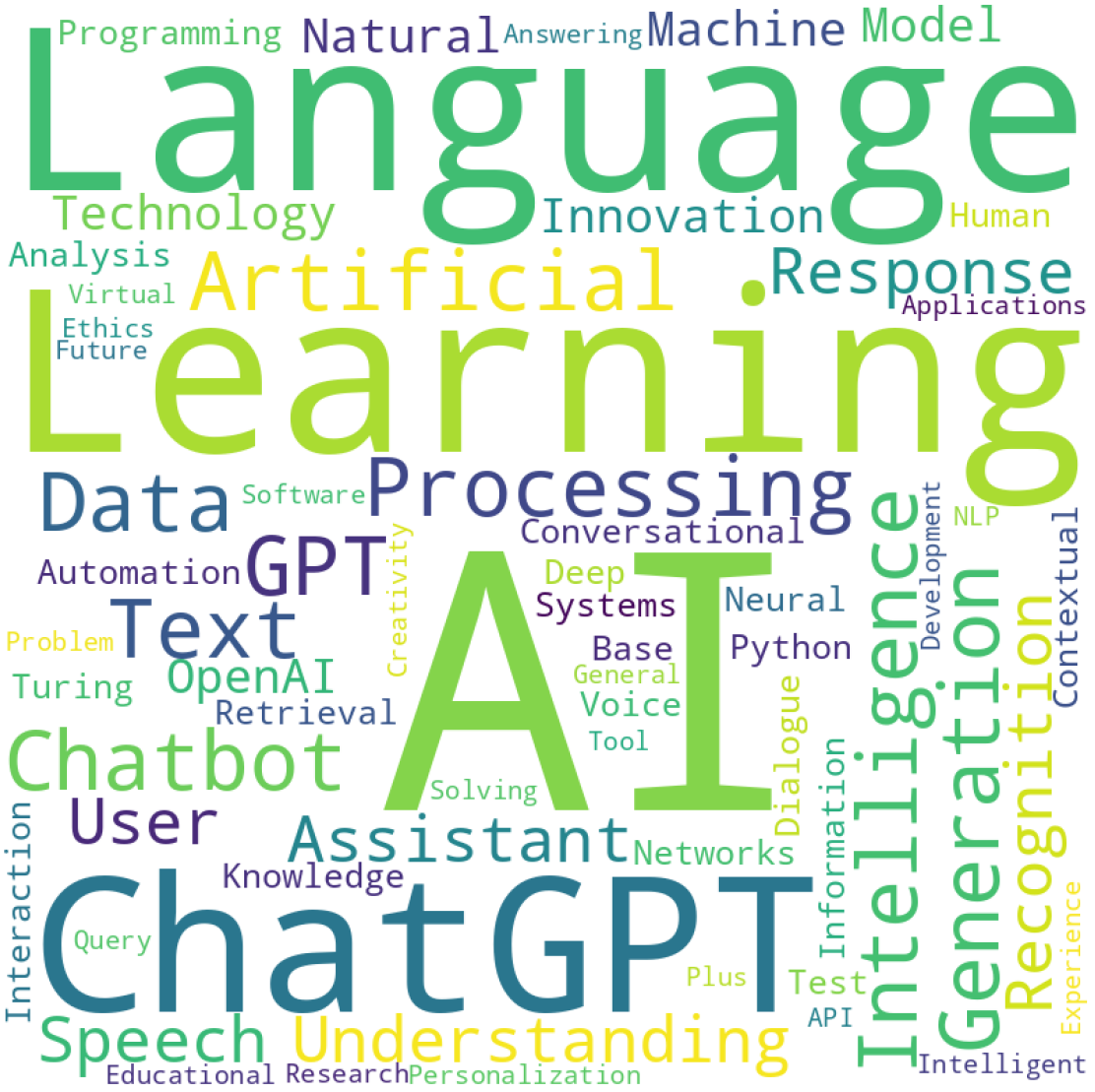
Introduction to Data Science and data Mining

STATISTICA NUMERICA - A.Y. 2022/2023

1 – Data science

A VERY SHORT TERMINOLOGY

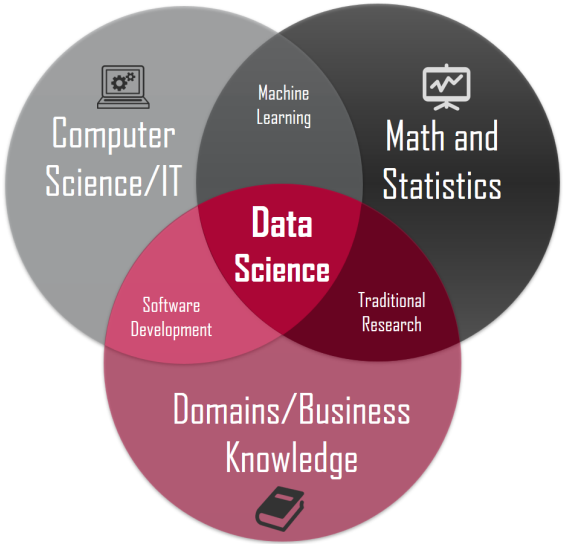
What do you know about Data Science?

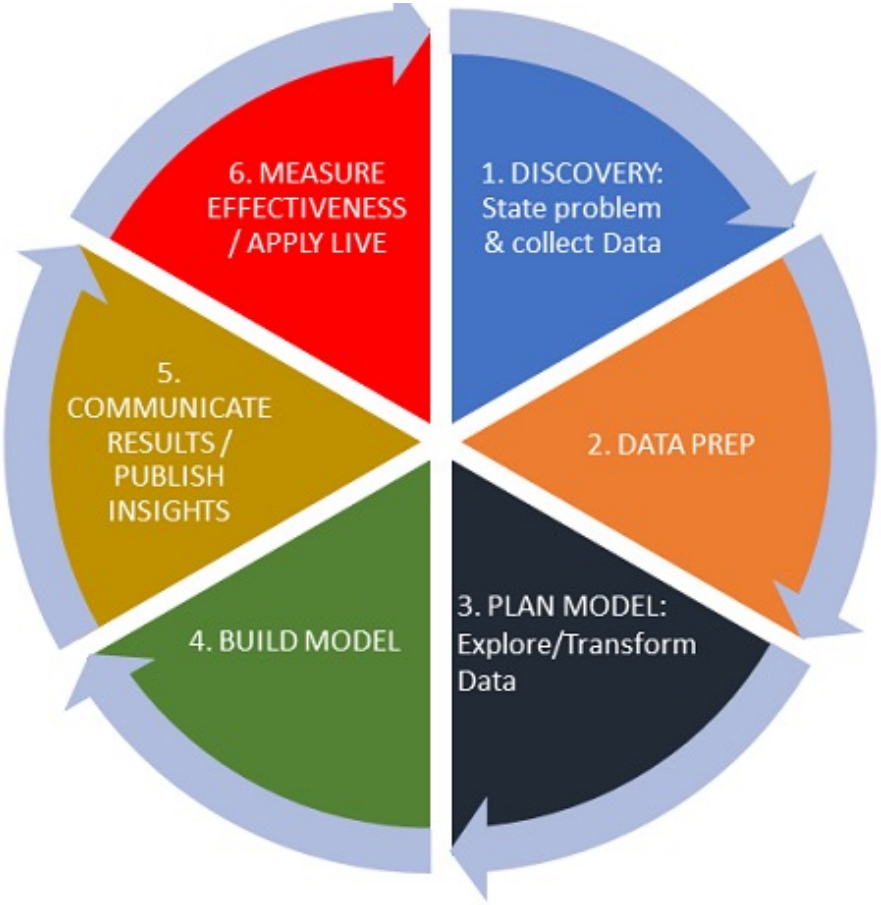


Data Science

DS IS A MULTIDISCIPLINARY SUBJECT

- It starts from user-specified objectives
- It exploits algorithms to extract patterns and models, with a mathematic approach





Data Science life cycle

Why/What Data Mining

We have a lot of ***data*** (a collection of raw value elements)

It is quite easy to extract ***information*** (the result of collecting and organizing data) such as

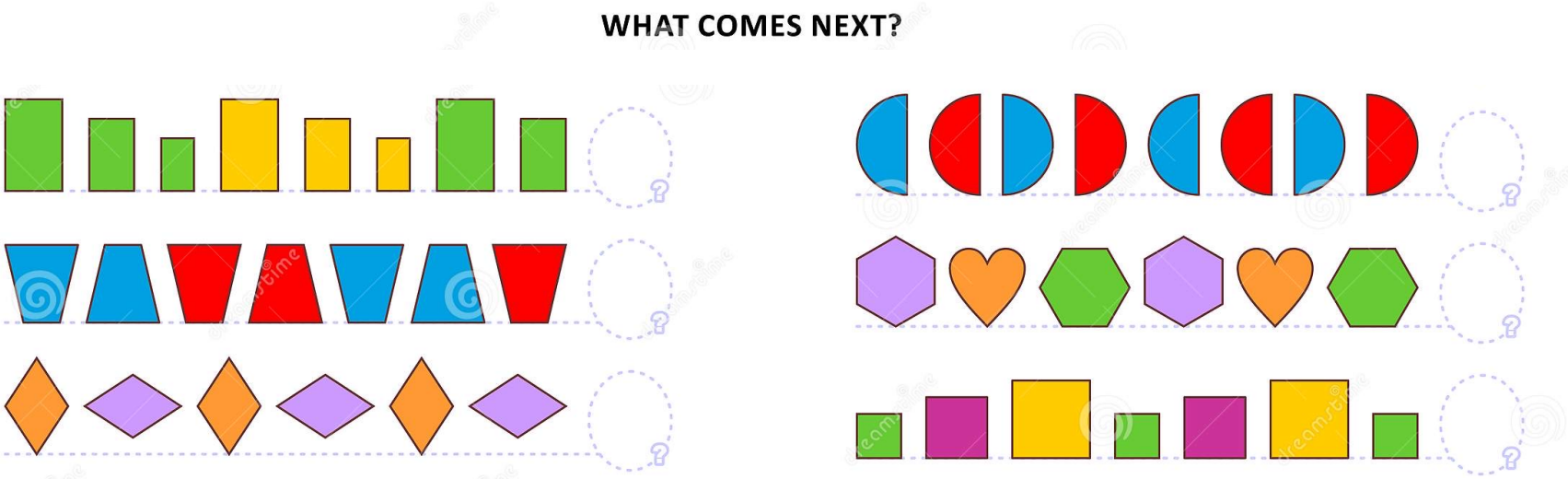
- relationships between data items
- context and meaning

We aim to something higher than information:
we need to shift from data to knowledge

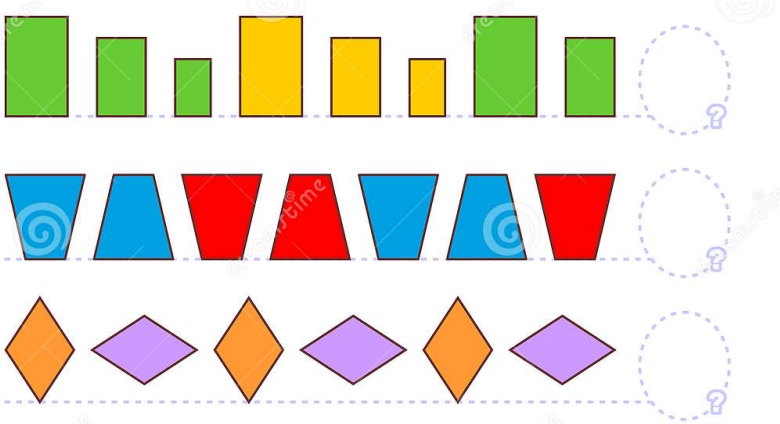
Knowledge: understanding information based on recognizing patterns

Curiosity: the DM name is wrong, because we are not mining for *data*, which are already there and available, but for *patterns!*

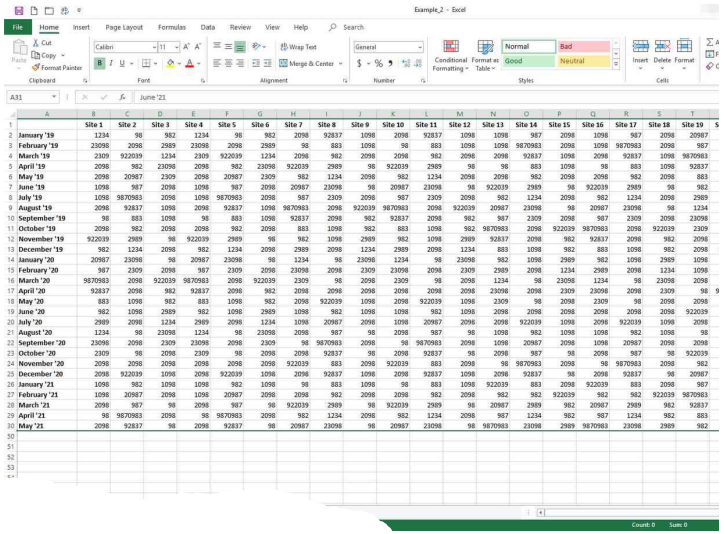
Toy examples of «mining for patterns»



Why are we interested in patterns?
Because they allow for (future) *predictions*!



VS



How can we mine for patterns on huge data sets?
With algorithms running on computers!

Data structures

There are three common types of data structures, for data analysis:

1. **Unstructured data** is information that either does not have a predefined data model or is not organized in a pre-defined manner.
2. **Semi-structured data** is a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contain tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data.
   Examples include JSON and XML.
3. **Structured data** is data that adheres to a pre-defined data model and is therefore straightforward to analyze. Structured data conforms to a tabular format with relationship between the different rows and columns.
   Examples: Excel files or SQL databases. Each of these have structured rows and columns that can be sorted.

Structured data

A **comma-separated values** (**CSV**) file is a delimited text file that uses a comma to separate values. Each line of the file is a data record.

- Each record consists of one or more fields, separated by commas.
- A CSV file typically stores tabular data (numbers and text) in plain text, in which case each line will have the same number of fields.
- CSV is a common data exchange format that is widely supported by consumer, business, and scientific applications.

A **database** is an organized collection of data and it can handle very complicated queries.

- Relational databases became dominant in the 1980s. These model data as rows and columns in a series of tables, and the vast majority use SQL for writing and querying data.
- In the 2000s, non-relational databases became popular, referred to as NoSQL because they use different query languages.

Characteristics of data sets

1. Dimensionality (number of columns)
   - the difference between having a small or a large (hundreds, thousands, . . .) of attribute is also qualitative
   - see the curse of dimensionality, later
2. Sparsity (when there are many missing values)
   - Nulls in disguise (beware!): a widespread bad habit is to store zero or some special value when a piece information is not available
3. Resolution has a great influence on the results
   - the analysis of too detailed data can be affected by noise
   - the analysis of too general data can hide interesting patterns

A good understanding of the data set is required!

Data types

Data Type		*Description*	*Examples*	*Descriptive statistics allowed*
Categorical	Nominal	The values are a set of labels, the available information allows to distinguish a label from another Operators: $=$ and $\neq$	zip code, eye color, sex, ...	mode, entropy, contingency, correlation, χ^2 test
	Ordinal	The values provide enough information for a total ordering Operators: $<>\leq\geq$	hardness of minerals, non–numerical quality evaluations (bad, fair, good, excellent)	median, percentiles, rank correlations
Numerical	Interval	The difference is meaningful Operators: $+-$	Calendar dates, temperatures in centigrades and Fahrenheit	average, standard deviation, Pearson's correlation, F and t tests
	Ratio	Have a univocal definition of 0 Allow all the mathematic operations on numbers	Kelvin temperatures, masses, length, counts	geometric mean, harmonic mean, percentage variation

The “description” and “descriptive statistics” columns are *incremental,*
i.e. the properties described in a row are added to the properties described in the rows above

Data preprocessing

Data preprocessing is a data mining technique that transforms raw data into an understandable format.

It is an important step in the DM process, above all in case of:

- Presence of noise and outliers
- Missing values
- Presence of duplicates and data inconstencies
- Curse of dimensionality (when dimensionality is very high)

It is tipycally a long step!

“garbage in, garbage out” (**GIGO**)
flawed or nonsense input data produces nonsense output

2 - Algorithms for DS

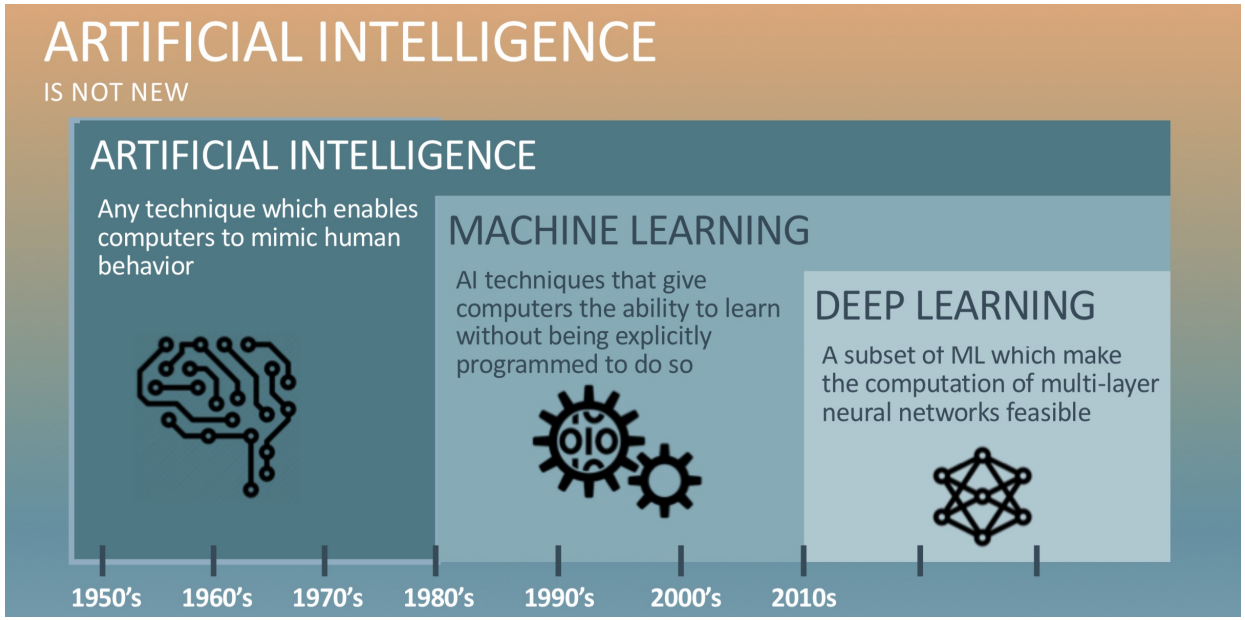
HOW TO LEARN FROM DATA

Brief history of Artificial Intelligence

In the beginning was the (pure) *statistics.* Since the 18th century, great developments in:

- Descriptive statistics
- Inferential statistics
- Statistical models

Since late ’50s of XX century:



Example of AI

Soybean diseases [Michalski and Chilausky, 1980]

Soybean (Large) Data Set

Download: Data Folder, Data Set Description

Abstract: Michalski's famous soybean disease database

Data Set Characteristics:	Multivariate	**Number of Instances:**	307	**Area:**	Life
Attribute Characteristics:	Categorical	**Number of Attributes:**	35	**Date Donated**	1988-07-11
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	120988

diagnosis of soybean diseases (19 different diseases)

AI approach:
starting from the expert knowledge, computer scientists tried to translate (formalised) the classification rules into code.

Examples:

If leaf condition = normal and stem condition = abnormal
and stem cankers = below soil line and canker lesion color = brown

then: diagnosis is rhizoctonia root rot

If leaf malformation = absent and stem condition = abnormal
and stem cankers = below soil line and canker lesion color = brown

then: diagnosis is rhizoctonia root rot

Results:
the diagnosis accuracy obtained by the rules alone, without expert assistance, is 72%

Problems:

1. the elicitation of rules from expert is difficult and time consuming (35 attributes = huge number of “if” cases!);

   If leaf condition = normal and stem condition = abnormal
   and stem cankers = below soil line and canker lesion color = brown

   then: diagnosis is rhizoctonia root rot

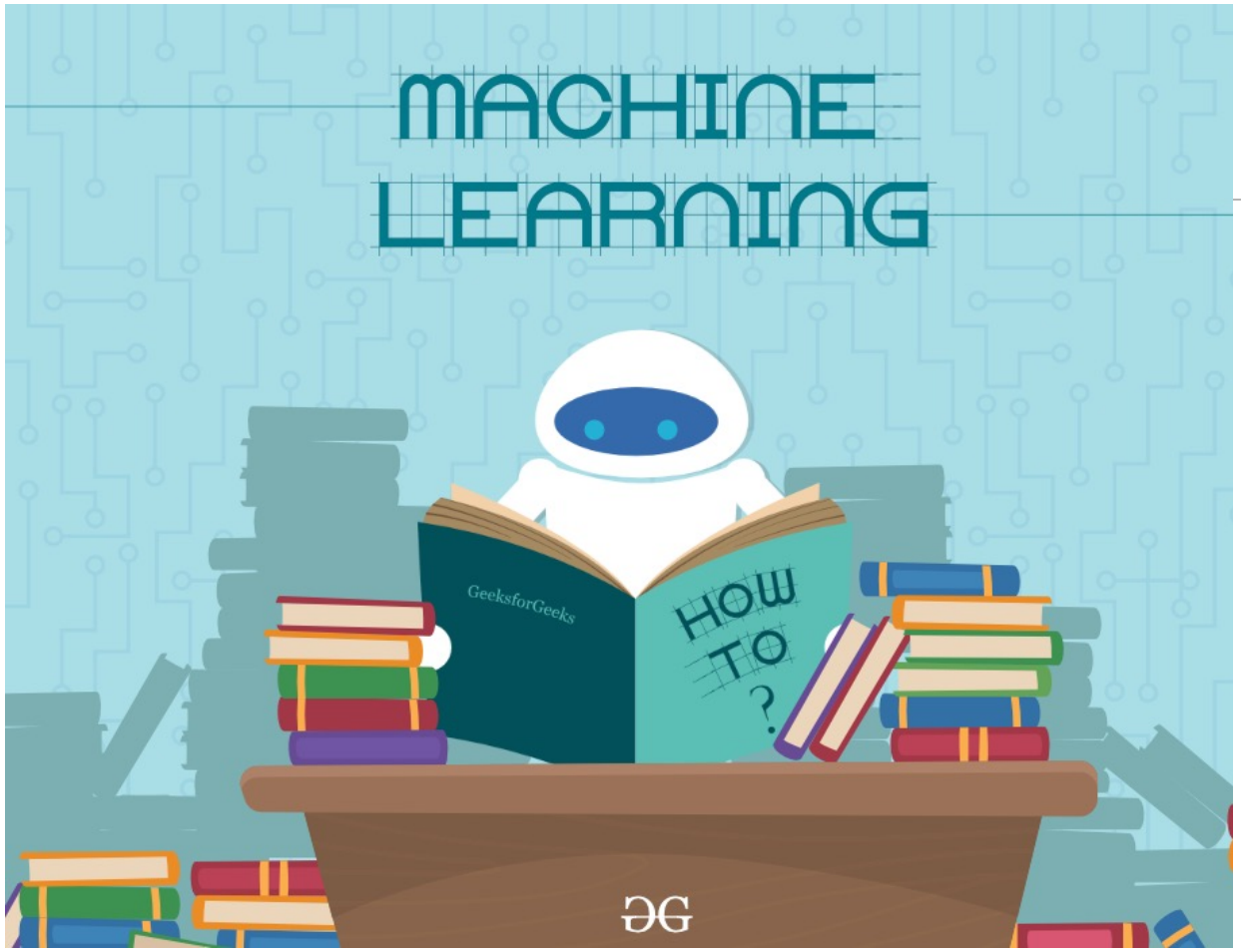
   If leaf malformation = absent and stem condition = abnormal
   and stem cankers = below soil line and canker lesion color = brown

   then: diagnosis is rhizoctonia root rot

2. the rules are not independent, and they should be carefully checked;

3. the achieved accuracy (72%) is not satisfactory.

Conclusion:
the rules are not able to capture all the expert knowledge *by being told*



Alternative approach

The same data set has been processed by a machine learning algorithm, to generate its classification rules.

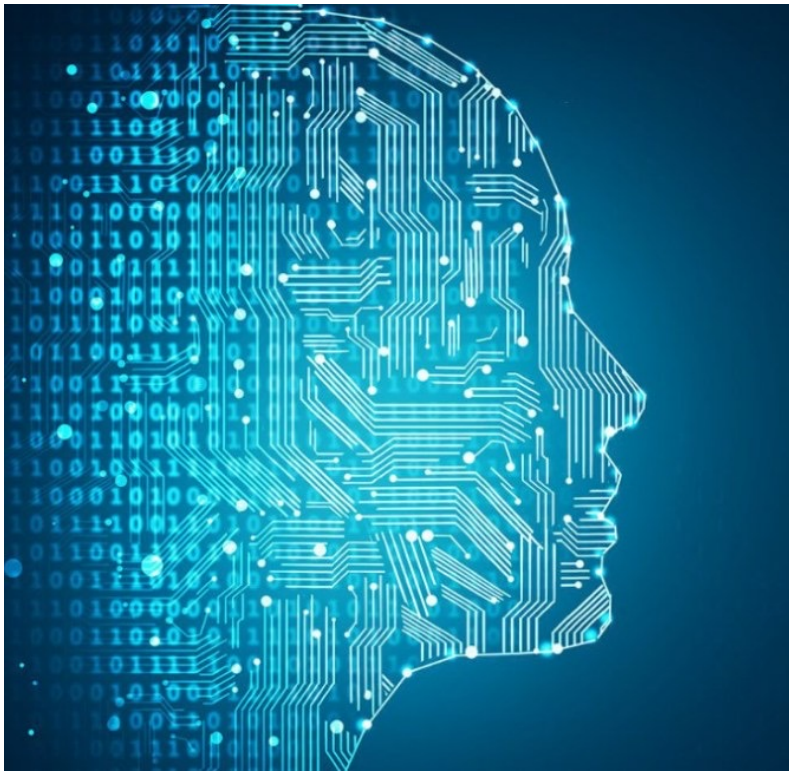
The new set of rules got an accuracy of 97.5%, comparable with that of a junior expert.

Machine Learning

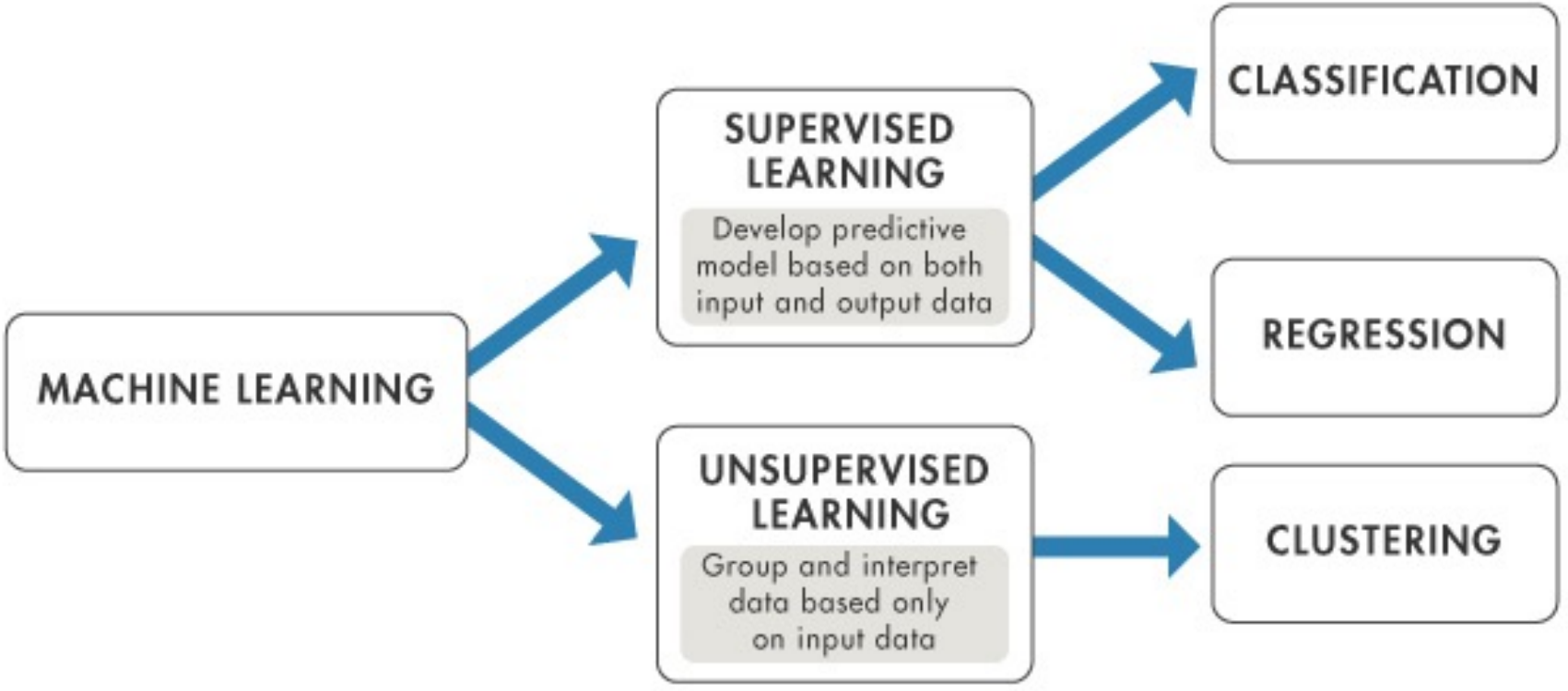
Machine Learning is a field of study that gives computers the ability to learn without being explicitly programmed (*learning from examples)*

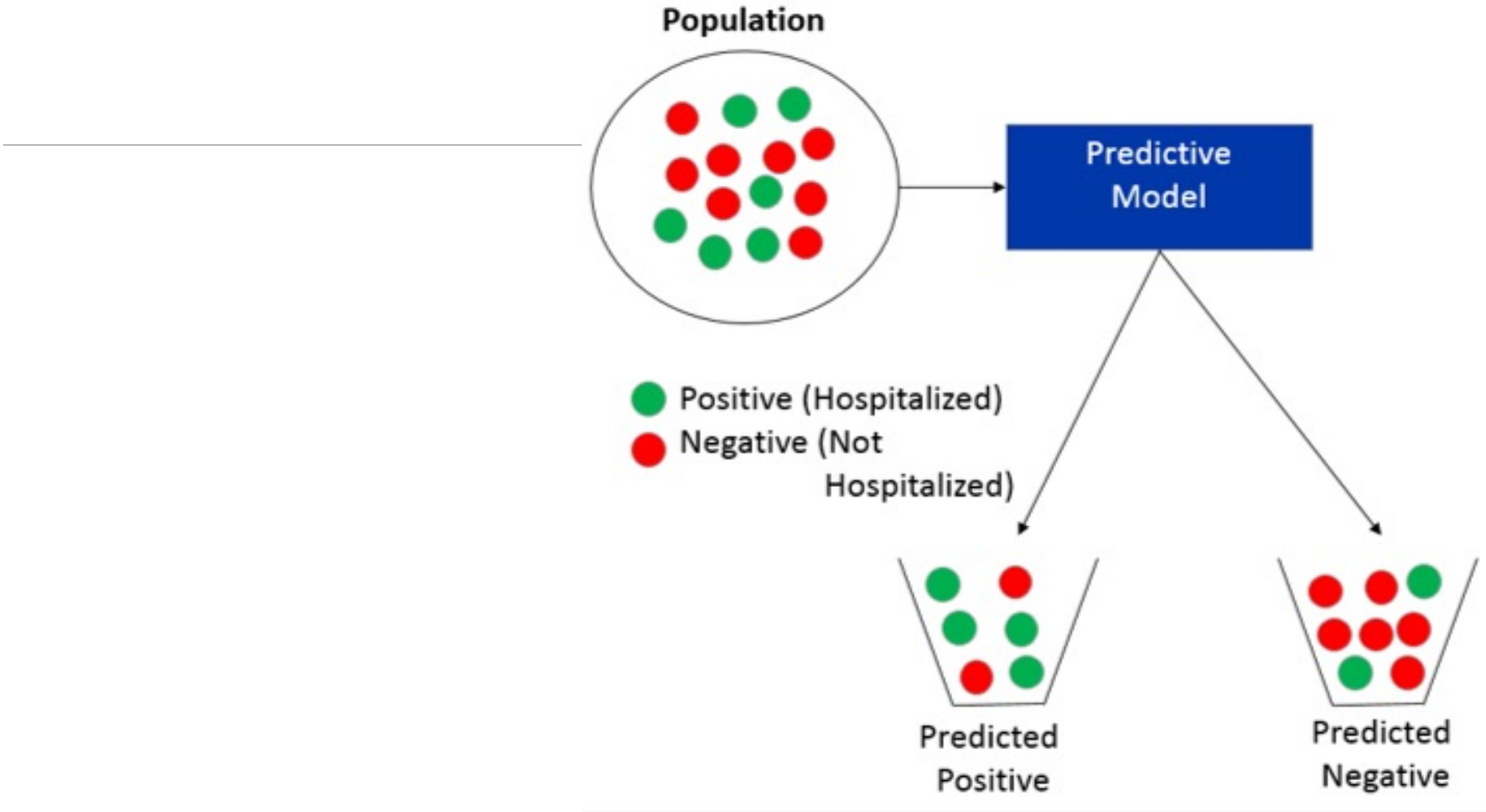
In particular*:*

- ML was born in 80s
- Since the early ’90s we talk about ‘’Data Mining processes with Machine Learning’’



Algorithms for approximations (i.e. for learning from available data)

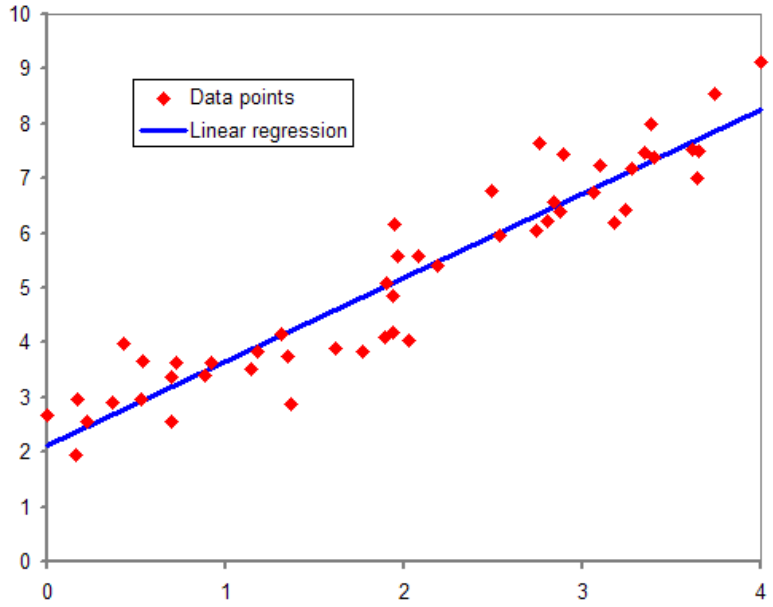



Population
Predictive Model
Positive (Hospitalized)
Negative (Not Hospitalized)
Predicted Positive
Predicted Negative

Regression

regression, value estimation

Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features').



x = independent variable
y = dependent variable,
$y \approx f(x)$

Given all the data points (x_i , y_i), we look for f

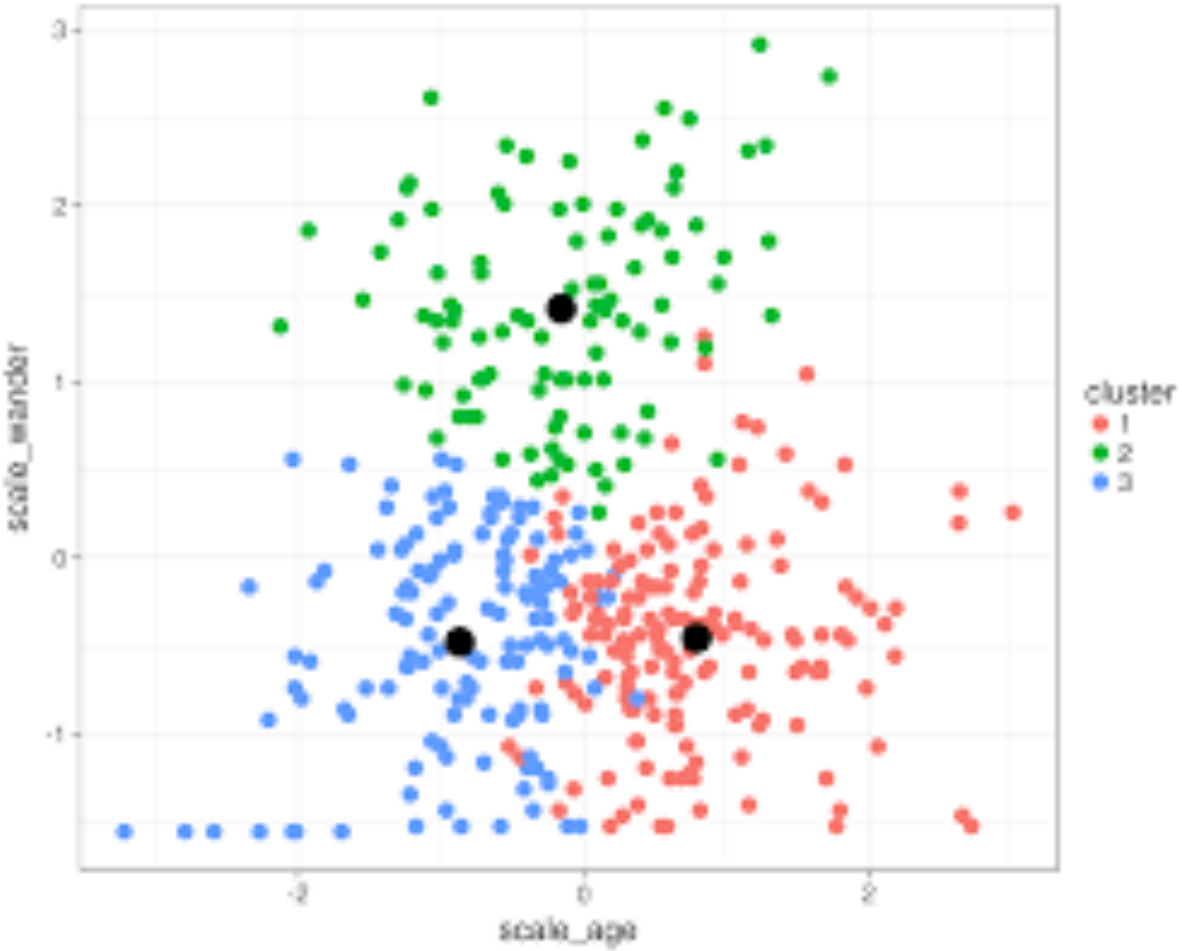
Remarks

- given a set of numeric attribute values for an individual, it estimates the value of another numeric attribute
- it is related to *classification*, but the methods are completely different
- is primarily used
  - for prediction and forecasting
  - to infer causal relationships between two variables

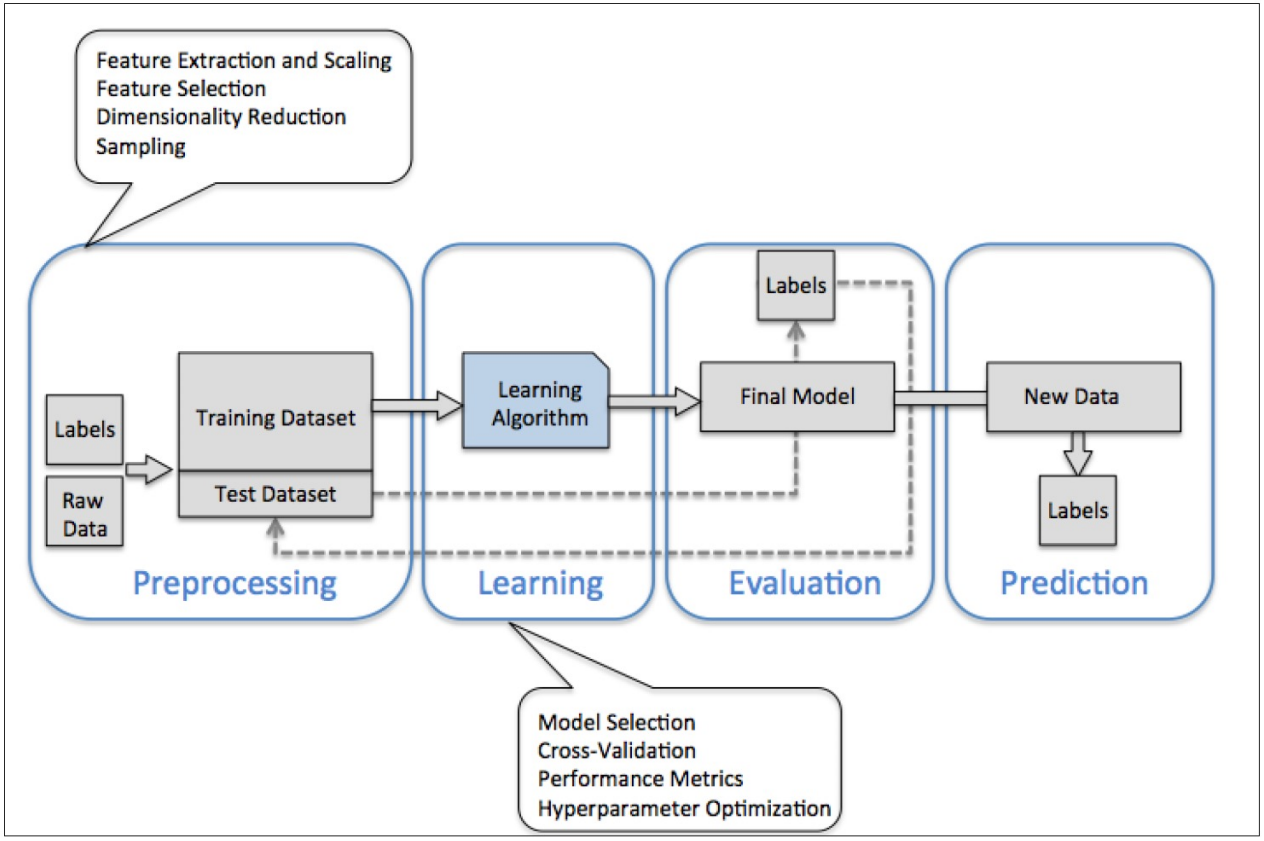
Clustering

clustering , cluster analysis

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups.



Steps for ML


Feature Extraction and Scaling
Feature Selection
Dimensionality Reduction
Sampling
Labels
Raw Data
Training Dataset
Test Dataset
Learning Algorithm
Labels
Final Model
New Data
Labels
Preprocessing
Learning
Evaluation
Prediction
Model Selection
Cross-Validation
Performance Metrics
Hyperparameter Optimization

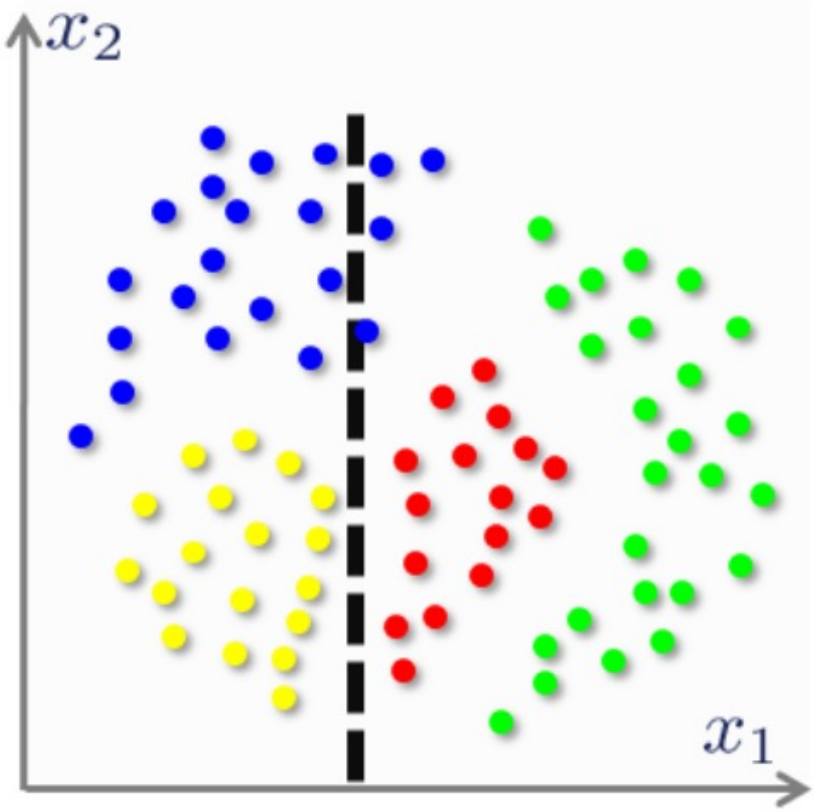
	H	J	F
1	Age	Annual Salar	Gender
2	55	$141,604	Female
3	59	$99,975	Male
4	50	$163,099	Female
5	26	$84,913	Female
6	55	$95,409	Male
7	57	$50,994	Male
8	27	$119,746	Female
9	25	$41,336	Male
10	29	$113,527	Male
11	34	$77,203	Female
12	36	$157,333	Female
13	27	$109,851	Female
14	59	$105,086	Male
15	51	$146,742	Female
16	31	$97,078	Male
17	41	$249,270	Female
18	65	$175,837	Female
19	64	$154,828	Female
20	64	$186,503	Male
21	45	$166,331	Male
22	56	$146,140	Male
23	36	$151,703	Female
24	59	$172,787	Male
25	37	$49,998	Male
26	44	$207,172	Male
27	41	$152,239	Male
28	56	$98,581	Female
29	43	$246,231	Male
30	64	$99,354	Male
31	63	$231,141	Male
32	28	$54,775	Male

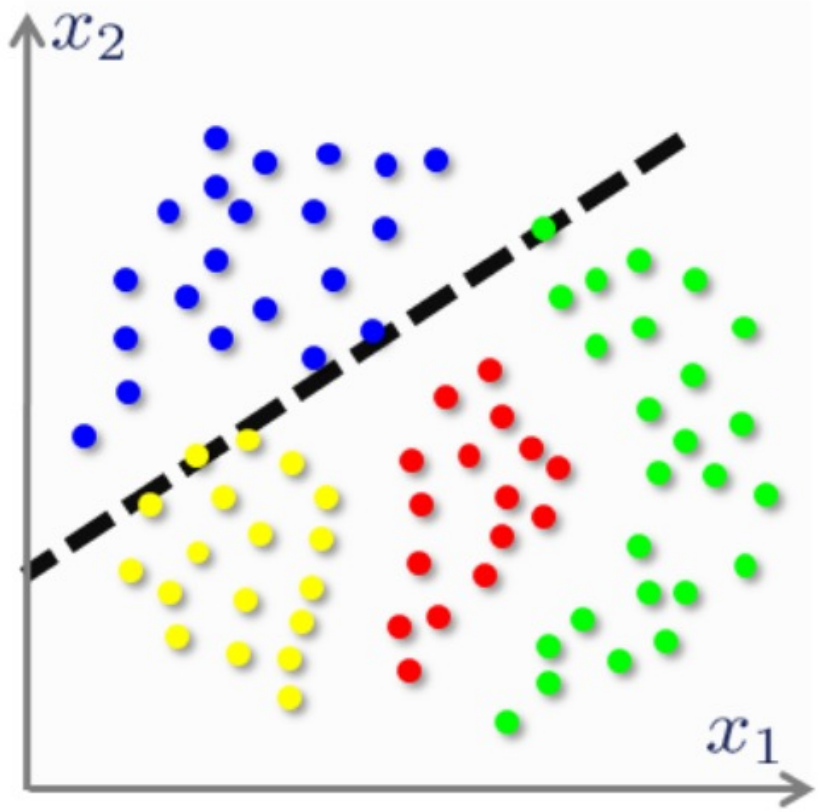
Starting problem

We have many observations (rows)

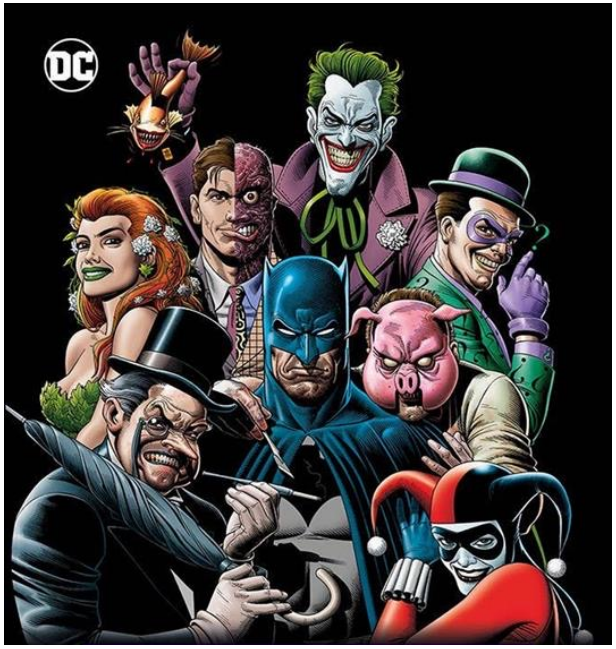
We need to find the hidden pattern (assuming it does exist!) between three columns:

- 2 (or more) column with numeric or categorical entries
- 1 column with classes


x_2
x_1


x_2
x_1

Example of decision tree

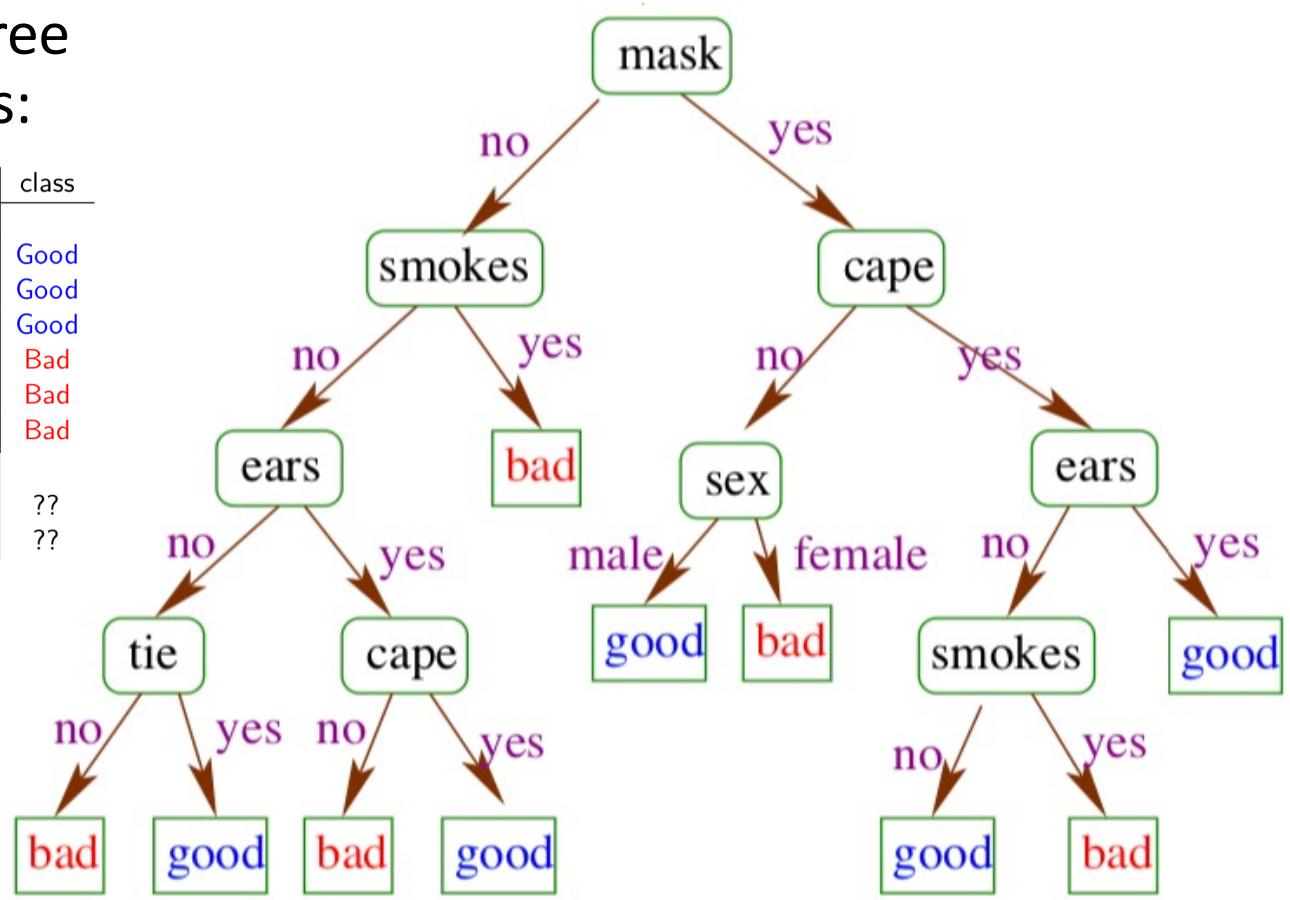


We want to identify characters as good or bad, from their appearance based on:

	sex	mask	cape	tie	ears	smokes	class
			training data				
batman	male	yes	yes	no	yes	no	Good
robin	male	yes	yes	no	no	no	Good
alfred	male	no	no	yes	no	no	Good
penguin	male	no	no	yes	no	yes	Bad
catwoman	female	yes	no	no	yes	no	Bad
joker	male	no	no	no	no	no	Bad

After training a decision tree on the available 6 samples:

	sex	mask	cape	tie	ears	smokes	class
			training data				
batman	male	yes	yes	no	yes	no	Good
robin	male	yes	yes	no	no	no	Good
alfred	male	no	no	yes	no	no	Good
penguin	male	no	no	yes	no	yes	Bad
catwoman	female	yes	no	no	yes	no	Bad
joker	male	no	no	no	no	no	Bad
			test data				
batgirl	female	yes	yes	no	yes	no	??
riddler	male	yes	no	no	no	no	??



Testing:

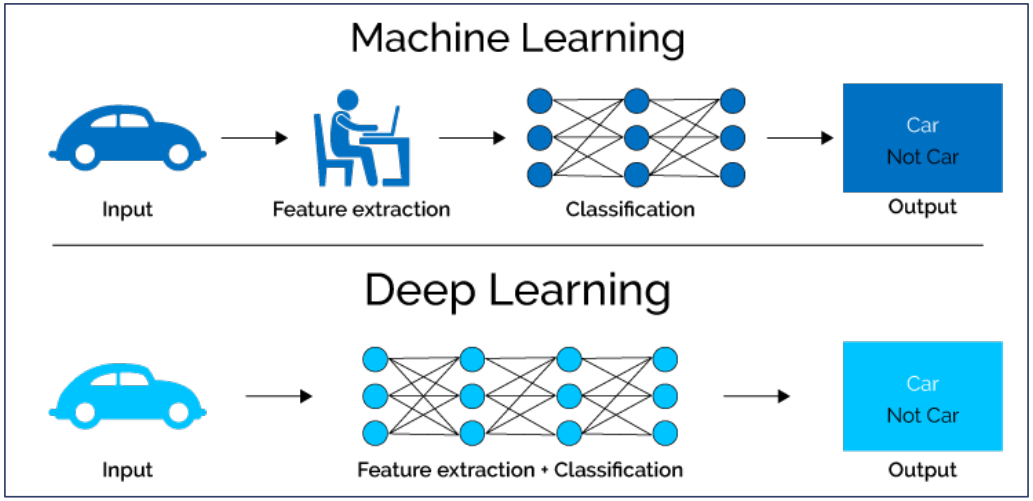
- Batgirl:
  good (correctly classified)
- Riddler:
  good (uncorrectly classified)

3 – Deep Learning

NNs for Deep Learning

Neural Networks are schemes accepting as input (also) no-structured data such as images, sounds and signals.

NNs embed the feature selection step, in their architectures, and allow for a *deeper* learning than ML.



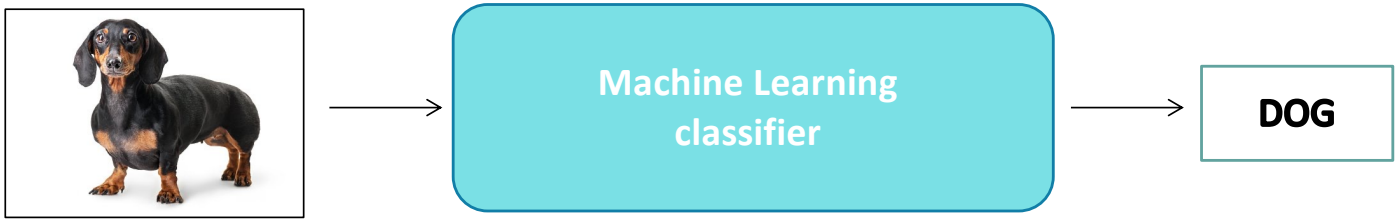
The two main deep learning architectures used in classification are Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).

ML definition: what does it mean?

''Machine Learning is that branch of computer science
that gives computers
the *ability to learn on their own*
without being explicitly programmed to do so''

Let's try to understand it with an example of image recognition.
In our setting:

- we have many images (containing cats or dogs) already labelled;
- we want to automatically classify new images as "CAT" or "DOG", through a ML program.



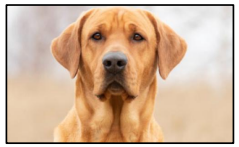
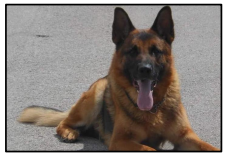
What we have:

1. Dog images

2. Cat images

3. ML theory

We need to train the ML classifier with the original labelled images

1.  DOG → ????????
2.  DOG → ???????
3.  DOG → ??????
4.  DOG → ?????
5. ...

6.  CAT → ????

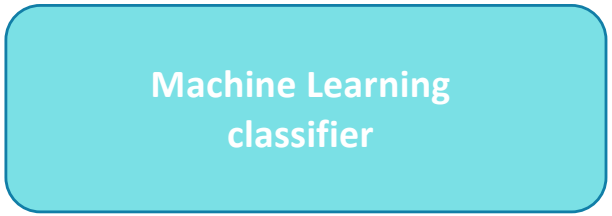
7.  CAT → ???

8.  CAT → ??

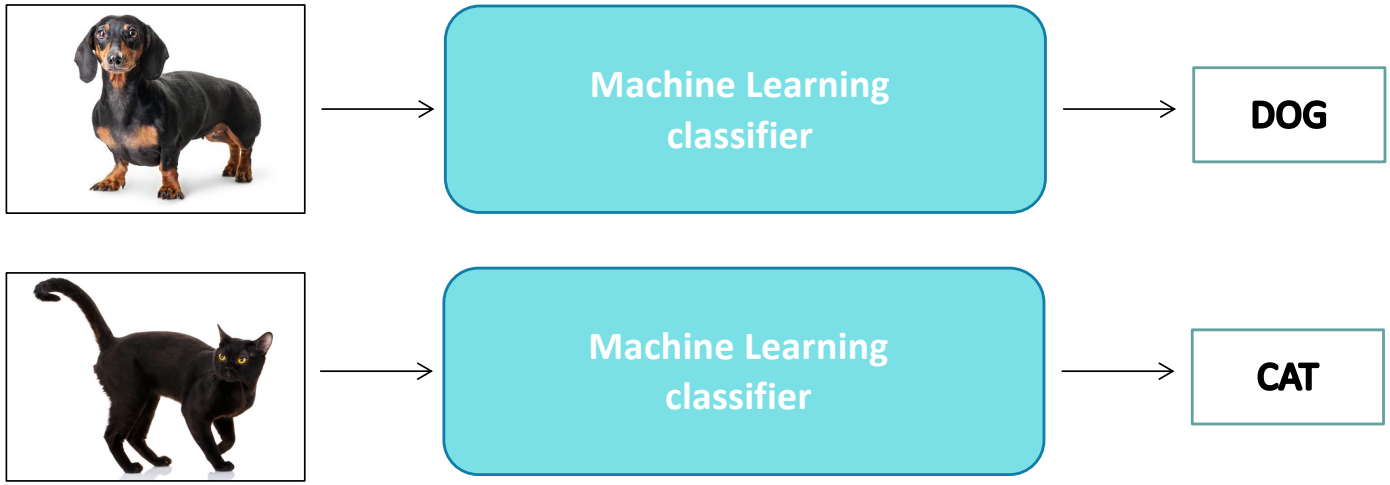
9.  CAT → ?

10. ...

After this phase, the Machine Learning classifier will be defined:



Thus, it will be usable on new, unseen images:



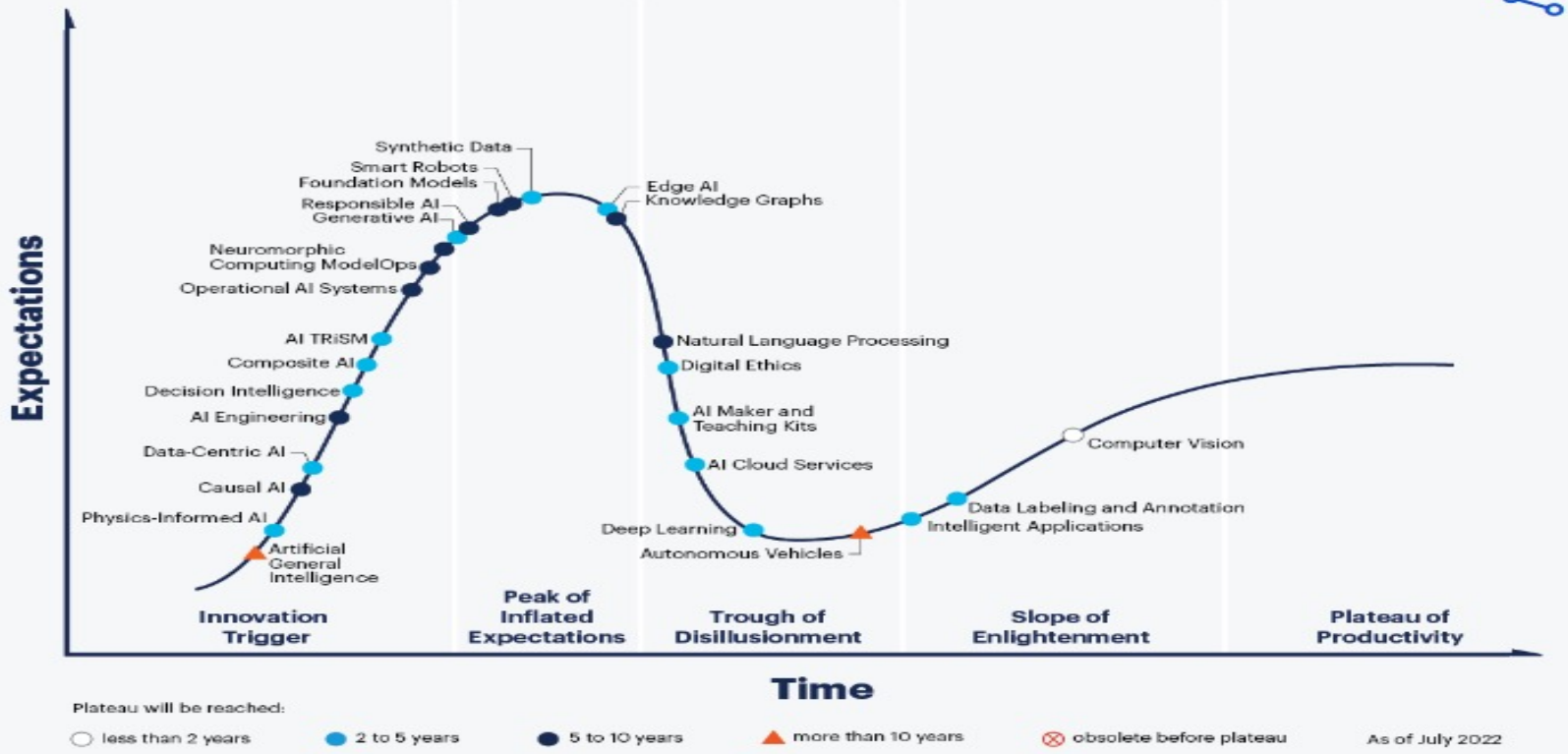
BUT pay attention, because there may be errors:

- on some unseen images

- on unseen objects (other than cats or dogs)

- on more than 1 subject

Hype Cycle for Artificial Intelligence, 2022



Plateau will be reached:

○ less than 2 years ● 2 to 5 years ● 5 to 10 years ▲ more than 10 years ⊗ obsolete before plateau

As of July 2022

gartner.com

Gartner