

INTRODUCTION TO MACHINE LEARNING

CHAP 1 **Introduction**

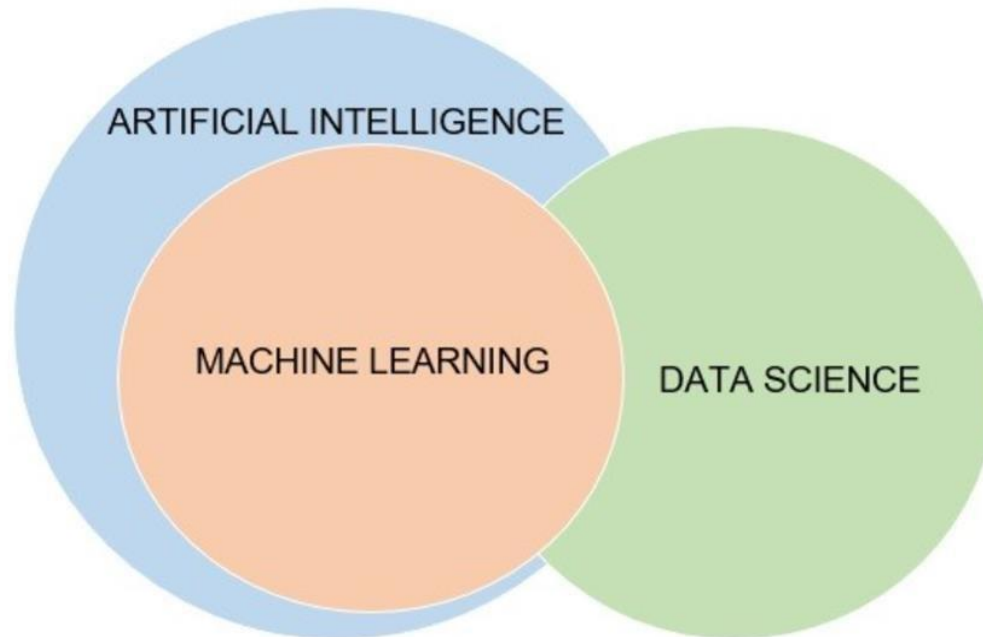
Jean Tabet
j.tabet@ul.edu.lb

→ GOALS OF THE COURSE

- Provide a general introduction to data science, data mining and statistical learning,
- Understand the kind of problems each architecture of a ML technique is useful for,
- Grading : Participation in class as well as a final project will be graded,
- Main background needed: basic notions in probabilities and statistics, programming skill.

→ GOALS OF THE COURSE

- Understand different tools behind data science,
- Distinguish between different approaches to data modeling,
- Understand what are the different problems encountered in machine learning.



➔ MAIN REFERENCES

Classical Machine Learning Textbooks:

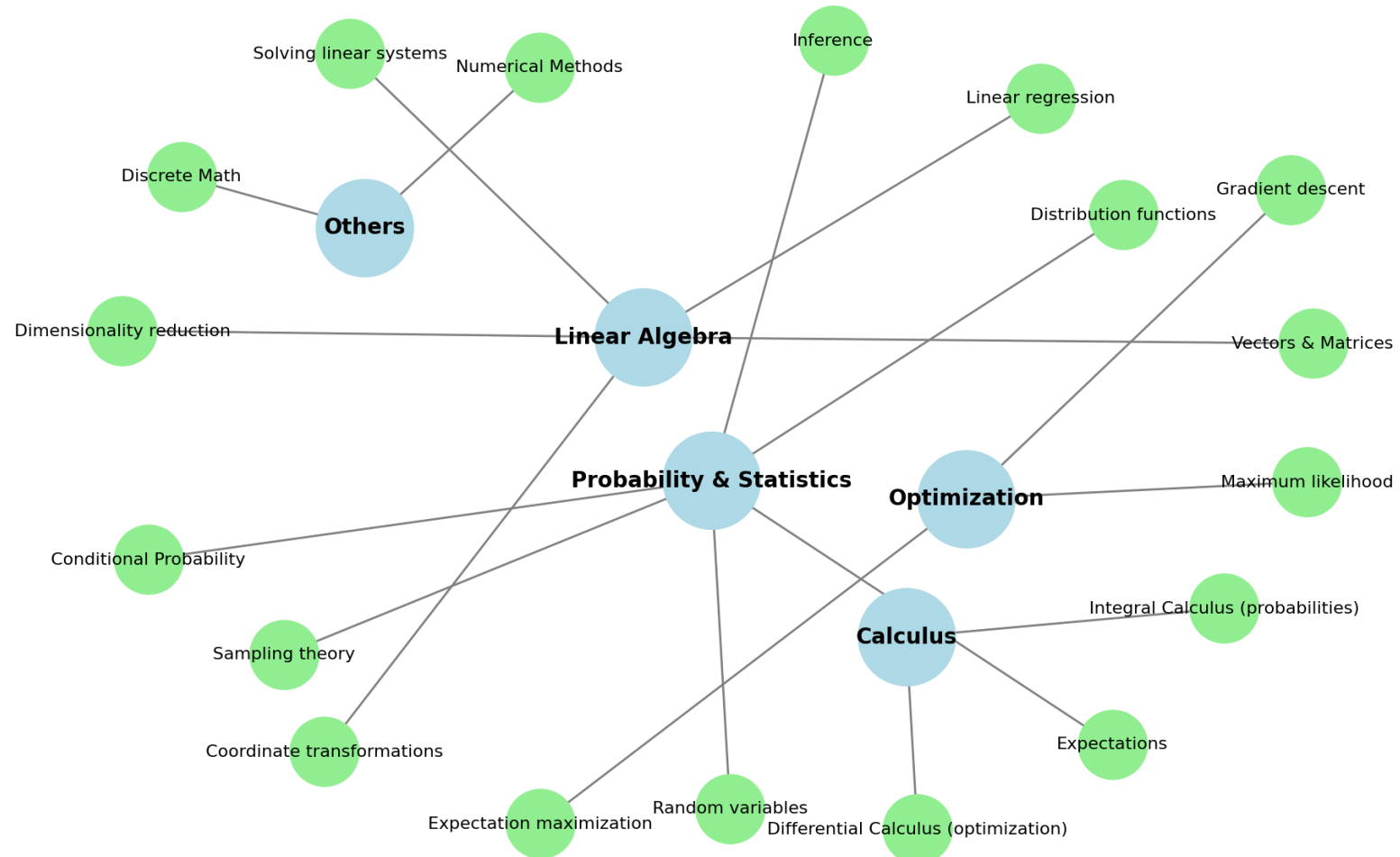
- Elements of statistical learning (ESL), [Hastie](#) et al., Springer
- An introduction to statistical learning (ISLR), [Hastie](#) et al., Springer
Both books are available online.
- Pattern recognition and Machine Learning, [Bishop](#), Springer

A lot of courses and tutorials on the web:

- Online courses Coursera, Andrew Ng ([CS229: Machine Learning \(stanford.edu\)](#)), DataCamp
- STATQUEST (Youtube)

➡ DATA SCIENCE

Mathematics for Data Science - Mind Map

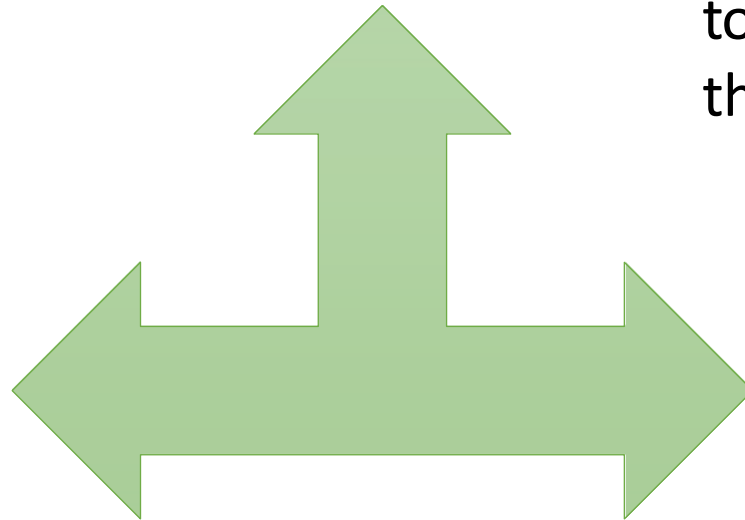


→ DATA SCIENCE

Statistician

Know and understand the techniques of modeling, data analysis, inference... knowing how to exploit regularities "Hidden" in the data, providers of knowledge.

Computer
scientist



Business
knowledge

Control the tools for manipulate the data, develop new strategies to manage the information, ...

Any analysis is part of a domain... that you need to know in order to decode and exploit the results.

“Data mining is the art of extracting knowledge from data. Data can be stored in data warehouses, in distributed databases or on the Internet. Data mining is not limited to the processing of data structured in the form of digital tables; it offers means for addressing corpus in natural language (text mining), images (image mining), sound (sound mining) or video and in this case, we speak more generally of multimedia mining.”



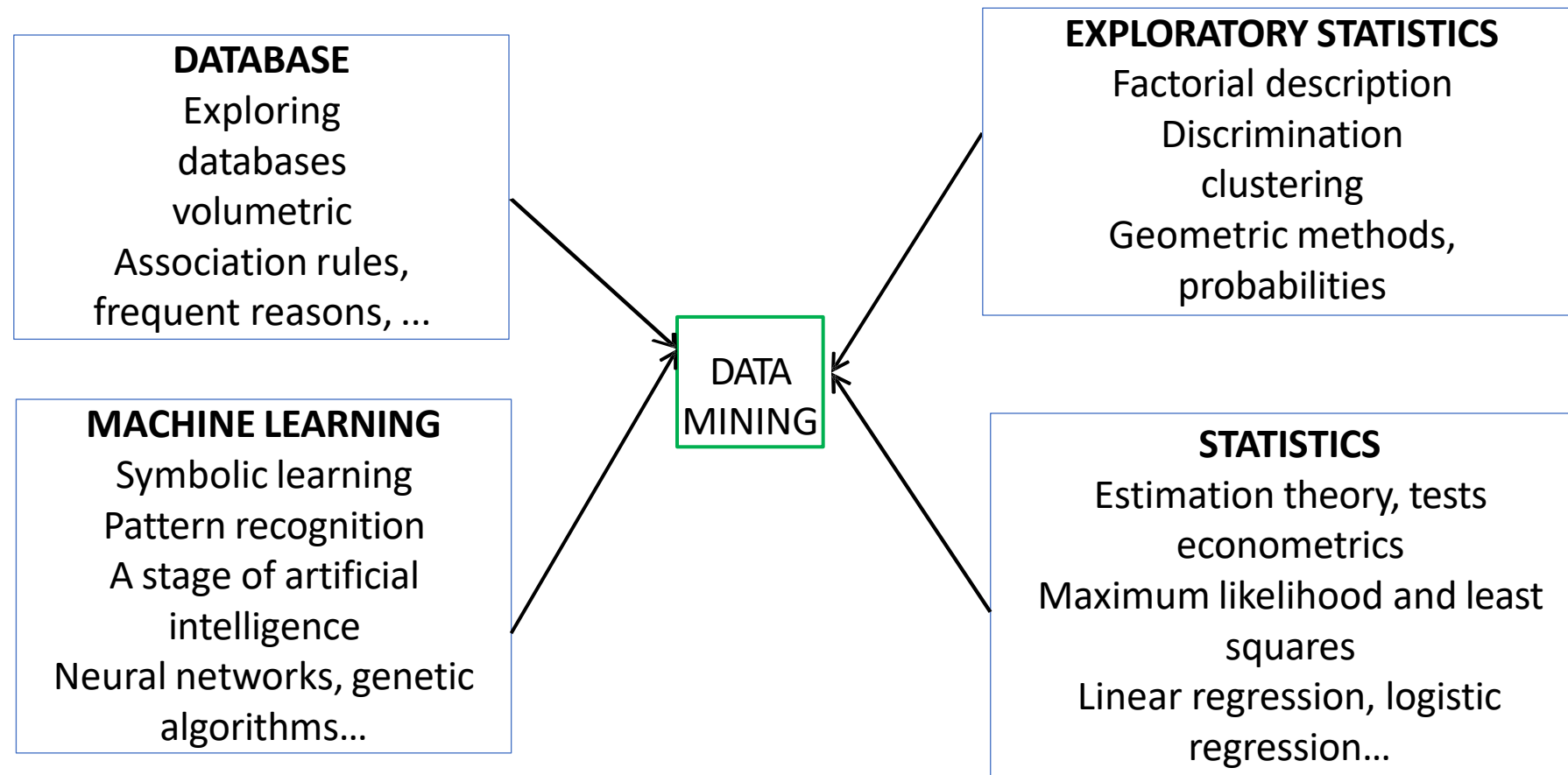
DATA MINING = KNOWLEDGE DISCOVERY IN DATABASES (KDD)

“Data science is the study of the generalizable **extraction of knowledge from data (object)**, yet the key word is science. It incorporates varying elements and builds on techniques and theories from many fields, including:

Signal processing, mathematics, probability models, machine learning, statistical learning, computer programming, data engineering, pattern recognition and learning, visualization, uncertainty modeling, data warehousing, and high-performance computing...”

→ Dual **statistical** and **computer** skills

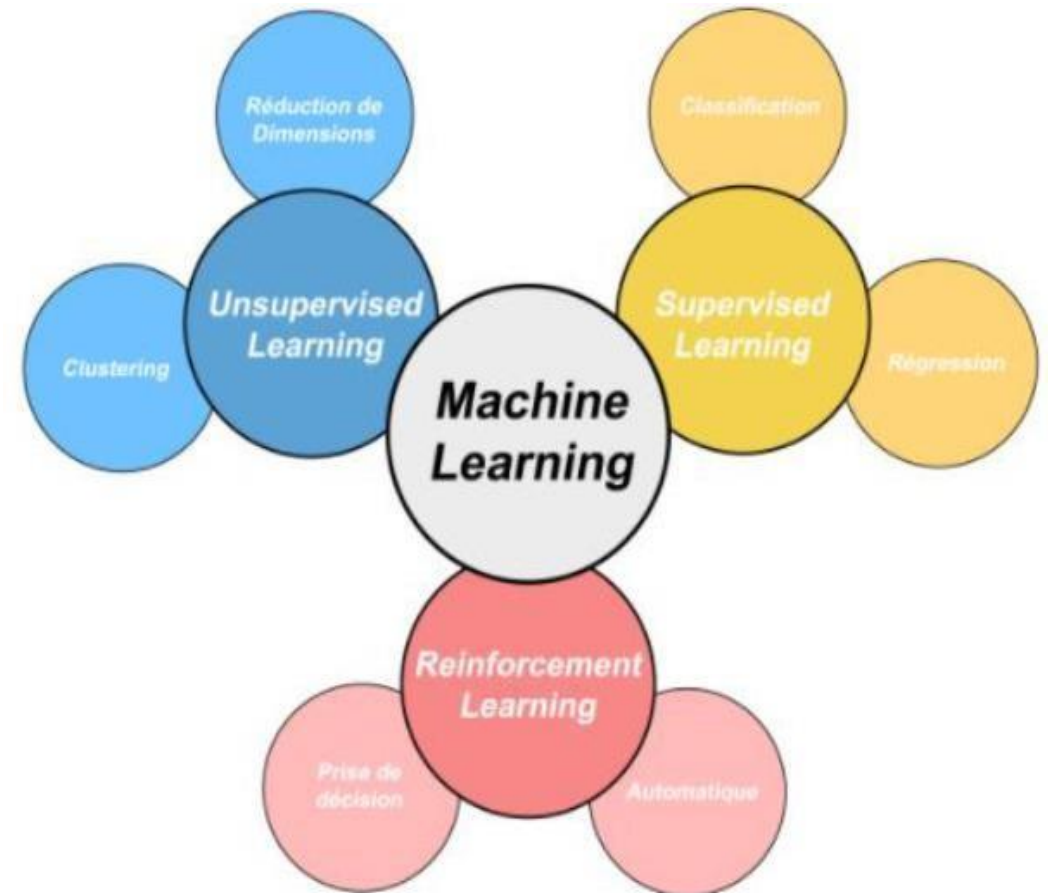
→ DATA MINING = UNIONS OF VARIOUS DISCIPLINES



➔ MACHINE LEARNING

"Give the ability of machines to learn without explicitly programming them " Arthur Samuel, 1959

- Implementation of algorithms having for objective of obtaining predictive analysis from data, for a specific purpose,
- Is said also 'Statistical Learning' because most of the algorithms used are in reality of statistical models developed from data;



➔ MACHINE LEARNING

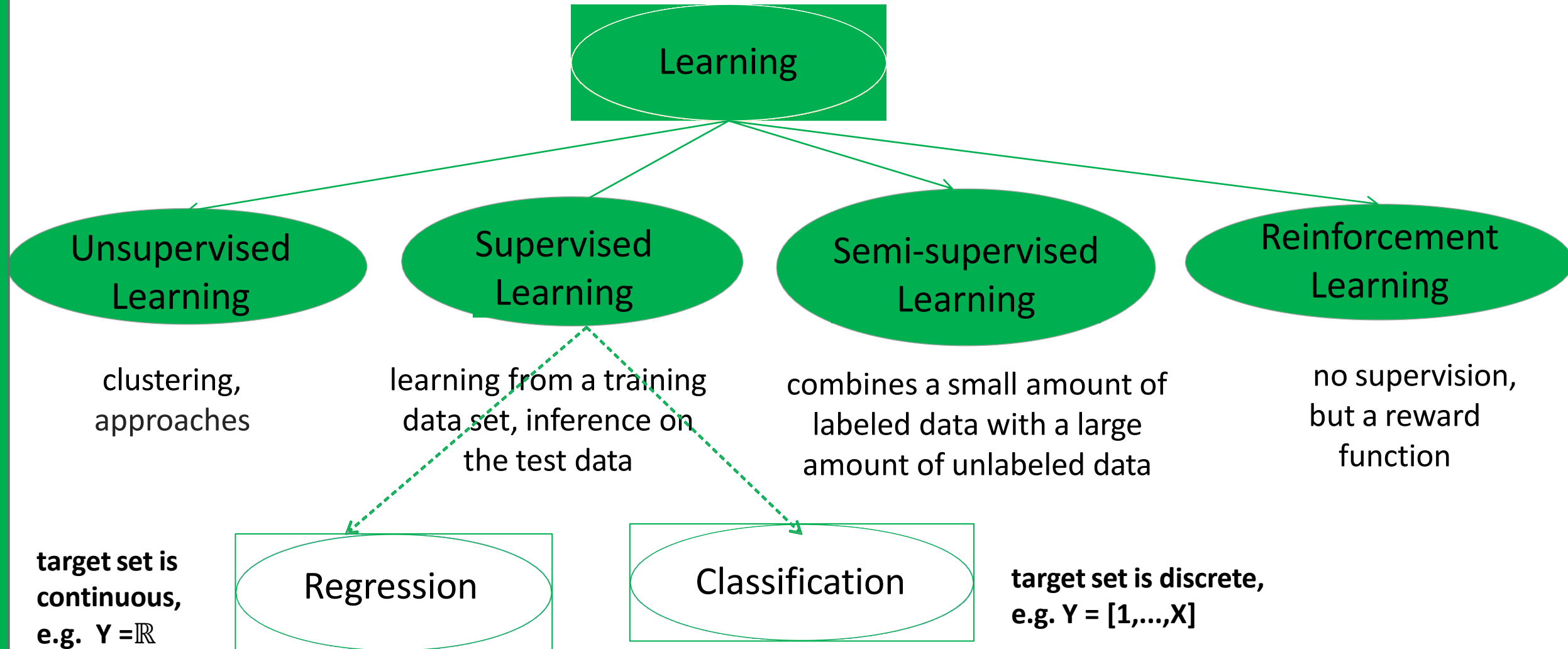
- Methods for predicting, detecting, or grouping data samples based on a model.
- The model must be learned with data.
- Methods can be geometrical(or not) and the model is based on distance metrics or linear boundaries.

➔ PROBLEMS THAT ML SOLVES

- Face recognition
- Fake news detection
- Sentiment analysis
- Recommender systems
- Fraud detection
- Effective web search
- Social networks recognize friends from photos or suggest friends
- Email spam detection
- Handwriting recognition

Ect.;

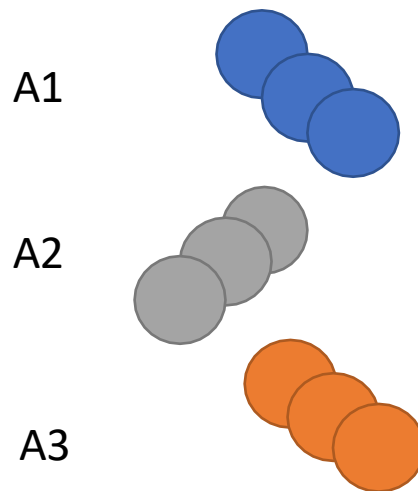
→ MACHINE LEARNING



➔ MACHINE LEARNING

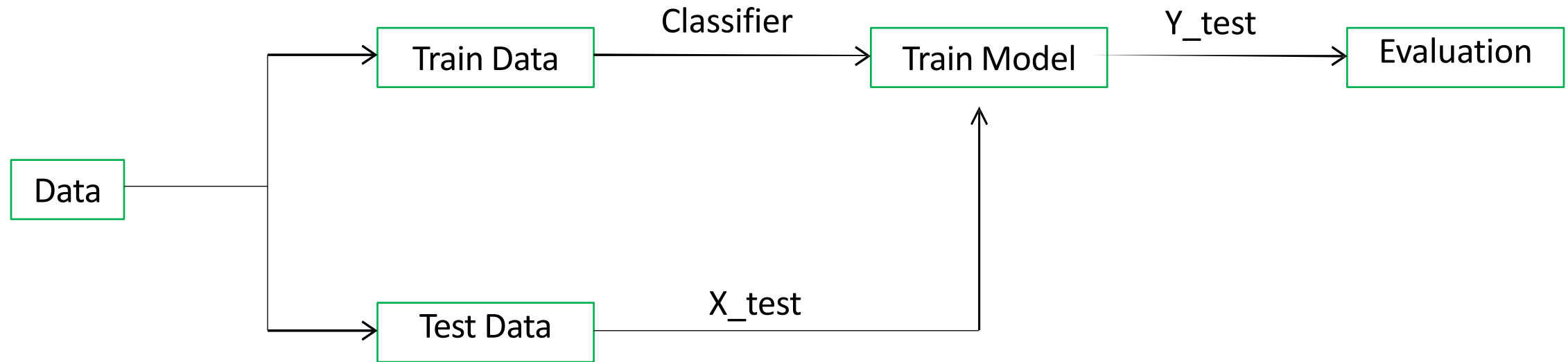
Supervised learning : knowledge of output : learning with the presence with an expert / teacher,

- Data Is labelled with a class or value,
- Goal: predict a class or value label (Neural Network, Support Vector Machine, Decision Trees, Bayesian classifiers...),
- Examples:
 - Face recognition
 - Spam detection



➔ MACHINE LEARNING

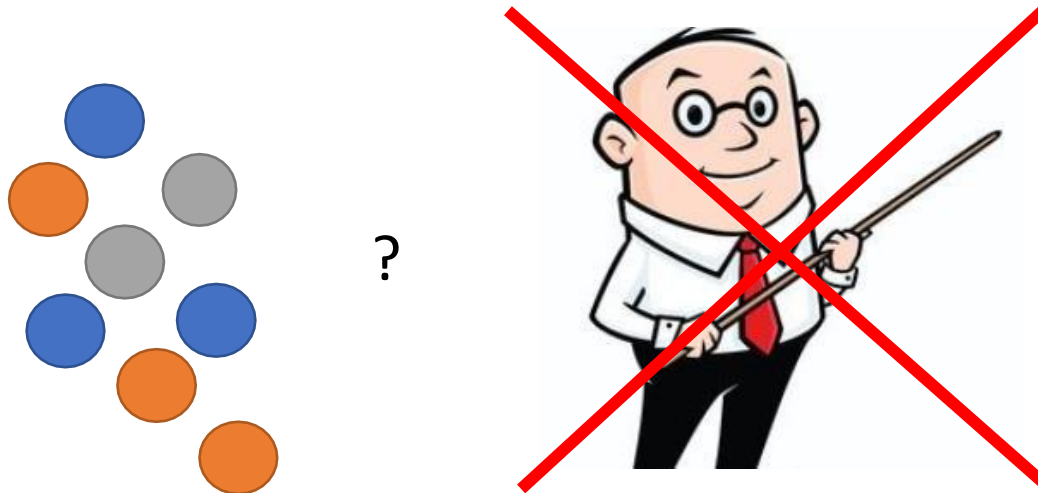
Supervised - Classifiers



→ MACHINE LEARNING

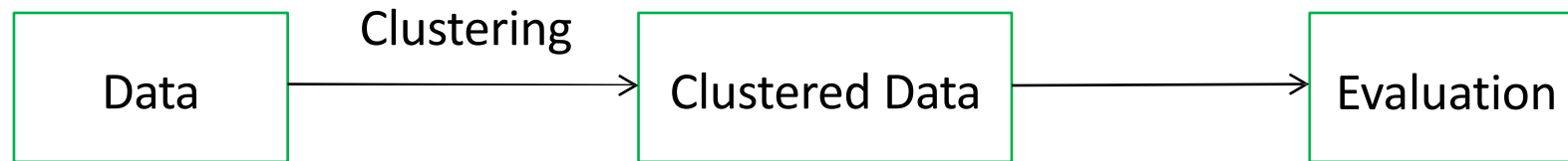
Unsupervised learning: no knowledge of class output or value,

- Data is unlabeled with a class and value are unknown,
- Goal: determine data patterns – groupings,
- Self-guided learning algorithm (k-means, clustering approaches...),
- Example: user behavior analysis.



➔ MACHINE LEARNING

Unsupervised - Clustering

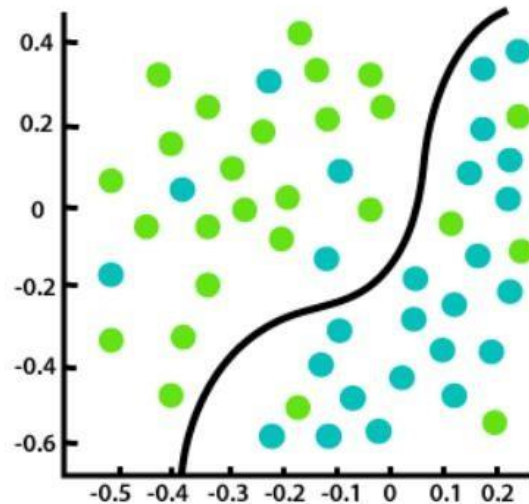


→ REGRESSION OR CLASSIFICATION

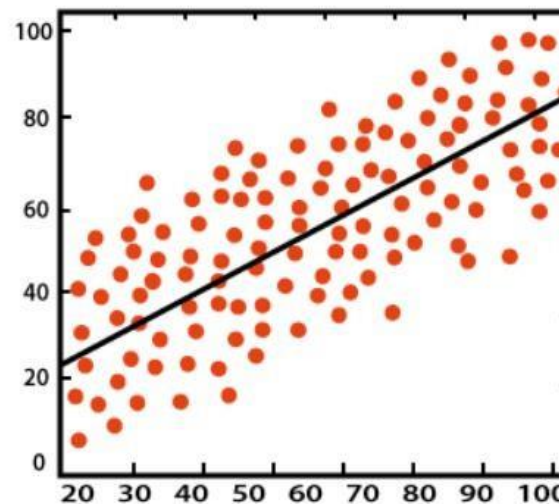
Another distinction that will help you in choosing a machine learning algorithm is **the type of output that is expected from our program**: Is it a **continuous value** (a number) or a **discrete value** (a category)? The first case is called a regression (like in fraud detection systems), the second a classification (like in web data classification).

Ex:

if a mail is a
spam or not



Classification



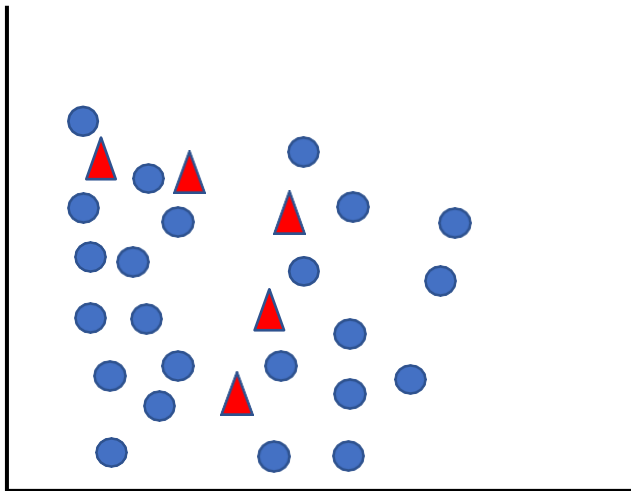
Regression

Ex:

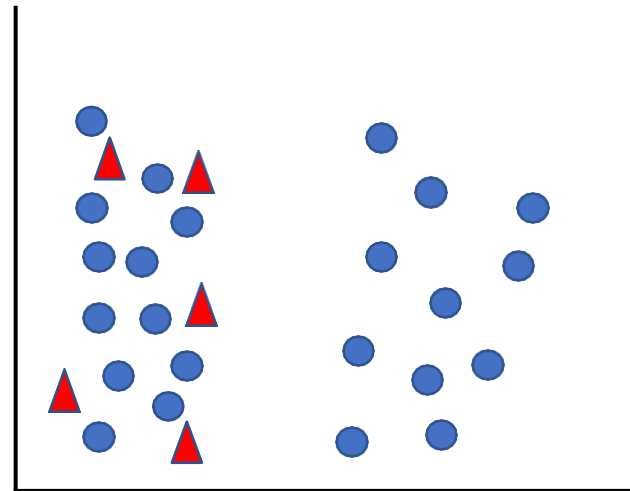
cost of a web
advertisement

→ MACHINE LEARNING

Semi-supervised learning takes as input **some annotated data** and **others not**. These are very interesting methods that take advantage of both worlds (supervised and unsupervised), but of course bring their share of difficulties.



Labeled and unlabeled data



Semi-supervised learning

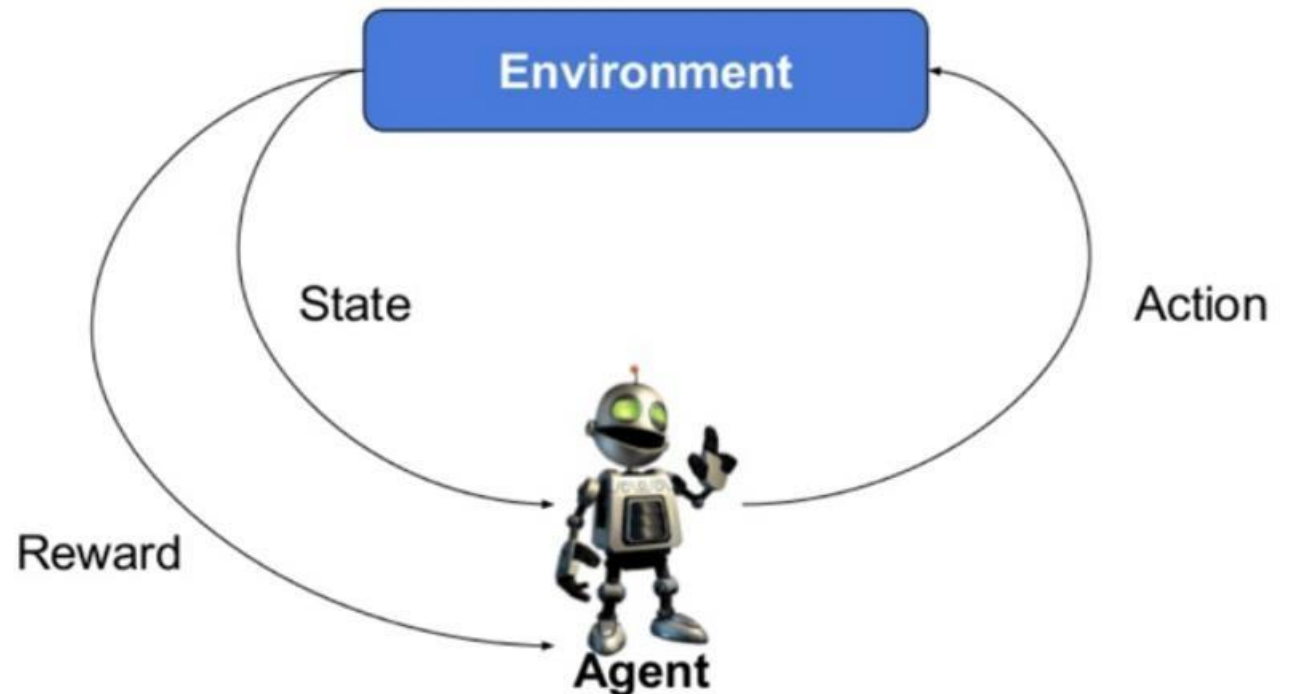
➔ MACHINE LEARNING

Reinforcement learning is based on an **experience/reward** cycle and improves performance with each iteration. An analogy often cited is that of the dopamine cycle: a "good" experience increases dopamine and therefore increases the probability that the **agent** repeats the experience.

Example:

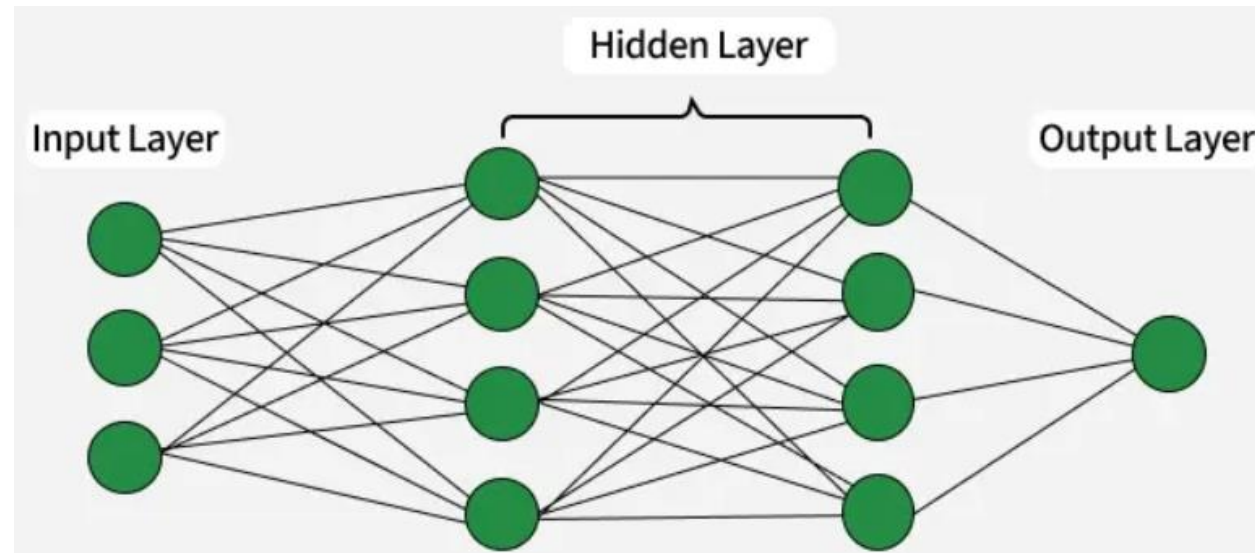
Intelligent gaming programs (e.g., Chess, ...)

Autonomous Vehicles: Making driving decisions like steering and braking



→ DEEP LEARNING

- Uses artificial neural networks to learn patterns from data.
- Batch processing for big data.
- Better for complex tasks like image and text processing, Image recognition, NLP, speech recognition.
- Needs GPU and high-performance systems



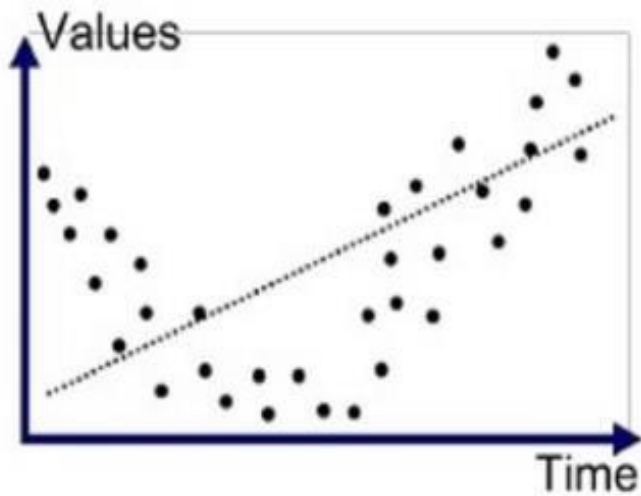
→ SUMMARY

Machine learning problems can be differentiated according to two criteria:

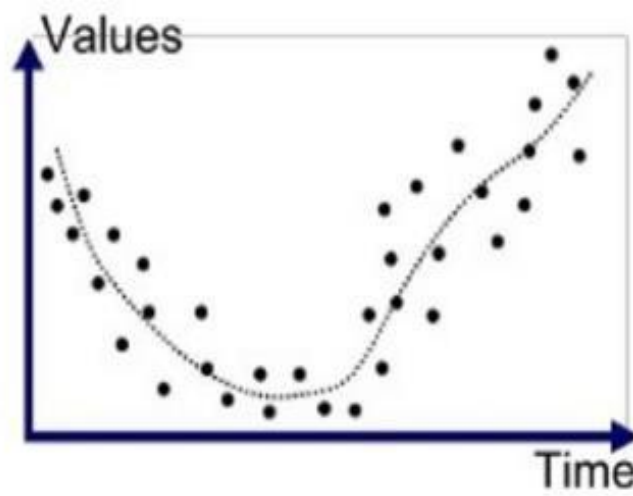
- Is the data you have annotated or not? If so, you are dealing with a **supervised learning** problem. Otherwise, you will be forced to use an **unsupervised learning** algorithm.
- What type of outcome do you want to predict? If it is a number (for example the cost per click of an advertisement), it is a **regression** problem. If it is rather a discrete value, a category (for example the type of an email) then it is a problem of **classification**.
- Learning is usually a two-step process consisting in a **training** and an **inference** step.

→ INCONSISTENCIES IN DATA

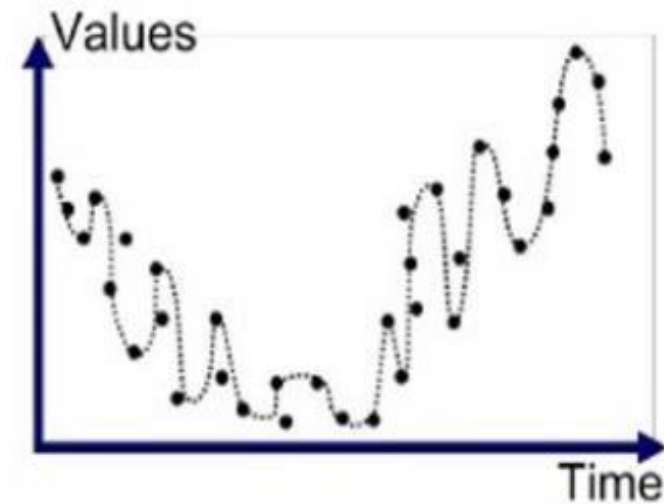
In the training phase, a ML model may or may not generalize perfectly:



Underfitted



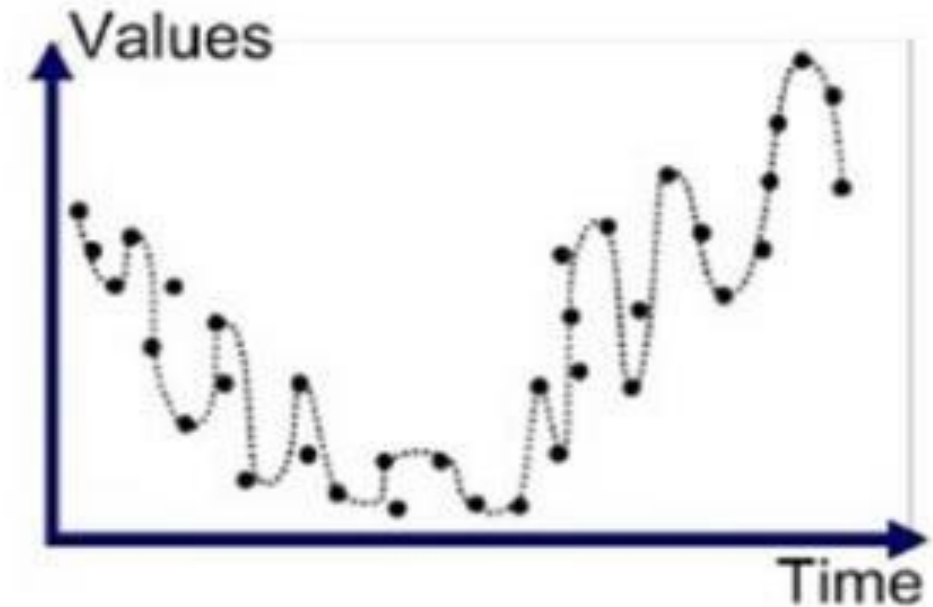
Good Fit/Robust



Overfitted

→ OVERFITTING

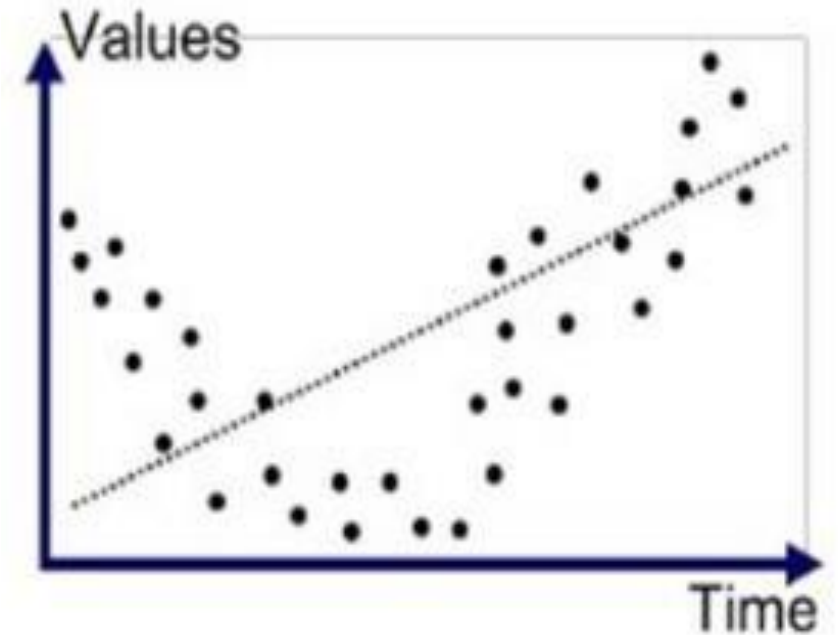
- The model does not give any error compared to the dataset, so it should be perfect! However, it doesn't generalize well from our training data to unseen new data.
- This is known as **overfitting**, and it's a common problem in machine learning and data science.



Overfitted

→ UNDERFITTING

- The predictive model generated during the learning phase does not adapt well to the training set.
- The predictive model does not even manage to capture the correlations of the training set. Therefore, the cost of error in the learning phase remains great.
- In this case, we say that the model suffers from **underfitting**. He is also said to be suffering from a severe Bias (bias).



Underfitted

➔ PERFORMANCE TUNING

- The most important iteration for a ML systems as it helps to improve their performances.
- The generalization error is the accuracy of classification or the error in forecast of regression model.
 - Mean Squared error,
 - Mean absolute error,
 - Precision, recall and accuracy,
 - Confusion matrix etc.

→ PERFORMANCE TUNING

How can model performance be improved ?

- Fetching the data to improve performance: fetching more data!
- Switching ML algorithms: try the right choice of algorithm!
- Ensemble learning: ensembling the performance of multiple algorithms!