & examples*. www.tableau.com/learn/articles/data-visualization (Accessed March 31, 2023).

Tableau. (n.d.). *What is data visualization? Definition & examples*. www.tableau.com/learn/articles/data-visualization (Accessed March 31, 2023).

Unzueta, D. (n.d.). *Data visualization theory: An introduction*. Towards Data Science. https://towardsdatascience.com/data-visualization-theory-an-introduction-a077c0d80498 (Accessed March 31, 2023).

Why Change. (n.d.). *A context model in 5 minutes*. https://why-change.com/2021/02/09/a-context-model-in-5-minutes/ (Accessed March 31, 2023).

Wikipedia. (n.d.). *Context model*. https://en.wikipedia.org/wiki/Context_model (Accessed March 31, 2023).

CHAPTER 7

Data Analysis

Understand How Descriptive, Predictive, and Prescriptive Analytics Can Support the Organizational Decision Processes

This chapter covers how to use different types of analytics to answer different types of questions and support different types of decisions. It also introduces analytical methods and tools that can facilitate interaction with data scientists and gain enough understanding of the different techniques to choose the best action forward.

LEARNING GOALS:

L7.1 Use descriptive analytics to summarize and explore your data
L7.2 Use predictive analytics to forecast and anticipate future outcomes or behaviors
L7.3 Use prescriptive analytics to optimize and recommend actions or solutions

The chapter case is about MAERSK, Logistics, Vietnam, where 2018 saw big shifts both internally and externally affecting how the business should be composed going forward.

Descriptive, predictive, and prescriptive are the three levels of analytics techniques that form the base of data-driven decision-making. This is associated with project management methods relevant to each level of analytics. Most organizations approach their analytics in either a “ticket” or project format depending on the organizational structure and data usage maturity.

DESCRIPTIVE ANALYTICS

Descriptive analytics is the classical form of data analysis and is even to some extent bound by legislation through financial reporting. The company financial report is a fundamental descriptive analysis. As soon as we are required to look into other parts of the business, we mostly use the same analytical techniques as in financial reporting. Some of the most common are described in the following section.

DOI: 10.4324/9781003457787-7

Methods

Descriptive analysis is the process of summarizing, visualizing, and exploring the data to understand its main characteristics and patterns. It is an important step in the data analysis process, as it helps to gain insights, identify problems, and prepare the data for further analysis.

Before the full analysis can start, the data to be analyzed must be prepared. This happens in three steps as shown in Figure 7.1:

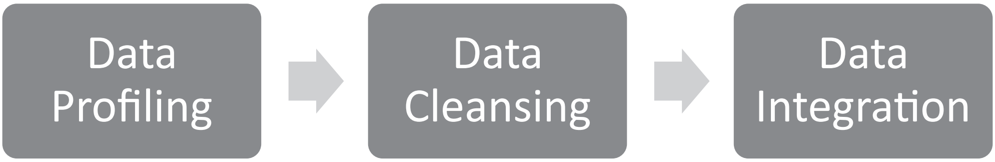


FIGURE 7.1 Preparing for data analysis

Data profiling is the process of examining the data and collecting metadata about its structure, content, and quality. Data profiling helps to understand the data types, formats, ranges, distributions, frequencies, uniqueness, completeness, accuracy, and consistency of the data. Data profiling also helps to identify errors, anomalies, outliers, missing values, duplicates, and inconsistencies in the data. Data profiling can be done using various tools and techniques, such as summary tables, frequency tables, cross tables, data dictionaries, data quality rules, and data quality reports.

Data cleansing is the process of correcting or removing errors, anomalies, outliers, missing values, duplicates, and inconsistencies from the data. However, it is important to note here that a version of the original data without any manipulation should be retained. What might seem like an error initially might be useful for experiments later. Data cleansing helps to improve the data quality and reliability and makes it suitable for further analysis. Data cleansing can be done using various tools and techniques, such as validation rules, transformation rules, standardization rules, matching rules, deduplication rules, imputation methods, outlier detection methods, and cleansing reports.

Four classic ways of analyzing data and making it ready for further analysis are:

Outlier detection: Outlier detection is the process of identifying and handling extreme or abnormal values in the data that deviate significantly from the rest of the data. Outlier detection helps to improve the data quality and accuracy and to avoid misleading or erroneous analysis results. Outlier detection can be done using various tools and techniques, such as boxplots, z-scores, interquartile range, standard deviation, distance-based methods, density-based methods, clustering-based methods, and isolation forest.

Normalization: Normalization is the process of scaling the values of a variable in the data to a specific range, such as [0, 1] or [–1, 1]. Normalization helps to make the data more comparable and consistent and to reduce the effect of different scales or units on the analysis results. Normalization can be done using various tools and techniques, such as min–max normalization, decimal scaling normalization, and softmax normalization.

Standardization: Standardization is the process of scaling the values of a variable in the data to have a mean of zero and a standard deviation of one. Standardization helps to make the data more compatible and homogeneous and to reduce the effect of outliers or skewness on the analysis results. Standardization can be done using various tools and techniques, such as z-score standardization, mean normalization, and unit vector standardization.

Imputation: Imputation is the process of replacing missing values in the data with plausible values that can preserve the statistical properties and relationships of the data. Imputation helps to handle incomplete data and to avoid losing information or biasing the analysis results. Imputation can be done using various tools and techniques, such as deletion methods, mean imputation, median imputation, mode imputation, constant imputation, random imputation, regression imputation, k-nearest neighbors' imputation, expectation–maximization imputation, and multiple imputation. Recently systems that generate synthetic data have been coming online using a combination of the imputation methods.

Then in the end you do data integration:

Data integration is the process of combining data from different sources and formats into a unified and consistent view. Data integration helps to enrich the data with additional information and attributes and makes it more comprehensive and complete for further analysis. Data integration can be done using various tools and techniques, such as extraction–transformation–loading (ETL) tools, data warehouses, data marts, data lakes, data pipelines, schema matching, record linkage, entity resolution, and data fusion. These tools are often the domain of the data engineer and are discussed in further detail in Chapter 8.

Descriptive statistical tools

Descriptive statistics are numerical measures that describe the main features of the data, such as its central tendency, variability, shape, and association. Descriptive statistics help summarize the data concisely and meaningfully and provide a basis for further analysis. Descriptive statistics can be calculated using various tools and techniques, such as mean, median, mode, standard deviation, variance, range, interquartile range, skewness, kurtosis, covariance, correlation coefficient, etc. It is recommended to pick up the statistics book if you are unsure about what each term covers. A few are also covered in the key terms section of this chapter.

Histograms: Histograms are graphical elements that show the frequency distribution of a single variable in the data. Histograms help to visualize the shape, spread, center, and outliers of the data. Histograms can convey information, such as bins – grouping of data, frequencies – how often something occurs, relative frequencies – relative occurrence, cumulative frequencies – the previously summarized, density – how close observations are, which also comes in the same variations as frequencies. Class (type groups) information as class intervals, boundaries, marks, and widths. Normal curves (normal distribution) can also be indicated by histograms with many observations.

Boxplots: Boxplots are graphical displays that show the five-number summary of a single variable or a comparison of multiple variables in the data. The five-number summary consists of the minimum value (Min), the first quartile value (Q1), the median value (Md), the third quartile value (Q3), and the maximum value (Max). Boxplots help visualize the data's center, spread, outliers, and symmetry. Boxplots often provide information about the data with information such as whiskers, interquartile range, fences, outliers, means, standard deviations, and confidence intervals visually provided.

Scatterplots: Scatterplots are visualizations that show the relationship between two variables in the data. Scatterplots help to visualize the direction, strength, form, and outliers of the association. Scatterplots are a foundational tool for visualizing the relationship between two continuous variables. Scatterplots show how changes in one variable are associated with changes in the other variable. The pattern of the points reveals the type of relationship (linear, non-linear, positive, negative, etc.). Different groups can be plotted in different colors/symbols to visually compare their distributions.

Correlation analysis: Correlation analysis is the process of measuring and testing the strength and direction of the linear relationship between two variables in the data. Correlation analysis helps to quantify the degree of association and to identify potential causal factors or predictors. Correlation analysis can be used to validate what is hypothesized in the scatterplot or other visualization, or screen larger data sets that don't have obvious relations.

As descriptive analysis is often the precursor of more advanced analysis, some of the processes that lead in that direction are also covered here. Working with features and dimensions will prepare for your regression analysis.

Feature selection: Feature selection is the process of selecting a subset of relevant and useful features or variables from the data that can improve the performance and efficiency of the modeling techniques. Feature selection helps to reduce the dimensionality, complexity, and noise of the data and to avoid overfitting and multicollinearity problems. Feature selection can be done using various tools and techniques, such as filter methods (FM), wrapper methods (WM), embedded methods (EM), ranking methods (RM), scoring methods (SM), information criteria (IC), and feature importance measures (FIM).

Feature engineering: Feature engineering is the process of creating new features or variables from existing ones or from external sources that can enhance the predictive power and interpretability of the modeling techniques. Feature engineering helps to capture more information, insights, and patterns from the data and to represent it in a more suitable and meaningful way for further analysis. Feature engineering can be done using various tools and techniques, such as transformation methods™, encoding methods (EM), discretization methods (DM), aggregation methods (AM), interaction methods (IM), domain knowledge methods (DKM), and text mining methods (TMM).

Among the feature engineering methods are One-hot encoding and Binning:

One-hot encoding: One-hot encoding is the process of transforming categorical variables in the data into binary variables (0 or 1) that indicate the presence or absence of each category. One-hot encoding helps to make the data more suitable and efficient for certain modeling techniques, such as logistic regression, neural networks, and support vector machines. One-hot encoding covers techniques such as dummy variables, indicator variables, and sparse matrices. Let's say you have a dataset with a categorical feature "Color" that takes on the values "Red", "Blue", and "Green". If you use this feature as in a regression model, the model might incorrectly interpret it as an ordinal variable (i.e., Red < Blue < Green). To avoid this, you can use one-hot encoding to create three new binary features: "Is_Red", "Is_Blue", and "Is_Green". Each of these features would take on a value of 1 or 0, indicating the presence or absence of the corresponding color. This way, the regression model can correctly interpret the categorical data.

Binning: Binning is the process of grouping the values of a continuous variable in the data into discrete intervals or bins. Binning helps to reduce the noise and complexity of the data and to capture the underlying patterns or trends of the data. Binning can be done using various tools and techniques, such as equal-width binning, equal-frequency binning, quantile binning, k-means binning, entropy-based binning, and decision tree-based binning. Suppose you have a continuous feature "Age" in your dataset. If you include "Age" as a raw feature in your regression model, the model might be too sensitive to slight changes in age, which could lead to overfitting. To mitigate this, you can use binning to create a new categorical feature "Age_Group" with categories like "Child", "Teenager", "Adult", and "Senior". This can make the model more robust and easier to interpret.

Dimensionality reduction: Dimensionality reduction is the process of reducing the number of features or variables in the data while preserving as much information as possible. Dimensionality reduction helps to improve the performance and efficiency of the modeling techniques and to avoid overfitting (good in test, bad in real life) and multicollinearity problems. Multicollinearity is where several variables suggest the same and you therefore cannot tell what the cause for an effect is.

Regression Analysis: There are several types of regression analysis, including linear regression, nonlinear regression, logistic regression, polynomial regression, and more. The choice of regression type depends on the nature of your response variable (continuous, binary, count, etc.) and the relationship between the predictors and the response variable.

Linear regression is a statistical model used to understand the relationship between two or more variables. It estimates the linear relationship between a dependent variable (also known as the response or outcome variable) and one or more independent variables (in descriptive statistics referred to as explanatory variables). Linear regression is typically used when your dependent variable is continuous (James et al., 2021).

Logistic regression is a statistical model used for binary classification problems. It estimates the probability of an event occurring (such as pass/fail, win/lose, alive/dead) based on a given set of predictors. Unlike linear regression, the dependent variable in logistic regression is binary. Logistic regression is primarily used as a predictive methodology but can also be used for grouping in descriptive statistics (James et al., 2021).

The list might seem overwhelming at first, but it is terms and concepts that everyone who wants to understand the data analysis process needs to have an understanding of. The rule of thumb is that 80% of the time an analyst spends on data work is on the above topic. That's also why analysis for decision-making isn't a quick thing initially. The data needs to be understood and prepared.

TICKET FUNCTION

A ticket function is to get queries from the business function and provide reports. It is a process or system that allows data analysts or BI professionals to handle requests, issues, or tasks related to data analysis or reporting. A ticket function can help organize, prioritize, track, and communicate the status of data-related work. It can also provide insights into the performance, efficiency, and quality of the data analysis/BI department.

Often the analysts associated with an apartment that runs a ticket function are closely linked to the IT department, where tickets have been a part of the process for years with tools like Jira and Zendesk.

The tasks that can be handled efficiently in a ticket system must be well structured and "self-contained" by nature. That is why it is mostly in the realm of descriptive analytics and for companies where the data foundation is reasonably mature. Therefore, it would often consist of reports and ad hoc analysis primarily within financial or other similarly structured data.

If the data is in place, a ticket system can help by reducing the amount of potential friction between the data department and the rest of the organization by having clear service level agreements (SLAs) that manage expectations and highlight resource allocations, priorities, and performance. The ticket system will also naturally generate process data and ensure flows. This data can then again be shown to stakeholders for decisions on the size and organization of the BI/Analytics function.

Tickets are usually placed in a central queue and then distributed to the available analysts. This also indicates that there are several analysts in a centrally located function.

As soon as the queries become more complex a project methodology must be put in place. CRISP-DM has been a standard for a long time.

CRISP-DM

CRISP-DM stands for *Cross-Industry Standard Process for Data Mining*. It is a widely used and well-established methodology for data analysis projects that aims to provide a structured and systematic approach to the entire data mining process. Data mining is the explorative and modeling part of data analysis that encompasses decision-making as well.

CRISP-DM consists of six main phases:

- Business understanding
- Data understanding
- Data preparation
- Modelling
- Evaluation, and
- Deployment.

Each phase has several tasks and outcomes that guide the data analyst from defining the problem to delivering the solution (IBM, 2024).

The CRISP-DM methodology is designed to be flexible and adaptable to different types of data, domains, and objectives. It is not a rigid set of rules, but rather a framework that can be customized and refined according to the specific needs and characteristics of each project. The CRISP-DM methodology also encourages iteration and feedback loops between the phases,

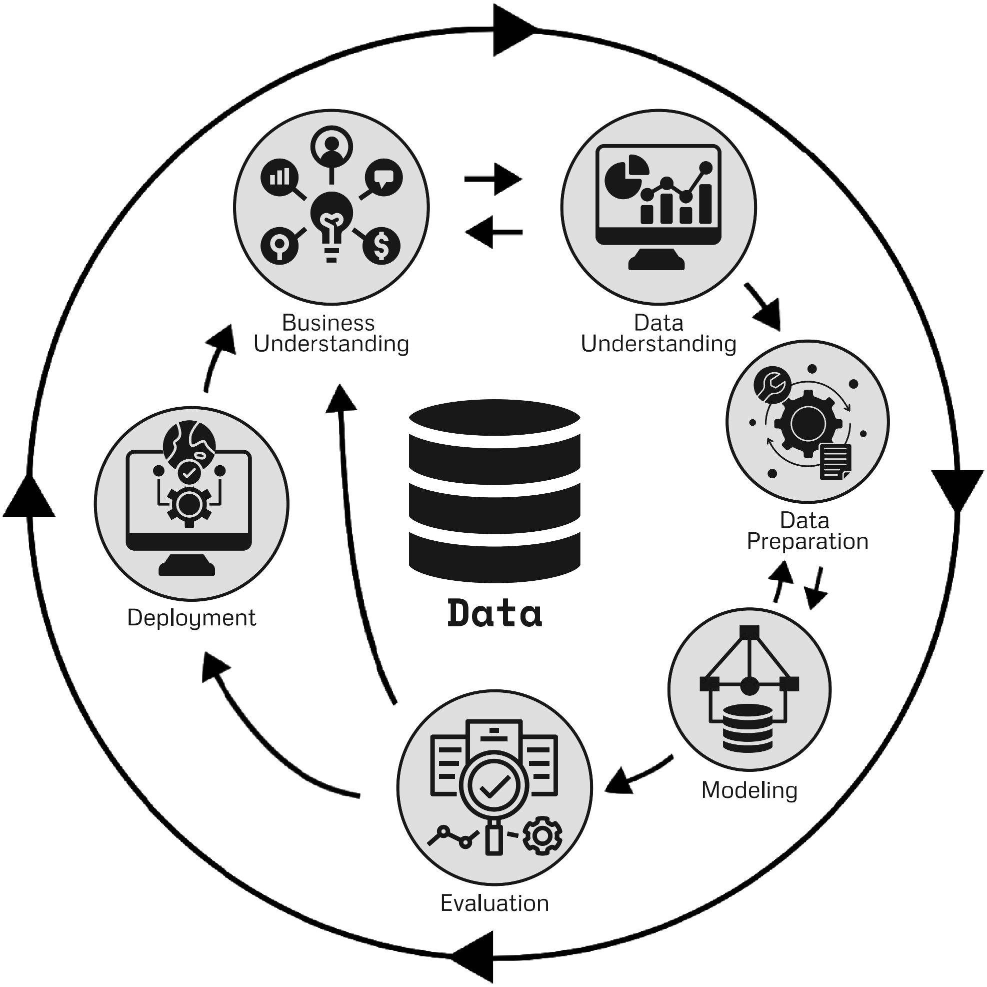


FIGURE 7.2 Cross industry standard process for data mining (CRISP-DM)

allowing the data analyst to revisit and revise previous steps as new insights or challenges emerge (Mariscal et al., 2010).

Descriptive analytics is primarily used in the data understanding and data preparation phases:

Business understanding: This is the first phase of CRISP-DM, where the data analyst defines the problem and the objectives of the data analysis project. The data analyst needs to understand the business context, the stakeholders, and the expected outcomes of the project. The data analyst also needs to formulate the data mining goals and plan the project scope and resources. In this phase, interviews, surveys, brainstorming sessions, observation, process mapping, and the creation of a project charter are the primary working methods.

Data understanding: This is the second phase of CRISP-DM, where the data analyst collects and explores the data that is relevant to the problem and the objectives. The data analyst needs to identify the data sources, acquire the data, and check the data quality and completeness. The data analyst also needs to perform descriptive analytics to summarize, visualize, and understand the main characteristics and patterns of the data. This includes methods and techniques like data profiling, data cleansing, data integration, descriptive statistics, histograms, various charts, and potentially correlation analysis.

Data preparation: This is the third phase of CRISP-DM, where the data analyst transforms and prepares the data for modeling. The data analyst needs to select the most relevant and useful features and records from the data, create new features or variables from existing ones, and handle missing values and outliers. The data analyst also needs to normalize, scale, encode, or discretize the data to make it suitable for different modeling techniques. Some of the methods and techniques used in this phase are feature selection, feature engineering, imputation, outlier detection, normalization, standardization, one-hot encoding, binning, and dimensionality reduction.

Modeling: This is the fourth phase of CRISP-DM, where the data analyst builds and trains predictive models using various algorithms and techniques. The data analyst needs to choose the appropriate modeling technique based on the type of problem (classification, regression, clustering, etc.), the type of data (numeric, categorical, text, etc.), and the evaluation criteria (accuracy, precision, recall, etc.). The data analyst also needs to tune the model parameters and compare different models to select the best one. Some of the methods and techniques used in this phase are: linear regression, logistic regression, multivariate regression, decision tree, random forest, k-means clustering, k-nearest neighbors (KNN), support vector machine (SVM), neural network (NN), deep learning (DL), cross-validation (CV), grid search (GS), and ROC curve. These techniques will be described further in the following chapters on predictive analytics.

Evaluation: This is the fifth phase of CRISP-DM, where the data analyst evaluates and validates the performance and quality of the selected model. The data analyst needs to test the model on new or unseen data and measure its accuracy and robustness. The data analyst also needs to assess whether the model meets the business objectives and expectations and whether it provides actionable insights or recommendations. Some of the methods and techniques used in this phase are: confusion matrix (CM), accuracy score (AS), precision score (PS), recall score (RS), F1 score (F1S), mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), r-squared score (R2S), lift chart (LC), gain chart (GC), cost-benefit analysis (CBA), and sensitivity analysis (SA).

Deployment: This is the sixth and final phase of CRISP-DM, where the data analyst deploys and delivers the model and its results to the stakeholders or end-users. The data analyst needs to communicate and present the model's findings and implications clearly and understandably. The data analyst also needs to provide documentation and support for using or maintaining the model. Some of the methods and techniques used in this phase are reports, dashboards, slideshows, infographics, storytelling, user manuals (UMs), application programming interfaces (APIs), web services (WSs), cloud computing (CC), containers (CTs), pipelines (PLs), feedback loops (FLs), and monitoring systems (MSs). These topics are explained further in the next chapter (Chapter 8) with the modern data stack.

PREDICTIVE ANALYTICS

Predictive analytics are applied by most companies every year. Not in the accounting process, but in the budgeting process. There are, however, many other places where we can move into this proactive mode of using data and analytics for decision-making.

Methods

Finding trends and presenting them "objectively" is the core of predictive analytics.

Predictive analysis is the process of building and applying models that can learn from the data and make predictions or recommendations for future or unknown situations. It is an advanced and powerful type of data analysis that can provide valuable insights, solutions, and opportunities for various domains and applications.

Some key concepts need to be introduced now as we are moving into uncertainty in our analysis (James et al., 2021). Training and test datasets, model fitting, and Bias-variance.

Training and test datasets: In predictive analytics (machine learning), a common task is the study and construction of algorithms that can learn from and make predictions on data. These input data used to build the model are usually divided into multiple datasets. Two (at least) datasets are commonly used in different stages of the creation of the model: training and test sets. The model is initially developed on a training dataset, which is a set of examples used to fit the parameters of the model. The fitted model is then used to predict the responses for the observations in a second dataset called the test dataset. The test dataset provides an unbiased evaluation of a final model fit on the training dataset.

Model fitting: The model fitting is a measure of how well a machine learning model generalizes to similar data to that on which it was trained. A well-fitted model produces more accurate outcomes. A model that is overfitted matches the data too closely. A model that is underfitted doesn't match closely enough. During the fitting process, you run an algorithm on data for which you know the target variable, known as "labeled" data, and produce a machine-learning model. Then, you compare the outcomes to real observed values of the target variable to determine their accuracy. Next, you use that information to adjust the algorithm's standard parameters, also called parameter tuning, to reduce the level of error, making it more accurate in uncovering patterns and relationships between the rest of its features and the target.

Bias–variance tradeoff: Bias and variance are the reasons machine learning models make prediction errors. In general, we want to have the lowest bias and variance possible, but in most cases, you can't decrease one without increasing the other; this is called the bias–variance trade-off. Bias is the difference between our actual and predicted values. Bias is the simple assumptions that our model makes about our data to be able to predict new data. Variance is the effect of a model that is too sensitive to noise present in the data.

Regression models have already been introduced in the descriptive analytics section, but where the focus there was to describe the historical information, it is now used to aid the "guesses" for the future.

Linear regression: Linear regression is a modeling technique that assumes a linear relationship between a dependent variable (also called response or outcome variable) and one or more independent variables (here as predictor variables) in the data. Linear regression helps to estimate the value of the dependent variable based on the values of the independent variables and to measure the strength and direction of the relationship. Some common techniques in linear regression are ordinary least squares (the most basic), gradient descent, ridge regression (L2), lasso regression (L1), elastic net regression (combining L1 and L2), and polynomial regression. L1 and L2 are used to limit the problems with overfitting which means the predictions are good for the test data, but not with new data.

Logistic regression: Logistic regression is a modeling technique that assumes a logistic or sigmoid function between a dependent variable (also called response or outcome variable) and one or more independent variables (also called predictor variables) in the data. Logistic regression helps to predict the probability of the dependent variable being in a certain category or class based on the values of the independent variables and to classify the data into different categories or classes. Logistic regression can be done using various tools and techniques, such as maximum likelihood estimation, gradient descent, regularization methods (L1, L2, or combined), and multinomial logistic regression.

Multivariate regression: Multivariate regression is a modeling technique that assumes a linear or nonlinear relationship between multiple dependent variables (also called response or outcome variables) and one or more independent variables (also called explanatory or predictor variables) in the data. Multivariate regression helps to estimate the values of multiple dependent variables based on the values of the independent variables and to measure the strength and direction of the relationships. Multivariate regression comes in different forms such as multivariate linear regression (MLR), multivariate nonlinear regression (MNR), multivariate logistic regression (MLR), multivariate adaptive regression splines (MARS), and partial least squares regression (PLSR) (James et al., 2021).

Decision tree: A decision tree is a modeling technique that represents a hierarchical structure of rules or conditions that split the data into different groups or classes based on the values of one or more features or variables in the data. A decision tree helps to classify or regress the data into different groups or classes based on the rules or conditions and to visualize the logic and process of the classification or regression. Decision tree can be done using various tools and techniques, such as entropy, information gain, gini index, classification and regression tree (CART), chi-square automatic interaction detection (CHAID), quick, unbiased, efficient statistical tree (QUEST), conditional inference tree (CIT), and C5.0 algorithm.

Random forest: Random forest is a modeling technique that combines multiple decision trees into an ensemble model that can improve the accuracy and robustness of the prediction or classification. Random forest helps to reduce the variance, bias, and overfitting problems of single decision trees by using bootstrap sampling, feature bagging, and majority voting methods. Some relevant terms related to random forest are bootstrap aggregating (bagging), random subspace method (RSM), out-of-bag error (OOB), variable importance measures (VIM), permutation feature importance (PFI), partial dependence plots (PDP), and local interpretable model-agnostic explanations (LIME).

K-means clustering: K-means clustering is a modeling technique that partitions the data into k groups or clusters based on the similarity or distance between the data points. K-means clustering helps to discover the hidden structure, patterns, and trends in the data and to segment the data into different groups or clusters.

K-nearest neighbors (KNN): KNN is a modeling technique that predicts or classifies the data based on the similarity or distance between the data points and their k nearest neighbors. KNN helps to capture the local structure and complexity of the data and to adapt to different types of data.

Support vector machine (SVM): SVM is a modeling technique that finds the optimal hyperplane or boundary that separates the data into different classes or groups based on the values of one or more features or variables in the data. SVM helps to maximize the margin or distance between the hyperplane and the closest data points and to minimize the classification error.

Neural network (NN): NN is a modeling technique that mimics the structure and function of the biological neural network in the brain. NN consists of multiple layers of interconnected nodes or units that can process and transmit information based on the values of one or more features or variables in the data. NN helps to learn complex and nonlinear relationships and patterns in the data and to perform various tasks such as classification, regression, clustering, dimensionality reduction, and natural language processing.

Deep learning (DL): DL is a branch of neural networks that uses multiple layers of nodes or units to learn high-level features or representations from the data. DL helps to handle large-scale and complex data and to perform various tasks such as image recognition, speech recognition, natural language processing, computer vision, natural language generation, machine translation, text summarization, sentiment analysis, and generative adversarial networks.

Cross-validation (CV): CV is a technique that evaluates the performance and generalization ability of a model by splitting the data into multiple subsets or folds and using some of them for training and some of them for testing. CV helps to avoid overfitting and underfitting problems and to select the optimal model parameters and hyperparameters.

Grid search (GS): GS is a technique that searches for the optimal combination of model parameters and hyperparameters by trying out all possible values in a predefined grid or range. GS helps to optimize the performance and accuracy of the model and to avoid underfitting and overfitting problems.

ROC curve: The ROC curve is a graphical display that shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR) of a binary classifier at different threshold levels. ROC curve helps to evaluate the performance and quality of the classifier and to compare different classifiers.

The techniques and methods described above are for the most part advanced techniques that should be handled by trained data scientists or analysts but being familiar with the concepts will help you ask explorative questions when someone claims their predictions are based on some specific technique.

DATA-PROJECT PROJECT MANAGEMENT

In this section, we will cover some approaches to projects that will enable data-driven decision-making. Decision-making in project management has traditionally been very data-driven with a focus on managing deliverables and deviations along with strong factual planning. A way to become a better leader with data-based decisions at the core can be through employing techniques from both traditional and modern project management.

Data projects are sometimes considered to be like regular IT projects. They include "information technology", they are time-bound, and they have distinct input and output. That would imply that a traditional project management approach could be used for data projects. There are, however, some important differences that need to be considered (Miller, 2023). Data projects are:

- Often collaborative and explorative by nature
- Based on dynamic digital assets often owned by the business
- Closely linked to the core business decisions of multiple stakeholders
- Short in time, but with strategic impact.

That means the data leader (project manager) needs to adopt the traditional approach.

Agile manifesto

The agile manifesto is the starting point for agile project management. It is built on four foundational ideas (Agile Manifesto, 2001):

- **Individuals and interactions** over processes and tools
- **Working software (solutions)** over comprehensive documentation
- **Customer collaboration** over contract negotiation
- **Responding to change** over following a plan.

Those where a response to the traditional project management methodologies were quite rigid in structure and had a focus on deterministic goalsetting; those that were hard to handle in the IT development process that had a high component of exploration, just like data projects.

Data projects require constant communication and feedback between the data team and the stakeholders, such as the customers, users, or managers. The data team should therefore prioritize the needs and expectations of the stakeholders and adjust their solutions accordingly. The data team also needs to work together effectively, sharing their skills and knowledge, and resolving any conflicts or issues as they require at least three different skill sets: data engineering, data science, and data analysis.

The manifesto states that data projects should deliver working solutions that provide value to the stakeholders, rather than spending too much time on documenting every detail of the data process. This is, however, a place where data products divert a bit from the pure software development projects. The data team should, however, focus on producing reliable and relevant data outputs, such as reports, dashboards, or models, that can be used by the stakeholders to make decisions or take actions. The data team should also test and validate their solutions regularly and incorporate feedback from the stakeholders to improve them.

Data projects involve close collaboration between the data team and the customers/stakeholders, rather than relying on fixed contracts or specifications. The data team should focus on understanding the goals and challenges of the customers and co-create solutions that meet their needs and preferences. The data team should also seek continuous input and feedback from the customers throughout the data project and use it to adapt the solutions/analyses accordingly.

Data projects should embrace change as an opportunity to learn and improve, rather than sticking to a rigid plan. The data team should be flexible and responsive to changes in the data sources, the business environment, or the customer requirements. The data team should therefore use an iterative and incremental approach to deliver solutions that can be modified or updated easily.

Structure

Agile (data) projects have a hierarchy of elements starting with themes of the minimum viable product (MVP) that is composed of some high-level road map elements broken into epics, which again are made from several user stories that define tasks in a backlog (Aha!, 2024). The special thing is that the lower-level elements are filled and prioritized as the project evolves.

Themes are broad and high-level goals or objectives that guide the vision and direction of the data project. For example, a theme could be "improve customer satisfaction" or "increase sales revenue".

The themes are broken into **epics** which are large chunks of work that cannot be completed in a single sprint and therefore need to be broken up.

User **stories** are short and specific descriptions of the features or functionalities that the customers want or need from the data solution. They are written from the perspective of the customers, and usually follow the format of "As a <role>, I want <feature>, so that <benefit>". For example, a user story could be "As a marketing manager, I want to see the conversion rate of different campaigns so that I can optimize my budget allocation".

Items in the backlog are **tasks** or activities that need to be done to complete the user stories and deliver the MVP. They are prioritized and assigned to the data team members, who work on them in short iterations or sprints. For example, an item in the backlog could be "collect and clean data from various sources" or "build and test a regression model".

The MVP is a version of the data solution that provides the minimum amount of functionality and value to the customers while allowing for feedback and learning. For example, an MVP could be a simple dashboard that shows key metrics or a basic model that predicts outcomes.

Process

A common framework for managing agile data projects is Scrum. Scrum employs an iterative approach, where the project is broken down into short, repeated work cycles called sprints that are typically 2–4 weeks long.

Each sprint starts with sprint planning, where the team selects user stories from the prioritized product backlog to work on during that sprint. The sprint goal is to deliver a potentially releasable increment of the product by the end of the sprint. The team holds daily Scrum meetings, which are short sync-up sessions to discuss the work completed, the next steps, and any blocking issues. This allows the team to track progress, coordinate efforts, and address obstacles quickly.

At the end of each sprint, two important meetings are held: the sprint review and the sprint retrospective. The sprint review meeting demonstrates the new increment to stakeholders and collects feedback. This allows stakeholders to provide input on the solution and validate that it aligns with their needs. The sprint retrospective meeting is for the team to inspect itself and create plans for improvements in the next sprint. Team members discuss what went well, what can be improved, and how to optimize team collaboration and efficiency.

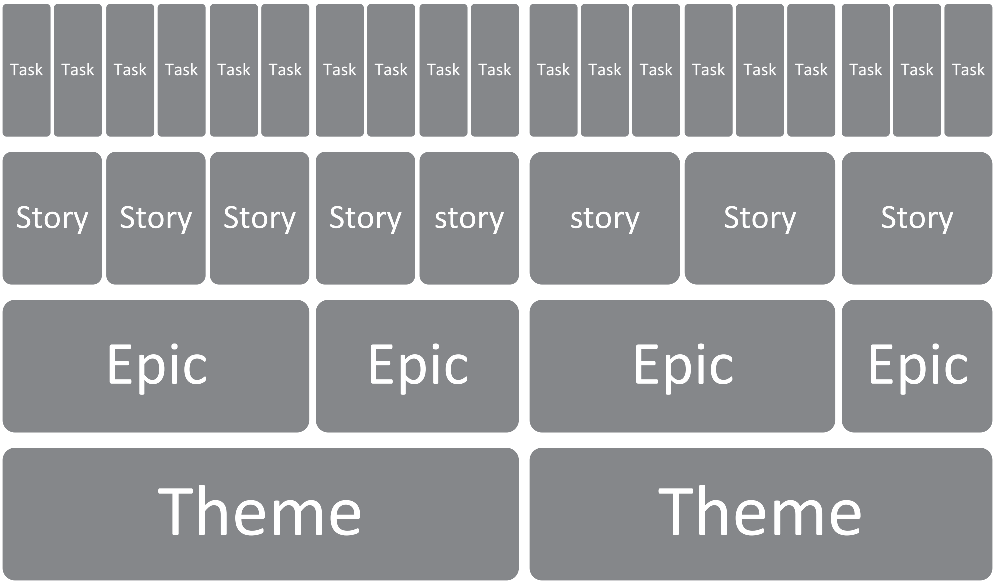


FIGURE 7.3 Task, story, epic, theme

Some common agile techniques used in Scrum data projects include:

- Task boards to visualize work status and progress
- Continuous integration to frequently merge and test new code
- Test-driven development to write tests before coding features
- Pair programming for real-time peer review and knowledge sharing
- Automated testing to verify correctness and catch issues early
- Short iterations and small batch sizes to deliver value faster.

This agile and collaborative framework promotes creativity, flexibility, and continuous improvement. The focus is on early and frequent inspection of actual working solutions rather than rigid project plans. There is also an emphasis on face-to-face communication and accountability across the self-organizing team. With the sprint retrospective driving process improvements, the data team can progressively enhance both the product and their teamwork. This sets up the project for adapting to evolving customer needs and priorities in an efficient way (Agile Business Consortium, 2014).

People

In agile projects, there are some distinct roles with responsibilities. That is also the case when the project is an analytics project that is to enable data-driven decision-making.

Product owner

The product owner is the voice of the customer and is accountable for ensuring the product creates value for end-users. They maintain and communicate the product vision, represent business and user needs, and decide feature priority.

Specific responsibilities include:

- Grooming the product backlog by writing clear user stories, estimating effort, and prioritizing based on value
- Clearly articulating acceptance criteria for work completion
- Validating solutions meet business objectives and user needs
- Making timely decisions on product direction and trade-offs
- Securing stakeholder buy-in and funding for the product.

Scrum master

The Scrum master is responsible for enabling teams to be productive, and empowered, and develop high-performance agile ways of working. They guide teams on Scrum theory, rules, and values while helping resolve impediments.

Typical duties involve:

- Coaching teams on agile and Scrum principles and practices
- Leading process improvements identified in sprint retrospectives
- Protecting teams from interference and distraction
- Facilitating important Scrum events like sprint planning, stand-ups, reviews, and retrospectives
- Tracking key metrics on team health and velocity to guide optimization.

Development team

The development team builds the product incrementally across sprints. Teams are cross-functional with all the expertise needed to take user stories from idea to implementation without relying on external groups.

Team responsibilities cover:

- Estimating the level of effort for backlog items
- Designing, developing, testing, and deploying agreed-upon features
- Executing sprint tasks to meet the committed sprint goal
- Inspecting work and processes to identify areas for improvement
- Collaboratively managing their own work and team composition.

When analytics projects grow to focus on prescriptive analytical results there is a need for further security as the decisions made from the analytics will be focusing on agreeing or disagreeing with the proposed solution. Therefore, additional steps are often needed in the project management that will be discovered in the end.

PRESCRIPTIVE ANALYTICS

The most advanced and complex form of analytics is prescriptive analytics. This is where our analysis can recommend a course of action. This does require a sound data foundation in terms of quality, amount, and structure. It is also here, however, that the organization moves from good to best in class when making data-driven decisions.

Methods

Prescriptive analysis is the process of using data to determine an optimal course of action. It goes beyond descriptive and predictive analysis by not only providing insights into what is likely to happen in the future but also suggesting the most appropriate actions to take based on those predictions. Prescriptive analysis uses data from a variety of sources, such as statistics, machine learning, and data mining, to identify possible future outcomes and show the best option.

Some of the common methods and techniques used for prescriptive analysis are:

Simulation: Simulation is a method that uses mathematical models and computer programs to imitate the behavior and dynamics of a real-world system or process. Simulations help to test different scenarios and assumptions and to evaluate the impact of different decisions or actions on the system or process. Simulation can be done using various tools and techniques, such as system dynamics models, discrete-event models, agent-based models, Monte Carlo methods, Markov chain methods, cellular automata methods, and simulation software (Yin & McKay, 2018).

Graph analysis: Graph analysis is a method that uses graph theory and network science to model and analyze complex relationships and interactions among entities in the data. Graph analysis helps to find the optimal paths, flows, or connections between entities and to optimize the network structure and performance. Graph analysis uses concepts such as nodes, edges, weights, directions, adjacency matrices, incidence matrices, graph algorithms, shortest path algorithms, spanning tree algorithms, etc. (Quinto, 2020).

Complex event processing: Complex event processing is a method that uses data streams and rules to detect and analyze patterns of events that occur in real-time or near real-time. Complex event processing helps to monitor and respond to situations that require immediate action or intervention. Complex event processing can be done using various tools and techniques, such as event sources, event sinks, event processors, event filters, event aggregators, event transformers, event correlators, event patterns, event rules, event queries, and complex event processing engines (Geisler, 2020).

Neural networks: Neural networks are a technique that mimics the structure and function of the biological neural network in the brain. Neural networks consist of multiple layers of interconnected nodes or units that can process and transmit information based on the values of one or more features or variables in the data. Neural networks help to learn complex and nonlinear relationships and patterns in the data and to perform various tasks such as classification, regression, clustering, dimensionality reduction, and natural language processing (Nielsen, n.d.).

Recommendation engines: Recommendation engines are a technique that uses data mining and machine learning to provide personalized suggestions or recommendations to users based on their preferences, behavior, or feedback. Recommendation engines help to increase customer satisfaction, loyalty, retention, and revenue (Walia, 2022).

Heuristics: Heuristics is a technique that uses rules of thumb or shortcuts to find approximate solutions or answers to complex problems or questions. Heuristics help to simplify the problem or question and to reduce the time and effort required to find a solution or answer (Pinheiro & McNeill, 2014).

Just like predictive analytics, these techniques and methods are complex to implement and require a significant skill level but understanding them on a conceptual level will enable interaction with the data scientist who would be leading the usage of these tools. As they start recommending courses of action and/or driving decisions independently it's important to have proper guardrails[1] in place.

PROJECT MANAGEMENT OF BIG DATA AND ML PROJECTS

Prescriptive analytics leverages data and models to recommend optimal actions or decisions for meeting business objectives. Prescriptive solutions can have tremendous value but require careful planning and requirements analysis due to their complexity.

The introduction and requirements-gathering phases focus on fully understanding the business needs, challenges, and metrics that the prescriptive solution aims to address. Structured workshops with key stakeholders help define the scope. What decisions should the solution optimize? What recommendations should it provide? What metrics reflect success? How will users interact with the system?

Detailed business requirements documents (BRDs) capturing this information in granular specificity are created. Each requirement is cataloged with descriptions, acceptance criteria, priorities for implementation, and traceability. The rigorous analysis clarifies technical, functional, and operational requirements from the end-user workflow to the backend data infrastructure.

The project manager partners with business analysts to facilitate exhaustive information gathering from subject matter experts via interviews and requirement-gathering sessions. Review cycles ensure documentation accuracy before final sign-off. Prototyping and visualization can illustrate planned functionality. Supplementing natural language requirements with models, diagrams, and use cases improves precision.

Careful requirements analysis reduces costly late-stage rework. It aligns stakeholder expectations on scope and functionality early. Traceability matrices tracing each element across requirement documents, system design specifications, test plans, and deployment procedures support verification and auditing later. This phase is intensive but laying this structured foundation prepares subsequent teams to engineer, build, and implement a solution matching business needs (Landau, 2022).

System design phase

With requirements established, the system design phase translates specifications into detailed technical architectures and plans.

Design starts by decomposing high-level requirements into subsystem components and modules. Architects define schema and data flows, application logic and integrations, user interfaces, and experiences. Rigorous documentation captures every facet of specifications that can be handed off for development confidently.

Database and infrastructure design provides the foundation. Schema and datatypes optimize storage and performance. Cloud vs on-prem vs hybrid decisions balance costs, capabilities, and compliance. API specifications enable the integration of predictive models, external data sources, and downstream data consumption applications.

Application design finalizes workflows, business logic, and experiences. Workflow diagrams visualize end-to-end data pipelines, calculations, rules engines, and process automation capabilities. Application frontends wireframe an intuitive, adopted interface and interaction model for target users leveraging modern UX standards.

With comprehensive specifications and infrastructure designs completed, reliability engineering ensures resilience and high availability even under peak loads. Disaster recovery plans balance recovery point and recovery time objectives with cost tolerance. Rigorous security reviews identify vulnerabilities for remediation a priori.

Each design document faces extensive peer review, especially at subsystem intersections, to ensure seamless integration and prevent gaps or disconnects across blueprints from multiple architects. Solution leads reconcile feedback until all stakeholders sign off on completed packages ready for development.

Meticulous system design ensures what gets built matches both requirements and infrastructure constraints. It also facilitates dividing workstreams allowing large teams to progress in parallel during project execution.

The implementation and testing phases

With comprehensive designs in hand, distributed component teams begin constructing defined architecture pieces – services, pipelines, models, databases, user interfaces (UIs), and integrations.

Code development tasks progress in parallel by applying agile best practices like test-driven development and continuous integration allowing rapid incremental progress. Automated testing frameworks validate code quality on every commit. Teams' demo functionality frequently seeks feedback driving refinements grounded in working software.

On the data side, pipeline code ingests, combines, cleans, transforms, and stores input data sets. Data engineers' script complex ETL processes and apply quality checks and profile outputs to catch issues early. Once data warehouses are populated, data analysts utilize BI tools to build visualization reporting dashboards.

Data scientists leverage pre-processed data to train, evaluate, optimize, and deploy predictive and prescriptive models. Rigorous statistical analysis ensures model fairness, explainability, and performance. Integration code seamlessly connects model APIs to downstream decision logic code that applies predictions to recommend actions meeting business KPIs.

Frontend engineers build responsive experiences for business users providing contextualized insights. Usability testing with target personas affirms workflows are intuitive driving refinements. Comprehensive training prepares both developers and users alike.

As all components are near completion, end-to-end testing verifies cohesive system functioning. Automation tests exercise critical user journeys and data pipelines validating conformance to specifications. Performance testing across staging environments with production-scale data gauges responsiveness, scalability, and resilience even under peak loads. Verified components are now prepared for delivery.

The deployment and maintenance phase

With extensively tested components cleared for release, the focus shifts to transitioning users and technology to production environments. Meticulous planning minimizes downtime and business disruption during migration.

Operational readiness reviews verify support processes, monitoring and alerting rules, technical documentation, and contingency plans are in place for go-live. Content libraries provide self-service troubleshooting. Admin training equips IT teams to manage ongoing ops.

Phased rollout plans initially direct a portion of traffic to new systems allowing testing real-world functioning before full cutover. Feature flags that can toggle functionality quickly contain any unforeseen issues. Staged database migrations prevent data loss.

To smooth adoption, change management campaigns prompt users of imminent changes and new capabilities. Brief test groups provide final usability feedback to refine help guides and user manuals. Training workshops demonstrate practical application, so users gain competency independently.

During support transition, Issue tracking and resolution SLAs transfer to ops teams for production incidents. However, an on-call escalation chain retains designer expertise to diagnose complex bugs. Ongoing maintenance contracts fund continuous improvements, technical debt reduction, and new capability development.

Post-deployment operational reviews assess system performance across critical KPIs like uptime, latency, data quality, and end-user productivity. Quarterly milestones evaluate adoption, utility, and ROI determining if designs require modifications. Feedback loops remain tied to the ultimate business value delivered.

The maintenance phase sustains the ongoing enhancement of a trusted solution delivering durable value. Continued funding secures the expanding potential of analytics generating exponential business impacts over time.

Managing the project lifecycle

Successfully delivering prescriptive analytics solutions requires structured project management spanning planning, execution, governance, and close down. Skilled project leaders proficiently juggle competing priorities to drive on-time, on-budget, beneficial implementations.

Initiation starts by chartering the project – securing executive sponsorship, assigning project leads, and cordoning key parameters like budget, timeline, and resourcing needs. Structured project plans segment efforts into discrete work packages with logical predecessors and dependencies. Realistic schedules balance parallel streams while setting the pace to control scope creep.

Resource management acquires requisite skills and capacities across each project phase while optimizing costs. Virtual talent bench models supply and demand to the right size teams when needs spike. Tools like Atlassian facilitate work allocation, progress dashboards, and transparency across large, distributed groups.

Financial governance ensures discipline tracking commitments against actuals. Reason codes explain deviations in spend rates, billable hours, or case counts. Variance triggers mid-course corrections guiding projects back on track. Standardized accounting separates capital and operating expenses monitoring the total cost of ownership.

Compliance and risk management implement controls fulfilling security, regulatory, and ethical best practices, especially in regulated industries like financial services and healthcare. Partner security reviews assess solution designs. Ethics boards evaluate data sourcing, predictive model fairness, and downstream impact early when issues are easier to address. Procurement protocols confirm vendor assessments and contracting meet corporate policies.

Change control processes manage modifications to scope, designs, or requirements. Change review boards carefully consider proposed adjustments providing transparency on attendant impacts to budget, timeline, and resourcing before approving. Traceability back to business requirements reminds teams of downstream effects.

As closing milestones approach, transition planning hands solutions over to permanent operations teams. Acceptance testing formally verifies user satisfaction. Retrospective evaluations summarize successes and lessons learned across process and product dimensions providing recommendations to improve future implementations and maximize business return on investment.

Operationalizing the project

Managing massive amounts of data: As the name suggests, big data is big. Most companies are increasing the amount of data they collect daily. Eventually, the storage capacity a traditional on-premises server solution can provide will be inadequate, which worries many business leaders. To handle this challenge, companies are migrating their IT infrastructure to the cloud.

The data itself presents another challenge to businesses. There is a lot, but it is also diverse because it can come from a variety of different sources. A business could have analytics data from multiple websites, sharing data from social media, user information from CRM software, email data, and more. None of this data is structured the same but may have to be integrated and reconciled to gather necessary insights and create reports.

Data silos occur when different departments or teams within an organization store their own data in separate systems or formats. This can lead to inconsistencies in the data and make it difficult to integrate and analyze. Poor data quality can also be a problem, as it can lead to inaccurate insights and decisions.

Lack of coordination to steer big data/AI initiatives: Big data projects often require collaboration between different teams with different skill sets, such as IT, analytics, and business operations. Without proper coordination and communication, these initiatives can become disjointed and fail to deliver value.

Big data/AI/prescriptive analytics projects require specialized skills such as data science, machine learning, and programming. Finding people with these skills can be difficult and expensive.

CONCLUSION

The chapter covers different types of analytics – descriptive, predictive, and prescriptive – and how they can be used to support organizational decision-making.

Descriptive analytics involves summarizing, visualizing, and exploring data to understand its characteristics and patterns. Methods include data profiling, cleansing, integration, descriptive statistics, visualizations like histograms and boxplots, correlation analysis, feature selection and engineering, imputation, outlier detection, normalization, standardization, encoding, binning, and dimensionality reduction. Descriptive analysis is used in the data understanding and preparation phases of the CRISP-DM methodology for data mining projects.

Predictive analytics involves building and applying models to learn from data and make predictions or recommendations about future outcomes or behaviors. Methods include linear regression, logistic regression, multivariate regression, decision trees, random forests, clustering algorithms like k-means, classification algorithms like KNN and SVM, neural networks and deep learning, cross-validation, grid search, and ROC analysis.

Prescriptive analytics involves using data to determine the optimal course of action, not just providing insights into likely future outcomes but also the best actions to take. Methods include graph analysis, simulation, complex event processing, neural networks, recommendation engines, and heuristics like hill climbing algorithms, dynamic programming, and genetic algorithms.

It's important to note that, depending on the utilization and types of decisions made of the various techniques, they can be placed in different categories.

Data project management discusses how to adapt traditional project management approaches for data projects, which tend to be more exploratory, collaborative, iterative, and more closely linked to business decisions.

Within project management, agile principles often make sense. It has a focus on valuing individuals over processes, working solutions over documentation, customer collaboration over contracts, and responding to change over rigid plans. Data projects require constant user feedback and cross-functional teamwork.

In the Scrum framework, projects are broken into sprints delivering releasable product increments. Meetings like sprint planning, stand-ups, reviews, and retrospectives enable transparency.

For prescriptive analytics, rigorous requirements analysis is key. The system design phase translates specifications into technical architectures. Implementation involves concurrent building of components like data pipelines, models, UIs, and integrations. Testing verifies cohesive functioning before phased deployment. Prescriptive analytics can often be linked with data products.

CASE

Maersk in Vietnam

Maersk is a global leader in integrated container logistics and operates an extensive network in Vietnam. The company first established operations in Vietnam in 1994 as Maersk Line, providing ocean freight services. Over the past three decades, Maersk has expanded significantly within Vietnam to offer a comprehensive range of logistics solutions.

Maersk Vietnam comprises several business units including Maersk Line, APM Terminals, Damco, and Maersk Container Industry. Its headquarters is in Ho Chi Minh City with additional offices in Hanoi, Da Nang, and other major ports.

The company employs over 1,000 people in Vietnam and maintains relationships with a large network of local partners and suppliers.

Maersk's core operations in Vietnam involve ocean shipping, port terminal operations, land-based container transportation, air freight, customs brokerage, warehousing, and distribution services. Key import commodities handled include electronics, garments, footwear, and agricultural products, while key exports include textiles, coffee, seafood, and manufactured goods.

Maersk possesses a strong advantage in Vietnam due to its global connectivity and integrated offerings. The company operates direct services from Vietnam to all major markets worldwide.

However, in 2019 they were facing significant uncertainty.

- Vietnam's GDP grew 7.0% in 2019, up from 6.8% in 2018, driven by strong manufacturing output and exports. This supported the demand for Maersk's shipping and logistics services.
- However, trade growth slowed to 7.6% in 2019, down from 12.2% in 2018, as Vietnam's key export markets like China and the United States faced economic headwinds. This moderated growth for Maersk's ocean shipping volumes.
- Vietnam's container throughput rose 6.4% year-on-year to about 20 million TEUs in 2019 across all ports. Maersk maintained a market share of around 20% of Vietnam's container volume.
- Infrastructure constraints, especially at main ports like Hai Phong and Ho Chi Minh City, resulted in congestion and delays. This posed operational challenges for Maersk.
- Rising labor costs in Vietnam increased operating expenses for Maersk's labor-intensive warehouse and trucking operations.
- Overall, Maersk continued to benefit from Vietnam's growth prospects and development needs. However, trade tensions and infrastructure gaps remained key challenges in 2019.

They needed to make some decisions on how to structure the business going forward. The plan had been to move towards becoming an end-to-end logistics services provider, which was different from working only with the shipping agents who already knew the details of the shipping industry. They were also getting busy with the requests by companies looking to move business from China as they saw a budding trade war between China and the USA.

You are to provide suggestions for:

- Analyses that could guide the direction for Maersk in Vietnam in 2019 (pre-Covid)?
- How a project could be structured?

After a decision was made, they needed to structure the continuous follow-up on the business progression. Ms. Nguyen Phu Thuy Van was Trade and Marketing

Manager responsible for various tasks related to the ongoing commercial operations. Which type of analysis and analytics setup would you recommend to ensure that she could help guide the process?

Would your recommendations have caught the Covid-19 pandemic and if yes, how?

KEY TERMS

Acceptance testing: User validation.
Agile manifesto: Values individuals over processes.
Backlogs: Prioritized lists of tasks.
Boxplots: Data visualization with quartiles and outliers.
Change control: Managing modifications.
Clustering: Grouping similar data points.
Compliance: Security and regulatory controls.
Correlation analysis: Measure linear relationship between variables.
Cross-validation: Evaluating models on subset splits.
Data cleansing: Correcting or removing errors and inconsistencies.
Data profiling: Examining data structure, content, and quality.
Decision trees: Models with hierarchical rule splits.
Deep learning: Multi-layer neural networks.
Deployment: Transition to production.
Descriptive analytics: Summarizing and visualizing data to explore main characteristics.
Event processing: Analyzing real-time data streams.
Feature engineering: Creating new variables from existing ones.
Financial governance: Budget discipline and tracking.
Histograms: Graphical data frequency distributions.
Implementation: Building system components.
Imputation: Replacing missing values with estimates.
Neural networks: Models inspired by biological neurons.
Normalization: Scaling data to a standard range.
Predictive analytics: Building models to forecast future outcomes.
Prescriptive analytics: Recommending optimal actions from data.
Project initiation: Securing sponsorship and resources.
Project planning: Activity sequencing and timelines.
Random forest: Ensemble of decision trees.
Recommendation engine: Suggest personalized recommendations.
Regression: Model the relationship between variables.
Requirements analysis: Detailed scoping with users.
Resource management: Skills and staffing allocation.
Retrospectives: Evaluating outcomes and improvements.
Simulation: Imitating real-world system dynamics.
Sprints: Short, fixed work cycles.

Standups: Daily progress meetings.
System design: Technical architecture plans.
Testing: Validating against specifications.
Transition planning: Support readiness and training.
User stories: Specific feature requests.

REVIEW QUESTIONS

1. What type of analytics focuses on exploring data characteristics and patterns?
2. What process examines data quality and metadata?
3. What visualization shows data frequency distributions?
4. What analysis measures the strength of the relationship between variables?
5. What involves creating new features from existing data?
6. What method replaces missing values?
7. What scale data have a mean of 0 and a standard deviation of 1?
8. What builds models to predict future outcomes?
9. What type of model assumes linear variable relationships?
10. What ensemble method combines multiple decision trees?
11. What technique groups similar data points?
12. What is a key benefit of deep learning models?
13. What evaluates model quality on subset data splits?
14. What leverages data to recommend optimal actions?
15. What technique imitates real-world system dynamics?
16. What analyzes real-time data streams?
17. What suggests personalized recommendations?
18. Which Agile Manifesto principle values working software?
19. What are short descriptions of desired features?
20. What are prioritized lists of work tasks?
21. What are short, fixed development cycles?
22. What meetings track daily progress?
23. What phase elicits detailed user requirements?
24. What phase designs system architecture?
25. What verifies user acceptance?

Answers to review questions

1. Descriptive analytics focuses on summarizing, visualizing, and exploring data to understand its main characteristics and patterns. This is covered in the "Descriptive analytics" section early in the passage.
2. Data profiling is the process of examining the data quality, structure, content, and metadata. Data profiling is discussed in the "Methods" sub-section under "Descriptive analytics".
3. Histograms are graphical displays that show the frequency distributions of variables. Histograms are covered in the "Methods" sub-section under "Descriptive analytics".

4 Correlation analysis measures and tests the strength and direction of the linear relationship between two variables. Correlation analysis is discussed in the "Methods" sub-section under "Descriptive analytics".

5 Feature engineering is the process of creating new features or variables from existing ones to enhance modeling performance. Feature engineering is covered in the "Methods" sub-section under "Descriptive analytics".

6 Imputation methods replace missing values in data with plausible estimates to handle incomplete data. Imputation is discussed in the "Methods" sub-section under "Descriptive analytics".

7 Standardization is the process of scaling data to have a mean of zero and a standard deviation of one. Standardization is discussed in the "Methods" sub-section under "Descriptive analytics".

8 Predictive analytics focuses on building and applying models that can learn from data to make predictions about future or unknown situations. This is covered in the "Predictive analytics" section.

9 Linear regression assumes a linear relationship between the dependent and independent variables. Linear regression is discussed in the "Methods" sub-section under "Predictive analytics".

10 Random forest combines multiple decision trees into an ensemble model. Random forest is discussed in the "Methods" sub-section under "Predictive analytics".

11 Clustering algorithms group or segment data points based on similarity or distance. Clustering is covered in the "Methods" sub-section under "Predictive analytics".

12 A key benefit of deep learning models is their ability to handle large-scale and complex data. This is mentioned in the overview paragraph of "Deep learning" in the "Methods" sub-section under "Predictive analytics".

13 Cross-validation evaluates model quality by testing on subset splits of the data. Cross-validation is discussed in the "Methods" sub-section under "Predictive analytics".

14 Prescriptive analytics uses data to determine optimal courses of action, not just predict future outcomes. This is covered in the "Prescriptive analytics" section.

15 Simulation uses mathematical models to imitate real-world system behaviors and dynamics. Simulation is discussed in the "Methods" sub-section under "Prescriptive analytics".

16 Complex event processing analyzes real-time data streams to detect situations needing immediate action. Event processing is covered in the "Methods" sub-section under "Prescriptive analytics".

17 Recommendation engines use data to provide personalized suggestions to users. Recommendation engines are discussed in the "Methods" sub-section under "Prescriptive analytics".

18 The manifesto principle "Working software over comprehensive documentation" values delivering working solutions rather than detailed documents. This principle is listed in the "Agile manifesto" section.

19 User stories are short, specific descriptions of features or functions desired by users. User stories are explained in the "Structure" sub-section.

20 Backlogs are prioritized lists of tasks needed to complete user stories and deliver functionality. Backlogs are also covered in the "Structure" sub-section.

21 Sprints are short, repeated work cycles to deliver releasable product increments. Sprints are discussed in the "Process" sub-section.

22 Daily stand-up meetings allow teams to track progress and coordinate efforts. Standups are mentioned in the "Process" sub-section.

23 The requirements analysis phase focuses on understanding detailed user needs and expectations. This phase is described in the "Project management of big data machine learning projects" section.

24 The system design phase creates plans for the technical architecture. This phase is covered in the "Project management of big data machine learning projects" section.

25 Acceptance testing verifies that the delivered solution satisfies user needs. Acceptance testing is discussed in the "Managing the project life cycle" section.

NOTE

[1] Process for stopping an algorithmic decision.

BIBLIOGRAPHY

Agile Business Consortium. (2014). DSDM Agile Project Framework Handbook. Retrieved from DSDM Agile Project Framework Handbook. agilebusiness.org

Agile Manifesto. (2001). Manifesto for Agile Software Development. https://agilemanifesto.org/

Aha! software. (2024). Themes vs. Epics vs. Stories vs. Tasks in Scrum. Aha! Agile Guide. www.aha.io/roadmapping/guide/agile/themes-vs-epics-vs-stories-vs-tasks

Geisler, S. (2020). Complex event processing (CEP). In: Schintler, L., McNeely, C. (Eds.) *Encyclopedia of Big Data*. Cham: Springer. https://doi.org/10.1007/978-3-319-32001-4_276-1

IBM. (2024). *CRISP-DM help overview*. www.ibm.com/docs/en/spss-modeler/saas?topic=guide-introduction-crisp-dm

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). New York: Springer.

Landau, P. (2022). *Requirements gathering: A quick guide*. www.projectmanager.com/blog/requirements-gathering-guide

Mariscal, G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, *25*(2), 137–166. https://doi.org/10.1017/S0269888910000032.

Miller, S. (2023, 8). *Top 8 challenges of big data and how to solve them*. Capterra. www.capterra.com/resources/challenges-of-big-data/

Nielsen, M. (n.d.). Neural networks and deep learning. http://neuralnetworksanddeeplearning.com/

Pinheiro, C. A. R., & McNeill, F. (2014). *Heuristics in analytics: A practical perspective of what influences our analytical world*. Wiley Online Library. www.wiley.com/en-gb/Heuristics+in+Analytics%3A+A+Practical+Perspective+of+What+Influences+Our+Analytical+World-p-9781118347607

Quinto, B. (2020). Graph analysis. In: *Next-generation machine learning with Spark*. Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4842-5669-5_6

Walia, M. S. (2022). *A comprehensive guide on recommendation engines In 2022*. Analytics Vidhya. www.analyticsvidhya.com/blog/2022/01/a-comprehensive-guide-on-recommendation-engines-in-2022/

Yin, C., & McKay, A. (2018). Introduction to modeling and simulation techniques. In: *Proceedings of ISCIIA 2018 and ITCA 2018. The 8th International Symposium on Computational Intelligence and Industrial Applications*, Binjiang International Hotel, Tengzhou, Shandong, China, November 2–6, 2018.

CHAPTER 8

Data Infrastructure

How to Build and Manage a Modern Data Stack

Data is the lifeblood of any organization that wants to make informed and effective decisions. However, data alone is not enough. You also need a robust and reliable data infrastructure that can collect, store, process, analyze, and deliver data efficiently and securely. Data infrastructure is a collection of hardware, software, networks, protocols, standards, policies, and procedures that enable the acquisition, storage, processing, analysis, and dissemination of data.

"I want us to make data-driven decisions. We have the data (I think). What now?". This could easily be said by many in the senior executive management layer to employees when they are handed the initial AI or data project. That is where this chapter becomes relevant. One of the basic requirements for making data-driven decisions is to have access to the data in the right quality, in the right formats, at the right times. This means we need to bring in some data engineering skills and understanding.

In this chapter, you will learn how to build and manage a modern data stack that can support your data-driven decision-making needs. You will start by understanding where your data comes from, both internally and externally. You will then explore the main components of a data system architecture that can handle different types of data sources. You will also learn how to choose appropriate tools and technologies for each component of the architecture. Next, you will learn how to define and prioritize your requirements for a data infrastructure project using FURPS (Functionality, Usability, Reliability, Performance, Supportability) and MoSCoW (Must have, Should have, Could have, Won't have) methods. You will also learn how to design and document a data infrastructure solution that meets your requirements.

There are roughly speaking three roles in the data setup that need to be filled. Chapter 2 elaborated on this.

- The analyst who understands the business and some of the math/statistics
- the data scientist who understands the math/statistics and some of the technology, and lastly
- the data engineer who understands the technology and a little of the math/stat.

In this chapter, we will be focusing on what the data engineer will consider their primary domain. The objective is to cover enough to work efficiently with the data engineer, but the deep technical implementations are out of scope as that would require several books and is a field that moves very rapidly.

DOI: 10.4324/9781003457787-8

LEARNING OBJECTIVES:

L8.1 Explain the importance and benefits of having a well-designed and managed data infrastructure for data-driven decision-making

L8.2 Identify and describe the main components of a data system architecture, such as ingestion, storage, analysis, and governance layers

L8.3 Compare and contrast different types of data sources, such as internal and external, structured and unstructured, batch and streaming, etc.

L8.4 Evaluate and select appropriate tools and technologies for each component of the data system architecture, such as ETL/ELT, data warehouse, data lake, data mesh, etc.

L8.5 Apply best practices for ensuring the quality, reliability, scalability, and performance of the data infrastructure, such as data validation, monitoring, testing, backup, etc.

L8.6 Define and prioritize the functional and non-functional requirements for a data infrastructure project using FURPS and MoSCoW methods

L8.7 Design and document a data infrastructure solution that meets the requirements and aligns with the business goals and strategy

The chapter case is about Intelligent Machines/bKash which is a Fintech company in Bangladesh. They needed to have an infrastructure that could support the scaling of the business in a market of potentially more than 100 million customers.

WHERE IS THE DATA

In Chapter 5 the different data locations were covered in detail. It is, however, relevant to revisit them as they form the foundation for the data engineering that we are talking about in this chapter.

Internal data sources

Internal data sources are the data generated and stored within your organization. They can provide valuable insights into your business processes, performance, customers, products, and services. Internal data sources can be classified into four main types: master data, transactional data, metadata, and log data. Each type of data has its characteristics, purposes, and challenges. In this section, you will learn about each type of internal data source and where you can find them in your organization.

Master data is the core data that defines the key entities and concepts of your organization, such as customers, products, suppliers, locations, etc. Master data is usually stored in a centralized database or system that ensures consistency and accuracy across different applications and departments. Master data is essential for providing a single source of truth and a common language for your organization.

Transactional data is the data that records the activities and events that occur in your organization, such as sales, orders, payments, deliveries, etc. Transactional data is usually stored in operational systems or databases that support the day-to-day functions of your organization. Transactional data is important for measuring and monitoring your business performance and outcomes.

Metadata is the data that describes the properties and characteristics of other data, such as definitions, formats, schemas, relationships, etc. Metadata is usually stored in a separate system or repository that provides information about the structure and meaning of your data. Metadata is useful for facilitating the understanding and integration of your data across different sources and systems.

Log data is the data that captures the actions and behaviors of users or systems in your organization, such as clicks, views, errors, exceptions, etc. Log data is usually stored in files or streams that record the details and timestamps of each event. Log data helps analyze and optimize the user experience and system performance of your organization.

Master data

Master data is one of the most important types of internal data sources for your organization. Master data represents the core business entities and concepts that are shared and used across different applications and departments. For example, master data can include information about your customers, such as their names, addresses, preferences, etc. Master data can also include information about your products, such as their names, descriptions, prices, etc. Master data can also include information about your suppliers, locations, employees, and any other relevant entities for your organization.

Master data is usually stored in a centralized database or system that ensures consistency and accuracy across your organization. This means that any changes or updates to the master data are reflected in all the applications and departments that use it. This also means that there is no duplication or discrepancy in the master data across different sources and systems. Having a centralized and consistent master data can provide many benefits for your organization, such as:

- Improving the quality and reliability of your data for decision-making
- Reducing the cost and complexity of data integration and maintenance
- Enhancing the efficiency and effectiveness of your business processes and operations
- Increasing customer satisfaction and loyalty by providing personalized and relevant services
- Supporting the compliance and governance of your data with relevant regulations and standards.

Master data is essential for providing a single source of truth and a common language for your organization. A single source of truth means that there is only one version of the master data that is authoritative and trusted by all the stakeholders. A common language means that there is a clear and consistent definition and understanding of the master data across your organization. Having a single source of truth and a common language can help you to:

- Align your business goals and strategy with your data
- Communicate and collaborate effectively with your internal and external partners
- Innovate and create new value from your data
- Adapt and respond to changing market conditions and customer needs.

In this chapter, you will learn more about how to design and implement a master data management (MDM) system that can help you create and maintain high-quality and reliable master data for your organization. You will also learn about the best practices and challenges of managing your master data in a dynamic and complex environment.

Transactional data

Transactional data is another important type of internal data source for your organization. Transactional data captures the activities and events that occur in your organization, such as sales, orders, payments, deliveries, etc. Transactional data reflects the transactions and interactions that your organization has with its customers, suppliers, partners, and other stakeholders. Transactional data is usually stored in operational systems or databases that support the day-to-day functions of your organization. For example, transactional data can include information about your sales transactions, such as the date, time, amount, product, customer, etc. Transactional data can also include information about your order fulfillment, such as the status, location, tracking number, etc.

Transactional data is important for measuring and monitoring your business performance and outcomes. Transactional data can provide you with valuable insights into your business processes, operations, and results. For example, transactional data can help you to:

- Analyze and optimize your sales performance and revenue
- Evaluate and improve your customer satisfaction and retention
- Identify and resolve any issues or errors in your order fulfillment
- Track and manage your inventory and supply chain
- Monitor and control your costs and expenses.

How to collect and store transactional data from different sources and systems is what will be covered in the section about the ingestion layer in the data architecture. How to process and store the data is covered when focusing on the storage layer. Transactional data in some cases can have a volume and veracity so that it must be streamed to the storage layer as it is generated. How to deliver transactional data to different users and applications is part of the analysis layer in the data system architecture.

Metadata

Metadata is another type of internal data source for your organization. Metadata is the data that describes the properties and characteristics of other data, such as definitions, formats, schemas, relationships, etc. Metadata provides information about the structure and meaning of your data.

For example, metadata can include information about the columns, types, constraints, and keys of a table in a database. Metadata can also include information about the source, quality, lineage, and usage of a data set.

Metadata is usually stored in a separate system or repository that provides a catalog or directory of your data. This system or repository is called a metadata management system. A metadata management system can help you to store, access, and update your metadata in a centralized and consistent way. The system can also help you to share and communicate your metadata with other users and applications.

Metadata is useful for facilitating the understanding and integration of your data across different sources and systems. Metadata can help you to:

- Discover and explore your data assets and resources
- Document and annotate your data with relevant information and context
- Validate and verify the accuracy and completeness of your data
- Transform and harmonize your data to meet different needs and standards
- Link and relate your data to other data sources and systems.

Log data

Log data is another type of internal data source for your organization. Log data captures the actions and behaviors of users or systems in your organization, such as clicks, views, errors, exceptions, etc. Log data reflects the user experience and system performance of your organization. For example, log data can include information about the web pages that your customers visit, the time they spend on each page, the actions they take, etc. Log data can also include information about the status, errors, and exceptions of your systems, such as CPU usage, memory consumption, response time, etc.

Log data is usually stored in files or streams that record the details and timestamps of each event. Log data can be generated by various sources and systems, such as web servers, applications, databases, devices, etc. Log data can be stored in different formats and levels of granularity, such as plain text, JSON, XML, etc. Log data can also be compressed or encrypted to save space or enhance security.

Log data helps analyze and optimize the user experience and system performance of your organization. Log data can help you to:

- Understand and improve your customer behavior and satisfaction
- Identify and troubleshoot any issues or problems in your systems
- Optimize and enhance your system efficiency and scalability
- Detect and prevent any security threats or attacks on your systems
- Audit and comply with any regulations or standards for your systems.

External data sources

Many times, internal sources must be supplemented by external sources to reach their full value potential. These sources can be from business partners as either suppliers or customers that provide forecasting information that plans and adjustments to plans can be built upon.

It can also be marketplaces that e.g., validate customer information or provide leads. Often benchmark information can also be garnered for marketplaces in the form of reports or statistics. They, however, often come at a significant cost, which makes the public sources interesting.

Public data sources are government or NGO/ association-controlled data that can be accessed free of charge. They do, however, often come with restrictions on usage and the organizations providing them might not present them in an entirely unbiased fashion.

Partners

Partner data is the data that you obtain from your business partners, such as suppliers or customers, that can help you plan and adjust your business activities. For example, partner data can include information about your supplier's inventory, delivery, and pricing. Partner data can also include information about your customer's demands, preferences, and feedback.

Partner data can provide you with many benefits for your decision-making, such as:

- Improving your collaboration and coordination with your partners
- Enhancing your forecasting and planning accuracy and efficiency
- Reducing your risks and uncertainties in your supply chain and market
- Increasing your customer satisfaction and loyalty by providing tailored and timely services
- Creating new value and opportunities from your partner relationships.

However, partner data also comes with some challenges and limitations, such as:

- Ensuring the quality and reliability of the partner data
- Protecting the security and privacy of the partner data
- Integrating and harmonizing the partner data with your internal data
- Managing the cost and complexity of obtaining and maintaining the partner data
- Balancing the benefits and risks of sharing your data with your partners.

You can use partner data to improve your supply chain management in several ways.

Improving your collaboration and coordination with your partners. By sharing information such as inventory, delivery, pricing, demand, preferences, and feedback, you can align your expectations and goals with your partners and work together to optimize your supply chain performance. For example, you can use partner data to synchronize your production and distribution schedules, negotiate better terms and conditions, and resolve any issues or conflicts quickly and effectively.

Enhancing your forecasting and planning accuracy and efficiency. By using partner data to understand the current and future trends and patterns of your supply chain, you can make better decisions and reduce uncertainties and risks. For example, you can use partner data to anticipate the changes in customer demand, adjust your inventory levels, and allocate your resources accordingly.

Reducing your costs and wastes in your supply chain. By using partner data to identify and eliminate any inefficiencies and redundancies in your supply chain processes, you can save time and money and improve your profitability. For example, you can use partner data to streamline your order fulfillment, reduce your transportation and storage expenses, and minimize your inventory obsolescence and spoilage.

Increasing your customer satisfaction and loyalty by providing tailored and timely services. By using partner data to understand and meet the needs and expectations of your customers, you can deliver high-quality products and services that enhance their experience and satisfaction. For example, you can use partner data to customize your products and services, offer flexible delivery options, and provide proactive customer support.

In some cases, your direct access to data is not enough. That is when you turn to marketplaces.

Marketplaces

Data marketplaces are the platforms or services that allow you to buy and sell data from various sources and providers. Data marketplaces can help you access and leverage data that is not available or generated within your organization. For example, data marketplaces can include information about market trends, customer behavior, competitor analysis and intelligence, social media sentiment, etc.

One of the ways that you can use data marketplaces to improve your decision-making is by complementing and enriching your internal data with external data. External data can provide you with additional information and context that can enhance the quality and value of your internal data. For example, you can use external data to fill in the gaps or missing values in your internal data or to add new dimensions or features to your internal data. By doing so, you can increase the completeness and diversity of your data and improve your data analysis and interpretation.

Another way that you can use data marketplaces to improve your decision-making is by validating and verifying your internal data with external data. External data can provide you with independent and objective sources that can help you check the accuracy and reliability of your internal data. For example, you can use external data to confirm or refute your assumptions or hypotheses based on your internal data or to compare and contrast your results or findings with other sources. By doing so, you can increase the confidence and credibility of your data and reduce the errors and biases in your data.

You can also use data marketplaces to improve your decision-making by benchmarking and comparing your performance with external data. External data can provide you with relevant and timely indicators that can help you measure and evaluate your performance against

TABLE 8.1 Benefits and challenges of data marketplaces

Benefits of data marketplaces	Challenges and limitations of data marketplaces
Expanding your data sources and variety	Ensuring the security and privacy of the data
Enhancing your data quality and reliability	Protecting the intellectual property and ownership
Increasing your data value and usability	Integrating and harmonizing with your internal data
Creating new insights and opportunities from data	Evaluating and selecting data providers and products
Reducing your data acquisition and maintenance costs	Balancing the benefits and risks of using external data

your competitors or industry standards. For example, you can use external data to assess your strengths and weaknesses, identify your opportunities and threats, and set your goals and targets based on your performance. By doing so, you can increase the effectiveness and efficiency of your performance and improve your competitive advantage and market position.

Data marketplaces can also be used to improve your decision-making by discovering and exploring new trends and patterns with external data. External data can provide you with novel and diverse perspectives that can help you uncover new insights and knowledge from your data. For example, you can use external data to discover new customer needs and preferences, identify new market segments and niches, or predict future scenarios and outcomes based on current trends. By doing so, you can increase the innovation and creativity of your decision-making and create new value and opportunities for your organization.

Finally, you can use data marketplaces to improve your decision-making by innovating and creating new value from external data. External data can provide you with unique and valuable resources that can help you to develop new products and services, or enhance existing ones, based on your data. For example, you can use external data to customize or personalize your products and services for different customers or markets or to integrate or combine different features or functions from different sources. By doing so, you can increase the differentiation and attractiveness of your products and services and improve your customer satisfaction and loyalty.

Recently a new version of digital marketplace services has sprung up. They provide synthetic data based either on your proprietary data or from other companies. The benefit of synthetic data created from internal data is that you get the volume of data you need for machine-learning algorithms that would otherwise take years to gather. When getting synthetic external data or providing your data externally to sell the benefits, you are now sharing private data, but only the structure of it.

Public sources

Public data sources are data that are collected and published by government or non-governmental organizations (NGOs) and associations and can be accessed largely free of charge by anyone. Public data sources can cover a wide range of topics, such as health, education, environment, economy, society, and more.

Public data sources can provide you with many benefits for your decision-making (Table 8.2):

TABLE 8.2 Benefits and challenges of public data sources

Benefits of public data	Challenges with public data
Enhancing your data quality and reliability	Ensuring the security and privacy of the data
Expanding your data sources and variety	Protecting the intellectual property and ownership of the data
Increasing your data value and usability	Integrating and harmonizing the data with your internal data
Creating new insights and opportunities	Evaluating and selecting the right data providers and products
	Balancing the benefits and risks of using external data

One of the ways that you can use public data sources to improve your decision-making is by complementing and enriching your internal data with external data. External data can provide you with additional information and context that can enhance the quality and value of your internal data. For example, you can use external data to fill in the gaps or missing values in your internal data or to add new dimensions or features to your internal data. By doing so, you can increase the completeness and diversity of your data and improve your data analysis and interpretation.

Another way that you can use public data sources to improve your decision-making is by validating and verifying your internal data with external data. External data can provide you with independent and objective sources that can help you check the accuracy and reliability of your internal data. For example, you can use external data to confirm or refute your assumptions or hypotheses based on your internal data or to compare and contrast your results or findings with other sources. By doing so, you can increase the confidence and credibility of your data and reduce the errors and biases in your data.

You can also use public data sources to improve your decision-making by benchmarking and comparing your performance with external data. External data can provide you with relevant and timely indicators that can help you measure and evaluate your performance against your competitors or industry standards. For example, you can use external data to assess your strengths and weaknesses, identify your opportunities and threats, and set your goals and targets based on your performance. By doing so, you can increase the effectiveness and efficiency of your performance and improve your competitive advantage and market position.

Public data sources can also be used to improve your decision-making by discovering and exploring new trends and patterns with external data. External data can provide you with novel and diverse perspectives that can help you uncover new insights and knowledge from your data. For example, you can use external data to discover new customer needs and preferences, identify new market segments and niches, or predict future scenarios and outcomes based on current trends. By doing so, you can increase the innovation and creativity of your decision-making and create new value and opportunities for your organization.

Some examples of public data sources that you can use to support your decision-making are:

- Data.gov[1] – From science and research to manufacturing and climate, Data.gov is one of the most comprehensive public data sources around the globe. Datasets are available in typical formats such as CSV, JSON, and XML.
- US Census Bureau[2] – For demographical data on US inhabitants, this public source is extremely useful. The source of census bureaus are federal, state, and local governments.
- Data.gov.uk[3] – Similar to Data.gov's source for US data, there's also one for the entire United Kingdom. Reports contain public information about everything from crime and justice to defense and government spending.
- UK Data Service[4] – A perfect complement to Data.gov.uk is the UK Data Service, a search engine for recent datasets on social media trends, politics, finance, and international relations.
- European Union Open Data Portal[5] – With almost 14,000 datasets available, EUROPA is one of the best public providers in the EU for insights on energy, education, commerce, and agriculture.

- Open Data Network[6] – This source allows users to look for public information using a robust search engine. Apply advanced filters to searches and pull information on everything from public safety and finance to infrastructure and housing.
- NASA Exoplanet Archive[7] – Public datasets covering planets and stars gathered by NASA's space exploration missions.
- UN Comtrade Database[8] – Statistics compiled and published by the United Nations on international trade.

These are just some of the many public sources that are available for you to use for free. You can find more public sources online or through platforms like Tableau, G2, or Forbes. By using public sources wisely & responsibly, you can improve your decision-making & achieve better outcomes for yourself and others.

All these data sources are, however, not relevant if we cannot store and access them in a proper system or architecture. That is what the following section is about.

DATA SYSTEM ARCHITECTURE

Data sources are essential for any data-driven organization, but they are not enough to enable effective data analysis and decision-making. Data sources need to be stored, processed, and accessed in a proper system or architecture that can handle the volume, variety, and velocity of data. This system or architecture is known as the data stack.

At the core, we make a separate data stack for analysis and data transformation to ensure that source data is not corrupted or the systems that "own" the data are not loaded by analysis requests that could easily adversely affect performance.

The data stack consists of different layers and components that perform different functions and tasks related to data management and analytics. The data stack can vary depending on the specific needs and goals of each organization, but there are some common elements and patterns that can be identified. In this section, we will describe the modern data stack, which is a collection of cloud-native tools and technologies that are designed to work with large-scale data in a scalable, efficient, and cost-effective way (MongoDB, n.d.).

The modern data stack has four main layers: ingestion layer, storage layer, analysis/exposure layer, and support/governance layer. Each layer has its purpose and challenges and requires different tools and techniques to operate. We will discuss each layer in detail in the following sections.

Ingestion layer (ETL/ELT)

The ingestion layer is the first layer of a data architecture that is responsible for collecting, importing, and processing data from various sources into a target system. The ingestion layer performs several functions, among which are:

- Connecting to external data sources and extracting data from them
- Transforming and validating data according to the business rules and requirements
- Loading data into the storage layer, such as a data lake or a data warehouse

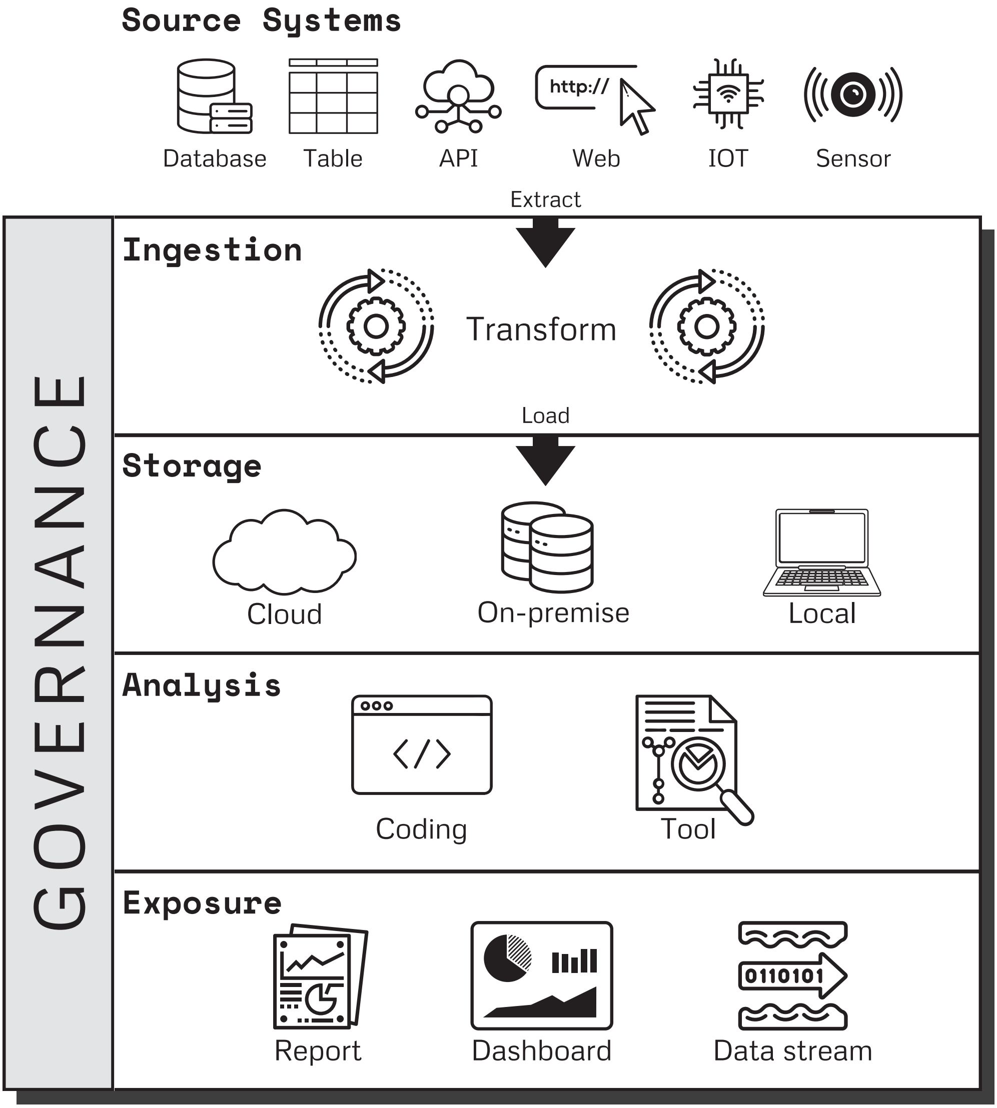


FIGURE 8.1 The modern data stack illustrated

- Scheduling and monitoring data ingestion jobs and pipelines
- Handling errors and exceptions during data ingestion.

The ingestion layer can use different methods and tools to perform these functions, depending on the type, volume, velocity, and variety of data. In general, you can either stream data sources, automatically load at specified intervals (batch load), or load ad hoc which is also a form of batch loading.

Streaming

Ingestion through streaming happens when data is collected and processed in real-time or near-real-time, as it is generated or updated by the source systems. Streaming data is often relevant when the data is continuously produced, and you need to be able to take quick actions. Many companies generate data from sales, production, and logistics continuously, but don't need to take quick actions based on it. Then batch loading is more cost-efficient and easier to manage.

Streaming ingestion can enable faster and more responsive analytics and applications that require timely insights from data but some of the common challenges with streaming data ingestion are:

- Handling the high volume and velocity of data may require scaling the ingestion pipeline using distributed systems and frameworks, such as Apache Spark or Apache Flink, which can parallelize the ingestion and processing of data.
- Dealing with the variety and complexity of data, which may require transforming and validating data according to different schemas and formats, such as JSON, XML, CSV, etc.
- Ensuring the reliability and consistency of data may require implementing error handling and retry mechanisms to handle failures and exceptions during data ingestion, as well as applying deduplication and watermarking techniques to avoid duplicate or out-of-order data.
- Securing and monitoring the data ingestion process may require encrypting data in transit and at rest, authenticating and authorizing access to data sources and target systems, logging and auditing all data ingestion activities and events, and implementing data quality checks at ingest.

Some of the tradeoffs that you should be aware of if you need to handle streaming data in your ingestion layer are that streaming ingestion may incur higher operational costs and complexity than batch ingestion, as it requires more resources and tools to handle the continuous flow of data. Streaming ingestion may also compromise the accuracy and completeness of data, as it may not capture all the changes or updates that occur in the source systems, or it may introduce noise or errors due to network latency or failures. Note that not all systems are natively built to provide streaming data, but rather event data (changes in a state), and the stream is there only when there are changes in values. When working with streaming ingestion you might also get more challenges with data governance and compliance, as it may expose sensitive or personal data to unauthorized parties or violate regulatory requirements for data retention or deletion if these issues are not handled in the design phase. Think in terms of activity/location data that could show where employees or customers are and what they do.

The choice between batch and streaming ingestion depends on various factors, such as the nature of the data, the data volume, the processing requirements, and the need for real-time analysis.

If your data is generated or updated at a high speed and requires low latency processing, streaming ingestion is more appropriate. If your data is generated or updated at a low speed and can tolerate high-latency processing, batch ingestion is more appropriate.

If your data is large and requires efficient and cost-effective processing, batch ingestion is more appropriate. If your data is small and requires fast and responsive processing, streaming ingestion is more appropriate.

If your data is structured or semi-structured and requires complex transformations and validations before loading, ETL-based batch (ad hoc) ingestion is more appropriate. If your data is unstructured or raw and requires minimal transformations or transformations on-demand after loading, ELT-based streaming ingestion is more appropriate.

Data that is consistent and complete and requires high accuracy and completeness for analysis will often make more sense to batch ingest. If your data analysis is not time-sensitive and can be performed offline or on a schedule, batch ingestion is often also more appropriate as batch ingestion is simpler and less costly (Gomede, 2023).

Batch and delta load

Batch ingestion is a method of collecting and processing data in batches or chunks at regular intervals or on demand. Batch ingestion is suitable for data that has high volume and low velocity, such as historical data, transactional data, files, and so on. Batch ingestion often provides the most efficient and cost-effective processing and storage of large amounts of data.

In practical terms, batch ingestion can be done by using the following steps:

TABLE 8.3 Batch loading of data

1	Identify the data sources and the target system for data ingestion	Data sources can be databases, files, APIs, or other systems that generate or store data. The target system can be an SQL database, a data lake, or a data warehouse that can store and analyze data.
2	Define the frequency and schedule of data ingestion	Depending on the business needs and the availability of data, data ingestion can be performed daily, weekly, monthly, or on-demand. For example, a company may want to ingest sales data every day, but customer feedback data every month.
3	Extract the data from the data sources using tools or scripts that can connect to the data sources and retrieve the data	For example, Apache Sqoop is a tool that can extract data from relational databases and transfer it to Hadoop, a distributed system that can store and process large-scale data.
4	Transform the data according to the business rules and requirements using tools or scripts that can manipulate and validate the data	For example, AWS Glue and Azure Data Factory are tools that can transform data using programming scripts, and apply various transformations such as filtering, joining, aggregating, cleansing, etc. (see Chapter 7 for details about transformations).
5	Load the data into the target system using tools or scripts that can transfer and store the data	For example, Apache Nifi is a tool that can load data into various target systems such as Hadoop, Amazon S3, Azure Blob Storage, etc.
6	Load the data into the target system using tools or scripts that can transfer and store the data	For example, Apache Airflow is a tool that can schedule and monitor data ingestion workflows using a web interface.

References: https://sqoop.apache.org/; https://hadoop.apache.org/; https://aws.amazon.com/glue/; www.python.org; www.scala-lang.org/; https://nifi.apache.org/; https://aws.amazon.com/s3/; https://azure.microsoft.com/en-us/services/storage/blobs/; https://airflow.apache.org/.

A version of the batch load is called delta load, which is a method of ingesting where only the changes or updates to the existing data in the target system, rather than the entire data set, are transmitted. Delta load is suitable for incremental or differential data ingestion, where only the new or modified records are transferred from the source systems. Delta load can enable faster and more accurate synchronization of data between the source and target systems. It is more complicated to set up than the full batch copying, but it reduces the system load. Choices about data consistency will, however, have to be made as it can potentially have an impact on the data lineage

ETL and ELT

ETL and ELT are two common methods of data integration that involve extracting data from various sources, transforming it to a suitable format, and loading it to a target system for analysis. The main difference between them is the order and location of the transformation process.

ETL stands for extract, transform, and load. In this method, data is extracted from the source systems, such as databases, files, ERP, CRM, etc., and then transformed on a separate processing server before loading it to the target system, such as a data warehouse. The transformation process can include cleaning, filtering, validating, aggregating, joining, and enriching the data according to the business rules and requirements. Chapter 7 provides details on the specific techniques used for transformation. ETL is more suitable for structured and semi-structured data that require complex transformations and validations before loading. ETL can also ensure data quality and consistency across different sources.

ELT stands for extract, load, and transform. In this method, data is extracted from the source systems and loaded directly to the target system without any transformation. The target system, such as a data lake, stores the raw data as it is from the sources. The transformation process is done later within the target system itself, using its processing power and scalability. The transformation process can be done on-demand or in batches, depending on the analytical needs. ELT is more suitable for unstructured and raw data that requires minimal transformations or transformations on-demand after loading. ELT can also handle large volumes of data and support real-time or near-real-time analysis.

Some examples of ETL and ELT processes are:

- A classic ETL example is the reporting system developed by companies for collecting data to make business decisions. For example, a retail company may use ETL to extract sales data from different stores, transform it to a common format apply business logic, and load it to a data warehouse for generating pre-made reports and dashboards.
- An ELT example is the big data analytics system that leverages cloud computing and distributed storage. For example, a social media company may use ELT to extract user-generated content from various platforms, load it to a data lake in its raw form, and transform it later using tools like Spark or Hadoop for sentiment analysis or recommendation systems.

Security in the ingestion layer

Security is a crucial aspect of data ingestion, as it ensures the confidentiality, integrity, and availability of data from various sources to the target systems. To achieve secure and reliable data ingestion, the ingestion layer needs to follow some best practices, such as:

- Encrypting data in transit and at rest
    - Data should be encrypted both when it is transferred from the source systems to the target systems and when it is stored in the target systems. Encryption keys and certificates should be used to encrypt and decrypt data, and they should be securely managed and rotated. Encryption can also prevent unauthorized access or tampering of data by malicious actors.
- Authenticating and authorizing access to data sources and target systems
    - Data sources and target systems should have proper authentication and authorization mechanisms to verify the identity and permissions of the data ingestion agents or users. Credentials, tokens, or API keys should be used to authenticate access, and they should be securely stored and refreshed.
- Implementing API security best practices
    - If data ingestion involves using APIs to access data sources or target systems, API security best practices should be followed to ensure secure and efficient data exchange. For example, HTTPS protocol should be used to encrypt the communication between the API client and server, input parameters should be validated to prevent injection attacks, API quotas and throttling limits should be checked to prevent denial-of-service attacks, etc.
- Logging and auditing all data ingestion activities and events
    - Data ingestion activities and events should be logged and audited to monitor the performance, status, and errors of the data ingestion process. Logging and auditing can also help identify and troubleshoot any issues or anomalies during data ingestion.
- Implementing error handling and retry mechanism
    - Data ingestion can encounter failures or exceptions due to various reasons, such as network issues, source system unavailability, data format mismatch, etc. Error handling and retry mechanisms should be implemented to handle these failures or exceptions gracefully and resume the data ingestion process.
- Testing and validating data quality and integrity
    - Data quality and integrity are essential for ensuring accurate and reliable data analysis. Data quality and integrity should be tested and validated during or after the data ingestion process. Data quality and integrity tests can include checking for missing values, duplicates, outliers, inconsistencies, etc.

By following these best practices, the ingestion layer can ensure secure and reliable data ingestion from various sources to the target systems for analysis.

Good ingestion practices will help in managing the storage layer structure where data is placed in the correct locations in the right formats enabling the retrieval for (further) transformation and analysis.

Storage layer

Data storage systems are the backbone of the modern data stack, as they store and organize the data for analysis. There are three main types of data storage systems: data warehouses, data lakes, and data meshes. Each of them has different features and benefits, depending on the data structure, schema, scalability, performance, and cost.

Data warehouses: Centralized repositories of structured and processed data that are optimized for analytical queries. They use a predefined schema that defines the data model and the relationships between the tables. Data warehouses are highly scalable, as they can handle large volumes of data and support concurrent users. They also offer high performance, as they use techniques such as indexing, partitioning, and compression to speed up the queries. However, data warehouses can be costly to maintain and update, as they require a lot of resources and expertise. They also have limited flexibility, as they cannot easily accommodate new or changing data sources or formats. If they are connected to a data lake, they are usually referred to as lake-houses.

Data lakes: Distributed collections of raw and unprocessed data that are stored in their native format. They use a schema-on-read approach that allows the users to define the schema at the time of analysis. Data lakes are very flexible, as they can store any type of data, such as structured, semi-structured, or unstructured data. They also have low cost, as they use cheap and scalable storage systems such as Cloud or Hadoop. However, data lakes can have low quality, as they do not enforce any validation or standardization of the data. They also have low performance, as they require a lot of processing and transformation to make the data ready for analysis.

Data meshes: Decentralized networks of domain-specific data platforms that are owned and governed by different teams or units within an organization. They use a schema-on-write approach that ensures the data is consistent and interoperable across the domains. Data meshes are very agile, as they enable faster and easier data sharing and collaboration among the teams. They also have high quality, as they apply common standards and policies for data governance and security. However, data meshes can be complex to implement and manage, as they require a lot of coordination and communication among the teams. They also have high overheads, as they involve a lot of duplication and replication of the data.

A data swamp is a term used to describe a data lake that has become unmanageable and unusable due to poor data quality, governance, and organization. A data swamp can result from a lack of processes and standards for ingesting, storing, and analyzing the data in the data lake. Data in a data swamp is difficult to find, manipulate, and extract value from. A data swamp can also arise from a data warehouse that has hybrid models with unstructured and ungoverned data (Lauer, 2021). To avoid or fix a data swamp, you need to implement proper data governance and curation measures, such as defining the data model, schema, metadata, quality, security, and lifecycle. You also need to have clear roles and responsibilities for managing and accessing the data, such as a chief data officer or a product owner. Many of these things are in the governance/support layer.

TABLE 8.4 Comparison of storage systems in the modern data stack

Data storage system	Advantages	Disadvantages	Best practices	Challenges
Data warehouse	– Optimized for analytical queries – Highly scalable and performant – Supports concurrent users	– Costly to maintain and update – Limited flexibility for new or changing data sources or formats	– Define a clear and consistent data model and schema – Use ETL (Extract, Transform, Load) processes to ingest and process data – Monitor and optimize the performance and resource utilization	– Handling complex and diverse data sources and formats – Ensuring data freshness and accuracy – Balancing the trade-off between performance and cost
Data lake	– Flexible for any type of data – Low-cost and scalable storage – Supports multiple analytical tools and frameworks	– Low data quality and reliability – Low performance and efficiency – Difficult to find and access relevant data	– Define and enforce data quality standards and validation rules – Use ELT (Extract, Load, Transform) processes to ingest raw data and transform on–demand – Use a data catalog or metadata management system to organize and discover data	– Managing data security and governance across multiple domains and users – Processing and transforming large volumes of raw data – Integrating data from different sources and formats
Data mesh	– Agile and collaborative data sharing across teams – High data quality and interoperability – Supports domain-specific data platforms and needs	– Complex to implement and manage – High overhead and duplication of data – Requires a lot of coordination and communication among teams	– Define a common data model, schema, metadata, governance, and security standards across domains – Use a federated query engine or API to access data from different domains – Establish clear roles and responsibilities for data ownership and stewardship	– Aligning the goals and incentives of different teams and units – Ensuring data consistency and compatibility across domains – Resolving data conflicts and issues

From the storage layer, the data is pulled into analytics tools and data products. Sometimes by programming and sometimes using visualization tools like PowerBI, Tableau, etc.

Analysis/exposure layer

When the data has been extracted from the sources through the ingestion layer it is stored for analysis in the storage layer. To make the data useful for driving decisions it must now be transformed to insights and presented in a way that enables decision support. This is what the DECAS model terms as analytics and one of the two additions that makes decision-making data-driven.

Programming languages

There are currently 2–3 programming languages that are popular for data analysis. Python has its strengths in machine learning and R in statistical learning. Julia is an upcoming programming language that is developed to be the best of both worlds.

Python is a versatile and widely adopted programming language known for its simplicity and readability. It has a rich ecosystem of libraries, making it an excellent choice for general-purpose data analysis and machine-learning tasks. Some of the most important packages for data analysis in Python are:

- NumPy: A library for working with multidimensional arrays and matrices, providing high-performance numerical computation and linear algebra operations.
- Pandas: A library for data manipulation and analysis, offering data structures and operations for handling tabular, time series, and multidimensional data.
- Scikit-learn: A library for machine learning, providing a range of supervised and unsupervised learning algorithms, as well as tools for model selection, evaluation, and preprocessing.
- Matplotlib: A library for data visualization, offering various types of plots, such as line, bar, scatter, histogram, etc., as well as interactive features and customization options.

Python is an ideal choice for data science projects that require a wide range of data manipulation, analysis, and machine learning capabilities. Its simplicity, extensive library ecosystem, and strong community support make it particularly suitable for beginners and interdisciplinary collaborations. Python is often preferred for tasks such as exploratory data analysis, web scraping, natural language processing, and building machine learning models. However, Python has some limitations in terms of performance, concurrency, and advanced techniques. Python's interpreted nature can result in slower execution speed compared to compiled languages. Python's global interpreter lock (GIL) restricts multi-threading. Python's object-oriented programming (OOP) paradigm may require more effort to master complex concepts like parallel processing or memory optimization.

R is a statistically focused programming language that was originally developed for statisticians. It has a strong focus on data analysis and visualization, offering a wide range of statistical methods and graphical techniques. Some of the most important packages for data analysis in R are:

- Tidyverse: A collection of packages that share a common philosophy of data manipulation and visualization, based on the principles of tidy data and functional programming.
- Dplyr: A package for data manipulation, providing a consistent set of verbs for filtering, selecting, grouping, summarizing, and joining data frames.
- Ggplot2: A package for data visualization, based on the grammar of graphics, allowing users to create complex and elegant plots using layers of aesthetic mappings and geometric objects.

R's statistical focus and visualization capabilities make it the language of choice for statisticians and researchers. R offers a comprehensive set of statistical methods and graphical techniques that can handle complex and specialized data analysis problems. R is often preferred for tasks such as hypothesis testing, regression modeling, clustering analysis, or creating publication-quality plots. However, R has some drawbacks in terms of scalability, performance, and general-purpose programming. R can struggle with large datasets or computationally intensive tasks due to its memory management issues, which means the entire dataset must be able to fit in the RAM. R can have lower performance than other languages due to its vectorized nature or lack of optimization. R can be less intuitive or consistent than other languages due to its functional programming style or multiple ways of doing the same thing. However, the newest research is published as R-packages.

Julia is a relatively new programming language that was designed for high-performance numerical computing. It combines the ease of use and expressiveness of scripting languages with the speed and efficiency of compiled languages. Some packages for data analysis to start with in Julia are:

- DataFrames: A package for working with tabular data, providing data structures and operations similar to those in Pandas or Dplyr.
- Plots: A package for data visualization, providing a high-level interface to various backends, such as GR, PyPlot, or Plotly.
- MLJ: A package for machine learning, providing a unified framework for composing, fitting, tuning, and evaluating machine learning models from various sources.
- Flux: A package for deep learning, providing a flexible and intuitive way to define and train neural networks using automatic differentiation. Similar to the Tensorflow package.

Julia's high-performance numerical computing makes it the language of choice for engineers and scientists. Julia combines the ease of use and expressiveness of scripting languages with the speed and efficiency of compiled languages. Julia is often preferred for tasks such as numerical simulation, optimization, or differential equations. Julia also supports multiple dispatch, metaprogramming, and macros, which enable powerful abstractions and code generation. However, Julia has some challenges in terms of maturity, stability, and compatibility. Julia is still a relatively new language and may have fewer libraries or resources than other languages. That is highlighted by the most popular package in January 2024 only having 10,000-star ratings (https://juliapackages.com/packages), whereas the NumPy python has more than 25,000 stars and has branched into more than 9,000 versions.

No matter which language is chosen, the output will be embedded in another product. That could be back into the source systems where data originated or new system/data products. This can be a complicated process and visualization tools are therefore often used to facilitate the interpretation of data.

Visualization tools

While the programming languages can provide visualization in the analysis/exposure layer, other tools are more useful/faster for this purpose. Visualization tools are software applications that allow users to create and interact with graphical representations of data. They can help users to explore, understand, and communicate data insights intuitively and engagingly. Visualization tools can also support data-driven decision-making by enabling users to discover patterns, trends, outliers, and relationships in data.

The three major ones are PowerBI from Microsoft, Tableau from Salesforce, and Looker Studio from Google.

Common for the tools is that they only provide simple statistical and ML analytical tools, but all are good at facilitating the explorative data analysis.

They have some common features and functionalities, such as:

- Connecting to various data sources, such as databases, files, web services, and cloud platforms
- Providing a drag-and-drop interface for building charts, dashboards, and reports without requiring coding skills
- Offering a range of visualization options, such as bar charts, line charts, pie charts, maps, scatter plots, and more
- Supporting interactivity, such as filtering, sorting, slicing, and drilling down into data
- Allow users to share and collaborate on their visualizations with others, either online or offline.

However, each tool also has its strengths and weaknesses, depending on the user's needs and preferences.

PowerBI: A business analytics service that is part of the Microsoft ecosystem. It integrates well with other Microsoft products, such as Excel, SharePoint, and Azure cloud services. It also has a large and active community of users and developers who provide support and resources. PowerBI has a free version for individual users and a paid version for organizations. It is suitable for users who want a low-cost and easy-to-use tool that can leverage the power of Microsoft. Be aware that PowerBI in the desktop version does not support Mac computers and is being discontinued. The online version has limitations in functionality even though it is a focus for Microsoft to upgrade it during 2024 and 2025.

Tableau: A data visualization software that is owned by Salesforce. It is known for its high performance, flexibility, and aesthetics. It can handle large and complex data sets and create stunning and sophisticated visualizations. Tableau has a desktop version for creating and publishing visualizations, a server version for hosting and managing them, and an online version for accessing them from anywhere. It is suitable for users who want a powerful and professional tool that can handle most visualizations. It supports more visualization types than PowerBI.

Looker Studio: A business intelligence platform that is owned by Google. It is based on Google's own modeling language called LookML which defines the data structure, logic, and calculations. It can connect to any SQL database and generate SQL queries on the fly. It also integrates with Google Cloud Platform and other Google services, such as BigQuery, Google Analytics, and Sheets. Looker has a web-based interface that can create and embed visualizations in any application. It is suitable for users who want a scalable and customizable tool that can support advanced analytics and data science. That it requires you to learn another modeling language has kept it from being very broadly supported, but the link with the entire Google Analytics estate is a big benefit.

In summary, visualization tools are essential for data analysis and data-driven decision-making. They can help users explore, understand, and communicate data insights intuitively and engagingly. PowerBI, Tableau, and Looker are three of the most popular and widely used visualization tools in the market. They have some common features and functionalities, but also some differences that make them suitable for different use cases and scenarios. You should choose the tool that best fits your data needs and communication goals. Chapter 6 explores data visualization in much further detail.

Governance and support layer

The governance and support layer of the modern data stack is the layer that ensures the quality, security, and compliance of the data and the data processes. It has three main components: Data governance, data security, and data quality.

Data governance: Data governance is the set of policies, standards, and procedures that define how data is collected, stored, accessed, and used in an organization. It also involves the roles and responsibilities of the data stakeholders, such as data owners, data stewards, data consumers, and data analysts. Data governance helps to ensure that data is accurate, consistent, reliable, and trustworthy. It also helps to align data with the business goals and strategies and to comply with the legal and ethical regulations. What should be kept where, how, and in which format?

Data security: Data security is the set of measures that protect data from unauthorized access, modification, or deletion. It also involves the encryption, masking, and anonymization of sensitive data, such as personal information, financial data, or health records. Data security helps to prevent data breaches, data leaks, and data theft, and to safeguard the privacy and confidentiality of the data. Essentially who should have access to what, when.

Data quality: Data quality is the degree to which data meets the expectations and requirements of the data users. It also involves the validation, verification, cleansing, and enrichment of data, such as removing duplicates, correcting errors, filling missing values, and adding metadata. Data quality management helps to improve the usability, reliability, and value of the data, and to avoid data errors, data inconsistencies, and data anomalies. It is the control mechanism on top of the load and transformation process in the ingestion layer.

The governance and support layer of the modern data stack is the layer that ensures the quality, security, and compliance of the data and the data processes. It includes data governance, data security, and data quality components, which help to manage, protect, and improve the data. This is done by controlling the other three layers of the modern data stack.

Understanding the data stack components is, however, only the beginning. If it is not built and provided for you, you will have to have it built. It can be a daunting task and good consultants will probably be required but to ensure that they propose and build the right setup you will have to provide some initial system requirements.

CREATING SYSTEM REQUIREMENTS

System requirements are the specifications and conditions that define what a system should do, how it should do it, and how well it should do it. They are essential for designing, developing, testing, and deploying a system that meets the needs and expectations of the stakeholders, such as the users, customers, developers, and managers. System requirements can also help to evaluate the feasibility, cost, and risk of a system project, and to ensure its quality, reliability, and performance.

System requirements can be classified into different types and categories, depending on their nature, scope, and priority. Two of the most common and widely used methods for classifying system requirements are FURPS and MoSCoW. These methods can help to organize, prioritize, and communicate system requirements clearly and consistently.

In the following sections, FURPS and MoSCoW will be explained along with how they can be applied to data system requirements. We will also provide some examples and best practices for using these methods in data engineering and data-driven decision-making.

FURPS+

FURPS is an acronym that stands for Functionality, Usability, Reliability, Performance, And Supportability. These are the five main categories of system requirements that describe the features and characteristics of a system. FURPS can help to identify, define, and prioritize system requirements in a systematic and structured way. FURPS can also help to evaluate and compare different system alternatives and solutions. The + was added to highlight the importance of considering constraints, interface requirements, and business rules (Prieto-González et al., 2016).

Functionality is the category of system requirements that specifies what the system should do, or the functions and capabilities that the system should provide. Functionality requirements can be expressed as use cases, user stories, scenarios, or specifications. They can also be classified into functional and non-functional requirements, depending on whether they describe the behavior or the quality of the system.

Usability is the category of system requirements that specifies how the system should interact with the users, or the ease of use and user satisfaction that the system should provide. Usability requirements can be expressed as user interface design, user experience design, user feedback, or user testing. They can also be classified into user and stakeholder requirements, depending on whether they describe the needs or the expectations of the users. This is made more explicit in the FURPS+ upgraded model, where the interface elements are separate.

Reliability is the category of system requirements that specifies how the system should perform under normal and abnormal conditions, or the dependability and robustness that the system should provide. Reliability requirements can be expressed as availability, fault tolerance, recoverability, or maintainability. They can also be classified into operational and environmental requirements, depending on whether they describe the conditions or the constraints of the system.

Performance is the category of system requirements that specifies how the system should respond to the user requests and the system load, or the efficiency and effectiveness that the system should provide. Performance requirements can be expressed as speed, scalability, capacity, or throughput. They can also be classified into quantitative and qualitative requirements, depending on whether they describe the measurable or the perceptible aspects of the system.

Supportability is the category of system requirements that specify how the system should evolve and adapt to the changing needs and expectations of the users and the stakeholders, or the extensibility and flexibility that the system should provide. Supportability requirements can be expressed as compatibility, modularity, interoperability, or portability. They can also be classified into evolutionary and adaptive requirements, depending on whether they describe the changes or the adjustments of the system.

TABLE 8.5 Examples of functionality requirements for a modern data stack or a data product

- The system should be able to ingest data from multiple sources, such as files, databases, web services, and cloud platforms.
- The system should be able to transform, clean, and enrich the data using various methods, such as SQL, Python, R, or Spark.
- The system should be able to store and manage the data in a secure, scalable, and cost-effective way, using various technologies, such as data warehouses, data lakes, or data meshes.
- The system should be able to expose and analyze the data using various tools, such as dashboards, reports, charts, or machine learning models.
- The system should be able to govern and support the data and the data processes, using various components, such as data catalog, data lineage, data quality, or data security.

TABLE 8.6 Examples of usability requirements for a modern data stack or a data product

- The system should have a user-friendly and intuitive interface that allows users to access, explore, and manipulate the data with minimal effort and training.
- The system should have a consistent and coherent design that follows the best practices and standards of data visualization and data communication.
- The system should have a responsive and adaptive design that supports different devices, platforms, and browsers and adjusts to different screen sizes, resolutions, and orientations.
- The system should have a personalized and customizable design that allows users to configure, customize, and save their preferences, settings, and views.
- The system should have a feedback and support mechanism that allows users to report issues, request features, and receive help and guidance.

TABLE 8.7 Examples of reliability requirements for a modern data stack or a data product

- The system should be available and accessible at all times, or have a high uptime and low downtime.
- The system should be able to handle errors, failures, and exceptions gracefully, or have a high fault tolerance and low failure rate.
- The system should be able to restore its normal operation quickly and easily or have high recoverability and low recovery time.
- The system should be able to update, upgrade, and repair its components and functionalities, or have a high maintainability and low maintenance cost.

TABLE 8.8 Examples of performance requirements for a modern data stack or a data product

- The system should be able to process the data and deliver the results promptly or have a high speed and low latency.
- The system should be able to handle the increasing amount and complexity of the data and the users or have a high scalability and low scalability limit.
- The system should be able to store and manage large and diverse data sets or have a high-capacity and low-capacity limit.
- The system should be able to handle concurrent and parallel requests and operations or have a high throughput and low throughput limit.

TABLE 8.9 Examples of supportability requirements for a modern data stack or a data product

- The system should be able to work with different versions and formats of the data and the tools or have high compatibility and low compatibility issues.
- The system should be able to add, remove, or modify its components and functionalities, or have a high modularity and low coupling.
- The system should be able to communicate and integrate with other systems and applications or have a high interoperability and low integration cost.
- The system should be able to migrate and deploy to different environments and platforms or have a high portability and low migration cost.

In summary, FURPS is a method for classifying system requirements into five main categories: Functionality, Usability, Reliability, Performance, and Supportability. These categories describe the features and characteristics of a system that meet the needs and expectations of the users and the stakeholders. FURPS can help to identify, define, and prioritize system requirements in a systematic and structured way. FURPS can also help to evaluate and compare different system alternatives and solutions. FURPS can be applied to system requirements for a modern data stack or a data product, as shown by the examples above.

Most of the examples provided are high-level requirements, but the more specific you can get, maybe through use-cases, the better a solution you will get.

MoSCoW

MoSCoW is an acronym that stands for "Must have", "Should have", "Could have", and "Won't have**"**. These are the four levels of priority that can be assigned to system requirements, depending on their importance and urgency. MoSCoW can help to prioritize and balance the system requirements. MoSCoW can also help to manage the scope, time, and resources of a system project, and to ensure its delivery and quality. It links well with the idea of a backlog in the agile development methods, where the most important functions are prioritized, and if time and resources allow it the "could have" functionalities are added.

MoSCoW works as follows:

- Must have: These are the system requirements that are essential and critical for the system to function and to meet the objectives and expectations of the users and the stakeholders. They are the minimum and non-negotiable requirements that must be delivered and satisfied, otherwise the system will fail or be rejected. They are the highest priority requirements that should be addressed and implemented first, and should not be compromised or sacrificed for any reason.
- Should have: These are the system requirements that are important and beneficial for the system to perform and to enhance the value and satisfaction of the users and the stakeholders. They are the desirable and optimal requirements that should be delivered and fulfilled unless some valid reasons or constraints prevent them. They are the high-priority requirements that should be addressed and implemented next, and should only be compromised or sacrificed with careful consideration and justification.
- Could have: These are the system requirements that are useful and nice for the system to improve and to increase the appeal and delight of the users and the stakeholders. They are the optional and additional requirements that could be delivered and met if there are enough time and resources available, and if they do not affect the other requirements. They are the low-priority requirements that could be addressed and implemented later and could be compromised or sacrificed with minimal impact and regret.
- Won't have: These are the system requirements that are irrelevant and unnecessary for the system to operate and to satisfy the needs and expectations of the users and the stakeholders. They are the excluded and rejected requirements that won't be delivered or considered, either because they are out of scope, out of budget, out of time, or out of alignment with the goals and strategies of the system. They are the lowest priority requirements that should not be addressed or implemented at all and should be removed or postponed to future releases or versions.

MoSCoW can be applied to system requirements for a modern data stack or a data product, as shown by the examples below. The examples are based on the FURPS categories and examples given in the previous answer, but they are not exhaustive or definitive, and they may vary depending on the context and situation of the system project.

Must have: These are the system requirements that are essential and critical for a modern data stack or a data product, such as:

- The system should be able to ingest data from multiple sources, such as files, databases, web services, and cloud platforms. (Functionality)

- The system should have a user-friendly and intuitive interface that allows users to access, explore, and manipulate the data with minimal effort and training. (Usability)
- The system should be available and accessible at all times, or have a high uptime and low downtime. (Reliability)
- The system should be able to process the data and deliver the results promptly, or have a high speed and low latency. (Performance)
- The system should be able to work with different versions and formats of the data and the tools or have high compatibility and low compatibility issues. (Supportability)

Should have: These are the system requirements that are important and beneficial for a modern data stack or a data product, such as:

- The system should be able to transform, clean, and enrich the data using various methods, such as SQL, Python, R, or Spark. (Functionality)
- The system should have a consistent and coherent design that follows the best practices and standards of data visualization and data communication. (Usability)
- The system should be able to handle errors, failures, and exceptions gracefully, or have a high fault tolerance and low failure rate. (Reliability)
- The system should be able to handle the increasing amount and complexity of the data and the users or have a high scalability and low scalability limit. (Performance)
- The system should be able to add, remove, or modify its components and functionalities, or have a high modularity and low coupling. (Supportability)

Could have: These are the system requirements that are useful and nice for a modern data stack or a data product, such as:

- The system should be able to expose and analyze the data using various tools, such as dashboards, reports, charts, or machine learning models. (Functionality)
- The system should have a responsive and adaptive design that supports different devices, platforms, and browsers and adjusts to different screen sizes, resolutions, and orientations. (Usability)
- The system should be able to restore its normal operation quickly and easily or have high recoverability and low recovery time. (Reliability)
- The system should be able to store and manage large and diverse data sets or have a high capacity and low capacity limit. (Performance)
- The system should be able to communicate and integrate with other systems and applications or have a high interoperability and low integration cost. (Supportability)

Won't have: These are the system requirements that are irrelevant and unnecessary for a modern data stack or a data product, such as:

- The system should be able to govern and support the data and the data processes, using various components, such as data catalog, data lineage, data quality, or data security.

(Functionality) (Note: This requirement may be out of scope or out of budget for a system project that focuses on data engineering and data-driven decision-making, and not on data governance and data support. However, this requirement may be relevant and necessary for a system project that has a different scope or budget, or that involves data governance and data support as part of its objectives and expectations.)

- The system should have a personalized and customizable design that allows users to configure, customize, and save their preferences, settings, and views. (Usability) (Note: This requirement may be out of time or out of alignment with the goals and strategies of a system project that prioritizes data exploration and data analysis, and not data personalization and data customization. However, this requirement may be useful and nice for a system project that has more time or that involves data personalization and data customization as part of its value and satisfaction.)
- The system should be able to update, upgrade, and repair its components and functionalities, or have a high maintainability and low maintenance cost. (Reliability) (Note: This requirement may be out of budget or out of time for a system project that has a limited or fixed duration and cost, and not a continuous or variable one. However, this requirement may be important and beneficial for a system project that has a longer or flexible duration and cost, or that requires a high maintainability and low maintenance cost as part of its dependability and robustness.)
- The system should be able to handle concurrent and parallel requests and operations or have a high throughput and low throughput limit. (Performance) (Note: This requirement may be out of scope or out of alignment with the goals and strategies of a system project that targets a small or moderate number of users and requests, and not a large or massive one. However, this requirement may be useful and nice for a system project that targets a larger or higher number of users and requests, or that requires a high throughput and low throughput limit as part of its efficiency and effectiveness.)
- The system should be able to migrate and deploy to different environments and platforms or have a high portability and low migration cost. (Supportability) (Note: This requirement may be out of budget or out of time for a system project that has a specific or predefined environment and platform, and not a generic or flexible one. However, this requirement may be useful and nice for a system project that has a more diverse or dynamic environment and platform, or that requires a high portability and low migration cost as part of its extensibility and flexibility.)

MoSCoW is a method for prioritizing system requirements into four levels of priority: Must have, Should have, Could have, and Won't have. These levels of priority describe the importance and urgency of system requirements, and how they affect the delivery and quality of the system. MoSCoW can help to prioritize and balance the system requirements simply and effectively. MoSCoW can also help to manage the scope, time, and resources of a system project, and to ensure its delivery and quality. MoSCoW can be applied to system requirements for a modern data stack or a data product, as shown by the examples above.

FURPS and MoSCoW are often combined for added value in Table 8.10.

Try to fill it with the information from the previous example.

TABLE 8.10 Combining FURPS and MoSCoW

	Must have	Should have	Could have	Won't have
Functional				
Usability				
Reliability				
Performance				
Supportability				

SUMMARY

Having a robust data infrastructure is imperative for organizations aiming to become data-driven and make decisions based on facts rather than intuitions. This chapter provides a comprehensive overview of how to design, implement, and manage modern data architectures that allow efficient ingestion, storage, processing, analysis, and delivery of data at scale.

A high-level data infrastructure comprises four key layers:

1 Ingestion layer: Responsible for extracting data from various sources like operational databases, cloud platforms, social media APIs, devices, etc., and ingesting it into the system. Supports both batch and real-time, streaming ingestion. Handles tasks like data validation, transformation, scheduling, monitoring, and error handling.

2 Storage layer: Stores and organizes ingested data for downstream analytics and applications. Three main architectures are:

- Data warehouses: Optimized for analytical workloads via SQL.
- Data lakes: Flexible storage layer that can handle both structured and unstructured data
- Data meshes: Decentralized data platforms owned by domains rather than centralized IT.

3 Analysis layer: Responsible for processing, analyzing, and transforming stored data into insights using tools like BI platforms, data science notebooks, dashboards, and machine learning models.

4 Governance layer: Focuses on managing the quality, security, privacy, and compliance of data via metadata management, access controls, encryption, etc.

To build an appropriate data infrastructure, robust requirements gathering is crucial early on. Two useful prioritization techniques are FURPS and MoSCoW:

FURPS classifies requirements into Functionality, Usability, Reliability, Performance, and Supportability. Helps identify technical and quality attributes.

MoSCoW labels requirements as Must Have, Should Have, Could Have, or Won't Have. Allows orderly prioritization of most critical vs. nice-to-have features.

By taking a holistic, proactive approach, the challenges can be adequately mitigated.

A reliable data infrastructure forms the backbone of data-driven decision-making in modern organizations. Carefully designing architectures, choosing appropriate technologies, gathering precise requirements, and proactively addressing technical and data quality challenges can lead to high ROI and sustained competitive advantage for companies.

CASE: BKASH, A FINTECH COMPANY IN BANGLADESH AND INTELLIGENT MACHINES

bKash is a leading mobile financial service provider (mobile fintech) in Bangladesh, offering a broad spectrum of financial solutions to its users. The services include sending money, mobile recharge, payment, cash out, adding money, bill payment, etc. The bKash app, which is available in both Bangla and English, makes these transactions fast, easy, and safe. The platform has been recognized as the nation's no. 1 brand for five consecutive years, reflecting its popularity and widespread use.

In addition to individual users, bKash also caters to businesses by providing them with fast, easy, and safe transaction solutions. Whether it's for personal use or business, bKash has revolutionized the way financial transactions are conducted in Bangladesh, making it a vital part of the country's digital economy.

It has, however, not always been easy to run an operation like this in a country known for cheap garments and floods. As the operations scaled and reached thousands of outlets where people could conduct business there was a significant need for smart and intelligent solutions. At that time Md. Oli Ahad, CEO of the young local IT company, Intelligent Machines reached out to them. He had an idea of using mobile devices with apps that could take pictures with geo-location tagging to validate the outlets quickly, safely, and efficiently.

During the next few months, they got the specifications made for a solution where the bKash agents could go out and validate the marketing materials at the outlets. If the outlets had all banners and point-of-sale (POS) marketing materials in place there would be a bonus for the checker and the outlet. They also had checks in place to automatically do image recognition, validate the location against the database of outlets, and link to the salary system for logging the bonuses. In addition, the management of bKash needed reports at set intervals to track the developments.

Provide a suggestion for a FURPS/MoSCoW matrix that describes the solution that Md. Oli Ahad from Intelligent Machines provided along with suggestions for what else could be provided.

The solution was a great success and enabled bKash to scale their business to 800,000 outlets without having to significantly increase the number of staff employed to check the outlets. This of course also requires a significant and complicated data stack. They decided early on, on a modern cloud solution.

Based on the information in the case, describe the data stack in the "modern data stack" framework.

KEY TERMS

Analytics layer: Transforms data into insights and intelligence through reporting, visualization, and advanced analytics.

Batch ingestion: Ingesting data in batches or chunks at regular intervals. Efficient for high-volume, low-velocity data.

Continuous integration: Merging developer code changes frequently into a shared code repository automatically to enable rapid testing and deployment.

Data catalog: An organized inventory of available data sets with enriching usage statistics and business context.

Data governance: Policies and guidelines for managing the security, quality, and lifecycle of data assets within an organization.

Data infrastructure: The technology, systems, and processes for ingesting, storing, managing, processing, analyzing, and delivering data.

Data lake: Flexible storage layer that can store large volumes of structured and unstructured data.

Data SLAs (service level agreements): Commitments for metrics like data accuracy, timeliness, and availability, delivered as per business needs.

Data warehouse: Centralized repository optimized for analytical queries for business intelligence.

DevOps: Integrating software development (Dev) and IT operations (Ops) to enable faster release cycles and ensure quality.

ELT (extract, load, transform): Ingestion process to extract raw data, load it directly into a target system, then transform.

ETL (extract, transform, load): Ingestion process to extract data, transform it, then load into target database or warehouse.

ETL/ELT pipelines: Automated workflows to schedule, sequence, and monitor batch data integration jobs for ingestion into warehouses.

FURPS: Method to classify functional and non-functional system requirements (Functionality, Usability, Reliability, Performance, and Supportability).

Ingestion layer: Component for collecting and importing data from sources into a target system.

Integration testing: Testing interactions between components and modules that have been unit tested.

Mesh: Decentralized data architecture with domain-oriented self-serve data platforms.

Metadata management: Tools, processes, and governance for standardized definitions, usage stats, and system-level metadata.

MoSCoW: Requirements prioritization technique (Must Have, Should Have, Could Have, Won't Have).

Observability: Enabling deep inspection and understanding of systems based on aggregated metrics, logs, and traces.

Streaming ingestion: Ingesting streaming data in real-time or near real-time from source systems. Enables low latency processing.

Test automation: Automating testing activities by scripting and replaying test scenarios without manual effort.

Unit testing: Testing isolated software components like functions, and classes individually using sample input/output.

Use case: High-level depiction of user goals and interactions with a system to achieve a desired outcome.

User acceptance testing: Getting feedback from target users to ensure software meets business and compliance needs.

REVIEW QUESTIONS

Here are 20 review questions for the chapter with answers provided at the end:

1 What are the four main layers of a modern data architecture?
2 What layer of the data architecture focuses on collecting and importing data from sources?
3 What ingestion method works well for large volumes of low-velocity data?
4 What ingestion method enables real-time processing with low latency?
5 What is the difference between ETL and ELT for data ingestion?
6 Which data storage system optimizes storage and querying for analytics use cases?
7 Which data storage system can store large volumes of structured, semi-structured, and unstructured data?
8 What type of data architecture decentralizes data platforms across domains?
9 What layer of the architecture transforms data into insights and intelligence?
10 What layer of the architecture manages the quality, security, and compliance of data?
11 What requirements-gathering technique focuses on Functionality, Usability, Reliability, Performance, and Supportability?
12 What requirements prioritization method uses the MoSCoW categories?
13 What are two common challenges faced while implementing data infrastructures?
14 What technique can help schedule, sequence, and monitor batch data integration jobs?
15 What specifies metrics like data accuracy, availability, and timeliness committed to business users?
16 What type of testing focuses on testing isolated software components?
17 What type of testing verifies interactions between integrated components?
18 What type of testing gets feedback from end users to ensure software meets needs?
19 What approach integrates software development and IT operations activities?
20 What enables deep inspection of systems using aggregated metrics, logs, and traces?

Answers to review questions

1 Ingestion, Storage, Analysis, Governance
2 Ingestion layer
3 Batch ingestion
4 Streaming ingestion
5 ETL transforms before loading, ELT transforms after loading
6 Data warehouse
7 Data lake
8 Data mesh
9 Analysis layer
10 Governance layer
11 FURPS
12 MoSCoW
13 Data integration, Scalability
14 ETL/ELT Pipelines

15 Data SLAs
16 Unit Testing
17 Integration Testing
18 User Acceptance Testing
19 DevOps
20 Observability

NOTES

[1] www.data.gov/
[2] www.census.gov/
[3] https://data.gov.uk/
[4] www.ukdataservice.ac.uk/
[5] https://data.europa.eu/euodp/en/home
[6] www.opendatanetwork.com/
[7] https://exoplanetarchive.ipac.caltech.edu/
[8] https://comtrade.un.org/

BIBLIOGRAPHY

Gomede, E. (2023, August 3). *Batch vs. streaming data ingestion: Choosing the right approach for efficient data processing*. Medium. https://medium.com/@evertongomede/batch-vs-streaming-data-ingestion-choosing-the-right-approach-for-efficient-data-processing-8fa492299dd4

Lauer, C. (2021, December 28). *What is a data swamp? Reasons for why you should avoid it*. Medium. https://medium.com/codex/what-is-a-data-swamp-38b1aed54dc6

Marr, B. (2018, February 26). *Big Data and AI: 30 amazing (and free) public data sources for 2018*. Forbes. www.forbes.com/sites/bernardmarr/2018/02/26/big-data-and-ai-30-amazing-and-free-public-data-sources-for-2018/

MongoDB. (n.d.). *What is a data stack? Modern data stack explained*. MongoDB. www.mongodb.com/basics/data-stack (Accessed January 16, 2024).

Pickell, D. (209, March 15). *50 best open data sources ready to be used right now*. G2 Learning. https://learn.g2.com/open-data-sources

Prieto-González, L., Tamm, G., & Stantchev, V. (2016). Towards a software engineering approach for Cloud and IoT. *Services in Healthcare*, 9789, 439–452. Conference paper. https://doi.org/10.1007/978-3-319-42089-9_31

Tableau.com (n.d.). Free public data sets for analysis. www.tableau.com/learn/articles/free-public-data-sets

CHAPTER 9

Data Ethics

How to Ensure the Data Practices Are Responsible, Secure, and Legal

Imagination is what sets the limits for what you *can* do with data analytics but that is not the same as what you *should* do and are *allowed* to do. Imagine these three dilemmas:

- On your company website you gather anonymous tracking data, while in your call center, you are logging who people are and what they ask. You can combine the two sources of data and get a deep insight into preferences. You can, but would it compromise personal privacy?
- Your HR system has thousands of CVs along with data on who was hired. You can optimize the hiring by filtering the ones who are never being hired. If you have a bias in your hiring it will be replicated in the filtering, but you never hire anyone that doesn't fit the company anyway. Correct, and right?
- Rebates towards your B2B[1] customers are given based on how likely they are to pay on time. Based on all your customer data, a neural network algorithm decides the rate. As neural networks are "black box" models you cannot tell how the result came about. Should you use it even if it is more accurate than before?

Data ethics is the study and practice of how to collect, use, and share data in a responsible, secure, and legal way. It is important because data has a significant impact on the lives of individuals, organizations, and society, and it can be used for good or evil purposes. Data ethics aims to ensure that data practices respect the values, rights, and interests of the data subjects[2] and stakeholders and that they contribute to the common good.

In this chapter, you will learn about the ethical, legal, and social implications of data practices from a holistic stakeholder perspective. You will also learn how to protect the privacy, security, and rights of your data subjects and stakeholders using policies, standards, and technologies. You will explore the principles and frameworks of data ethics, such as the FAIR principles, the OECD principles, and the ACM code of ethics, and how to apply them to design and implement responsible, secure, and legal data practices. You will also get the tools to develop a data ethics strategy and a data ethics policy for a specific organization or project and demonstrate the skills and competencies of a data ethicist. Finally, you will reflect on the current and emerging challenges and opportunities of data ethics, and how to anticipate and prepare for the future of data ethics.

DOI: 10.4324/9781003457787-9

LEARNING GOALS:

L9.1 Explain the ethical implications of data practices for different stakeholders
L9.2 Identify the legal implications of data practices for different jurisdictions
L9.3 Analyze the social implications of data practices for various aspects
L9.4 Evaluate the policies, standards, and technologies that can help mitigate the risks of data practices
L9.5 Develop a data ethics strategy and a data ethics policy for a specific organization or project

This chapter is structured into three initial sections that cover the ethical, legal, and societal implications of using data for decision-making and then move into policy creation and implementation of those policies. Various tools can be used to support the implementation and they to some extent overlap with the data governance tools covered in the previous chapters.

Chapter case: Amnesty International LGBTQ+, NGO, Belgium

ETHICAL RISK IN DATA USAGE FOR DECISION-MAKING

Ethical data usage is the practice of collecting, storing, and using data in a way that respects the rights and interests of the data subjects and users. Data-driven decision-making can have ethical implications for different stakeholders, such as data collectors, data users,[3] and society at large.

This can have both positive and negative impacts. A positive example is Apple's commitment to privacy, which minimizes personal data collection, processes much of the data on the user's device instead of in the cloud, provides transparency reports, and gives the user significant control over their data. This is even despite many of the companies that are dependent on their ecosystem (Apple, 2023). A negative example is the Facebook and Cambridge Analytica scandal, which involved the unauthorized harvesting and misuse of the personal data of millions of Facebook users for political purposes (Cadwalladr & Graham-Harrison, 2018). This has had a potential direct impact on the elections held and politics being conducted around the world.

Ethical data usage is based on the following five principles:

- **Fairness:** Data should be used for legitimate and beneficial purposes, and not for causing harm or discrimination to individuals or groups. Data should also be collected and processed in a way that minimizes biases and ensures representativeness and accuracy.
- **Accountability:** Data users should be responsible for the outcomes and impacts of their data usage and be able to explain and justify their decisions and actions. Data users should also comply with the relevant laws and regulations and adhere to the ethical standards and codes of conduct of their profession and organization.
- **Transparency:** Data users should be open and honest about their data collection and usage practices and provide clear and accessible information to the data subjects and the

public. Data users should also allow the data subjects to access, correct, and delete their data, and to opt out of data collection and usage if they wish.

- **Privacy:** Data users should respect the privacy and confidentiality of the data subjects, and protect their data from unauthorized access, use, and disclosure. Data users should also use appropriate methods of encryption, anonymization, and aggregation to reduce the risks of data breaches and re-identification.
- **Consent**: Data users should obtain the informed and voluntary consent of the data subjects before collecting and using their data and respect their preferences and choices. Data users should also inform the data subjects of the purpose, scope, duration, and potential benefits and risks of their data usage, and provide them with the opportunity to withdraw their consent at any time.

Ethical risk categories

Data-driven decision-making is the process of using data to inform and support decisions in various domains and contexts. While data-driven decision-making can offer many benefits, such as efficiency, accuracy, and innovation, it can also pose many ethical risks, such as bias, discrimination, manipulation, exploitation, and harm.

TABLE 9.1 Ethical risks in data-driven decision making

Bias	Discrimination	Manipulation	Exploitation	Harm

Bias: The deviation from objectivity or fairness in data collection, analysis, or interpretation. Bias can result from human errors, assumptions, preferences, or prejudices, as well as from technical limitations, such as data quality, availability, or representativeness. Bias can lead to inaccurate or misleading outcomes that can affect the validity and reliability of data-driven decisions. For example, a hiring algorithm that is trained on a biased data set that favors certain demographic groups over others can result in unfair and discriminatory hiring practices (Dastin, 2018).

Discrimination: The unjust or prejudicial treatment of individuals or groups based on their characteristics or attributes, such as age, gender, race, ethnicity, religion, disability, or sexual orientation. Discrimination can occur when data-driven decisions are based on biased or incomplete data, or when data is used to target, exclude, or disadvantage certain individuals or groups. For example, a credit scoring system that uses data to assess the risk and eligibility of borrowers can result in discriminatory lending practices that deny or limit access to credit for certain individuals or groups (O'Neil, 2016).

Manipulation: The intentional or unintentional influence or control of individuals or groups through the use of data or information. Manipulation can occur when data is used to persuade, deceive, or coerce individuals or groups to act or think in certain ways, or when data is used to alter or distort the reality or perception of individuals or groups. For example, a social media platform that uses data to personalize the content and ads that users see can result in the manipulation of users' preferences, opinions, and behaviors (Zuboff, 2019).

Exploitation: The unfair or unethical use of data or information for one's benefit or advantage, often at the expense or harm of others. Exploitation can occur when data is used to extract value, profit, or power from individuals or groups without their consent, knowledge, or compensation, or when data is used to violate or infringe the rights or interests of individuals or groups. For example, a data broker that collects and sells personal data of individuals or groups without their consent or awareness can result in the exploitation of their privacy, security, and identity (Sadowski, 2020).

Harm: The negative or adverse impact or consequence of data-driven decisions on individuals, groups, or society. Harm can be physical, psychological, emotional, social, economic, or environmental, and can be intentional or unintentional, direct or indirect, immediate or delayed, or reversible or irreversible. For example, a self-driving car that uses data to navigate and operate can cause harm to passengers, pedestrians, or other road users in case of a malfunction, error, or accident (Calo, 2017).

Frameworks for risk mitigation

One of the ways to address the ethical risks of data-driven decision-making is to use methods and frameworks that can help assess and mitigate these risks. Some of the methods and frameworks that have been proposed or used are:

TABLE 9.2 Frameworks for risk mitigation

Ethical impact assessment	Data protection impact assessment	Value-sensitive design	Ethical data governance

Ethical impact assessment: This is a tool that aims to evaluate whether a specific data or AI project is aligned with the values, principles, and guidance set by the UNESCO Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2021). It also aims to ensure transparency by requiring information about the project, its methods, and outcomes be available to the public. The ethical impact assessment covers the entire AI life cycle and includes ex-ante and ex-post requirements. It can be used by procurers, developers, or users of AI systems, in the public or private sectors, who wish to develop AI ethically and comply with international standards (UNESCO, 2021).

Data protection impact assessment: This is a tool that aims to identify and minimize the data protection risks of a project involving personal data. It is required by the General Data Protection Regulation (GDPR) for projects that are likely to result in a high risk to the rights and freedoms of individuals (European Commission, 2016). It helps to assess the necessity, proportionality, and compliance of the data processing, and to identify and implement measures to address the risks. It also helps to consult and inform the relevant stakeholders, such as data subjects, data protection authorities, or data protection officers (European Commission, 2016).

Value-sensitive design: This is a framework that aims to integrate human values into the design of technology. It is based on the premise that technology is not value-neutral, but rather can support or hinder certain values, such as privacy, autonomy, or justice. It involves identifying the direct and indirect stakeholders of the technology, the values that are relevant to them, and the potential value conflicts or trade-offs. It also involves designing the technology to respect and promote the values and evaluating the impacts of the technology on the values (Friedman et al., 2017).

Ethical data governance: This is a framework that aims to establish and enforce the rules and norms for the collection, storage, analysis, and sharing of data. It involves defining the roles and responsibilities of the data actors, such as data owners, data stewards, data custodians, or data users. It also involves setting the standards and policies for data quality, security, privacy, and ethics, and monitoring and auditing the compliance and performance of the data actors. It also involves engaging and empowering the data subjects and the public, and ensuring their rights and interests are respected and protected (Kitchin, 2014).

Ethical data usage

Ethical data usage is not without challenges and limitations, as there are often trade-offs, conflicts, uncertainties, and gaps that need to be considered and addressed.

Ethical data usage may require balancing different values, interests, and objectives, such as privacy, security, accuracy, efficiency, innovation, and social good. For example, enhancing the privacy of data subjects may reduce the utility or quality of the data for analysis, or improving the security of data may increase the cost or complexity of data processing. Ethical data usage may also involve weighing the benefits and risks of data usage for different stakeholders, such as data subjects, data users, and society at large. For example, using data for public health or security purposes may benefit society, but may also pose risks to the privacy or autonomy of the data subjects.

Ethical data usage may encounter conflicts between different values, principles, or norms, such as fairness, accountability, transparency, and consent. For example, ensuring the fairness of data-driven decisions may require revealing the logic or criteria of the algorithms, but this may conflict with the accountability of the data users who may want to protect their intellectual property or trade secrets. Similarly, obtaining the consent of the data subjects may require providing them with clear and comprehensible information about the data usage, but this may conflict with the transparency of the data users who may want to avoid disclosing sensitive or confidential information.

Ethical data usage may face uncertainties about the outcomes and impacts of data usage, as well as the preferences and expectations of the data subjects and the public. For example, the outcomes and impacts of data usage may be unpredictable, dynamic, or context-dependent, and may vary across different domains, scenarios, or time frames. Likewise, the preferences and expectations of the data subjects and the public may be diverse, ambiguous, or evolving, and may depend on various factors, such as culture, education, or awareness.

Ethical data usage may suffer from gaps between the theory and practice of data ethics, such as the gap between the ethical principles and guidelines and their implementation and enforcement, or the gap between the ethical awareness and competence and their education and training. For example, the ethical principles and guidelines may be vague, incomplete, or inconsistent, or may not reflect the latest developments or challenges of data usage. Alternatively, the implementation and enforcement of the ethical principles and guidelines may be inadequate, ineffective, or inconsistent, or may lack the necessary resources, tools, or mechanisms. Similarly, the ethical awareness and competence may be low, uneven, or outdated, or may not match the level or complexity of data usage. Alternatively, the education and training of ethical awareness and competence may be insufficient, irrelevant, or inaccessible, or may lack the appropriate content, methods, or incentives.

LEGAL RISKS IN DATA USAGE FOR DECISION-MAKING

Making decisions based on data is regulated in different ways in different jurisdictions, such as the European Union, the United States, and China. How to comply with relevant laws and regulations, such as the GDPR, the CCPA, and the CSL can be a real challenge. A positive example of how legislation can benefit data practices positively is the GDPR and Data Protection Act of the European Union, which is a comprehensive regulation that protects the personal data of EU citizens and residents and imposes obligations and penalties on data collectors and users (European Commission, 2023).

A negative example is Project Nightingale and Google, which involved the secret transfer of health data of millions of Americans from Ascension, a healthcare provider, to Google, without the consent or knowledge of the patients or doctors (Copeland & Mattioli, 2019).

In this section, we will divide the legislation into three categories. Personal data, health data, and AI/algorithmic data usage.

Personal/privacy data legislation

GDPR and the CCPA are examples of legal frameworks for data protection and privacy. They are laws that regulate how personal data is collected, processed, stored, and shared by different entities, such as businesses, governments, or individuals. They also grant rights to individuals to control their data, such as the right to access, delete, or correct their data.

GDPR stands for General Data Protection Regulation, and it is a law that applies to the European Union (EU) and the European Economic Area (EEA). It was adopted in 2016 and became effective in 2018. It is one of the most comprehensive and strict data protection laws in the world, and it has influenced many other countries to adopt similar laws or standards. GDPR aims to protect the fundamental rights and freedoms of individuals in the EU and the EEA, especially their right to privacy and data protection. It imposes obligations on data controllers and data processors, such as obtaining consent, providing transparency, ensuring security, and reporting breaches. It also gives individuals the right to access, rectify, erase, restrict, modify, and object to the processing of their personal data. GDPR also establishes a supervisory authority for each member state, as well as a European Data Protection Board, to enforce the law and impose fines or sanctions for non-compliance. The fines that can be given are also very substantial.

CCPA stands for California Consumer Privacy Act, and it is a law that applies to California, USA. It was enacted in 2018 and became effective in 2020. It is one of the first and most significant state-level data protection laws in the US, and it has inspired other states to propose or enact similar laws. CCPA aims to protect the privacy rights and consumer protection of individuals in California, especially their right to know, access, and delete their personal data. It

ILLUSTRATION 9.1 AI generated image by Canva

imposes obligations on businesses that collect, sell, or share personal information of California residents, such as providing notice, honoring opt-out requests, and ensuring security. It also gives individuals the right to access, delete, and opt out of the sale or sharing of their personal information, as well as the right to non-discrimination for exercising their rights. CCPA also establishes the California Attorney General as the primary enforcer of the law, as well as a private right of action for individuals to sue for damages in case of data breaches.

Health and privacy information

HIPAA, which stands for the Health Insurance Portability and Accountability Act of 1996 is a federal law in the United States that regulates the privacy and security of health information. It was one of the first legislations that tried to protect personal data. HIPAA applies to covered entities, such as healthcare providers, health plans, healthcare clearinghouses, and their business associates, who use or disclose protected health information (PHI) for certain purposes. HIPAA also grants rights to individuals to access, correct, and control their personal health information (PHI).

There are similar laws in other places around the world that aim to protect the privacy and security of health information. For example, in the European Union (EU), there is the General Data Protection Regulation (GDPR), which applies to the processing of personal data as introduced in the previous section also including health data, by any entity that operates in the EU or offers goods or services to individuals in the EU. GDPR imposes obligations on data controllers and data processors, such as obtaining consent, providing transparency, ensuring security, and reporting breaches. It also gives individuals the right to access, rectify, erase, restrict, modify, and object to the processing of their personal data. In Canada, there is the Personal Information Protection and Electronic Documents Act (PIPEDA), which applies to the collection, use, and disclosure of personal information, including health information, by any organization that engages in commercial activities or operates across provincial or national borders. PIPEDA requires organizations to obtain consent, provide notice, limit collection, ensure accuracy, protect security, and be accountable for the personal information they handle. It also gives individuals the right to access, correct, and challenge the handling of their personal information. In Australia, there is the Privacy Act 1988, which applies to the handling of personal information, including health information, by most Australian government agencies, all private sector and not-for-profit organizations with an annual turnover of more than $3 million, all private health service providers, and some small businesses. The Privacy Act sets out 13 Australian Privacy Principles (APPs) that regulate the collection, use, disclosure, storage, security, and access of personal information. It also gives individuals the right to complain, seek compensation, and request access or correction of their personal information.

These are just some examples of the legal frameworks for data protection and privacy in different regions of the world. There may be other laws or regulations that apply to specific sectors, jurisdictions, or situations that are not covered here, so you should always seek legal advice regarding your industry and jurisdiction. Most regulation is however shaped similarly, and for many, it will be sufficient to adhere to the European GDPR act as it is the most comprehensive.

AI and algorithmic legislation

The EU AI Act and the US presidential decree are two examples of legal frameworks that aim to regulate the development and use of artificial intelligence (AI) in their respective jurisdictions. They are still in development in most cases, but with the rapid development in AI over the last couple of years the legislative work has also been sped up.

The EU AI Act is a regulation that lays down harmonized rules on AI and amends certain Union legislative acts. It was proposed by the European Commission in April 2021, approved finally in February 2024, and is being implemented currently (2024) in national legislation. The AI Act defines AI as

> software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.
>
> (Feingold, 2023)

The AI Act classifies AI systems into four categories of risk: prohibited, high-risk, limited-risk, and minimal-risk, and imposes different obligations and requirements for each category. The AI Act also establishes a governance framework that involves national competent authorities, a European Artificial Intelligence Board, and the Commission (European Commission, 2021).

The US presidential decree is an executive order that establishes new standards for AI safety and security, protects Americans' privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more. It was issued by President Biden in October 2023 and is the first comprehensive federal action on AI governance. The decree directs various agencies to take actions that adhere to eight guiding principles and priorities: AI must be safe and secure, AI must respect privacy and civil liberties, AI must advance equity and civil rights, AI must empower workers and consumers, AI must foster innovation and competition, AI must uphold democratic values and institutions, AI must promote international cooperation and standards, and AI must be accountable and transparent (The White House, 2023).

Other countries have also proposed or implemented legislation or initiatives related to AI governance. For example, in Brazil, a non-permanent jurisprudence commission of the Brazilian Senate presented a report with studies on the regulation of AI, including a draft for the regulation of AI, in December 2022. The draft is based on three central pillars: guaranteeing the rights of people affected by the system, classifying the level of risk, and predicting governance measures for companies that provide or operate the AI system. However, the legislation regarding artificial intelligence is still in the process of being drafted in many jurisdictions which means that it is a field anyone embarking on using AI systems for decision support will have to monitor closely. An example of a decision that is likely to be regulated is hiring decisions where bias in the selection must be documented. The bias would come from training the AI algorithms with biased data. The International Association of Privacy Professionals has developed a tracker for the development of legislation globally (International Association of Privacy Professionals, n.d.). It is recommended to use that to get up-to-date information about the state of AI legislation globally.

SOCIAL RISK IN DATA USAGE FOR DECISION-MAKING

Data-driven decision-making can have social implications for many areas. Human dignity, autonomy, justice, and democracy, and how to balance the benefits and risks of data use for the common good can be challenging. A positive example is Toronto's Sidewalk Labs, which is a project that aims to create a smart city that uses data and technology to improve urban life, such as mobility, sustainability, and affordability (Sidewalk Labs, 2023). A potentially negative example is China's social credit system, which is a system that uses data and algorithms to monitor and evaluate the behavior and trustworthiness of citizens and organizations, and reward or punish them accordingly (Botsman, 2017).

There are four main dimensions of social risk: privacy, discrimination, manipulation, and accountability.

Privacy

Privacy is the *right to control one's personal information and to limit its access and use by others*. Privacy is essential to maintain human dignity, autonomy, and the right to self-expression, as well as for the protection of other rights, such as freedom of expression, association, and political

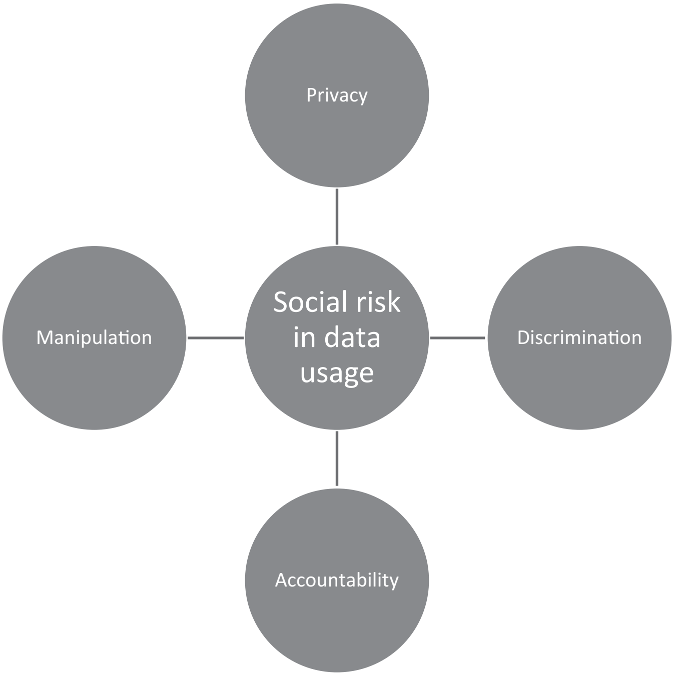


FIGURE 9.1 Social risk of data usage

participation. Data-driven decision-making may threaten privacy in several ways, such as collecting, processing, and storing large amounts of personal data, often without the consent, knowledge, or awareness of the individuals concerned. This is done both by large corporate entities and governments. By combining, linking, or inferring personal data from different sources, such as online platforms, sensors, cameras, or biometrics, these organizations can create detailed and comprehensive profiles of individuals and groups. This will enable them to provide or withdraw privileges based on otherwise private information. What's often termed as "big tech"[4] and some of the larger Chinese tech conglomerates are capable of this. This can also happen through sharing, transferring, or selling personal data to third parties, such as advertisers, marketers, insurers, or governments, for various purposes, such as targeting, profiling, or surveillance. Finally hacking, or leaking personal data, either intentionally or unintentionally that exposes individuals to identity theft, fraud, or harassment is a risk that cannot be neglected.

Some measures have been put in place to protect privacy in data-driven decision-making. From the official side the implementation of data protection and privacy laws and regulations, such as the GDPR in the EU or the CCPA in California in the US, that set standards and rules for the collection, processing, storage, and sharing of personal data, and grant rights to individuals to access, delete, or correct their data. This was discussed in detail in the previous section on legal risks.

Companies can adopt privacy-by-design and privacy-by-default principles, which ensure that privacy is embedded and prioritized in the design and development of data and

AI systems and that the default settings are the most privacy-friendly ones. Instead of reacting to privacy risks or invasions when they happen, companies actively build processes and procedures to prevent them from occurring in the first place (Hiscock, n.d.).

Applying these privacy-enhancing technologies, such as encryption, anonymization, or differential privacy enables the use of data without revealing the identity or sensitive information of the individuals involved. Synthetic data generated based on private information is also a way of protecting individuals as their information is not used but relevant analysis can still be made.

Governments can also use education and empowering individuals and communities, such as through digital literacy, awareness campaigns, or advocacy groups, that enable them to understand and exercise their privacy rights and choices and to challenge or resist privacy violations. Globally such digital literacy programs have been rolled out even into primary schools.

Discrimination

Discrimination is the *unfair or unequal treatment of individuals or groups based on their characteristics, such as race, gender, age, disability, or religion*. Data-driven decision-making may cause or increase discrimination in several ways, such as:

In the context of data analysis and machine learning, several critical considerations arise:

- **Data bias:** Existing biases, stereotypes, or prejudices in the data, whether they are historical, cultural, institutional, or introduced by human decisions or actions, can be reflected, reproduced, or amplified. When machine learning algorithms are trained on such biased data, the results they produce will also be biased.
- **Algorithmic bias:** The process of data analysis or interpretation can create, reinforce, or perpetuate new biases, stereotypes, or prejudices. These can take various forms, such as algorithmic biases, statistical biases, or confirmation biases. There are always elements of choice in the algorithm selection.
- **Impact on outcomes:** The data profiles or scores assigned to individuals or groups can affect, influence, or determine their outcomes or opportunities in various domains. These domains can include education, employment, health, or justice, and the scores can take various forms, such as credit scores, risk scores, or social scores.

These considerations underscore the importance of careful and ethical data handling and algorithm design. Some of the possible ways to prevent or mitigate discrimination in data-driven decision-making are easy to implement on the surface but quickly become complicated technically. An example is biases of AI systems might be easy to identify but filtering data training data for that bias is complicated. In the context of anti-discrimination and human rights, several key actions can be taken.

- **Legal implementation:** Anti-discrimination and human rights laws and regulations, like the UK's Equality Act or the US's Civil Rights Act, are implemented. These laws prohibit discrimination based on protected characteristics and pr**o**vide individuals with the right to challenge or seek redress for discrimination.

- **Principles of fairness, equality, and diversity:** Principles that ensure data and AI systems are inclusive, representative, and respectful of the diversity and complexity of individuals and groups are adopted. These principles aim to prevent the creation or exacerbation of existing inequalities or injustices.
- **Fairness-enhancing technologies:** Technologies that enhance fairness, such as auditing, testing, or debiasing, are applied. These technologies enable the detection, measurement, and correction of biases or discrimination in data and AI systems, and the evaluation and improvement of their fairness and accuracy.
- **Education and empowerment:** Efforts are made to educate and empower individuals and communities. This can be achieved through digital literacy, awareness campaigns, or advocacy groups, enabling them to understand and exercise their anti-discrimination rights and choices and to challenge or resist discrimination.

These actions highlight the importance of proactive measures in promoting fairness and equality in the use of data and AI systems.

Manipulation

Manipulation is the deliberate or covert influence or control of individuals or groups by others, such as governments, corporations, or individuals, for their interests or agendas. Manipulation is detrimental to human autonomy, agency, and democracy, as well as to the trust and legitimacy of society. Data-driven decision-making can potentially facilitate manipulation in several ways:

- **Exploitation and persuasion**: Data and AI can be used to exploit, persuade, or nudge individuals or groups to behave, think, or feel in certain ways. This can be achieved by tailoring, personalizing, or optimizing messages, content, or recommendations, or by creating or manipulating emotions, opinions, or beliefs.
- **Deception and misinformation**: Data and AI can be used to deceive, mislead, or confuse individuals or groups about reality, truth, or facts. This can be done by fabricating, falsifying, or distorting information, evidence, or images, or by creating or spreading misinformation, disinformation, or propaganda.
- **Coercion and intimidation**: Data and AI can be used to coerce, threaten, or intimidate individuals or groups to comply, conform, or cooperate. This can be achieved by monitoring, tracking, or surveilling individuals or groups, or by creating or enforcing incentives, sanctions, or punishments.

These potential manipulations underscore the importance of ethical considerations in the use of data and AI for decision-making. Because it is possible some people can act adversely to the use of data and AI in decision-making.

Some of the possible ways to counter manipulation in data-driven decision-making are in the realm of information and communication:

- **Legal implementation:** Laws and regulations, such as the US's Freedom of Information Act or the EU's Audiovisual Media Services Directive, are implemented. These set standards and rules for the quality, accuracy, and transparency of information and communication, and grant individuals the right to access, verify, or correct information.

- **Principles of transparency, accountability, and explainability:** Principles ensuring that data and AI systems are open, understandable, and verifiable by individuals and society are adopted. These systems can be questioned, challenged, or appealed to by individuals and society.
- **Manipulation-detecting technologies:** Technologies for detecting manipulation, such as fact-checking, verification, or authentication, are applied. These enable the identification, assessment, and validation of the sources, methods, and purposes of data and AI systems, and the detection and exposure of manipulation or deception.
- **Education and empowerment:** Efforts are made to educate and empower individuals and communities. This can be achieved through digital literacy, awareness campaigns, or advocacy groups, enabling them to understand and exercise their information and communication rights and choices and to challenge or resist manipulation.

These actions highlight the importance of proactive measures in promoting transparency and accountability in the use of data and AI systems.

Accountability

Accountability is *the obligation or responsibility of individuals or groups to answer for or justify their actions or decisions*, especially to those who are affected by them. Accountability is essential for human trust, as well as for the quality and reliability of the society. Data-driven decision-making may challenge or undermine accountability by:

- **Role obscurity:** The roles, responsibilities, and liabilities of the actors or stakeholders involved in data and AI systems, such as data providers, data processors, data users, data subjects, or data regulators, can become obscured, complicated, or dispersed. This can create gaps or conflicts between them.
- **Reduced human involvement:** The human involvement, oversight, or control in data and AI systems can be reduced, limited, or removed, especially with the use of automated, autonomous, or self-learning data and AI systems. This can create risks or uncertainties for human safety, security, or welfare.
- **Power shift:** The power, influence, or authority of data and AI systems over individuals and society can increase, shift, or transfer. This can occur when data and AI systems are used to make or support decisions or actions that have significant or irreversible impacts or consequences for individuals and society.

These potential challenges underscore the importance of ethical considerations in the use of data and AI systems. Some of the possible ways to ensure or enhance accountability in data-driven decision-making are:

In the context of data and AI governance, several key actions are undertaken:

- **Legal implementation:** Laws and regulations, such as the EU's AI Act or the US's presidential decree, are implemented. These set standards and rules for the development and use of data and AI systems, and grant individuals the right to oversee, control, or challenge these systems.

- **Principles of accountability, responsibility, and liability:** Principles ensuring that data and AI systems are designed and developed with human values and interests in mind are adopted. The actors or stakeholders involved in these systems are held accountable, responsible, and liable for their actions or decisions, and their impacts or consequences.
- **Accountability-enhancing technologies**: Technologies that enhance accountability, such as logging, monitoring, or auditing, are applied. These enable the recording, tracking, and reviewing of the actions or decisions of data and AI systems, and the assessment and improvement of their performance and outcomes.
- **Education and empowerment**: Efforts are made to educate and empower individuals and communities. This can be achieved through digital literacy, awareness campaigns, or advocacy groups, enabling them to understand and exercise their data and AI governance rights and choices and to challenge or appeal against data and AI systems.

These actions highlight the importance of proactive measures in promoting accountability and responsibility in the use of data and AI systems.

In conclusion, there are four main dimensions of social risk: privacy, discrimination, manipulation, and accountability. Within each of these dimensions, several challenges and remedies have been shown. It's summarized in the figure below, that a company uses to assess the risk of social harm in data-driven decision-making.

TABLE 9.3 Dimensions of social risk from data-driven decision-making

Privacy		**Discrimination**	
Risk	**Remedy**	**Risk**	**Remedy**
• Reidentification and deanonymization • Opacity and secrecy of profiling • Data Exploitation	• Legal implementation • Principles of privacy by design • Collection and retention minimization • Awareness programs	• Data bias • Algorithmic bias • Impact on outcomes	• Legal implementation • Principles of fairness, equality, and diversity • Fairness enhancing technologies • Education and empowerment
Manipulation		**Accountability**	
Risk	**Remedy**	**Risk**	**Remedy**
• Exploitation and persuasion • Deception and misinformation • Coercion and intimidation	• Legal implementation • Principles of transparency, accountability, and explainability • Manipulation-detecting technologies • Education and empowerment	• Role obscurity • Reduced human involvement • Power shift	• Legal implementation • Principles of accountability, responsibility, and liability • Accountability-enhancing technologies: • Education and empowerment

POLICIES, STANDARDS, AND TECHNOLOGIES TO MITIGATE RISKS

A way to mitigate some of the risks mentioned in the previous sections is to use policies, standards, and technologies, such as data governance and protection policies, as well as data quality and data security tools. An example of how bad practices can hurt a company is the Equifax data breach, which involved the exposure of personal and financial data of nearly 150 million Americans, due to the lack of adequate data security measures and practices (Fruhlinger, 2021).

Legal risk mitigation

Data-driven decision-making requires data ethics, which is not only a moral obligation, but also a legal requirement. Organizations that collect, store, process, and share data are subject to legal liabilities, reputational damage, and loss of trust if they breach, misuse, discriminate, or harm data subjects. Therefore, it is essential to mitigate legal risks by implementing effective policies, standards, and technologies that ensure data protection, privacy, and security.

Policies for legal risk mitigation

Policies are the rules and guidelines that govern the data life cycle within and outside an organization. They should reflect the ethical principles and values of the organization, as well as comply with relevant laws and regulations, such as the GDPR in the EU, or the CCPA in the US (Bennett & Raab, 2020). Policies should also specify the roles and responsibilities of data owners, stewards, custodians, and users, and the procedures for data quality, access, consent, retention, and disposal. Policies should be communicated and enforced across the organization and reviewed and updated regularly to reflect changes in the data environment and legal landscape (Kitchin, 2014).

Standards for legal risk mitigation

Standards are the technical specifications and best practices that ensure data is consistent, interoperable, and secure. They can be internal or external, depending on the source and scope of the data. Internal standards are developed by the organization to meet its own data needs and objectives, such as data formats, naming conventions, metadata, and documentation. External standards are adopted by the organization to align with industry or sector norms, such as data models, protocols, and frameworks. Standards can also be voluntary or mandatory, depending on the level of compliance and accountability required by the organization or its stakeholders. Standards should be applied and monitored throughout the data life cycle, and evaluated and improved periodically to ensure data quality, reliability, and usability (Janssen et al., 2017).

Technologies for legal risk mitigation

Technologies are the tools and systems that enable data collection, storage, processing, and sharing. They should be designed and deployed in a way that respects data ethics and minimizes legal risks. For example, technologies should incorporate privacy by design and by default, which means that data protection and privacy are embedded in the design and operation of the technologies, and that the default settings are the most privacy-friendly ones (Cavoukian, 2011). Technologies should also support data governance and accountability, such as data audits, logs, and reports, that demonstrate compliance with policies and standards (OECD, 2019).

Discrimination risk mitigation

Discrimination occurs when data is used to unfairly or unjustly treat or exclude individuals or groups based on their personal characteristics, such as race, gender, age, or disability. Discrimination can have negative impacts on the data subjects, as well as on the trust, reputation, and social responsibility of data users. Therefore, it is essential to mitigate discrimination risk by implementing effective policies, standards, and technologies that ensure data fairness, transparency, and accountability.

Policies for discrimination risk mitigation

Policies are the rules and guidelines that govern how data is collected, stored, processed, and shared within and outside an organization. Policies should reflect the ethical principles and values of the organization, as well as comply with relevant laws and regulations, such as the Equality Act 2010 in the UK, or the Civil Rights Act 1964 in the US (Bennett & Raab, 2020). Policies should also specify the criteria and objectives for data use, the methods and measures for data analysis, and the safeguards and remedies for data misuse. Policies should be communicated and enforced across the organization and reviewed and updated regularly to reflect changes in the data environment and legal landscape (Kitchin, 2014).

Standards for discrimination risk mitigation

Standards are the technical specifications and best practices that ensure data is consistent, interoperable, and secure. They can be internal or external, depending on the source and scope of the data. Internal standards are developed by the organization to meet its own data needs and objectives, such as data quality, validity, and reliability. External standards are adopted by the organization to align with industry or sector norms, such as data ethics frameworks, codes of conduct, and certification schemes. Standards can also be voluntary or mandatory, depending on the level of compliance and accountability required by the organization or its stakeholders. Standards should be applied and monitored throughout the data life cycle, and evaluated and improved periodically to ensure data fairness, transparency, and accountability (Janssen et al., 2017).

Technologies for discrimination risk mitigation

Technologies are the tools and systems that enable data collection, storage, processing, and sharing. They should be designed and deployed in a way that respects data ethics and minimizes discrimination risk. For example, technologies should incorporate fairness by design and by default, which means that data is collected and analyzed in a way that avoids or mitigates bias, prejudice, and stereotyping. Technologies should also implement transparency by design and by default, which means that data is processed and presented in a way that is clear, understandable, and explainable. Technologies should also support accountability by design and by default, which means that data is used and reported in a way that is responsible, auditable, and responsive (Mittelstadt et al., 2016).

Manipulation risk mitigation

While data-driven decision-making can enhance efficiency, productivity, and innovation for individuals and organizations, it also presents ethical dilemmas, such as the potential for manipulation. Manipulation refers to the use of data to sway people's actions or thoughts without their complete understanding or agreement. This can adversely affect the autonomy, dignity, and well-being of those whose data is used, and can harm the trust, reputation, and social accountability of those using the data. As such, it's crucial to reduce the risk of manipulation by adopting robust policies, standards, and technologies that guarantee the accuracy, integrity, and respect of data.

Policies for manipulation risk mitigation

Policies serve as the guiding principles that dictate the collection, storage, processing, and dissemination of data within an organization and beyond. These policies should embody the organization's ethical values and principles, and adhere to pertinent laws and regulations, such as India's Consumer Protection Act 2019 or the US's Federal Trade Commission Act 1914, both of which safeguard consumers from manipulation. The policies should clearly define the intent and extent of data usage, the constraints of data analysis, and the entitlements and obligations of data subjects and users. These policies must be disseminated and implemented throughout the organization and are routinely revised and updated to accommodate shifts in the data ecosystem and legal framework.

Standards for manipulation risk mitigation

Standards refer to the technical guidelines and best practices that guarantee data consistency, interoperability, and security. They can either be internal or external, contingent on the data's source and scope. Internal standards are formulated by the organization to fulfill its specific data requirements and goals, such as ensuring data accuracy, completeness, and relevance. Conversely, external standards are embraced by the organization to conform to industry or sector standards, like data ethics frameworks, codes of conduct, and certification schemes. Depending on the degree of compliance and accountability demanded by the organization or its stakeholders, standards can be either voluntary or obligatory. It's crucial that these standards are implemented and supervised throughout the data life cycle and are periodically assessed and enhanced to maintain data accuracy, integrity, and respect.

Technologies for manipulation risk mitigation

Technologies refer to the tools and systems that facilitate the collection, storage, processing, and dissemination of data. They should be engineered and utilized in a manner that upholds data ethics and reduces the risk of manipulation. For instance, technologies should embody the principle of accuracy by design and by default, implying that data should be gathered and analyzed in a manner that prevents or rectifies inaccuracies, distortions, and fabrications. Similarly, technologies should adhere to the principle of honesty by design and by default, ensuring that data is processed and displayed in a manner that is truthful, equitable, and balanced. Lastly, technologies should espouse the principle of respect by design and by default, meaning that data should be used and reported in a manner that is respectful, considerate, and sensitive.

Accountability risk mitigation

Accountability risks refer to the potential negative consequences of using data and technology in ways that are unethical, unfair, or harmful to individuals or society. Accountability risks can arise from various sources, such as data quality, data privacy, data security, algorithmic bias, human oversight, and legal compliance. To mitigate these risks, organizations need to adopt and implement appropriate policies, standards, and technologies that ensure the ethical use of data and technology and protect the interests and rights of the company stakeholders, such as shareholders, employees, customers, partners, and regulators.

Policies for mitigating accountability risks

Policies are high-level guidelines and principles that define the goals, values, and expectations of an organization regarding the ethical use of data and technology. Policies can help establish a culture of data ethics and foster trust and transparency among stakeholders. Policies can also provide a framework for identifying, assessing, and addressing accountability risks and issues and ensuring the accountability and responsibility of the organization and its agents. For example, a data ethics policy can specify the criteria and processes for data collection, analysis, and sharing, as well as the roles and responsibilities of data owners, data users, and data stewards. A data ethics policy can also outline the ethical principles and values that guide the design, development, and deployment of data-driven technologies, such as fairness, privacy, security, and explainability. Policies should be aligned with the relevant laws and regulations, as well as the best practices and standards in the industry and the society. Policies should also be reviewed and updated regularly to reflect the changing needs and expectations of the stakeholders and the evolving nature of data and technology (Młodziejewska & Soller, 2023).

Standards for mitigating accountability risks

Standards are the specific rules and requirements that operationalize the policies and ensure the consistent and compliant implementation of the ethical use of data and technology. Standards can help define the technical and procedural specifications and criteria for data quality, data privacy, data security, algorithmic fairness, algorithmic explainability, and human oversight. Standards can also help measure and monitor the performance and impact of data and technology on the stakeholders and society and provide evidence and assurance of the accountability and responsibility of the organization and its agents. For example, a data quality standard can specify the minimum level of accuracy, completeness, timeliness, and relevance of data, as well as the methods and tools for data validation, verification, and cleaning. A data privacy standard can specify the types and categories of data that can be collected, stored, processed, and shared, as well as the consent mechanisms, access controls, and data protection measures that need to be applied. A data security standard can specify the encryption, authentication, and authorization techniques and technologies that need to be used to prevent unauthorized or malicious access, use, or disclosure of data. An algorithmic fairness standard can specify the metrics and methods for assessing and mitigating the potential bias and discrimination in data and algorithms, as well as the mechanisms for ensuring the diversity and inclusion of the stakeholders in the data and technology life cycle. An algorithmic explainability standard can specify the level and format of transparency and interpretability that need to be provided to the stakeholders regarding the data and algorithms, as well as the mechanisms for enabling feedback and recourse. A human oversight standard can specify the degree and mode of human involvement and intervention in the data and technology processes, as well as the mechanisms for ensuring the accountability and responsibility of the human actors (Data Ethics Framework, 2018).

Technologies for mitigating accountability risks

Technologies are the tools and systems that enable and support the implementation of policies and standards and facilitate the ethical use of data and technology. Technologies can help automate and optimize the data and technology processes, as well as enhance the capabilities and competencies of the human actors. Technologies can also help detect and prevent accountability risks and issues and provide alerts and notifications to the stakeholders and the regulators. For example, a data quality technology can help automate the data validation, verification, and cleaning processes, as well as provide data quality indicators and reports. A data privacy technology can help automate the consent management, access control, and data protection processes, as well as provide data privacy audits and alerts. A data security technology can help automate the encryption, authentication, and authorization processes, as well as provide data security logs and notifications. An algorithmic fairness technology can help automate the bias and discrimination detection and mitigation processes, as well as provide fairness scores and explanations. An algorithmic explainability technology can help automate the transparency and interpretability generation and presentation processes, as well as provide explainability scores and feedback. A human oversight technology can help automate the human involvement and intervention processes, as well as provide oversight dashboards and recommendations (Berkowitz, 2020).

DATA ETHICS STRATEGY AND POLICY

Companies and organizations that collect, store, analyze, or share data have an ethical responsibility to develop a thoughtful data ethics strategy and policy. These outline the organization's approach, priorities, and rules regarding ethical data practices that respect relevant laws, stakeholders' rights, and societal values.

An effective data ethics strategy aligns with the organization's overall vision and business objectives while upholding important ethical principles around areas like privacy, security, transparency, and avoiding bias or harmful impacts. The strategy helps guide decisions around data collection, storage, usage, and sharing. It is brought to life through a formal data ethics policy.

The data ethics policy codifies important aspects of the strategy into specific rules, procedures, roles, and responsibilities. It covers topics like:

- What types of data will be collected and for what purposes
- How consent will be obtained from data subjects
- How data will be protected and secured
- Who has access to data and under what controls or limitations
- What governance processes will oversee data practices and address ethical concerns
- How transparency will be ensured around data practices
- How potential biases will be avoided in data and analytics.

The policy should be informed by legal requirements, industry best practices, stakeholder expectations, and an ethical risk assessment of the organization's data practices. As a baseline, many organizations adopt existing ethical frameworks like the OECD Privacy Principles or Fair Information Practice Principles.

Once developed, the policy must be clearly communicated to data subjects, employees, partners, and other relevant stakeholders. Ongoing training helps embed policy rules into actual practices. Audits, accountability mechanisms, and continuous improvement processes help evolve policies over time as risks, regulations, and societal expectations shift.

A positive example of someone taking data ethics seriously is IBM's AI Ethics Policy. IBM has published detailed principles and practices around trust and transparency in AI that guide its internal strategy and policies. Their AI ethics policy stresses transparency, explainability, and fairness in AI systems. For example, AI systems should be transparent about the data and models used and should provide explanations about how they arrive at decisions or predictions. Additionally, IBM aims to identify and remove potential biases embedded in data and algorithms (IBM, 2023).

On the other hand, Uber has faced scrutiny over data practices that some perceive as overly intrusive or opaque. For example, the company tracked and stored location data about riders and drivers for an extended time. Its surge pricing algorithm that sets ride prices during busy times has been criticized as manipulative. And a 2016 data breach compromised the personal information of 57 million Uber customers and drivers. While Uber has since updated some data policies and protections, these incidents highlight the ethical risks around poor data governance (Isaac, 2017).

Data ethics strategy, encoded into an actionable policy, can help organizations collect, analyze, and share data responsibly. Ongoing governance, training, and improvement processes are key to turning policies into ethical data practices.

DATA ETHICS SKILLS AND COMPETENCIES: DATA LITERACY

Practicing good data ethics requires a specialized skillset and mindset. Data ethicists need competencies in areas like ethical reasoning, decision-making, leadership, and communication.

Ethical reasoning: This involves carefully considering the impacts of data programs through different ethical frameworks to identify risks, harms, biases, and unintended consequences. It also entails exploring alternative options and interventions that could mitigate ethical issues.

Ethical decision-making: This applies ethical reasoning along with good judgment to make principled choices about data practices. This includes deciding what types of data to collect, what analyses are appropriate, what algorithms are fair, and how insights will be used.

Ethical leadership: This involves championing important data ethics principles throughout an organization. Leaders institute governance practices, assess risks, develop safeguards, communicate priorities and incentives, and promote an ethical culture around data.

Ethical communication: This entails transparently explaining data sources, models, uncertainties, and possible issues to both data subjects and decision-makers. Communication should enable oversight, understanding, and informed consent.

These interrelated competencies require ongoing education and training. Various emerging roles also focus specifically on data ethics, including positions like Chief Ethics Officer, data ethicist, algorithm auditor, and public interest technologist.

Cathy O'Neil is a positive role model as a data scientist and author focused on ethical analytics. Her expertise spans statistics, algorithms, and developing standards that address bias. She founded ORCAA, a company that audits algorithms to assess fairness and accuracy based on mathematical principles. Her skills in ethical reasoning help identify problematic assumptions or inferences. She is an ethical leader, giving talks and writing books that call attention to data ethics considerations.

On the other hand, Aleksandr Kogan exemplified how not to do data analytics ethically. He collected data from millions of Facebook users for political consulting firm Cambridge Analytica under the guise of academic research. This deception and his violation of Facebook's terms demonstrate flawed ethical reasoning and decision-making. His lack of consent and transparency further signify deficiencies in ethical communication and leadership around research ethics and data privacy. The scandal provoked global outrage regarding technology and data ethics.

Specialized skills in ethical reasoning, judgment, communication, and leadership can help data practitioners address ethical challenges responsibly. Fostering these professional competencies is key for both individuals and organizations hoping to operate ethically as they leverage data and AI for data-driven decision-making.

DATA ETHICS CHALLENGES AND OPPORTUNITIES

Rapid technological developments create new possibilities but also new ethical questions around ownership, control, comprehension, and advocacy issues in data systems. Proactive policies, education, technologies, and collective action can help address these complex challenges.

Data ownership: This refers to who rightfully possesses and profits from data. Tensions arise from companies monetizing users' data without compensation to those individuals. New models like data cooperatives and initiatives enabling people to own and sell their data on their own terms help shift this power dynamic.

Data sovereignty: This represents peoples' rights to self-governance over their data. However, governments and companies often demand access to data in ways that infringe on privacy and autonomy. Advancing individual rights, encryptions, community-controlled data trusts, and policy limitations on surveillance are approaches to ensure greater sovereignty.

Data literacy: This entails the skills and awareness to properly understand and evaluate data systems. As data-driven technologies grow more complex and opaque, massive gaps in public comprehension create risks of misinformation, manipulation, and disempowerment. Formal and informal education initiatives to improve data literacy at all ages and sectors are essential.

Data activism: This comprises actions and campaigns to promote data ethics and address related social justice issues. Activists raise awareness of problems, advocate for better safeguards, develop countermeasures, organize resistance efforts, and prompt reform through protests and political engagement.

An example of someone who is trying to make a positive impact is Invisibly who provides a platform for individuals to choose what personal data they share, limit access permissions, and receive payments when their data gets used by companies. This startup aims to shift the data ownership balance back towards users. Invisibly also hopes to improve data literacy by making data sharing more transparent for users.

On the other hand, Clearview AI covertly collected billions of photos to power a facial recognition tool marketed to police departments and private companies. They violated terms of use agreements and the expectations of those people whose images they took without consent. They enabled the adoption of a controversial technology with many data ethics concerns about privacy, bias, and surveillance overreach. Their business practices and impacts present serious data ethics challenges around ownership, sovereignty, literacy, and activism.

Evolving data systems create new ethical dilemmas at the cutting edge of science and society. However, proactive efforts around empowering users, providing education, setting limitations, and enabling collective action can help maximize data's benefits while minimizing harm.

CONCLUSION

This chapter aimed to explain the ethical, legal, and social implications of data practices from a holistic stakeholder perspective. It also covered protecting the privacy, security, and rights of data subjects and stakeholders using policies, standards, and technologies.

Ethical implications goals:

- Explain ethical impacts on stakeholders like data subjects, collectors, users, and society
- Apply data ethics principles and frameworks like FAIR, OECD, and ACM code of ethics
- Assess potential consequences and monitor/audit practices for accountability.

Legal implications goals:

- Identify relevant laws and regulations in jurisdictions like the EU, US, and China
- Comply with data protection laws like GDPR, CCPA, and CSL
- Mitigate legal risks using governance policies, security standards, and privacy-preserving technologies.

Social implications goals:

- Analyze impacts on human dignity, autonomy, justice, democracy, and the common good
- Prevent discrimination, manipulation, and accountability issues
- Empower individuals and communities through literacy programs and advocacy groups.

The chapter content reflected these goals. The ethical implications section covered topics like fairness, transparency, consent, and assessing stakeholder impacts. The legal implications section introduced major privacy laws and approaches to ensure compliance using organizational policies, industry standards, and privacy-enhancing technologies. And the social implications section analyzed issues like discrimination, manipulation, and lack of accountability enabled by opaque algorithms and data systems, proposing mitigation approaches like new public policies, algorithmic auditing methods, and public engagement.

Additional sections dove deeper into building ethics policies and governance practices within companies to codify responsible data usage rules. And the importance of developing specialized skills and thinking like ethical reasoning, judgment, leadership, and communication were emphasized to embed strong data ethics competencies across roles and organizations.

The chapter concluded by identifying newly emerging challenges as rapid technological change outpaces ethics and policy. Key areas highlighted were shifting dynamics around data ownership, sovereignty, literacy and activism. Scenarios like people losing control over their data provenance and usage, opaque algorithms manipulating users, and activism efforts prompting backlash and regulation were presented as crucial frontier issues requiring proactive problem-solving by policymakers, companies, researchers, and advocates alike.

Overall, this chapter equipped readers to actively assess and address the ethical, legal, and social impacts of data programs by learning key concepts, issues, laws, frameworks, governance best practices, methodologies, skill building, and forward-looking concerns that define modern data ethics. Readers should now be empowered to make principled, responsible decisions managing or overseeing data and AI systems in political, corporate, research, advocacy, or other contexts. They should be able to construct sound policies, practices, risk assessments, and impact evaluations for ethical data usage that respects people's rights and humanity's shared values.

- The challenges and limitations of ethical data usage, such as trade-offs, conflicts, uncertainties, and gaps.
- The best practices and recommendations for ethical data usage, such as codes of conduct, standards, guidelines, and policies.

CASE

Amnesty International case

Amnesty International (AI) has a clear standpoint on data ethics. They consider current targeted advertising practices that rely on indiscriminate corporate surveillance and profiling to be inherently incompatible with human rights and data protection principles established in the General Data Protection Regulation (GDPR) and the Charter of Fundamental Rights of the European Union.

They believe that the Digital Services Act (DSA) should impose stricter limits on the targeting of online advertising based on the processing of personal data. They urge the co-legislators to consider restrictions on targeted advertising based on invasive tracking practices, such as cross-site tracking and tracking based on sensitive data or other personal data that could lead to discriminatory outcomes.

Furthermore, Amnesty International views facial recognition technology (FRT) for identification as a mode of mass surveillance and as such, is a violation of the right to privacy. They insist that any interference with the right to privacy must always be legitimate, necessary, and proportionate.

Those are strong and clear standpoints that strive to protect the rights of people around the world. At the same time, Amnesty International is dependent on donor contributions and obtaining data about human rights from countries that aren't necessarily open to sharing the data.

- Where does AI have **ethical** challenges in the way they work and how can they potentially mitigate them?
- Where does AI have **legal** challenges in the way they work and how can they potentially mitigate them?
- What should a data ethics strategy contain for AI?
- How could data literacy be a cornerstone in the work of AI?

Additional sources:

Bibliography

Amnesty International Position on the Proposals for a Digital Services. . . . www.amnesty.eu/news/amnesty-international-position-on-the-proposals-for-a-digital-services-act-and-a-digital-markets-act/

Amnesty International Policy Recommendations on Technology and Human. . . . www.amnestyusa.org/updates/technology-human-rights-recommendations-june-2021/

Ten questions we're asking about ethics, data . . . Amnesty International. https://citizenevidence.org/2020/11/10/ethics-data-open-source/

KEY TERMS

Accountability: Data users should be responsible for the outcomes and impacts of their data usage.

Bias: Deviation from objectivity or fairness in data collection, analysis or interpretation.

CCPA: California Consumer Privacy Act that protects the privacy rights of California residents.

Competencies: Skills, knowledge, and capabilities regarding data ethics.

Consent: Data users should obtain informed and voluntary consent from data subjects before collecting and using their data.

Data ethics: The study and practice of how to collect, use, and share data in a responsible, secure, and legal way.

Data subjects: The individuals whose personal data is being collected, stored, used, or shared.

Discrimination: Unjust or prejudicial treatment of individuals or groups based on their characteristics.

Exploitation: Unfair use of data for one's benefit or advantage, often harming others.

Fairness: Data should be used for legitimate purposes and not to cause harm or discrimination.

GDPR: General Data Protection Regulation that protects the personal data of EU citizens and residents.

Governance: Processes to oversee data practices and address ethical concerns.

HIPAA: Health Insurance Portability and Accountability Act that regulates the privacy of health information in the US.

Judgment: Making principled choices about data practices.

Literacy: Skills and awareness to understand and evaluate data systems.

Manipulation: Influencing individuals or groups through data in ways that undermine autonomy or democracy.

Policies: Rules and guidelines governing data practices in an organization.

Policy: Formal rules, procedures, roles, and responsibilities regarding data practices.

Privacy: Data users should protect the confidentiality of data subjects and limit unauthorized access to their data.

Reasoning: Carefully considering data impacts through ethical frameworks.

Stakeholders: The various individuals, groups, or organizations who have an interest or concern in the data practices.

Standards: Technical specifications and best practices ensuring consistent, interoperable, and secure data.

Strategy: High-level approach and priorities regarding data ethics.

Technologies: Tools and systems that enable data collection, storage, processing, and sharing.

Transparency: Data users should be open about their data practices and allow data subjects access to their data.

REVIEW QUESTIONS

1. What does data ethics refer to?
2. Who are data subjects?
3. What are some key ethical principles for data usage?
4. What are some potential ethical risks of data-driven decision making?
5. What is bias in the context of data analysis?
6. What are some legal frameworks that regulate data practices?
7. How can policies help mitigate legal risks in data usage?
8. What are some standards that can ensure ethical data practices?
9. What technologies can help reduce manipulation risks?
10. What are the key elements of a data ethics strategy?
11. What does a data ethics policy codify?
12. What competencies are important for data ethicists?
13. How can ethical reasoning help identify data issues?

14 What are some emerging challenges around data ethics?
15 How can collective action help address data ethics concerns?

Answers to review questions

1 What does data ethics refer to? Data ethics is the study and practice of how to collect, use, and share data in a responsible, secure, and legal way. (Introduction, paragraph 1)
2 Who are data subjects? Data subjects are the individuals whose personal data is being collected, stored, used, or shared. (Ethical risk in data usage for decision-making, paragraph 1)
3 What are some key ethical principles for data usage?
4 Some key ethical principles are fairness, accountability, transparency, privacy, and consent. (Ethical risk in data usage for decision-making, paragraph 2)
5 What are some potential ethical risks of data-driven decision-making? Some risks are bias, discrimination, manipulation, exploitation, and harm. (Ethical risk categories, Table 9.1)
6 What is bias in the context of data analysis? Bias is the deviation from objectivity or fairness in data collection, analysis, or interpretation. (Ethical risk categories, Table 9.1)
7 What are some legal frameworks that regulate data practices? Some examples are GDPR, CCPA, HIPAA, and emerging AI/algorithmic regulations. (Legal risks in data usage for decision-making)
8 How can policies help mitigate legal risks in data usage? Policies establish rules and guidelines aligned with laws and ethics principles. (Policies for legal risk mitigation)
9 What are some standards that can ensure ethical data practices? Standards provide technical specifications and best practices for consistent, interoperable, and secure data. (Standards for legal risk mitigation)
10 What technologies can help reduce manipulation risks? Technologies for detecting manipulation like fact-checking and transparency generation can help. (Technologies for manipulation risk mitigation)
11 What are the key elements of a data ethics strategy? A strategy aligns with business objectives and ethical principles, guiding data practices. (Data ethics strategy and policy, paragraph 1)
12 What does a data ethics policy codify? A policy codifies rules, procedures, roles, and responsibilities around ethical data practices. (Data Ethics Strategy and Policy, paragraph 2)
13 What competencies are important for data ethicists? Important competencies are ethical reasoning, decision-making, leadership, and communication. (Data ethics skills and competencies – data literacy, paragraph 1)
14 How can ethical reasoning help identify data issues? It involves considering impacts through ethical frameworks to identify risks, harms, biases, and unintended consequences. (Data ethics skills and competencies – data literacy, paragraph 2)
15 What are some emerging challenges around data ethics? Some challenges are around data ownership, sovereignty, literacy, and activism. (Data ethics challenges and opportunities, paragraph 1)
16 How can collective action help address data ethics concerns? Collective action through campaigns, protests, and political engagement can prompt reform around data ethics. (Data Ethics Challenges and Opportunities, paragraph 5)

NOTES

[1] Business to business.

[2] The persons or organizations providing the data.

[3] Those that make decisions/take actions based on the data gathered.

[4] Meta (Facebook), Alphabet (Google), Baidu, etc.

BIBLIOGRAPHY

Apple (2023) *Privacy*. intego.com/mac-security-blog/apple-security-and-privacy-in-2023-the-year-in-review/#:~:text=In%20January%202023%2C%20Apple%20added,password%20without%20this%20hardware%20key (Accessed November 14, 2023).

Bennett, C. J., & Raab, C. D. (2020). *The governance of privacy: Policy instruments in global perspective*. Cambridge, MA: MIT Press.

Berkowitz, J. (2020). AI and data ethics: 5 principles to consider. Spark. www.adp.com/spark/articles/2020/08/ai-and-data-ethics-5-principles-to-consider.aspx#:~:text=It%20means%20thinking%20through%20what,and%20what%20might%20go%20wrong

Botsman, R. (2017). Big data meets Big Brother as China moves to rate its citizens. *Wired*, October 21. www.wired.com/story/chinese-government-social-credit-score-privacy-invasion/ (Accessed November 14, 2023).

Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*, March 17. www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election (Accessed November 14, 2023).

Calo, R. (2017). Artificial intelligence policy: A primer and roadmap. *UC Davis Law Review*, *51*(2), 399–435.

Cavoukian, A. (2011). *Privacy by design: The 7 foundational principles*. Toronto: Information and Privacy Commissioner of Ontario.

Copeland, R., & Mattioli, D. (2019). Google's "Project Nightingale" gathers personal health data on millions of Americans. *The Wall Street Journal*, November 11. www.wsj.com/articles/google-s-secret-project-nightingale-gathers-personal-health-data-on-millions-of-americans-11573496790 (Accessed November 14, 2023).

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, October 10. www.reuters.com/article/idUSKCN1MK0AG/

Data Ethics Framework. (2018). www.gov.uk/government/publications/data-ethics-framework

en.wikipedia.org. (n.d.). https://en.wikipedia.org/wiki/Health_Insurance_Portability_and_Accountability_Act

ENISA. (2015). *Privacy and data protection by design*. Heraklion: European Union Agency for Network and Information Security.

European Commission. (2016). *Data protection impact assessments*. https://ec.europa.eu/newsroom/article29/items/611236/en

European Commission. (2021). *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. https://eur-lex.europa.eu/legal-content/en/HIS/?uri=CELEX:52021PC0206

European Commission (2023) *Data protection in the EU*. https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en (Accessed November 14, 2023).

Feingold, S. (2023). *The EU's Artificial Intelligence Act, explained*. World Economic Forum. www.weforum.org/agenda/2023/06/european-union-ai-act-explained/

Friedman, B., Kahn Jr, P. H., & Borning, A. (2017). Value sensitive design and information systems. In: *Early engagement and new technologies: Opening up the laboratory*. London: Springer, pp. 55–95.

Fruhlinger, J. (2021). Equifax data breach FAQ: What happened, who was affected, what was the impact? CSO Online, January 8. www.csoonline.com/article/567833/equifax-data-breach-faq-what-happened-who-was-affected-what-was-the-impact.html (Accessed November 14, 2023).

Health Information & Privacy: FERPA and HIPAA. CDC. (n.d.). www.cdc.gov/phlp/publications/topic/healthinformationprivacy.html

Health Insurance Portability and Accountability Act of 1996 (HIPAA). (n.d.). www.cdc.gov/phlp/publications/topic/hipaa.html.

Hill, K. (2020). The secretive company that might end privacy as we know it. *The New York Times*, January 18. www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html (Accessed November 14, 2023).

HIPAA. (n.d.). HHS.gov. www.hhs.gov/programs/hipaa/index.html

HIPAA versus State Laws. HealthIT.gov – ONC. (n.d.). www.healthit.gov/topic/hipaa-versus-state-laws

Hiscock, R. (2024). The 7 principles of privacy by design. *OneTrust Blog*. www.onetrust.com/blog/principles-of-privacy-by-design/

IBM. (2023). *IBM's principles for trust and transparency*. www.ibm.com/policy/trust-principles/#:~:text=We%20believe%20that%20government%20data,and%20equitable%20and%20prioritize%20openness.&text=Clients%20are%20not%20required%20to,of%20IBM's%20solutions%20and%20services.&text=IBM%20client%20agreements%20are%20transparent (Accessed November 14, 2023).

International Association of Privacy Professionals. (n.d.). *Global AI legislation tracker*. https://iapp.org/resources/article/global-ai-legislation-tracker/

Invisibly. (2023). *Invisibly: Own your data*. www.invisibly.com/learn-blog/ (Accessed November 14, 2023).

Isaac, M. (2017). Uber's C.E.O. plays with fire. *The New York Times*, April 23. www.nytimes.com/2017/04/23/technology/travis-kalanick-pushes-uber-and-himself-to-the-precipice.html (Accessed November 14, 2023).

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2017). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, *34*(4), pp. 258–268.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. London: SAGE Publications.

Microsoft. (2023). *Microsoft privacy statement*. https://privacy.microsoft.com/en-gb/privacystatement (Accessed November 14, 2023).

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 1–21.

Młodziejewska, M., & Soller, H. (2023). Putting data ethics into practice. McKinsey. www.mckinsey.com/capabilities/mckinsey-digital/our-insights/tech-forward/putting-data-ethics-into-practice

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown.

OECD. (2019). *Enhancing access to and sharing of data: Reconciling risks and benefits for data re-use across societies*. Paris: OECD Publishing.

Rosenberg, M., Confessore, N., & Cadwalladr, C. (2018). How Trump consultants exploited the Facebook data of millions. *The New York Times*, March 17. www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html (Accessed November 14, 2023).

Sadowski, J. (2020). *Too smart: How digital capitalism is extracting data, controlling our lives, and taking over the world*. Cambridge, MA: MIT Press.

Sidewalk Labs. (2023). *Sidewalk Toronto*. www.tomorrow.city/sidewalk-toronto-the-vision-behind-googles-failed-city/ (Accessed November 14, 2023).

Summary of the HIPAA Privacy Rule. HHS.gov. (n.d.). www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

The White House. (2023). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. UNESCO. www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. New York: Public Affairs.

CHAPTER 10

Perspectives on Decision-Making Using Generative AI

Data-driven decision-making is the process of using data to inform or support decisions in various contexts. Data can provide valuable insights or evidence that can help decision-makers achieve their goals or solve their problems. However, data alone is not enough to make effective decisions. Data needs to be analyzed, interpreted, communicated, or presented in a meaningful way that can facilitate understanding or action.

Traditionally, data analysis or presentation has been done by humans using various methods or tools such as statistics, visualization, reporting, storytelling, etc. However, with the rapid growth and complexity of data, human capabilities or capacities may not be sufficient or scalable to handle the massive amount of data available. Moreover, human biases or preferences may also influence or distort the way data is analyzed or presented.

This is where generative artificial intelligence (AI) comes in. Generative AI is a branch of AI that aims to generate new data or content from existing data or content. Generative AI can create novel or realistic outputs such as texts, images, audio, or videos that can mimic or augment human creativity or expression. One of the most common forms of generative AI is natural language generation (NLG), which is the process of generating natural language texts from non-textual inputs such as data, images, audio, or videos (multimodal models).

Generative AI can create new forms of data-driven decision-making by generating insights or content from data that can enhance or complement human analysis or presentation. For example, generative AI can produce summaries, reports, stories, headlines, captions, reviews, recommendations, or feedback from data that can help decision-makers understand or communicate the key points or messages of the data. Generative AI can also generate new data or content that can expand or diversify the options or alternatives for decision-making. For example, generative AI can create new products, designs, logos, names, or slogans from data that can inspire or stimulate decision-makers to explore or discover new possibilities or opportunities.

However, generative AI is not a magic bullet that can solve all the problems or challenges of data-driven decision-making. Generative AI also comes with opportunities and risks that need to be carefully considered and managed. For example, generative AI can offer benefits such as enhancing creativity, personalization, automation, scalability, accuracy, diversity, and ethicality in decision-making. But it can also pose threats such as introducing biases, errors, inconsistencies, uncertainties, ambiguities, or harm in decision-making, as the results are not better than what the models are trained upon.

DOI: 10.4324/9781003457787-10

Therefore, it is important to develop a responsible and ethical approach to using generative AI for decision-making. This means following some best practices and guidelines that can ensure the quality and reliability of the generative AI systems and their outputs. It also means assessing and mitigating the potential impacts or consequences of the generative AI systems and their outputs on the decision-makers, the decision subjects, and the society at large.

In this chapter, we will explore how generative AI and NLG can create new forms of data-driven decision-making by generating insights or content from data. We will also evaluate the opportunities and risks of using generative AI for decision-making and develop a responsible and ethical approach to using generative AI for decision-making.

LEARNING GOALS:

L10.1 Define generative artificial intelligence (AI) and natural language generation (NLG) and explain how they can create new forms of data-driven decision-making by generating insights or content from data

L10.2 Identify the main components and steps of a generative AI system and describe how they work together to produce outputs from inputs

L10.3 Compare and contrast different types of generative AI techniques such as text generation, image generation, audio generation, and video generation, and give examples of their applications in various domains

L10.4 Evaluate the opportunities and risks of using generative AI for decision-making, such as enhancing creativity, personalization, automation, scalability, accuracy, diversity, and ethicality

L10.5 Recognize the potential sources and impacts of model training biases in generative AI and how they can affect the quality and reliability of the generated outputs and the decision-making process

L10.6 Develop a responsible and ethical approach to using generative AI for decision-making by following best practices and guidelines such as transparency, accountability, fairness, privacy, security, and human oversight.

GENERATIVE ARTIFICIAL INTELLIGENCE (AI) AND NATURAL LANGUAGE GENERATION (NLG)

In this section, we define generative AI and NLG and explain how they can create new forms of data-driven decision-making by generating insights or content from data. It will also provide a brief history and background of generative AI and NLG and their evolution over time.

The history of generative AI and NLG

Generative AI and NLG are not new concepts. They have been an active area of research since the 1950s and 1960s when researchers first began exploring the possibilities of artificial intelligence (AI). At that time, AI researchers were focused on developing rule-based systems that could simulate human thinking and decision-making. One of the earliest examples of generative AI was ELIZA, a chatbot developed by Joseph Weizenbaum in 1966, that could generate responses based on a set of predefined rules and patterns (Dale, 2023). ELIZA was one of the first examples of natural language processing (NLP), which is the branch of AI that deals with the analysis and generation of natural language texts.

However, rule-based systems had their limitations. They required a lot of manual effort to create and maintain, they were not very flexible or adaptable to new situations or domains, and they could not handle complex or ambiguous language phenomena. Therefore, researchers started to look for alternative approaches that could overcome these challenges. One of these approaches was based on the idea of using statistical models to learn from data and generate outputs based on probabilities. This approach was inspired by the work of Claude Shannon, who published his paper "A Mathematical Theory of Communication" in 1948, which introduced the concept of n-grams (Dale, 2023). N-grams are sequences of n words that can be used to estimate the likelihood of the next word given the previous words. For example, given the sentence "The cat is", an n-gram model can predict that the most probable next word is "on" or "in", based on the frequency of these words in a large corpus of text.

Statistical models became more popular and powerful with the advent of machine learning, which is the branch of AI that deals with the creation and application of algorithms that can learn from data and improve their performance over time. Machine learning techniques such as hidden Markov models (HMMs), neural networks, and support vector machines (SVMs) were applied to various NLP tasks such as speech recognition, machine translation, text summarization, and text generation. For example, in 1984, Raj Reddy and his team at Carnegie Mellon University developed a speech recognition system called HARPY, which used HMMs to recognize words from speech signals. In 1988, Rollo Carpenter created Cleverbot, a chatbot that used information retrieval techniques to generate responses by finding how a human had responded to the same question in a conversation database (Dale 2023).

However, statistical models also had their limitations. They required a lot of data to train and tune, they were not very interpretable or explainable, and they could not capture the long-term dependencies or the semantic and pragmatic aspects of natural language. Therefore, researchers started to look for more advanced approaches that could overcome these challenges. One of these approaches was based on the idea of using deep learning, which is a subfield of machine learning that deals with the creation and application of artificial neural networks that have multiple layers of processing units. Deep learning techniques such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and generative adversarial networks (GANs) were applied to various NLP tasks such as image captioning, text style transfer, text summarization, and text generation. For example, in 2014, Oriol Vinyals and his team at Google developed a neural image captioning system called NIC, which used CNNs to encode images into feature vectors and RNNs to decode them into natural language captions. In 2016, Ian Goodfellow and his team at OpenAI introduced GANs, which are composed of two neural networks: a generator that tries to produce realistic outputs from random inputs, and a discriminator that tries to distinguish between real and fake outputs.

Deep learning models became more popular and powerful with the development of large-scale pre-trained language models (LMs), which are neural networks that are trained on massive amounts of text data to learn the general patterns and structures of natural language. These LMs can then be fine-tuned or adapted to specific NLP tasks or domains using smaller amounts of task-specific data. Some examples of LMs are BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), XLNet (eXtreme Language Model), T5 (Text-To-Text Transfer Transformer), etc. These LMs use a special type of neural network architecture called Transformer, which was introduced by Ashish Vaswani and his team at Google in 2017. Transformers use attention mechanisms to learn how to focus on the most relevant parts of the input or output sequences.

LMs have achieved state-of-the-art results on various NLP tasks such as question answering, natural language inference, sentiment analysis, text summarization, and text generation. For example, in 2018, OpenAI released GPT, which was a large-scale LM that could generate coherent and diverse texts from a given prompt. In 2019, Google released BERT, which was a large-scale LM that could encode both the left and right context of a given word or sentence and improve the performance of various NLP tasks. In 2020, OpenAI released GPT-2 and GPT-3, which were larger and more powerful versions of GPT that could generate high-quality texts on various topics and domains. In 2021, OpenAI released DALL-E, which was a large-scale LM that could generate realistic images from natural language descriptions.

Table 10.1 and Figure 10.1 summarize the main milestones in the history of generative AI and NLG:

With an understanding of where we come from let’s move into the main AI components and steps of a generative AI system and how they work together to produce outputs from inputs.

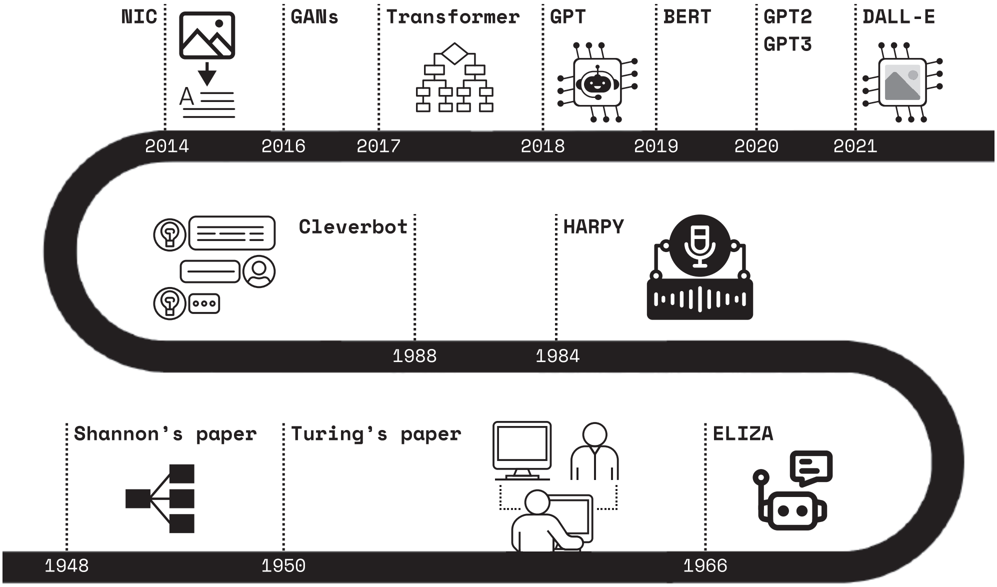


FIGURE 10.1 Timeline for NLG and generative AIs

TABLE 10.1 History of generative AI

Year	Milestone	Description
1948	Shannon's paper	Claude Shannon publishes his paper "A Mathematical Theory of Communication", which introduces the concept of n-grams.
1950	Turing's paper	Alan Turing publishes his paper "Computing Machinery and Intelligence", which introduces the Turing Test.
1966	ELIZA	Joseph Weizenbaum developed ELIZA, the first chatbot that uses rule-based natural language generation.
1984	HARPY	Raj Reddy and his team at Carnegie Mellon University developed HARPY, a speech recognition system that uses hidden Markov models.
1988	Cleverbot	Rollo Carpenter created Cleverbot, a chatbot that uses information retrieval techniques to generate responses.
2014	NIC	Oriol Vinyals and his team at Google developed NIC, a neural image captioning system that uses convolutional neural networks and recurrent neural networks.
2016	GANs	Ian Goodfellow and his team at OpenAI introduce generative adversarial networks, which are composed of two neural networks: a generator and a discriminator.
2017	Transformer	Ashish Vaswani and his team at Google introduced Transformer, a neural network architecture that uses attention mechanisms.
2018	GPT	OpenAI releases GPT, a large-scale pre-trained language model that uses Transformer to generate coherent and diverse texts.
2019	BERT	Google releases BERT, a large-scale pre-trained language model that uses a bidirectional Transformer to encode both the left and right context of a given word or sentence.
2020	GPT-2 and GPT-3	OpenAI releases GPT-2 and GPT-3, larger and more powerful versions of GPT that can generate high-quality texts on various topics and domains.
2021	DALL-E	OpenAI releases DALL-E, a large-scale pre-trained language model that can generate realistic images from natural language descriptions.

Components and steps of an AI system

A (generative) AI system is a system that uses artificial intelligence to create new content, such as text, images, music, audio, and videos, from existing content, such as data, images, audio, or videos. Both a regular and a generative AI system typically consists of the following components and steps:

Data collection: This is the process of gathering the data that will be used to train and evaluate the AI model. The data can be obtained from various sources, such as online databases, websites, social media platforms, sensors, cameras, microphones, etc. The data should be relevant and representative of the domain and the task that the AI system aims to perform. For example, if the generative AI system is designed to generate natural language texts from images, the data should consist of pairs of images and corresponding texts that describe the images.

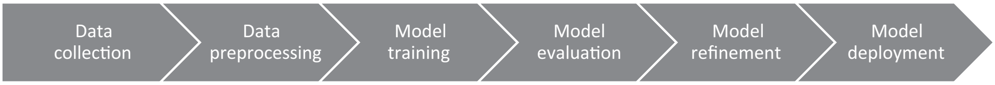


FIGURE 10.2 Process of creating an AI model

Data preprocessing: This is the process of cleaning and transforming the data into a suitable format for the AI model. The data preprocessing steps may include removing noise, outliers, duplicates, or irrelevant information from the data; resizing, cropping, or augmenting the images; converting the audio or videos into numerical features; tokenizing, lemmatizing, or stemming the texts; encoding the data into vectors or matrices; splitting the data into training, validation, and test sets; etc. The data preprocessing steps may vary depending on the type and quality of the data and the requirements of the AI model. This step can be very costly which is why LLMs are coming with this already done.

Model training: This is the process of learning the patterns and relationships in the data by using a machine learning algorithm. The machine learning algorithm can be supervised, unsupervised, or semi-supervised. Supervised learning means that the algorithm learns from labeled data, which means that each input has a corresponding output. Unsupervised learning means that the algorithm learns from unlabeled data, which means that there are no outputs given for the inputs. Semi-supervised learning means that the algorithm learns from a combination of labeled and unlabeled data. The machine learning algorithm can be based on various techniques, such as hidden Markov models (HMMs), neural networks (NNs), generative adversarial networks (GANs), variational autoencoders (VAEs), transformers (TRFs), etc. The machine learning algorithm can be implemented using various frameworks or libraries, such as TensorFlow, PyTorch, Keras, Scikit-learn, etc.

Model evaluation: This is the process of measuring the performance and quality of the generative AI model by using various metrics and criteria. The model evaluation can be done on both quantitative and qualitative aspects. Quantitative evaluation means that the model is assessed by using numerical scores or values that reflect its accuracy, precision, recall, F1-score, perplexity, BLEU (Bilingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation), etc. Qualitative evaluation means that the model is assessed by using human judgments or feedback that reflect its coherence, fluency, relevance, diversity, creativity, etc. The model evaluation can be done on both the training and test sets. The training set is used to optimize the model parameters during the model training process. The test set is used to measure the generalization ability of the model on unseen data.

Model refinement: This is the process of improving or fine-tuning the generative AI model by using various methods or techniques. The model refinement can be done by adjusting or optimizing the model parameters, such as learning rate, batch size, number of epochs, number of layers, number of units, dropout rate, activation function, loss function, optimizer, etc. The model refinement can also be done by adding or modifying some features or components of the model, such as attention mechanisms, memory networks, beam search, temperature, top-k sampling, top-p sampling, etc. The model refinement can also be done by using some external knowledge or information to guide or constrain the model generation process, such as ontologies, dictionaries, rules, templates, etc. The model refinement aims to enhance or balance some aspects of the model performance or quality, such as accuracy, diversity, creativity, consistency, fairness, robustness, etc.

All the big cloud providers offer pre-trained models where the user goes directly to model refinement, sometimes called tuning, with their own data to ensure the robustness of the models in their specific use-cases.

Model deployment: This is the process of making the AI model available and accessible for use by end-users or applications. The model deployment can be done on various platforms or environments, such as cloud computing services (e.g., Google Cloud Platform, Amazon Web Services, Microsoft Azure, etc.), edge computing devices (e.g., smartphones, tablets, laptops, etc.), web browsers (e.g., Chrome, Firefox, Safari, etc.), application programming interfaces (APIs) (e.g., RESTful APIs, GraphQL APIs, etc.), software development kits (SDKs) (e.g., Android SDK, iOS SDK, etc.), etc. The model deployment should ensure that the generative AI model is secure, scalable, reliable, and user-friendly.

In the following example, the generative AI system goes through the following components and steps to generate an image:

- Data collection: The system collects pairs of text descriptions and corresponding images from various sources, such as online databases, websites, social media platforms, etc. The text descriptions are natural language sentences that describe the content or the scene of the images. The images are realistic photos that match the text descriptions.
- Data preprocessing: The system preprocesses the data by resizing and cropping the images to a fixed size (e.g., 256 x 256 pixels); converting the images into numerical tensors; tokenizing, lemmatizing, and encoding the text descriptions into numerical vectors; splitting the data into training and test sets; etc.
- Model training: The system trains a GAN model by using a machine learning algorithm based on neural networks. The GAN model consists of two neural networks: a generator and a discriminator. The generator takes a text description and a random noise vector as inputs and generates an image as output. The discriminator takes an image and a text description as inputs and predicts whether the image is real or fake. The generator tries to fool the discriminator by generating realistic images that match the text descriptions. The discriminator tries to distinguish between real and fake images by using the text descriptions as additional information. The system uses TensorFlow as the framework to implement the GAN model.
- Model evaluation: The system evaluates the GAN model by using various metrics and criteria. The system uses quantitative metrics such as inception score (IS), Fréchet inception distance (FID), and precision and recall (PR) to measure the realism, diversity, and coverage of the generated images. The system also uses qualitative methods such as human

evaluation or user feedback to measure the coherence, fluency, relevance, diversity, and creativity of the generated images.

- Model refinement: The system refines the GAN model by using various methods or techniques. The system adjusts or optimizes some parameters of the GAN model, such as learning rate, batch size, number of epochs, number of layers, number of units, dropout rate, activation function, loss function, optimizer, etc. The system also adds or modifies some features or components of the GAN model, such as attention mechanisms, conditional batch normalization, auxiliary classifiers, etc. The system also uses some external knowledge or information to guide or constrain the GAN model generation process, such as ontologies, dictionaries, rules, templates, etc. The system aims to enhance or balance some aspects of the GAN model performance or quality, such as realism, diversity, creativity, consistency, fairness, robustness, etc.
- Model deployment: The system deploys the GAN model on a cloud computing service (e.g., Google Cloud Platform) that provides an API for users or applications to access the GAN model. The users or applications can send text descriptions to the API and receive generated images from the API. The system ensures that the GAN model is secure, scalable, reliable, and user-friendly.

This section has described the main components and steps of a generative AI system and how they work together to produce outputs from inputs. In the next section, we will compare different types of generative AI techniques such as text generation, image generation, audio generation, and video generation, and give examples of their applications in various domains. The concept of agents or RAGs has also developed with generative AI as that is a core process in enhancing the capabilities of the generative AI.

Types and applications of generative AI techniques

Generative AI techniques are methods or algorithms that can generate new content from existing content, such as data, images, audio, or videos. Generative AI techniques can be classified into different types based on the type of content they generate, such as text generation, image generation, audio generation, and video generation. In this section, we will compare these types of generative AI techniques and give examples of their applications in various domains. Multimodal models are generative AI models that are capable of producing more than one of the following output types.

Text generation

Text generation is the process of generating natural language texts from both non-textual inputs, such as tabular data, images, audio, or videos, and text prompts. Text generation can be used for various purposes, such as summarization, translation, captioning, storytelling, dialogue, etc. Text generation can also be used to create new texts that are not based on any existing inputs, such as poems, songs, jokes, etc.

Text generation can be done by using various generative AI techniques, such as hidden Markov models (HMMs), neural networks (NNs), generative adversarial networks (GANs), variational autoencoders (VAEs), transformers (TRFs), etc. For example, HMMs can be used to generate texts based on n-grams, which are sequences of n words that can be used to estimate the likelihood of the next word given the previous words. NNs can be used to generate texts based on recurrent neural networks (RNNs), which are neural networks that can capture the sequential dependencies in the texts. GANs can be used to generate texts based on generative adversarial networks, which are composed of two neural networks: a generator that tries to produce realistic texts from random inputs, and a discriminator that tries to distinguish between real and fake texts. VAEs can be used to generate texts based on variational autoencoders, which are neural networks that can learn the latent representations of the texts and generate new texts from them. TRFs can be used to generate texts based on transformers, which are neural networks that use attention mechanisms to learn how to focus on the most relevant parts of the input or output sequences.

Some examples of applications of text generation are:

- Text summarization: This is the process of generating a concise and informative summary of a longer text, such as an article, a report, a book, etc. Text summarization can help users quickly understand the main points or messages of the text without reading the whole text. That way a decision-maker can more quickly consume vast amounts of information to base a decision on.
- Text translation: This is the process of generating a text in a different language from a given text in a source language, such as English, Spanish, French, etc. Text translation can help users to communicate or access information across different languages and cultures. This opens up information from countries that would otherwise have been difficult/lengthy to access for decision-making.
- Image captioning: This is the process of generating a natural language description of an image, such as a photo, a drawing, a painting, etc. Image captioning can help users to understand or appreciate the content or the scene of the image. Makes images searchable by text prompts and therefore potentially a way to enhance the data foundation for the decisions.
- Storytelling: This is the process of generating a coherent and engaging narrative or story from a given input, such as a prompt, a genre, a character, a setting, etc. Storytelling can help users to express or explore their creativity or imagination. For example, OpenAI's ChatGPT uses storytelling to generate stories from natural language prompts. Often decisions are made in a context. Stories can help highlight the context for a decision and therefore make it more relatable.
- Dialogue: This is the process of generating a natural language conversation or dialogue between two or more agents, such as humans, chatbots, virtual assistants, etc. Dialogue can help users interact or communicate with the agents for various purposes, such as entertainment, information, assistance, etc. In a creative process, the AI dialogue can be used to expand the potential solution field.

Image generation

Image generation is the process of generating realistic or artistic images from non-image inputs, such as data, texts, audio, or videos. Image generation can be used for various purposes, such as illustration, design, synthesis, manipulation, etc. Image generation can also be used to create new images that are not based on any existing inputs, such as paintings, cartoons, logos, etc.

Image generation can be done by using many of the same generative AI techniques as text generation, such as neural networks (NNs), generative adversarial networks (GANs), variational autoencoders (VAEs), transformers (TRFs), etc. For example, NNs can be used to generate images based on convolutional neural networks (CNNs), which are neural networks that can learn the features and structures of the images. GANs can be used to generate images based on generative adversarial networks, which are composed of two neural networks: a generator that tries to produce realistic images from random inputs, and a discriminator that tries to distinguish between real and fake images. VAEs can be used to generate images based on variational autoencoders, which are neural networks that can learn the latent representations of the images and generate new images from them. TRFs can be used to generate images based on transformers, which are neural networks that use attention mechanisms to learn how to focus on the most relevant parts of the input or output sequences.

Some examples of applications of image generation are:

- Image illustration: This is the process of generating an image that illustrates or visualizes a given input, such as text, audio, video, etc. Image illustration can help users to understand or communicate the input more vividly or expressively. For example, OpenAI's DALL-E uses image illustration to generate images from natural language descriptions.
- Image design: This is the process of generating an image that follows or satisfies some design criteria or specifications, such as a style, a theme, a color, a shape, etc. Image design can help users create or discover new or original designs for various purposes, such as logos, products, marketing images, architecture, etc.
- Image synthesis: This is the process of generating an image that combines or blends two or more images, such as a face, a background, a foreground, etc. Image synthesis can help users create or modify images for various purposes, such as entertainment, education, marketing, etc.
- Image manipulation: This is the process of generating an image that changes or alters some aspects of an existing image, such as the content, the style, the quality, the resolution, etc. Image manipulation can help users enhance or transform images for various purposes, such as editing, restoration, improvement, etc. This is also sometimes called "deep fakes" if they are used in a manipulative way.

Audio generation

Audio generation is the process of generating realistic or artistic audio from non-audio inputs, such as data, texts, images, or videos. Audio generation can be used for various purposes, such as speech, music, sound effects, etc. Audio generation can also be used to create new audio that is not based on any existing inputs, such as speeches, songs, melodies, noises, etc.

Generative AI techniques, such as neural networks (NNs), generative adversarial networks (GANs), variational autoencoders (VAEs), and transformers (TRFs), can be employed for audio generation. For instance, recurrent neural networks (RNNs), a type of NN, can generate audio by capturing sequential dependencies in the audio data. GANs, which consist of a generator network that attempts to create realistic audio from random inputs and a discriminator network that tries to differentiate between real and synthetic audio, can also be used for audio generation. VAEs, another type of NN, can generate audio by learning the latent representations of the audio data and creating new audio from these representations. Lastly, TRFs, which are NNs that utilize attention mechanisms to concentrate on the most relevant parts of the input or output sequences, can be used for audio generation. Some examples of applications of audio generation are:

- Speech generation: This is the process of generating natural language speech from non-speech inputs, such as data, texts, images, or videos. Speech generation can help users to understand or communicate the input more naturally or conveniently. This is showing promise in the call center business which is already using scripts for call center agents.
- Music generation: This is the process of generating musical sounds or compositions from non-musical inputs, such as data, texts, images, or videos. Music generation can help users to express or explore their creativity or emotions. The soundscape relevant to marketing materials can be created here instead of hiring a composer and recording artist.
- Sound effect generation: This is the process of generating sound effects that match or enhance a given input, such as a text, an image, a video, etc. Sound effect generation can help users to create or improve the audiovisual experience of the input.

Video generation

Video generation is the process of generating realistic or artistic videos from non-video inputs, such as data, texts, images, or audio. Video generation can be used for various purposes, such as animation, simulation, synthesis, manipulation, etc. Video generation can also be used to create new videos that are not based on any existing inputs, such as movies, cartoons, games, etc.

Many of the same generative AI techniques can be utilized for video generation as well. Some examples of applications of video generation are:

- Video animation: This is the process of generating an animated video that depicts or illustrates a given input, such as text, an image, an audio, etc. Video animation can help users to understand or communicate the input more dynamically or expressively. This could be relevant for manuals and assembly instructions or illustrating complex interactions.
- Video simulation: This is the process of generating a simulated video that models or predicts a given scenario, such as a physical phenomenon, a social behavior, a future event, etc. Video simulation can help users to explore or test the effects or outcomes of the scenario.
- Video synthesis: This is the process of generating a video that combines or blends two or more videos, such as a face, a background, a foreground, etc. Video synthesis can help users create or modify videos for various purposes, such as entertainment, education, research, etc.
- Video manipulation: This is the process of generating a video that changes or alters some aspects of an existing video, such as the content, the style, the quality, the resolution, etc. Video manipulation can help users enhance or transform videos for various purposes, such as editing, restoration, improvement, etc.

For the most part, the results of the generative text, image, audio, and video still need some human editing before being presented in a business context, but developments are very fast here. It does, however, require that you are asking the right questions and that you know what you want. Anyone working for customers, clients, or partners knows that is not a given thing.

Retrieval augmented generation to expand the models

Augmenting the generative AI models is one of the solutions to improve the predictability and accuracy of the AI models that wouldn't otherwise understand the context in which they have been prompted.

RAG and the use of agents together with generative AI models and in particular LLM generative AI models are some of the recent developments in the field of natural language generation (NLG). NLG is the process of generating natural language texts from non-textual inputs, such as data, images, audio, or videos. NLG can be used for various purposes, such as summarization, translation, captioning, storytelling, dialogue, etc.

RAG stands for retrieval-augmented generation, which is a technique that allows generative AI models to access external information from e.g., vector databases to augment their generation process. A vector database is a collection of documents or chunks of text that are embedded into numerical vectors using an embedding model. The embedding model is a machine learning model that can represent texts as high-dimensional vectors that capture their semantic meanings. RAG works by taking a query (such as a question or a prompt) and passing it to the embedding model to obtain an embedded query vector. Then, the embedded query vector is passed to the vector database to retrieve the top-k relevant documents or chunks of text that are closest to the query vector in terms of distance. These retrieved documents or chunks of text are then passed to the generative AI model along with the query to generate a response. The generative AI model can use the retrieved information to enhance or control the quality, diversity, relevance, and accuracy of the generated response. For example, if the query is "Who is the president of France?", the RAG technique can retrieve some documents or chunks of text that contain information about the president of France, such as his name, biography, political party, etc. This retrieved information can then be used by the generative AI model to generate a response, such as "The president of France is Emmanuel Macron, who is the leader of the La République En Marche! party and has been in office since 2017" (Mohandas & Moritz, 2023).

Agents are interactive systems that can generate natural language texts based on the user's input and feedback. Agents can use the language understanding power of generative AI models to make a plan on how to solve a given problem or task. The plan consists of a sequence of steps or actions that the agent can take to achieve the desired goal or outcome. The agent can also execute the actions by using various tools or APIs that can perform the actions. For example, if the agent needs to search for some information, it can use a search tool or API to query the internet or a vector database. If the agent needs to generate some text, it can use a generative AI model to produce the text. The agent can also interact with the user by asking questions, providing feedback, or requesting confirmation. The agent can use the user's input and feedback to update or refine its plan or actions. For example, if the user asks the agent "How can I learn Ray?", the agent can use a generative AI model to understand the user's query and make a plan to answer it. The plan may consist of the following steps or actions:

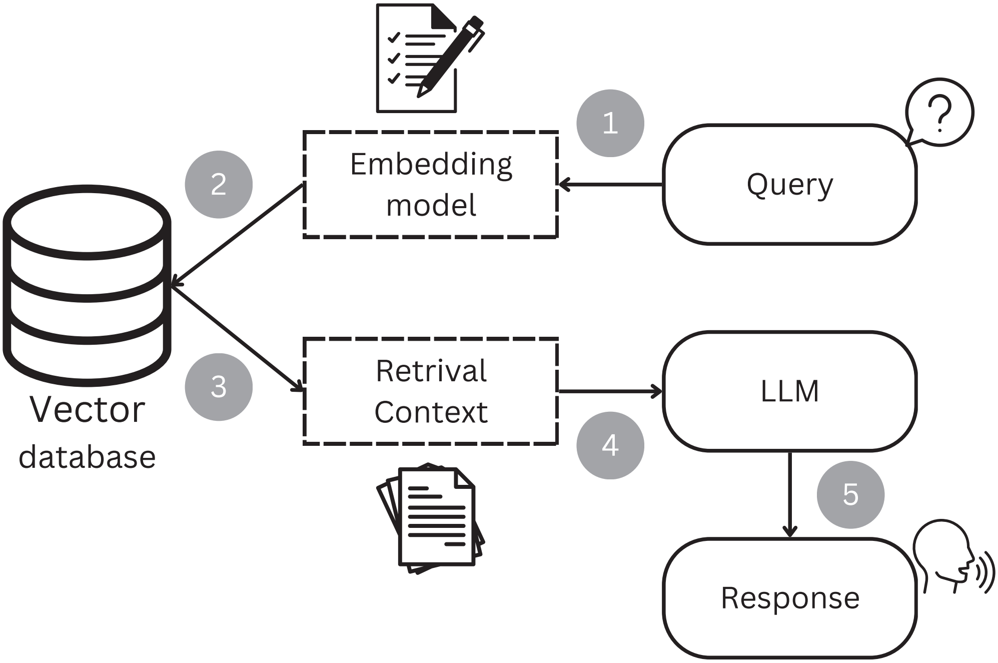


FIGURE 10.3 Retrieval augmented generation (RAG) for large language models

1 Search for Ray on the internet or a vector database and retrieve some information about it, such as its definition, features, benefits, etc.
2 Generate a text that summarizes the information and explains what Ray is and why it is useful.
3 Search for Ray tutorials or courses on the internet or a vector database and retrieve some links or resources that can help the user learn Ray.
4 Generate a text that provides the links or resources and suggests the user follow them.
5 Ask the user if the answer is satisfactory or if the user has any further questions.

The agent can also execute the actions by using various tools or APIs, such as a search tool or API to query the internet or a vector database, a generative AI model to generate the texts, and a dialogue tool or API to interact with the user. The agent can also use the user's input and feedback to update or refine its plan or actions. For example, if the user says "Yes, thank you. But can you also tell me how to install Ray?", the agent can use a generative AI model to understand the user's feedback and make a new plan to answer it. The new plan may consist of the following steps or actions:

- Search for Ray installation instructions on the internet or a vector database and retrieve some information about it, such as the requirements, the commands, the steps, etc.
- Generate a text that summarizes the information and explains how to install Ray on the user's device
- Ask the user if the answer is satisfactory or if the user has any further questions.

RAG and the use of agents together with generative AI models and in particular LLM generative AI models are some of the ways to create new forms of data-driven decision-making by generating insights or content from data. They can enhance the creativity, personalization, automation, scalability, accuracy, diversity, and ethicality of the generative AI models and their outputs. They can also enable more complex and interactive applications that can solve various problems or tasks in various domains. For example, RAG and agents can be used to create assistants that can answer questions, provide recommendations, give feedback, or perform actions for the users. They can also be used to create applications that can generate texts, images, audio, or videos from various inputs, such as prompts, genres, characters, settings, etc.

This section has compared and contrasted different types of generative AI techniques such as text generation, image generation, audio generation, and video generation, and given examples of their applications in various domains along with a method of enhancing the generation. In the next section, we will evaluate the opportunities and risks of using generative AI for decision-making, such as enhancing creativity, personalization, automation, scalability, accuracy, diversity, and ethicality.

Opportunities and risks of using generative AI for decision-making

Generative AI is a powerful and promising technology that can create new forms of data-driven decision-making by generating insights or content from data. However, generative AI also poses some opportunities and risks that need to be carefully considered and managed. In this section, we will evaluate some of the main opportunities and risks of using generative AI for decision-making, such as enhancing creativity, personalization, automation, scalability, accuracy, diversity, and ethicality.

Creativity: Generative AI can enhance the creativity of decision-making by producing novel or original outputs that can inspire or stimulate decision-makers to explore or discover new possibilities or opportunities. For example, generative AI can generate new products, designs, logos, names, or slogans from data that can help decision-makers create or innovate new solutions or strategies. However, generative AI can also pose a risk of reducing the creativity of decision-making by replacing or overshadowing human input or contribution. For example, generative AI can generate texts, images, audio, or videos that can influence or manipulate decision-makers to accept or follow the generated outputs without questioning or challenging them.

Personalization: Generative AI can enhance the personalization of decision-making by producing customized or tailored outputs that can match or satisfy the preferences or needs of decision-makers or decision subjects. For example, generative AI can generate recommendations, feedback, or guidance from data that can help decision-makers make better or more informed decisions. However, generative AI can also pose a risk of compromising the personalization of decision-making by violating or ignoring the privacy or consent of decision-makers or decision subjects. For example, it can generate texts, images, audio, or videos that can reveal or expose sensitive or confidential information about decision-makers or decision subjects without their permission or awareness.

Automation: Generative AI can enhance the automation of decision-making by producing fast or efficient outputs that can save or optimize the time or resources of decision-makers or decision subjects. For example, it can generate summaries, reports, stories, headlines, captions, reviews, or feedback from data that can help decision-makers understand or communicate the key points or messages of the data without reading or writing the whole data. However, generative AI can also pose a risk of undermining the automation of decision-making by introducing errors or inconsistencies in the generated outputs that can affect or impair the quality or reliability of the decision-making process. For example, it can generate texts, images, audio, or videos that can contain or propagate false or misleading information that can confuse or mislead decision-makers or decision subjects.

Scalability: Generative AI can enhance the scalability of decision-making by producing large or diverse outputs that can cover or address the scope or complexity of the decision-making problem or task. For example, it can generate texts, images, audio, or videos that can expand or diversify the options or alternatives for decision-making. However, generative AI can also pose a risk of overwhelming the scalability of decision-making by producing too many or too varied outputs that can exceed or challenge the capacity or capability of decision-makers or decision subjects. For example, it can generate texts, images, audio, or videos that can create or increase the cognitive or emotional load or stress of decision-makers or decision subjects.

Accuracy: Generative AI can enhance the accuracy of decision-making by producing precise or correct outputs that can reflect or support the facts or evidence of the decision-making problem or task. For example, it can generate texts, images, audio, or videos that can provide or verify the information or data that are relevant or necessary for decision-making. However, generative AI can also pose a risk of compromising the accuracy of decision-making by producing inaccurate or incorrect outputs that can contradict or undermine the facts or evidence of the decision-making problem or task. For example, it can generate texts, images, audio, or videos that can contain or generate biases, errors, inconsistencies, uncertainties, ambiguities, or harms that can affect or impair the quality or reliability of the decision-making process.

Diversity: Generative AI can enhance the diversity of decision-making by producing varied or inclusive outputs that can represent or respect the differences or perspectives of decision-makers or decision subjects. For example, it can generate texts, images, audio, or videos that can capture or express the cultural, linguistic, gender, racial, or other dimensions of diversity that are relevant or important for decision-making. However, generative AI can also pose a risk of reducing the diversity of decision-making by producing uniform or exclusive outputs that can ignore or marginalize the differences or perspectives of decision-makers or decision subjects. For example, it can generate texts, images, audio, or videos that can reflect or reinforce the stereotypes, prejudices, discriminations, or oppressions that can affect or harm the decision-makers or decision subjects.

Ethicality: Generative AI can enhance the ethicality of decision-making by producing fair or responsible outputs that can align or comply with the values or principles of decision-makers or decision subjects. For example, it can generate texts, images, audio, or videos that can follow or adhere to the ethical standards or guidelines that are established or accepted for decision-making. However, generative AI can also pose a risk of compromising the ethicality of decision-making by producing unfair or irresponsible outputs that can violate or conflict with the values or principles of decision-makers or decision subjects. For example, it can generate texts, images, audio, or videos that can infringe or harm the rights, interests, or welfare of decision-makers or decision subjects.

Generative AI model evaluation

The accuracy of a generative AI model is the degree to which the model can produce realistic or correct outputs from given inputs. The accuracy of a generative AI model can be evaluated by using various methods or techniques, depending on the type and purpose of the model and the output. Some of the common methods or techniques are:

Quantitative evaluation: This is the method of measuring the accuracy of the generative AI model by using numerical scores or values that reflect some aspects of the output quality, such as realism, diversity, coverage, coherence, fluency, relevance, etc. Quantitative evaluation can be done by using various metrics or criteria, such as inception score (IS), Fréchet inception distance (FID), precision and recall (PR), perplexity, BLEU, ROUGE, etc. Quantitative evaluation can be done by comparing the generated outputs with the ground truth outputs (if available) or with some reference outputs (such as human-generated outputs or outputs from other models). For example, FID is a metric that measures the distance between the feature distributions of the generated images and the real images (Betzalel et al., 2022). PR is a metric that measures the ratio of the generated texts that are relevant and novel to the input texts (Mohandas, & Moritz, 2023).

Qualitative evaluation: This is the method of measuring the accuracy of the generative AI model by using human judgments or feedback that reflect some aspects of the output quality, such as realism, diversity, coverage, coherence, fluency, relevance, etc. Qualitative evaluation can be done by using various methods or techniques, such as human evaluation, user feedback, user study, etc. Qualitative evaluation can be done by asking humans to rate, rank, or compare the generated outputs based on some criteria or questions. For example, human evaluation is a method that asks humans to rate the generated outputs on a scale from 1 to 5 based on some criteria, such as how realistic, coherent, or relevant the outputs are. User feedback is a method that asks users to provide comments or suggestions on the generated outputs, such as what they like or dislike about the outputs, or how they can improve the outputs.

The accuracy of a generative AI model can vary depending on the type and purpose of the model and the output. Therefore, it is important to choose the appropriate methods or techniques for evaluating the accuracy of the generative AI model. It is also important to consider the trade-offs and limitations of the methods or techniques, such as the validity, reliability, scalability, and interpretability of the evaluation results.

In most business contexts the qualitative evaluation will have to suffice as the quantitative evaluation will be too complex in the time available, but it must also be remembered that model evaluation is not a ‘one-off’ process. Due to the potential of model drift, and the fact that the model and the data might change over time, evaluations will have to be done at regular intervals. An example of a risky model drift is that the inflation changes in a country and the weights of the model change faster or slower than purchase behaviors. Marketing automation would then start running campaigns that did not match the consumers.

This section has evaluated some of the main opportunities and risks of using generative AI for enhancing decision-making. In the next section, we will develop a responsible and ethical approach to using generative AI for decision-making by following best practices and guidelines such as transparency, accountability, fairness, privacy, security, and human oversight.

Responsible and ethical approach to using generative AI for decision-making

This section will develop a responsible and ethical approach to using generative AI for decision-making by following best practices and guidelines. It will provide some principles and frameworks for designing, developing, deploying, and using generative AI systems responsibly and ethically. It will also suggest some methods and tools for assessing and mitigating the potential biases in generative AI systems.

Generative AI is a powerful and promising technology that can create new forms of data-driven decision-making by generating insights or content from data. However, generative AI also faces some challenges and limitations that need to be carefully considered and addressed. Some of the main challenges and limitations of generative AI are:

Data quality: This is the challenge of ensuring that the data used to train/tune and evaluate the generative AI models are accurate, complete, consistent, and relevant. Data quality can affect the performance and quality of the generative AI models and their outputs. For example, if the data contains (systematic) noise, outliers, duplicates, or irrelevant information, the generative AI models may learn or generate incorrect or misleading outputs. Therefore, it is important to use data preprocessing techniques, such as cleaning, transforming, encoding, or splitting the data, to improve the data quality (FP Team, 2023).

Data availability: This is the challenge of obtaining enough data to train and evaluate the generative AI models. Data availability can affect the scalability and generalization ability of the generative AI models and their outputs. For example, if the data is scarce, incomplete, or imbalanced, the generative AI models may not be able to learn or generate diverse or representative outputs. Therefore, it is important to use data augmentation techniques, such as generating, synthesizing, or expanding the data, to increase the data availability (Harvard Online, 2023).

Data ownership: This is the challenge of respecting and protecting the rights and interests of the data owners or providers. Data ownership can affect the ethicality and legality of the generative AI models and their outputs. For example, if the data is obtained or used without the permission or consent of the data owners or providers, the generative AI models may infringe or harm their rights. Therefore, it is important to use data governance techniques, such as establishing or following the data policies, standards, or agreements, to ensure data ownership (Tang, 2023).

Data privacy: This is the challenge of preserving and safeguarding the privacy and confidentiality of the data subjects or users. Data privacy can affect the personalization and security of the generative AI models and their outputs. For example, if the data is revealed or exposed to unauthorized or malicious parties, the generative AI models may violate or compromise the privacy or confidentiality of the data subjects or users. Therefore, it is important to use data protection techniques, such as encrypting, anonymizing, or deleting the data, to ensure data privacy. The current LLMs are trained on large parts of the available internet, which also includes information that is not there intentionally.

Model robustness: This is the challenge of ensuring that the generative AI models can handle or adapt to the changes or uncertainties in the data or the environment. Model robustness can affect the accuracy and stability of the generative AI models and their outputs. For example, if the data or the environment are noisy, ambiguous, or adversarial, the generative AI models may generate or suffer from errors, or inconsistencies. Therefore, it is important to test, debug, and verify model robustness, to ensure the model robustness.

Model accountability: This is the challenge of assigning and enforcing the responsibilities and liabilities of the generative AI models and their outputs. Model accountability can affect the fairness and ethicality of the generative AI models and their outputs. For example, if the generative AI models or their outputs are biased, discriminatory, or harmful, they may not be able to acknowledge or correct their outputs. Therefore, it is important to use model accountability techniques, such as monitoring, evaluating, or regulating the generative AI models or their outputs, to ensure model accountability.

Those are some of the challenges and limitations of generative AI that need to be carefully considered and addressed. By using data preprocessing, data augmentation, data governance, etc., we can improve the performance and quality of the generative AI models and their outputs. We can also ensure the responsible and ethical use of generative AI for decision-making.

Generative AI tools are powerful and promising technologies that can create new forms of data-driven decision-making by generating insights or content from data. However, generative AI tools also pose some ethical and social challenges and risks that need to be carefully considered and addressed. Therefore, it is important to follow some responsible and ethical best practices for using generative AI tools, such as:

- Understanding the limitations of AI models
- addressing biases
- maintaining transparency
- securing sensitive information
- ensuring accountability and explainability
- continuous monitoring
- educating and empowering users
- addressing legal and ethical concerns
- collaborating with experts and stakeholders, and
- establishing content guidelines.

These are vital steps towards mitigating risks associated with generative AI (Croak & Gennai, 2023). Those headlines come from one of the key industry players in generative AI, Google. There is at the same time significant activity in academia where Brian Spisak et al. outline 13 principles for using AI responsively (Spisak, Rosenberg, & Beilby, 2023). The principles or steps they propose are:

- Informed consent
- Aligned interests
- Opt in and easy exits
- Conversational transparency
- Debiased and explainable AI
- AI training and development
- Health and well-being – check the employee impact
- Data collection check
- Data sharing
- Privacy and security
- Third party disclosure
- Communication of changes in the collection of data
- Laws and regulations upholding must be communicated.

Most of the principles will be upheld if you follow the European GDPR privacy legislation and/or the AI Act.

These are some of the responsible and ethical best practices for using generative AI tools that can help to balance the benefits and risks of this potent technology. By following these best practices, we can improve the performance and quality of the generative AI tools and their outputs, as well as ensure the responsible and ethical use of generative AI for decision-making.

CASE: FACTIVE, MARKETING AUTOMATION, AUSTRIA

Factive is a firm dedicated to facilitating change in the world by assisting individuals and businesses. They substitute traditional data collection and research methods with dynamic data, which they believe is more precise, predictive, and efficient. This method allows for more focus on decision-making, which is crucial for change.

Factive provides a variety of tools designed to identify the quickest paths to growth. These tools assist in identifying the largest volumes of existing consumer behavior that a brand or product could become a part of, and the most relevant new propositions with the highest commercial potential. Instead of qualifying insights to speculate and hypothesize about future action, they enable rapid prototyping of scenarios and validation from large-scale dynamic data. This is in line with the design thinking methodologies. The core of their solution, Resonans, is used to test as wide a variety of solutions as possible. This is done by making variants of products and marketing messages to a big panel of internet respondents. They then evaluate whether it makes them want "to learn more", "to take action", or whether they "don't find it relevant".

Their offerings are designed for strategy, innovation, new product development (NPD), brand, and marketing development. Their main objective is to enable consultancies, specialists, and departments to have more confidence in their decisions' outcomes and allocate more time and resources to the implementation required to realize it. They offer Software as a Service (SaaS), white label, and project-based solutions (Factive website n.d.).

In 2022 ChatGPT was going mainstream and creating a rush in the demand for AI solutions. It was, however, not the boom that Thomas Thorstholm, managing partner at Factive had hoped. Now everyone thought they could just ask the chatbot to do the analysis Factive were providing. However, one key insight dawned on the Factive management when they started using generative AI themselves. They could use it for "divergent thinking", meaning making different solutions, instead of trying to make the "right solution". They provide the main solutions and let the AI generate all the variations, which would usually be discarded due to lack of time. That could be small target groups and solutions rarely selected

Question 1: How could Factive use generative AI to support divergent thinking
Question 2: Thomas doesn't think that advertising agencies will become irrelevant due to generative AI. What could be his reasons for thinking that? What do you think?

CONCLUSION

In closing, this chapter has explored the emergent domain of using generative artificial intelligence for enhanced data-driven decision-making. We have surveyed the history of generative AI techniques like natural language generation and analyzed their components, capabilities, and applications across modalities including text, image, audio, and video. While these technologies harbor immense potential for augmenting human creativity and knowledge work, we must also remain cognizant of the risks, from perpetuating biases to enabling misinformation.

By evaluating opportunities and challenges through a lens of responsible innovation, we can develop ethical frameworks to guide the design and deployment of generative AI systems. Industry leaders have an imperative to implement best practices around transparency, accountability, security, and alignment with stakeholder values. Policymakers must also play a role in establishing ambits of appropriate use. With diligence and collaboration, we can harness the power of generative AI to expand the frontiers of insight without compromising our shared values.

The path ahead will involve continuous co-creation between humans and intelligent machines. As generative models rapidly advance in sophistication, we must be proactive in directing their progress toward enlightened outcomes that uplift society. If we build upon the foundations established in this chapter with care and wisdom, these technologies hold profound potential to enhance decision-making and unlock new realms of human understanding.

KEY TERMS

Beam search: Technique to generate more fluent and coherent text by maintaining multiple candidate outputs.

BLEU: Metric for evaluating the similarity of generated to reference text.

Conditional batch normalization: Technique to control the style of generated images.

Data augmentation: Techniques like synthesis and expansion to increase the size and diversity of training data.

Data collection: Gathering relevant data to train and evaluate generative AI models.

Data preprocessing: Cleaning and formatting data to prepare it for training generative AI models.

Fréchet inception distance (FID): Metric for measuring the similarity of generated images to real images.

Generative AI: A branch of artificial intelligence focused on generating new data or content from existing inputs.

Inception score (IS): Metric for evaluating the quality and diversity of generated images.

Language model (LM): Neural network trained on large amounts of text data to learn linguistic patterns.

Model deployment: Making trained generative AI models available for use via APIs, web apps, etc.

Model evaluation: Assessing the performance and quality of trained generative AI models.

Model refinement: Improving generative AI models by adjusting parameters or adding new components.

Model training: Using machine learning algorithms to train generative AI models on data.

Natural language generation (NLG): Generative AI technique for producing text from non-textual inputs.

Ontologies: Formal representations of concepts and relationships used to inform generative models.

Perplexity: Metric for measuring fluency and predictability of generated text.

Precision and recall: Metrics for evaluating relevance and coverage of generated text.

Qualitative evaluation: Human judgment of generative model outputs based on criteria like realism, and coherence.

Quantitative evaluation: Numerical assessment of generative model accuracy using metrics.

Retrieval augmented generation (RAG): Technique to enhance NLG by retrieving relevant information from a database to inform text generation.

ROUGE: Metric for assessing the quality of text summarization by comparing it to reference summaries.

Rules: Logical constraints are provided to generative models to make outputs more consistent.

Templates: Predefined structures given to generative models to control the output format.

Top-k sampling: Generating text by sampling from the k most likely next tokens at each step.

Top-p sampling: Generating text by sampling from the smallest set of tokens whose cumulative probability exceeds p.

Transformer: Neural network architecture using attention mechanisms, commonly used for NLG models.

REVIEW QUESTIONS

1 What is generative AI and how does it differ from other branches of AI?
2 What is natural language generation and how has it evolved over the past few decades?
3 What are the main components and steps involved in developing a generative AI system?
4 What are the key types of generative AI and what are some examples of their applications?
5 How can retrieval augmented generation enhance the capabilities of natural language generation models?
6 What are some of the main opportunities for using generative AI for decision-making?
7 What risks or challenges need to be considered when using generative AI for decision-making?
8 How can the accuracy of a generative AI model be evaluated quantitatively and qualitatively?
9 What metrics are commonly used to assess the quality of generated text, images, and videos?
10 What techniques can be used to refine and improve the performance of a trained generative AI model?
11 What are some best practices for ethically and responsibly developing, deploying, and using generative AI systems?
12 How can data quality, privacy, security, fairness, accountability, etc. be addressed when working with generative AI?
13 What should be considered when deploying a trained generative AI model for real-world usage?
14 Why is it important to continuously monitor, evaluate, and update generative AI systems after deployment?

Answers to review questions

Here are potential answers to the review questions with references to the relevant sections of the textbook chapter:

1 Generative AI focuses on creating new data or content from existing inputs, as opposed to other branches like analysis, classification, or control (Introduction)
2 NLG has evolved from rule-based to statistical to deep learning models over 50+ years, with key milestones like ELIZA, GPT, BERT, and DALL-E (Generative AI and NLG section)
3 The main components are data collection, preprocessing, model training/evaluation/refinement, and deployment. Steps involve data gathering, cleaning, training ML models,

assessing performance, improving models, and deployment via APIs (Components and steps section)

4 Key types are text, image, audio, and video generation. Applications include summarization, translation, captioning, animation, and manipulation (Types and applications section)

5 RAG enhances NLG by retrieving relevant information from a database to inform text generation (Types and applications section)

6 Opportunities include enhanced creativity, personalization, automation, scalability, accuracy, and diversity (Opportunities and risks section)

7 Risks include biases, errors, privacy/security threats, and lack of oversight. Accuracy depends on data and model quality (Opportunities and risks section)

8 Accuracy can be evaluated quantitatively using metrics like IS, FID, PR, perplexity, etc., and qualitatively via human assessment of criteria like realism, and coherence (Opportunities and risks section)

9 Metrics include IS, FID for images; BLEU, ROUGE for text; and perplexity for language fluency (Opportunities and risks section)

10 Model refinement techniques include parameter tuning, adding new components like attention, and using knowledge sources like ontologies (Components and steps section)

11 Best practices include transparency, testing for biases, protecting rights/interests, ensuring accountability, and aligning with stakeholder values (Responsible and ethical approach section)

12 Methods include data governance, encryption, anonymization, fairness/bias mitigation techniques, monitoring, and evaluation (Responsible and ethical approach section)

13 Consider scalability, reliability, security, and ease of integration when deploying models via APIs, web apps, etc. (Components and steps section)

14 Continuous monitoring, user feedback incorporation, and model updating are critical after deployment to maintain performance and ethical compliance (Introduction, Conclusion)

BIBLIOGRAPHY

Betzalel, E., Penso, C., Navon, A., & Fetaya, E. (2022). A study on the evaluation of generative models. arXiv preprint arXiv:2206.10935. https://arxiv.org/abs/2206.10935

Croak, M., & Gennai, J. (2023, July 27). 3 emerging practices for responsible generative AI. *Google AI Blog*. https://blog.google/technology/ai/google-responsible-generative-ai-best-practices/

Dale, R. (2023). Navigating the text generation revolution: Traditional data-to-text NLG companies and the rise of ChatGPT. *Natural Language Engineering*, *29*(4), 1188–1197. https://doi/10.1017/S1351324923000347

Factive. (n.d.). Factive. www.factive.io/

FP Team. (2023, January 17). *Generative AI: Advantages, disadvantages, limitations, and challenges*. Fact Technology. https://fact.technology/learn/generative-ai-advantages-limitations-and-challenges/

Harvard Online. (2023, April 19). *The benefits and limitations of generative AI: Harvard experts answer your questions*. www.harvardonline.harvard.edu/blog/benefits-limitations-generative-ai

Kanjee, Z., Crowe, B., & Rodman, A. (2023). Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA*, *330*(1), 78–80. https://doi.org/10.1001/jama.2023.8288

Mohandas, G., & Moritz, P. (2023, October 25). Building RAG-based LLM applications for production. *Anyscale*. www.anyscale.com/blog/a-comprehensive-guide-for-building-rag-based-llm-applications-part-1

Spisak, B., Rosenberg, L. B., & Beilby, M. (2023, June 30). 13 principles for using AI responsibly. *Harvard Business Review*. https://hbr.org/2023/06/13-principles-for-using-ai-responsibly

Tang, Y. (2023, February 22). Current problems with generative AI: An exploration. *Dev Genius*. https://blog.devgenius.io/current-problems-with-generative-ai-an-exploration-747c4f150d9c

Index

Please note that page references to Figures will be in **bold**, while references to Tables are in *italics*.