Data-Driven Decision Making: *Penerapan Algoritma-Algoritma Sains Data*

Achmad Benny Mutiara

Medan, 16 November 2024

BAGIAN 1

DATA-DRIVEN DECISION MAKING

Apa itu Data-Driven Decision Making (DDDM)?

Data-Driven Decision Making mengacu pada *pengumpulan, analisis, pemeriksaan, dan interpretasi data yang sistematis, biasanya melalui penerapan metode dan teknik* ***analitik*** *atau* ***pembelajaran mesin****, untuk mencapai keputusan yang tepat.*

(Elgendy, Elragal, & Päivärintaa, 2022)

Apa itu Data-Driven Decision Making (DDDM)?

- Dalam konteks ini, data menjadi pondasi utama dalam menentukan arah strategi dan tindakan yang akan diambil oleh perusahaan atau organisasi.
- **Data-Driven Decision Making** melibatkan pengumpulan data yang relevan, analisis mendalam, dan interpretasi yang akurat untuk menghasilkan ***insight*** yang dapat diimplementasikan.
- **Pendekatan ini berbeda** dengan metode tradisional yang sering mengandalkan intuisi atau pengalaman subjektif. Dengan **Data-Driven Decision Making**, keputusan dibuat berdasarkan bukti nyata yang diperoleh dari data, sehingga mengurangi risiko kesalahan dan meningkatkan akurasi hasil.

Sejarah DDDM

- **Data-Driven Decision Making** bukanlah konsep baru. Hanya saja jumlah data yang tersedia dan alat untuk menganalisisnya baru.
- Itu juga berarti bahwa untuk memahami **DDDM**, kita harus kembali sedikit dan mengeksplorasi bagaimana keputusan awalnya terlihat telah dibuat dan bagaimana ketersediaan data dan alat berdampak padanya.

Sejarah DDDM

Tiga perspektif berbeda secara historis telah menjadi fokus ketika menggambarkan pengambilan keputusan:

- **Decision-making process**: berkaitan dengan struktur proses.
- **Decision-maker**: Orang yang membuat keputusan akhir akan membawa preferensi dan pengalaman pribadinya ke dalam proses pengambilan keputusan.
- **Decision quality (type)**: termasuk ketepatan waktu, akurasi, dan kebenaran Keputusan. Itu juga berarti bahwa konsekuensi dari keputusan dan waktu yang tersedia harus dipertimbangkan sebagai elemen

Sejarah DDDM: *Classical Decision Theories*

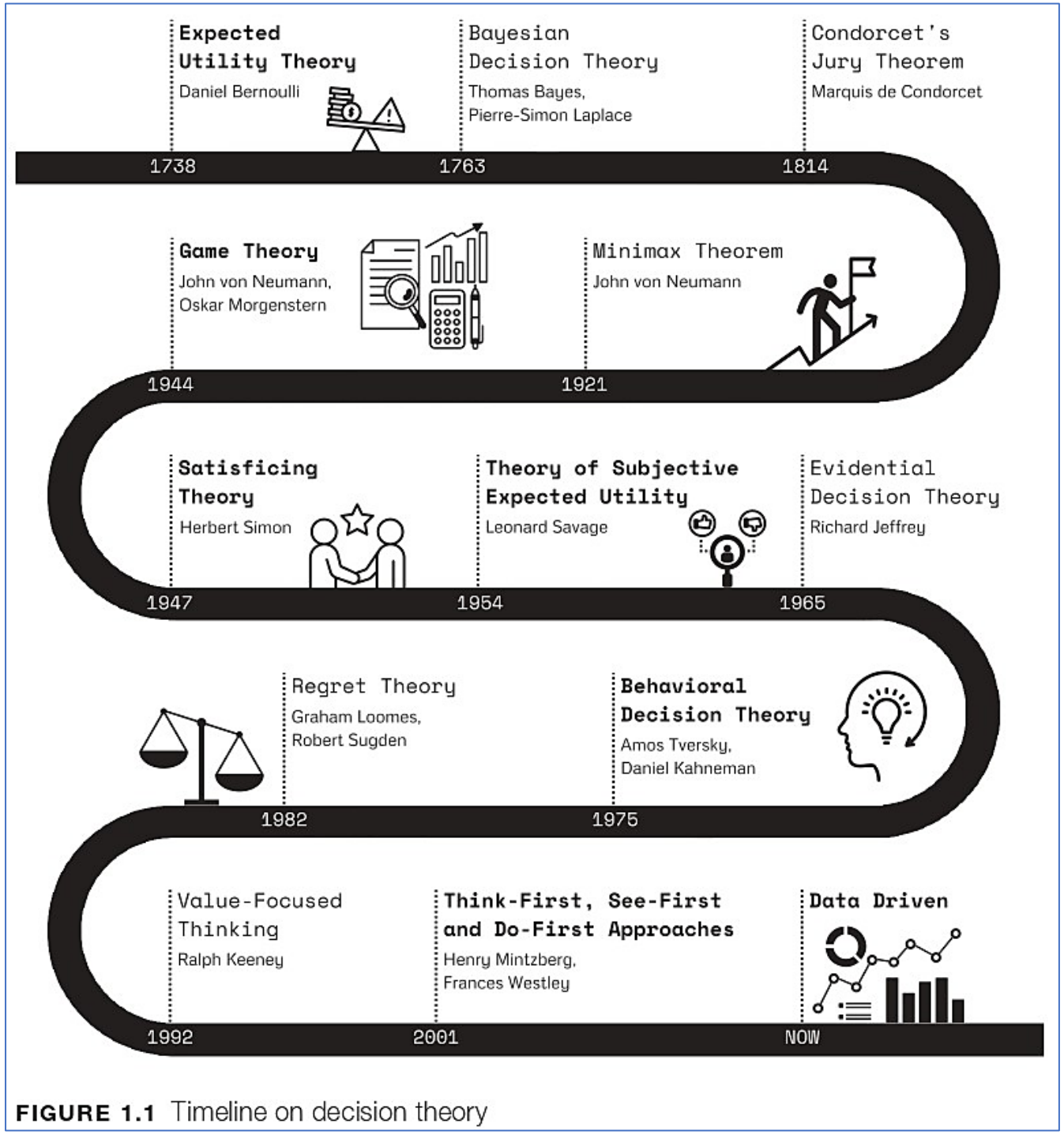


FIGURE 1.1 Timeline on decision theory

Sejarah DDDM: *Classical Decision Theories*

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.

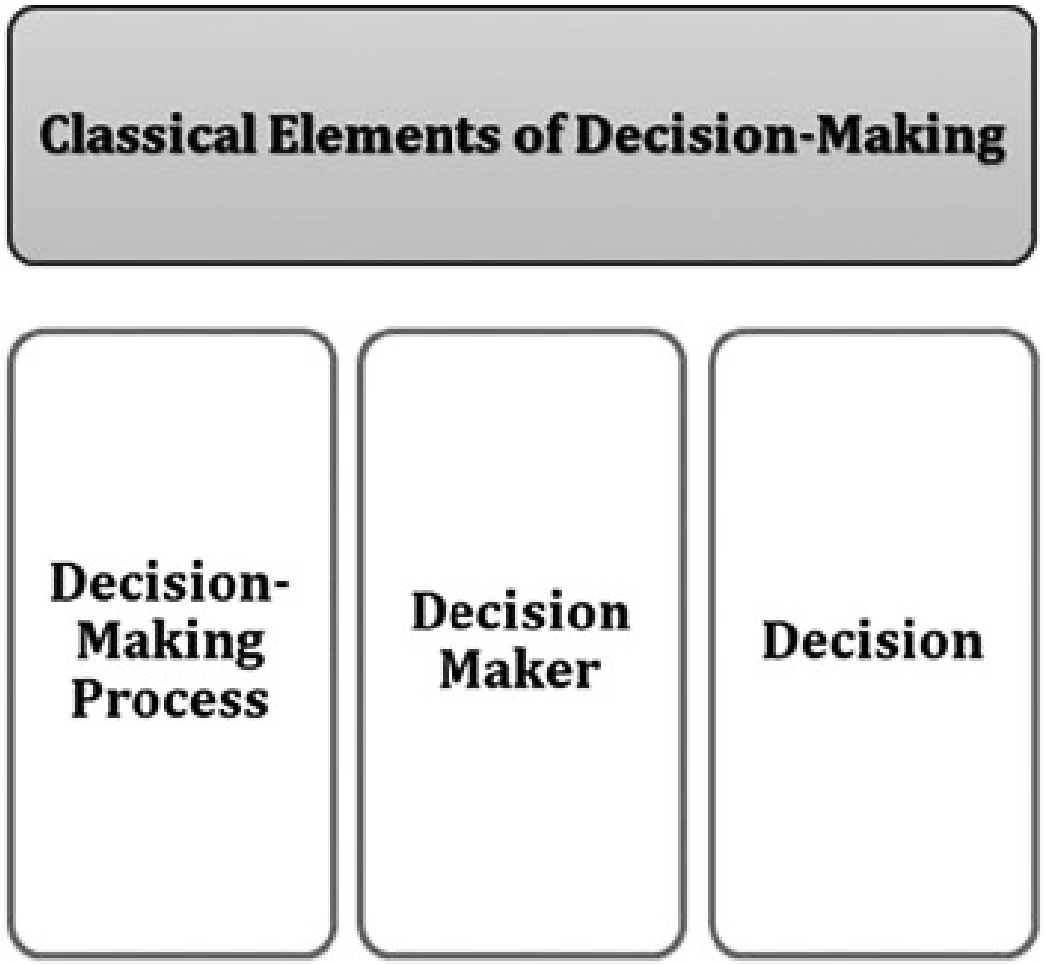
(Continued)

Sejarah DDDM:
Classical Decision Theories

TABLE 1.1 (Continued)

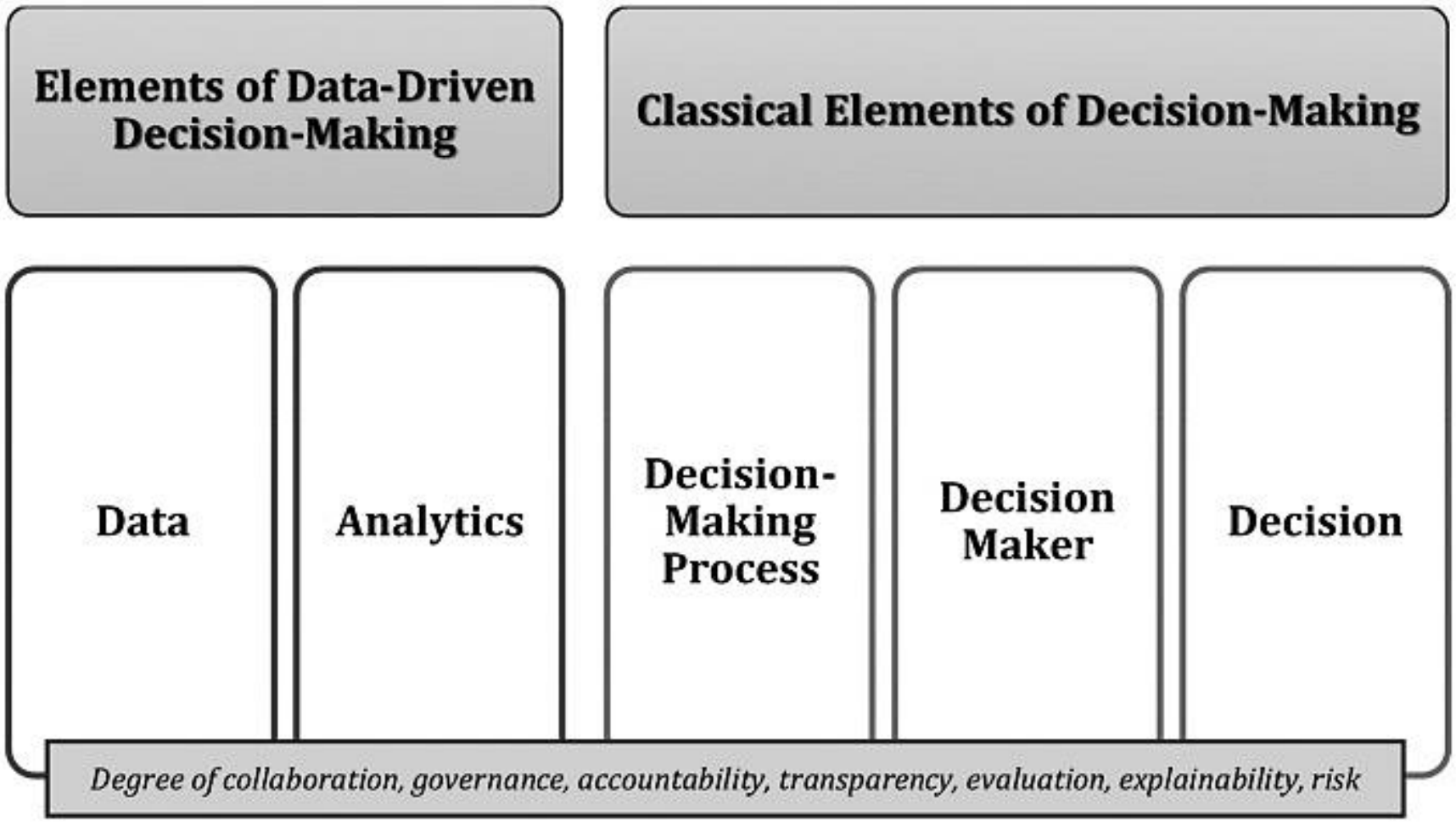
Year	Theory	Author(s)	Category	Description
1975	**Behavioral decision theory**	Amos Tversky, Daniel Kahneman	Decision-maker	A theory that incorporates psychological insights into decision-making, such as heuristics, biases, framing effects, and prospect theory.
1982	Regret theory	Graham Loomes, Robert 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.

Sejarah DDDM: *Classical Decision Theories*



The elements of classical decision making

Sejarah DDDM: *A Modern Decision Theory-DECAS*



The elements of data-driven decision making

Sejarah DDDM: *A Modern Decision Theory-DECAS*

- **DECAS**, atau teori yang meliputi: the **D**ecision-making process, d**E**cision maker, de**C**ision, d**A**ta, and analytic**S**
- **DECAS** (**modern data-driven decision theory**), perluasan dari teori keputusan klasik dengan mengusulkan **tiga klaim utama**:

1) **(big) data** dan **analitik (mesin)** harus dianggap sebagai elemen terpisah;
2) **kolaborasi** antara pengambil keputusan (manusia) dan analitik (mesin) dapat menghasilkan rasionalitas kolaboratif, melampaui rasionalitas terbatas yang didefinisikan secara klasik; dan
3) **Integrasi** yang bermakna dari elemen pengambilan keputusan klasik dengan data dan analitik dapat menghasilkan keputusan yang lebih terinformasi, dan mungkin lebih baik.

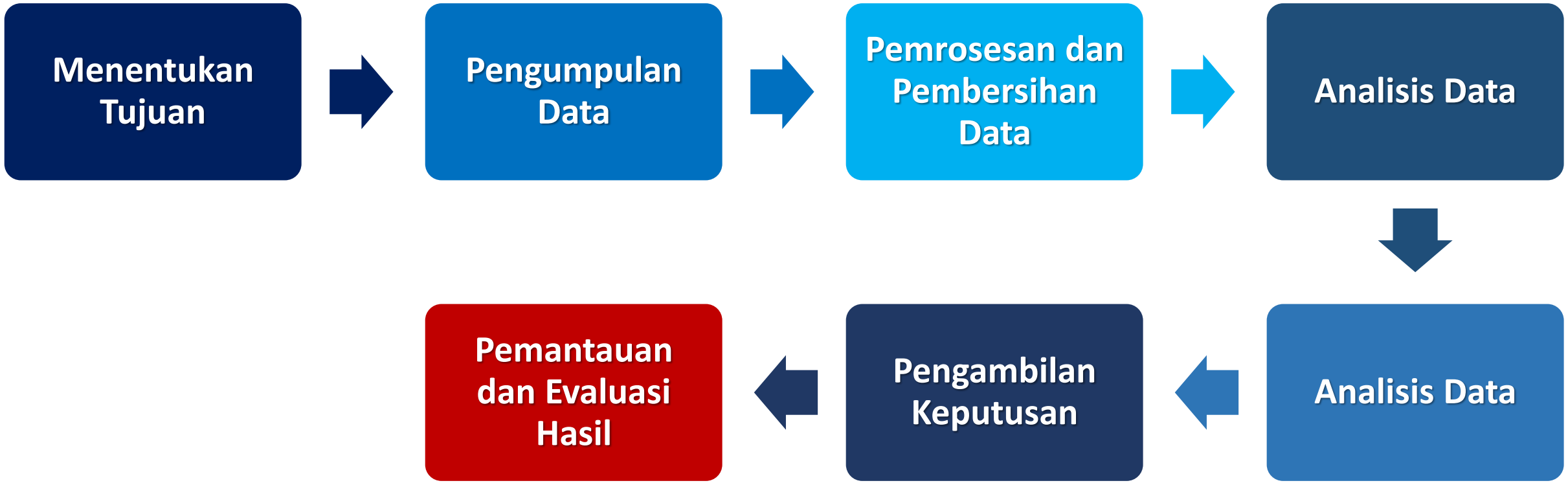
Manfaat Data-Driven Decision Making

Mengimplementasikan **Data-Driven Decision Making** membawa sejumlah manfaat signifikan bagi bisnis, antara lain :

Peningkatan efisiensi operasional	• Dengan menggunakan data yang akurat dan relevan, organisasi dapat mengurangi biaya operasional, mengoptimalkan penggunaan sumber daya, dan memperbaiki proses internal.
Inovasi	• Dengan data, kita dapat mengungkap tren baru dan peluang yang dapat mendorong inovasi produk dan layanan.
Continuous Improvement	• Evaluasi kinerja secara berkala dan menerapkan perbaikan secara ber-tahap.
Collaborative Decisions	• Semua anggota tim dapat berkontribusi berdasarkan data, bukan hanya pendapat pribadi.
Planning	• Analisis data dapat membantu memprediksi tren pasar dan perilaku konsumen. Dan juga perusahaan dapat menyusun rencana strategis yang lebih baik.

Proses DDDM

Untuk memanfaatkan **DDDM**, perusahaan dan organisasi perlu mengikuti proses yang terstruktur seperti berikut :



Proses DDDM

Untuk memanfaatkan **DDDM**, perusahaan dan organi-sasi perlu mengikuti proses yang terstruktur seperti berikut :

1. **Menentukan Tujuan:**

   Langkah pertama adalah mengidentifikasi dengan jelas masalah yang ingin dipecahkan atau tujuan yang ingin dicapai. Tanpa pemahaman yang jelas, proses selanjutnya mungkin tidak efektif.

2. **Pengumpulan Data:**

   Kumpulkan data yang relevan dari berbagai sumber, baik internal maupun eksternal. Data dapat berupa data transaksi, data pelanggan, data pasar, dan lain sebagainya.

3. **Pemrosesan dan Pembersihan Data:**

   Data yang dikumpulkan perlu diproses dan dibersihkan untuk memastikan akurasinya. Dalam hal ini melibatkan penghapusan kesalahan, inkonsistensi, dan data yang duplikat.

Proses Data-Driven Decision Making

4. **Analisis Data**:
   Gunakan berbagai teknik analisis data, seperti statistik deskriptif, analisis regresi, dan machine learning, untuk mengidentifikasi pola dan tren dalam data.
5. **Visualisasi Data**:
   Visualisasikan data dalam bentuk grafik, chart, atau dashboard untuk memudahkan pemahaman dan komunikasi.
6. **Pengambilan Keputusan**:
   Gunakan ***insight*** yang diperoleh dari analisis data untuk membuat keputusan bisnis yang strategis.
7. **Pemantauan dan Evaluasi Hasil**:
   Setelah keputusan diimplementasikan, penting untuk memantau hasilnya dan mengevaluasi hasil keputusan tersebut. Feedback ini dapat digunakan untuk melakukan perbaikan jika diperlukan.

Sumber Data untuk Data-Driven Decision Making

Data Internal

- Data internal merupakan data yang dihasilkan dari aktivitas bisnis sehari-hari dan tersimpan dalam sistem perusahaan.
- Contohnya seperti: Data Transaksi, Data Pelanggan, Data Operasional, Data Keuangan dan masih banyak lagi.

Data Eksternal

- Data eksternal merupakan data yang diperoleh dari sumber di luar perusahaan.
- Contoh data eksternal antara lain: laporan market research, publikasi industri, dan analisis kompetitor.

Data Terbuka

- Data yang tersedia secara publik dan gratis dapat dimanfaatkan untuk analisis,
- Contoh seperti : Data Social Media, Data Pemerintah, Data Organisasi Internasional, Dataset penelitian dan masih banyak lagi.

Tools dan Teknologi Pendukung

Untuk memanfaatkan sumber data secara maksimal, perusahaan atau organisasi dapat menggunakan berbagai tools dan teknologi berikut ini :

Business Intelligence (BI) Tools

- Software untuk visualisasi data dan dash-board interaktif.

Big Data Technologies

- Platform seperti Hadoop dan Spark untuk mengolah data dalam skala besar.

Machine Learning dan AI

- Algoritma yang dapat memprediksi tren dan pola dari data yang kompleks.

Cloud Computing

- Infrastruktur yang fleksibel dan skalabel untuk penyimpanan dan pemrosesan data.

BAGIAN 2

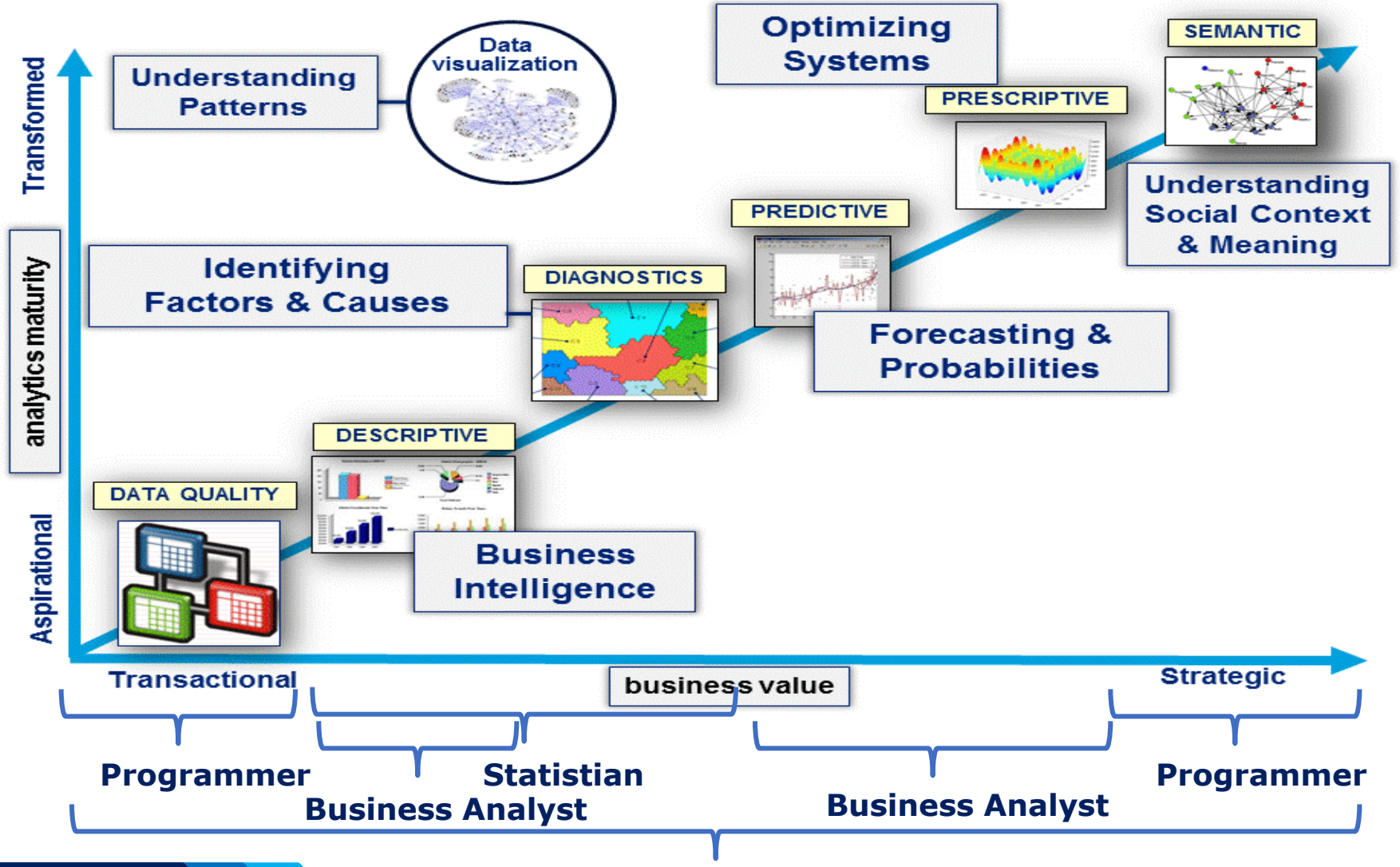
GAMBARAN UMUM SAINS DATA

Definisi Data Science

Definisi ***Data Science*** dari **NIST** (2018).

Data science is the extraction of useful knowledge directly from data through a process of discovery, or of hypothesis formulation and hypothesis testing.

Apa itu Sains Data ?



Tujuan Tugas yang Biasa Digunakan

01 **Descriptive:**
Menjelaskan keadaan bisnis saat ini melalui data historis.

02 **Diagnostic:**
Menjelaskan mengapa suatu masalah terjadi dengan melihat data historis.

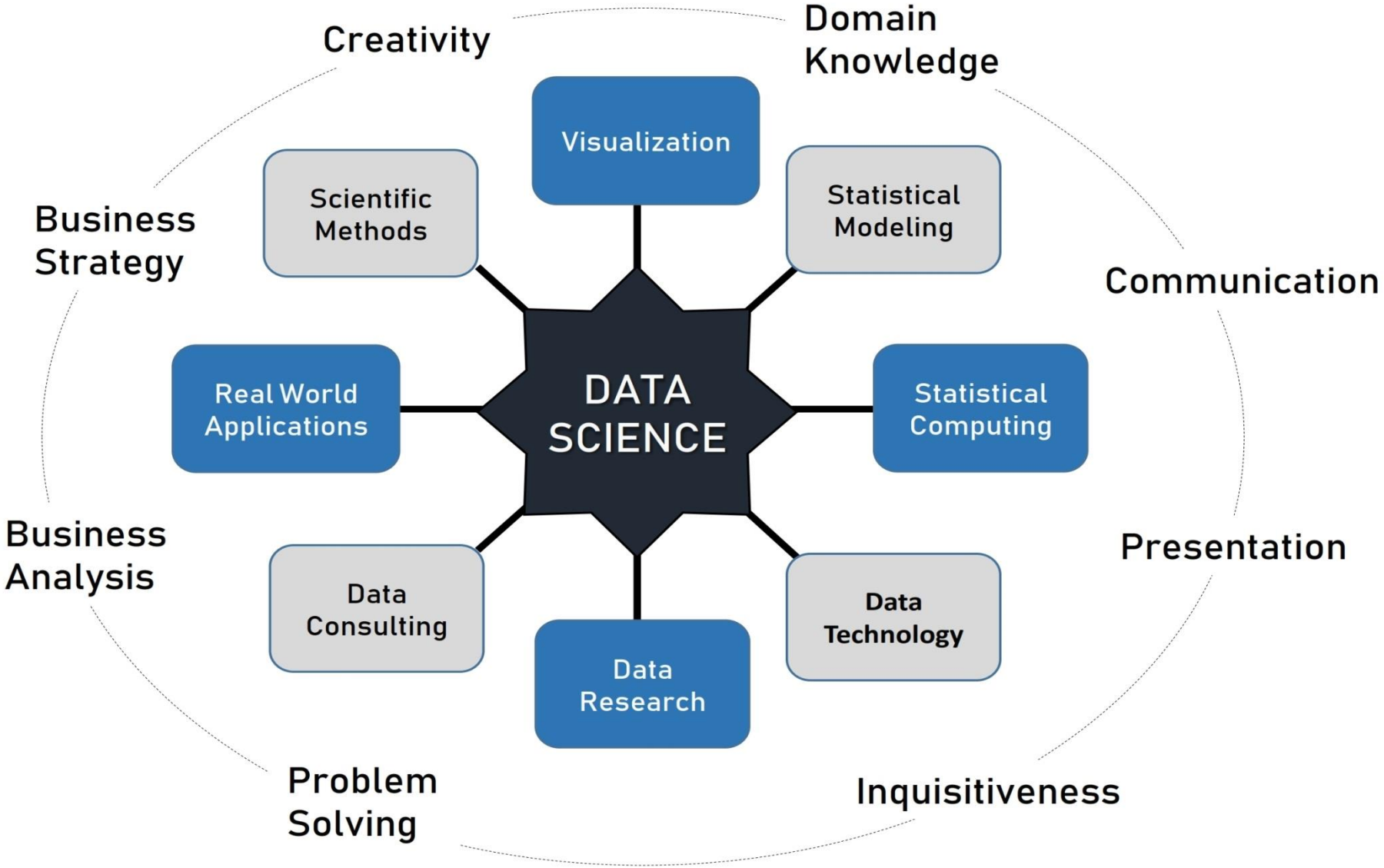
03 **Predictive:**
Memproyeksikan atau memprediksi hasil masa depan berdasarkan data historis.

04 **Prescriptive:**
Menggunakan hasil analitik prediktif dan pengetahuan lain dengan menyarankan upaya terbaik di masa depan.

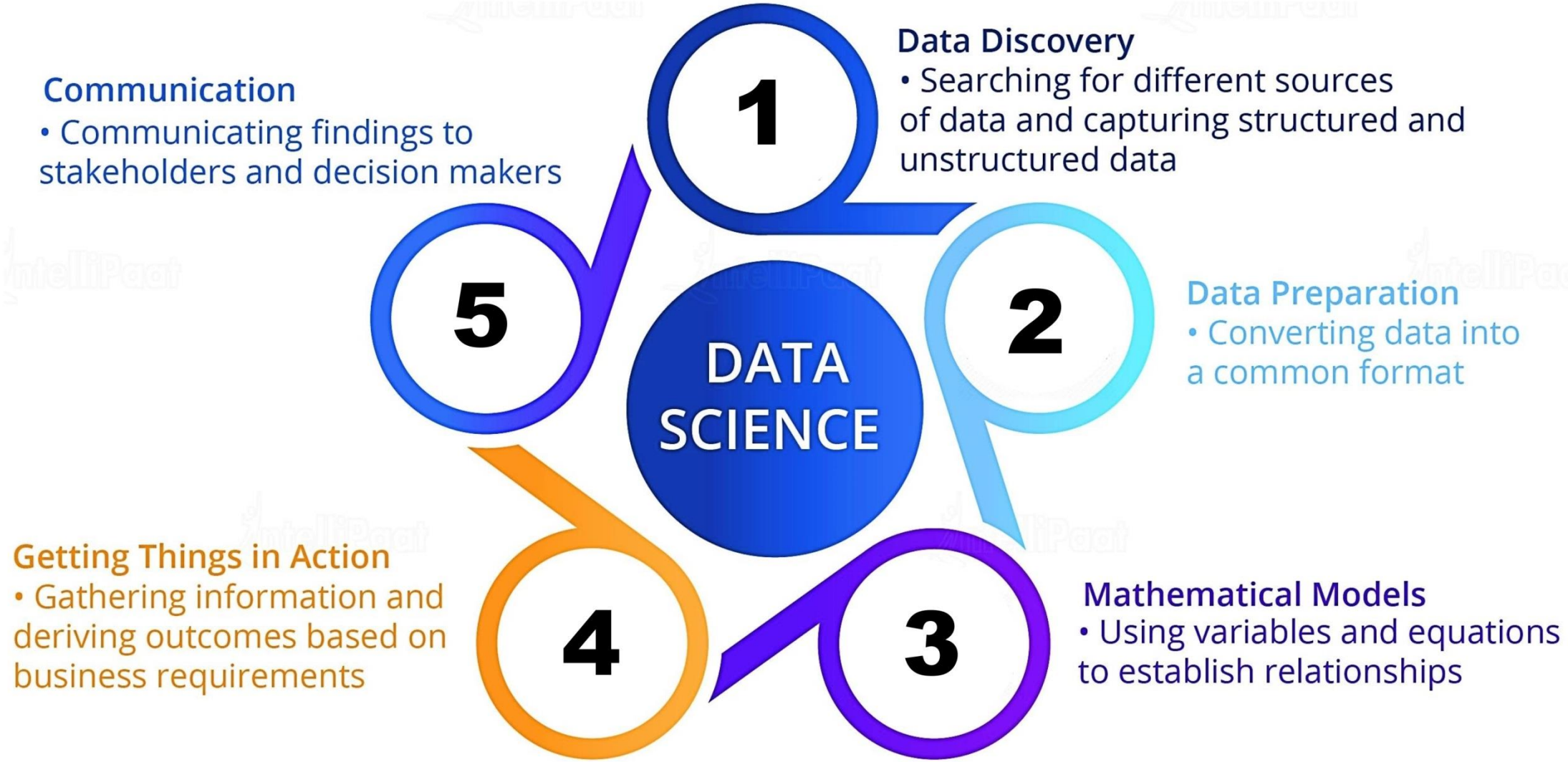
Jenis Tugas yang Dikembangkan

- Regression / Estimation
- Classification
- Clustering
- Association
- Anomaly Detection
- Sequence Mining
- Recommendation Systems

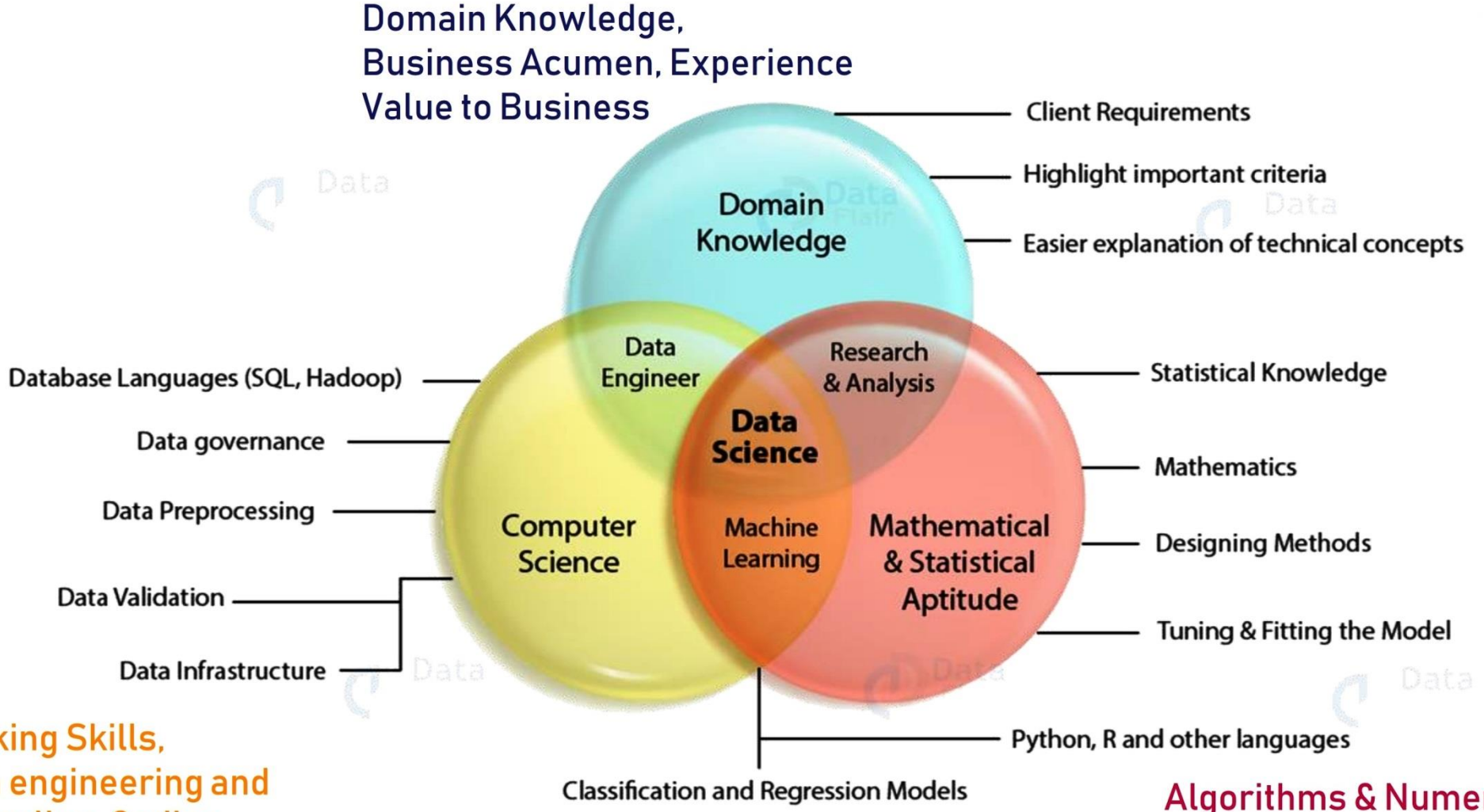
Sains Data: Multi-Disiplin



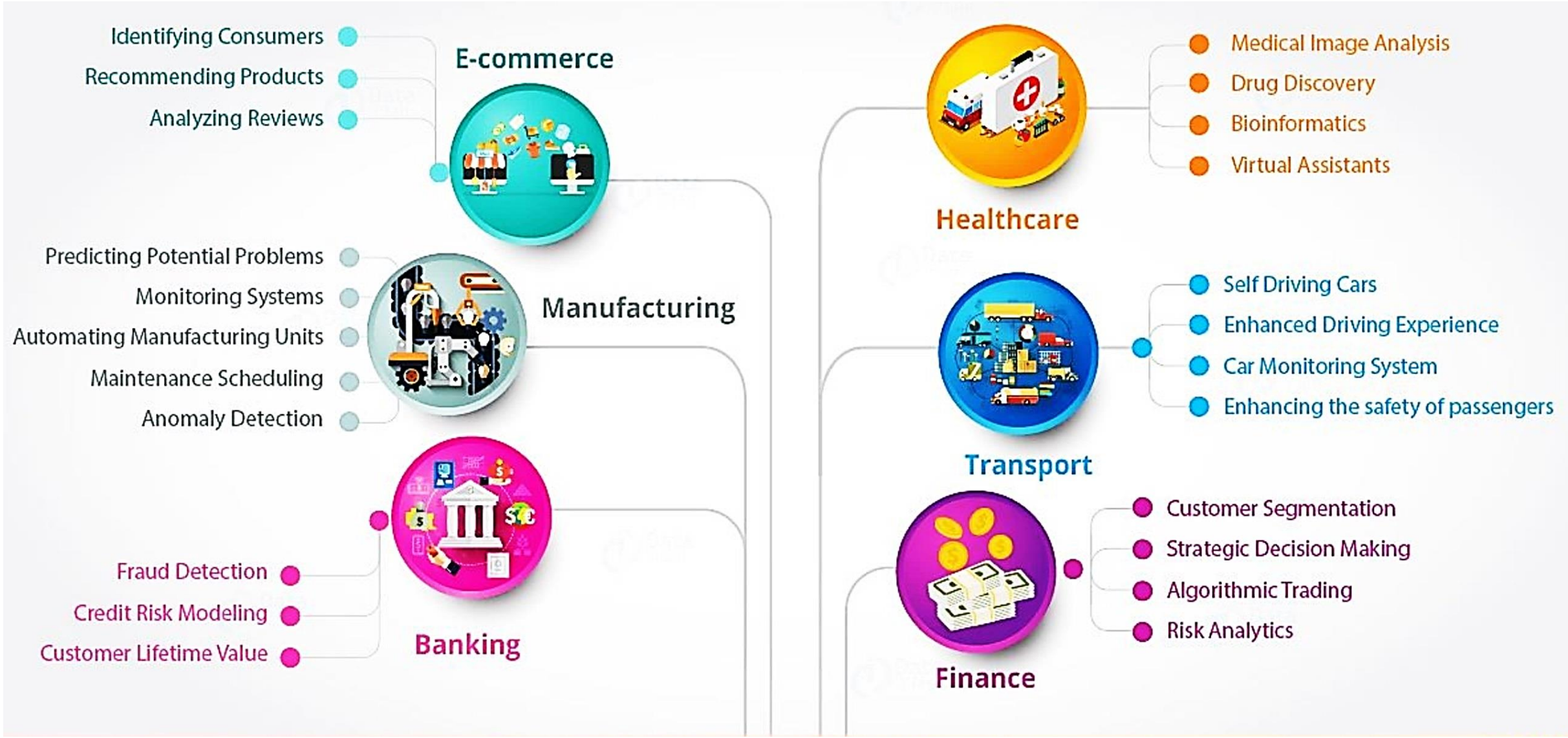
Siklus Hidup-nya



Komponen-Komponen-nya



Penerapan Utama Sains Data



Penerapan Sains Data

IN FINANCE	IN BANKING	IN HEALTH CARE
Risk Analytics	Risk Modeling	Secured & Shared Medical Records
Real-Time Analytics	Fraud Detection	Optimized & Integrated Care
Consumer Analytics	Customer Lifetime Value	Improved Monitoring
Consumer Data Management	Customer Segmentation	Automated Lab-work and Standardized Image Recognition
Personalized services	Recommendation engines	Drug interaction detection
Fraud Detection	Real-time Predictive Analytics	Personalized medicine
Algorithmic Trading		Communication assistance

Proses Sains Data

GET DATA

EXPLORE DATA

MODEL DATA

VISUALIZE RESULTS

STIMULATE RESULTS

USE RESULTS IN BUSINESS

BUSINESS APPLICATION

Definisi Data Scientist dari NIST

Definitions by **NIST** Big Data WG (NIST SP1500 - 2015)

- **A Data Scientist** is
  - *a practitioner with sufficient knowledge in the overlapping regimes of expertise in **business needs, domain knowledge, analytical skills, and programming and systems engineering** expertise to manage the end-to-end scientific method process through each stage in the **big data lifecycle**.*
- **Data science** *is the empirical synthesis of actionable knowledge and technologies required to handle data from raw data through the complete data lifecycle process.*

Data Scientist

DATA SCIENTIST

MATHS & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning : decision trees, random forests, logistic regression
- Unsupervised learning : clustering, dimensionality reduction
- Optimization : gradient descent and variants

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools eg. Flare, D3.js, Tabeau

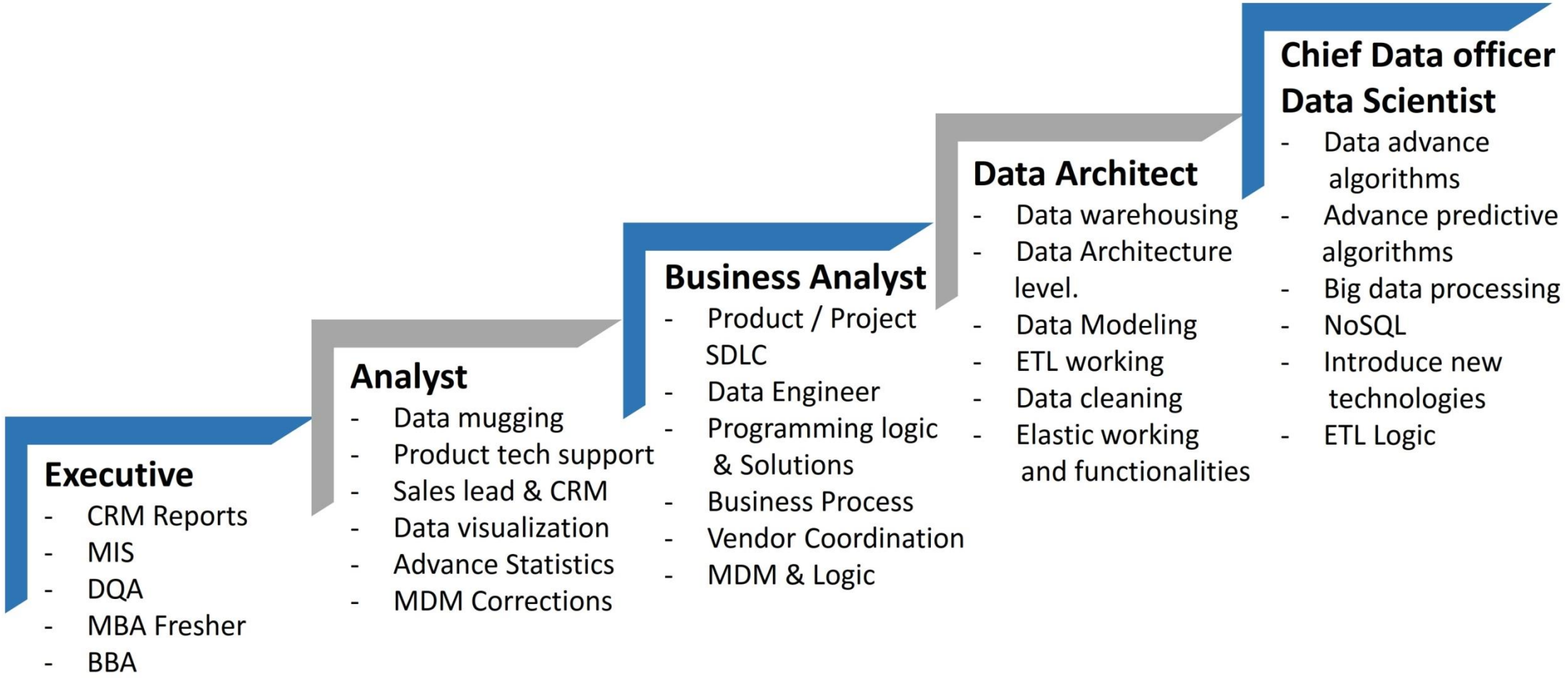
DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about the data
- Influence without authority
- Hacker mindset
- Problem Solver
- Strategic, proactive, creative, innovative and collaborative

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting Language e.g. Python
- Statistical computing package e.g. R
- Database SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom Reducers and experience with xaaS like AWS

Jenjang Karir



Data Scientist vs Data Analyst

FEATURES	DATA SCIENTIST	DATA ANALYST
Background	A Data scientist deals with various Data operations	A Data Analyst's role is related to data Cleaning, transforming & generating Inferences from data
Scope	Involved with several underlying Data procedures	Involvement is limited to small data And static inferences
Types of Data	Handles structured & unstructured data	Deals with structured data only
Skills	Possess knowledge of mathematics Statistics & machine learning algorithm	Has problem solving skills, Knowledge of basic statistics
Tools	Proficient in SAS, Python, R, TensorFlow, Hadoop, Spark	Knows Excel, SQL, R (in some cases), Tableau

https://data-flair.training/blogs/data-scientist-vs-data-analyst/

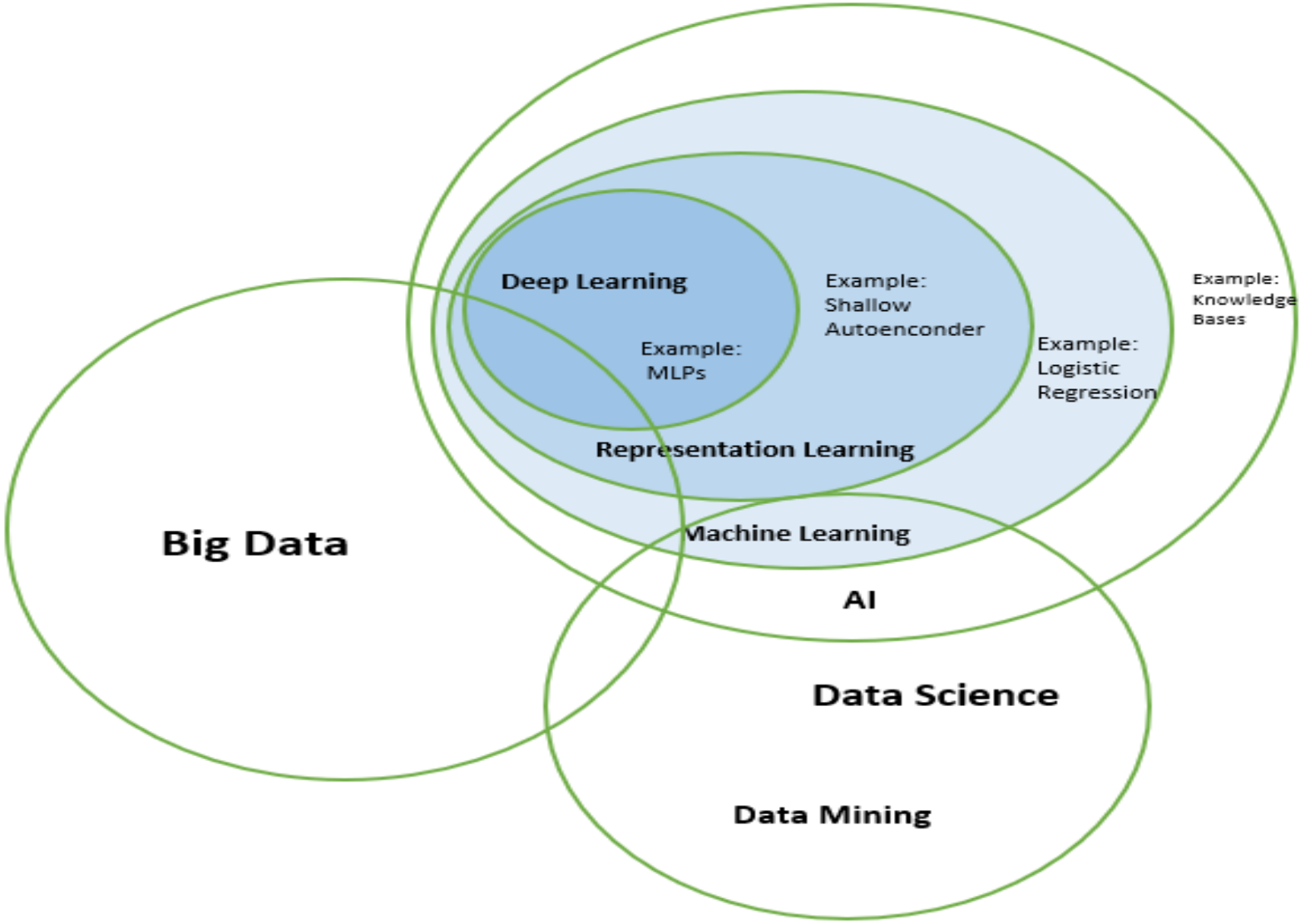
Data Scientist vs Statistician

	DATA SCIENTIST	STATISTICIAN
Meaning	Analyze and interpret complex digital data in order to assist decision-making in businesses.	An expert in the preparation and analysis of statistical data.
Task	Interpret and manage data & solve complex problems using expertise in a variety of data niches.	Analyze data with specific statistical software and developing theory of statistics.
Connected with	Mathematical Sciences, Computer Science, Modelling Statistics and analytics	Statistical data in various field like Economics, Biology, Public Health, Sports, Medicines, and etc.
Responsibilities	To provide new insights into the business and create solutions that enable enhanced business performance.	Develop and initiate innovative statistical techniques, issues and protocols.
Designation	Data Scientist, Data Engineer, Data Analyst, Data Architect, etc.	Statistician, Data Analyst, etc.

BAGIAN 3

TEKNIK MACHINE LEARNING DALAM DECISION MAKING

Hubungan DS-BD-AI-ML-DL Dewasa ini



Source: adaptation from Ian Goodfellow, et.al 2016 & and Matthew Mayo, 2016

Machine Learning

- A machine learning-based system is not explicitly programmed but learned from data.
- A machine learning algorithm infers patterns and relationships between different variables in a dataset and then uses that knowledge to generalize beyond the training dataset.

Terms used in the context of machine learning

1. Data mining
2. Features
3. Labels
4. Models
5. Accuracy, and Precision

Data Mining

- Definition: "**The application of specific algorithms for extracting patterns from data**."
- **data mining** is a **process** in which machine learning algorithms are utilized to extract potentially valuable patterns held within datasets.
- Data mining is exploratory analysis with unsupervised machine learning algorithms

The main application is text mining in a big data environment:

1. Text parsing
2. Sentiment analysis
3. Opinion mining
4. Natural language processing using a deep learning algorithm

Data Mining algorithms:

1. C4.5
2. k-means
3. Support vector machines
4. Apriori
5. EM
6. PageRank
7. AdaBoost
8. kNN
9. Naive Bayes
10. CART

Features

- a feature represents an independent variable
- In a tabular dataset, a row represents an observation and a column represents a feature.
- Features are also collectively referred to as dimensions.

Categorical Features

- It can take on one of a fixed number of discrete values with a name or a label.
- The values of a categorical feature have no ordering.
- gender is a categorical feature. It can take on only one of two values.

Numerical Features

- It can take on any numerical value
- numerical features have mathematical ordering
- discrete numerical feature: number of persons, number of rooms
- continuous numerical feature: temperature value

Labels

- label is a variable that a machine learning learns to predict
- **categorical label**: ex, the category of a news article is a categorical label
- **numerical label**: ex, price is a numerical label

Features					Label
Position	Experience	Skill	Country	City	Salary ($)
Developer	0	1	USA	New York	103100
Developer	1	1	USA	New York	104900
Developer	2	1	USA	New York	106800
Developer	3	1	USA	New York	108700
Developer	4	1	USA	New York	110400
Developer	5	1	USA	New York	112300
Developer	6	1	USA	New York	114200
Developer	7	1	USA	New York	116100
Developer	8	1	USA	New York	117800
Developer	9	1	USA	New York	119700
Developer	10	1	USA	New York	121600

Models

- A mathematical construct for capturing patterns within a dataset and estimates the relationship between the dependent and independent variables and has predictive capability. and can calculate or predict the value for the dependent variable when getting the values of the independent variables.
- Training a model is a compute-intensive task while using it is not as compute-intensive.
- A model is generally saved to disk so that it can be used without going to the training step again.

Training Data (80% of data)

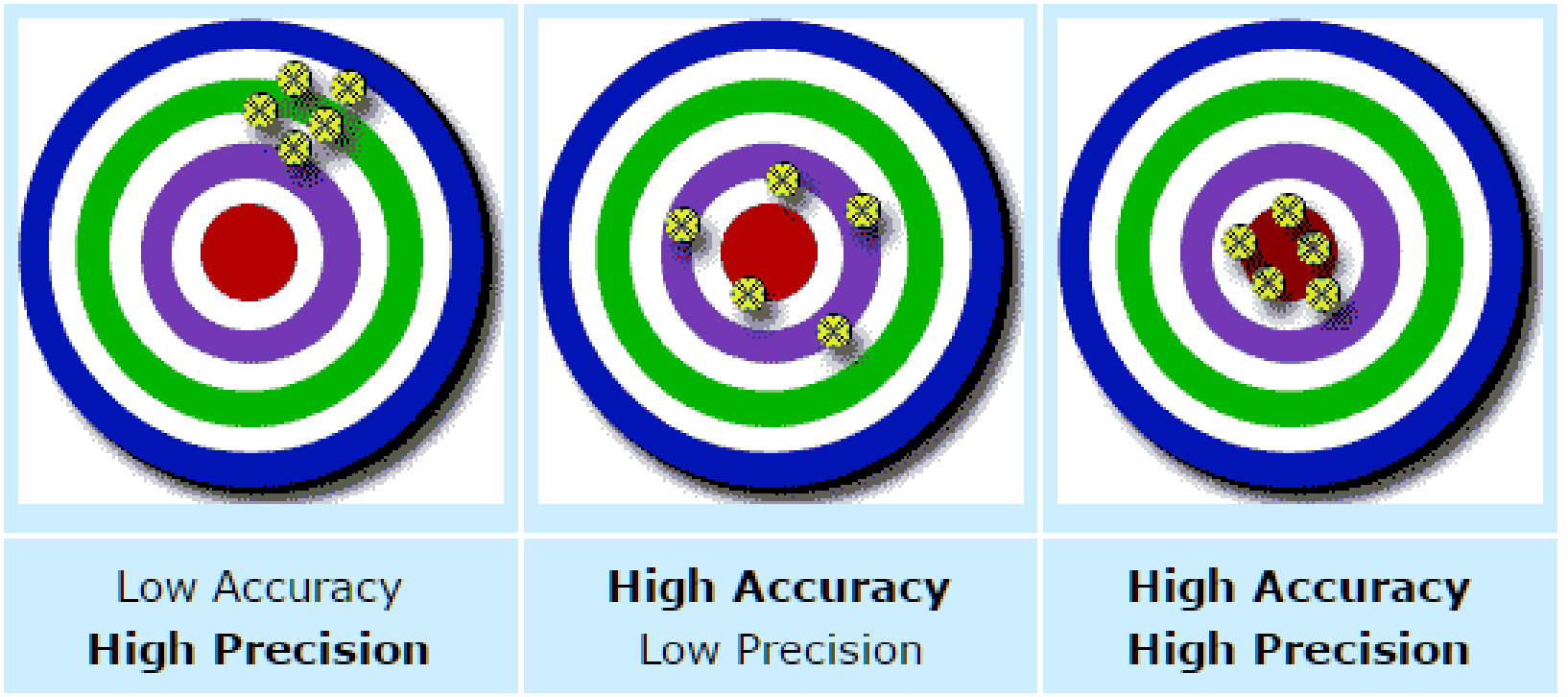
- The data used by a machine learning algorithm to train a model

Test Data (20% of data)

- The data used for evaluating the predictive performance of a model

Accuracy vs Precision

- **Accuracy** adalah seberapa dekat nilai terukur dengan nilai aktual (benar).
- **Precision** adalah seberapa dekat nilai terukur satu sama lain.

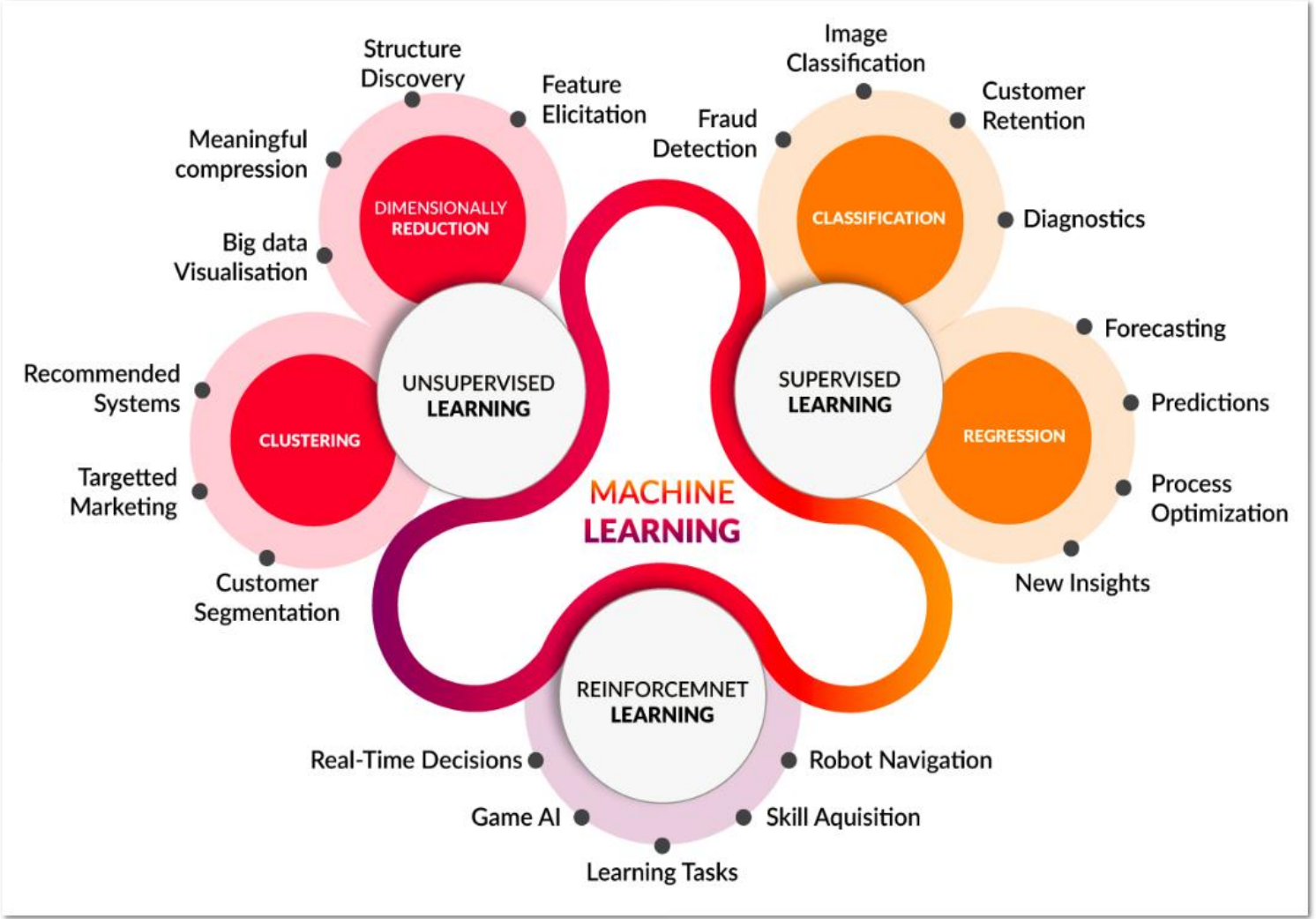


Low Accuracy **High Precision**	**High Accuracy** Low Precision	**High Accuracy** **High Precision**

Teknik-Teknik Machine Learning

Machine Learning terutama memiliki tiga jenis teknik pembelajaran:

- Supervised learning
- Unsupervised learning
- Reinforcement learning



Kategori tugas Machine Learning

1. Classification
2. Regression
3. Clustering
4. Anomaly detection
5. Association
6. Recommendation
7. Dimensionality reduction
8. Computer Vision
9. Text Analytics

Classification [supervised learning]

- Classification is concerned with building models that separate data into distinct classes. These models are built by inputting training data for which the classes are pre-labeled for the algorithm to learn from. The model is then used by inputting a different dataset for which the classes are withheld, allowing the model to predict their class membership based on what it has learned from the training set.

Binary classification examples (divide data to two options only)

- spam filtering is a classification task
- Tumor diagnosis can be treated as a classification problem.
- determining credit risk using personal information such as income, outstanding debt

Multi-class classification examples

- handwritten recognition of each character is a multi-class classification problem
- image recognition is a multi-class classification task
- Xbox Kinect360, which infers body parts and position

Regression [supervised learning]

- The goal is to predict a numerical label for an unlabeled observation

Regression algorithms: Linear regression, Decision trees

Examples

- home valuation
- Asset trading, and forecasting
- Sales or inventory forecasting

Features					Label
Position	Experience	Skill	Country	City	Salary ($)
Developer	0	1	USA	New York	103100
Developer	1	1	USA	New York	104900
Developer	2	1	USA	New York	106800
Developer	3	1	USA	New York	108700
Developer	4	1	USA	New York	110400
Developer	5	1	USA	New York	112300
Developer	6	1	USA	New York	114200
Developer	7	1	USA	New York	116100
Developer	8	1	USA	New York	117800
Developer	9	1	USA	New York	119700
Developer	10	1	USA	New York	121600

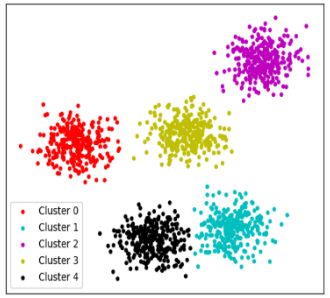
Anomaly Detection [supervised learning]

the goal is to find outliers, noise, deviations in a dataset

Anomaly detection applications

- In manufacturing, it is used for automatically finding defective products.
- In data centers, it is used for detecting bad systems.
- Websites use it for fraud detection.
- Detecting security attacks. Network traffic associated with a security attack is unlike normal network traffic. Similarly, hacker activity on a machine will be different from a normal user activity.

Clustering [unsupervised learning]

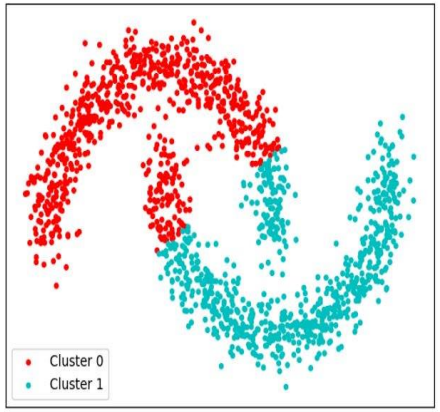


- The aim is to split a dataset into a specified number of clusters or segments.
- Elements in the same cluster are more similar to each other than to those in other clusters.

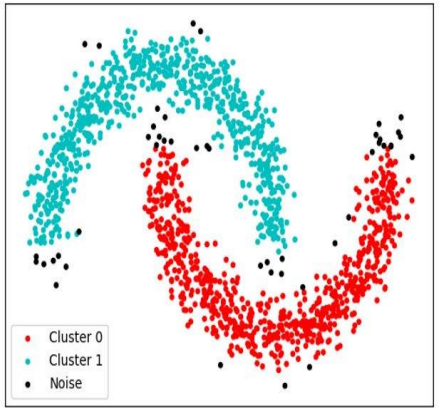
Clustering algorithms

1) k-means

- The number of clusters (k) must be given explicitly.
- Identify the best k cluster centers in an iterative manner
- Clusters are assumed to be spherical.



K-means



OPTICS

2) OPTICS /DBSCAN

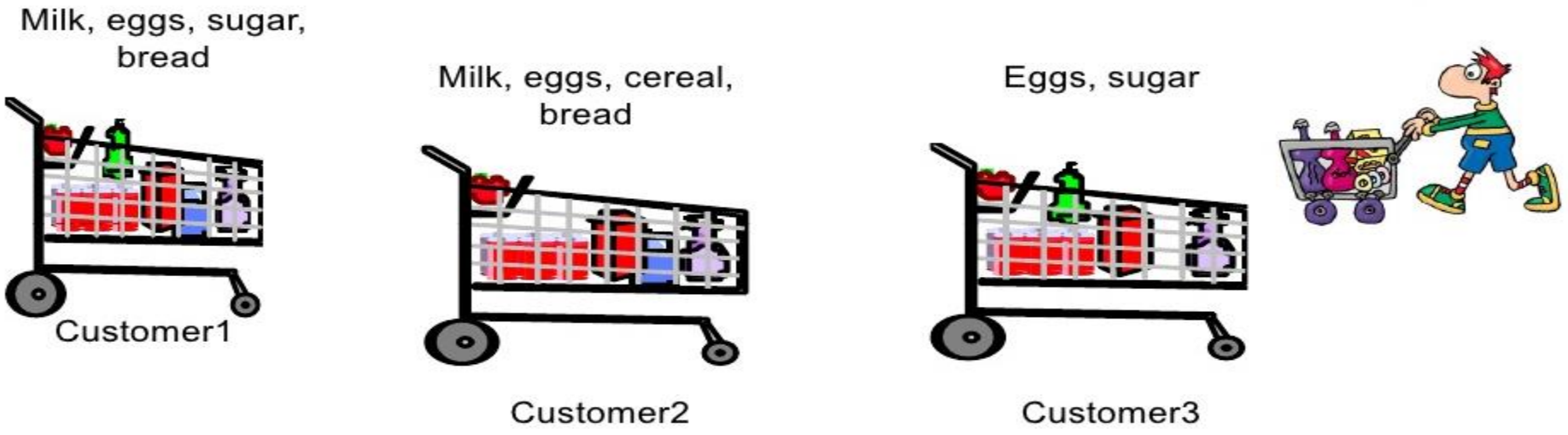
- it is a density-based clustering algorithm. represents clusters by its nature

Example

- creating customer segments, which can be targeted with different marketing programs

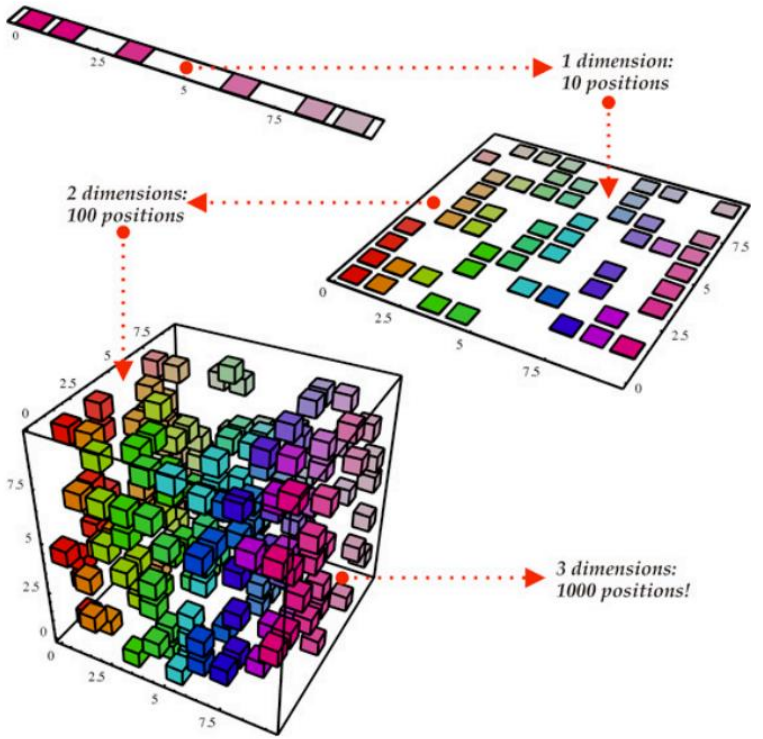
Association [unsupervised learning]

- Association is most easily explained by introducing market basket analysis, a typical task for which it is well-known.
- attempts to identify associations between the various items that have been chosen by a particular shopper and placed in their market basket and assigns support and confidence measures for comparison.
- The value of this lies in cross-marketing and customer behavior analysis.



Dimensionality Reduction [unsupervised learning]

- The goal in dimensionality reduction is to reduce the number of features in a dataset without significantly impacting the predictive performance of a model and this will reduce the computational complexity and cost of machine learning.



Recommendation [Reinforcement]

- The goal of a recommendation system is to recommend a product to a user based on past behavior to determine user preferences.
- Unlike Association, Recommendation focus on user behavior and suggest according to this user behavior.
- It is reinforcement because we not sure of result until user choose one of our recommendation, if not, then our recommendation was not correct.

Supervised machine learning algorithms

Classification

- **Two Class Classification**
  - Logistic Regression (Fast)
  - Decision Tree (Fast)
  - Decision jungle(Accurate)
  - SVM (Accurate) (>100 features)
  - Boosted Decision Tree (Fast - Large memory)
  - Bayes point machine (Fast)
- **Multi Class Classification**
  - Decision Tree (Fast)
  - Logistic Regression (Fast)
  - Random Forest (Accurate)
  - Gradient Boosting Tree (Accurate)
  - Naive Bayes (Big Data)
  - Decision jungle(Accurate)

Regression

- Linear Regression (Fast)
- Decision Tree (Fast)
- Random Forest (Accurate)
- Gradient Boosting Tree (Accurate)
- Ordinal regression
- Bayesian linear regression
- Boosted Decision Tree (Fast - Large memory)
- SGD Regressor (<100K rows)
- Lasso/ ElasticNet (Few Features)
- RidgeRegression
- SVR(kernel=linear/ rbf)
- EnsembleRegressors

Anomaly Detection

- One Class SVM (support vector machine)
- PCA based Anomaly Detection
- Time Series Anomaly Detection

Unsupervised machine learning algorithms

Clustering

- K-means Clustering
- K-modes (Categorical variables)
- DBScan (predict the number of clusters automatically)
- OPTICS (predict the number of clusters automatically)

Association

Apriori

Dimension Reduction

- PCA
- Singular value decomposition

Recommendation

- Matchbox Recommender

Computer Vision

OpenCV Library

Text Analytics (Supervised Learning)

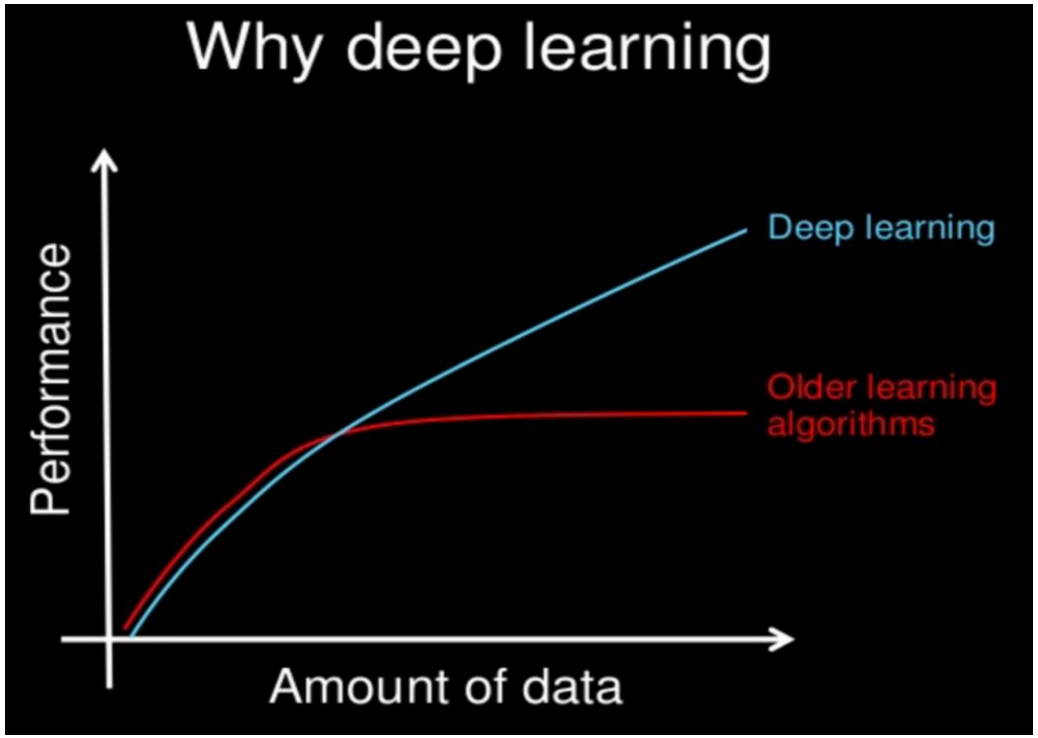
- Named Entity Recognition
- Sentimental Analysis

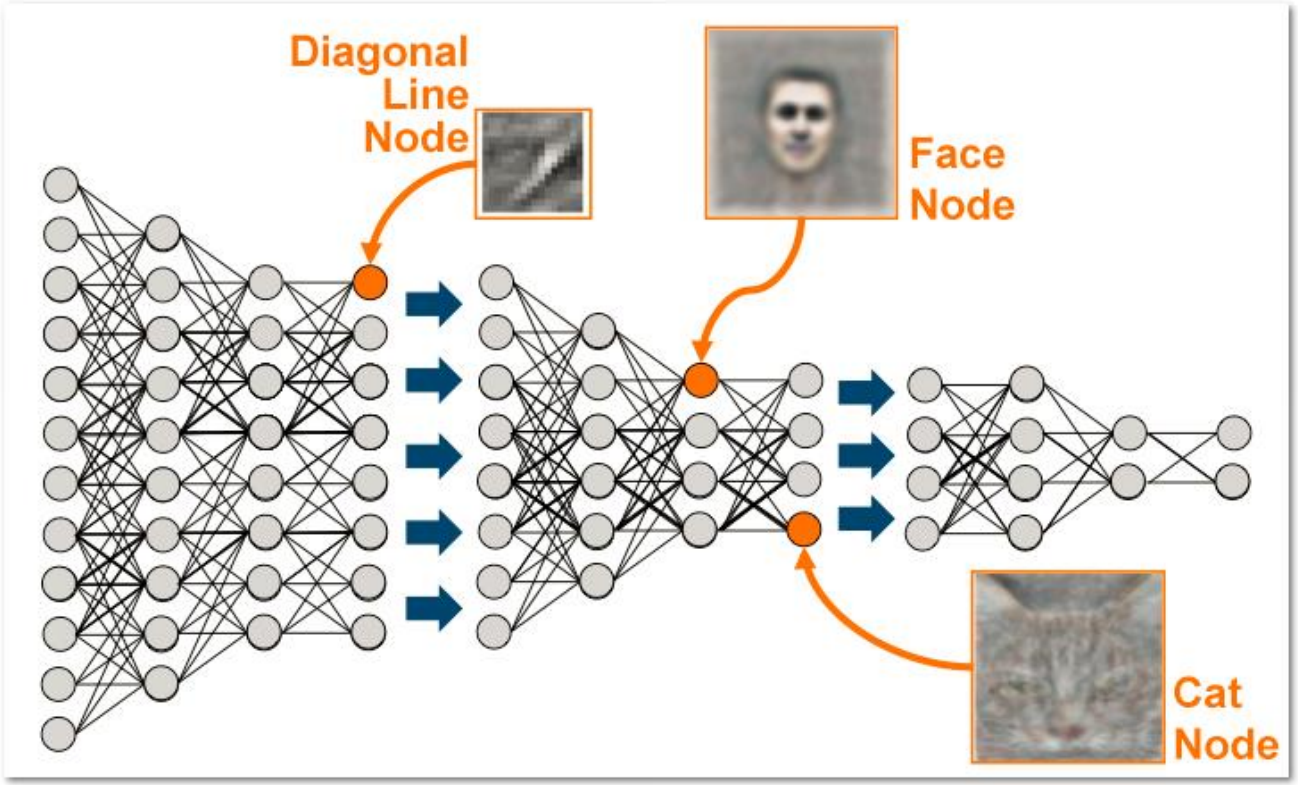
Neural Network

- The feedforward neural network algorithm uses a technique known as backpropagation to train a model.
  - During the training phase, prediction errors are fed back to the network.
  - The algorithm uses this information to adjust the weights of the edges connecting the nodes to minimize prediction errors.
  - This process is repeated until the prediction errors converge to a value less than a predefined threshold.
- Generally, a neural network with one layer is sufficient in most cases. If more than one hidden layers are used, it is recommended to have the same number of nodes in each hidden layer.
- Neural networks are better suited for classifying data that is not linearly separable.

Deep learning

- It consists of many layers/nodes of neural networks.
- Deep learning automatically finds out the features that are important for classification, but in Machine Learning we had to manually give the features
- Deep learning algorithms need a large amount of data to work perfectly.



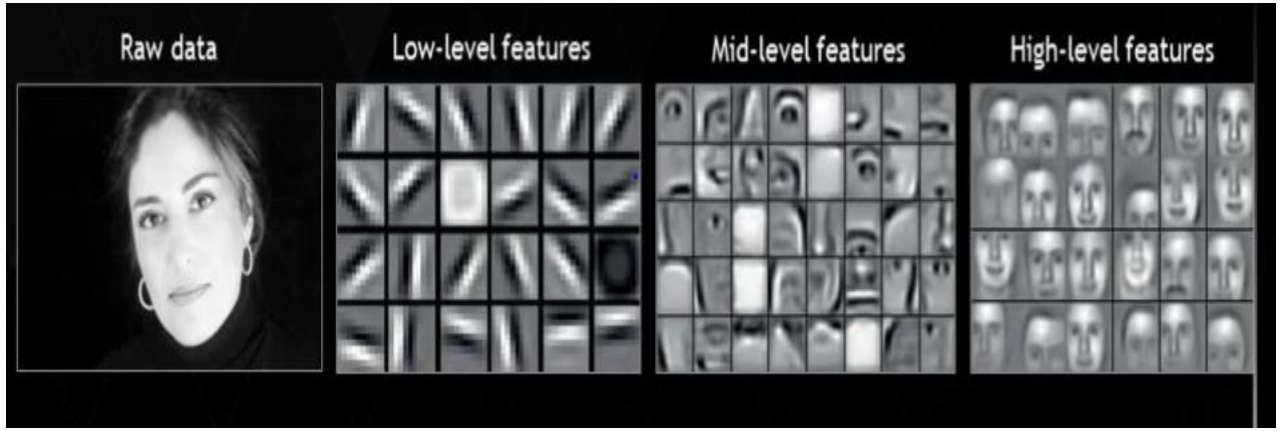


Traditional ML Algorithms

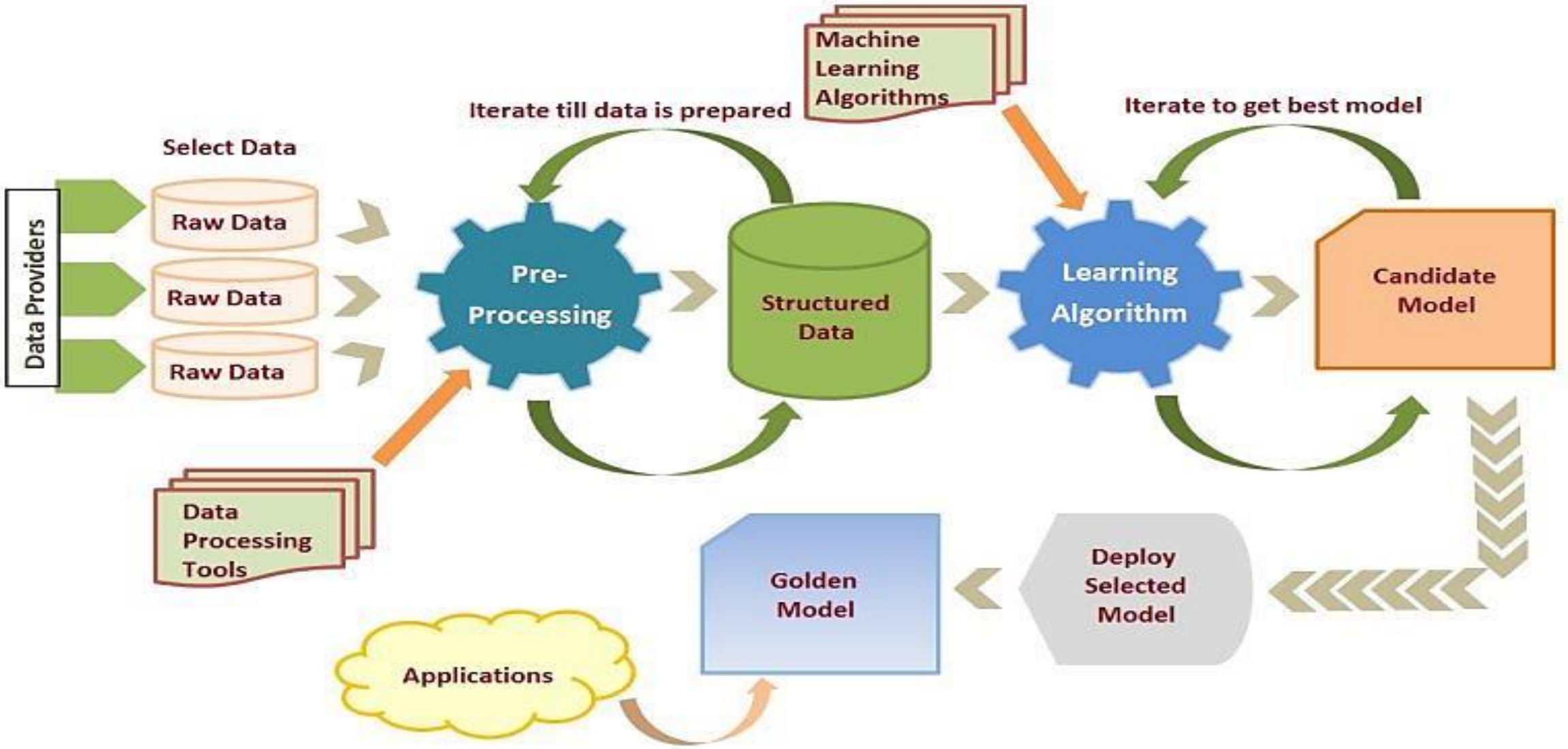
- can work on low-end machines
- break the problem down into different parts, solve them individually, and combine them to get the result.
- Takes much less time to train, ranging from a few seconds to a few hours.
- Gives us the selection rules, so it is easy to interpret, and safe to use in industry for interpretability.

Deep learning algorithms

- need a large amount of data to work perfectly
- heavily depend on GPUs
- solve the problem end-to-end.
- takes a long time to train. (like two weeks)
- It is excellent and is near human performance, but no guarantee to be always like this!, because most of the operations are hidden and we can't review selection logic!



Proses Machine Learning



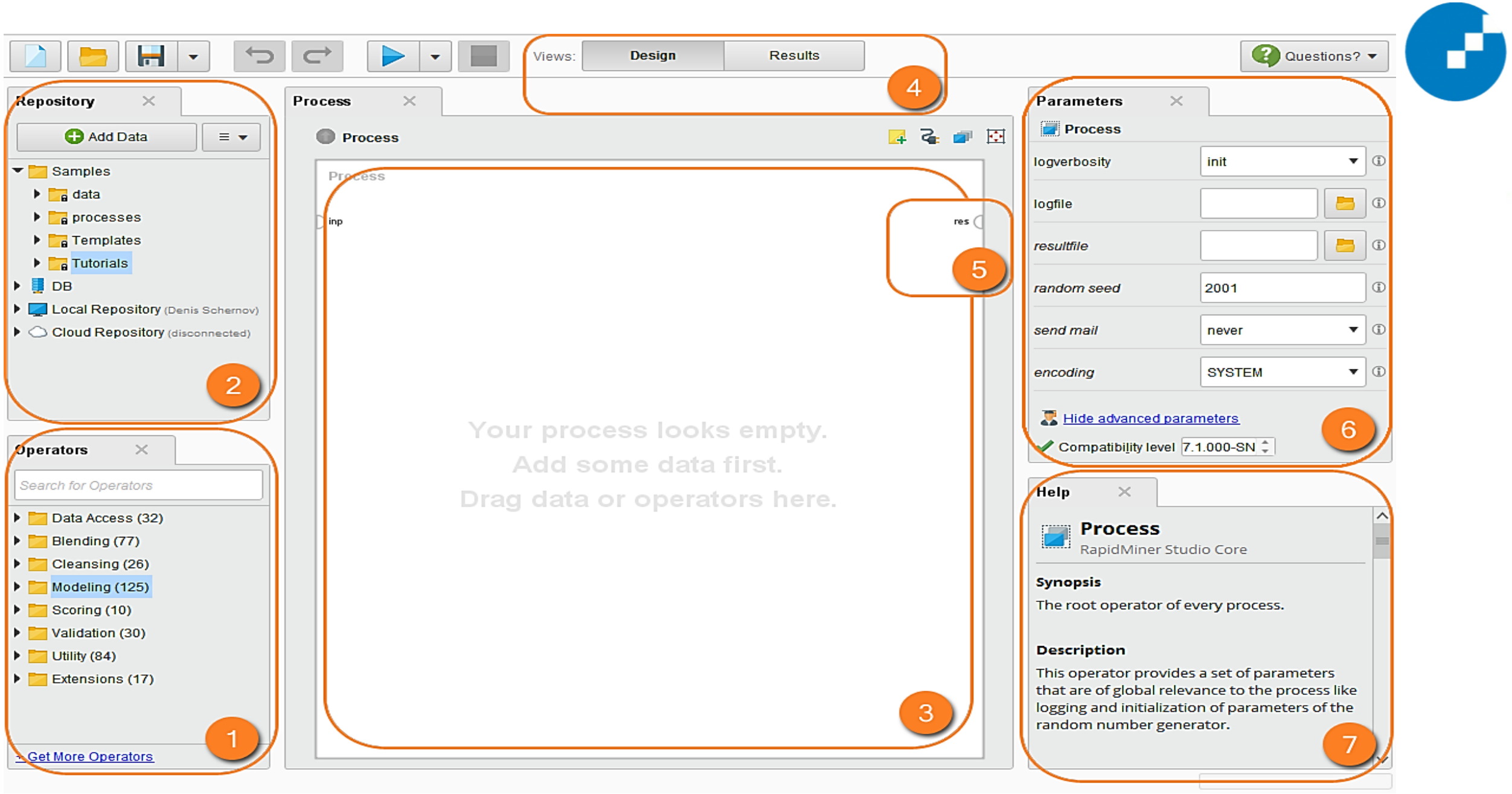
Tool Implementasi: Matlab

- Matlab https://www.mathworks.com/products/m
  - Komersial versi terakhir R2020a
  - Tersedia Toolbox: **AI, Data Science, and Statistics**
    - Statistics and Machine Learning Toolbox
    - Deep Learning Toolbox
    - Reinforcement Learning Toolbox
    - Text Analytics Toolbox
    - Predictive Maintenance Toolbox

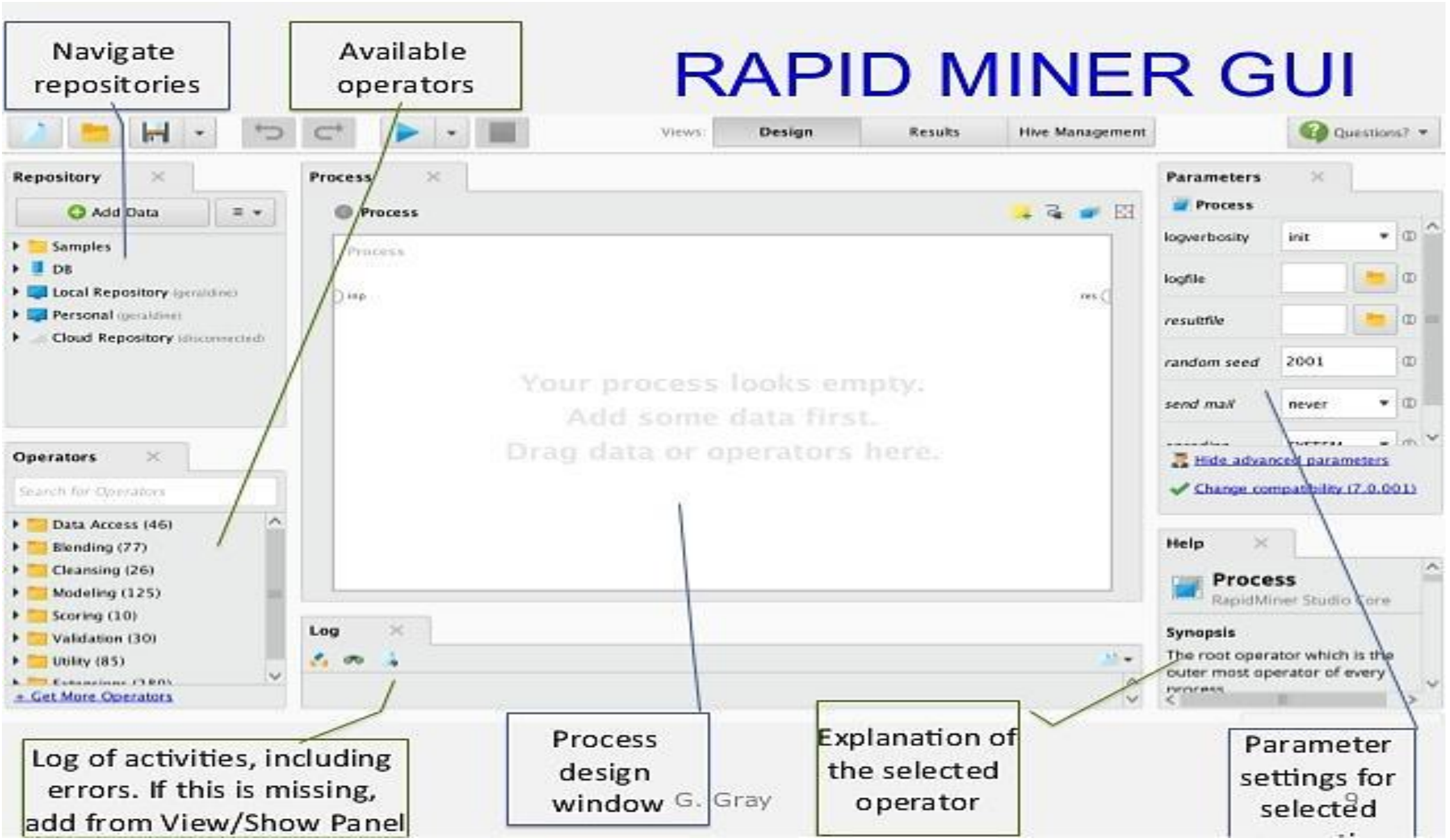
- Link buku Matlab: https://drive.google.com/drive/folders/1qHLqc2kYrI7REC2UClijIZhrzICmm8AF?usp=sharing
- Link buku Deep Learning with Matlab: https://drive.google.com/drive/folders/1QuU9tAMPF-XPwM4WmSBRiSYQoj8aA9Wg?usp=sharing

Tool Implementasi: RapidMiner

- RapidMiner https://rapidminer.com/
  - platform perangkat lunak data science
  - yang dikembangkan oleh perusahaan bernama sama dengan yang menyediakan lingkungan terintegrasi untuk **data preparation, machine learning, deep learning, text mining, and predictive analytics.**
  - Digunakan untuk bisnis dan komersial, juga untuk penelitian, pendidikan, pelatihan, *rapid prototyping*, dan pengembangan aplikasi serta mendukung semua langkah dalam proses **machine learning termasuk data preparation, results visualization, model validation and optimization**.
  - RapidMiner dikembangkan pada **open core model**. Dengan RapidMiner Studio **Free Edition**, yang terbatas untuk 1 prosesor logika dan 10.000 baris data, tersedia di bawah lisensi AGPL. **RapidMiner Studio 9.7 (https://my.rapidminer.com/nexus/account/index.html#downloads)** Harga komersial dimulai dari $2.500 dan tersedia dari pengembang.
- Link buku RapidMiner: https://drive.google.com/drive/folders/1ln2R4ryr2qj_Iwbk-ZZT_T9wTyvpuhaN?usp=sharing


Views:
Design
Results
Questions?
Repository
Add Data
Samples
data
processes
Templates
Tutorials
DB
Local Repository (Denis Schernov)
Cloud Repository (disconnected)
2
Operators
Search for Operators
Data Access (32)
Blending (77)
Cleansing (26)
Modeling (125)
Scoring (10)
Validation (30)
Utility (84)
Extensions (17)
Get More Operators
1
Process
Process
Process
inp
res
Your process looks empty.
Add some data first.
Drag data or operators here.
3
4
5
Parameters
Process
logverbosity
init
logfile
resultfile
random seed
2001
send mail
never
encoding
SYSTEM
Hide advanced parameters
Compatibility level
7.1.000-SN
6
Help
Process
RapidMiner Studio Core
Synopsis
The root operator of every process.
Description
This operator provides a set of parameters that are of global relevance to the process like logging and initialization of parameters of the random number generator.
7

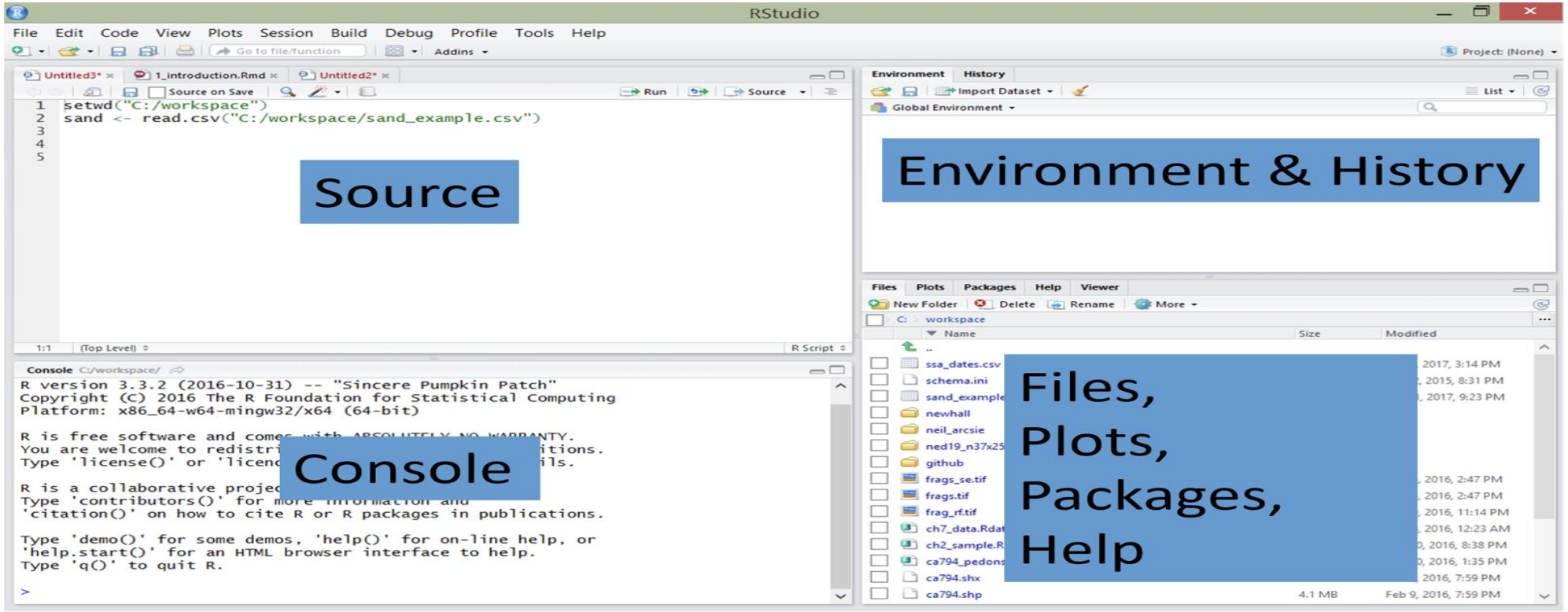
RAPID MINER GUI



Tool Implementasi: R-Studio

4 RStudio: An Integrated Development Environment (IDE) for R

RStudio is an integrated development environment (IDE) that allows you to interact with R more readily. RStudio is similar to the standard RGui, but is considerably more user friendly. It has more drop-down menus, windows with multiple tabs, and many customization options. The first time you open RStudio, you will see three windows. A forth window is hidden by default, but can be opened by clicking the **File** drop-down menu, then **New File,** and then **R Script.**. Detailed information on using RStudio can be found at at RStudio's Website.

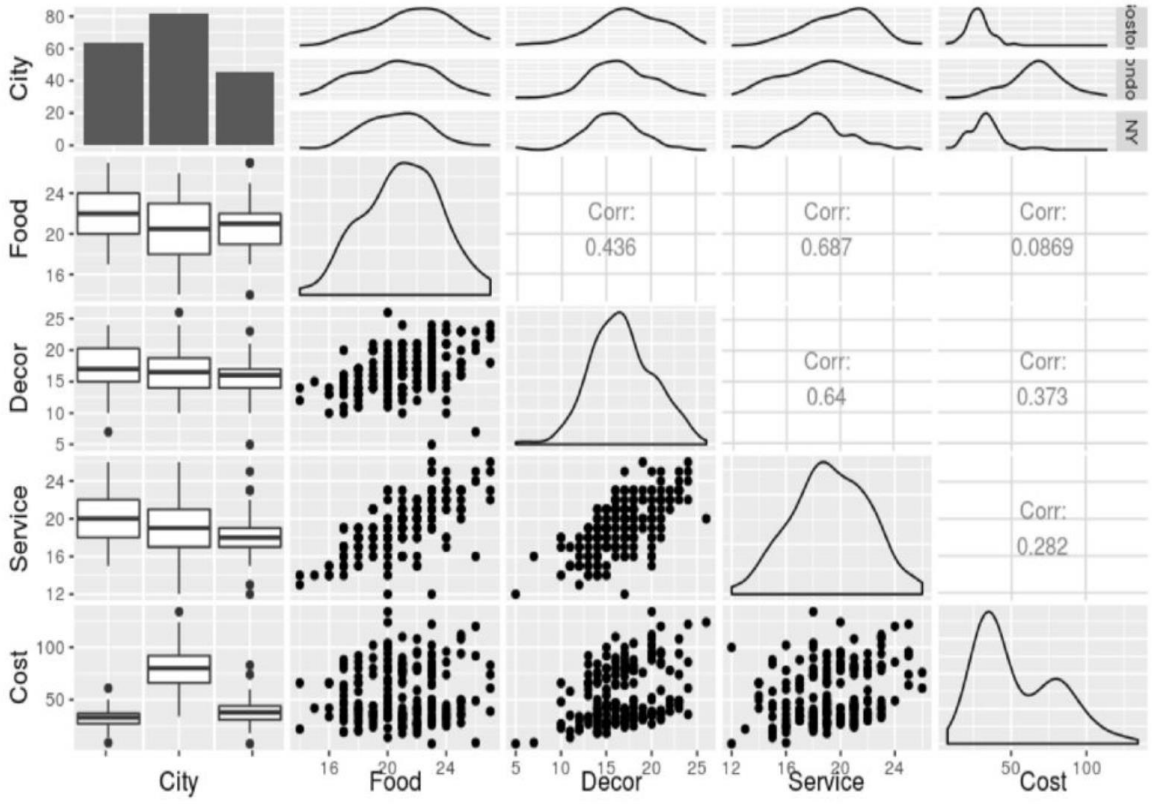


RStudio Windows / Tabs	Location	Description

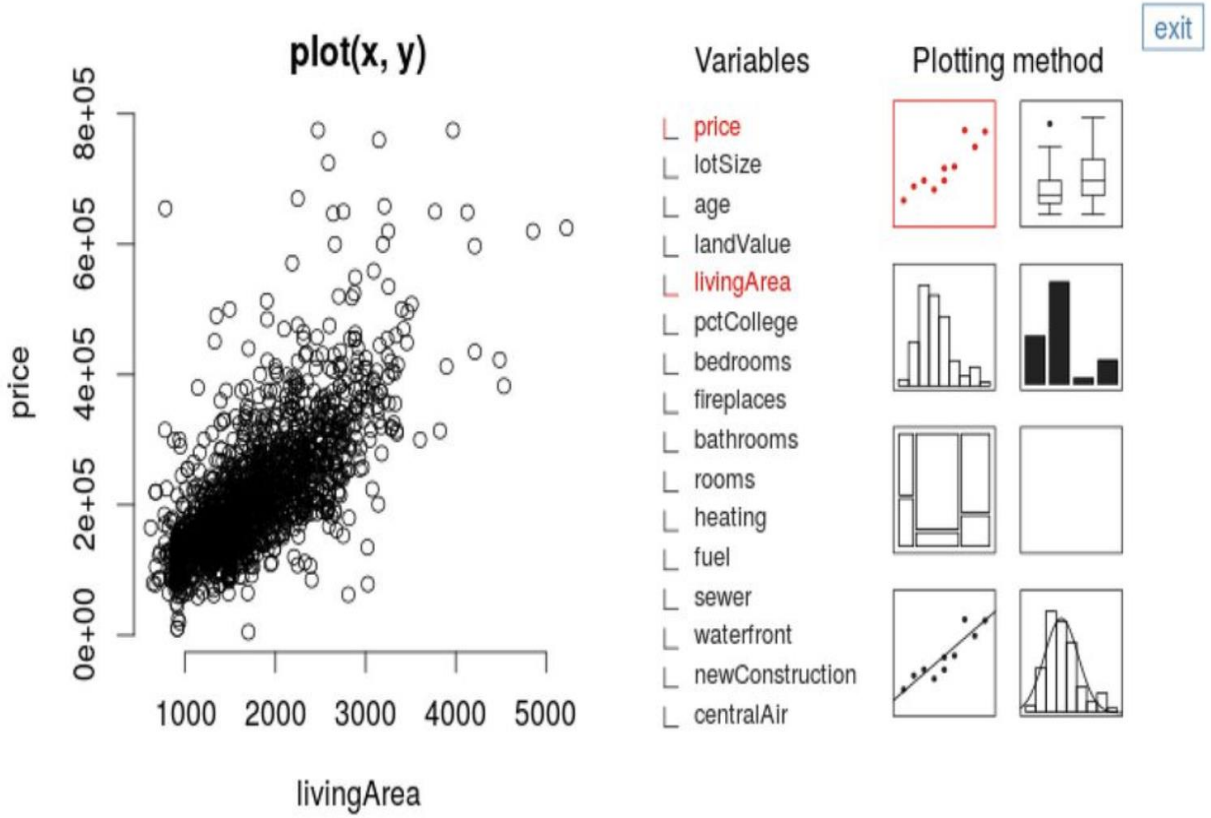
Mengapa Pakai R Language ?

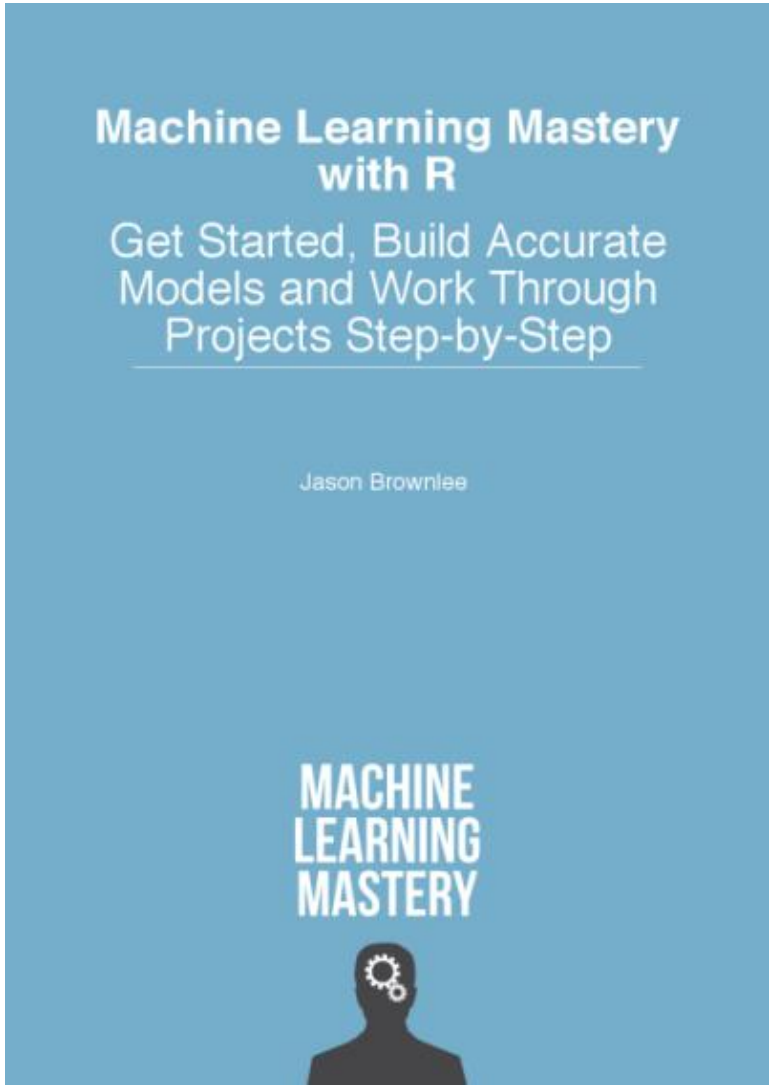
- R is a free, open-source software and programming language developed in 1995 at the University of Auckland as an environment for statistical computing and graphics (Ikaha and Gentleman, 1996).
- Since then R has become one of the dominant software environments for **data analysis** and is used by a variety of scientific disiplines, including soil science, ecology, and geoinformatics (Envirometrics CRAN Task View; Spatial CRAN Task View).
- R is particularly popular for its graphical capabilities, but it is also prized for it's GIS capabilities which make it relatively easy to

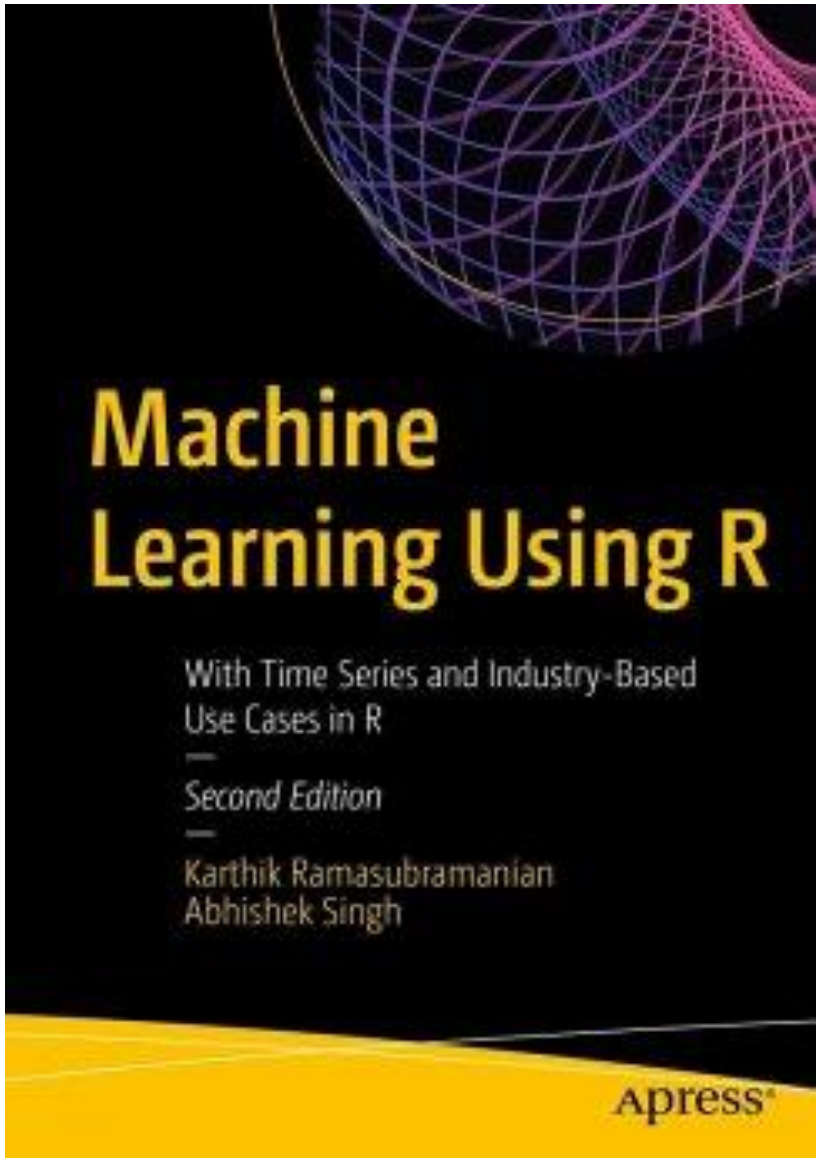
An example of graphing using the GGally package in R

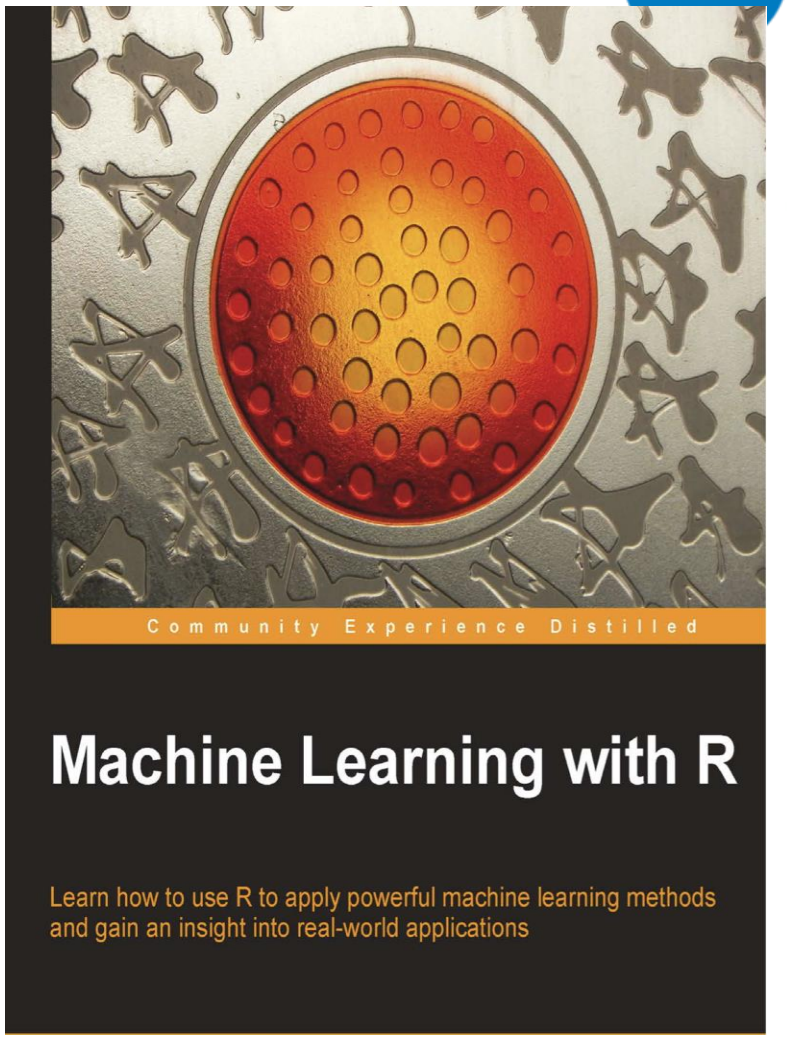


The openintro package edaPlot() command makes exploring data graphically easy to do. edaPlot(SaratogaHouses)




Machine Learning Mastery with R
Get Started, Build Accurate Models and Work Through Projects Step-by-Step
Jason Brownlee
MACHINE LEARNING MASTERY

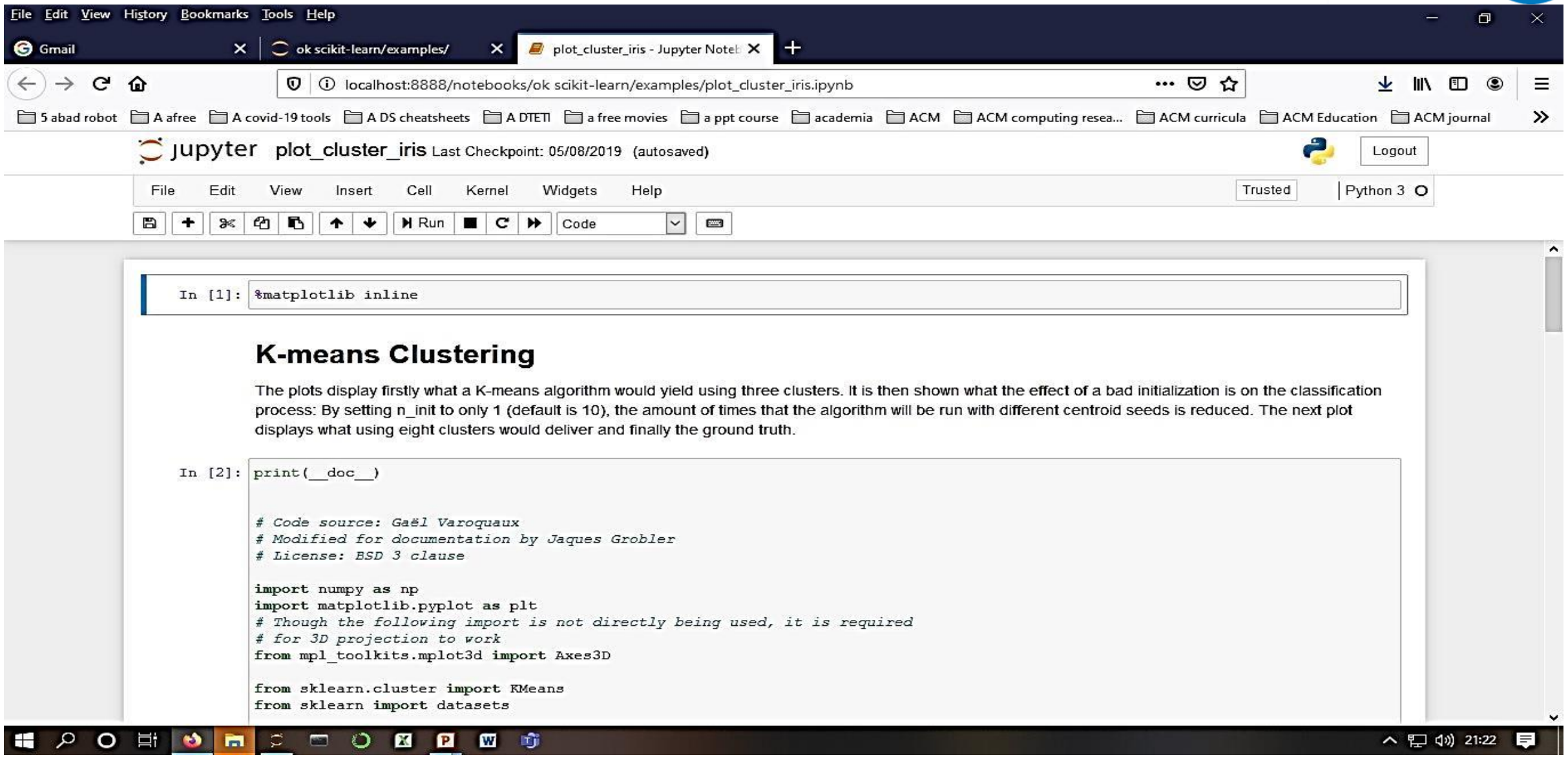

Machine Learning Using R
With Time Series and Industry-Based Use Cases in R
Second Edition
Karthik Ramasubramanian
Abhishek Singh
Apress


Community Experience Distilled
Machine Learning with R
Learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications

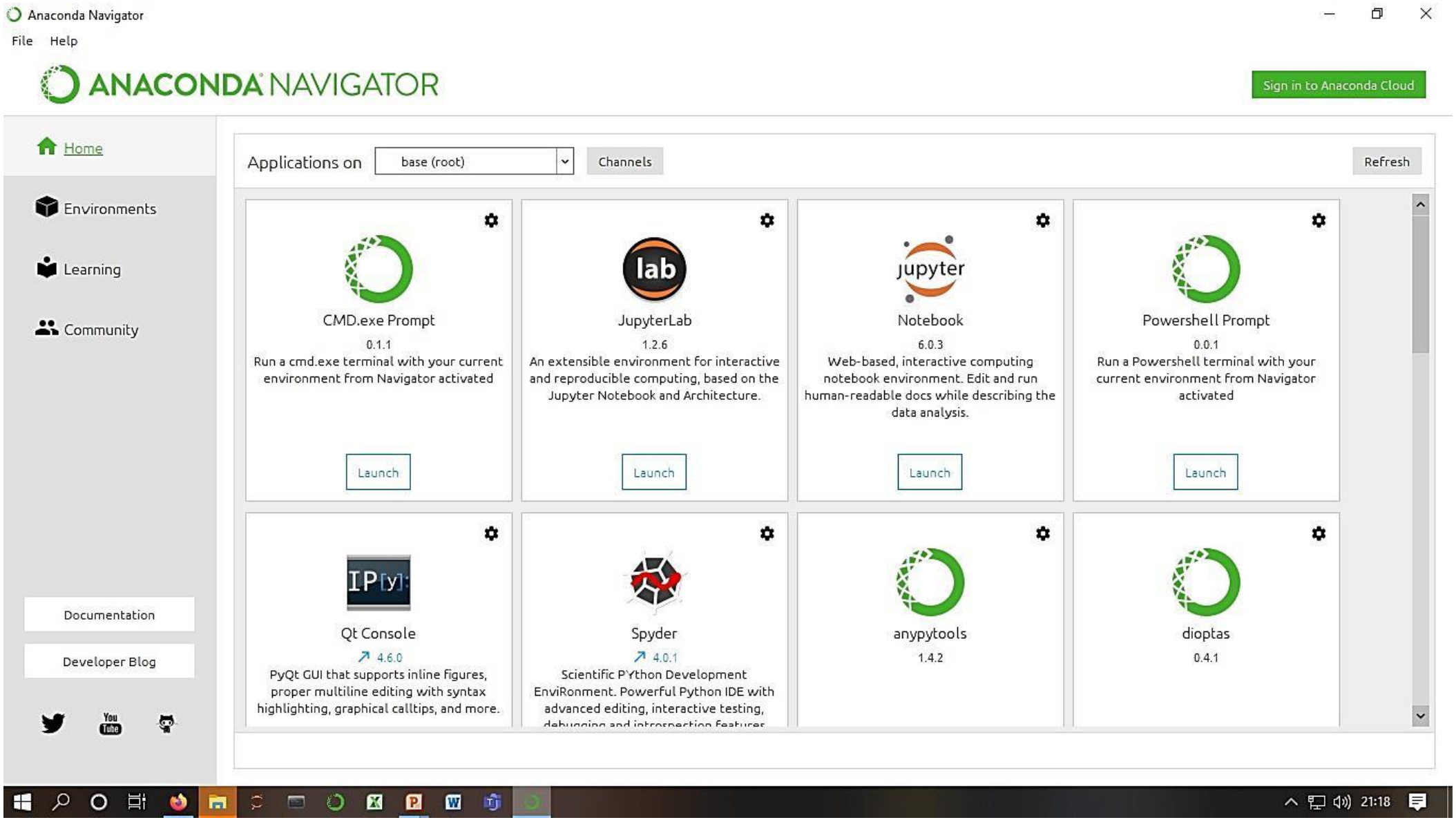
Tool Implementasi: Python, Jupyter, Anaconda

- Python
  - Versi 3.8.X
  - Tersedia IDE: **Spyder** https://www.spyder-ide.org/
  - Tool interactive: **Jupyter** *(Project Jupyter exists to develop open-source software, open-standards, and services for interactive computing across dozens of programming languages.)* https://jupyter.org/
  - Toolkit: **Anaconda** (*the open-source Individual Edition (Distribution) is the easiest way to perform Python/R data science and machine learning on a single machine. Developed for solo practitioners, it is the toolkit that equips you to work with thousands of open-source packages and libraries*) https://www.anaconda.com/
  - **Google Colab** Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with
    - Zero configuration required
    - Free access to GPUs
    - Easy sharing
    - Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work easier https://colab.research.google.com/notebooks/intro.ipynb

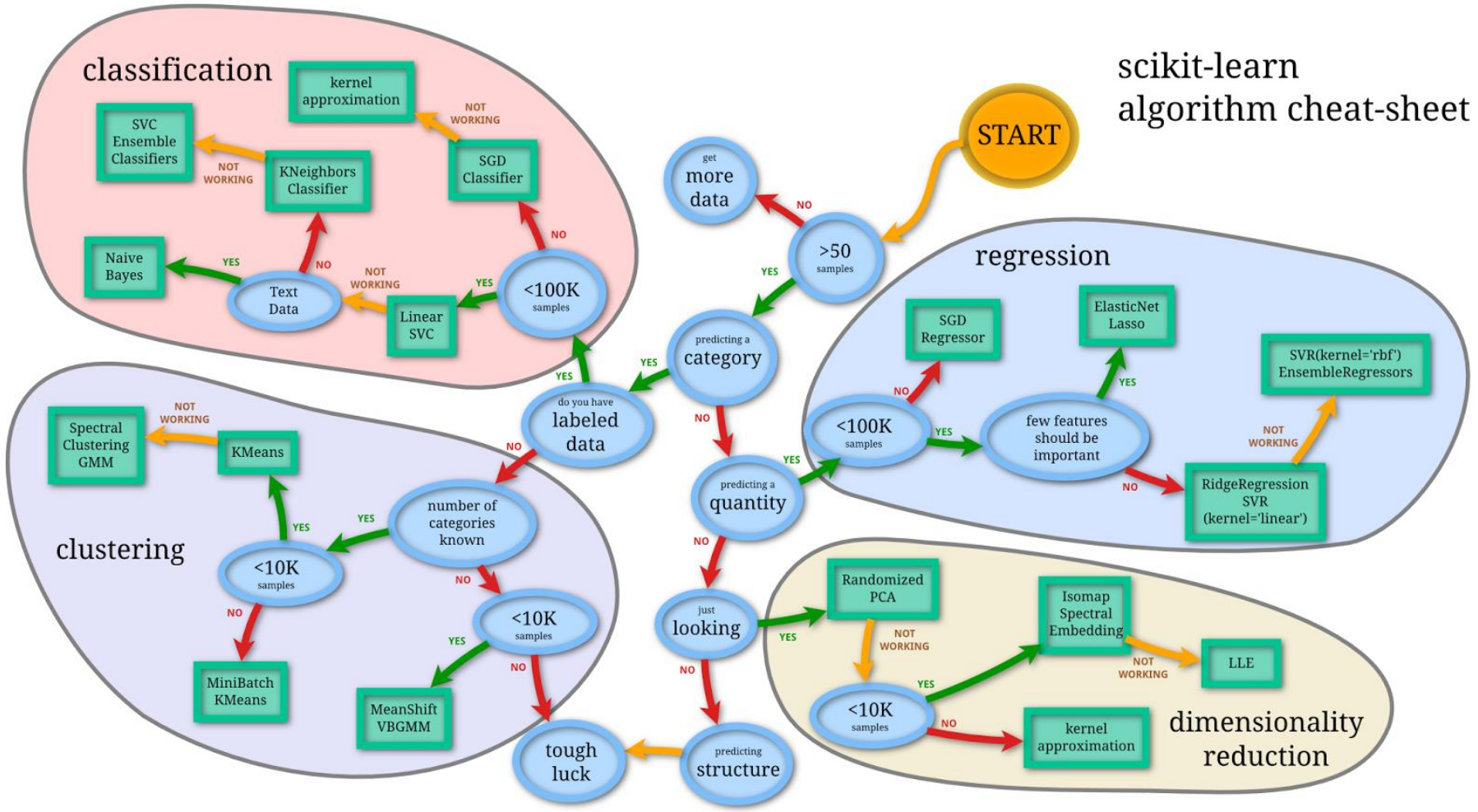
Jupyter


File Edit View History Bookmarks Tools Help
Gmail
ok scikit-learn/examples/
plot_cluster_iris - Jupyter Noteb
localhost:8888/notebooks/ok scikit-learn/examples/plot_cluster_iris.ipynb
5 abad robot
A afree
A covid-19 tools
A DS cheatsheets
A DTETI
a free movies
a ppt course
academia
ACM
ACM computing resea...
ACM curricula
ACM Education
ACM journal
jupyter plot_cluster_iris Last Checkpoint: 05/08/2019 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [1]: %matplotlib inline
K-means Clustering
The plots display firstly what a K-means algorithm would yield using three clusters. It is then shown what the effect of a bad initialization is on the classification process: By setting n_init to only 1 (default is 10), the amount of times that the algorithm will be run with different centroid seeds is reduced. The next plot displays what using eight clusters would deliver and finally the ground truth.
In [2]: print(__doc__)
# Code source: Gaël Varoquaux
# Modified for documentation by Jaques Grobler
# License: BSD 3 clause
import numpy as np
import matplotlib.pyplot as plt
# Though the following import is not directly being used, it is required
# for 3D projection to work
from mpl_toolkits.mplot3d import Axes3D
from sklearn.cluster import KMeans
from sklearn import datasets
21:22

Anaconda


Anaconda Navigator
File Help
ANACONDA NAVIGATOR
Sign in to Anaconda Cloud
Home
Environments
Learning
Community
Documentation
Developer Blog
Applications on
base (root)
Channels
Refresh
CMD.exe Prompt
0.1.1
Run a cmd.exe terminal with your current environment from Navigator activated
Launch
JupyterLab
1.2.6
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.
Launch
Notebook
6.0.3
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.
Launch
Powershell Prompt
0.0.1
Run a Powershell terminal with your current environment from Navigator activated
Launch
Qt Console
4.6.0
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.
Spyder
4.0.1
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing,
anypytools
1.4.2
dioptas
0.4.1
21:18

Library: scikit-learn



SCIKIT LEARN : https://scikit-learn.org/stable/

- Link buku-buku
  - **Big-data dan Data Science**: https://drive.google.com/drive/folders/18jbNHjUWsRor8W64oNDxggOd_yWqMzHs?usp=sharing
  - Deep Learning dan Machine Learning: https://drive.google.com/drive/folders/1hJ-E5OJhg35R7LC7_bHy99ccoY7CJ3nO?usp=sharing
  - **Python:** https://drive.google.com/drive/folders/1zqr5GPjQhP96XqKcWMxcmeWiMAAZ1iVx?usp=sharing

scikit-learn

- scikit-learn user guide, Mar 01, 2019 : https://drive.google.com/drive/folders/1rRsU6WdnPUlT3d9NcsTuk2f6N92PLZkc?usp=sharing

Terima Kasih