Outline

Outline

1. Introduction **1**
1.1. Objectives 1
1.2. A World of Data 2
1.3. Data Science 5

2. Machine Learning **10**
2.1. Machine Learning 10
2.2. Bio-inspired Learning 11
2.3. Model vs Algorithm 15
2.4. Machine Learning Methods Division 16

3. Data Science Process **22**
3.1. Stages of the Data Science Process 22
3.2. Stage 2: Preprocessing 24
3.3. Stage 3: Machine Learning 24
3.4. Stage 4: Evaluation 24
3.5. Stage 5: Dissemination 25

4. Applications **26**
4.1. General Applications 26
4.2. Unusual applications 28

5. Software **28**
5.1. Software 28
5.2. Scikit-learn 29

6. Conclusions **30**

1. Introduction

1.1. Objectives

Objectives

What is Machine Learning?

- Know what Data Science and Machine Learning are.
- Understand how the most commonly used techniques in Machine Learning work.
- Understand the difference between the different types of existing problems.
- Understand the main difficulties in Machine Learning.
- Know some practical examples of the application of Machine Learning techniques.

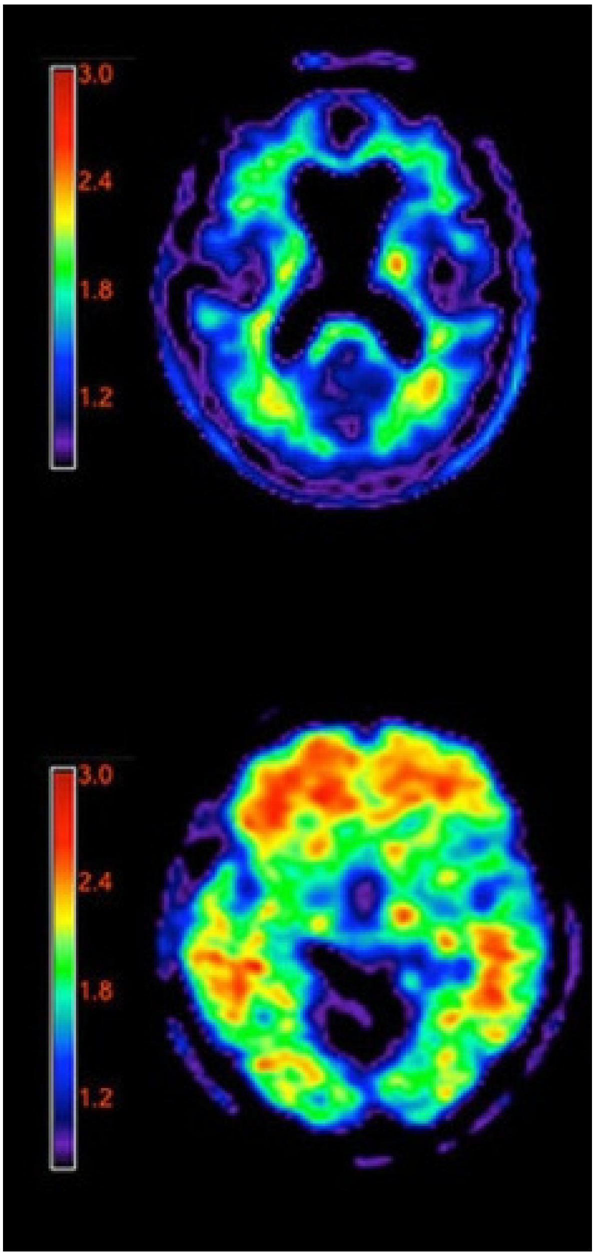
1.2. A World of Data

A world of data

Our world increasingly revolves around data:

- Science: astronomy, genomics, environment...
- Industry and Energy: sensor networks, Internet of things, wind farm management, demand forecasting, smart cities...
- Social sciences and humanities: digitised books, historical documents, social data...

A world of data



- Entertainment: recommender systems, digital content, multimedia searches...
- Medicine: medical imaging, hospital demand forecasting, expert systems...
- Finance and business: automated market transactions...

Data explosion

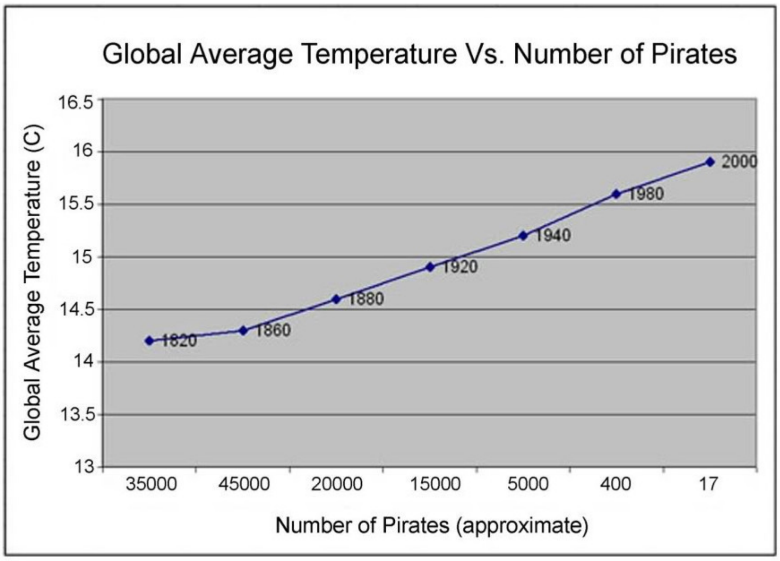
Data has been stored for decades and could not be processed until a few years ago:

- Database technologies.
- Cost of storage hardware.
- Increase in bandwidth.
- Increased processing capacity.
- Scientific software.

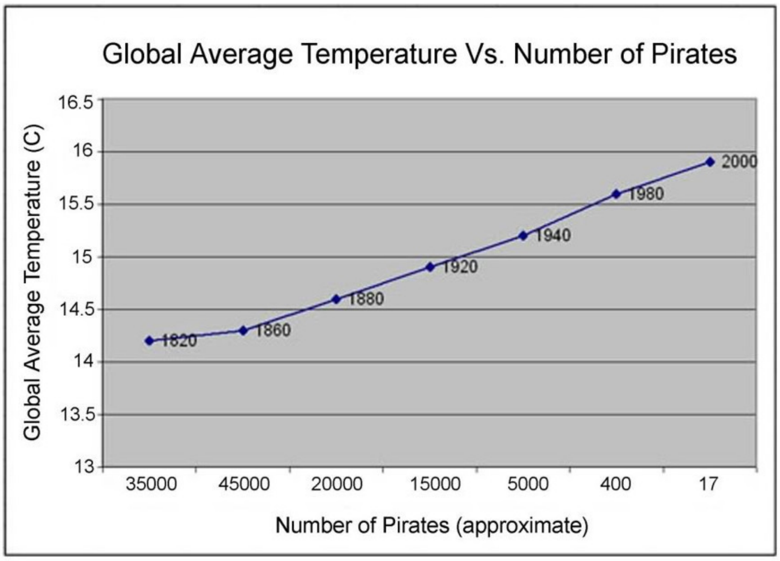
Computer scientists to the rescue

All this enables us to move from information to knowledge.

Caution



Caution

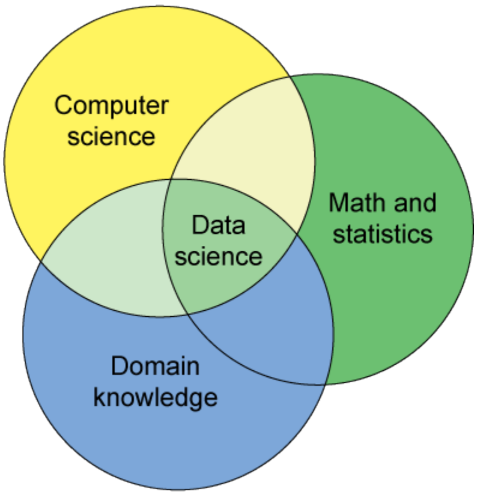


To stop global warming: let’s become pirates!

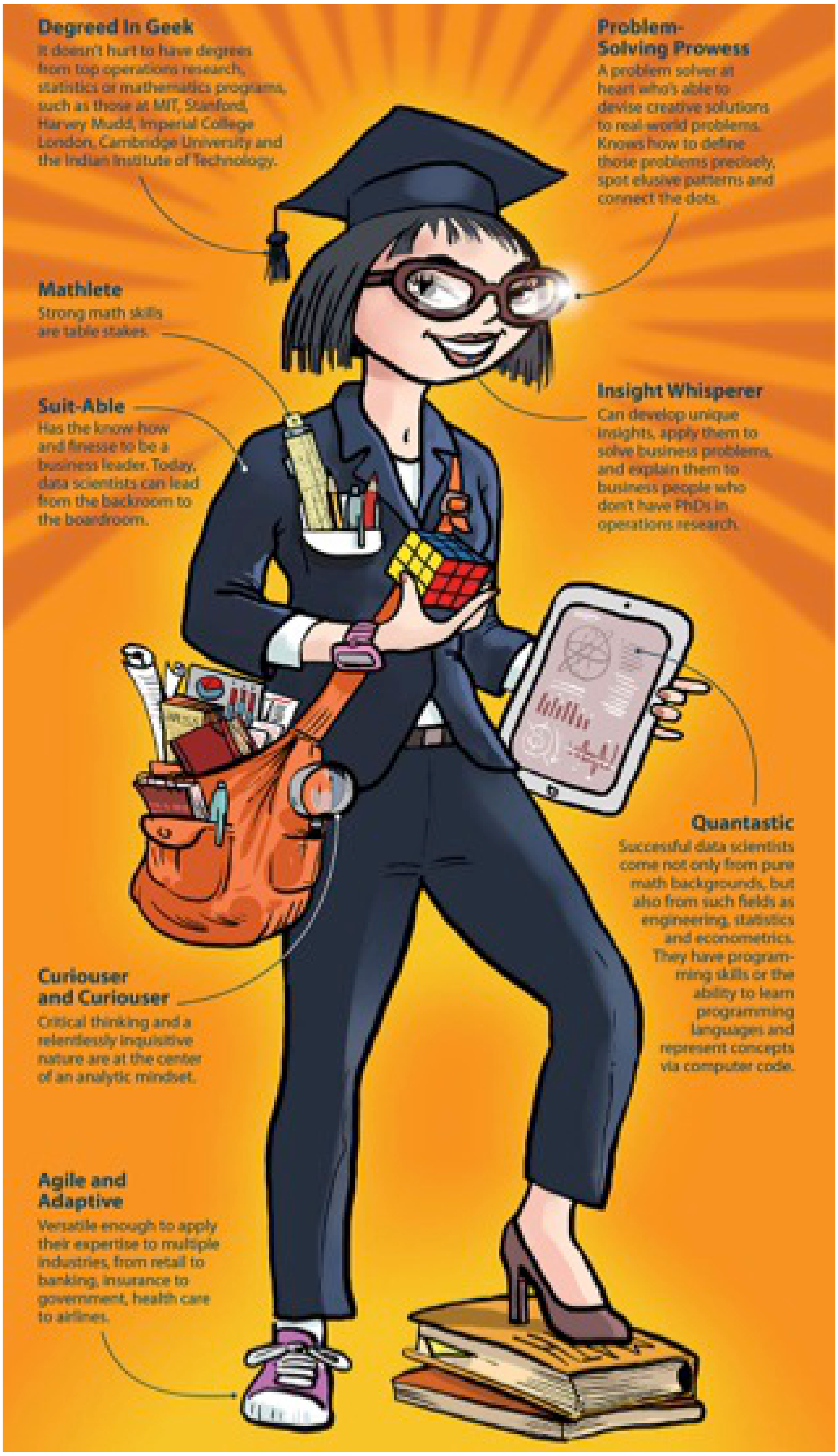
1.3. Data Science

Data Science

Definition
Knowledge area covering the skills associated with data processing.



Career prospects


Degreed In Geek
It doesn't hurt to have degrees from top operations research, statistics or mathematics programs, such as those at MIT, Stanford, Harvey Mudd, Imperial College London, Cambridge University and the Indian Institute of Technology.
Problem-Solving Prowess
A problem solver at heart who's able to devise creative solutions to real-world problems. Knows how to define those problems precisely, spot elusive patterns and connect the dots.
Mathlete
Strong math skills are table stakes.
Suit-Able
Has the know-how and finesse to be a business leader. Today, data scientists can lead from the backroom to the boardroom.
Insight Whisperer
Can develop unique insights, apply them to solve business problems, and explain them to business people who don't have PhDs in operations research.
Quantastic
Successful data scientists come not only from pure math backgrounds, but also from such fields as engineering, statistics and econometrics. They have programming skills or the ability to learn programming languages and represent concepts via computer code.
Curiouser and Curiouser
Critical thinking and a relentlessly inquisitive nature are at the center of an analytic mindset.
Agile and Adaptive
Versatile enough to apply their expertise to multiple industries, from retail to banking, insurance to government, health care to airlines.

- Data scientist: programming skills, Machine Learning, AI and specific knowledge.

- 2012: Called *"the sexiest job of the 21st century"* by the Harvard Business Review.
- 2022: Data Scientist: Still a Powerhouse, but No Longer the Unicorn.
- Salaries range from $80,000 to $300,000.
- Demanding profiles.

What is a data scientist?

José Antonio Guerrero: one of the world's best data scientists in the world (Kaggle Platform).

Its definition
This is a person with a background in mathematics, statistics and optimisation methods, with knowledge of programming languages, and who also has practical experience in analysing real data and building predictive models.

Of the three characteristics, perhaps the most difficult is the third; it is not for nothing that data modelling has sometimes been defined as an art. There are no golden rules here, and every dataset is a blank canvas.

Source: El Confidencial

KDD, Data mining and Data Science

- KDD → Knowledge Discovery from Databases.
- KDD is the whole process of extracting knowledge from databases.
- The term was introduced in 1989 to emphasise that knowledge is the end product of a data-driven discovery process.
- Data Mining (DM) is a stage in the KDD process, although informally, DM is associated with KDD.

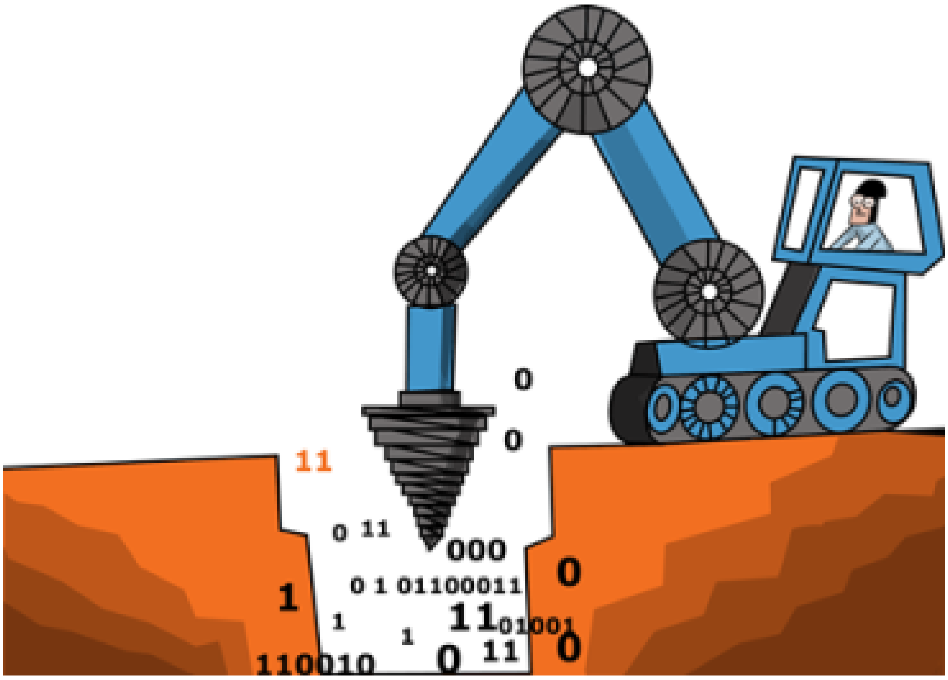
KDD, Data mining and Data Science

Data Mining/Machine Learning
DM/ML is the process of extracting patterns of information (implicit, non-trivial, unknown and potentially useful) from large amounts of data.

Contribution of the term Data Science
The term “Data Science” adds more activities, such as, for example, an emphasis on data visualisation or working with unstructured data (which is quite common in Big Data).

Data Science, ML, and Big Data are more current terms than DM and KDD.

Why do we analyse data?



- Summarise a large database.
- Visualise multi-dimensional data.
- Predict values.
- Explain existing data.

Data sources

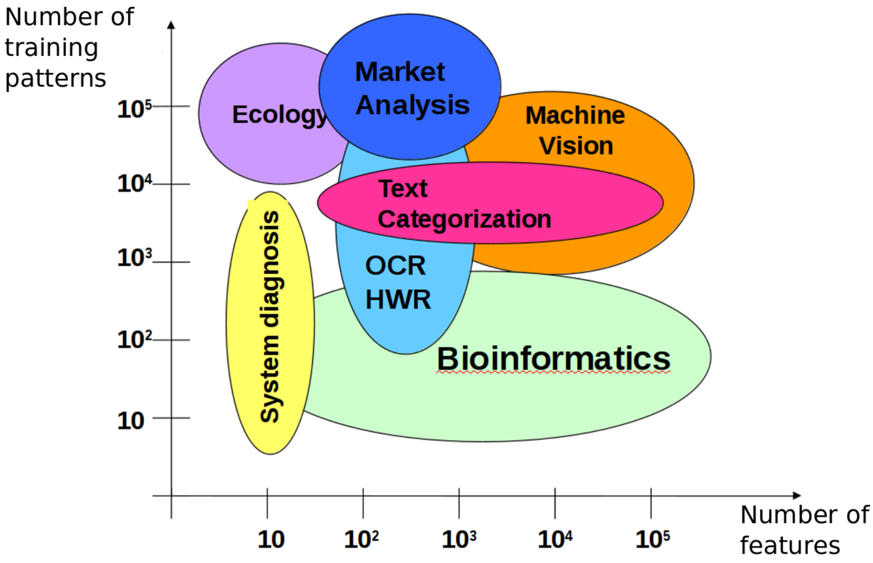
Data sources are very varied, even mixed, giving rise to disciplines such as information fusion:

- Relational databases.
- Spatial and/or temporal databases: mobile phone.
- Document databases.

- Multimedia databases: images, videos, sounds...
- The World Wide Web.

Often, all these data are unstructured (Big Data).

Data size



2. Machine Learning

2.1. Machine Learning

How to extract knowledge?

Data Science is about extracting knowledge from data using:

1. Classical statistical techniques.
2. Artificial Intelligence and Machine Learning.

Many Machine Learning methods rely on mathematical optimisation methods and statistical techniques.

However, they are often combined with Artificial Intelligence (AI) techniques to overcome the former's limitations in model training and design solutions to problems, create expert systems, etc.

Machine Learning

Machine Learning (sub-area of AI) ≡ Data Mining.

Machine Learning

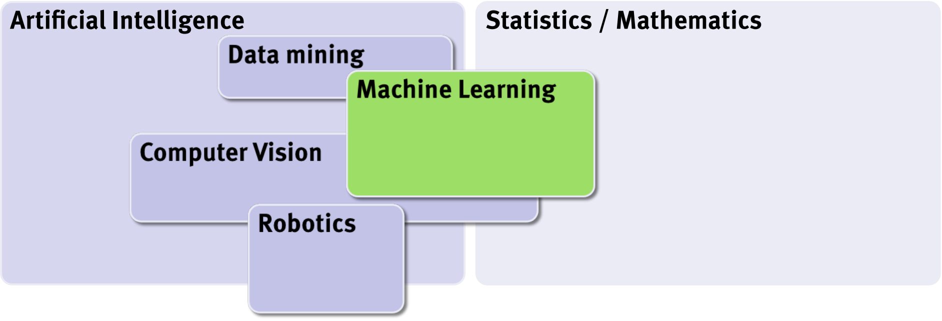
Machine Learning is defined as "the field of study that gives computers the ability to learn without being explicitly programmed".

Machine Learning amounts to “learning from data” in order to extract the necessary knowledge for different purposes.

This “learning from data” makes Machine Learning fall between different branches belonging to Artificial Intelligence, statistics and mathematics.

Machine Learning

Definition
An area of study that gives computers (machines) the ability to learn without being specifically programmed for the task at hand.

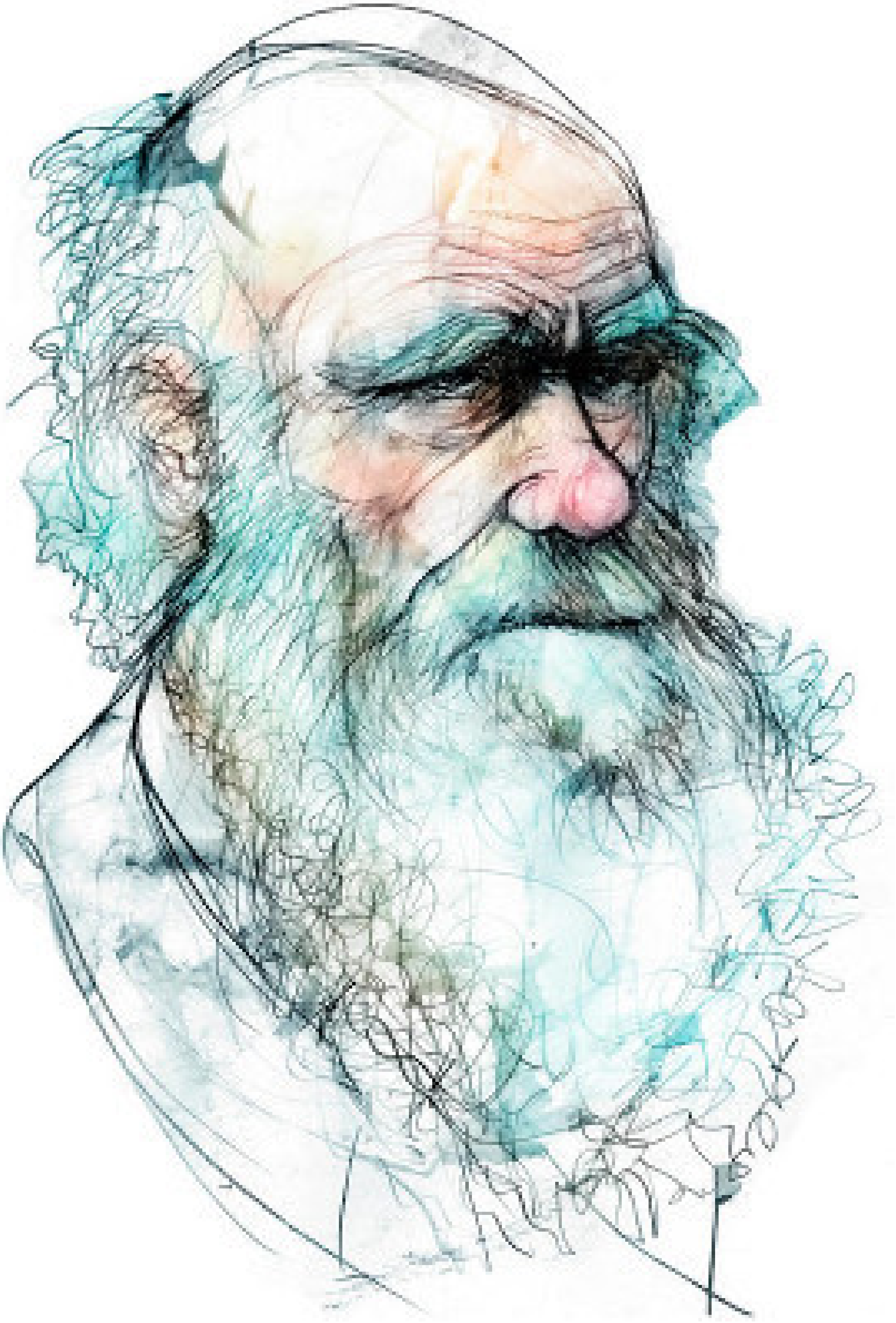


2.2. Bio-inspired Learning

Learning from nature

A specific area of AI, ML and optimisation tries to mimic processes occurring in nature to provide solutions to different optimisation problems. These are bio-inspired algorithms.

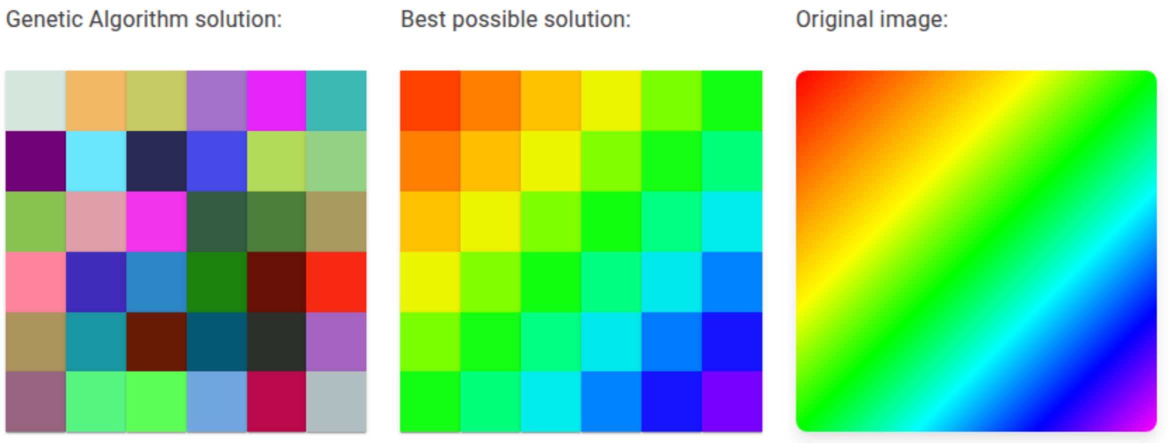
Evolutionary algorithms



- They are non-deterministic search algorithms, which incorporate the semantics of natural evolution into optimisation processes.
- Darwin: Species evolve according to their environment and the best adapted survive.
- Individuals of each species → Solutions to the problem.
- Design of mutation and crossover operators.

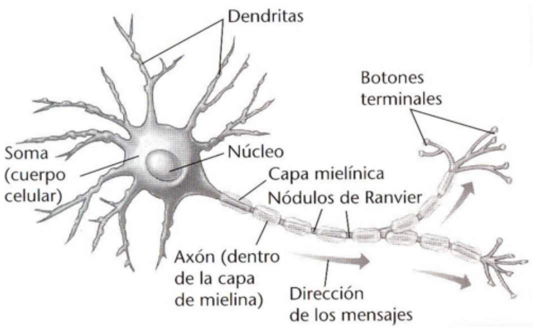
Genetic algorithms

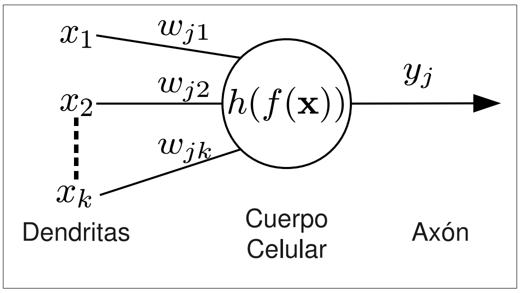
The https://geneticalgorithms.online/ allows you to solve different problems using genetic algorithms.



Artificial Neural Networks

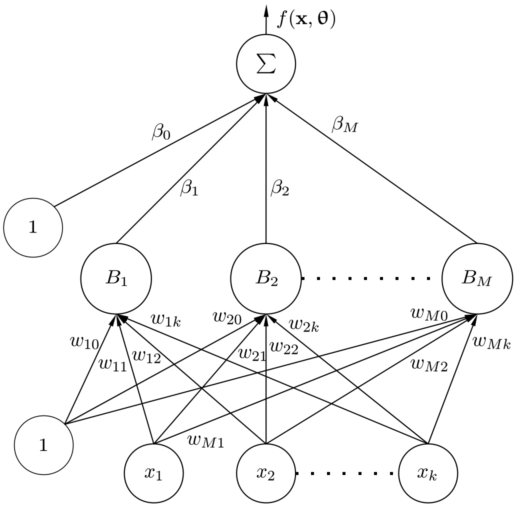
- Modelling technique based on the emulation of biological nervous systems.
- It combines many simple processing elements (neurons), highly interconnected and layered together.
- A neural network is a mathematical functional relationship between input and output variables.





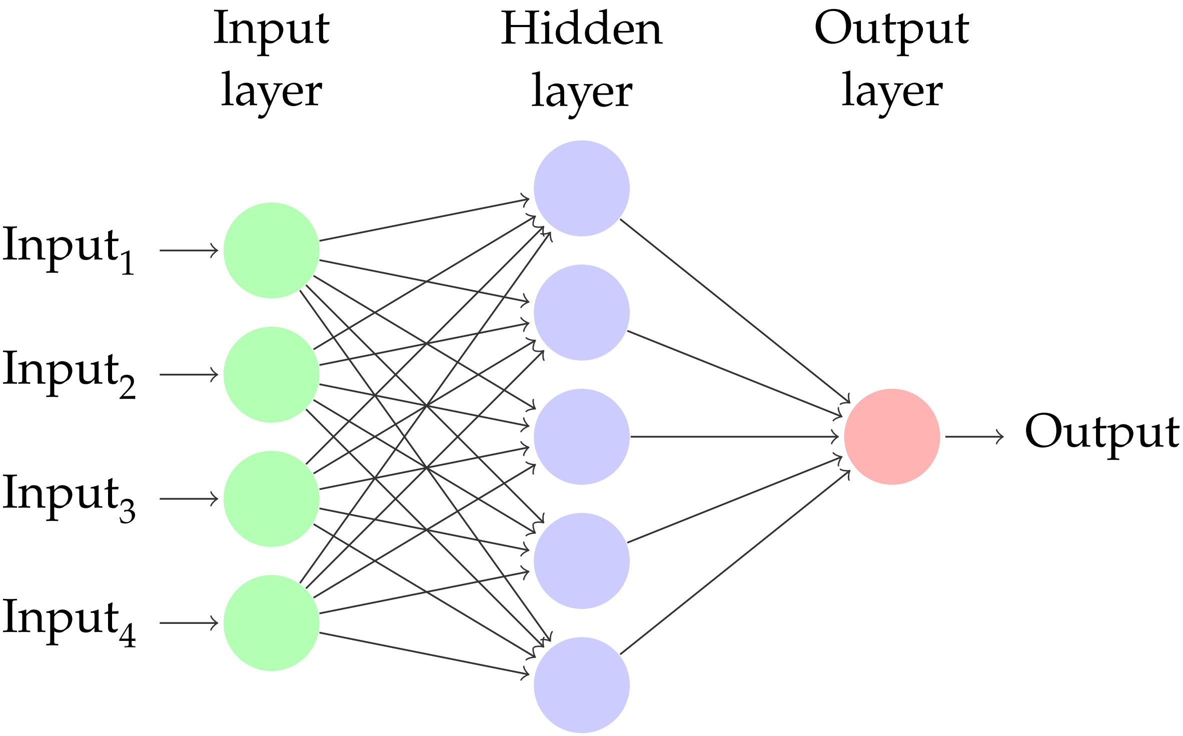
ANN for regression

$$f(\mathbf{x}, \boldsymbol{\theta}) = \beta_0 + \sum_{j=1}^{M} \beta_j B_j(\mathbf{x}, \mathbf{w}_j)$$



2.3. Model vs Algorithm

Difference between model and algorithm



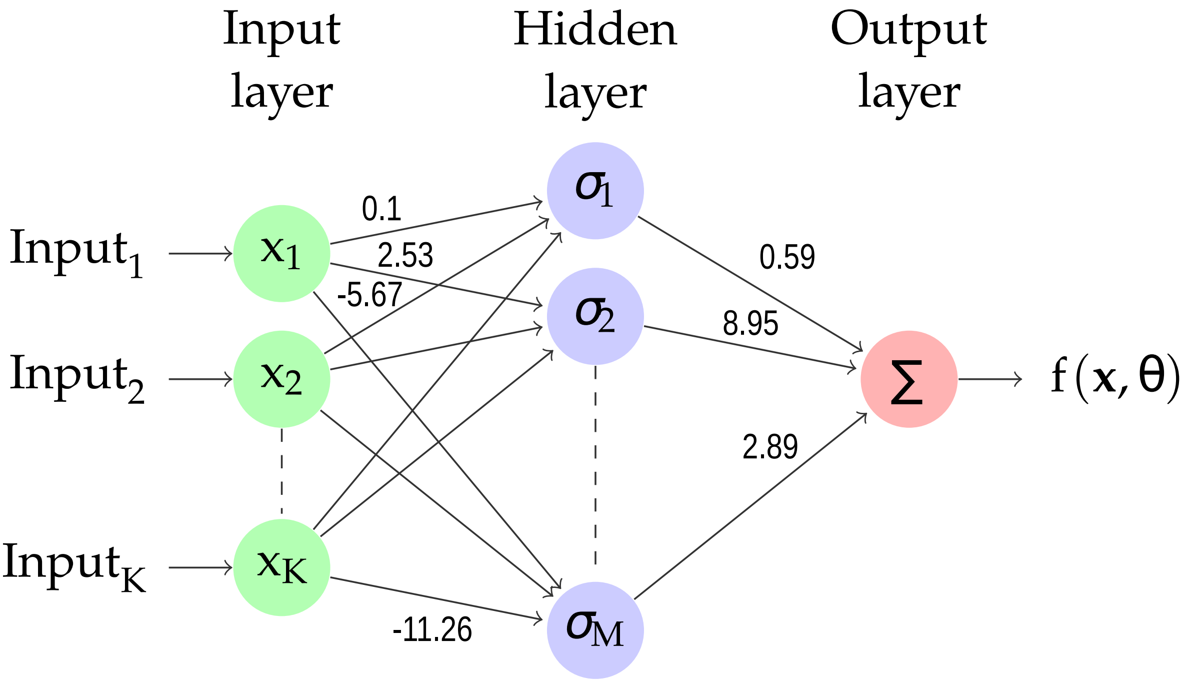
- A model is, in general, a function or structure which represents the knowledge underlying knowledge in a set of data.

- Depending on the objective of the problem, type of input variables, type of expected output complexity of the data, etc., there will be more or less appropriate.

Difference between model and algorithm

- An algorithm is a sequence of steps with a purpose.
- In Machine Learning, a learning algorithm ensures that the model learns from the data, or in other words, that the model fits the data.
- There are algorithms corresponding to the field of numerical optimisation, but others are also based on statistical learning or computational intelligence.

Trained model: neural network



- In the case of an artificial neural network (ANN), training a model consists of assigning weights to the network connections that minimise an error function, for example, the mean squared error.
- The resulting model depends on the input data, the model parameters (basis function, number of neurons and connections,...) and the learning algorithm.

2.4. Machine Learning Methods Division

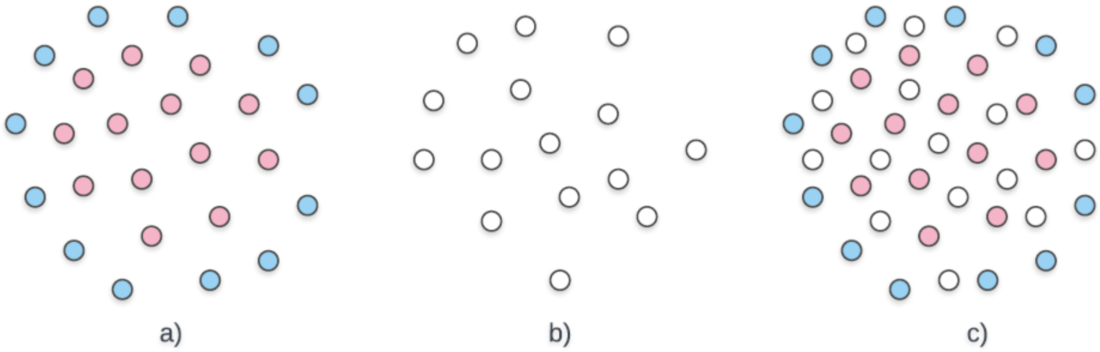
Division

Methods of Machine Learning can be divided into:

a) Supervised learning.

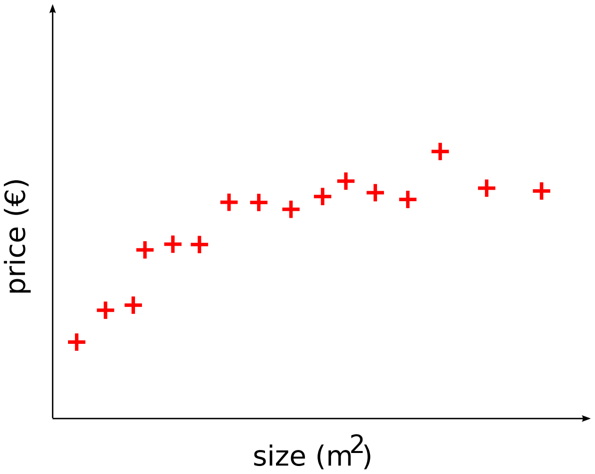
b) Unsupervised learning.

c) Semisupervised learning.



Supervised Learning: Regression

The output variable is a real value.



Given these facts, a friend has a 75 square metre house. How much could he expect to sell it for?

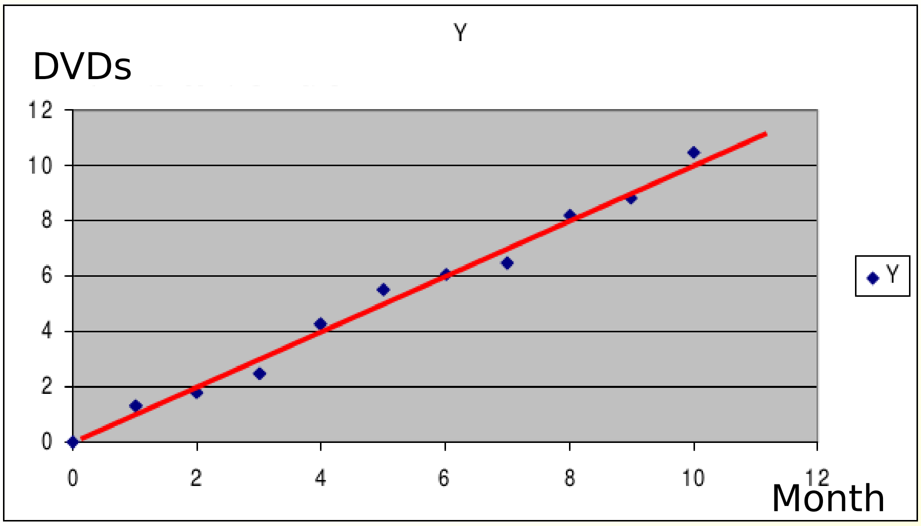
Supervised Learning: Regression

A large chain of household appliance shops wishes to optimise the operation of its warehouse by keeping sufficient stock of each product in order to be able to deliver the equipment purchased by its customers quickly.

Product	Feb	...	Oct	Nov	Dec	Jan
TV	20	...	52	14	139	74
Video	11	...	43	32	26	59
Fridge	50	...	61	14	5	28
Microwave	3	...	21	27	1	49
Discman	14	...	27	2	25	12
...	...	...	...	...	...	

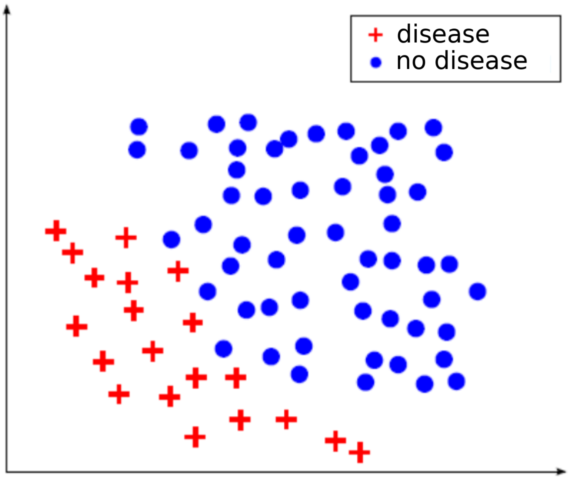
Supervised Learning: Regression

Knowledge gained → model that predicts what it will sell in January based on what it sold in previous months. January is based on what it sold in the previous months (February-December). This model can be built using time series:



Supervised Learning: Classification

The output variable is a categorical, discrete or nominal value.



Supervised Learning: Classification

- An online bank wants to obtain rules to predict which of those who apply for a loan will not repay it.
- The bank has an extensive database of previous loans granted to other customers and a history of what happened to those loans.

Supervised Learning: Classification

ID	Years	Euros	Salary	Own home	Overdue accounts	... ...	Repays credit
101	15	60000	2200	Yes	2	...	No
102	2	30000	3500	Yes	0	...	Yes
103	9	9000	1700	Yes	1	...	No
104	15	18000	1900	No	0	...	Yes
105	10	24000	2100	No	0	...	No
...	...	...	...	...	...		...

Supervised Learning: Classification

Knowledge gained → rule-based system:

- IF (overdue-accounts $>$ 0) THEN repays-credit $=$ no.
- IF (overdue-accounts $=$ 0) AND ((salary $>$ 2500) OR (years $>$ 10)) THEN repays-credit $=$ yes.

Supervised Learning: Classification

A visual introduction to Machine Learning (classification):

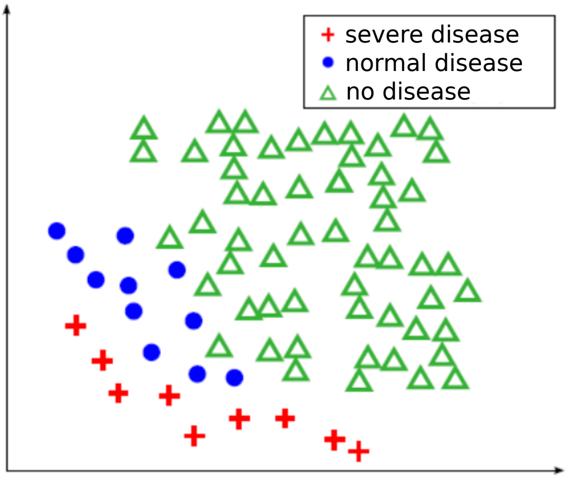
http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

Jointly realised by:

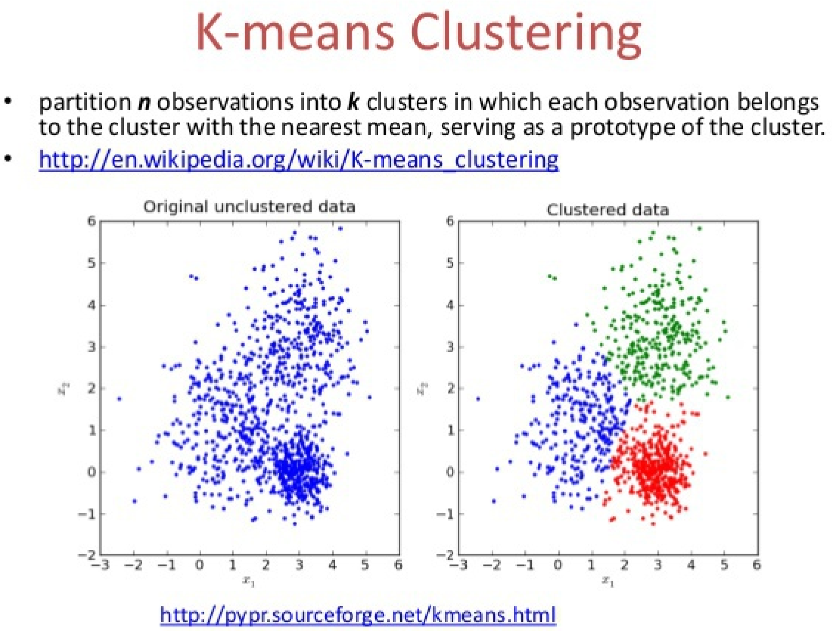
- Stephanie Yee: statistics expert.
- Tony Chu: data visualisation expert.

S. Learning: Ordinal Classification

The output variable is a categorical, discrete or nominal value, but in addition there is a non-quantifiable order between classes.



Unsupervised Learning: Clustering



K-means Clustering

- partition ***n*** observations into ***k*** clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
- http://en.wikipedia.org/wiki/K-means_clustering

http://pypr.sourceforge.net/kmeans.html

Unsupervised Learning: Clustering

The HR department of a company wants to categorise its employees into different groups to understand their behaviour better and treat them appropriately.

Id	Salary	Married	Car	Children	Rent / Prop	Syndicated	Leave	Seniority	Gender
1	1000	Y	N	0	Rent	N	7	15	M
2	2000	N	Y	1	Rent	Y	3	3	F
3	1500	Y	Y	2	Prop	Y	5	10	M
4	3000	Y	Y	1	Rent	N	15	7	F
5	1000	Y	Y	0	Prop	Y	1	6	M
...	...	...	...	...	...	...	...	...	...

Unsupervised Learning: Clustering

Knowledge gained → groups of employees:

	GROUP 1	GROUP 2	GROUP 3
Salary	1535	1428	1233
Married (N/Y)	77 %/23 %	98 %/2 %	0 %/100 %
Car (N/Y)	82 %/18 %	1 %/99 %	5 %/95 %
Children	0.05	0.3	2.3
Rent/Prop	99 %/1 %	75 %/25 %	17 %/83 %
Syndicated (N/Y)	80 %/20 %	0 %/100 %	67 %/33 %
On Leave	8.3	2.3	5.1
Seniority	8.7	8	8.1
Gender (M/F)	61 %/39 %	25 %/75 %	83 %/17 %

Unsupervised Learning: Clustering

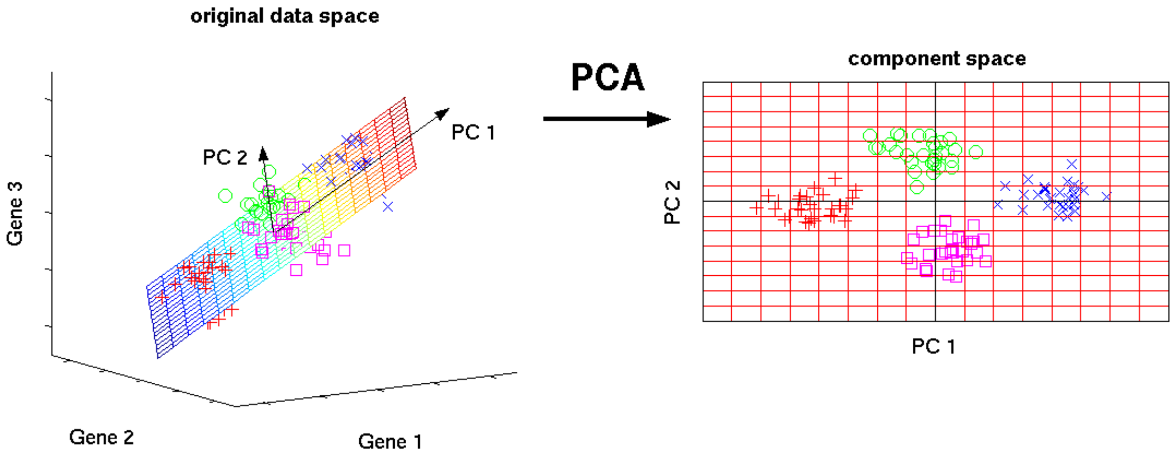
- Group 1: no children and rented housing. Few syndicated members. Many casualties.

- Group 2: without children and with a car. Highly syndicated. Few casualties. They are usually women and live in rented accommodation.

- Group 3: with children, married and with a car. Mostly male owners. Not very syndicated.

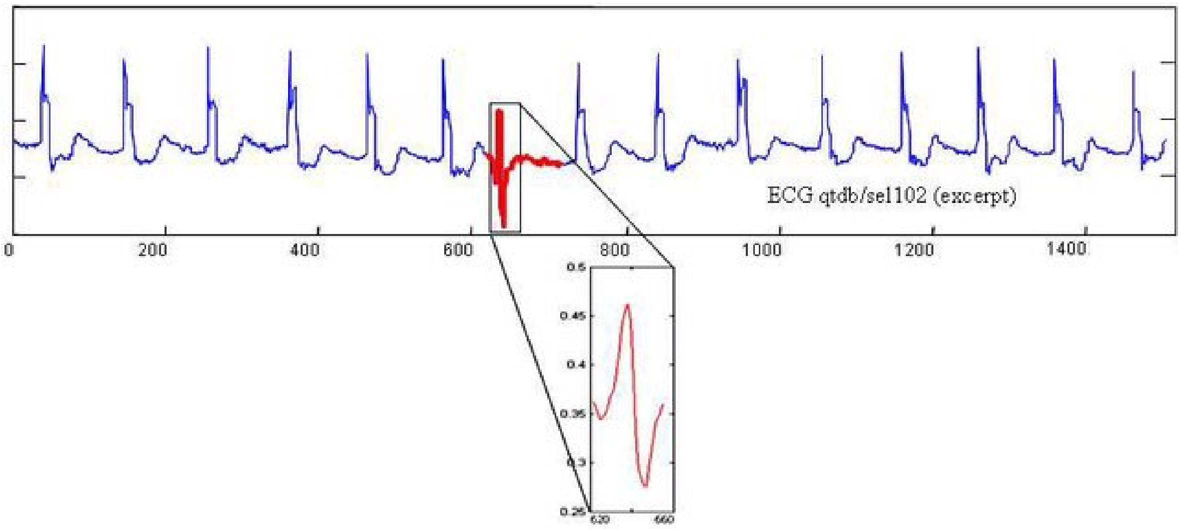
Dimensionality Reduction

This technique involves **reducing the number of variables** in a dataset while preserving the most relevant information. Its goal is to simplify analysis, reduce processing time, and facilitate the visualization of results.



Anomaly Detection

Anomaly detection is a technique that identifies unusual data or atypical behaviors within a dataset. These anomalies often represent significant events, such as fraud, technical failures, or unusual patterns.



Uns. Learning: Association rules

- A supermarket wants to obtain information about the shopping behaviour of its customers.
- It is thought that the service can be improved by placing certain products together, etc.
- Which products are usually placed together in shopping baskets? What is the probability that a person who buys product A will buy product B?

Uns. Learning: Association rules

Id	Eggs	Oil	Nappies	Wine	Milk	Butter	Salmon	Lettuce	...
1	Y	N	N	Y	N	Y	Y	Y	...
2	N	Y	N	N	Y	N	N	Y	...
3	N	N	Y	N	Y	N	N	N	...
4	N	Y	Y	N	Y	N	N	N	...
5	Y	Y	N	N	N	Y	N	Y	...
6	Y	N	N	Y	Y	Y	Y	N	...
7	N	N	N	N	N	N	N	N	...
8	Y	Y	Y	Y	Y	Y	Y	N	...

Uns. Learning: Association rules

Knowledge gained → association rules:

- IF nappies THEN milk = yes (100 %, 37 %).
- (a,b) = (confidence,support).
    - Confidence: percentage of times the rule is correct.
    - Support: percentage of occurrence of the rule in the data.

3. Data Science Process

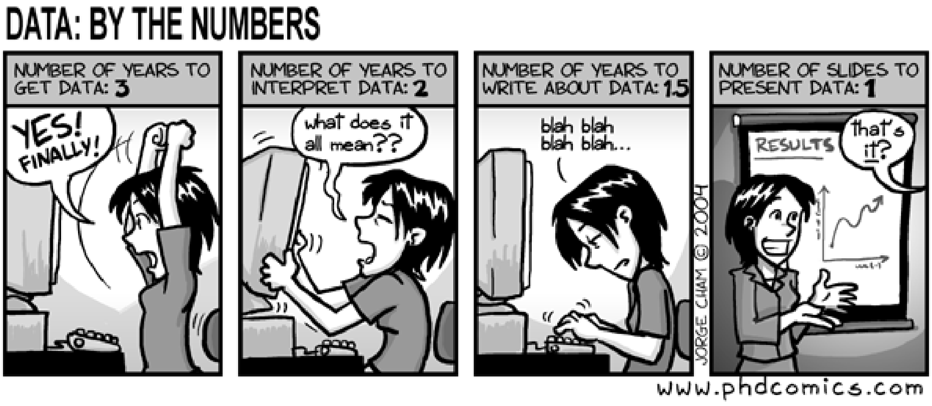
3.1. Stages of the Data Science Process

Stages of the Data Science process

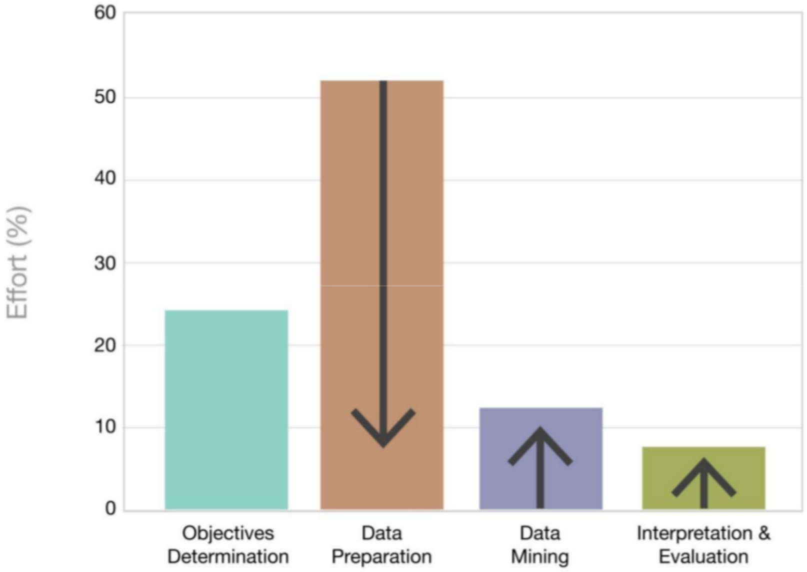
1. Integration and collection: Understanding the problem application domain, identifying a priori knowledge and creating a data warehouse.
2. Pre-processing: Data selection, cleaning, reduction and transformation.
3. Selection of the ML technique and its application.
4. Evaluation, interpretation and presentation of the results obtained.
5. Dissemination and use of the new knowledge.

Which stage takes the most effort?

Which stage in the Data Science process takes the most effort?


DATA: BY THE NUMBERS
NUMBER OF YEARS TO GET DATA: 3
YES! FINALLY!
NUMBER OF YEARS TO INTERPRET DATA: 2
what does it all mean??
NUMBER OF YEARS TO WRITE ABOUT DATA: 1.5
blah blah blah blah...
NUMBER OF SLIDES TO PRESENT DATA: 1
RESULTS
that's it?
JORGE CHAM © 2004
www.phdcomics.com

Which stage takes the most effort?



Time estimates in the analysis of a problem using Data Science techniques

3.2. Stage 2: Preprocessing

Stage 2: Preprocessing

Often, the results will depend more on the quality of the data in relation to the problem than on the ML part.

We will talk about noise in the data, relevance of variables... always with respect to an objective. A variable that we consider to be noise may be useful information for a different problem.

Although, this phase will be detailed in *Unit 3* of the subject, some of the tasks at this stage are:

- Feature data selection: feature extraction and feature selection.
- Data cleaning.
- Data transformation.
- Dimensionality reduction.

3.3. Stage 3: Machine Learning

Stage 3: Machine Learning

The objective of Data Mining/Machine Learning

To produce new knowledge by building/training a model based on the collected data, which is a description of the data and their relationships, and that serves to make predictions, better understand the data or explain past situations.

Stage 3: Machine Learning

Depending on the problem to be solved:

- What kind of knowledge are we looking for? Predictive, descriptive...
- Which technique is the most appropriate? Classification, regression, clustering, association rules, recommender system...
- What type of model? In classification: decision trees, neural networks, SVM...
- Other requirements: interpretability, fuzzy logic...
- Which learning algorithm is the most suitable? Gradient descent, evolutionary algorithm, analytical algorithm...

3.4. Stage 4: Evaluation

Stage 4: Evaluation

Evaluation, interpretation and presentation of results:

- During the analysis, we will generate several model hypotheses. Which are the most valid?
- The no free lunch theorem tells us that no one method will be the best in all cases.

- Validation criteria:
    - Accuracy.
    - Interpretability.
    - Novelty of the knowledge found.

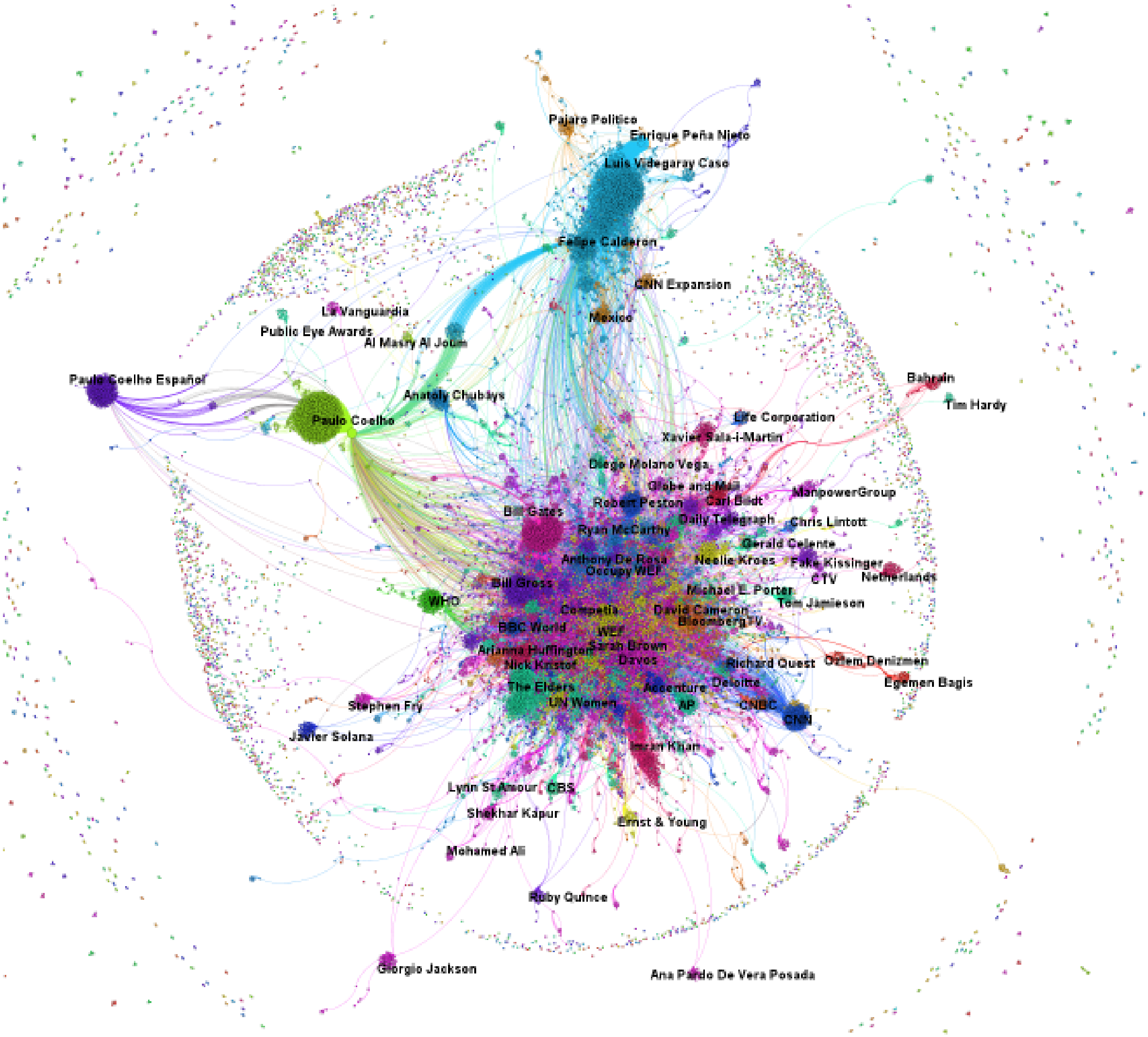
Stage 4: Evaluation

- Evaluation techniques / experimental designs → at least the data set is divided into two:
    - Train: extract knowledge (fitting the model).
    - Test: test the model with data not examined during training.
- Evaluation measures, depending on the task:
    - Classification: accuracy, geometric mean of sensitivities, ROC curve...
    - Regression: root mean squared error...
    - Clustering: cohesion and separation metrics...
    - Association rules: confidence, support...
- Can the model be interpreted and/or visualised (decision trees vs SVM/Neural Networks)?

3.5. Stage 5: Dissemination

Stage 5: Dissemination

- Contrast with prior knowledge.
- System observation.
- Feedback from the system (online learning).
- More and more artists are becoming involved in data visualisation.



4. Applications

4.1. General Applications

Applications

- Medicine:
  - Diagnosis of diseases (e.g. diagnosis of abdominal pain).
  - Predict whether a chemical compound causes cancer.
  - Predict whether a person could potentially have a disease from their DNA.
  - Characterisation of melanomas from dermoscopic images.
  - Diagnosis of Parkinson's disease through functional imaging.
- Finance and banking:

    - Obtaining fraudulent credit card patterns.
    - Delinquency forecasting (loans).

Applications

- Market analysis:
    - Market basket analysis.
    - Market segmentation.
- Computer Science:
    - Spam detection.
    - Product recommendation in e-shops (amazon.com).
    - Automatic classification of web pages for directories.
    - Character recognition, voice recognition, etc.
- Agronomy, farming:
    - Field production forecasting.
    - Analysis and prediction of efficiency in different types of livestock farms.
    - Weed maps.

Applications

- Renewable energies:
    - Prediction of electricity demand, gas demand, etc.
    - Design of wind farms.
- Predictive microbiology:
    - Estimation of growth parameters of microorganisms in food.
- Insurance and private health:
    - Identification of potentially expensive customers.
- Education:
    - e-learning (detection of abandonment, prediction of results...).
- Astronomy:
    - Classification of celestial bodies (SKYCAT).

4.2. Unusual applications

Applications

- *Seal Mobile* is trying to recognise a mobile phone user with accelerometer information (how you hold the phone and move it).
- The *Online Privacy Foundation* sponsored a competition to see if it is possible to predict whether someone is a psychopath by analysing their Twitter usage.
- *Marinexplore* and *Cornell University* are trying to identify whales in the ocean based on audio so that ships can avoid colliding with them.
- *Dunnhumby* and *hack/reduce* are trying to predict in advance whether a product launch will be successful or not.
- The *University of Oregon* seeks to determine the species of birds present in an area given an audio file.

Applications

- Archaeologists are using ML models on SAR (Synthetic Aperture Radar) data to detect buried ancient structures beneath desert sand, reconstructing 3D terrain with sub-meter accuracy.
- Researchers in Australia applied quantum machine learning to optimize semiconductor design, modeling electrical resistance in chips with higher precision than classical methods.
- Beekeepers are using recurrent autoencoders and sensors inside hives to detect anomalies in bee behavior, predicting swarming or disease events in advance.
- *MENACE* (Matchbox Educable Noughts and Crosses Engine) was a physical reinforcement learning system built from 304 matchboxes in the 1960s to learn how to play tic-tac-toe.
- *Google Deep Dream* uses convolutional neural networks to generate psychedelic, dream-like images by enhancing patterns recognized by the model.

5. Software

5.1. Software

Machine Learning software

- We live in a time of the explosion of Machine Learning in different applications.
- Even terms such as *Big data* or *Deep Learning* are becoming more than commonplace in the news headlines.
- This has led to the existence of many Machine Learning tools.

Machine Learning software

Two types:

- At the user level, software packages such as *Weka*, *Orange* or *Knime* allow us to load databases, preprocess them and apply Machine Learning algorithms without requiring extensive programming knowledge:
  - Easier to use.
  - Difficulties in automating processes.
- As programmers, we can use high-level languages (*R* or *Python*) together with different libraries to exploit the full potential of Machine Learning:
  - Somewhat more complex to use (requires programming knowledge).
  - Automation and extension are possible.

5.2. Scikit-learn

What is scikit-learn?

- It is a library that provides a comprehensive set of supervised and unsupervised learning algorithms through a consistent Python interface.
- Released under BSD license and distributed on many Linux systems, it favours commercial and educational use.
- This library is built on top of *SciPy* (Scientific Python), which must be installed before use, including:
  - *Numpy.*
  - *Matplotlib.*
  - *Sympy.*
  - *Pandas.*

Characteristics of scikit-learn

- This library focuses on data modelling rather than loading and manipulating data, for which we use *NumPy* and *Pandas.*
- Some of the things we can do with *scikit-learn* are:

- Clustering.
- Cross-validation.
- Test datasets.
- Dimensionality reduction.
- Ensemble methods.
- Feature selection.
- Parameter tuning.

Advantages of using scikit-learn

- Consistent interface to Machine Learning models.
- Provides many configuration parameters.
- Exceptional documentation.
- Very active development.
- Community.

6. Conclusions

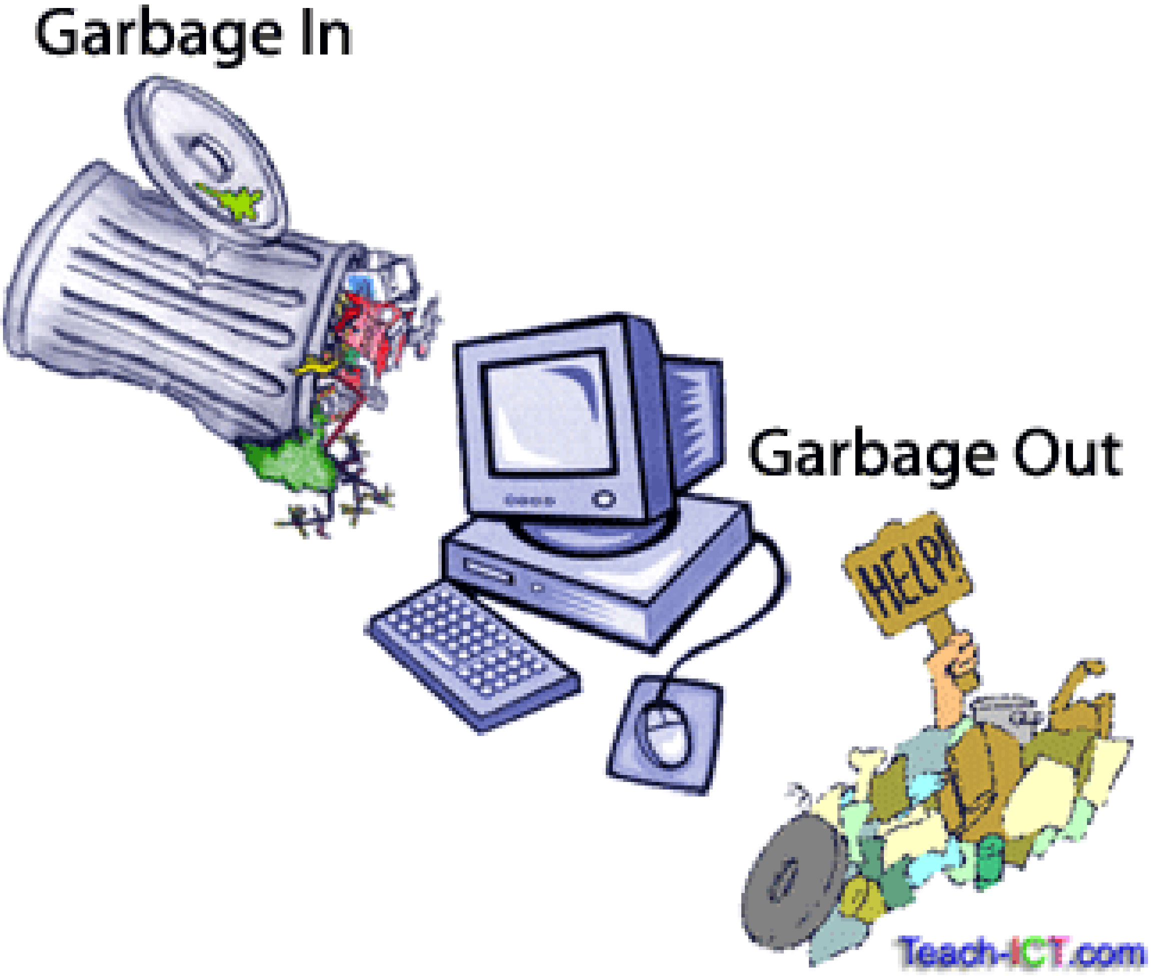
What have we seen?

- Introduction to Data Science and Machine learning.
- Overview of Machine Learning techniques.
- Application of these techniques to the real world.
- Some Machine Learning models.
- We have highlighted the importance of preprocessing in the Data Science process.
- We highlighted the importance of knowing the data and the models to use.

Key questions

- The more data, the more accurate the classification model will be.
- The selection of the best method is key to success or failure.
- The parameters of the methods have to be optimised.

- ML methods can be as good as the data, not better. Care must be taken in the process of data selection and data cleaning.



“Your analysis is as good as your data”

Key questions

- It is necessary to understand the variables in our model.

- Selecting an error function associated with the problem is vital.

- Ensure proper treatment of training and testing in order to analyse the behaviour of the model on generalisation data (and detect overfitting).

- Learning to deal with unstructured data (text, time series, images).

- Translating real-world problems to ML problems.

References

- Benítez-Iglésias, R., Escudero-Bakx, G., Kanaan-Izquierdo, S., Masip-Rodó, D. (2014). Inteligencia artificial avanzada. Editorial UOC.
- Bishop, CM. (2006). Pattern recognition and machine learning. Springer.
- Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow (Second Edition). O'Reilly.
- James, G.; Witten, D.; R Hastie, T.; Tibshirani, R.; Taylor, J. (2023). An Introduction to Statistical Learning with Applications in Python. Springer.
- Mohri, M., Talwalkar, A., Rostamizadeh, A. (2019). Foundations of machine learning. Cambridge: The MIT Press.

Questions?