Introdução à Aprendizagem Automática (IAA)

SUSANA BRÁS

SUSANA.BRAS@UA.PT

IAA – L1

- IAA
  - presentation
  - evaluation
- The AI process: Data, Model, Value
- Characteristics of a Data Professional
- Terminology
- Difference between unsupervised and supervised learning
- Difference between classification and regression
- Data Visualization
- Pre-Processing:
  - Data Cleaning:
    - missing values
    - incorrect entries: inconsistency detection, domain knowledge, outliers, noise and artifacts detection
    - scaling and normalization
  - Data sampling
  - Feature selection
    - Non-informative features (impact and reason to remove it)
    - Redundant features (impact and reason to remove it)
    - Wrapper methods (backward or forward selection)
    - Filter methods (e.g. hypothesis testing, entropy)
    - Unsupervised methods

IAA

1,5 h, TP class:

- Friday 14h
- Anf. IV
- Prof. Susana Brás

2h PL class:

- Tuesday, 15h
  - PL1: 4.2.07, prof. Susana Brás
  - PL2: 4.2.04, Prof. Filipe Silva
- Friday, 9h
  - PL2, 4.2.08, prof. João Rodrigues
- Friday, 11h
  - PL3, 4.2.08, prof. João Rodrigues
  - PL5, 4.2.27, prof. Ana Rocha

Contacts:

Susana Brás – susana.bras@ua.pt

João Rodrigues – jmr@ua.pt

Filipe Silva – fmsilva@ua.pt

Ana Rocha – aprocha@ua.pt

IAA

Goals:

Understanding the basics of Machine Learning (ML) algorithms (Regression, Classification, Clustering, Deep Learning):

- Design, implementation and optimization of models, by the evalutation of the topic, the question, the data characteristics.
- Apply ML algorithms to practical problems in different application domains.
- Formulate real-world problems in the context of ML terminology and choose the most relevant approach to solving the problem.
- Performance evaluation and comparison between models.
- Learn the concepts of representativeness, overfitting and generalization.
- Develop critical and analytical reasoning, approaching complex problems logically and systematically.
- Learn to work as a team, strengthen communication skills.

PL

1) Install Anaconda 3 for Python 3:

https://docs.anaconda.com/anaconda/install/

2) Learn how to use Jupyter Notebook (part of Anaconda)

https://www.dataquest.io/blog/jupyter-notebook-tutorial/

IAA – Core Concepts

C1 — Data Literacy

- ✓ Understanding the nature, quality, and limitations of data
- ✓ Identification of inconsistencies, missing values, and bias

C2 — Preparation & Representation

- ✓ Feature selection
- ✓ Scaling, normalization, encoding
- ✓ Awareness of the impact of representation decisions

C3 — Algorithmic Understanding

- ✓ Conceptual understanding (not just API usage)
- ✓ Relationship between objective, cost, optimization, and model behavior

C4 — Evaluation & Diagnosis

- ✓ Appropriate choice of metrics
- ✓ Identification of overfitting / underfitting
- ✓ Use of validation and learning curves

C5 — Critical & Ethical Thinking

- ✓ Careful language in interpretation
- ✓ Awareness of bias, social impact, and limitations
- ✓ Ability to question results

C6 — Technical Communication

- ✓ Clear justification of decisions
- ✓ Use of scientific Markdown
- ✓ Reproducibility of work

IAA – Evaluation

NF = 0.3*T1+0.3*T2+0.4*E

- Work 1 (T1):
  - **Individual** Assignment
  - A theoretical/technical written assignment where students, individually, must explore a data-related topic. At the end of the semester, the assignments will be compiled into a class portfolio on "Data Compliance, Quality, and Exploration.“
  - The use of AI tools is allowed; they should be mentioned in the document, with preference given to those that enable reference tracking (e.g., https://scite.ai/ , https://consensus.app/ , https://elicit.com/).
  - Assessment: Based on document quality, technical/scientific rigor, and critical analysis of the topic. Maximum 4 pages, IEEE format.
  - Mandatory document sections:
    - Critical Analysis - critical description of the selected topic, evidencing the student’s reflection on the acquired knowledge
    - Acknowledgement of used AI tools – including which tools had been used, the rationale behind the selection
  - There will be no presentation.
  - **Topic selection:** until 20/02/2026
  - **Submission**: until 13/03/2026.

IAA – Evaluation

Minimum grade in each component – 8 out of 20

NF = 0.3*T1+0.3*T2+0.4*E

- Work 2 (T2):
  - **Group** Assignment (3 members)
  - A practical assignment where students must solve a machine learning problem proposed by the group, using datasets available on web platforms and the tools taught in class.
  - Three scheduled submissions:
    - Submission 1 (07/03/2026) – Group definition, topic, title, goals.
    - Submission 2 (24/04/2026)– Preliminary results (selected algorithm, data processing methodology).
    - Submission 3 (29/05/2026) – Final document, code and presentation.
  - Presentation: During the last class of the semester.
  - Assessment:
    - Submissions 1 and 2: No quantitative assessment, only qualitative feedback if adjustments are needed.
    - Final Document: Assessed for quality, technical/scientific rigor, problem difficulty, and critical analysis. Maximum 6 pages, IEEE format.
    - Presentation: Evaluated for quality and discussion of the topic.
    - Peer Evaluation: Within the group.
    - Class Peer Evaluation: By classmates.
    - Document (40%), presentation (40%), Peer Evaluation (10%), Class Peer Evaluation (10%)

- Exam (E): **written exam**.

ECTS

1 ECTS	6 ECTS IAA	APPROXIMATELY	CLASSES	INDEPENDENT WORK
27 hours effort	162 hours of effort in 13 weeks	12.5 hours of effort per week	3.5 hours per week • 1.5 TP • 2 PL	5.67 hours a week • Class preparation • Autonomous tasks • Project development

A bit of history

1950, Alan Turing: "Computing Machinery and Intelligence" define the question "Can machines think?“ =>Turing test.

1956 –The field of Artificial Inteligente (AI) formally established at the conference in Dartmouth College.

1959, Arthur Samuel: “ Field of study that gives computers the ability to learn without being explicitly programmed ".

1998, Tom M. Mitchell: “ Can the computer program learn from experience ? “.

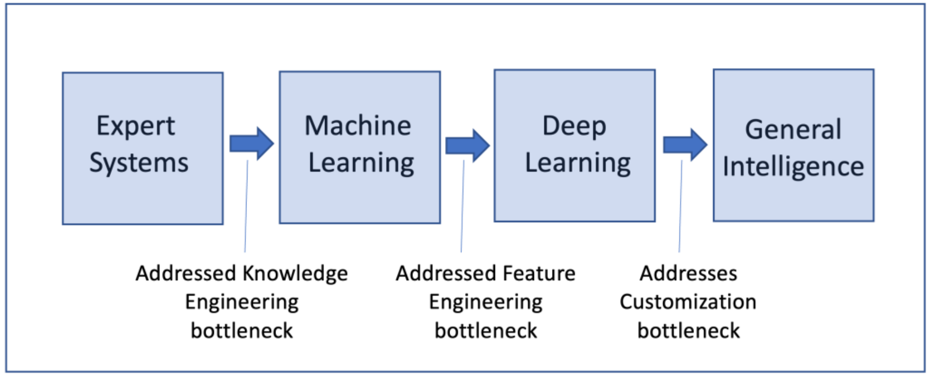


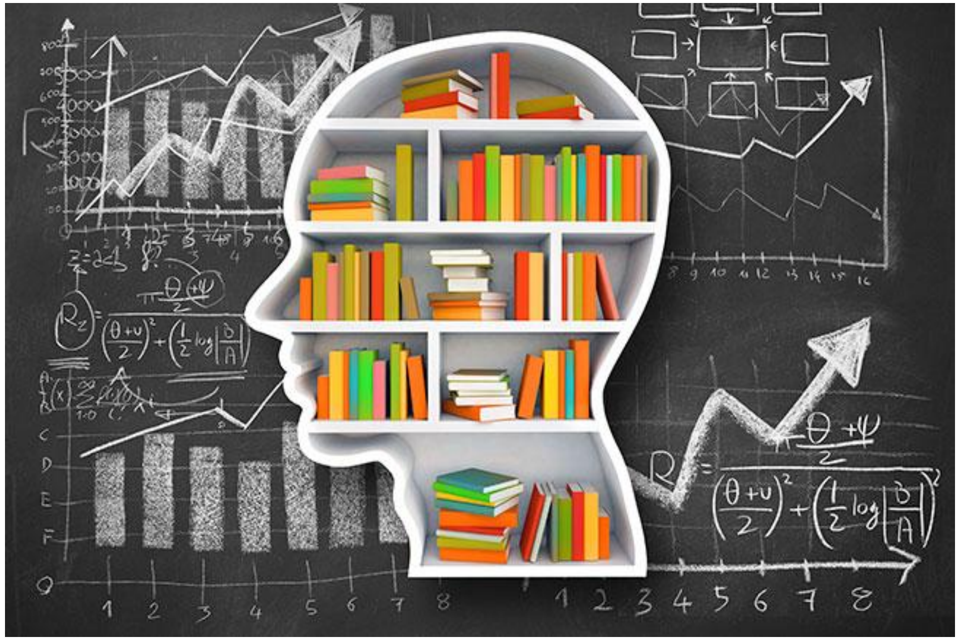
Figure. The history of artificial intelligence.

Table. The Paradigm Shifts in AI.

	DATA	EXEMPLAR	SCOPE	CURATION
Expert Systems	**Human**	**Rules**	**Follows**	**High**
Machine Learning	**+ Databases**	**Rules/networks**	**+ Discovers relationships**	**Medium**
Deep Learning	**+ Sensory**	**Deep neural networks**	**+ Senses relationships**	**Low**
General Intelligence	**+ Everything**	**Pre-trained deep neural networks**	**+ Understands the world**	**Minimal**

https://cacm.acm.org/research/the-paradigm-shifts-in-artificial-intelligence/

Important characteristic of a Data Professional



Important characteristic of a Data Professional

An important characteristic of a successful data professional is their ability to think **critically** and **analytically**. This skill involves approaching complex problems with a **logical** and **systematic** mindset, using data to **extract meaningful insights and make informed decisions**.

Important characteristic of a Data Professional

Problem Solving

Data Interpretation

Hypothesis Testing

Informed Decision-Making

Ethical Considerations

An important characteristic of a successful data professional is their ability to think **critically** and **analytically**. This skill involves approaching complex problems with a **logical** and **systematic** mindset, using data to **extract meaningful insights and make informed decisions**.

Adaptability

Identify Patterns and Trends

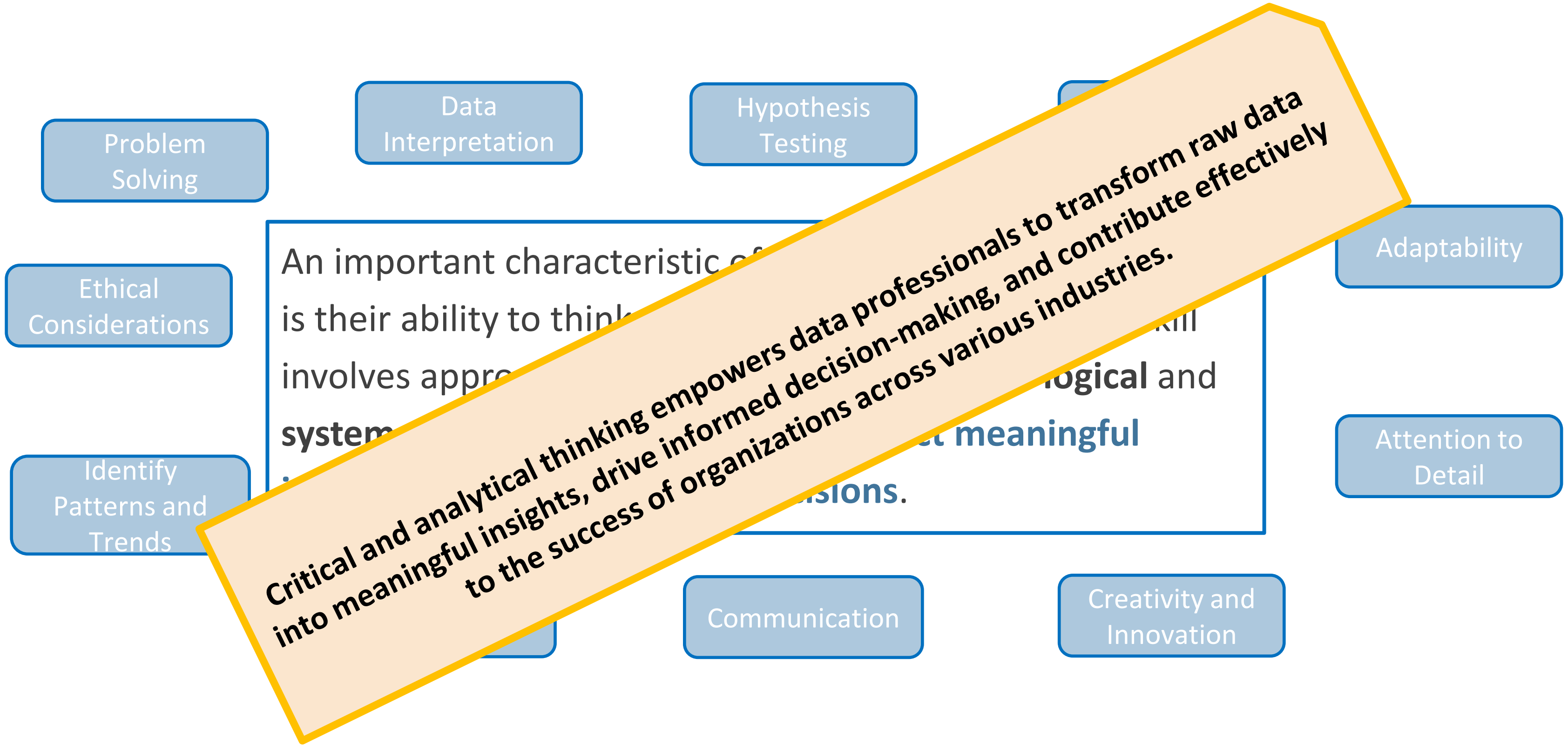
Attention to Detail

Continuous Learning

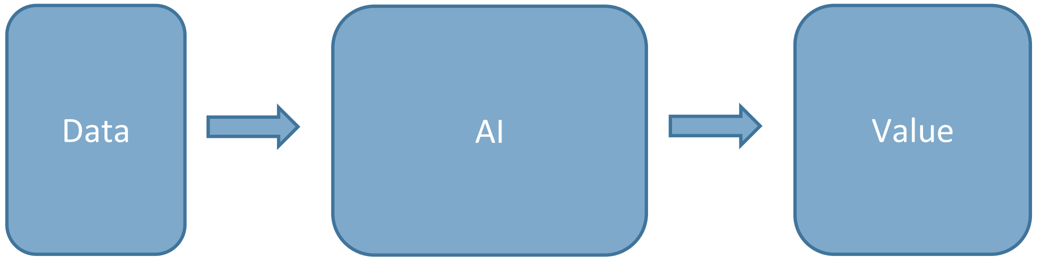
Communication

Creativity and Innovation

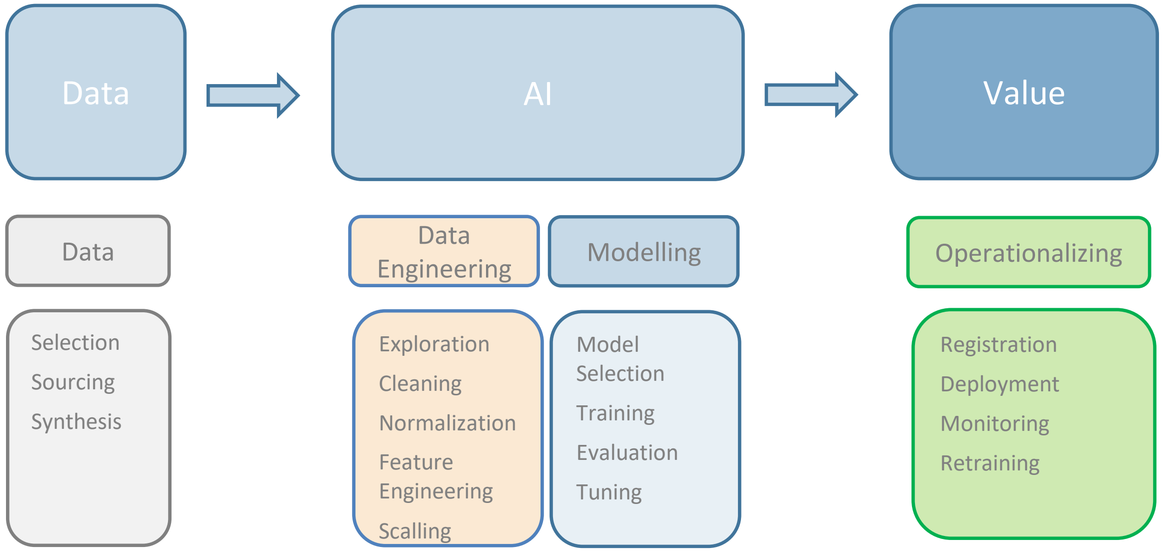
Important characteristic of a Data Professional



Pipeline

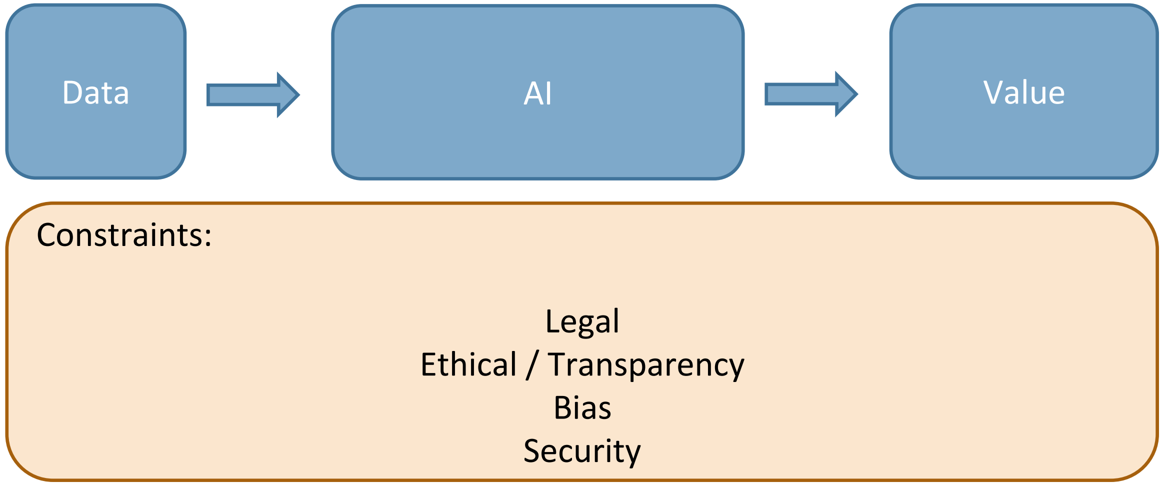


Pipeline



Data
AI
Value
Data
Data Engineering
Modelling
Operationalizing
Selection
Sourcing
Synthesis
Exploration
Cleaning
Normalization
Feature Engineering
Scalling
Model Selection
Training
Evaluation
Tuning
Registration
Deployment
Monitoring
Retraining

Pipeline


Data
AI
Value
Constraints:
Legal
Ethical / Transparency
Bias
Security

Data

Data – Data refers to a **collection of factual information**, **facts**, **statistics**, or **pieces of information** that are **collected**, **organized**, and **used** for various purposes. Data can take various forms, including numbers, text, images, audio, and more. *It serves as the raw material from which insights, patterns, and conclusions can be derived through analysis and interpretation.*

Data is a crucial resource in various fields, including science, business, healthcare, finance, and more. It is used to make informed decisions, uncover patterns and trends, develop insights, and drive innovation. The process of extracting meaningful information from data is known as data analysis, and it plays a vital role in many aspects of modern life.

Data

Photo by Luke Chesser na Unsplash

Data drives decisions, but not all data reflects reality.

- **Data-Driven Decisions:** Organizations increasingly rely on data to guide strategies, affecting workers, users, customers, etc.

- **Real-World Impacts:** From product to design to public health, biased data has measurable consequences on people's lives.

AI

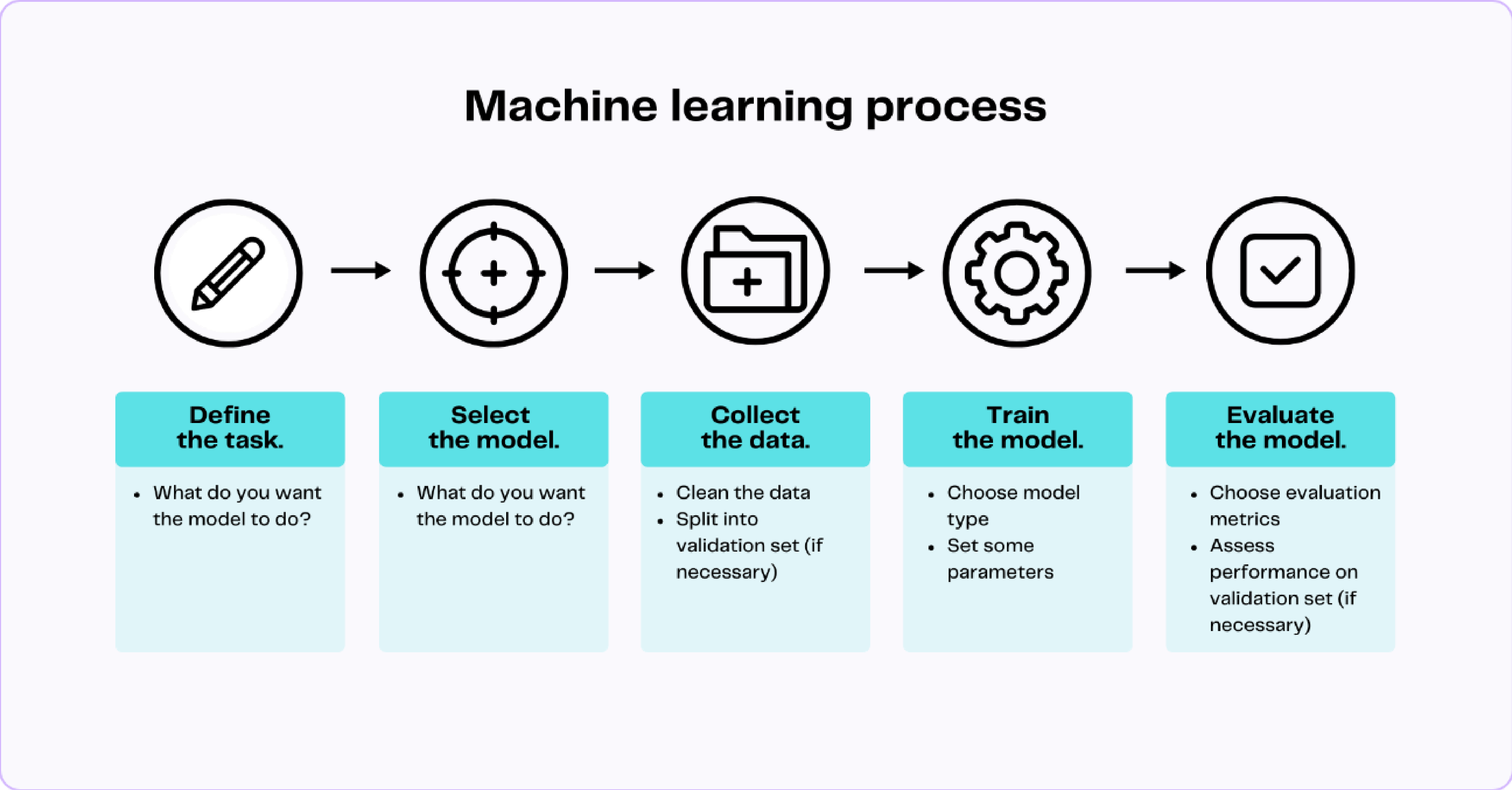
AI – AI stands for "Artificial Intelligence“. *It refers to the simulation of human intelligence processes by computer systems.* These systems can be broadly categorized into two main types:

Narrow or Weak AI: This type of AI is designed and trained for a specific task or a limited range of tasks. It operates within a defined domain and excels at performing the tasks it has been trained for.

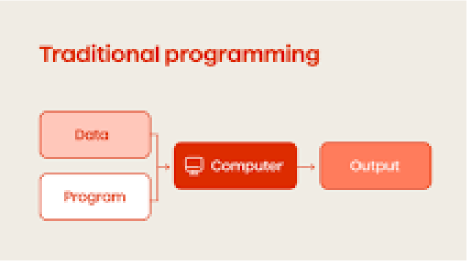
General or Strong AI: This type of AI, which **does not currently exist**, would possess human-like intelligence and be capable of understanding, learning, and performing any intellectual task that a human being can. It would exhibit a high level of adaptability across different domains and tasks.

AI has numerous real-world applications across industries such as healthcare (diagnosis and treatment planning), finance (algorithmic trading), automotive (self-driving cars), customer service (chatbots), and many others. It has the potential to transform various aspects of society and economy by **automating** tasks, enhancing decision-making.

Machine Learning Process



What is Machine Learning?

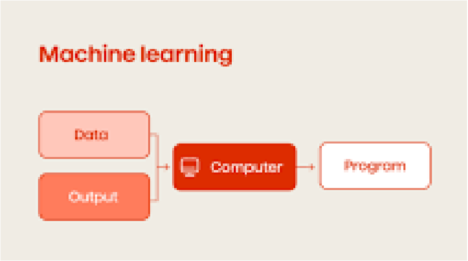


Traditional programming is about writing explicit instructions for the machine to follow.

- You provide the rules (logic) and data
- The computer processes these to produce output

Pros:
Easy to understand and debug
Control over every detail
Predictable behavior

Cons:
Not adaptive
Poor performance with massive or messy data
Time-consuming for complex problems



Machine Learning teaches the computer to learn from data.

- You provide the input and output
- The system figures out the rules by itself

Pros:
Handles complex and dynamic problems
Learns and improves over time
Can find hidden patterns

Cons:
Requires huge datasets
Hard to debug (black-box nature)
Computationally expensive

“A computer program is said to learn from experience if its performance improves with experience.”
(Tom Mitchell, simplified)

What is Machine Learning?

Factor	Traditional Programming	Machine Learning
Instruction Method	Explicit rules and logic	Learns patterns from data
Data Handling	Structured, predictable data	Large, often unstructured data
Outcome Predictability	Always the same result for same input	Predictions may vary as model adapts to new data
Flexibility	Limited to predefined conditions	Adjusts based on new data (self-improving)
Development Process	Linear: write, debug, deploy	Iterative: train, evaluate, tune, retrain
Transparency	Easy to trace and debug	Can be opaque, often needs explainable AI
Problem Complexity	Best for simple, well-defined tasks	Best for complex, data-rich tasks

Machine Learning Approaches

Supervised Learning

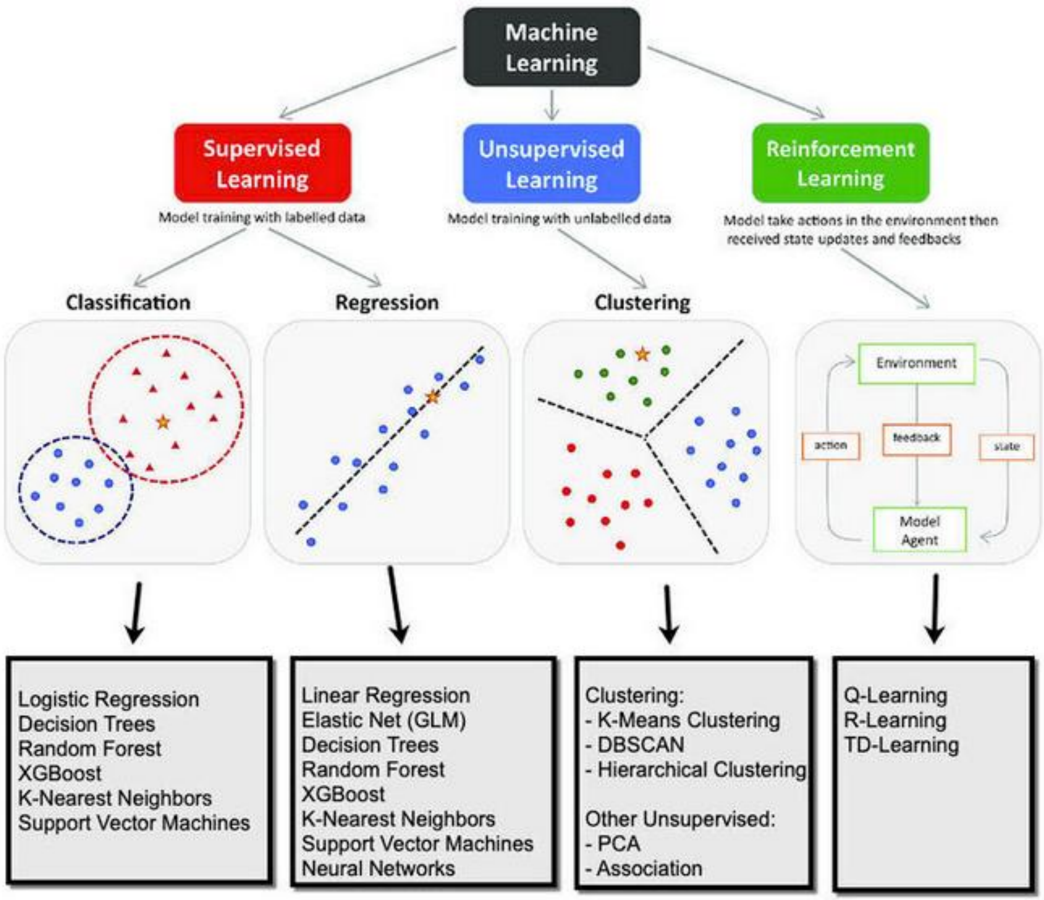
The data is labelled, and at the beginning, you know each observation in your dataset to which class belongs.

Example: given dataset with spam/not-spam labeled emails

Unsupervised Learning

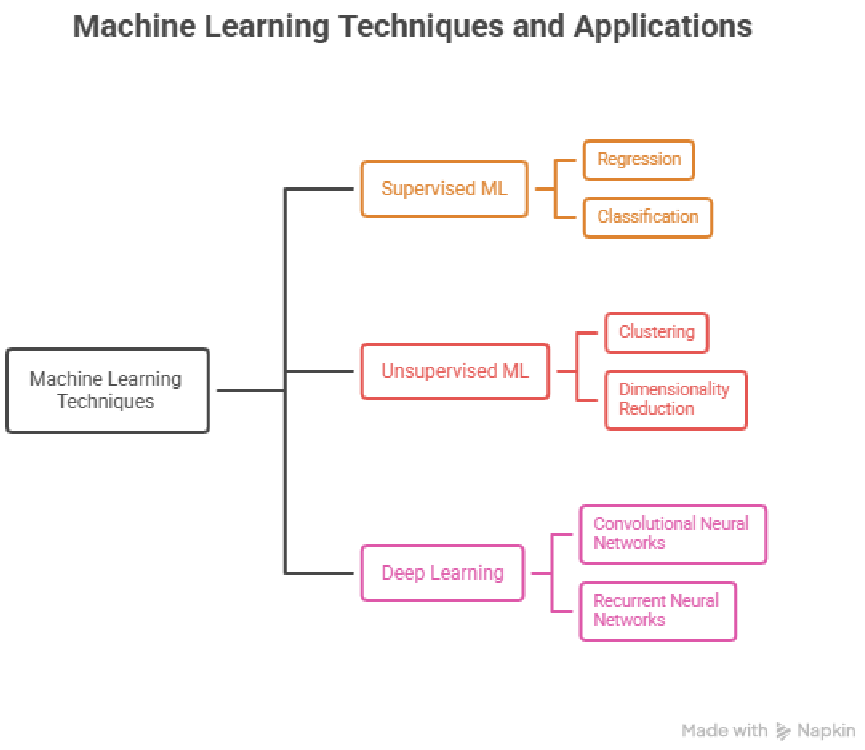
The data has no labels, and there is no association between the observations and a defined class. This methods are designed to explore and discover patterns in the data, allowing to find similar profiles in the observations.

Example: Group customers into segments based on purchase history.



Final Note:
Supervised learning focuses on prediction using labeled data, while unsupervised learning uncovers hidden patterns from unlabeled datasets.

IAA



Supervised learning

- Linear (univariate/ multivariate) regression
- Logistic regression
- Artificial Neural Networks (ANN)
- Support Vector Machines (SVM)
- Decision Tree (DT)
- Naive Bayes classifier
- k-Nearest Neighbor (k-NN) classifier

Unsupervised learning

- Clustering
- Principal components analysis (PCA)

Deep Learning

- CNN (Convolutional Neural Networks);
- RNN (Recurrent Neural Network)

Terminology

- The terms *sample*, *data point*, *observation*, or *instance* refer to a single, independent unit of data, such as a customer, patient, or compound. The term *sample* can also refer to a subset of data points, such as the training set sample. The text will clarify the appropriate context when this term is used.
- The *training set* consists of the data used to develop models while the *test* or *validation* sets are used solely for evaluating the performance of a final set of candidate models.
- The *predictors*, *independent variables*, *attributes*, or *descriptors* are the data used as input for the prediction equation.
- *Outcome*, *dependent variable*, *target*, *class*, or *response* refer to the outcome event or quantity that is being predicted.
- *Continuous* data have natural, numeric scales. Blood pressure, the cost of an item, or the number of bathrooms are all continuous. In the last case, the counts cannot be a fractional number, but is still treated as continuous data.
- *Categorical* data, otherwise known as *nominal*, *attribute*, or *discrete* data, take on specific values that have no scale. Credit status ("good" or "bad") or color ("red," "blue," etc.) are examples of these data.
- *Model building*, *model training*, and *parameter estimation* all refer to the process of using data to determine values of model equations.

Data Types

1. **Numeric (Quantitative) features**
   - Integer numbers
   - Floats (decimals) **-** temperature, height, weight, humidity, etc.

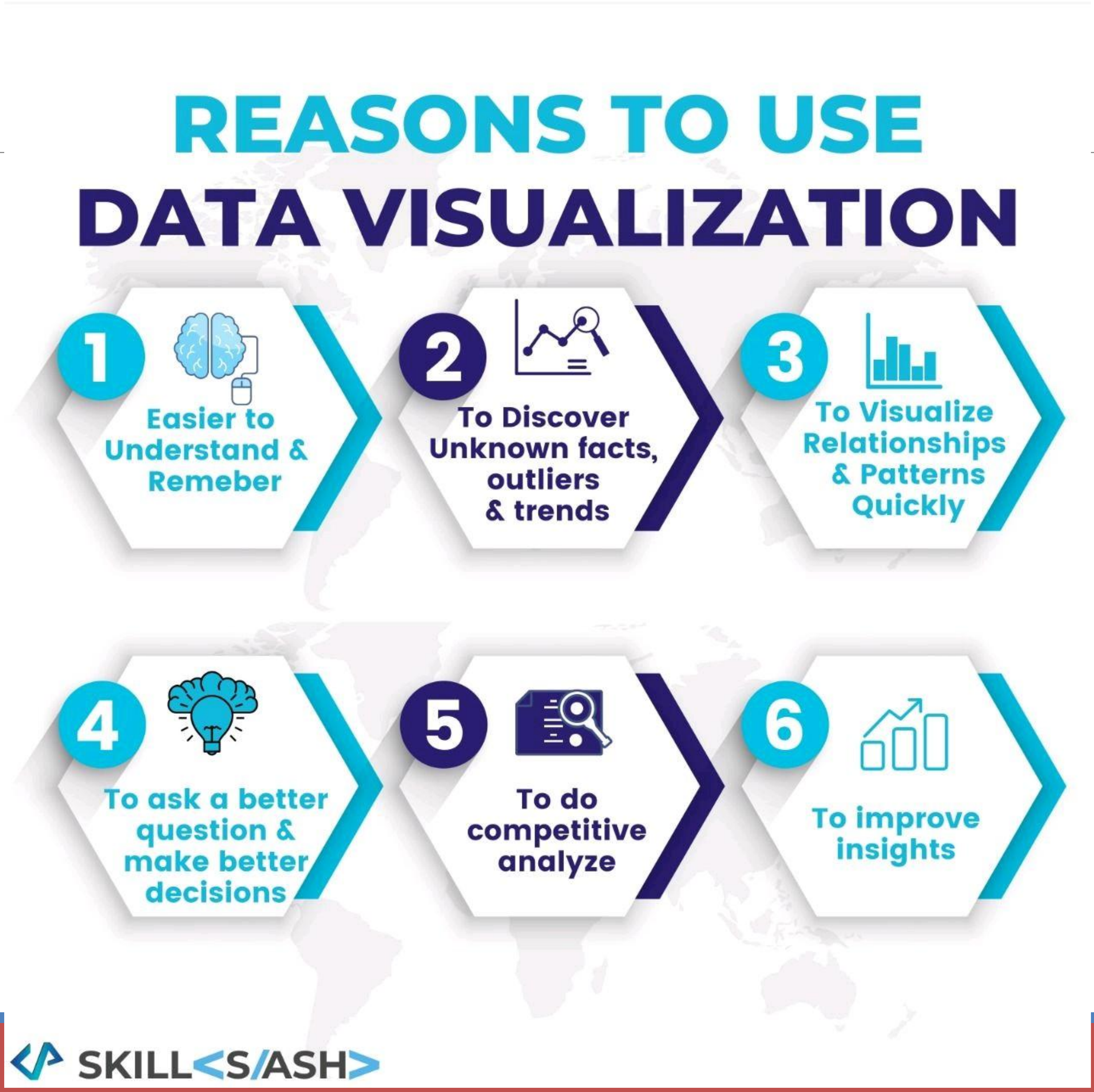
2. Boolean – True/False

3. Categorical features - gender, days of the week, seasons, country of birth, colors, etc.

How to deal with categorical features ? - One-hot encoding (1,0) transforms *n* categories into *n* features

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

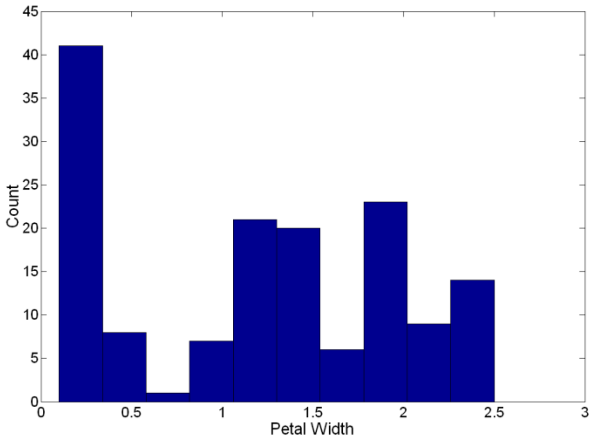

REASONS TO USE
DATA VISUALIZATION
1
Easier to Understand & Remeber
2
To Discover Unknown facts, outliers & trends
3
To Visualize Relationships & Patterns Quickly
4
To ask a better question & make better decisions
5
To do competitive analyze
6
To improve insights
SKILL<S/ASH>

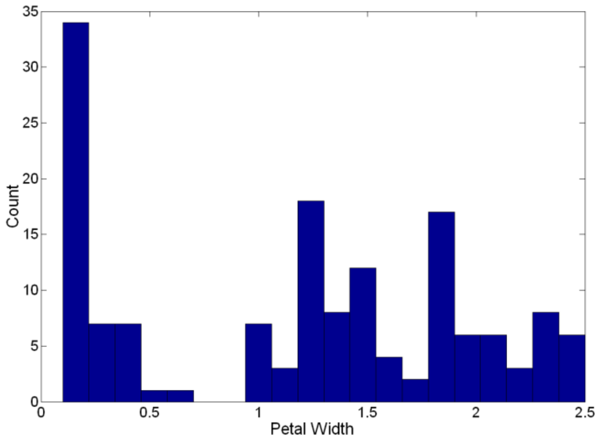
Data Visualization

Histograms

- Show the distribution of values of a single feature
- Divide the range of values of a single feature into bins and show bar plots of the number of examples in each bin.
- Histogram shape depends on the number of bins

Example: Petal Width (10 and 20 bins, respectively)

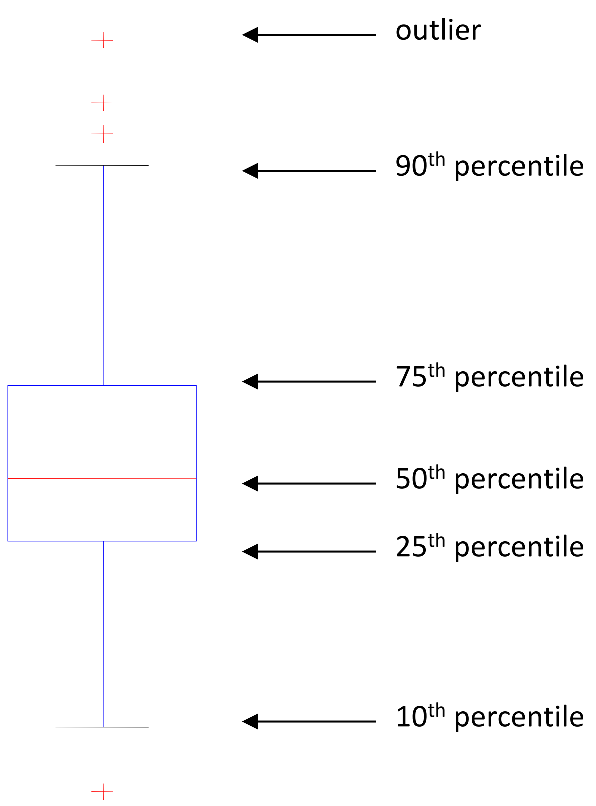




Data Visualization

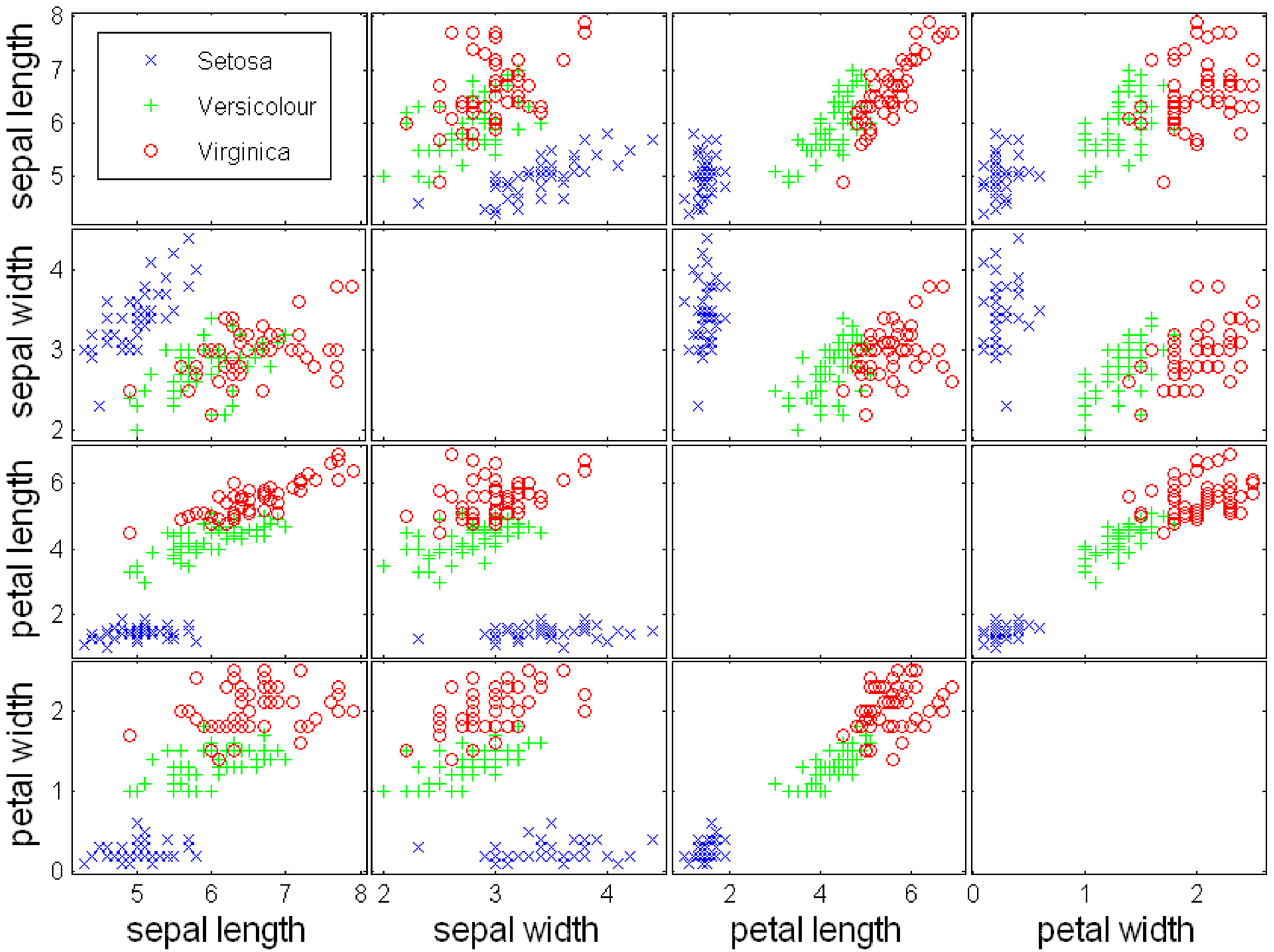
Box Plots

- Another way of displaying the distribution of data



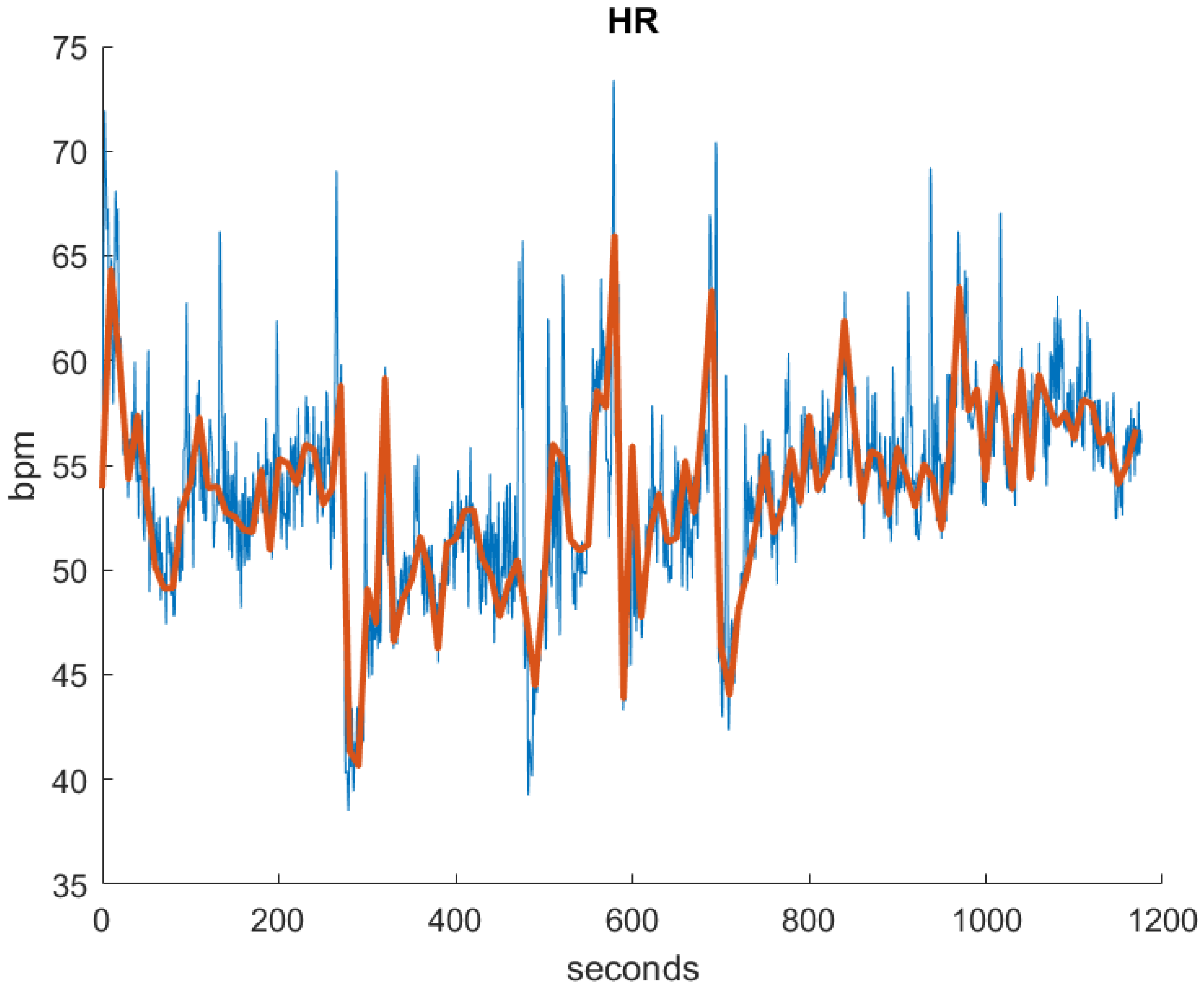
Data Visualization

Scatter Plot Array

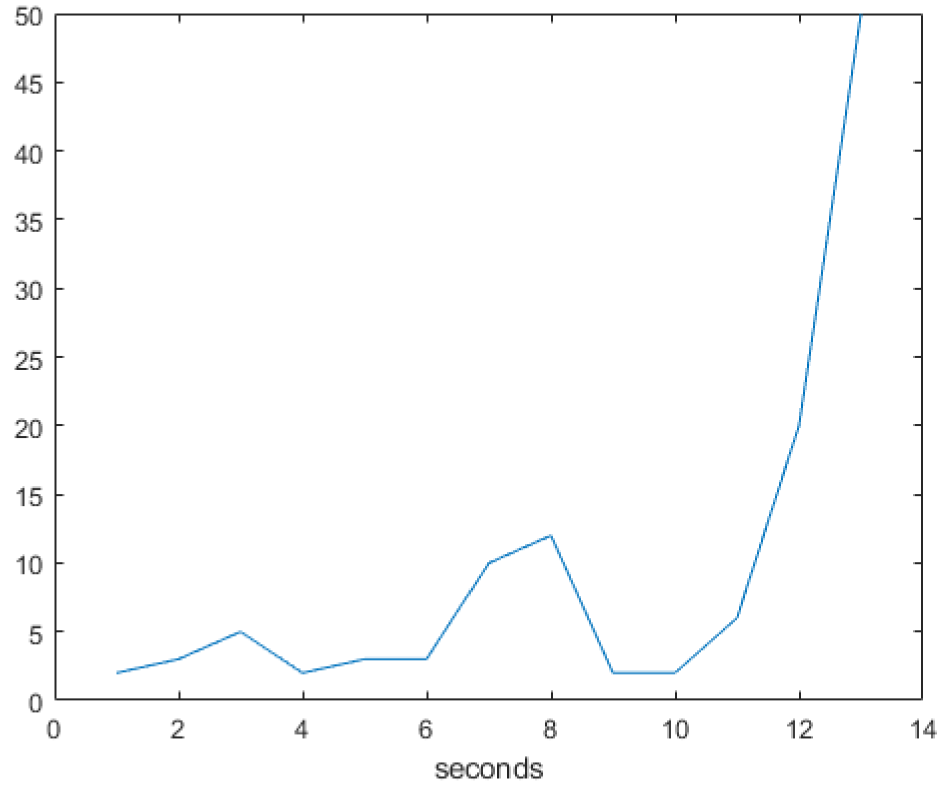


Data Visualization

Time series representation



Data Visualization - example



X = [2, 3, 5, 2, 3, 3, 10, 12, 2, 2, 6, 20, 50];

Central Tendency Measures:

mean = 9.2308

median = 3

mode = 2

Dispersion Measures:

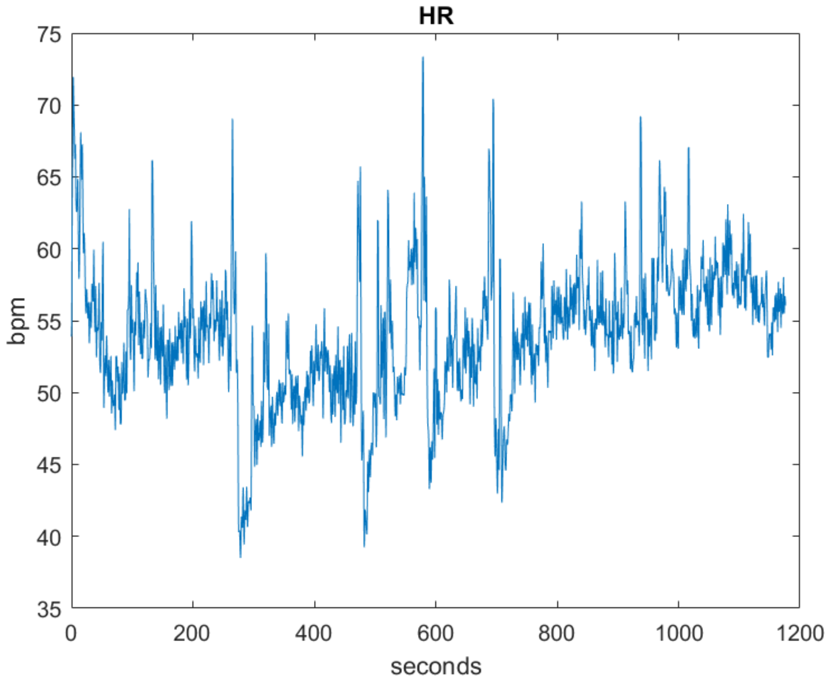
variance (var) = 178.3590

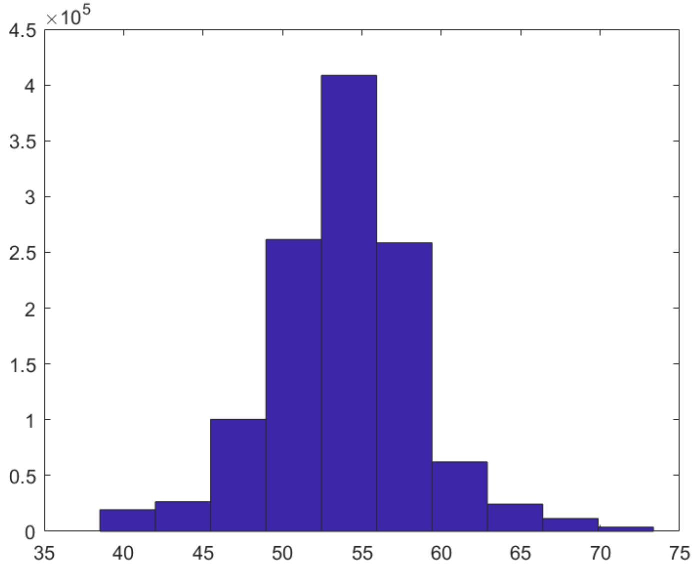
standard deviation (std) = 13.3551

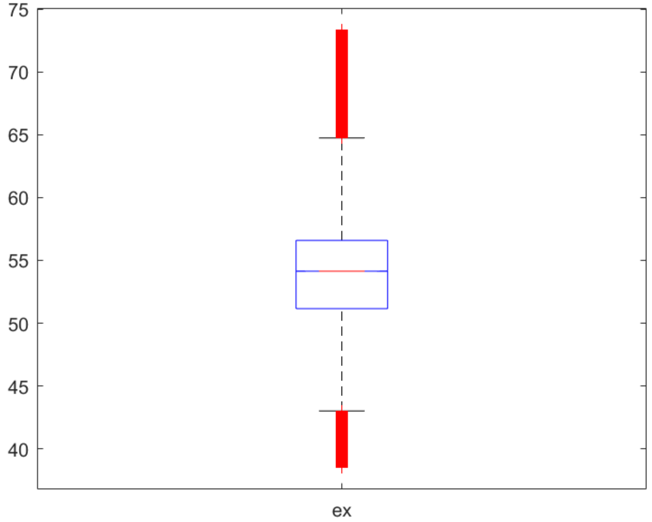
interquartile range (iqr) = 8.5000

range = 48

mean absolute deviation (mad) = 8.4734







Mean	53.91		Variance	22.01
Median	54.14		Std	4.69
Mode	47.81		iqr	5.43
			range	34.83
			mad	3.51

Data Processing Pipeline

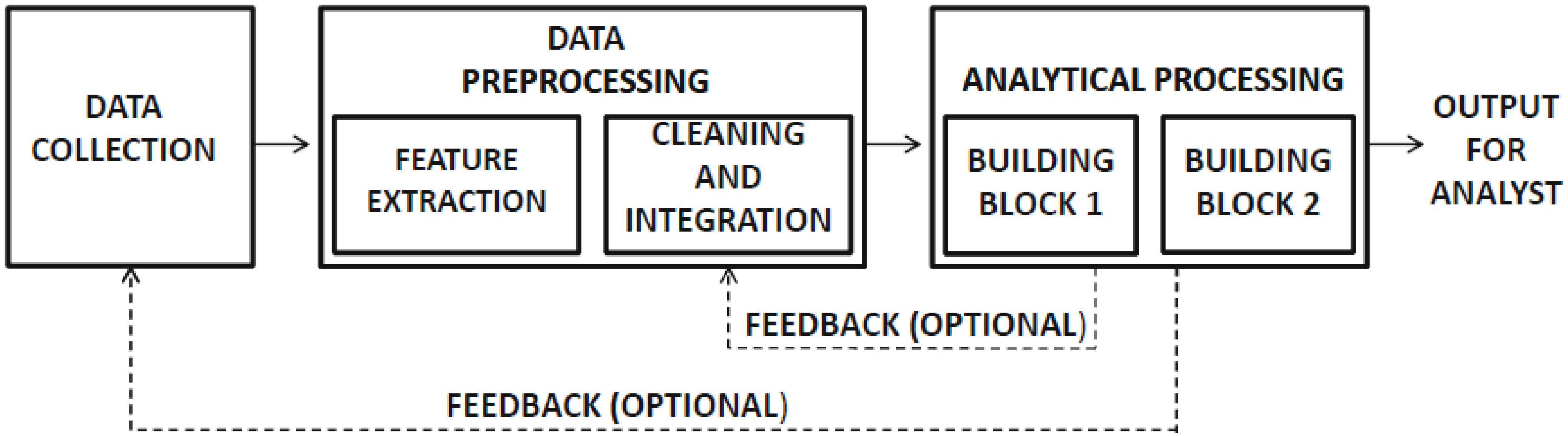


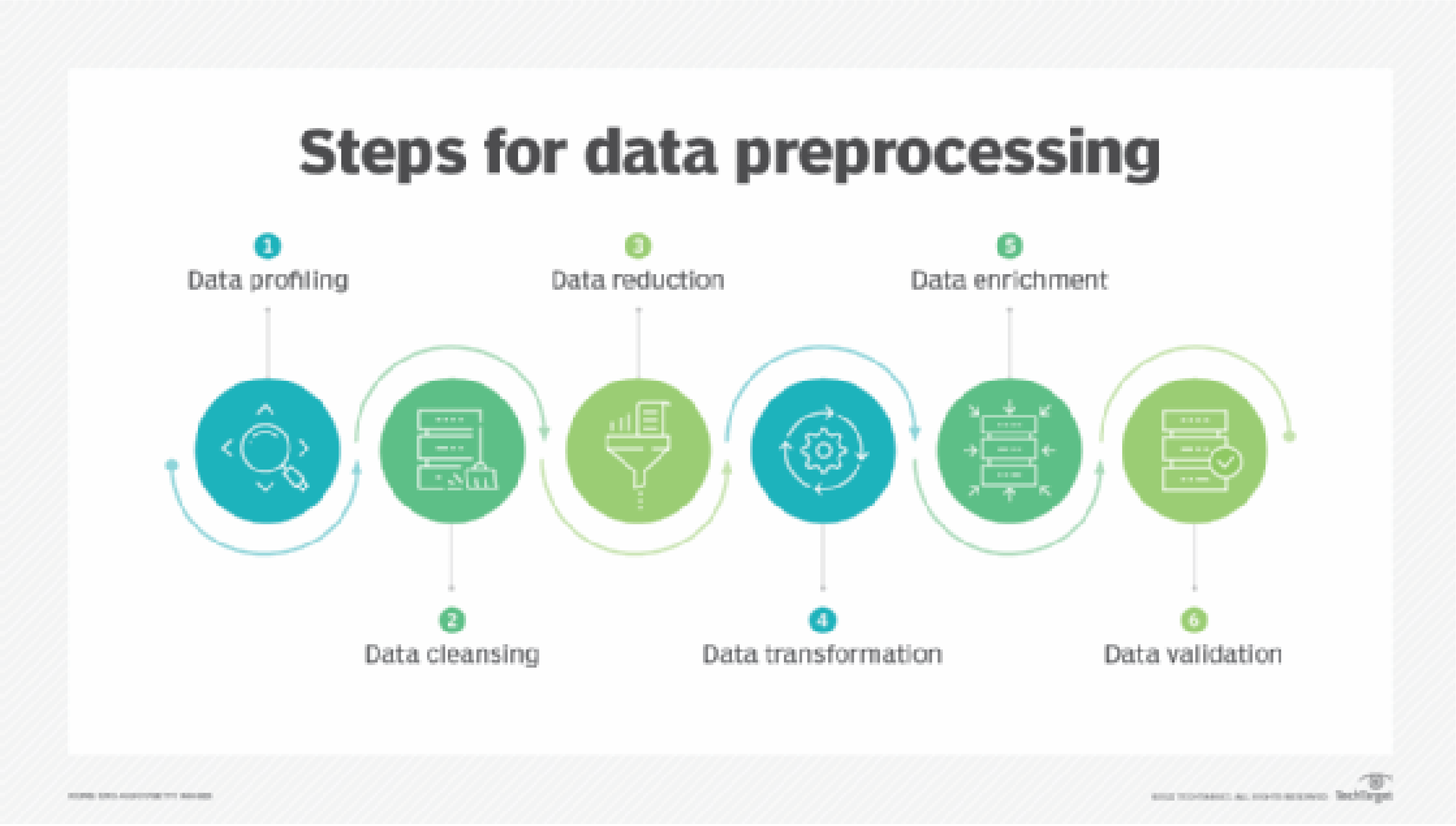
Figure 1.1: The data processing pipeline

Pre Processing – what?

Pre Processing – refers to cleaning, transforming, and organizing raw data into a structured and consistent format that is suitable for analysis and machine learning.

– It ensures the data is consistent, complete, and meaningful so that the model can learn effectively.

Key steps include handling missing values, smoothing noisy data, removing outliers and duplicates, normalizing data features, and transforming data into a uniform scale to improve the quality and accuracy of the downstream processes.



Steps for data preprocessing
1
Data profiling
2
Data cleansing
3
Data reduction
4
Data transformation
5
Data enrichment
6
Data validation

Pre Processing – why?

Imagine data collected to describe customer purchase data to predict customer buying behavior.

And your data is:

1. Some customers didn’t enter their age.
2. The gender column has entries like “Male”, “male”, and “M”.
3. A few purchase amounts are recorded as “-999” or “0”.
4. Some records are duplicated or missing essential fields.
5. Product categories vary in format — “Electronics”, “electronics”, “ELECTRONIC”.

What do you think about the data?

Pre Processing – why?

Imagine data collected to describe customer purchase data to predict customer buying behavior.

And your data is:

Missing values

1. Some customers didn't enter their age.
2. The gender column has entries like "Male", "male", and "M". Inconsistent entries
3. A few purchase amounts are recorded as "-999" or "0". Misleading entries
4. Some records are duplicated or missing essential fields. Redundant information
5. Product categories vary in format — "Electronics", "electronics", "ELECTRONIC".

Different data types

What do you think about the data?

Pre- Processing and Data preparation

1. Feature Extraction

- Extracting the right features is often a skill that requires an understanding of the specific application domain at hand.

2. Data cleaning

- The extracted data may have erroneous or missing entries.

3. Feature Selection and Transformation

- A variety of methods are used to either remove irrelevant features or transform the current set of features to a new data space that is more amenable for analysis

Pre- Processing and Data preparation

1. Feature Extraction

- Extracting the right features is often a skill that requires an understanding of the specific application domain at hand.

Pre- Processing and Data preparation

2. Data cleaning

- The extracted data may have erroneous or missing entries.

Pre- Processing and Data preparation

2. Data cleaning

- Missing data
- Incorrect entries
- Scaling and normalization

Pre- Processing and Data preparation

2. Data cleaning

- Missing data
  - Are this missing values informative?
    - Yes. So, use it.
    - No.
      1. Eliminate the entry.
      2. Estimate the value, by find the model that describes the data.
      3. Use robust data mining methods that deal with missing values.

When working with time series, that has a temporal dependency, estimation is the usual process.

Pre- Processing and Data preparation

2. Data cleaning

- Incorrect entries
  - Inconsistency detection
    - If data is available from diferente sources, identify inconsistencies. (e.g. HR information from ECG and pletismography)
  - Domain knowledge
    - If data has requisites, evaluate them, e.g.: negative blood pressure value.
  - Outlier detection
    - Boxplot
    - Data distribution
    - Clustering
    - Distance base methods
    - Information theory models

Pre- Processing and Data preparation

2. Data cleaning

- Scaling and normalization
  - Data may present different scales, and therefore it is not comparable.
  - Also, some methods present inconsistencies when features have different scales. (e.g. distance base methods, results will be biased to the attribute with higher magnitude)
- Solution
  - Standardization
  - Min-Max scaler

$$y_i = \frac{x_i^j - \mu_j}{\sigma_j}$$

$$y_i = \frac{x_i^j - min_j}{max_j - min_j}$$

- Disadvantage
  - The physical meaning is lost.
  - When data present outliers, careful is needed in the application of min-max scaler.

Pre- Processing and Data preparation

3. Feature Selection and Transformation

Pre- Processing and Data preparation

3. Feature Selection and Transformation

- Usually implemented in the pre-processing step.
- Some analytical models, include the feature selection or a strategy to deal with high dimensional data.

- Goal:
  - Reduce the size of the data through:
    - feature subset selection;
    - data transformation.

- Consequence:
  - When the size of the data is reduced, the algorithms are generally more efficient.
  - Elimination of irrelevant features or irrelevant records improve the quality of the data mining process.

Pre- Processing and Data preparation

3. Feature Selection and Transformation

- Data sampling
  - The records from the underlying data are sampled to create a much smaller database.
- Feature selection
  - A subset of features is selected and used on the analytical process.
  - It depends on the problem and scientific question.
- Data reduction with axis rotation
  - Information between features (correlation, variance) is used to build a projection of the data in a new data space with smaller dimension.
- Data reduction with type transformation
  - Convert the data type to reduce its dimension.

Pre- Processing and Data preparation

3. Feature Selection and Transformation

- Data sampling
  - The main advantage of sampling is that it is simple, intuitive, and relatively easy to implement.
  - Select some data points from the entire sample.

- **Biased sampling:** If it is known that some parts of the data have more importance than others, sampling may intentionally select those subsets.

- **Stratified sampling:** When rare events are present on data, traditional sampling may neglect them. So sampling by strata, independently selecting samples based on a predefined proportion may solve that problem.

Pre- Processing and Data preparation

3. Feature Selection and Transformation

- Feature selection
  - Feature selection is primarily focused on removing non-informative or redundant predictors from the model.

Low information quantity, that may be evaluated for example by variance of the predictor.

Collinearity is the technical term for the situation where a pair of predictor variables have a substantial correlation with each other. It is also possible to have relationships between multiple predictors at once (called multicollinearity).

Pre- Processing and Data preparation

3. Feature Selection and Transformation

- **Wrapper methods:** Methods to find the optimal predictors combination to maximize model performance.
  - Computationally more heavy.
  - Risk of over-fitting.
- **Filter methods:** Evaluate the predictor relevance, and only the ones that pass a criteria will be included in the model.
  - Computationally more effective.
  - Selection criteria is not related to model effectiveness.
  - Most filter methods performs a univariable evaluation, missing predictors relations.
- **Unsupervised feature selection:** Removal of noisy or redundant predictors, by clustering.

Pre- Processing and Data preparation

3. Feature Selection and Transformation

- **Filter methods:**
  - Evaluate the ability of the features to significantly discriminate conditions. E.g.:
    - Hypothesis testing.

Pre- Processing and Data preparation

3. Feature Selection and Transformation

- **Feature Importance:**
  - Ranking the importance of predictors allows to achieve better modeling performances.

  - Numeric predictors:
    - correlation statistic is the classic approach to quantifying the predictors relationship with the outcome.
      - If the relation is linear – Pearson correlation.
      - If the relationship is nearly linear or curvilinear - Spearman's correlation coefficient.
    - Nonlinear relations:
      - Locally weighted regression model (known more commonly as LOESS). This technique is based on a series polynomial regressions that model the data in small neighborhoods (similar to computing a moving average). The approach can be effective at creating smooth regression trends that are extremely adaptive.

Pre- Processing and Data preparation

3. Feature Selection and Transformation

- **Feature Importance:**
  - Previous techniques evaluate each predictor without considering the others. This can be potentially misleading.
    - If two predictors are highly correlated with the response and with each other, then the univariate approach will identify both as important.
      - As we have seen, some models will be negatively impacted by including this redundant information.
      - Pre-processing approaches such as removing highly correlated predictors can alleviate this problem.
    - The univariate importance approach will fail to identify groups of predictors that together have a strong relationship with the response.

Best Practices & Takeaways

- Always explore your data before analysis
- Preprocessing is often the most time-consuming step
- Document decisions (e.g., how you handled missing values)
- No “one-size-fits-all” solution