Aprendizagem Computacional
Machine Learning

Departamento de Engenharia Informática,
Universidade de Coimbra

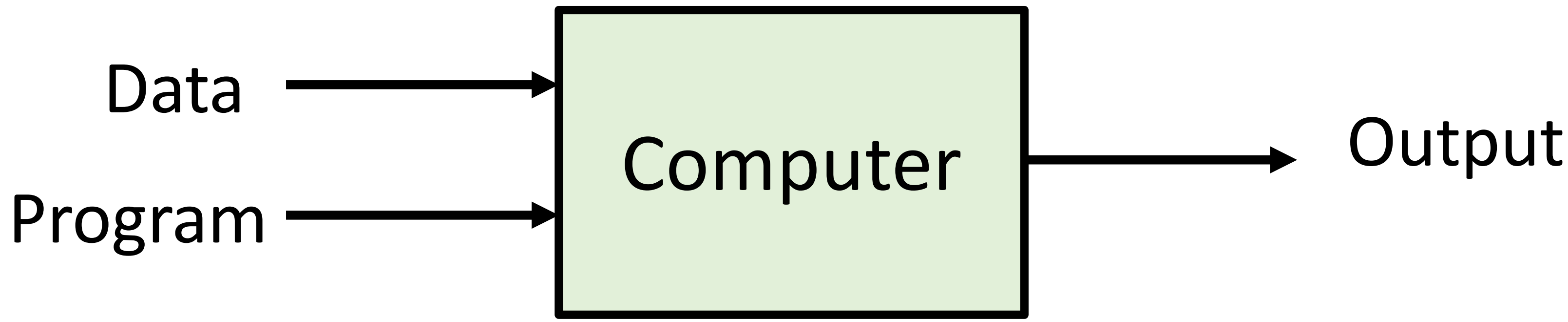
Catarina Silva, 2024

Contents

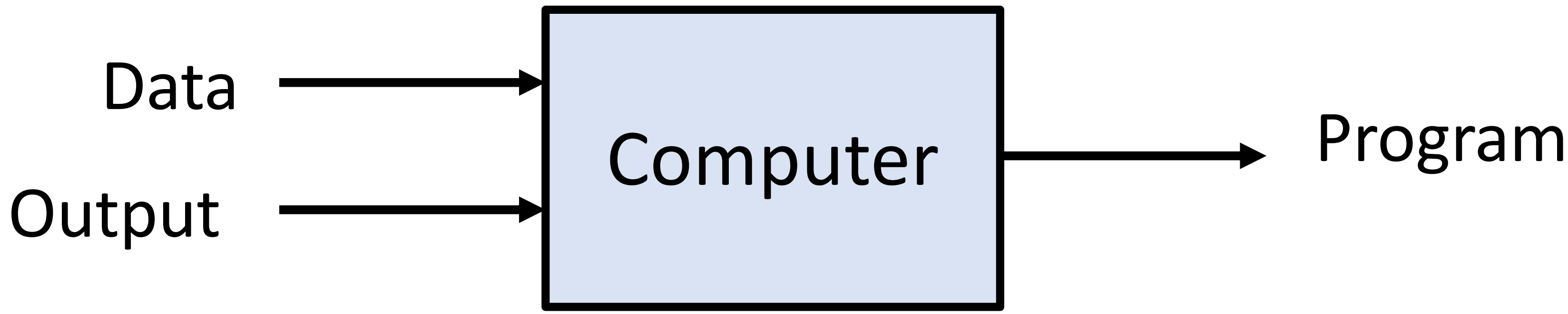
- What is Machine Learning
- Types of problems
- Types of data
- Types of learning
- ML workflow
    - Preprocessing
    - Modeling
    - Evaluation
- ML Challenges

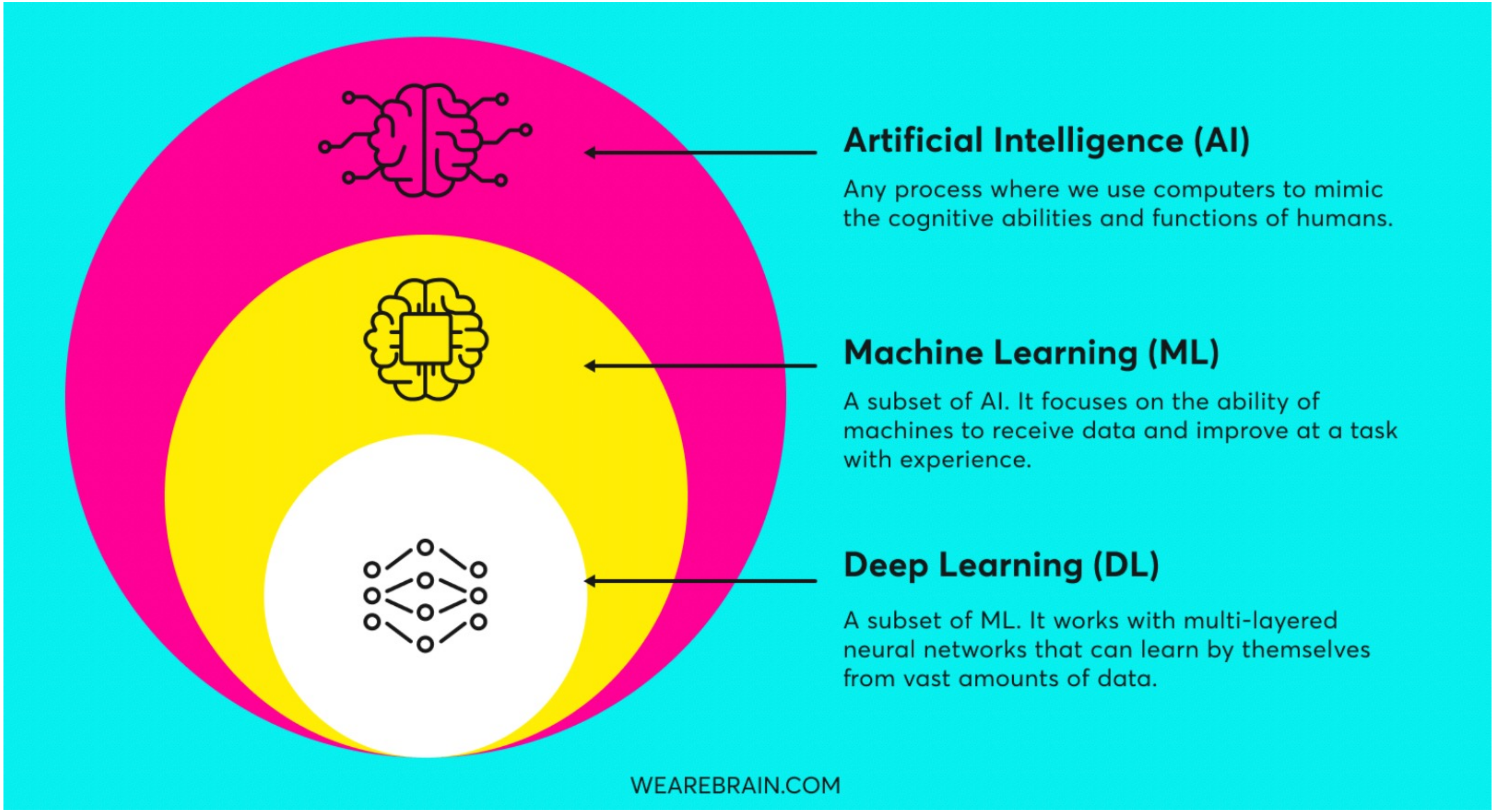
Traditional vs ML

- **Traditional Programming**



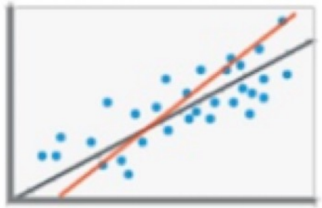
- **Machine Learning**



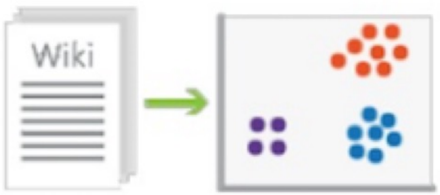

Artificial Intelligence (AI)
Any process where we use computers to mimic the cognitive abilities and functions of humans.
Machine Learning (ML)
A subset of AI. It focuses on the ability of machines to receive data and improve at a task with experience.
Deep Learning (DL)
A subset of ML. It works with multi-layered neural networks that can learn by themselves from vast amounts of data.
WEAREBRAIN.COM

Types of problems (1)

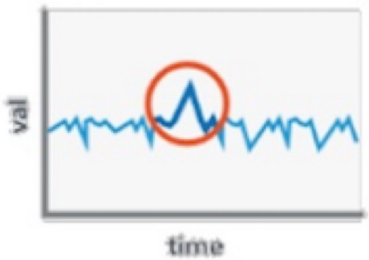
Classification
(supervised – predictive)



Regression
(supervised – predictive)

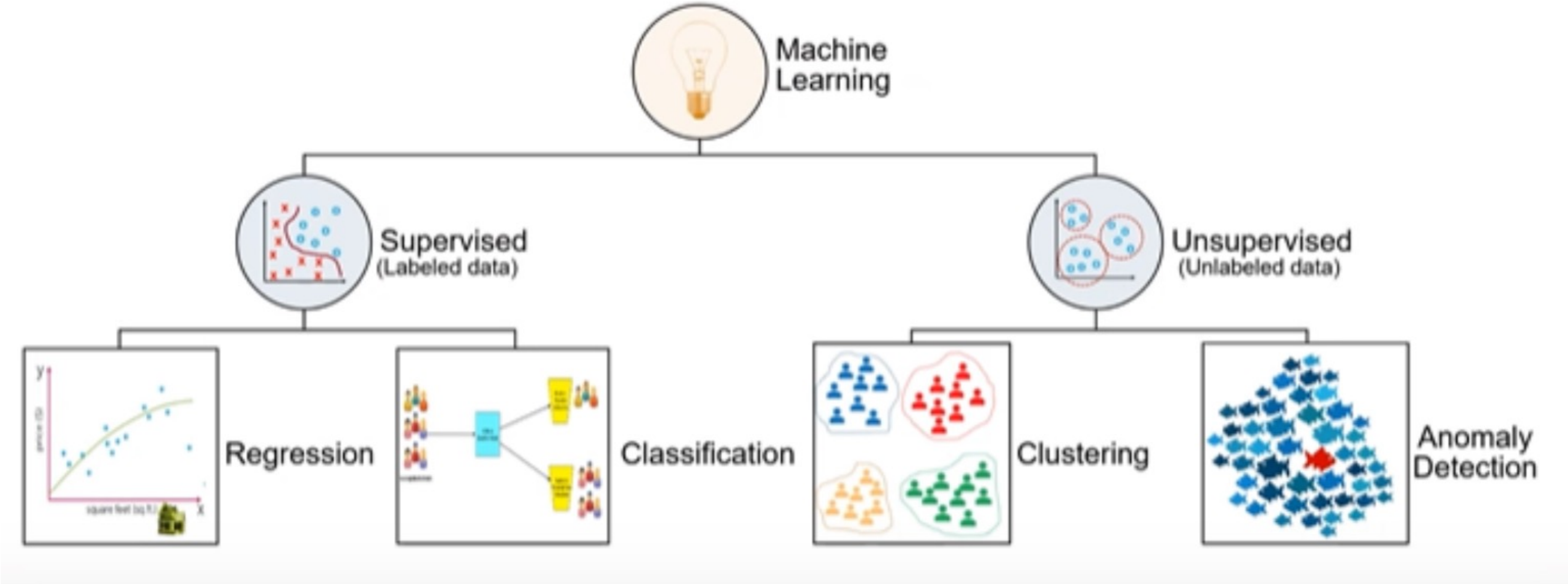


Clustering
(unsupervised – descriptive)



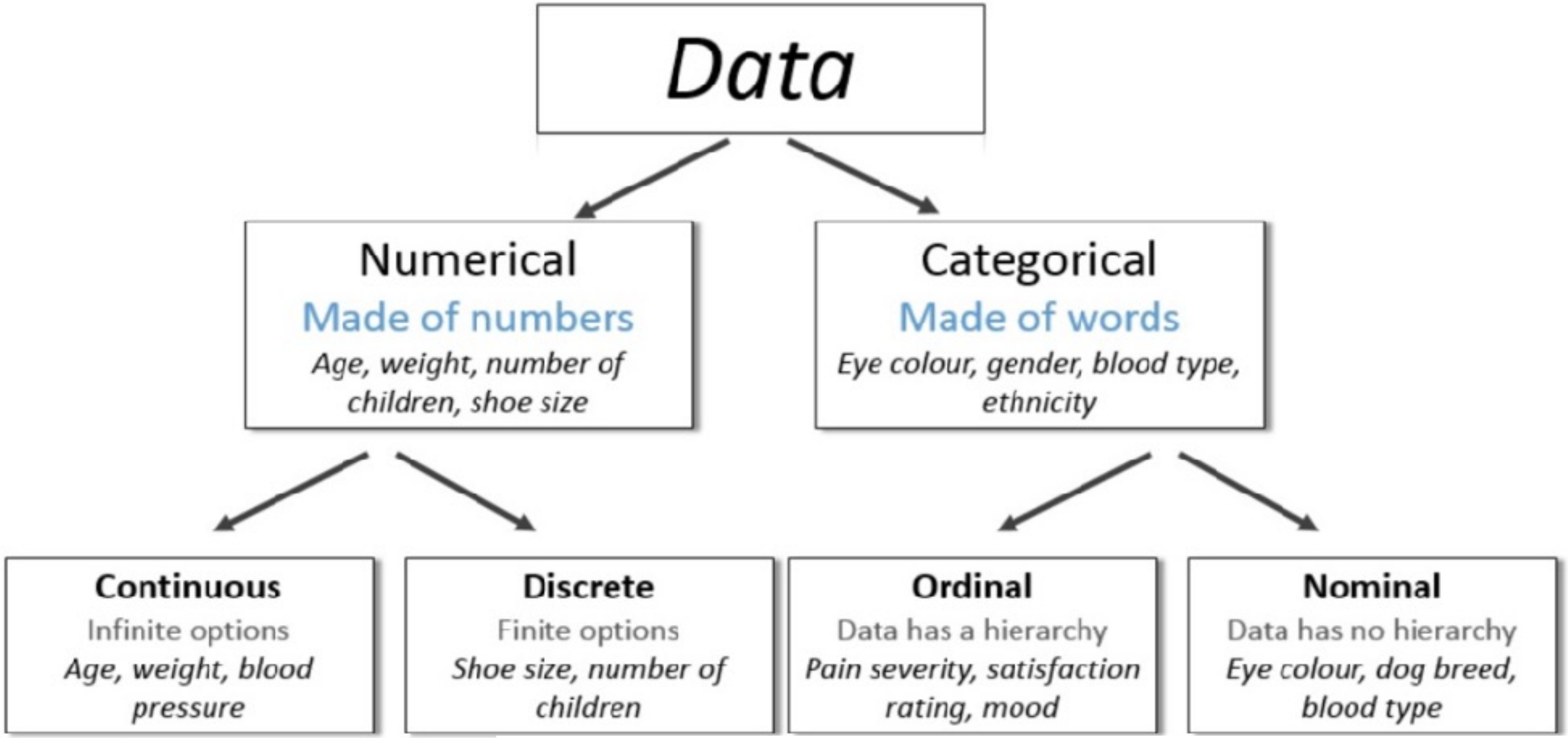
Anomaly Detection
(unsupervised– descriptive)

Types of problems (2)



www.altair.com

Types of data



www.medium.com

Types of learning (1)

Supervised learning: Learn from a previously labeled training set

Example: SPAM detector for previous SPAM examples

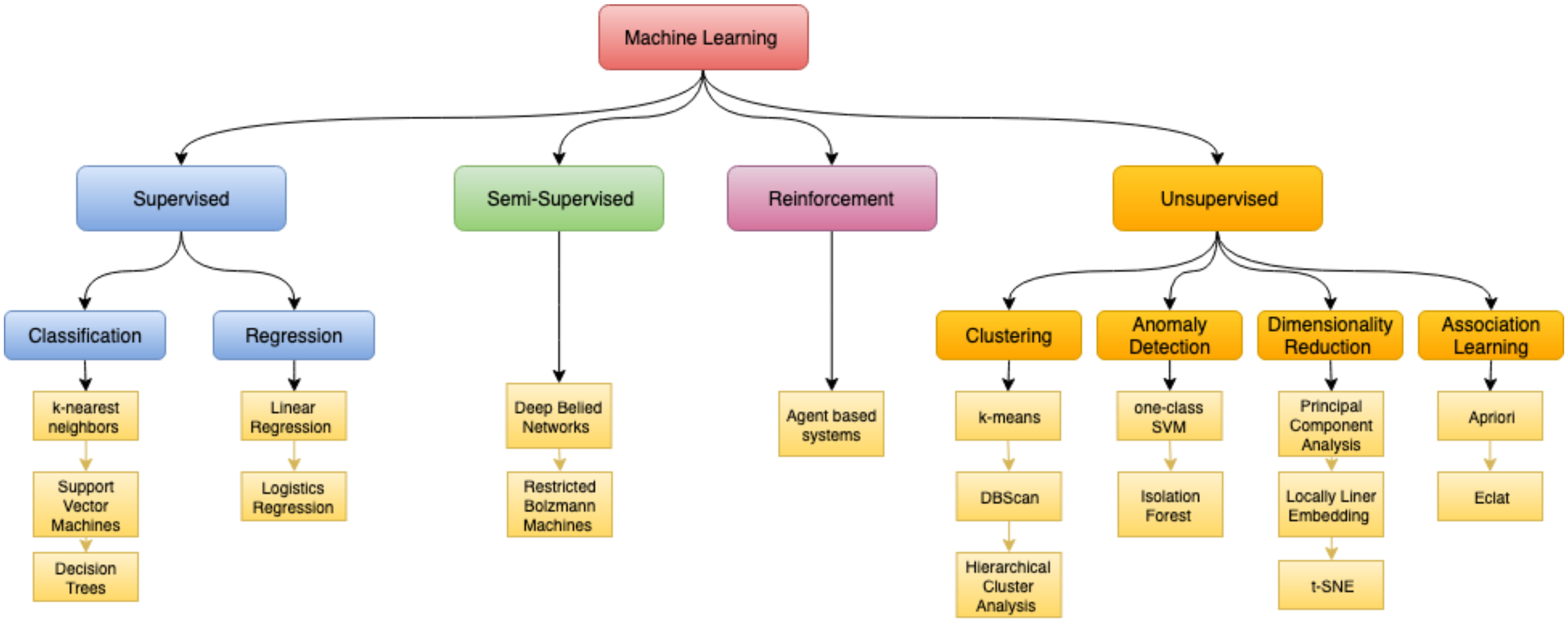
Unsupervised learning: Find patterns in unclassified data

Example: group documents based in the text

Reinforcement learning: Learn using the feedback or reward

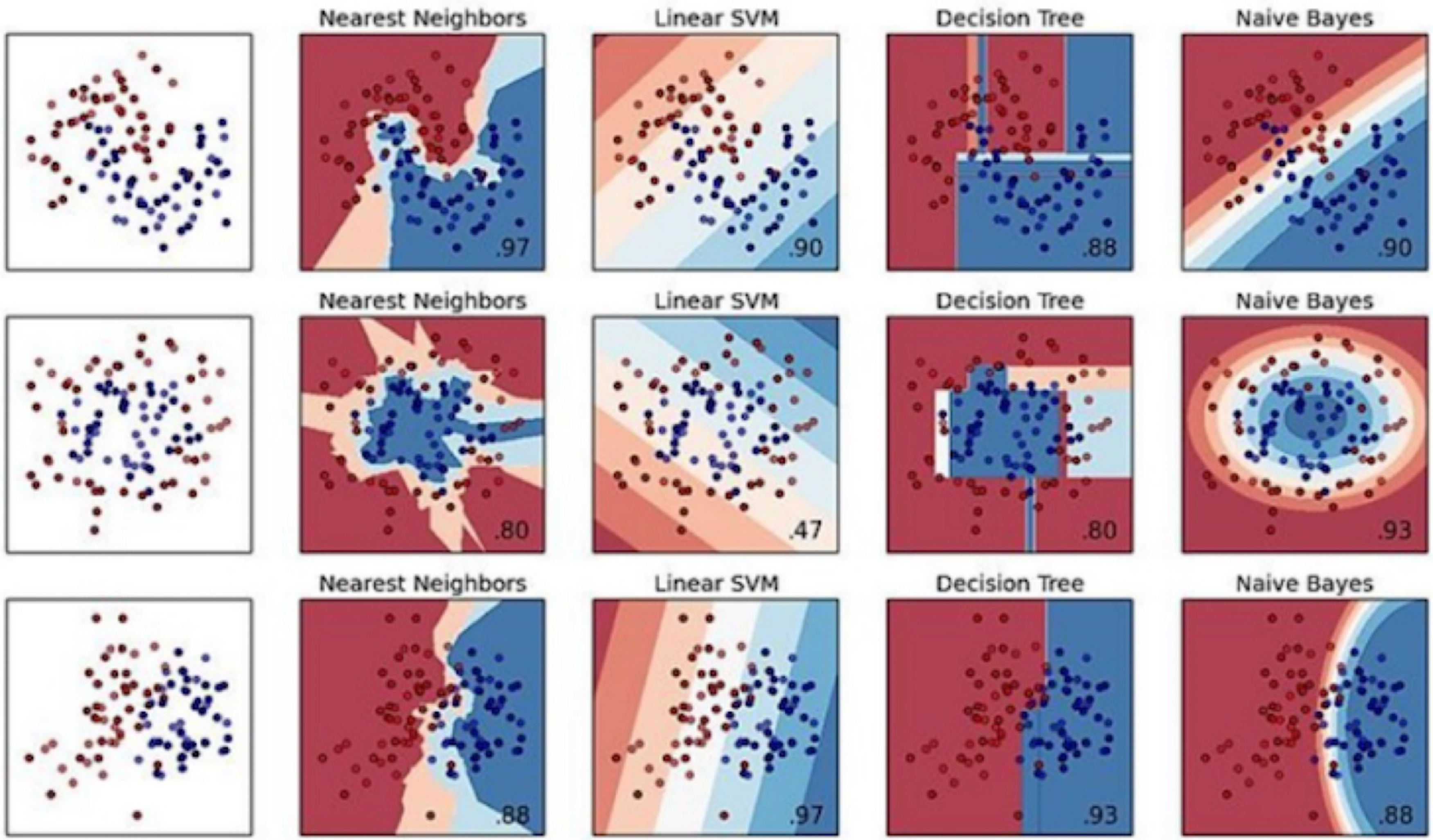
Example: Learn to play chess by previous games outcomes

Types of learning (2)

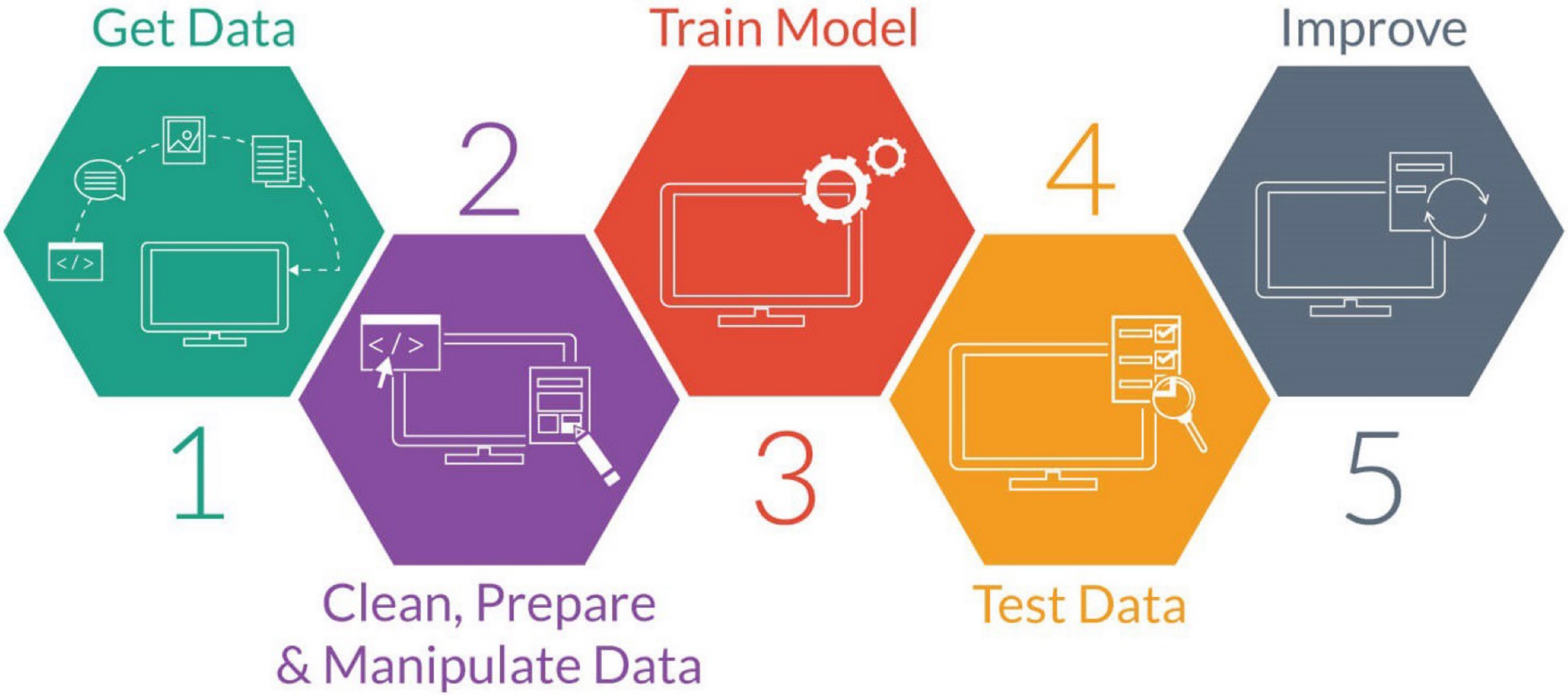


www.medium.com

Algorithms


Nearest Neighbors
Linear SVM
Decision Tree
Naive Bayes
.97
.90
.88
.90
Nearest Neighbors
Linear SVM
Decision Tree
Naive Bayes
.80
.47
.80
.93
Nearest Neighbors
Linear SVM
Decision Tree
Naive Bayes
.88
.97
.93
.88

ML Workflow



https://hazaq.me

Rules of thumb

- More data is better
- Later knowledge effects previous steps
- Expect to go backwards
- Data is never as you need it

ML Workflow – Preprocessing (1)

- Pre-processing aims to make the data **valid** and **consistent**, increasing its **quality** and also often putting it in a format where the algorithm can perform better

ML Workflow – Preprocessing (2)

- Existence of invalid data - cleaning
- Existence of too much data with too many repetitions or not informative - reduction
- Quantization and normalization
- Filtering, feature selection
- Feature extraction

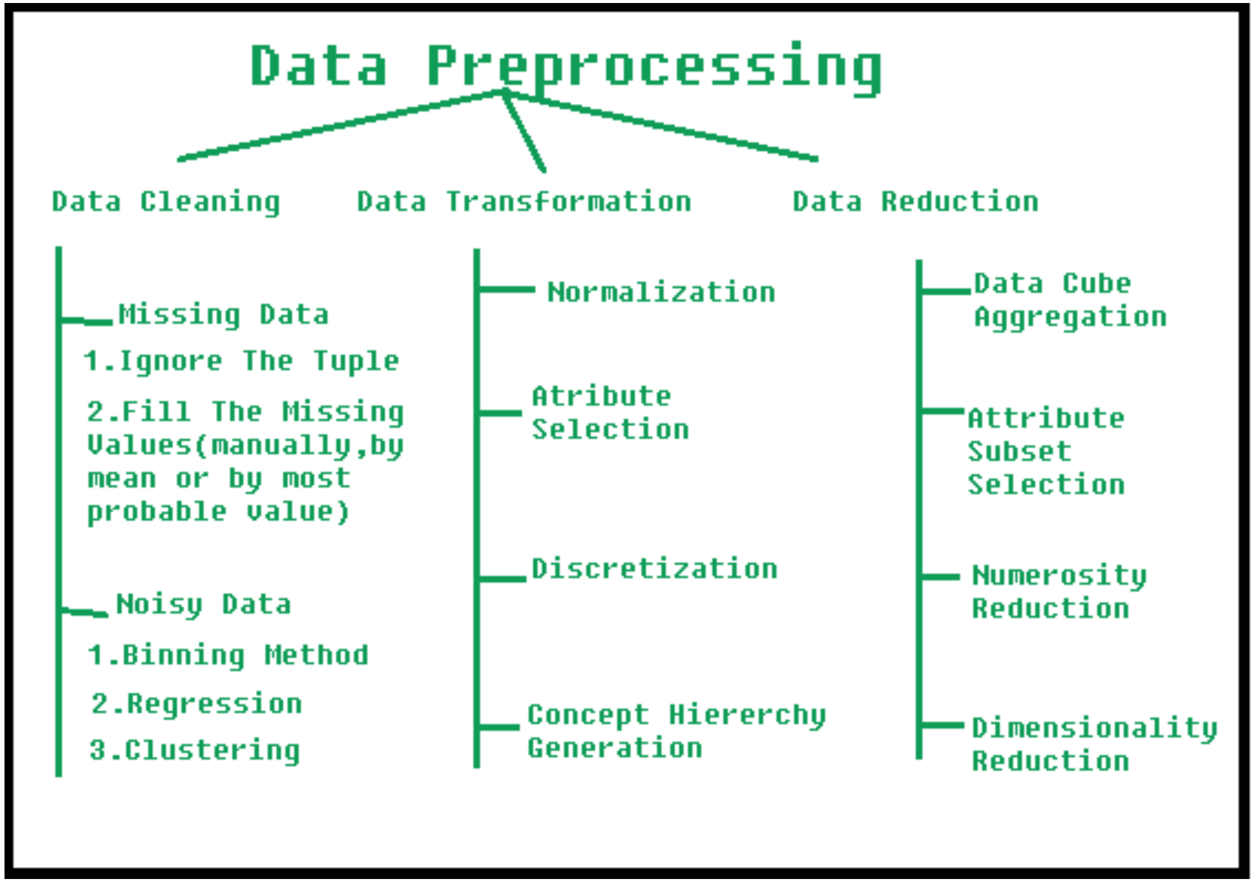
ML Workflow – Preprocessing (3)

- Data cleaning
  - Missing data (skip rows or fill values with averages)
  - Noisy data (implement outlier detectors)

- Data transformation
  - Normalization (scaling in a range, e.g. 0 to 1)
  - Attribute selection (choose the ones that contain information)
  - Discretization or binning (replacement of categories by values or ranges)
  - Hierarchical replacement of concepts (e.g. city/country)

- Data reduction
  - Data aggregation
  - Selecting a subset of attributes
  - Dimensionality reduction

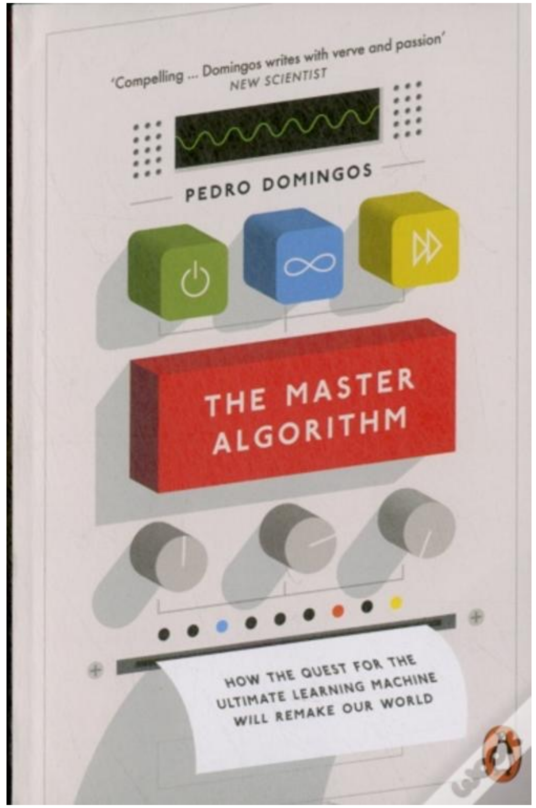
ML Workflow – Preprocessing (4)



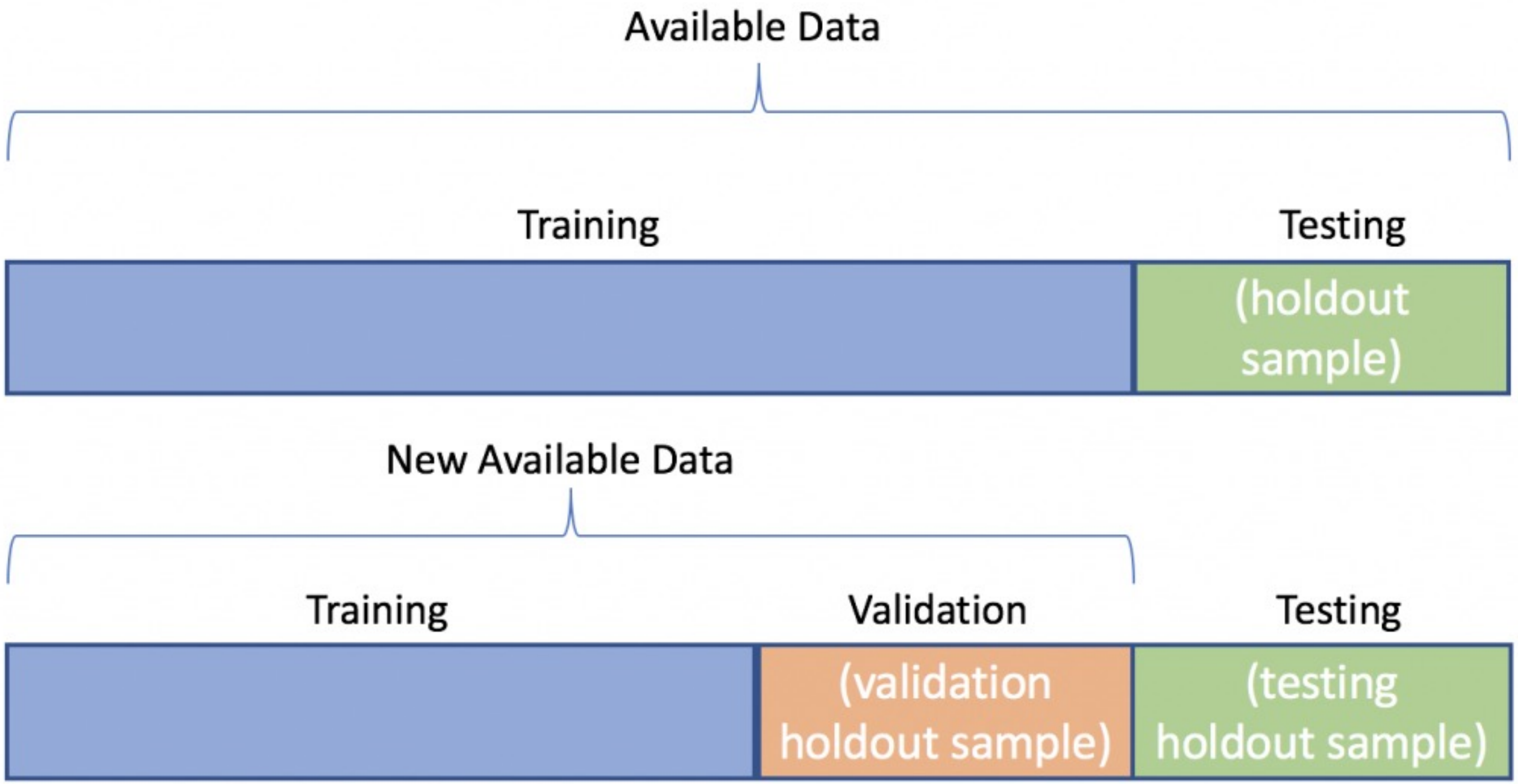
Fonte: geeksforgeeks

ML Workflow – Evaluation (1)

- A very important factor in machine learning is the evaluation of models in order to be able to compare different algorithms in different applications and choose, in an informed way, the best of them in **each situation**.
- No free lunch vs. The master algorithm



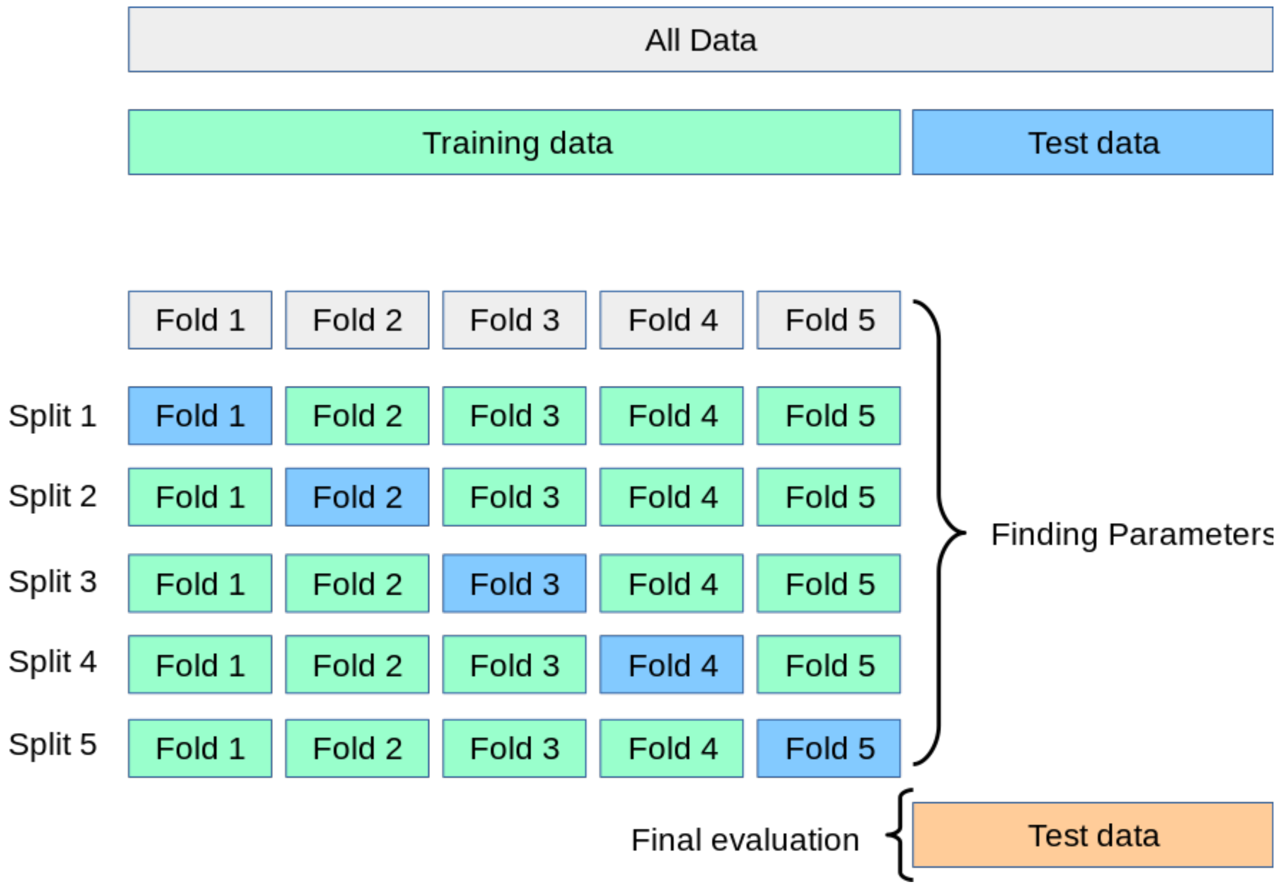
ML Workflow – Evaluation (2)



datascience.stackexchange.com

ML Workflow – Evaluation (3)

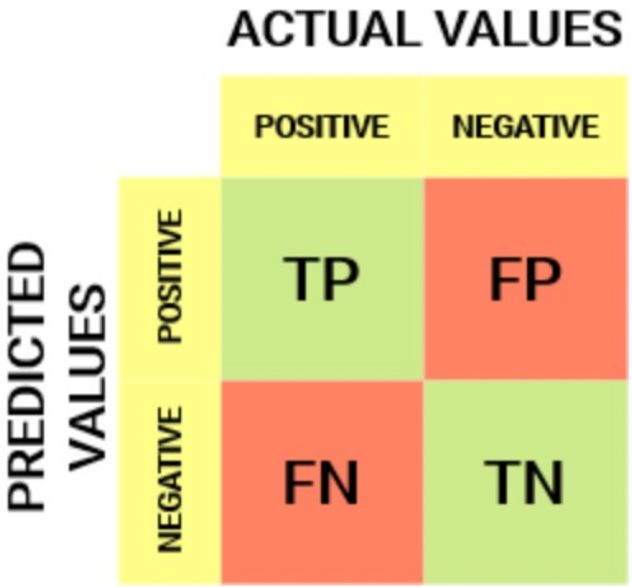
- K-fold Crossvalidation



www.scikitlearn.org

ML Workflow – Evaluation (4)

- Confusion matrix



True Positive (TP) : The predicted value matches the actual value. The actual value was positive and the model predicted a positive value

True Negative (TN): The predicted value matches the actual value The actual value was negative and the model predicted a negative value

False Positive (FP) – Type 1 error: The predicted value was falsely predicted. The actual value was negative but the model predicted a positive value

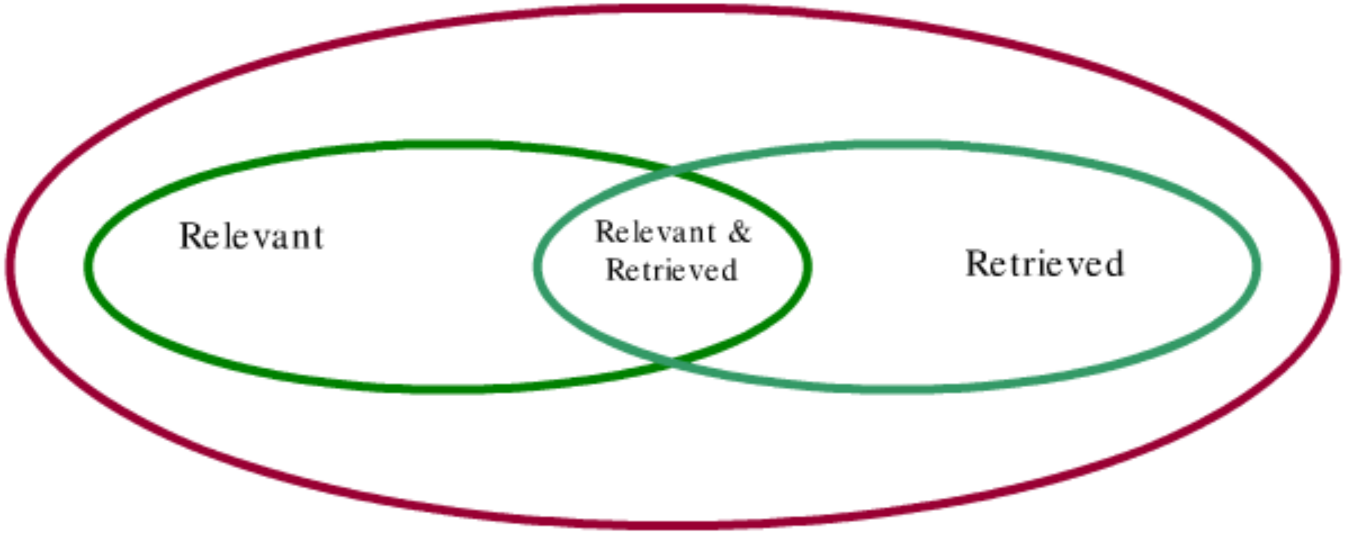
False Negative (FN) – Type 2 error: The predicted value was falsely predicted The actual value was positive but the model predicted a negative value

www.analyticsvidhya.com

Evaluation metrics

		Actual class: Positive	Actual class: Negative	
Predicted class	Positive	**TP:** True Positive	**FP:** False Positive (Type I Error)	**Precision:** $\frac{TP}{(TP + FP)}$
Predicted class	Negative	**FN:** False Negative (Type II Error)	**TN:** True Negative	**Negative Predictive Value:** $\frac{TN}{(TN+FN)}$
		Recall or Sensitivity: $\frac{TP}{(TP + FN)}$	**Specificity:** $\frac{TN}{(TN + FP)}$	**Accuracy:** $\frac{TP + TN}{(TP + TN + FP + FN)}$

https://www.analyticsvidhya.com



- Precision: the % of the retrieved items that are in fact relevant to the question (i.e., “correct”)

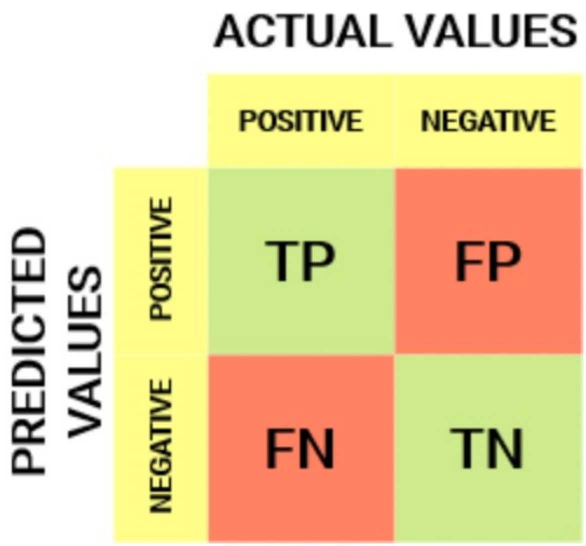
$$\text{Precision} = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}$$

- Recall: the % of items that are relevant to the question and were in fact retrieved

$$\text{Recall} = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}$$

Evaluation exercise

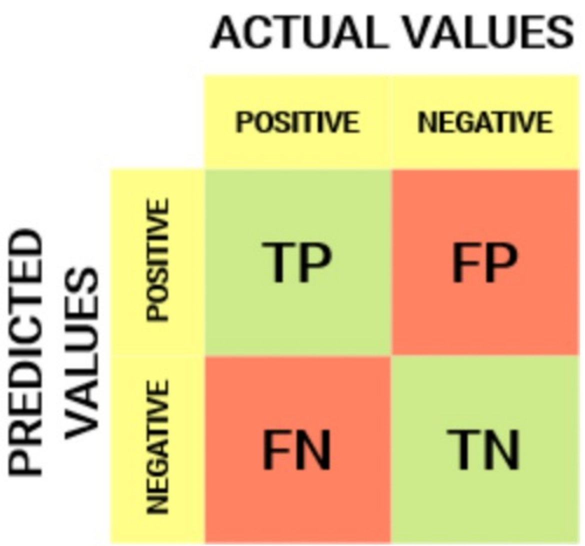
- Consider a data analytics system that is queried for a specific subject in a dataset of 1000 items where only 100 match the desired goal.
- The system returns 70 items, of which only 40 are correct.
- Construct the confusion matrix and calculate the evaluation metrics: accuracy, precision, recall, and F1



- TP =
- FP =
- FN =
- TN =

Evaluation exercise

- Consider a data analytics system that is queried for a specific subject in a dataset of 1000 items where only 100 match the desired goal.
- The system returns 70 items, of which only 40 are correct.
- Construct the confusion matrix and calculate the evaluation metrics: accuracy, precision, recall, and F1



- TP = 40
- FP = 30
- FN = 60
- TN = 870

Evaluation exercise

- Consider a data analytics system that is queried for a specific subject in a dataset of 1000 items where only 100 match the desired goal.
- The system returns 70 items, of which only 40 are correct.
- Construct the confusion matrix and calculate the evaluation metrics: accuracy, precision, recall, and F1

- TP = 40
- FP = 30
- FN = 60
- TN = 870

- Accuracy = (TP+TN)/(TP+FP+FN+TN) =
- Recall = TP/(TP+FN) =
- Precision = TP/(TP+FP) =
- F1 = 2*P*R/(P+R) =

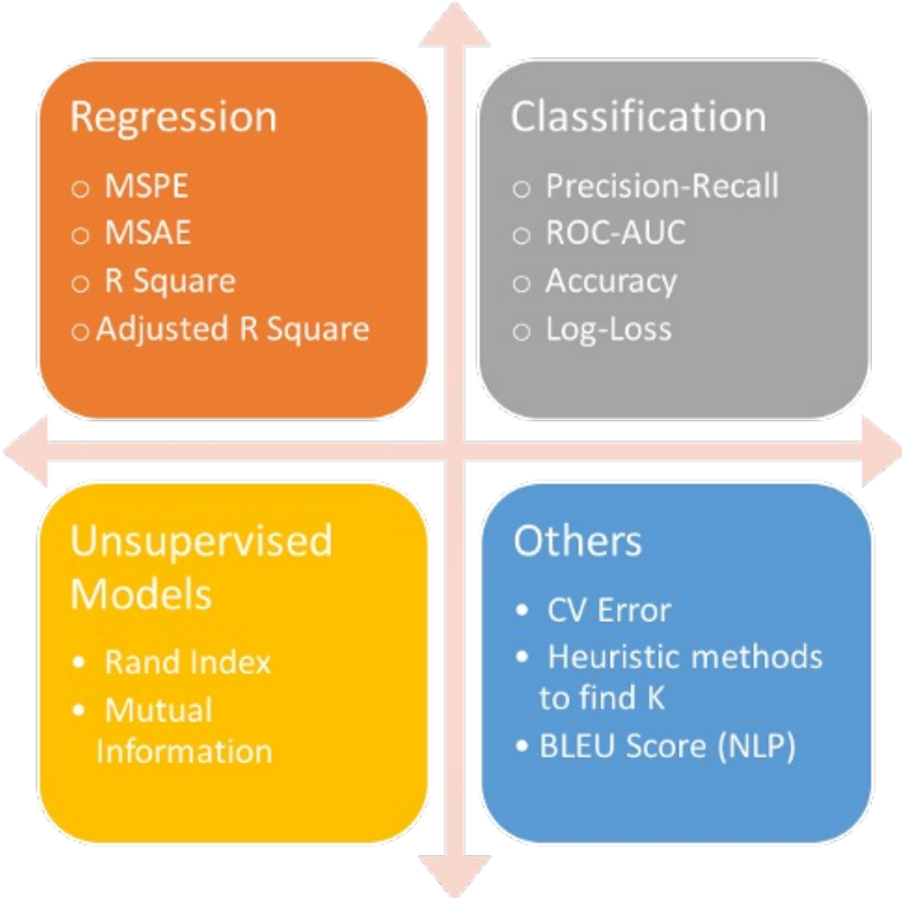
Evaluation exercise

- Consider a data analytics system that is queried for a specific subject in a dataset of 1000 items where only 100 match the desired goal.
- The system returns 70 items, of which only 40 are correct.
- Construct the confusion matrix and calculate the evaluation metrics: accuracy, precision, recall, and F1

- TP = 40
- FP = 30
- FN = 60
- TN = 870

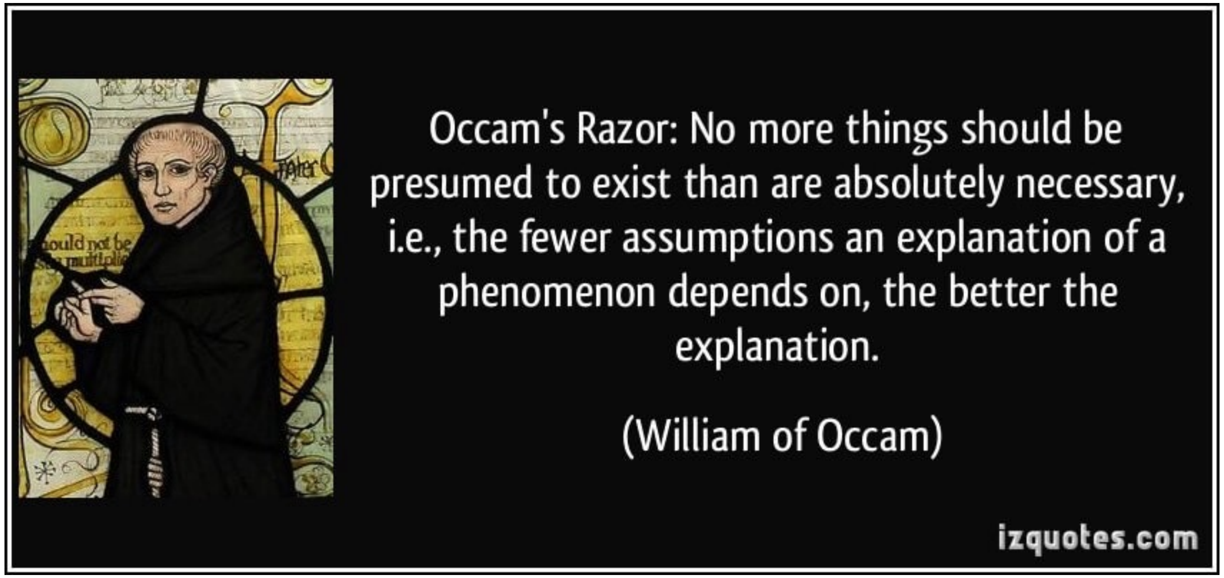
- Accuracy = (FP+FN)/(TP+FP+FN+TN) = 91%
- Recall = TP/(TP+FN) = 40%
- Precision = TP/(TP+FP) = 57%
- F1 = 2*P*R/(P+R) = 47%

Evaluation – different metrics



www.kdnuggets.com

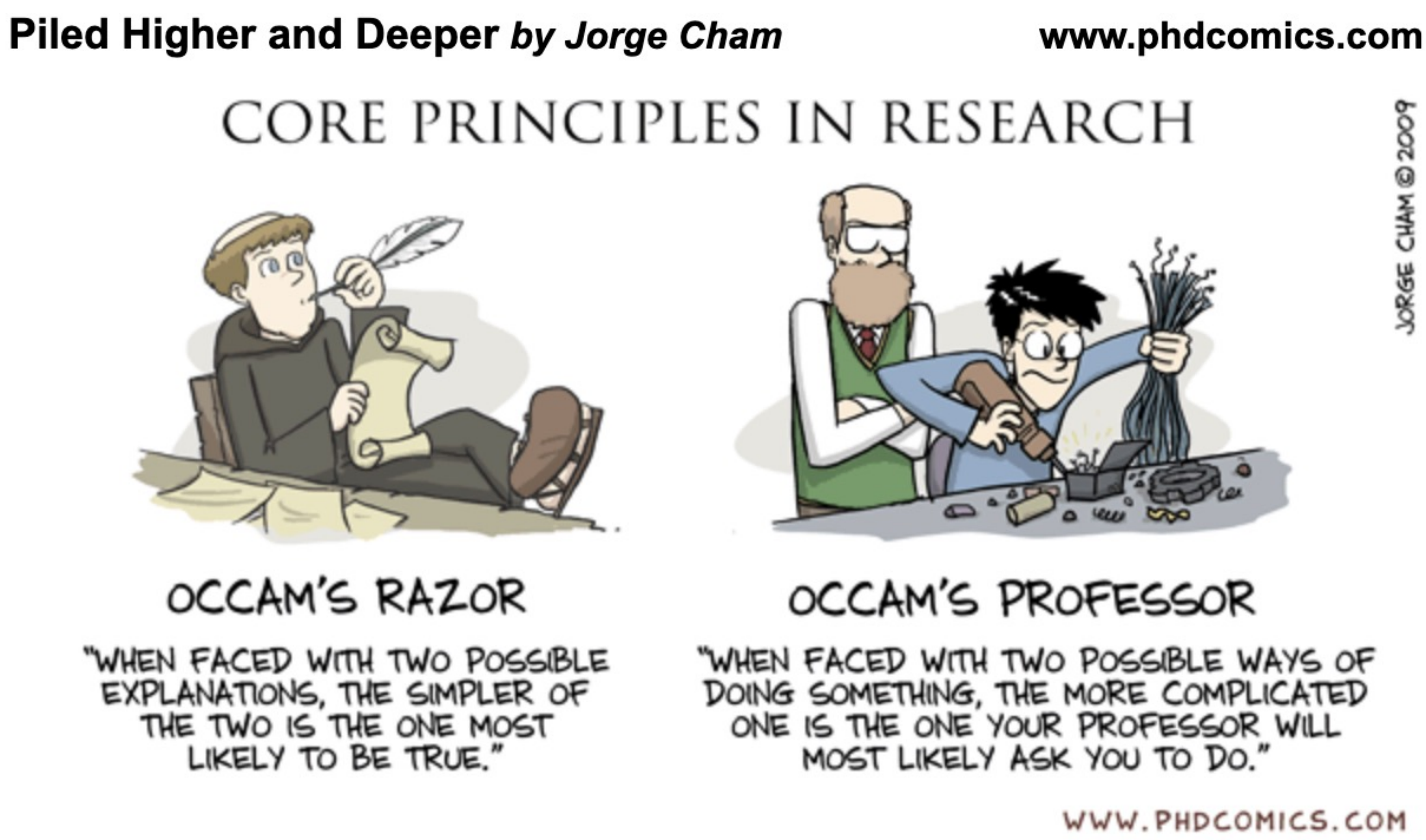
Which algorithm to choose?



Which algorithm to choose?

Occam's razor

- "All things being equal, the simplest solution tends to be the best one," or alternately, "the simplest explanation tends to be the right one." In other words, when multiple competing theories are equal in other respects, the principle recommends selecting the theory that introduces the fewest assumptions and postulates the fewest hypothetical entities. It is in this sense that Occam's razor is usually understood.
- Wikipedia



title: "Core Principles" - originally published 10/12/2009