fit@hcmus

University of Science – VNU-HCM
Faculty of Information Technology
CSC15105 – Text Mining and Application

Slot 01 - Introduction to Text Mining

Presenter:

Dr. LE Thanh Tung

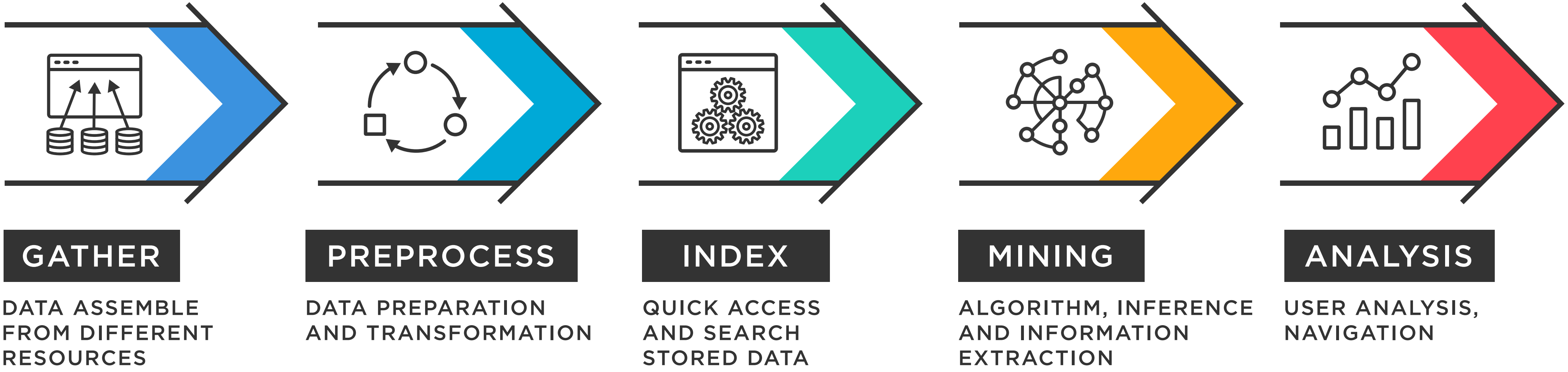
Content

1. Text Mining
2. Machine Learning
3. Evaluation Metrics
4. Programming Language

Introduction

- Text mining is the process of transforming unstructured **text** into a structured format to identify meaningful patterns and new insights

TEXT MINING INVOLVES A SERIES OF ACTIVITIES TO BE PERFORMED IN ORDER TO EFFICIENTLY MINE THE INFORMATION. THESE ACTIVITIES ARE:



GATHER

DATA ASSEMBLE FROM DIFFERENT RESOURCES

PREPROCESS

DATA PREPARATION AND TRANSFORMATION

INDEX

QUICK ACCESS AND SEARCH STORED DATA

MINING

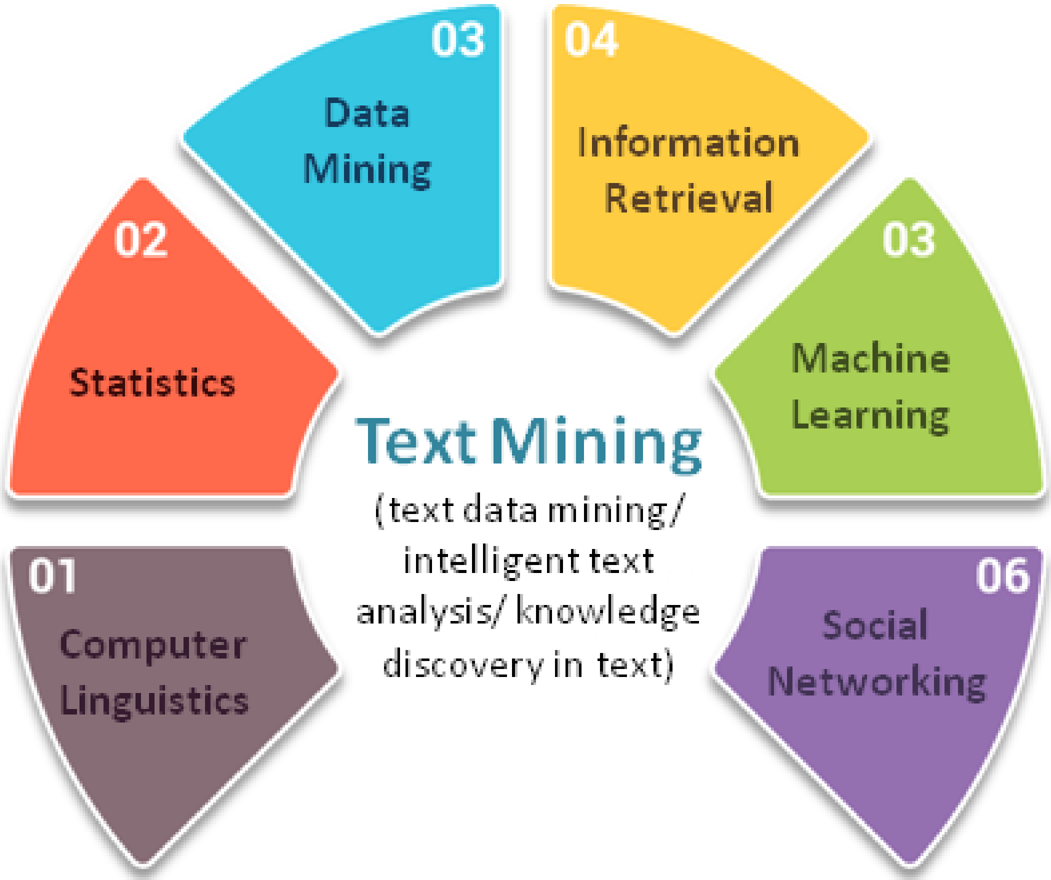
ALGORITHM, INFERENCE AND INFORMATION EXTRACTION

ANALYSIS

USER ANALYSIS, NAVIGATION

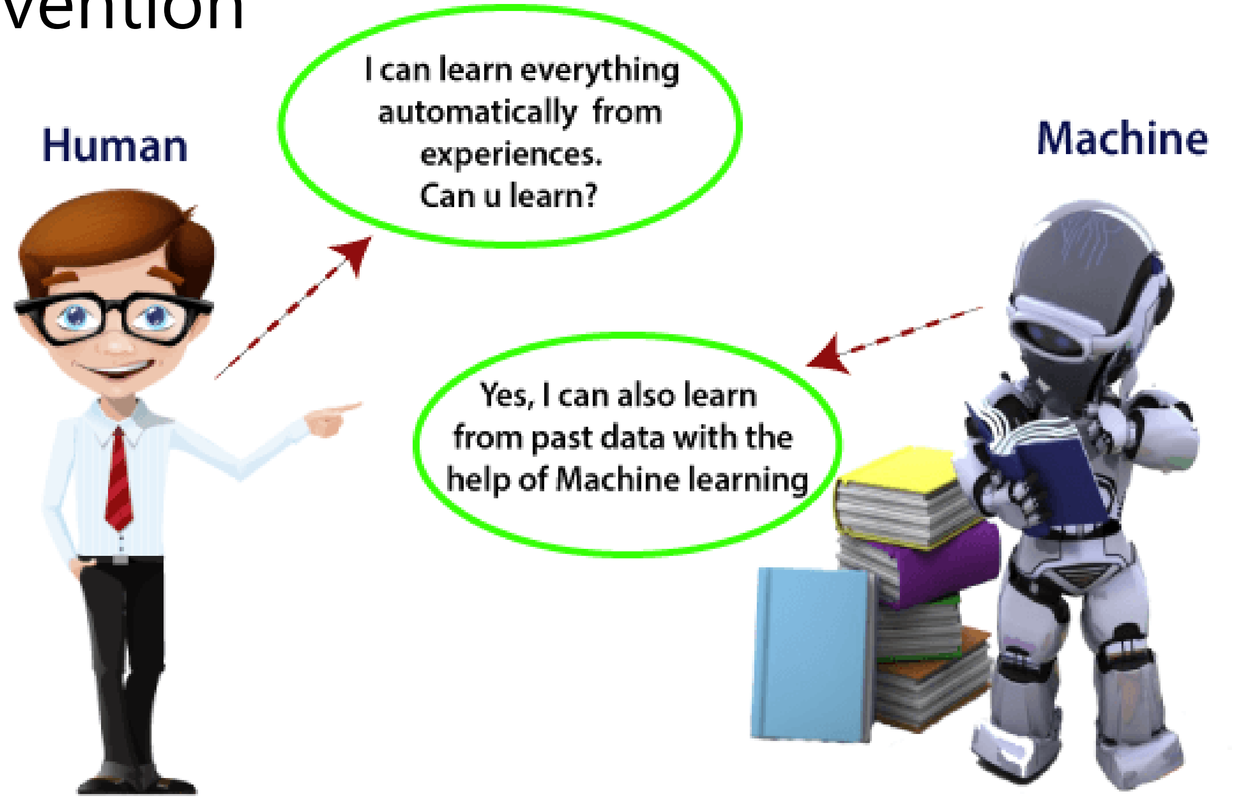
Introduction

- Gupta & Lehal (2009) have regarded text mining as new interdisciplinary area which is an amalgamation of data mining, information retrieval, machine learning, computer linguistic and statistics



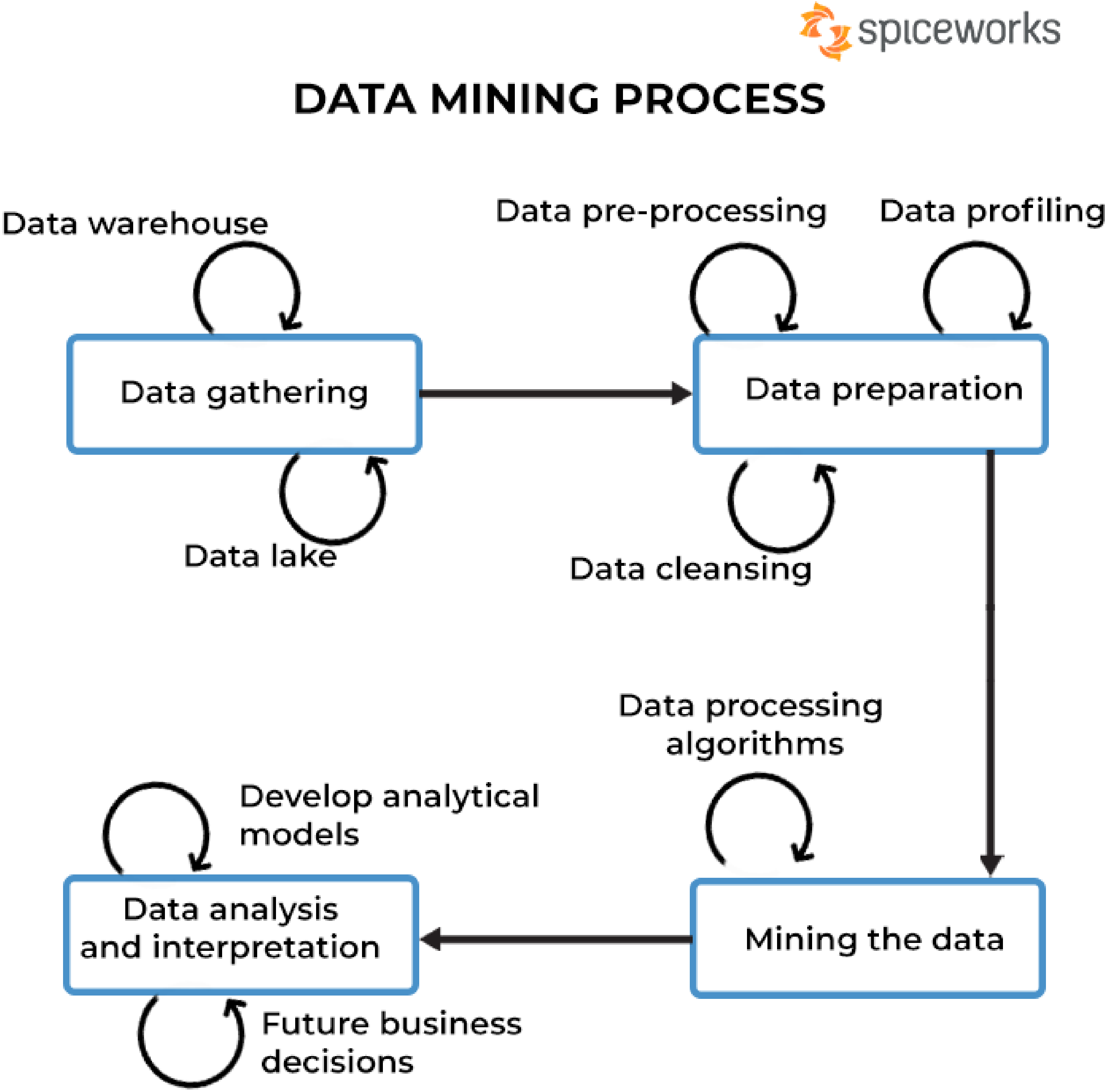
Introduction

- Machine Learning is a discipline of artificial intelligence (AI) that provides machines with the ability to automatically **learn from data and past experiences** while identifying patterns to make predictions with minimal human intervention


Human
I can learn everything automatically from experiences. Can u learn?
Machine
Yes, I can also learn from past data with the help of Machine learning

Introduction

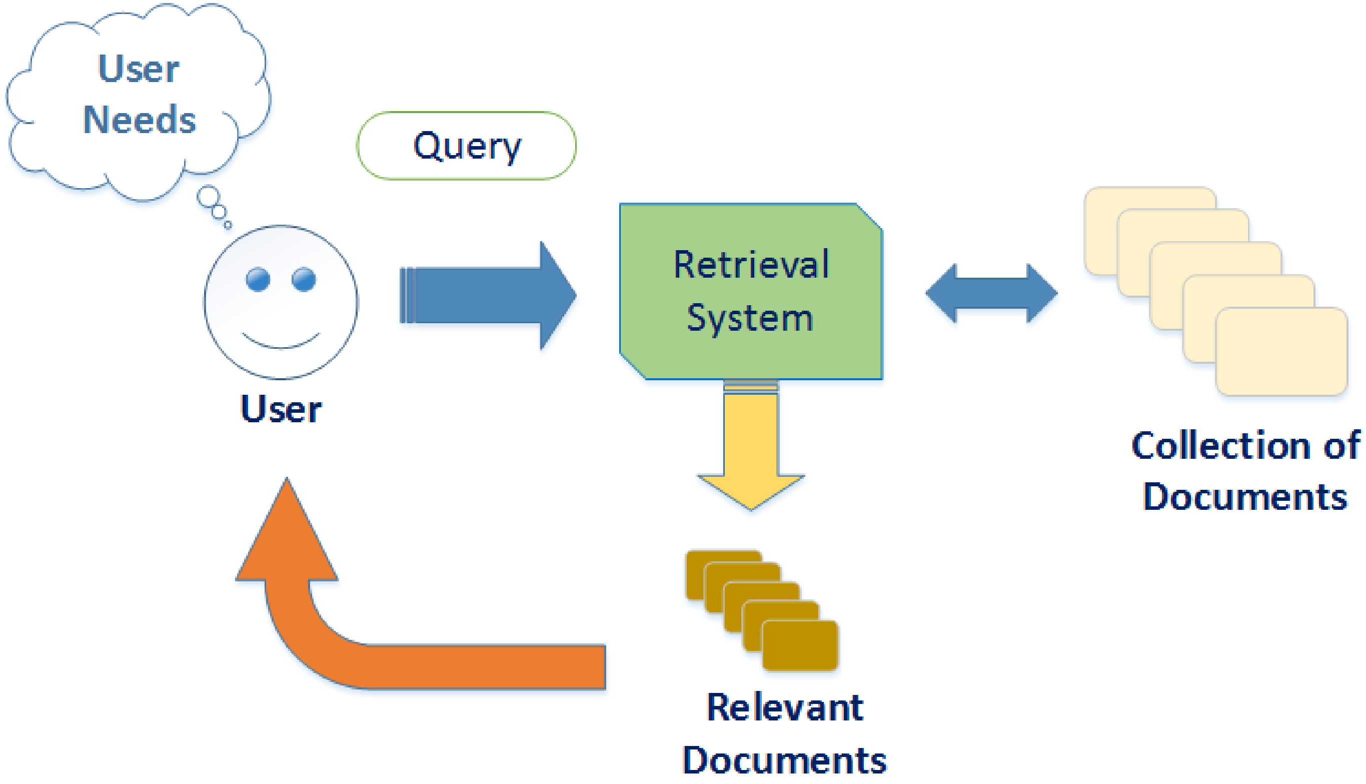
- Data mining is the process of analyzing a large batch of information to **discern trends and patterns**

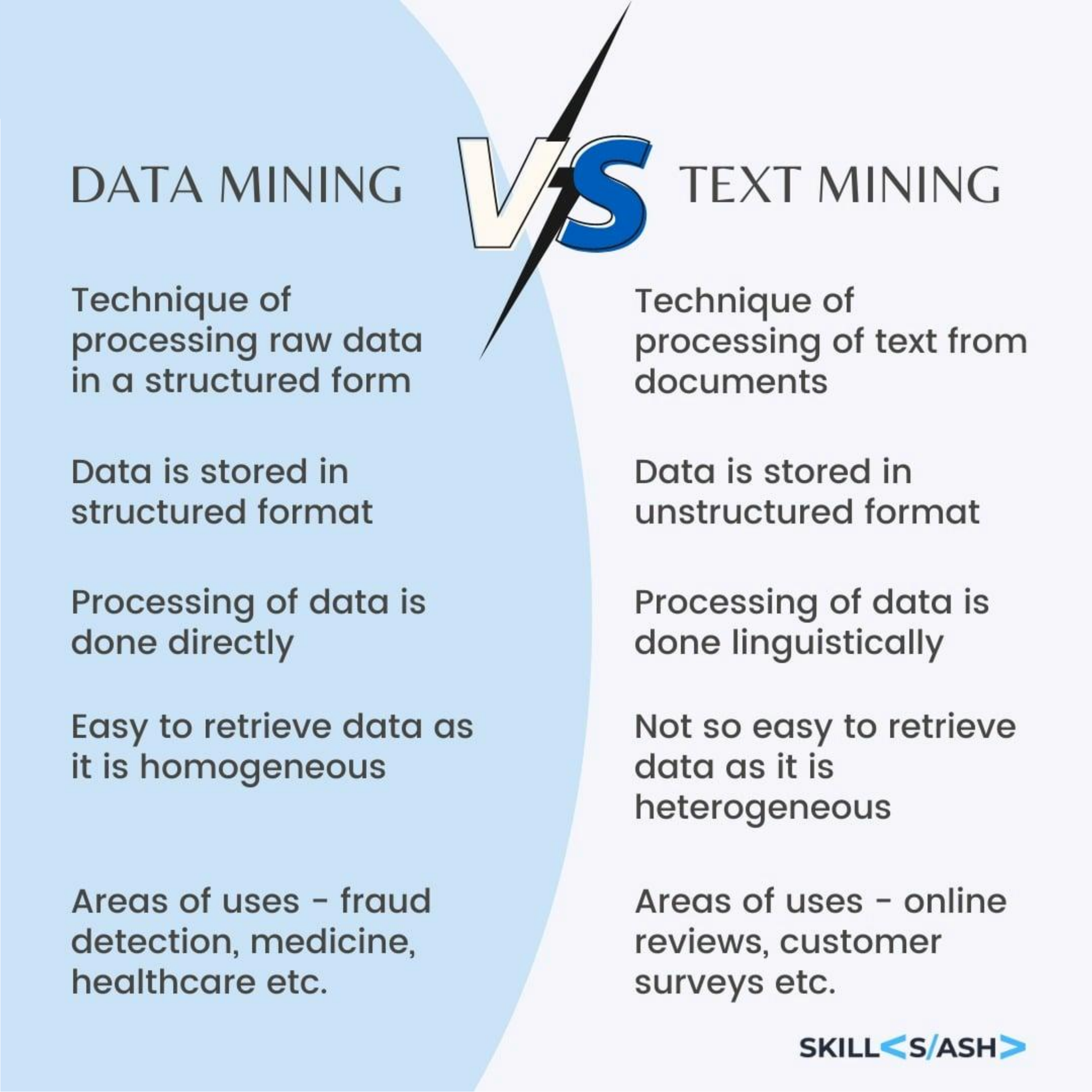


Introduction

- Information Retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)

- Two main approaches are matching words in the query against the database index (keyword searching) and traversing the database using hypertext or hypermedia links

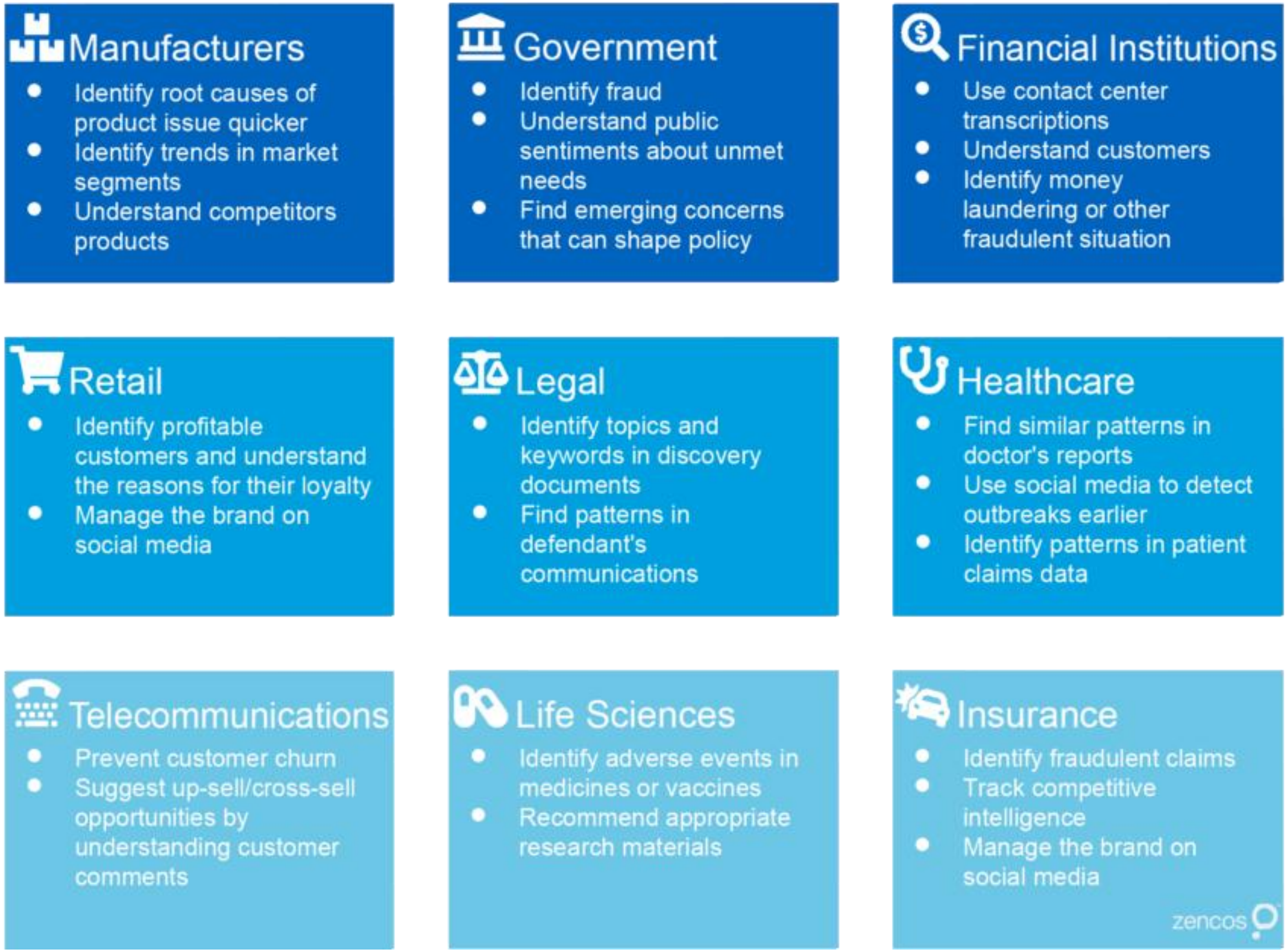




DATA MINING	TEXT MINING
Technique of processing raw data in a structured form	Technique of processing of text from documents
Data is stored in structured format	Data is stored in unstructured format
Processing of data is done directly	Processing of data is done linguistically
Easy to retrieve data as it is homogeneous	Not so easy to retrieve data as it is heterogeneous
Areas of uses - fraud detection, medicine, healthcare etc.	Areas of uses - online reviews, customer surveys etc.

SKILL<S/ASH>

Text Analytics Application



Manufacturers
- Identify root causes of product issue quicker
- Identify trends in market segments
- Understand competitors products

Government
- Identify fraud
- Understand public sentiments about unmet needs
- Find emerging concerns that can shape policy

Financial Institutions
- Use contact center transcriptions
- Understand customers
- Identify money laundering or other fraudulent situation

Retail
- Identify profitable customers and understand the reasons for their loyalty
- Manage the brand on social media

Legal
- Identify topics and keywords in discovery documents
- Find patterns in defendant's communications

Healthcare
- Find similar patterns in doctor's reports
- Use social media to detect outbreaks earlier
- Identify patterns in patient claims data

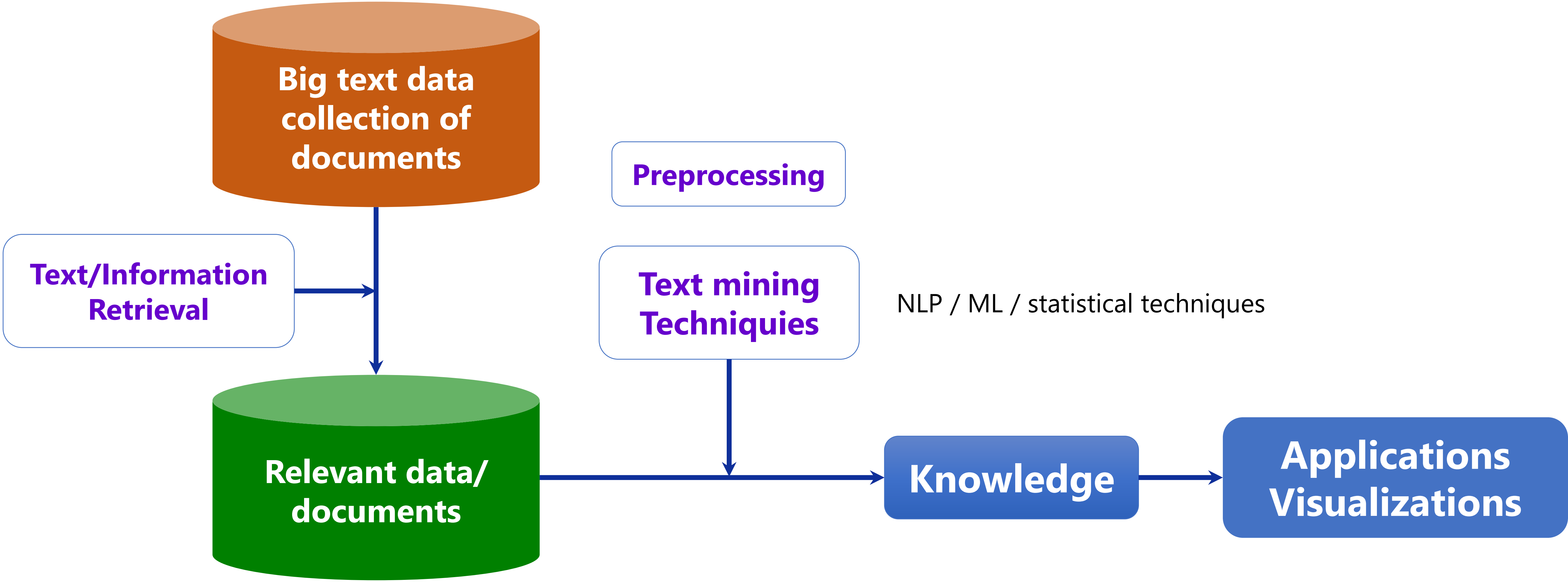
Telecommunications
- Prevent customer churn
- Suggest up-sell/cross-sell opportunities by understanding customer comments

Life Sciences
- Identify adverse events in medicines or vaccines
- Recommend appropriate research materials

Insurance
- Identify fraudulent claims
- Track competitive intelligence
- Manage the brand on social media

zencos

General Architecture



Big text data collection of documents

Text/Information Retrieval

Relevant data/ documents

Preprocessing

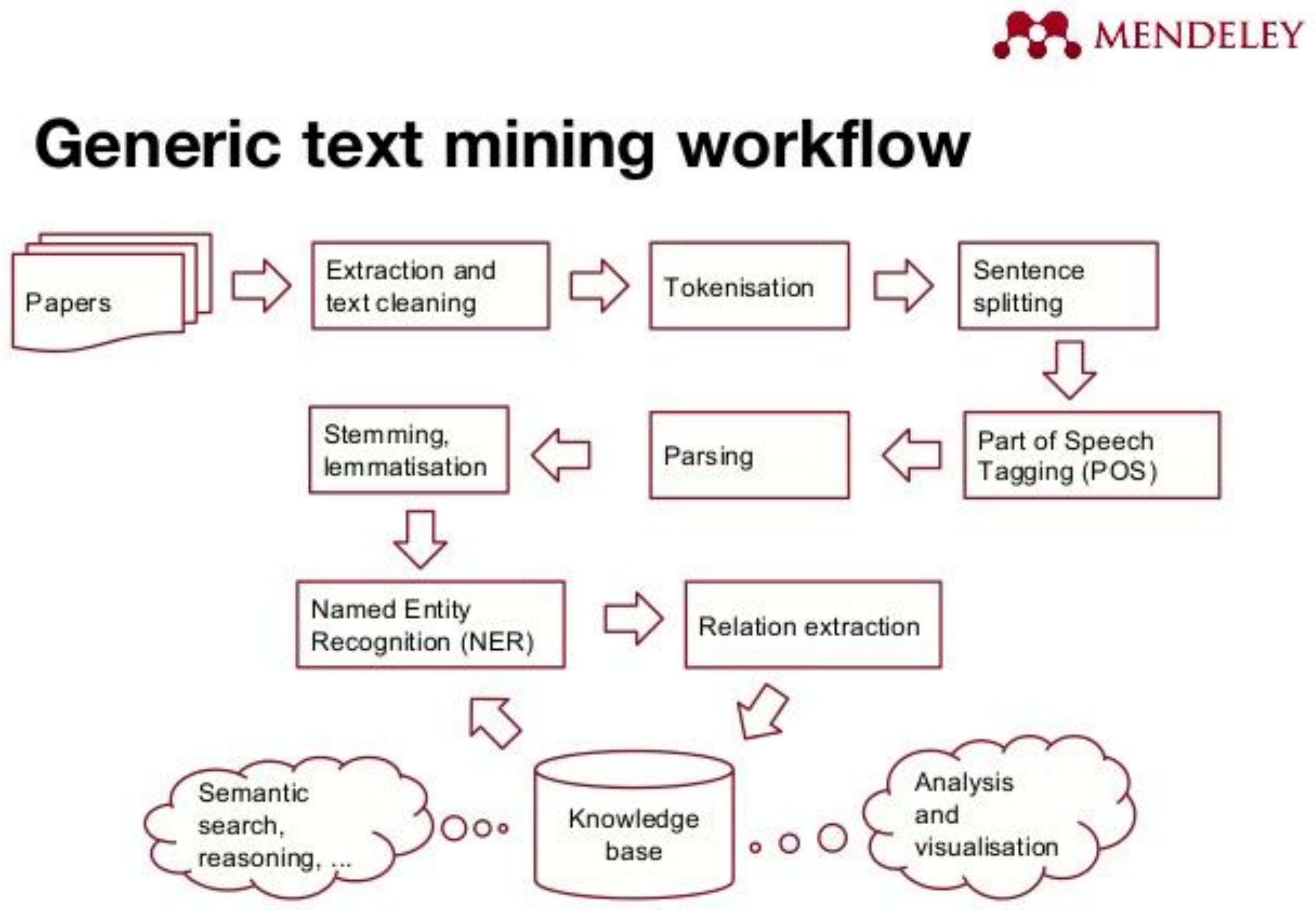
Text mining Techniquies

NLP / ML / statistical techniques

Knowledge

Applications Visualizations

General Architecture



MENDELEY

Generic text mining workflow

Papers

Extraction and text cleaning

Tokenisation

Sentence splitting

Part of Speech Tagging (POS)

Parsing

Stemming, lemmatisation

Named Entity Recognition (NER)

Relation extraction

Knowledge base

Semantic search, reasoning, ...

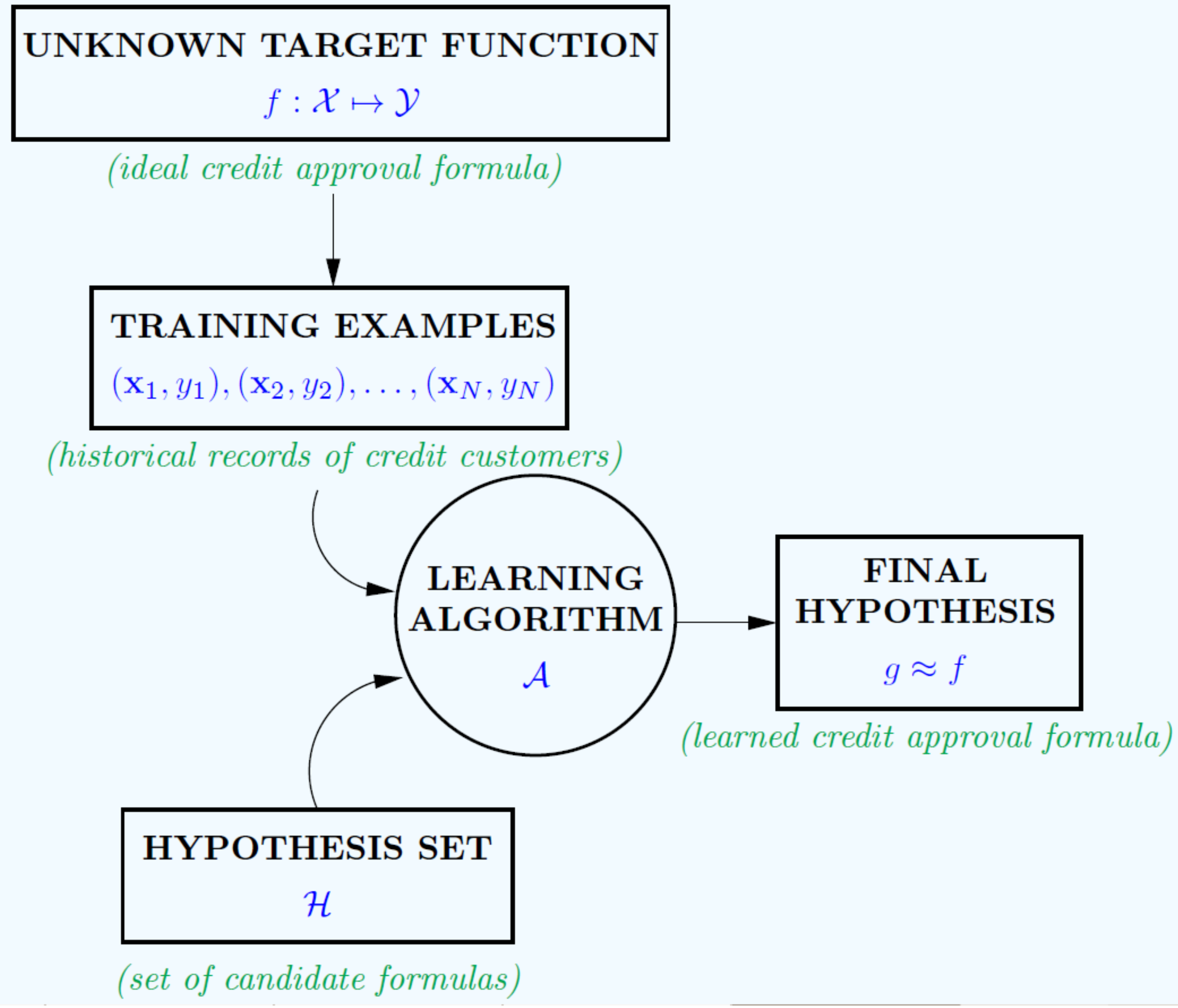
Analysis and visualisation

Introduction

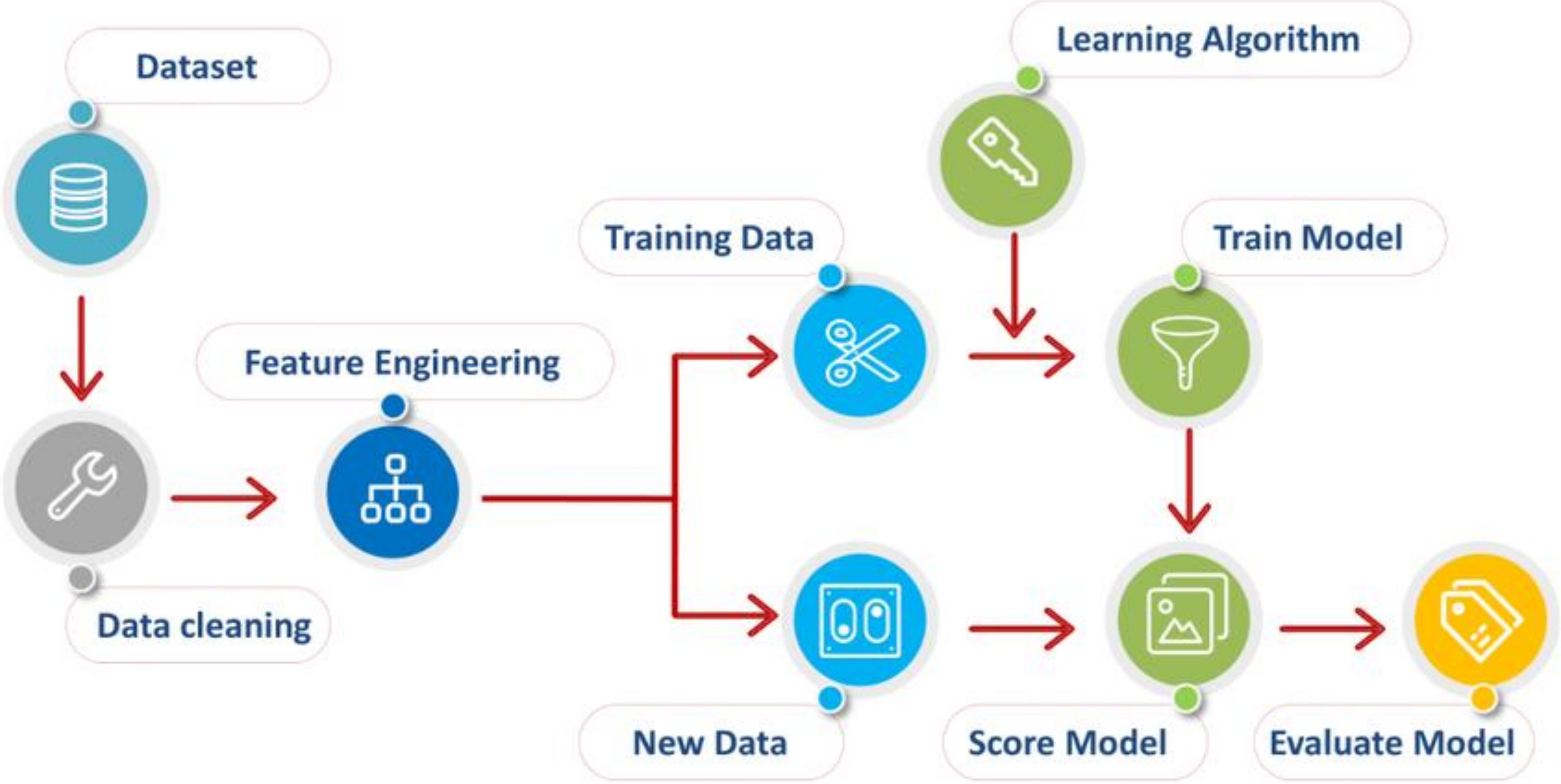
- Machine learning tasks
    - Build a **model** from some **data**
        - choose how to map raw data to feature vectors
        - choose a model form
        - choose parameter values in the model
    - **Test** or **validate** the model
        - Evaluate the model on unseen data to assess its performance

Learning Algorithms

- The data is formed by a specific distribution.
- The goal of learning is to find the data distribution which is generalized in this kind of data

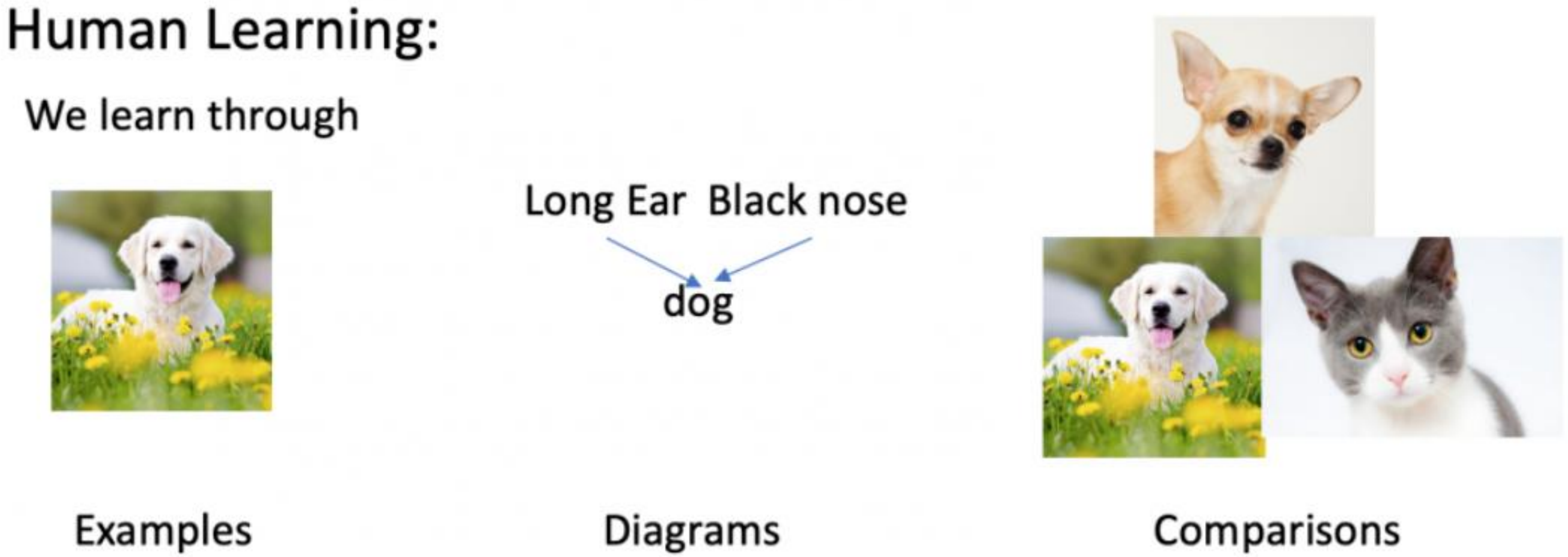

UNKNOWN TARGET FUNCTION
$f : \mathcal{X} \mapsto \mathcal{Y}$
(ideal credit approval formula)
TRAINING EXAMPLES
$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$
(historical records of credit customers)
LEARNING ALGORITHM
$\mathcal{A}$
FINAL HYPOTHESIS
$g \approx f$
(learned credit approval formula)
HYPOTHESIS SET
$\mathcal{H}$
(set of candidate formulas)

Introduction



Introduction

- Feature
    - A measurable property or characteristic of phenomenon being observed.

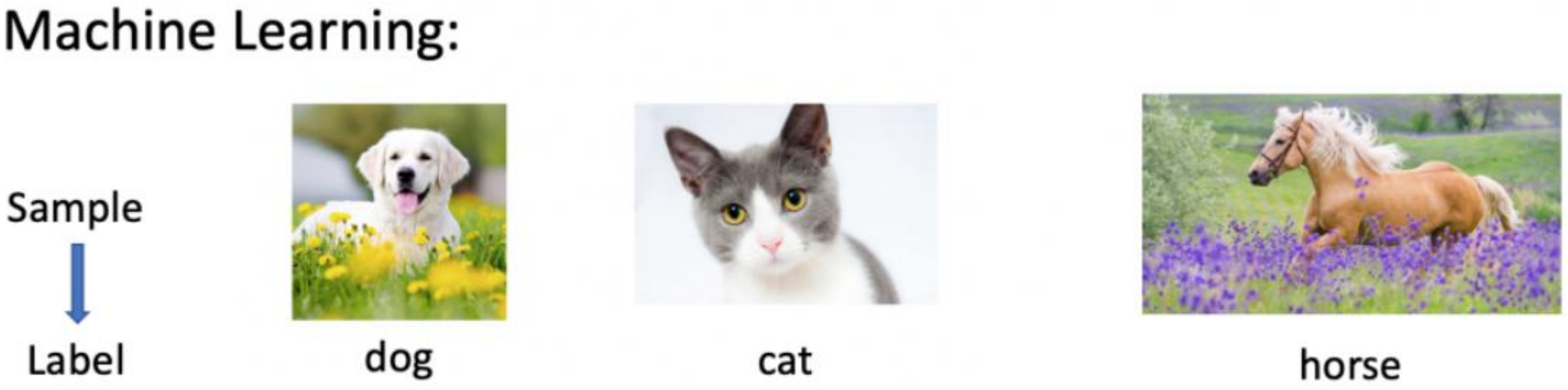


Human Learning:

We learn through

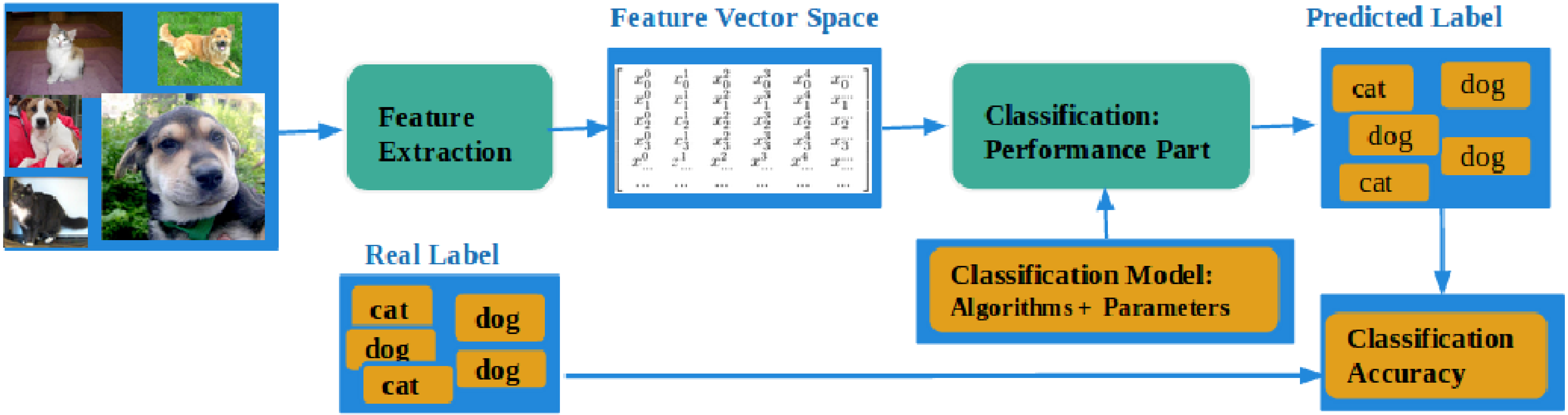
Introduction

- Feature
  - A measurable property or characteristic of phenomenon being observed.



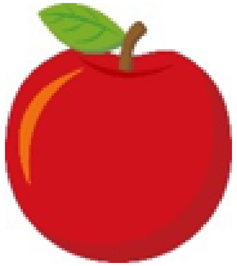
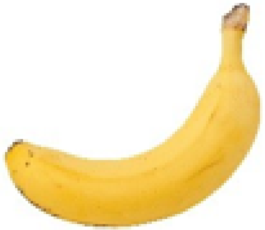
Introduction

- Feature
  - A measurable property or characteristic of phenomenon being observed.



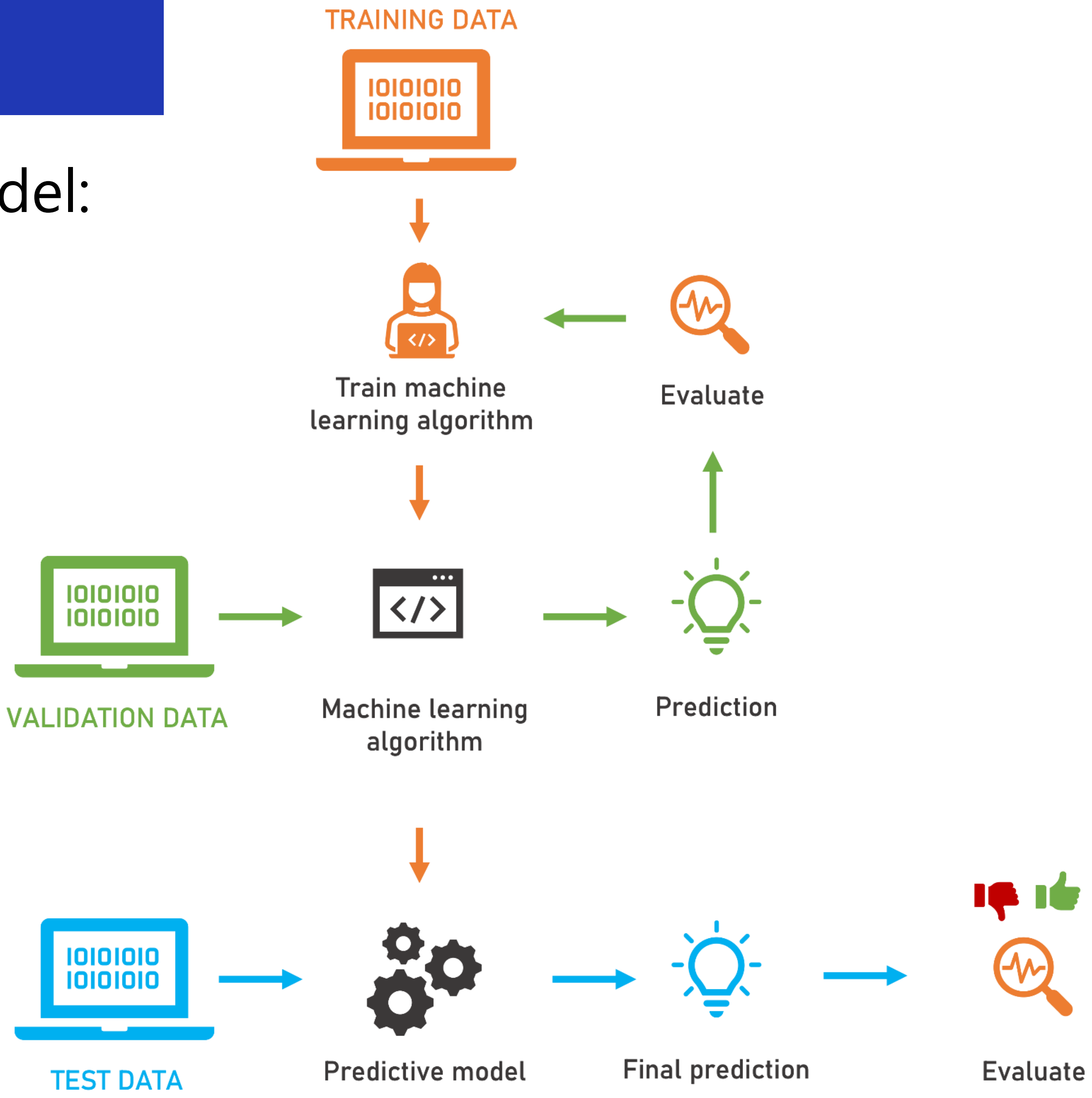
Introduction

- Feature
  - A measurable property or characteristic of phenomenon being observed.

	Color	Weight	my Rank	…
	Red	200g	1st	
	Yellow	300g	3rd	

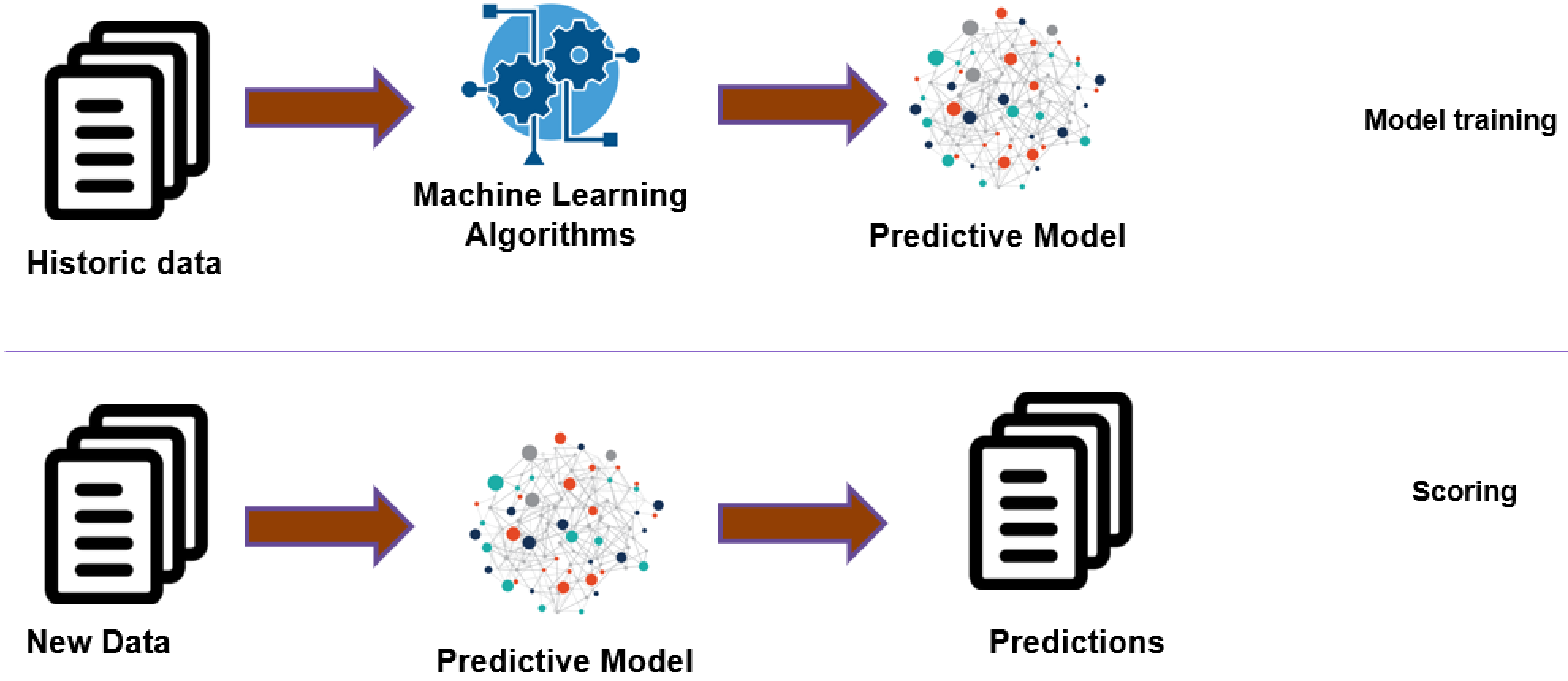
Predictive Tasks

- Build a Machine Learning model:
  - Training
  - Evaluating
  - Testing

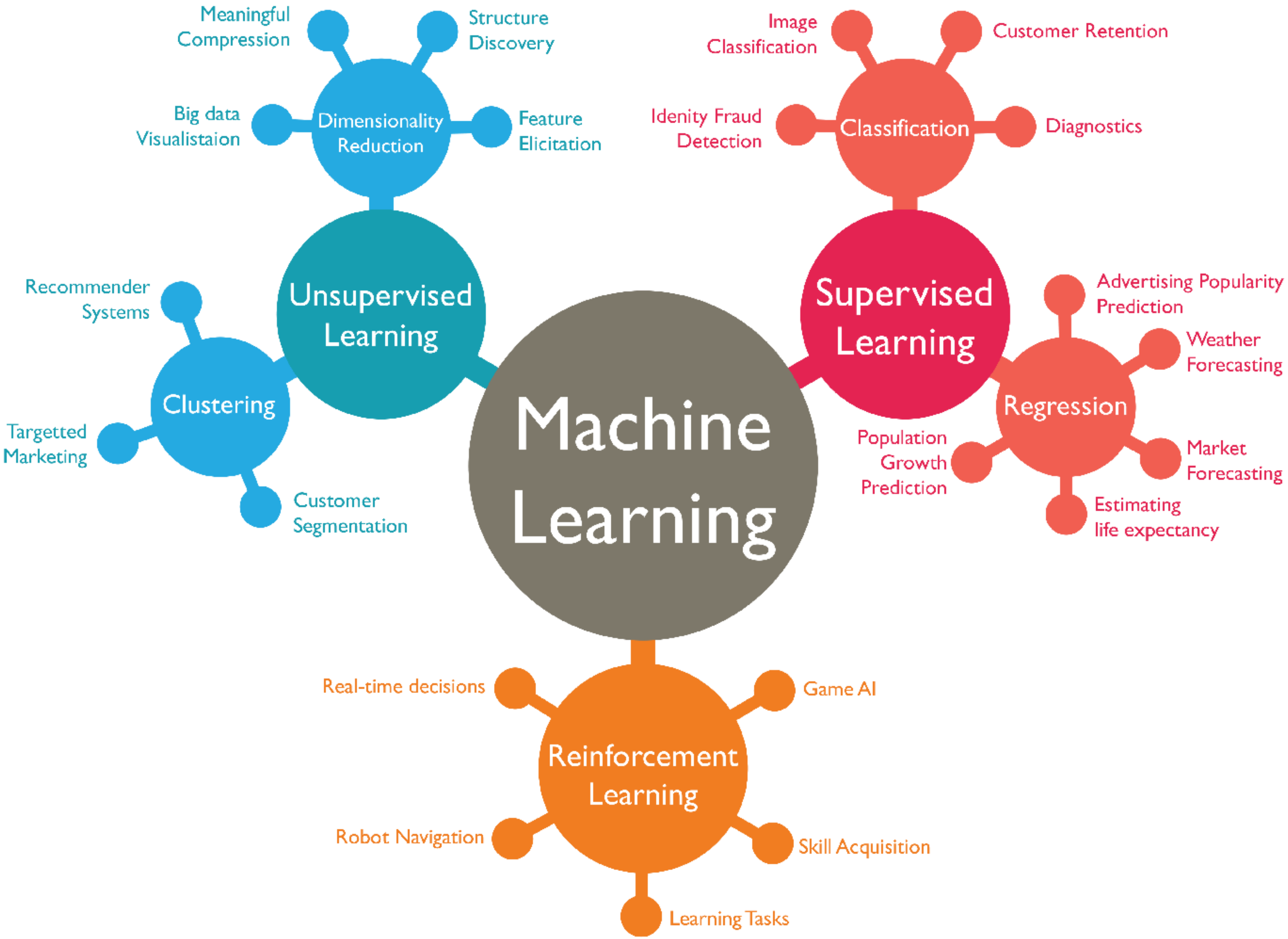


Learning Algorithms

- The goal of learning is to find the hidden pattern/distribution from historic data



Type of Machine Learning



Type of Machine Learning

Supervised

Input:

- Features
- Gold labels

Tasks: task-oriented

- Classification
- Regression

Unsupervised

Input:

- Features

Tasks: data-oriented

- Clustering
- Dimension Reduction

Reinforcement

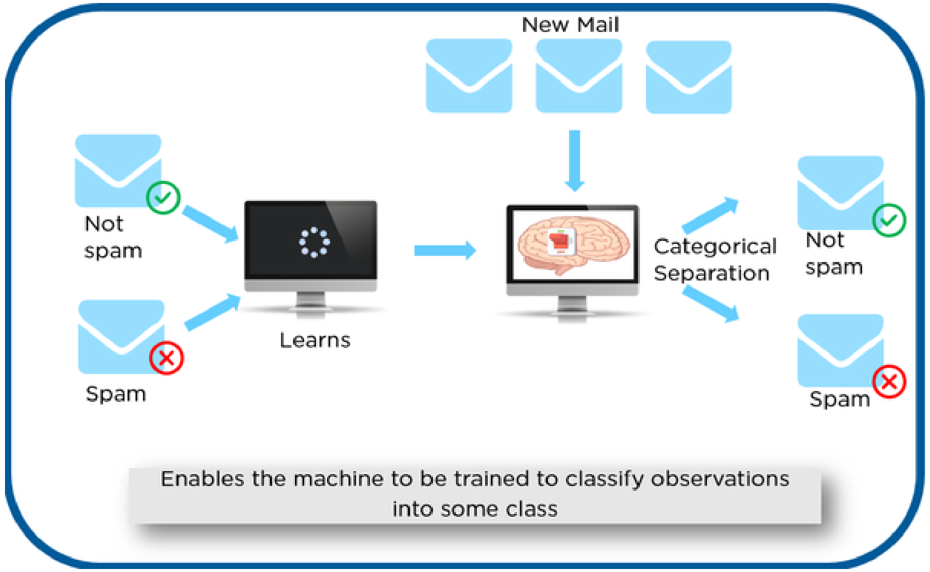
Input:

- Features
- Feed-back function

Tasks: environment

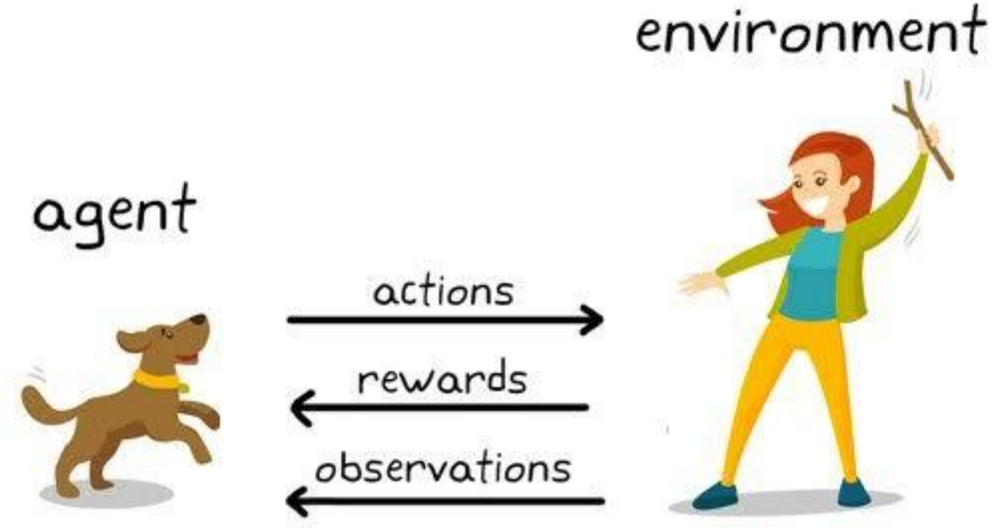
- Real-time Decision
- Learning tasks

Type of Machine Learning

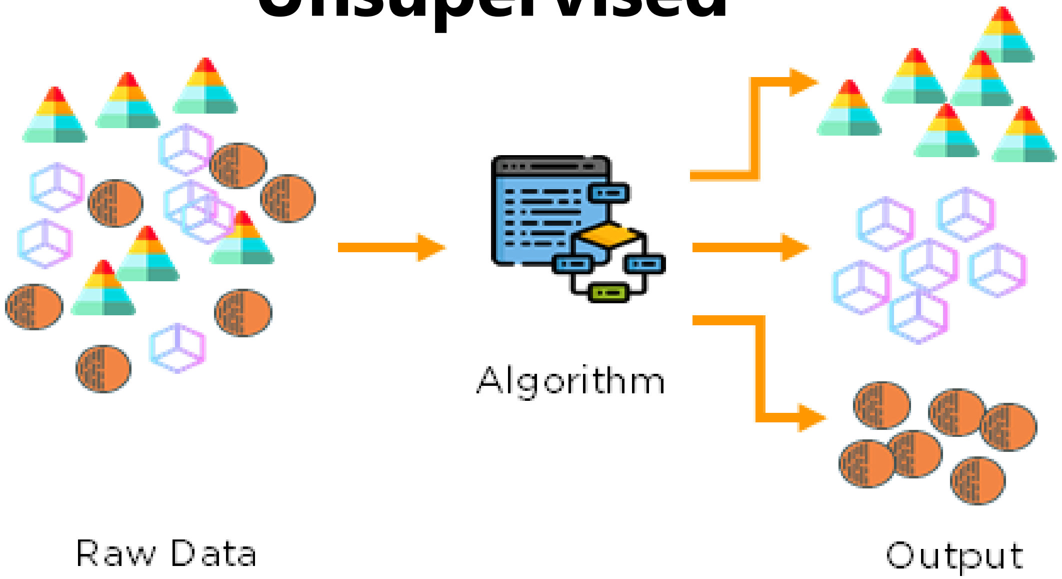


Enables the machine to be trained to classify observations into some class

Supervised



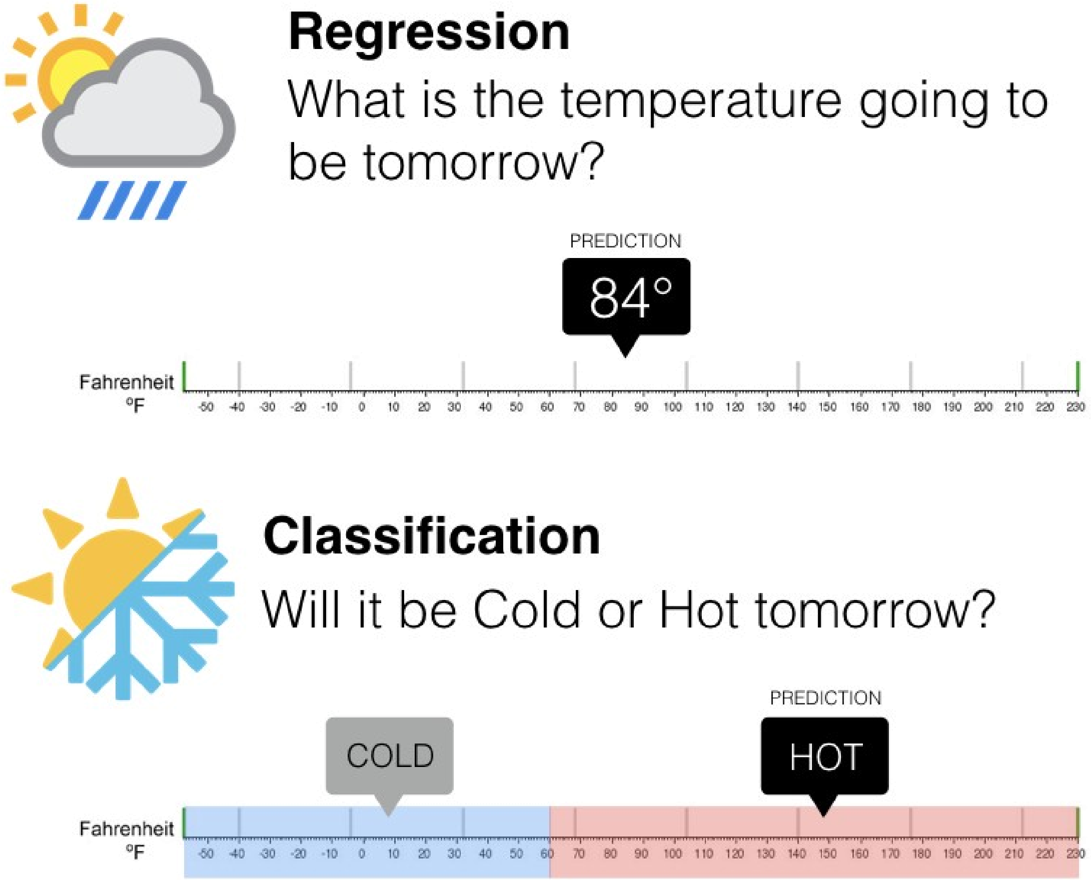
Unsupervised



Reinforcement

Supervised learning

- Learn from **labeled** data
  - Classification methods: predict a label
  - Regression methods: predict a quantity



Regression

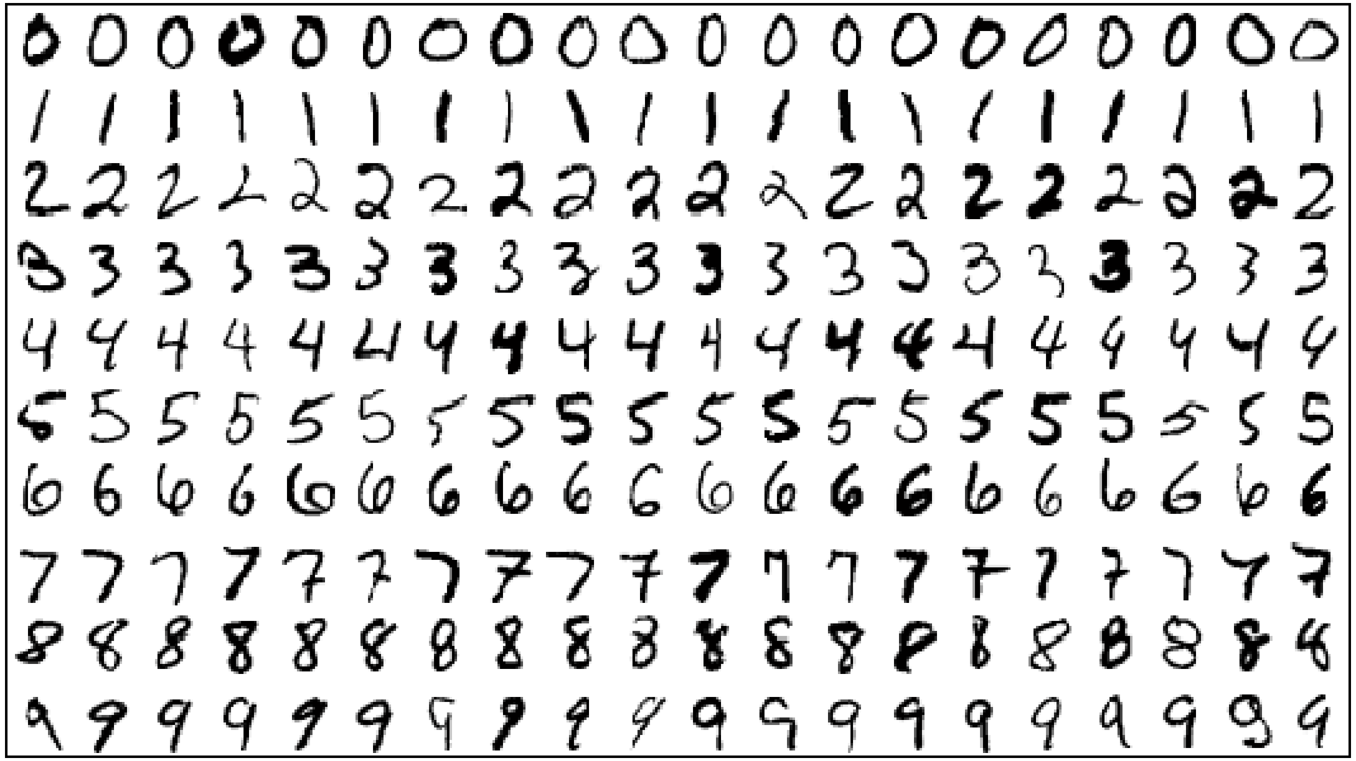
What is the temperature going to be tomorrow?

Classification

Will it be Cold or Hot tomorrow?

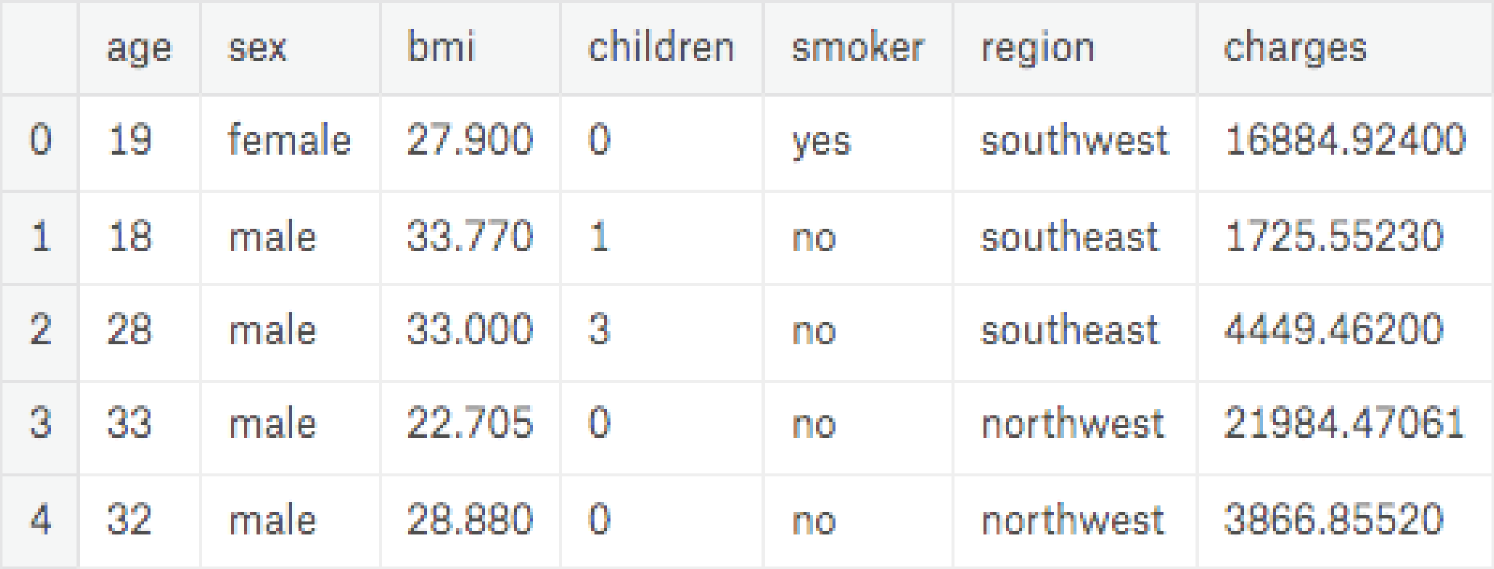
Supervised learning: Example

- Handwriting recognition:
  - Data: a set of handwriting images and their labels.
  - Goal: build a model to predict a digit for a given image.



Supervised learning: Example

- Medical Costs Analysis:
  - Data: containing medical records
  - Goal: Build a model to predict medical costs for a given individual based on their demographic and health-related information

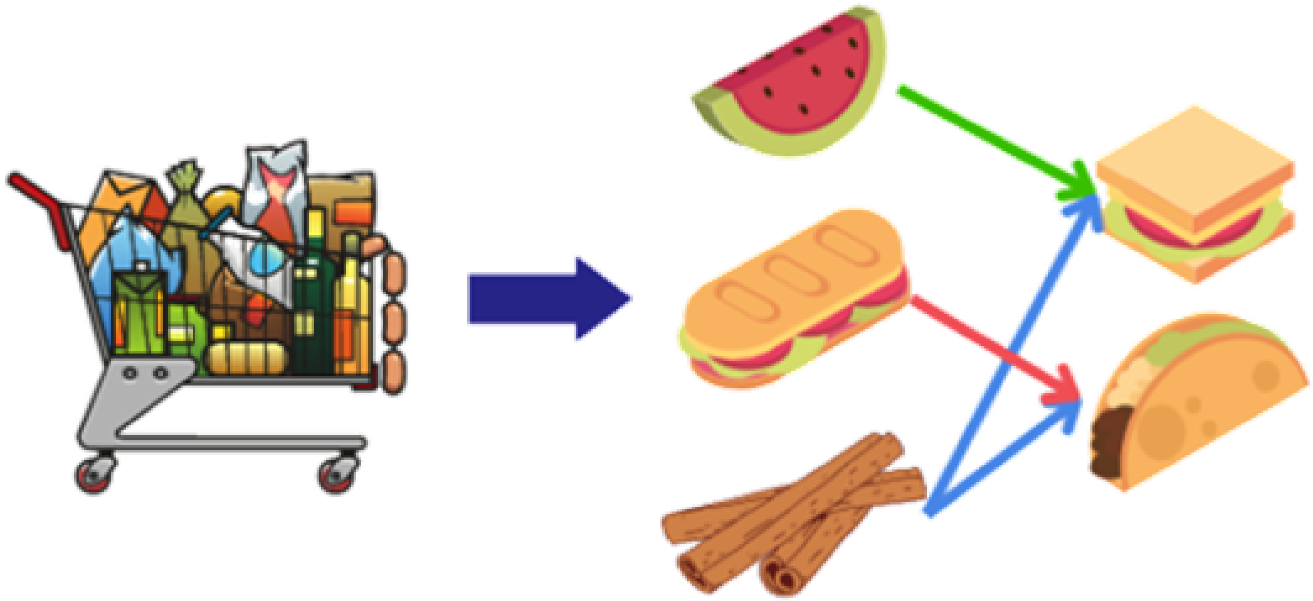


	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Unsupervised learning

- Learn from unlabeled data
  - Dimension reduction
  - Clustering
  - Association

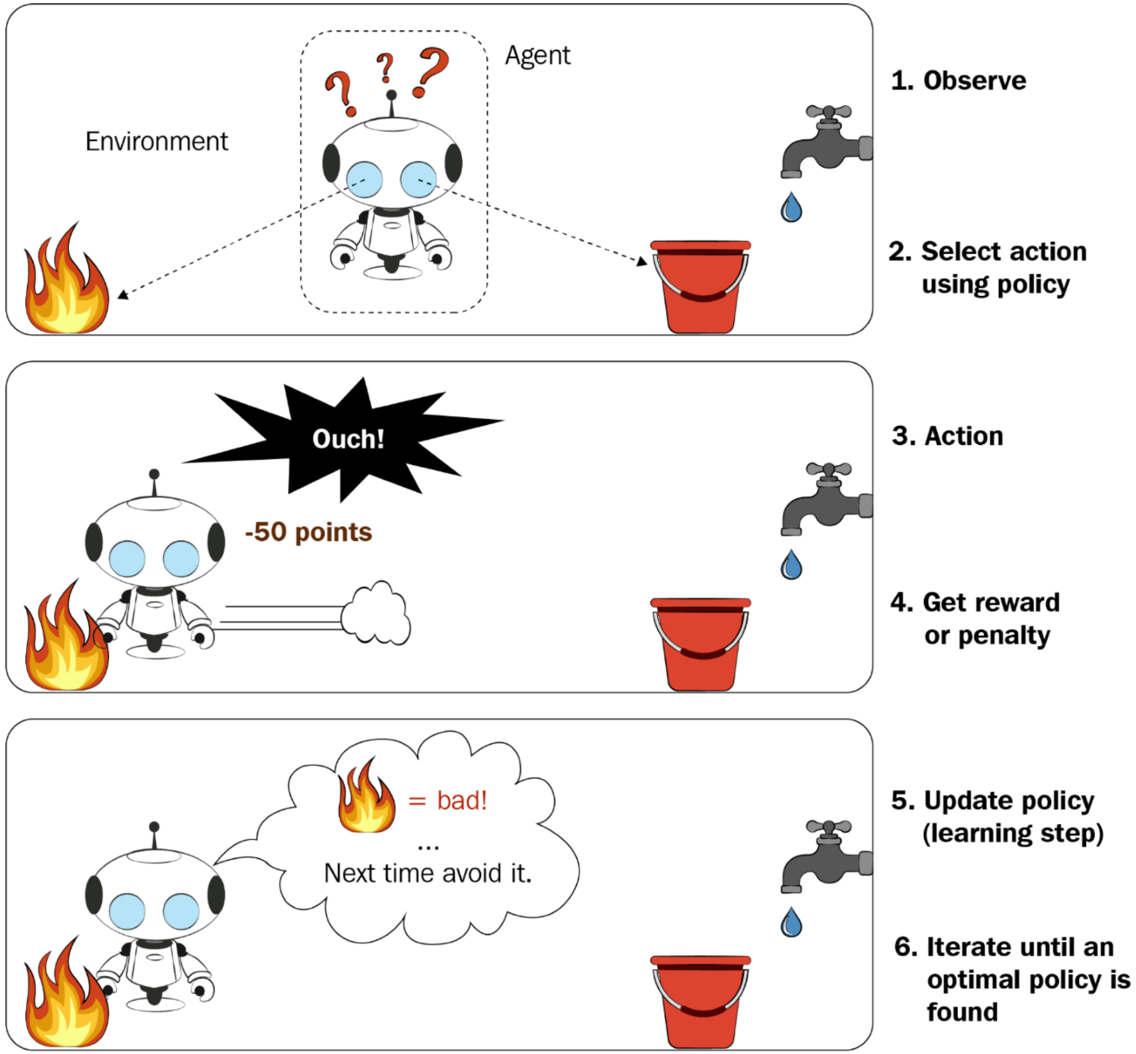
Association Rule Learning



"93% of people who purchased item A also purchased item B"

Reinforcement Learning

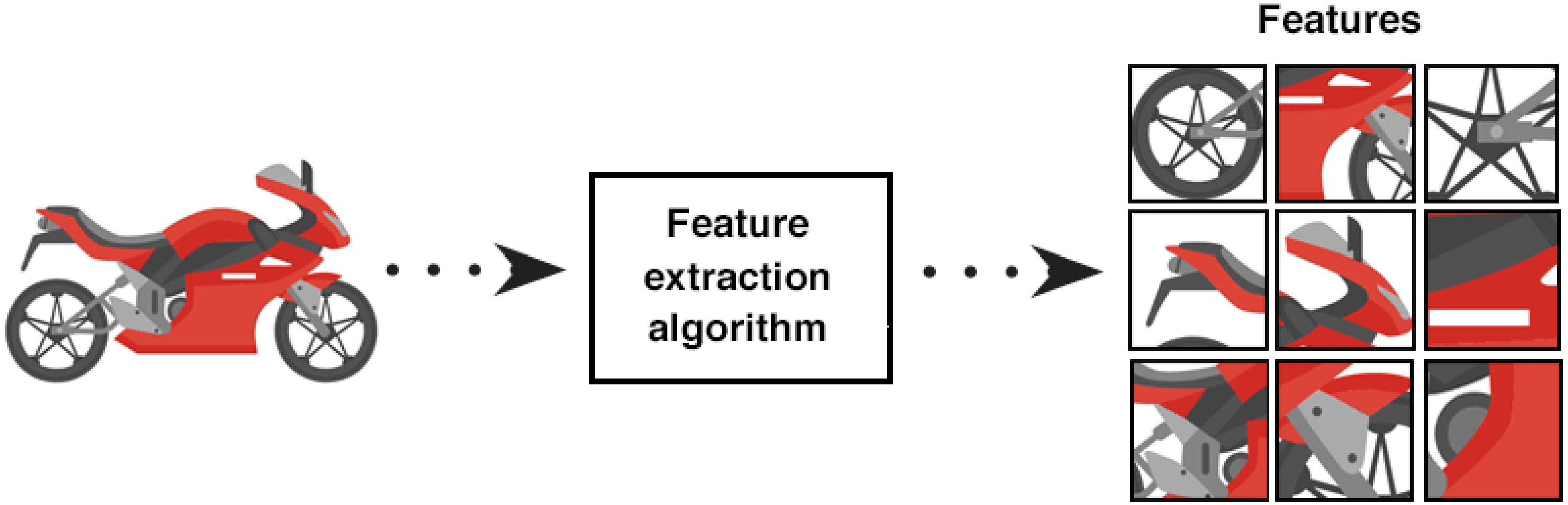
- Learn from mistakes



1. **Observe**
2. **Select action using policy**
3. **Action**
4. **Get reward or penalty**
5. **Update policy (learning step)**
6. **Iterate until an optimal policy is found**

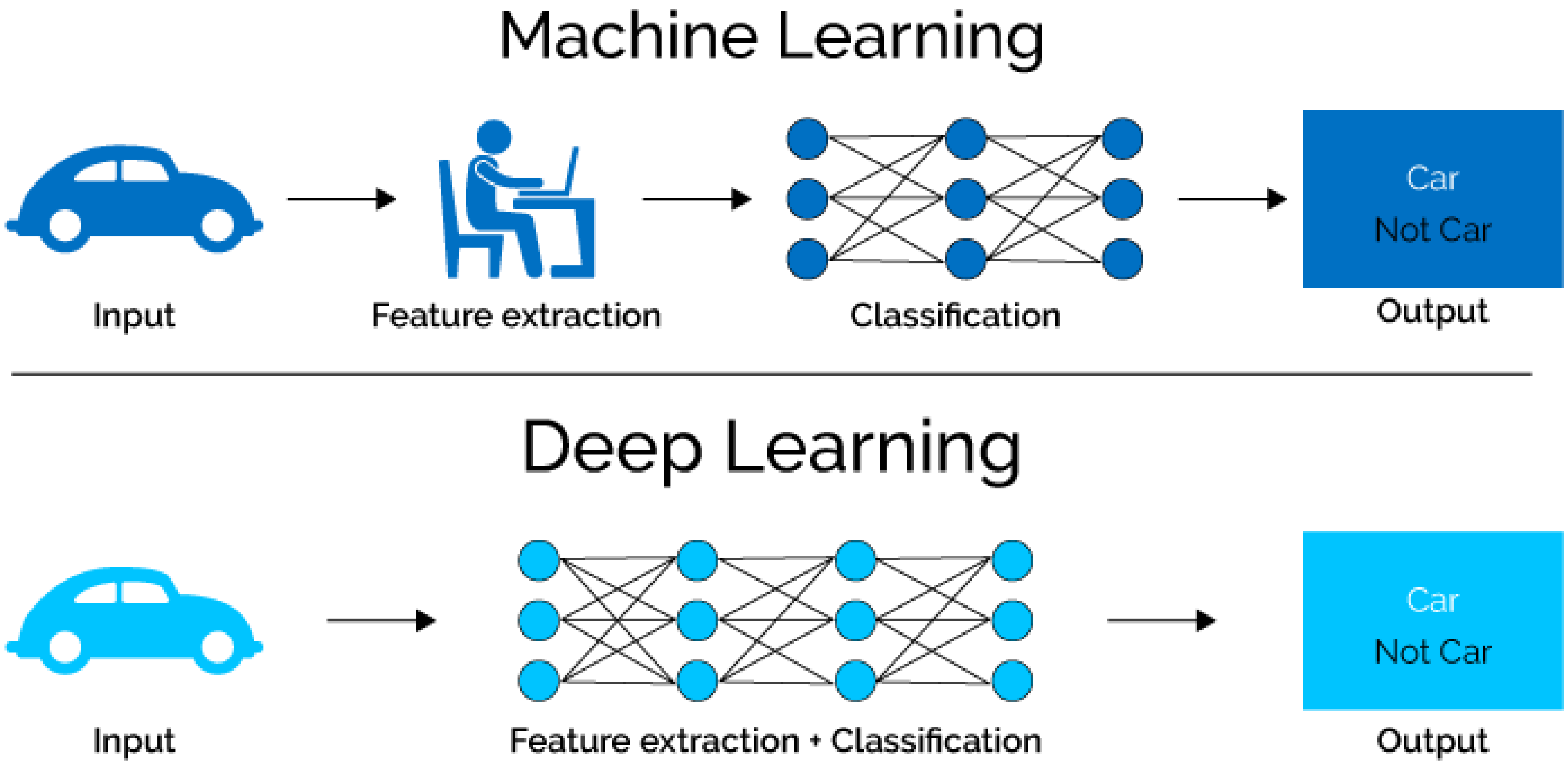
Feature Extraction

- Represent the data into the “learnable” object in computer



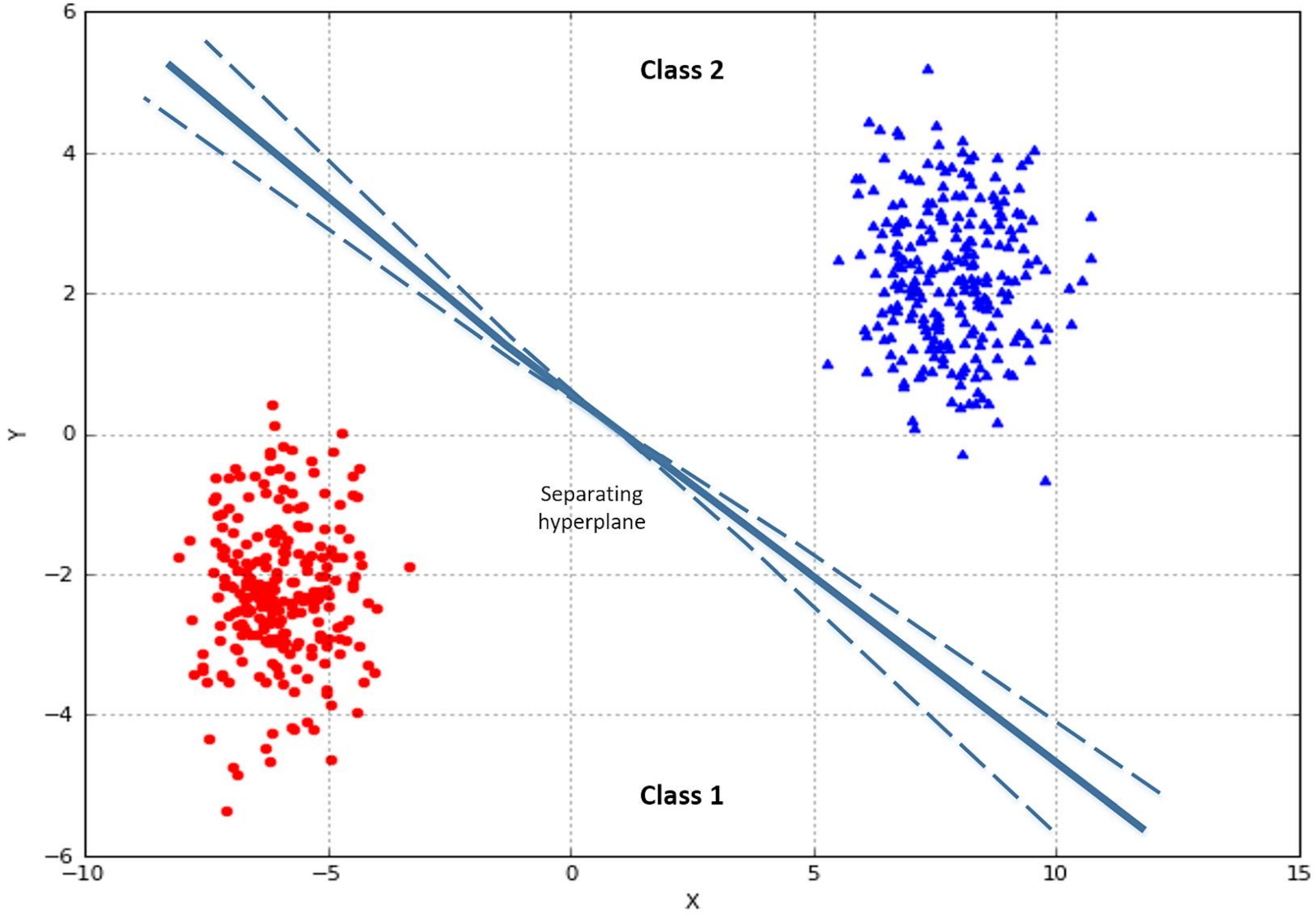
  - Image: pixels, rules, …
  - Audio: sampling, quantization
  - Text: ????

Feature Extraction



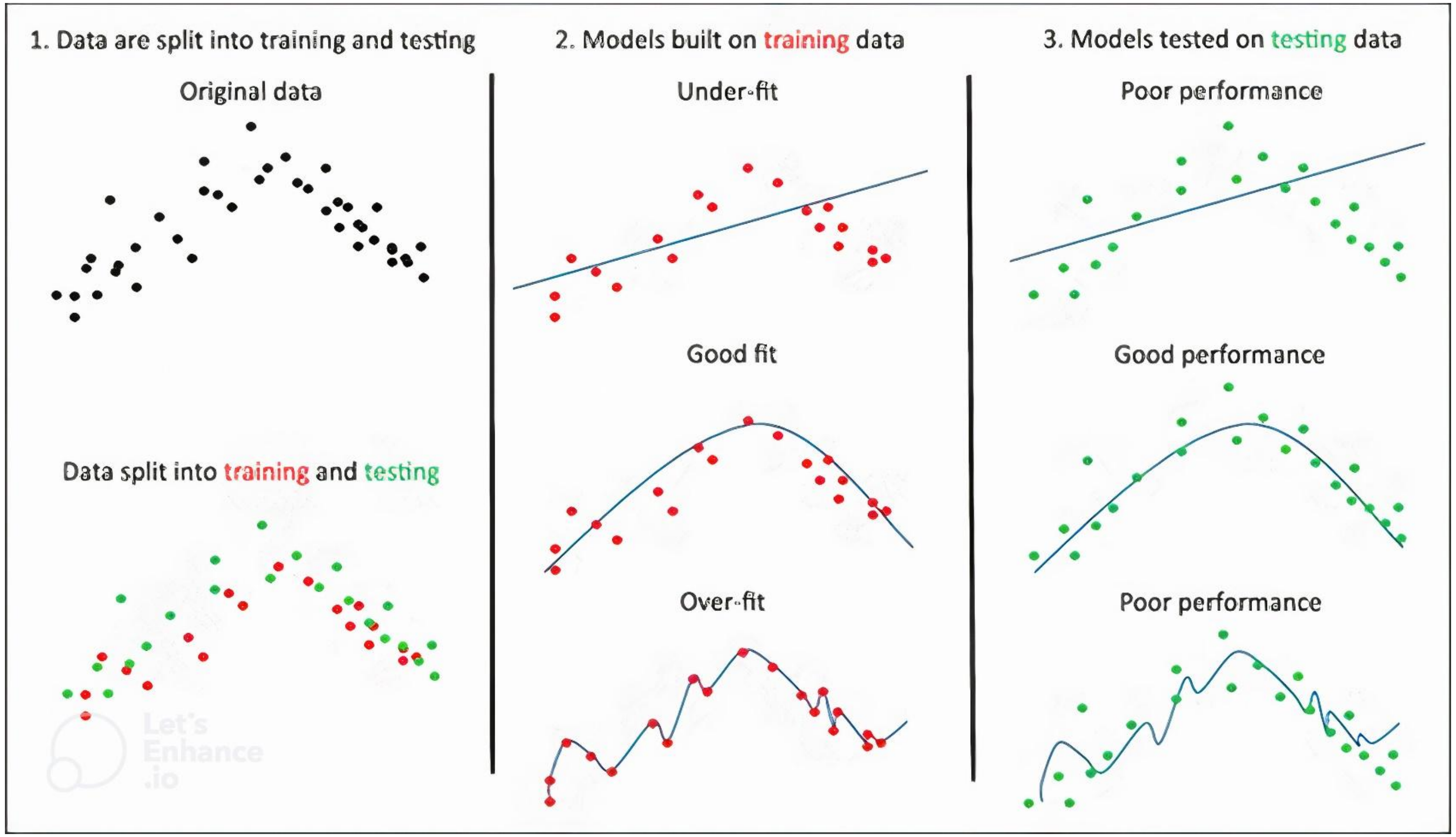
Evaluation

- How to choose the best “suitable” model



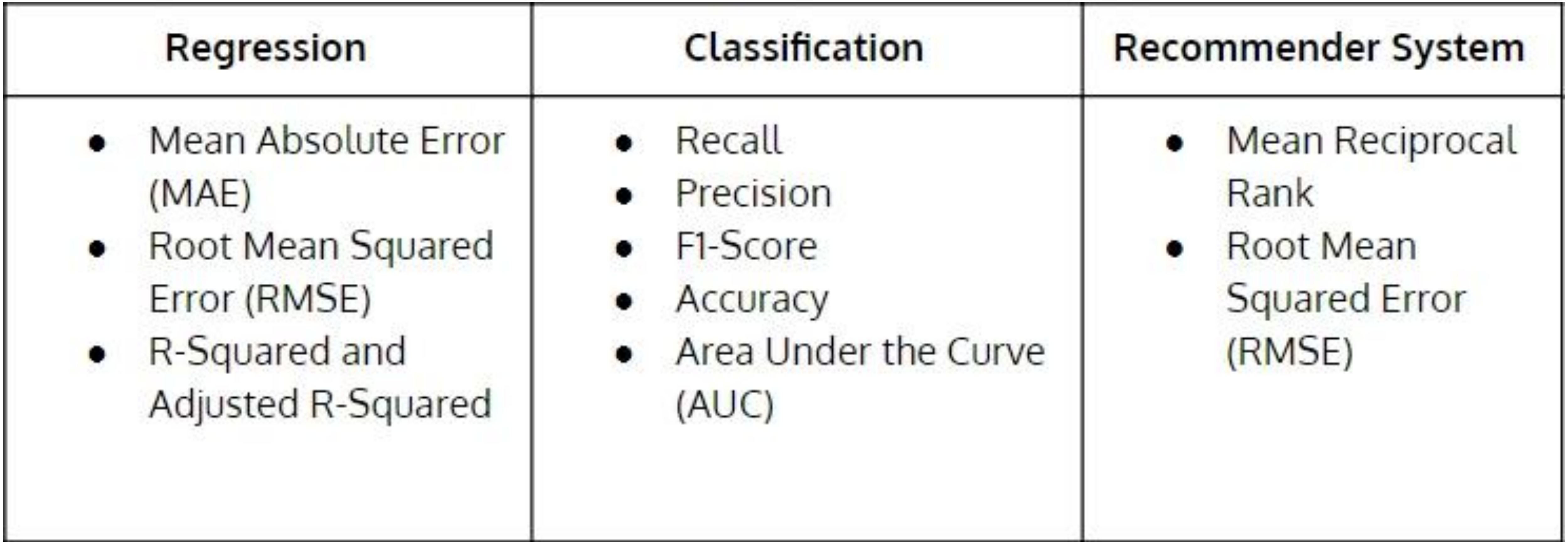
Evaluation

- Generalization refers to your model's ability to adapt properly to new, previously **unseen data**, drawn from the same distribution as the one used to create the mode


1. Data are split into training and testing
Original data
Data split into training and testing
2. Models built on training data
Under-fit
Good fit
Over-fit
3. Models tested on testing data
Poor performance
Good performance
Poor performance
Let's Enhance .io

Evaluation

- Based on some metrics that compares the model’s predicted labels with the known true labels



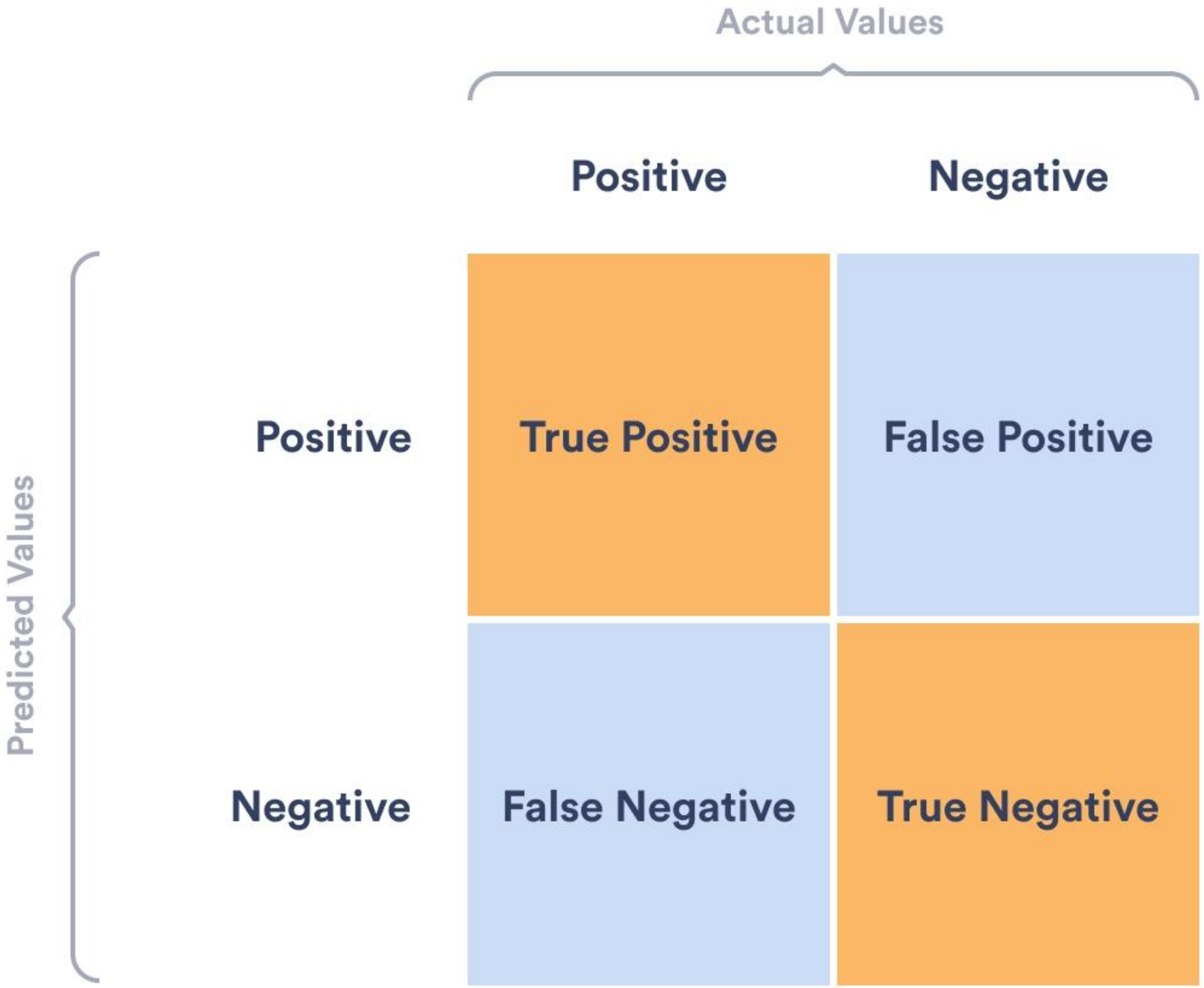
Regression	Classification	Recommender System
• Mean Absolute Error (MAE) • Root Mean Squared Error (RMSE) • R-Squared and Adjusted R-Squared	• Recall • Precision • F1-Score • Accuracy • Area Under the Curve (AUC)	• Mean Reciprocal Rank • Root Mean Squared Error (RMSE)

Evaluation: Classification

- A confusion matrix is a performance measurement technique
- Accuracy is a metric that generally describes how the model performs across all classes.

$$Accuracy = \frac{Correct\ prediction}{Total\ cases} * 100\%$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} * 100\%$$



Predicted Values \ Actual Values	Positive	Negative
Positive	True Positive	False Positive
Negative	False Negative	True Negative

Evaluation: Classification

- Precision:

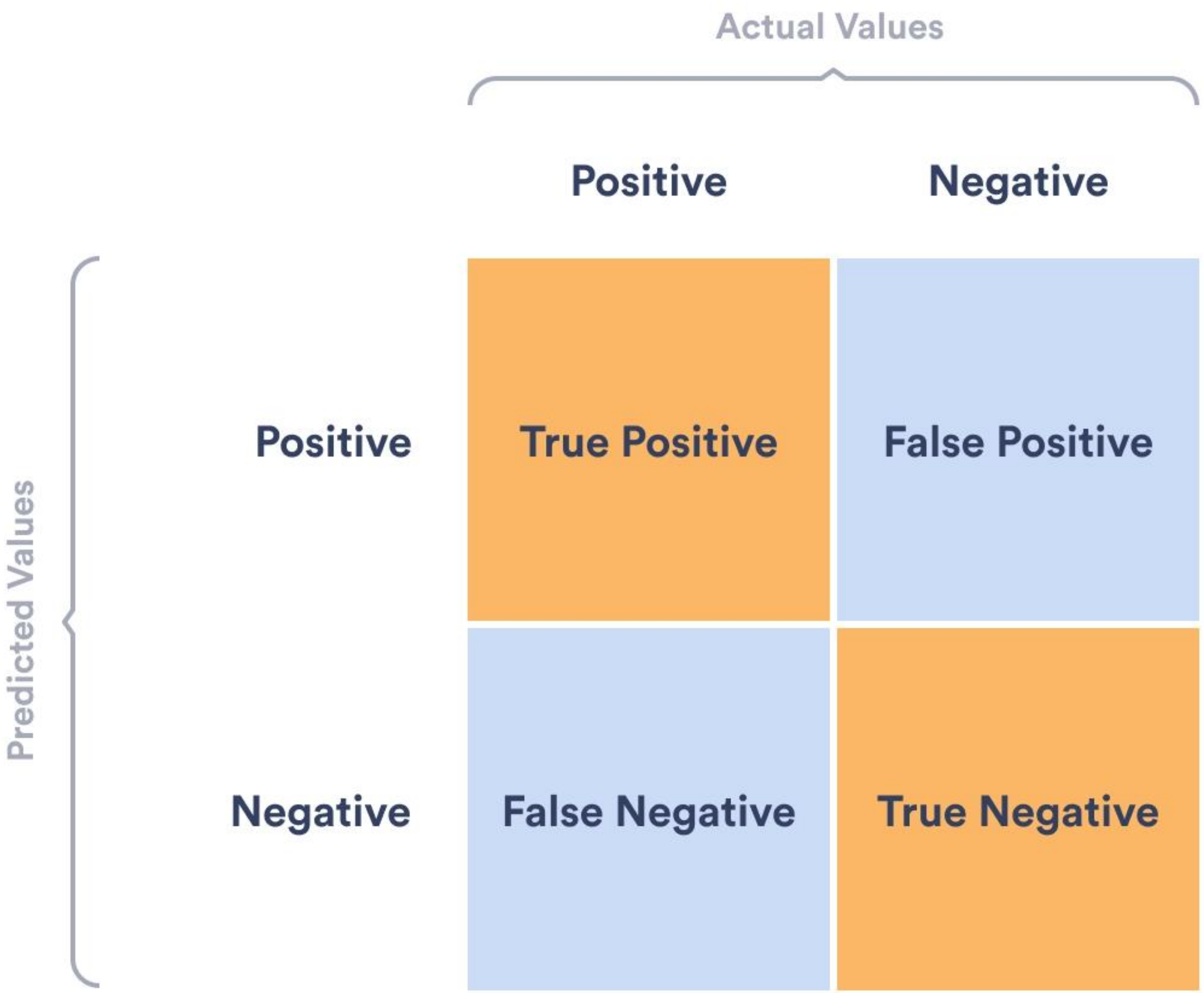
$$\text{Precision} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Positive}}$$

$$= \frac{\textit{True Positive}}{\textit{Total Predicted Positive}}$$

- Recall:

$$\text{Recall} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Negative}}$$

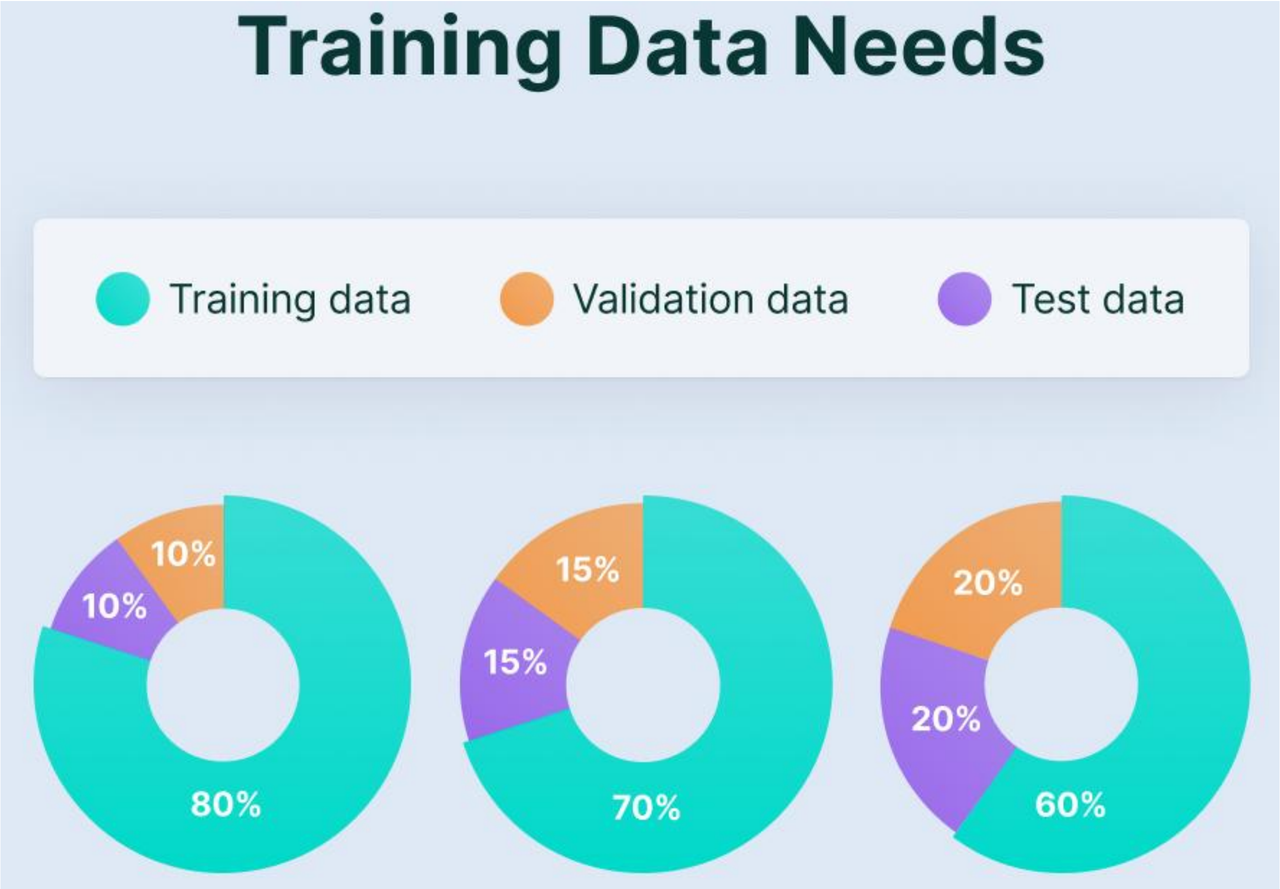
$$= \frac{\textit{True Positive}}{\textit{Total Actual Positive}}$$



		Actual Values	
		Positive	Negative
Predicted Values	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

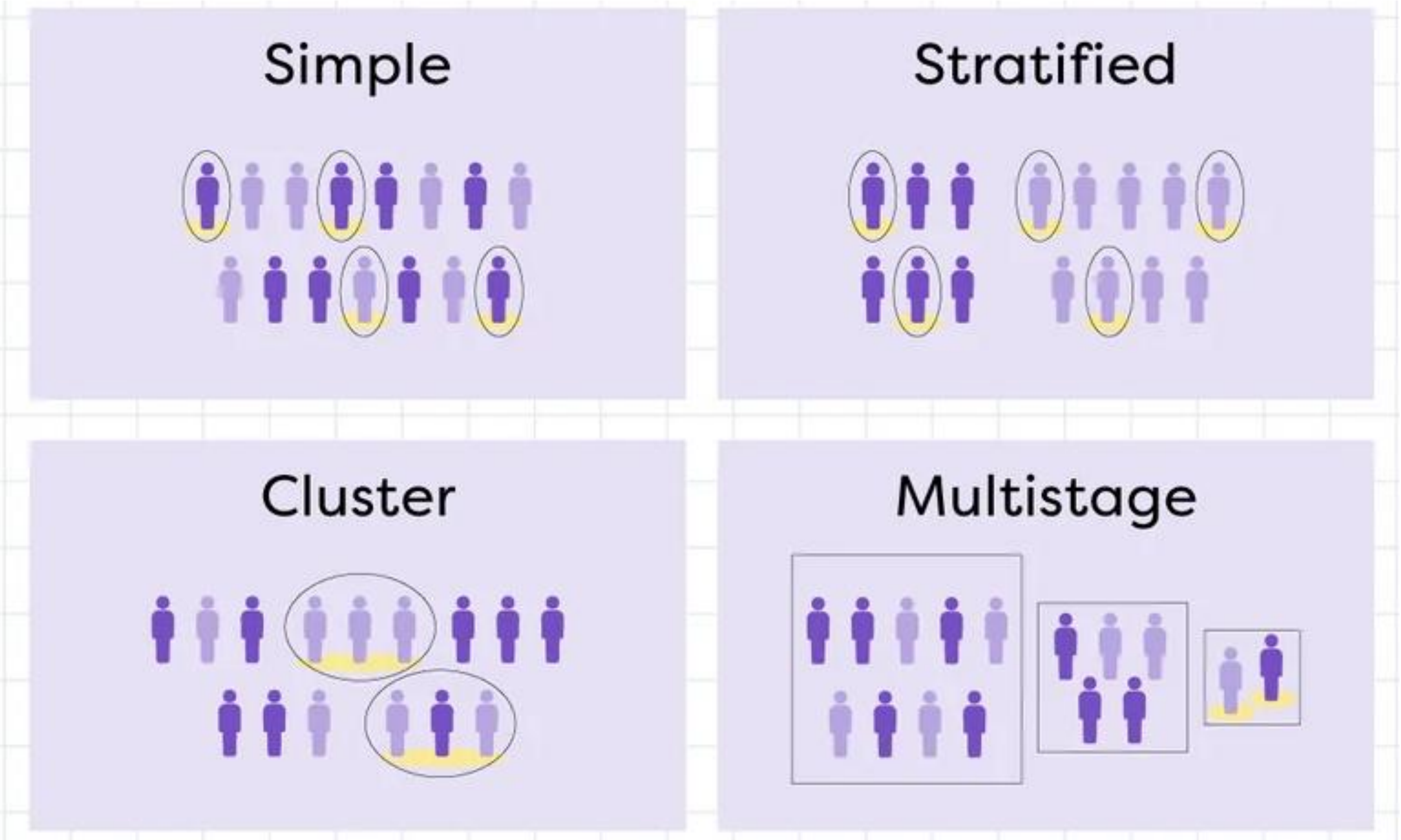
Data Splitting Techniques

- Kind of data:
  - Train: learn from the historical patterns to reveal the hidden models
  - Validating: choose the best/good model, tune the model's hyperparameters and configurations
  - Testing: evaluate the practical performance



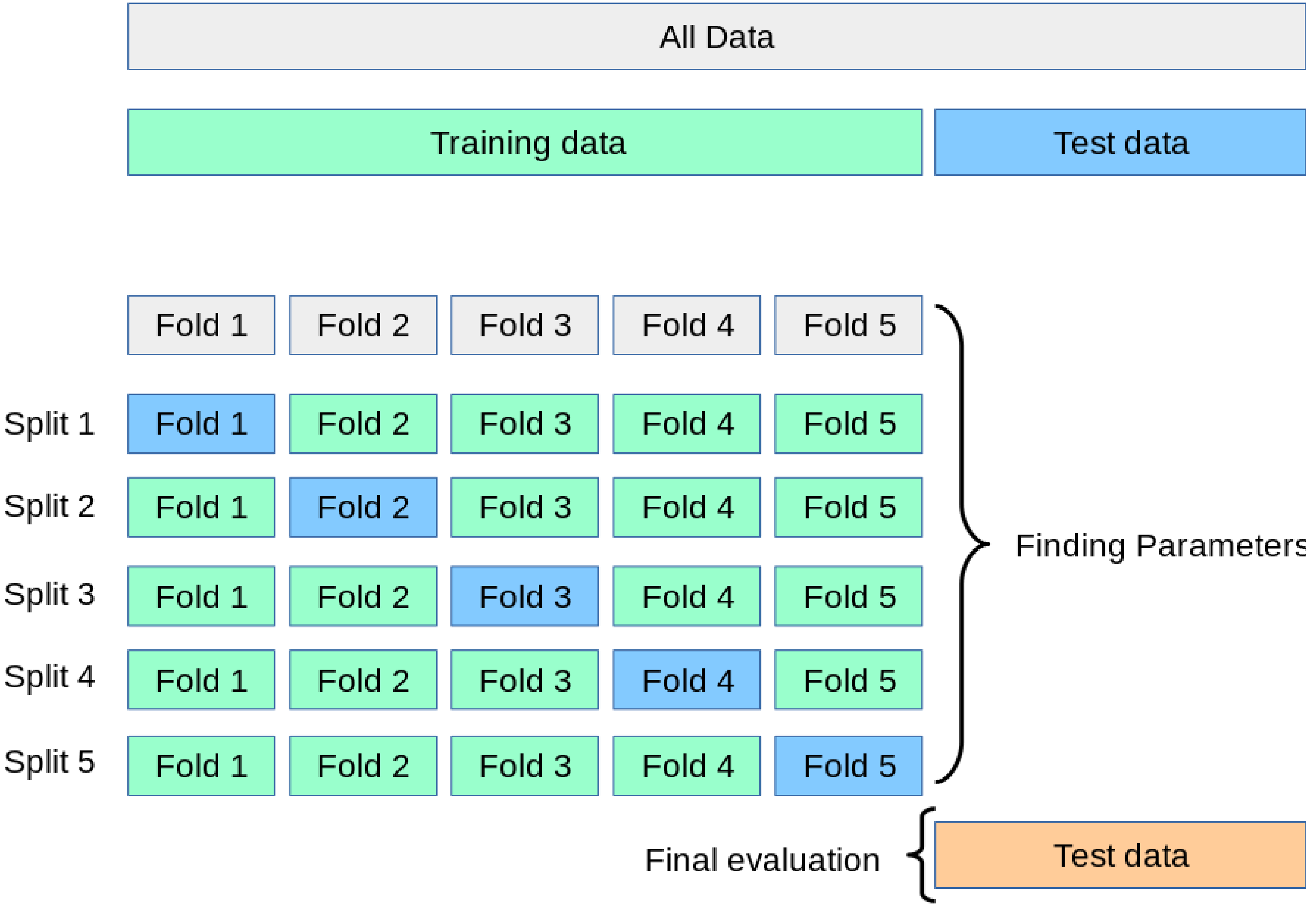
Data Splitting Techniques

- Random sampling
  - Simple
  - Stratified: random from each group (group division by label)
  - Cluster: group division by features
  - Multistage: divide by many kinds of random sampling



Data Splitting Techniques

- Cross-validation:
  - Cross over training data
  - Cross over all data
- This alleviates any bias occurring as selecting data in the training and validation sets.
- The goal of this technique is validating and testing, not training model.



All Data

Training data | Test data

Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Split 1: Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Split 2: Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Split 3: Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

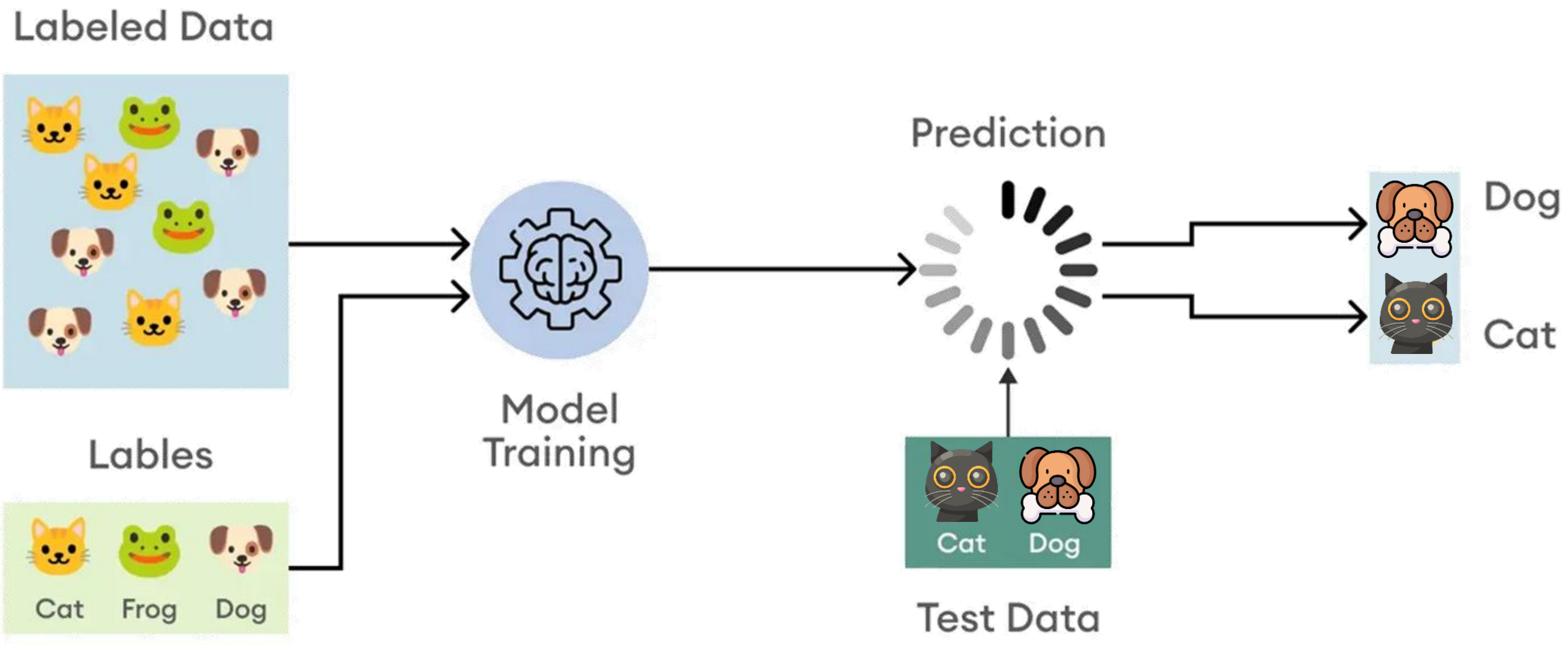
Split 4: Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Split 5: Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5

Finding Parameters

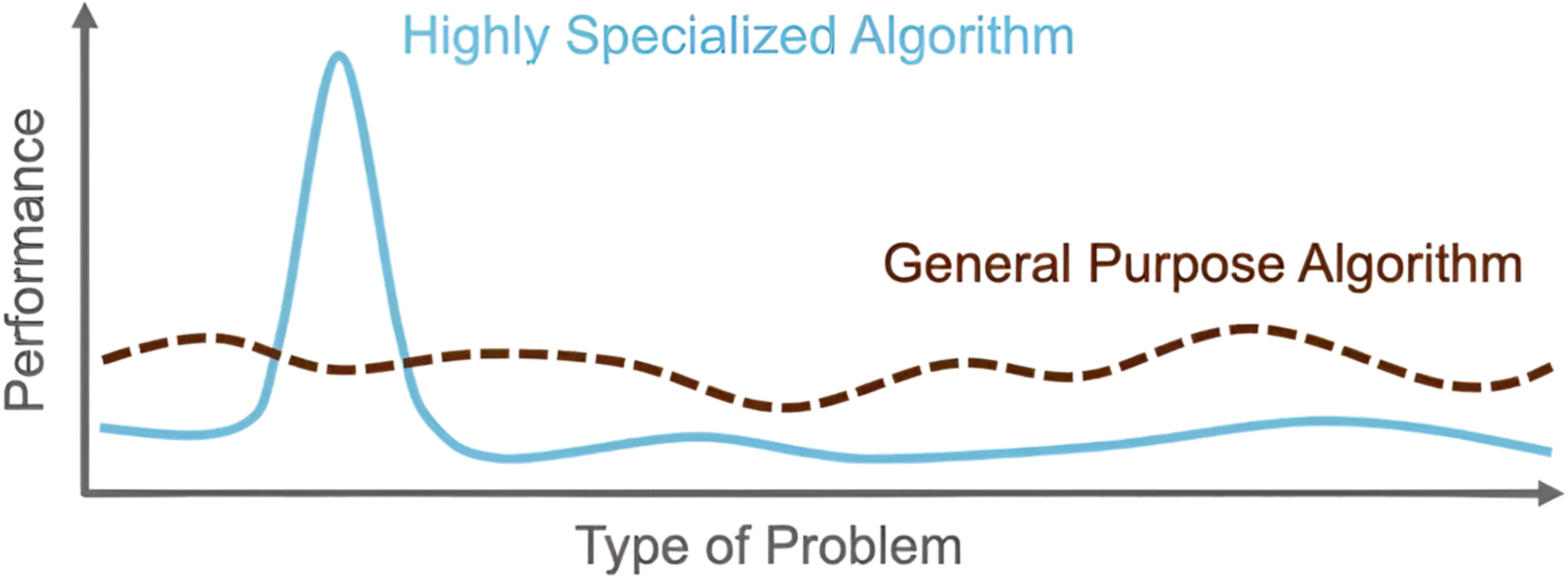
Final evaluation: Test data

Generalization Error



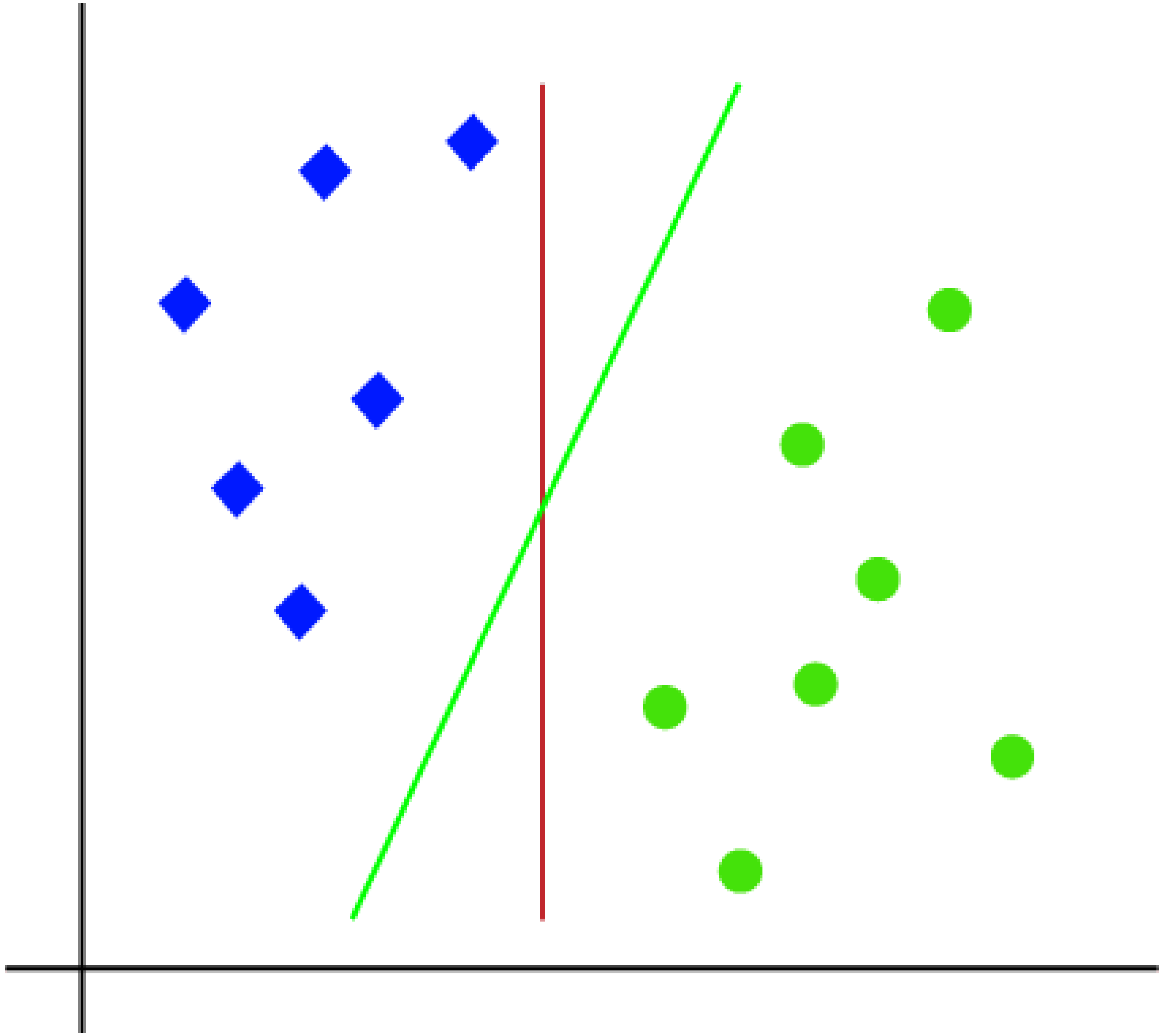
No free lunch theorem

- If an algorithm performs well on a certain class of problems, then it necessarily pays for that with degraded performance on the set of all remaining problems. (Wolpert and Macready)



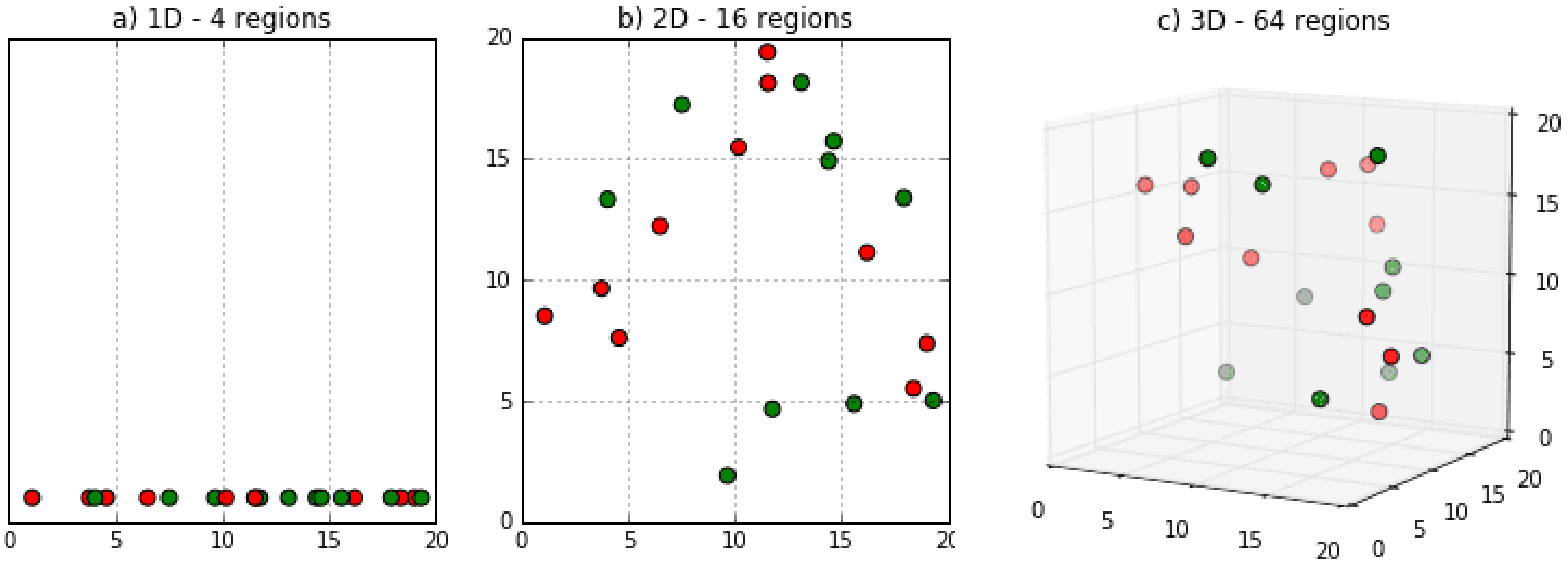
Ugly duckling theorem

- There is **no** privileged or **best feature** representation, and that even the notion of similarity between patterns depends implicitly on assumption that may or may not be correct. (Watanabe)



Curse of dimensionality

- Various phenomena that arise when analyzing data in high-dimensional spaces that do not occur in low-dimensional settings



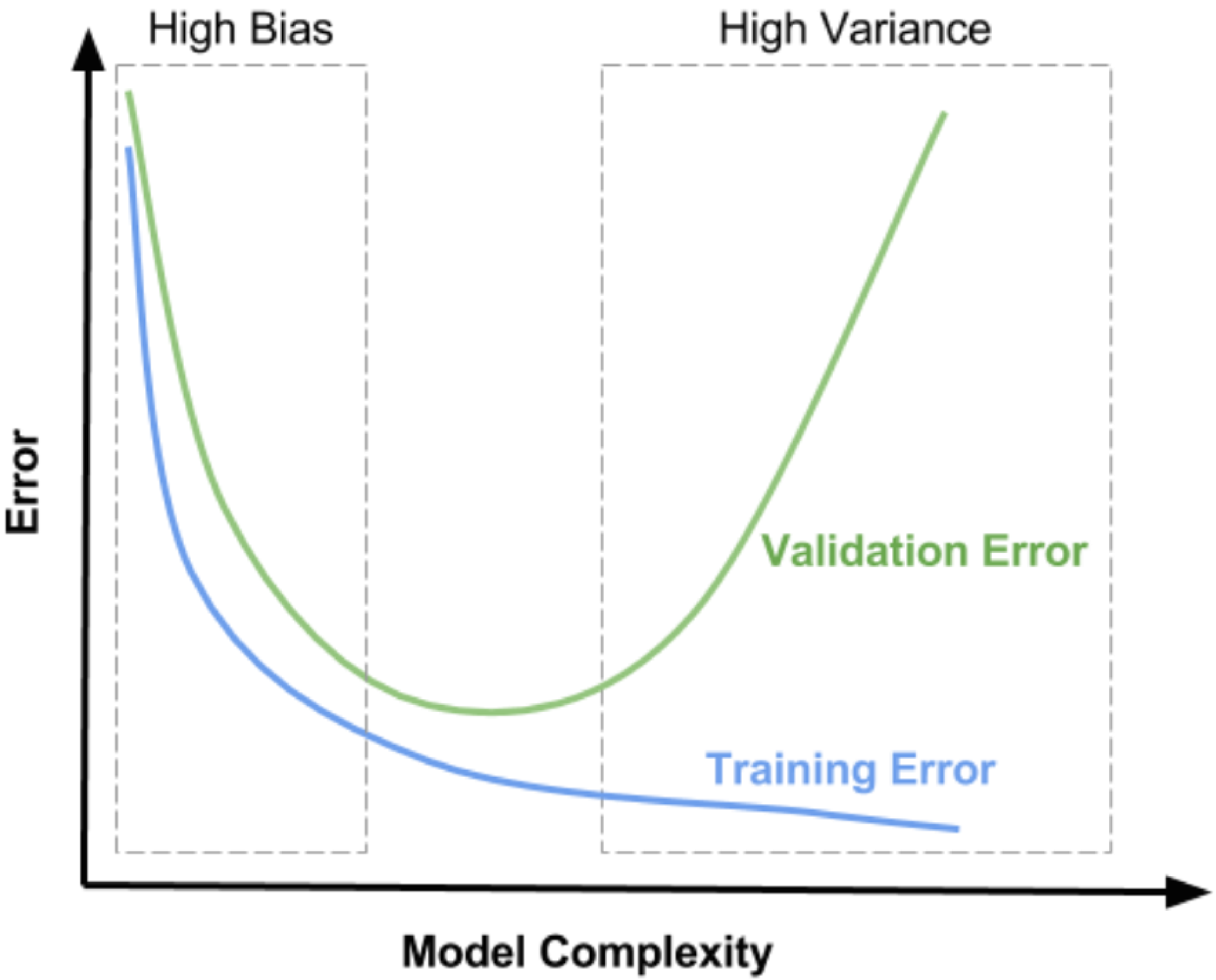
Bias & Variance

- Estimation Error

– Bias: difference between expected prediction and correct value

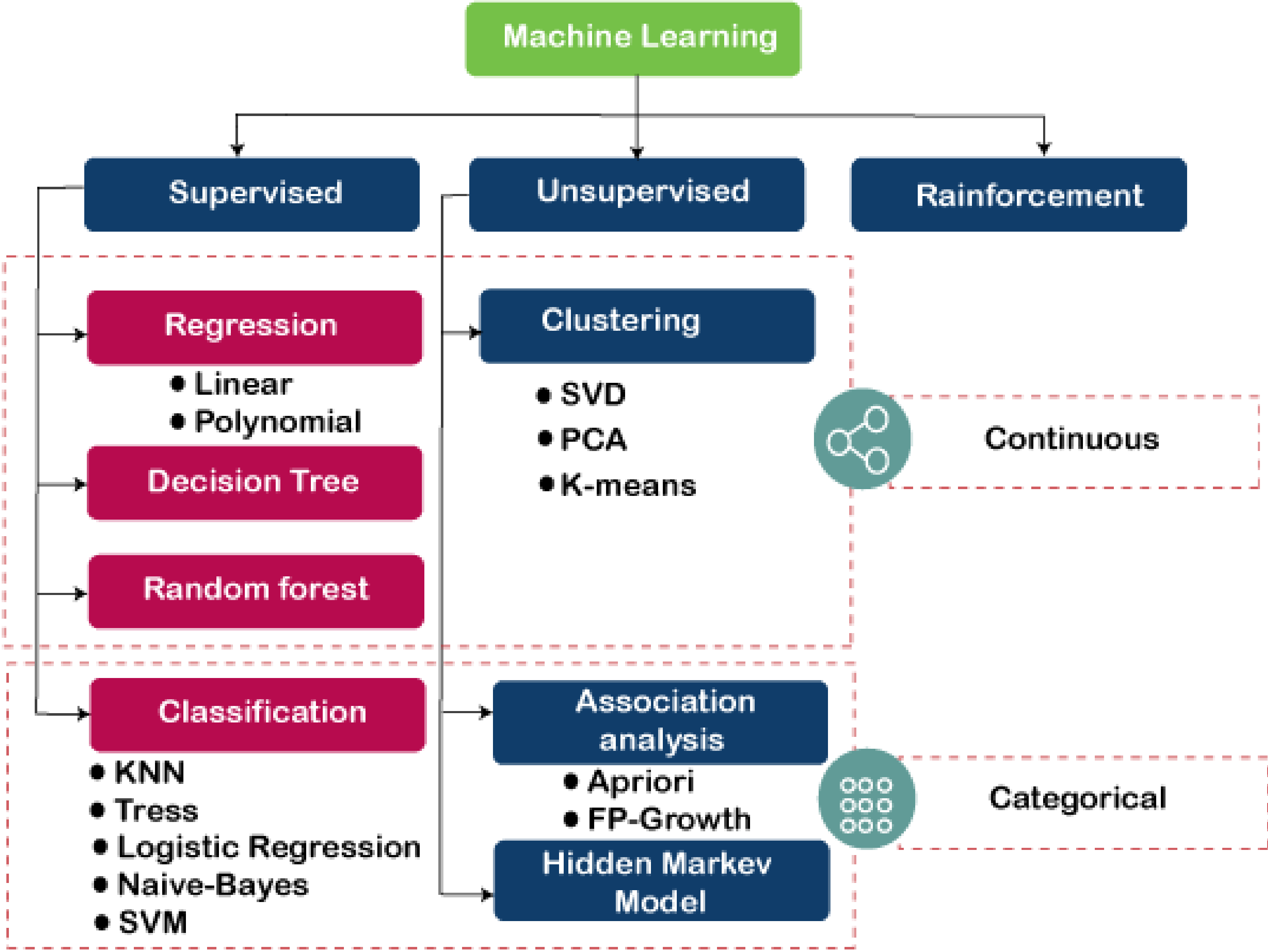
– Variance: variability of prediction for a given data point

– Trade-off



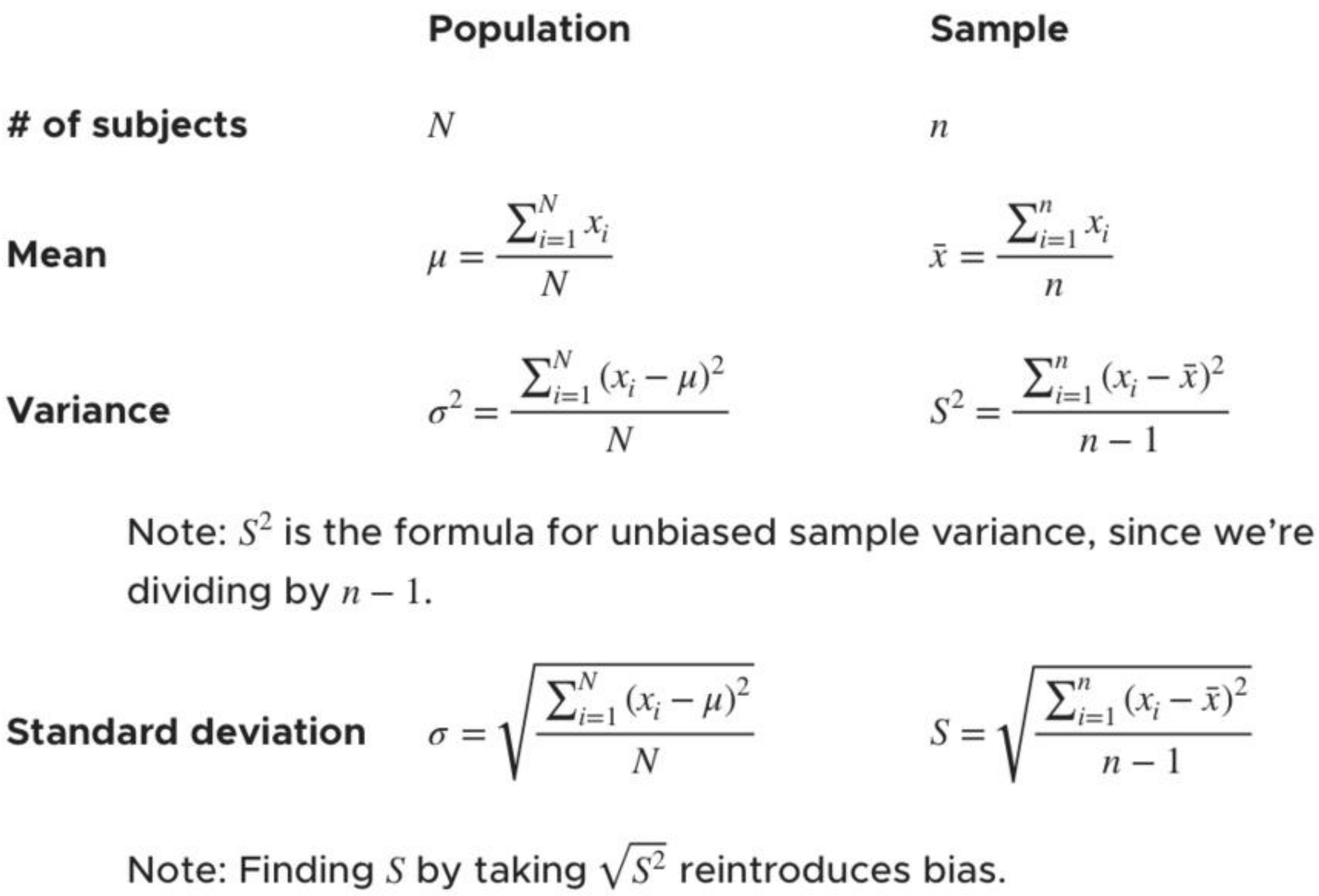
Machine Learning Algorithms

- Read more at:
  - https://www.javatpoint.com/machine-learning-algorithms
  - https://machinelearningcoban.com/



Basic Statistics

Reviews



	Population	Sample
# of subjects	N	n
Mean	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$
Standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$

Note: S^2 is the formula for unbiased sample variance, since we're dividing by $n - 1$.

Note: Finding S by taking $\sqrt{S^2}$ reintroduces bias.

Exercise

Given a sample: X = [10, 7, -22, 4, -9]

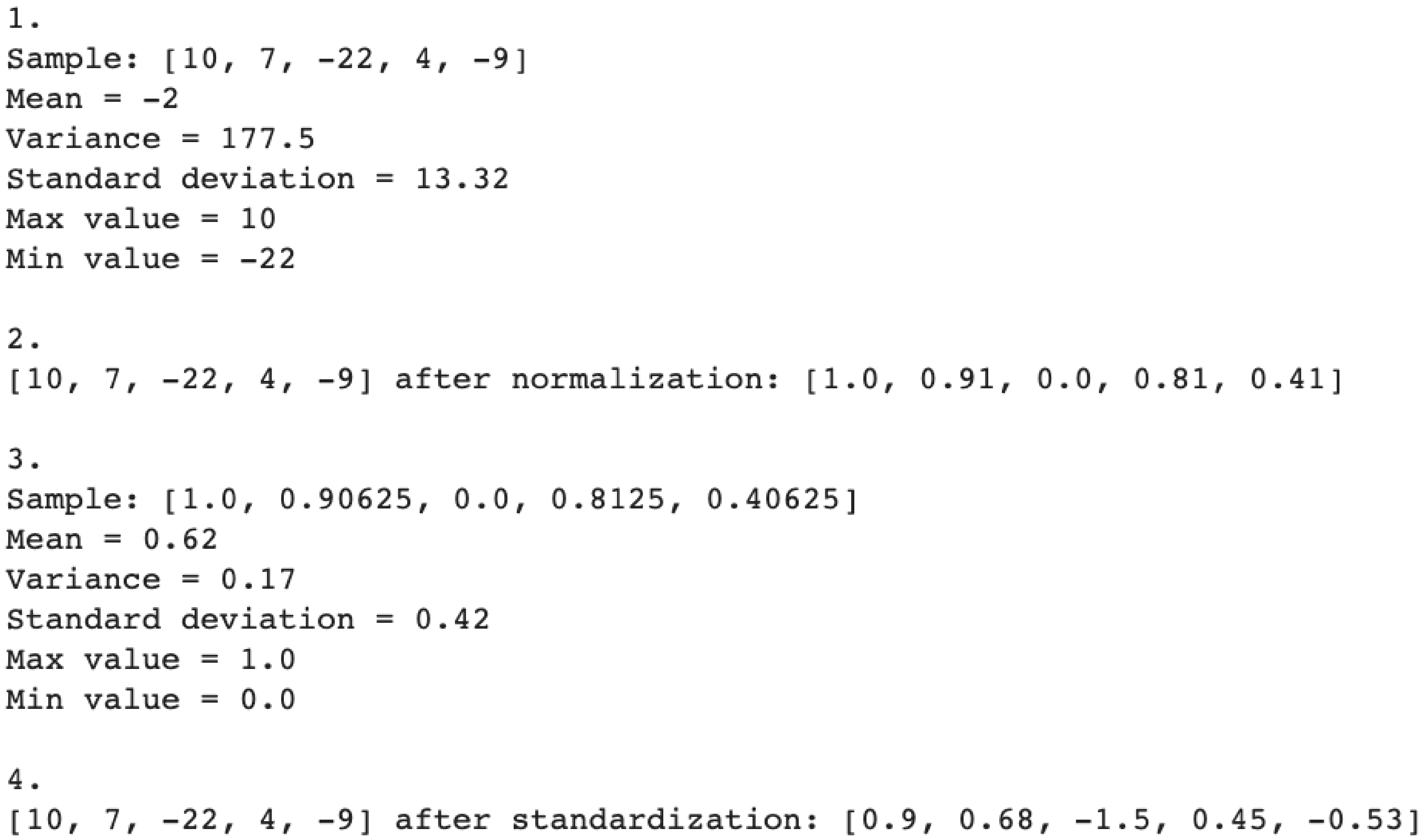
1. Calculate the mean, variance and standard deviation
2. Normalize the data points into the range [0, 1]
3. Calculate the mean, variance and standard deviation after normalizing
4. Standardize X s.t. the mean of X is 0 and standard deviation is 1

Note: round the value up with 2 digits after the floating point (e.g., 10.23)

Normalization: $X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$

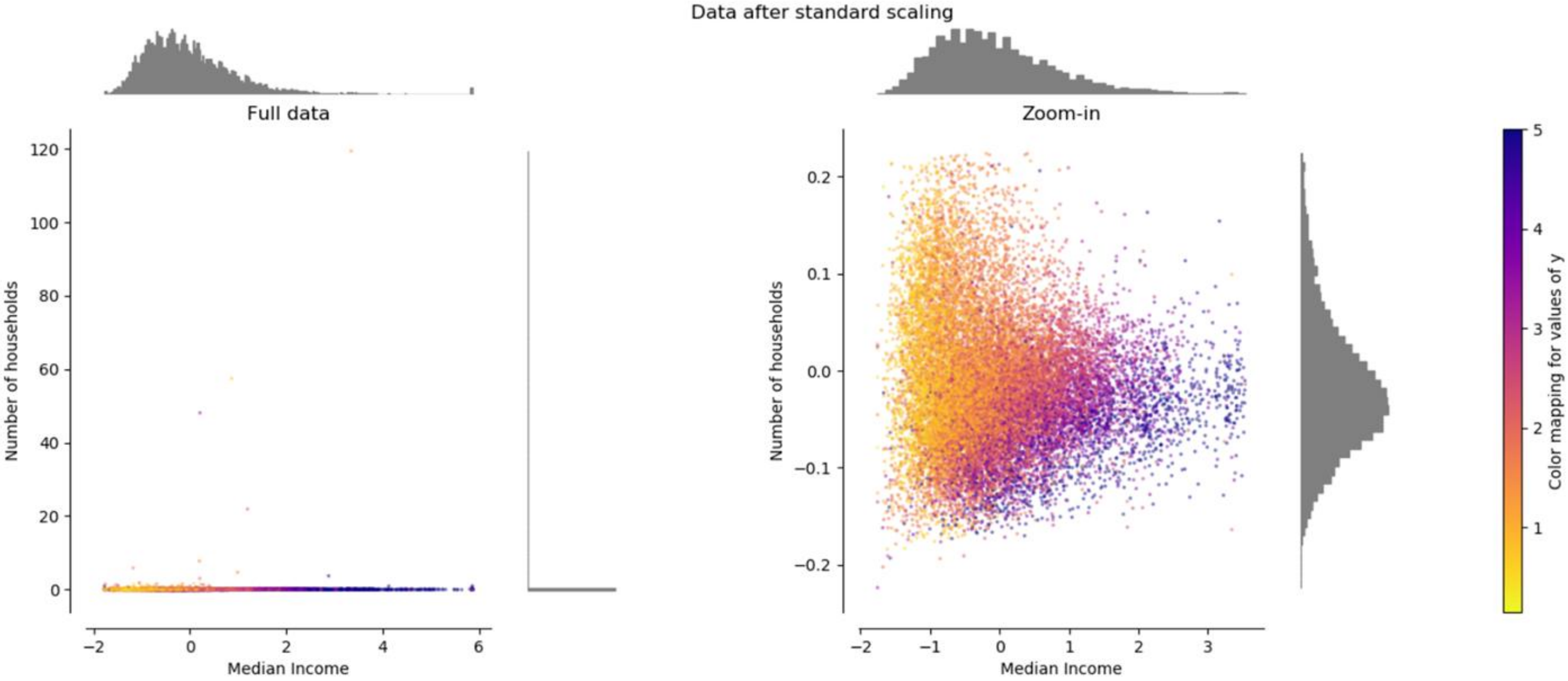
Standardization: $X_{new} = \frac{X - \bar{X}}{s}$

Exercise: Solution



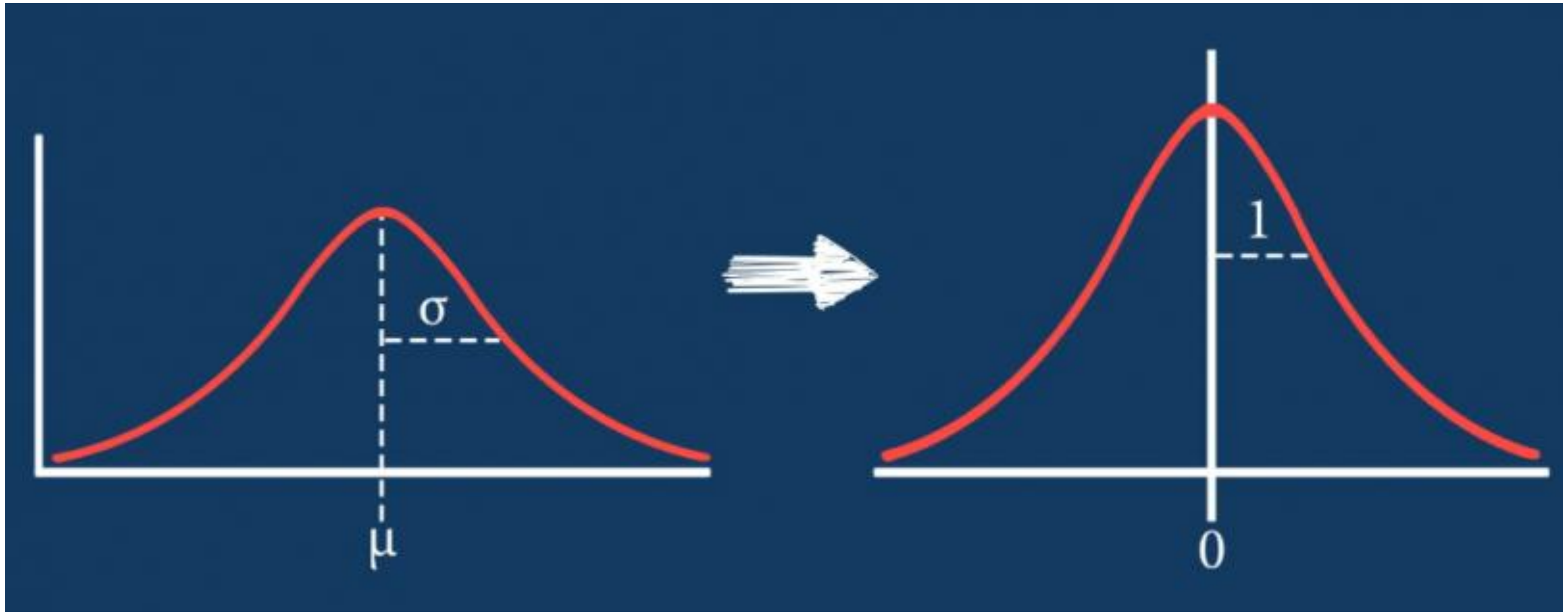
```
1.
Sample: [10, 7, -22, 4, -9]
Mean = -2
Variance = 177.5
Standard deviation = 13.32
Max value = 10
Min value = -22

2.
[10, 7, -22, 4, -9] after normalization: [1.0, 0.91, 0.0, 0.81, 0.41]

3.
Sample: [1.0, 0.90625, 0.0, 0.8125, 0.40625]
Mean = 0.62
Variance = 0.17
Standard deviation = 0.42
Max value = 1.0
Min value = 0.0

4.
[10, 7, -22, 4, -9] after standardization: [0.9, 0.68, -1.5, 0.45, -0.53]
```

Data Standardization



Source: https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html

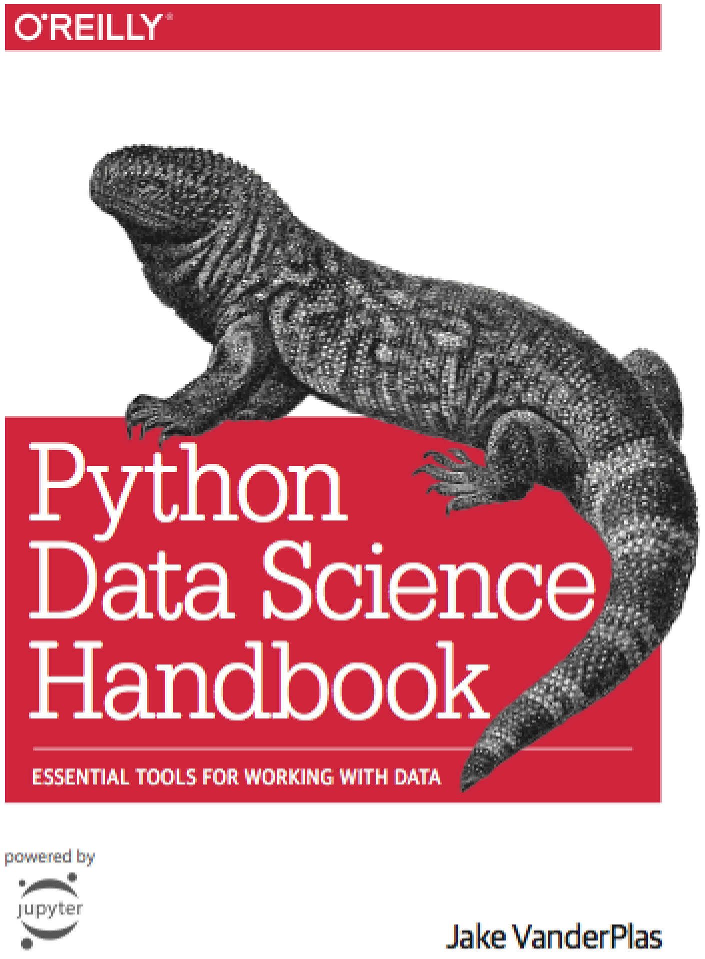
Data Standardization



- Needed before statistic-based algorithm (e.g. clustering, PCA, SVM, …)
- No need in regression and tree-based algorithms

Python: Introduction

Introduction


O'REILLY
Python Data Science Handbook
ESSENTIAL TOOLS FOR WORKING WITH DATA
powered by
jupyter
Jake VanderPlas

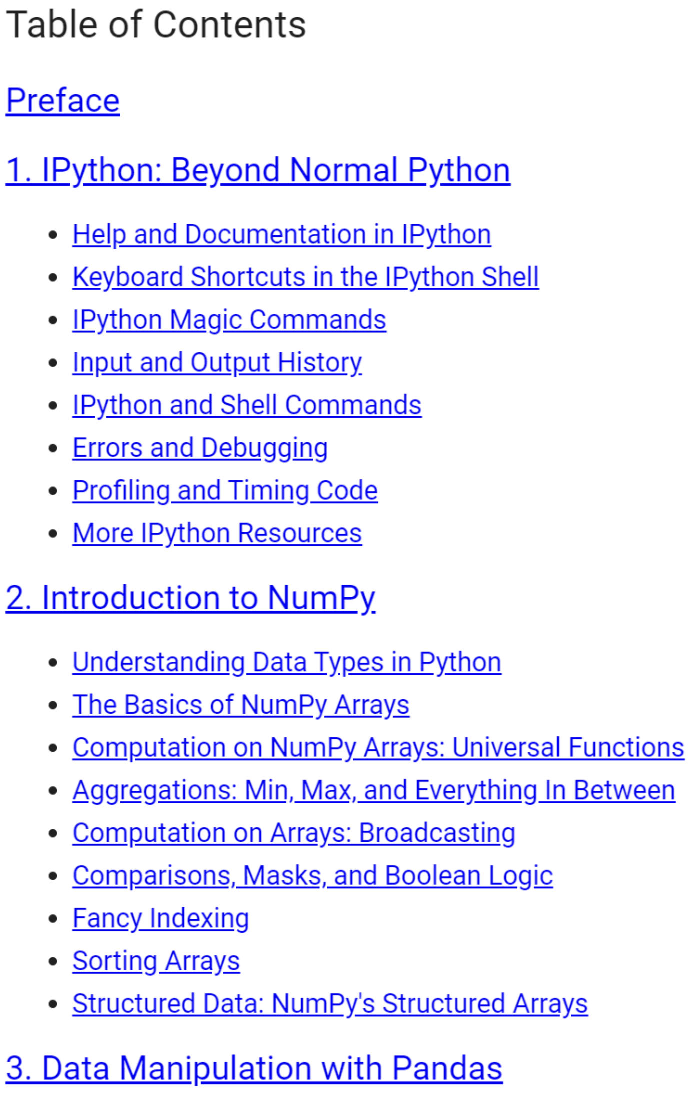


Table of Contents

Preface

1. IPython: Beyond Normal Python

- Help and Documentation in IPython
- Keyboard Shortcuts in the IPython Shell
- IPython Magic Commands
- Input and Output History
- IPython and Shell Commands
- Errors and Debugging
- Profiling and Timing Code
- More IPython Resources

2. Introduction to NumPy

- Understanding Data Types in Python
- The Basics of NumPy Arrays
- Computation on NumPy Arrays: Universal Functions
- Aggregations: Min, Max, and Everything In Between
- Computation on Arrays: Broadcasting
- Comparisons, Masks, and Boolean Logic
- Fancy Indexing
- Sorting Arrays
- Structured Data: NumPy's Structured Arrays

3. Data Manipulation with Pandas

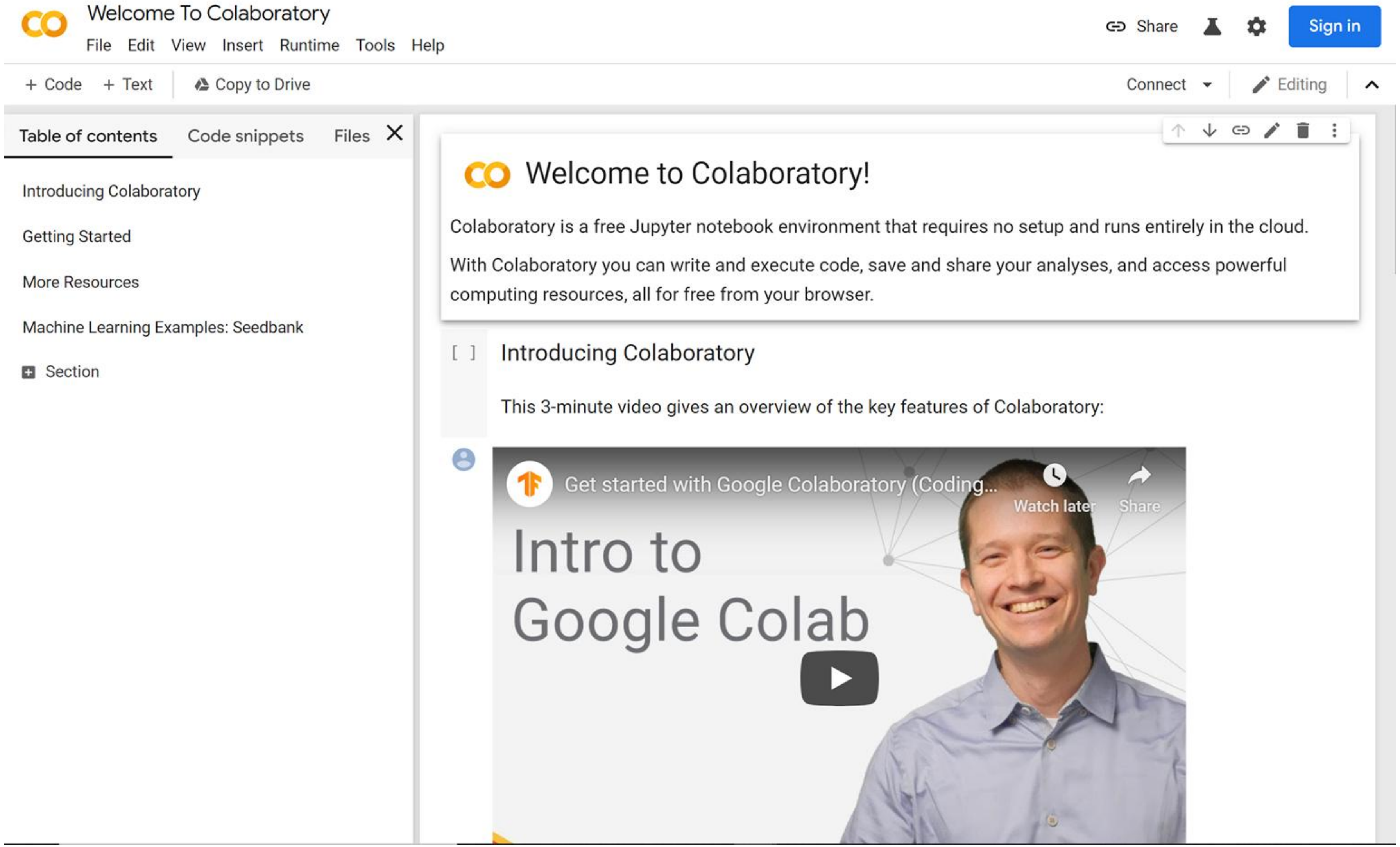
Link:

https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/Index.ipynb

Keras

Pytorch

Google Colab



Welcome To Colaboratory

Introducing Colaboratory
Getting Started
More Resources
Machine Learning Examples: Seedbank
Section

Welcome to Colaboratory!

Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud.

With Colaboratory you can write and execute code, save and share your analyses, and access powerful computing resources, all for free from your browser.

Introducing Colaboratory

This 3-minute video gives an overview of the key features of Colaboratory:

THANK YOU

for YOUR ATTENTION