VIETNAMESE SPRING SCHOOL
in
STATISTICS AND MACHINE LEARNING

CHAPTER I. Introduction

Agnès LAGNOUX

lagnoux@univ-tlse2.fr

LECTURE SUPPORTS

All the files of this lecture are available on my webpage :

`https://perso.math.univ-toulouse.fr/lagnoux/enseignements/`

You will find them at the bottom of the page.

LECTURE OUTLINE

1. Introduction
2. Supervised classification
   - Linear regression
   - k-nearest neighbors
   - Discriminant factor analysis
   - Naive Bayesian
   - Logistic regression
3. Unsupervised classification
   - Hierarchical clustering analysis
   - k-means

Introduction to learning

Supervised and unsupervised learning

Introduction to classification

Plan

Introduction to learning

Supervised and unsupervised learning

Introduction to classification

Introduction

Artificial intelligence (AI) and machine learning are all around us, from online recommendation systems and voice assistants to facial recognition and autonomous driving...

Questions arise...

- How do they really work ?
- How do they learn from our behaviors, preferences, and interactions ?

What is AI ?

AI is a field of computer science and mathematics that brings together a set of algorithmic techniques and theories for creating machines that mimic human intelligence.

Its aim is to reproduce intelligence in order to be able to solve complex problems. Humans attempt to replicate their intelligence in order to automate certain tasks.

This involves modelling human intelligence as a phenomenon, as could be done in physics, chemistry or biology. AI is a rapidly expanding field, using theory and applications in a wide range of fields, including probability theory, neuroscience, robotics, game theory, healthcare, and transportation.

What is AI ?

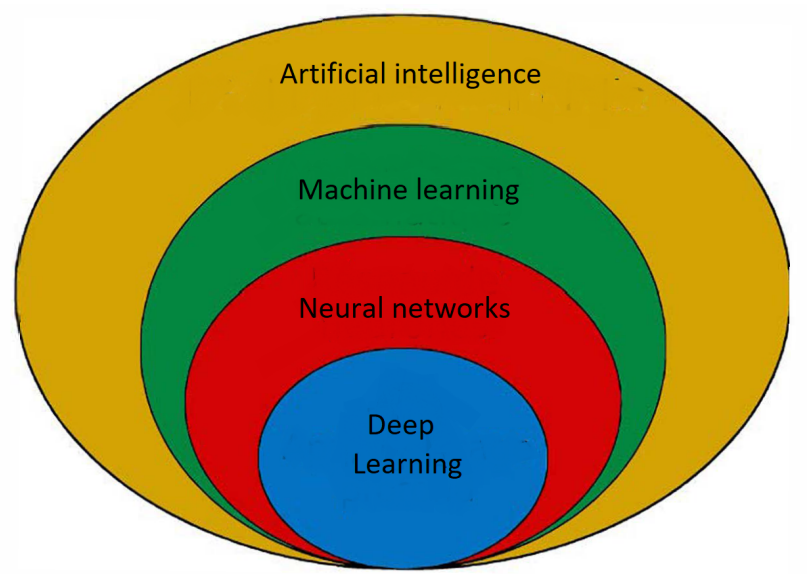
The aim of AI is to simulate human intelligence, and in particular to learn a wide range of tasks. There are two possible ways of learning :

- Rote learning consists in explicitly memorizing all possible examples so as to be able to play them back.
- The aim of generalizing learning is to extract implicit rules from a large number of examples in order to reapply them to new situations never encountered before.

Rote learning is relatively easy for a machine, as long as the examples are available. On the other hand, learning by generalization is difficult, as it requires the extraction of rules that are not explicitly mentioned in the examples. This challenge lies at the heart of machine learning.

What is AI ?

The field of AI is divided into several interlocking subfields. Not all AI techniques are necessarily machine learning. learning.



In particular, prediction rules can be supplied directly to the machine in the form of a sequence of if-then ? conditions. These systems, known as expert systems, were used extensively in the 1980s to solve given tasks with a high degree of accuracy.

What is machine learning ?

Machine learning is a sub-domain of AI, which involves learning from experience or from a database of implicit rules to answer to a given problem. This field focuses on the statistical analysis of training data. Historically, this branch has been defined as the development of machines capable of learning without having been explicitly programmed to learn a task.

It's important to note that there's a big difference between AI and machine learning. Machine learning is a form of AI that consists of a system that improves with experience, whereas AI can be a simple set of rules and heuristics.

Example : automatic recognition of handwritten numbers

These characters vary widely in shape, orientation and thickness... In this case, it is difficult to write an explicit list of rules for discriminating digits. So we're going to implement an algorithm that extracts implicit rules based on example data. These rules can then be applied to new digits in order to recognize them.

The phases of machine learning

Generally speaking, machine learning algorithms are divided into several phases.

- Training phase (or learning phase) : the chosen model is subjected to a large number of significant examples. The system then seeks to learn implicit rules based on this data (called training data). This training phase generally precedes the use of the model, although some systems continue to learn indefinitely if they have feedback on the results (this is called on-line learning).
- Inference phase : the trained model can be used on new inputs. Inputs provided during the inference phase can be processed even if they were not seen by the model during the learning phase. Indeed, thanks to the extraction of implicit rules, the model can generalize to unknown inputs.

Types of machine learning

Machine learning uses different types of learning, with supervised learning and unsupervised learning playing a prominent role. We will see that there are numerous algorithms using a variety of mathematical models.

Deep learning is a set of techniques that use neural networks to solve complex problems. These techniques are widely used, particularly in the fields of image processing, time series processing (speech recognition...).

Reinforcement learning consists in learning by interacting with the agent's environment. A system of rewards reinforces good choices and penalizes bad ones.

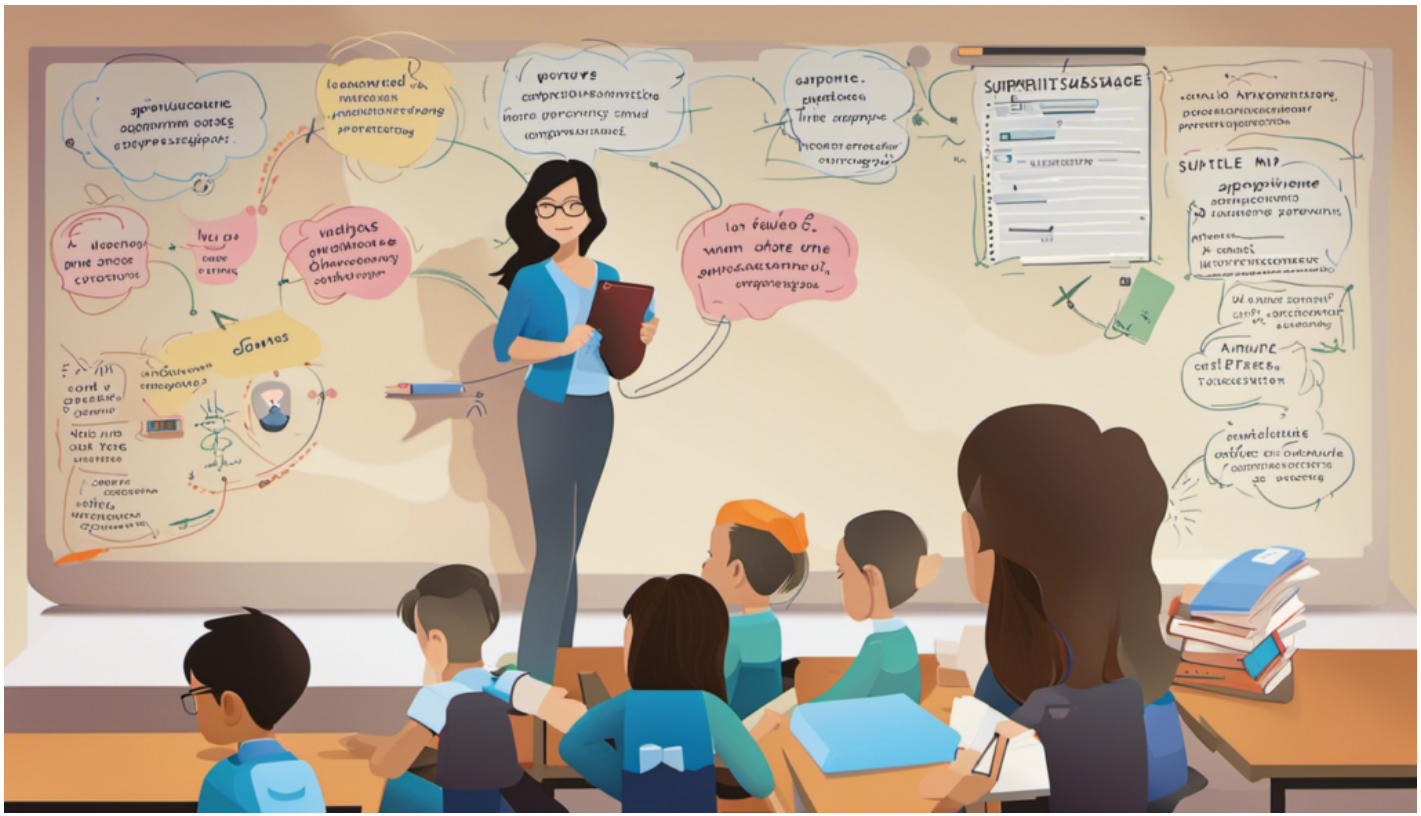
Plan

Introduction to learning

Supervised and unsupervised learning

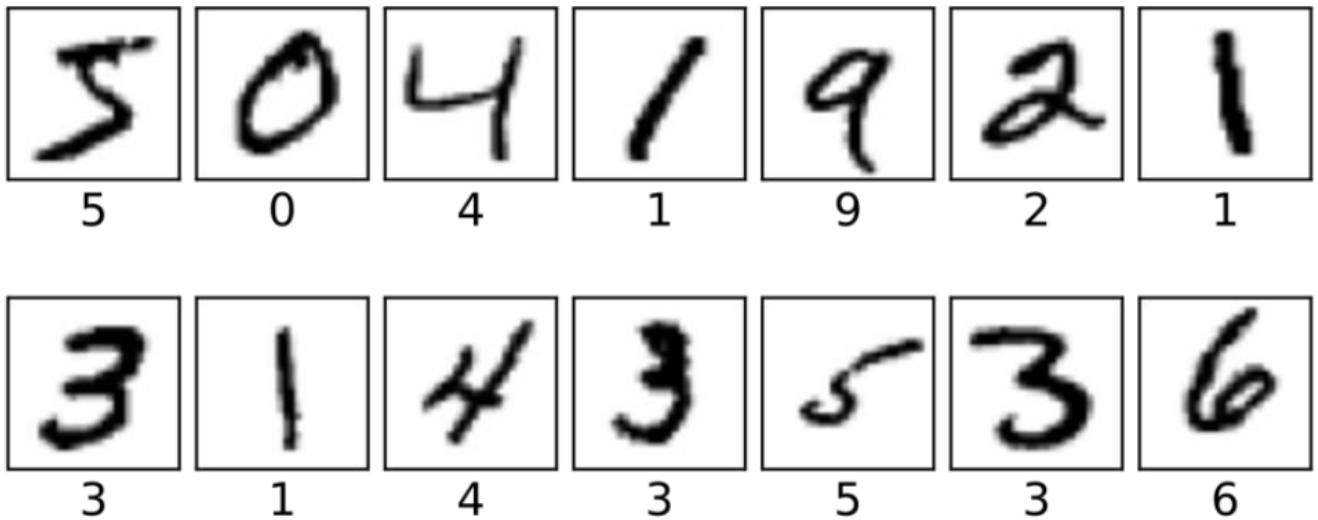
Introduction to classification

Supervised learning



Supervised learning

We speak of supervised learning when we have labeled training data, i.e. when we know the desired output. Noting the n inputs x_i and the associated target outputs y_i, we have the data set $(x_i, y_i)_{i=1,\cdots,n}$.



The aim is to train the chosen model so that it can correctly predict the output for unlabeled inputs.

Supervised learning

Supervised learning is generally used for regression or classification.

- Regression is used when the output to be predicted can take continuous values, i.e. a real variable.

  Example : an algorithm predicting the power consumption of an installation or the stock market price.

- Classification is the task of choosing a class (value) from all those possible.

  Example : An algorithm predicting the handwritten number on the input image, or an algorithm classifying a tumor as "benign" or "malignant".

Supervised learning

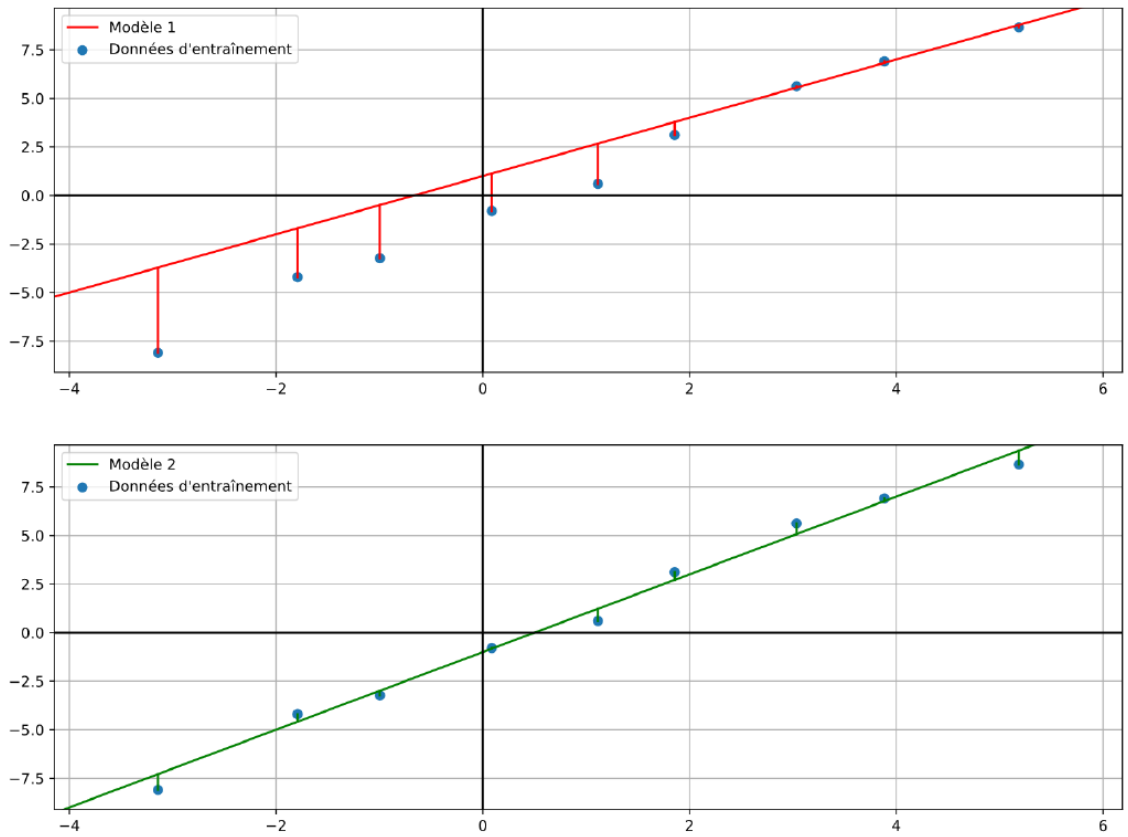


Figure – An example of linear regression

Supervised learning

It's important to note that these two families, classification and regression, are not exhaustive.

The most frequently used examples of models between regression and classification are found in automatic language processing, where words can be modeled as combinations of letters, giving rise to very high-dimensional vectors.

Classic supervised learning algorithms include linear regression, nearest neighbor algorithms, discriminant factor analysis, logistic regression, neural networks, decision trees, random forests and vector support machines.

Unsupervised learning

Unsupervised learning

Unsupervised learning is the term when unlabeled data are at stake. We therefore have input data for which we don't know the associated output. The data set is therefore $(x_i)_{i=1,\dots,n}$ and the aim of the system is to identify features common to the training data.

Unsupervised learning is mainly composed of clustering algorithms. These algorithms seek to separate the input data into a given number of groups.

☞ Each element in the group must have characteristics close to those of elements in the same group, but relatively distant from those of other groups.

Unsupervised learning

These algorithms group entries into families and automatically label them. For example, a clustering algorithm can be used to group patients together to predict possible reactions to certain treatments.

Among the most common unsupervised learning algorithms are the k-means algorithm, hierarchical ascending classification, principal component analysis, DBSCAN, singular value decomposition and some neural networks.

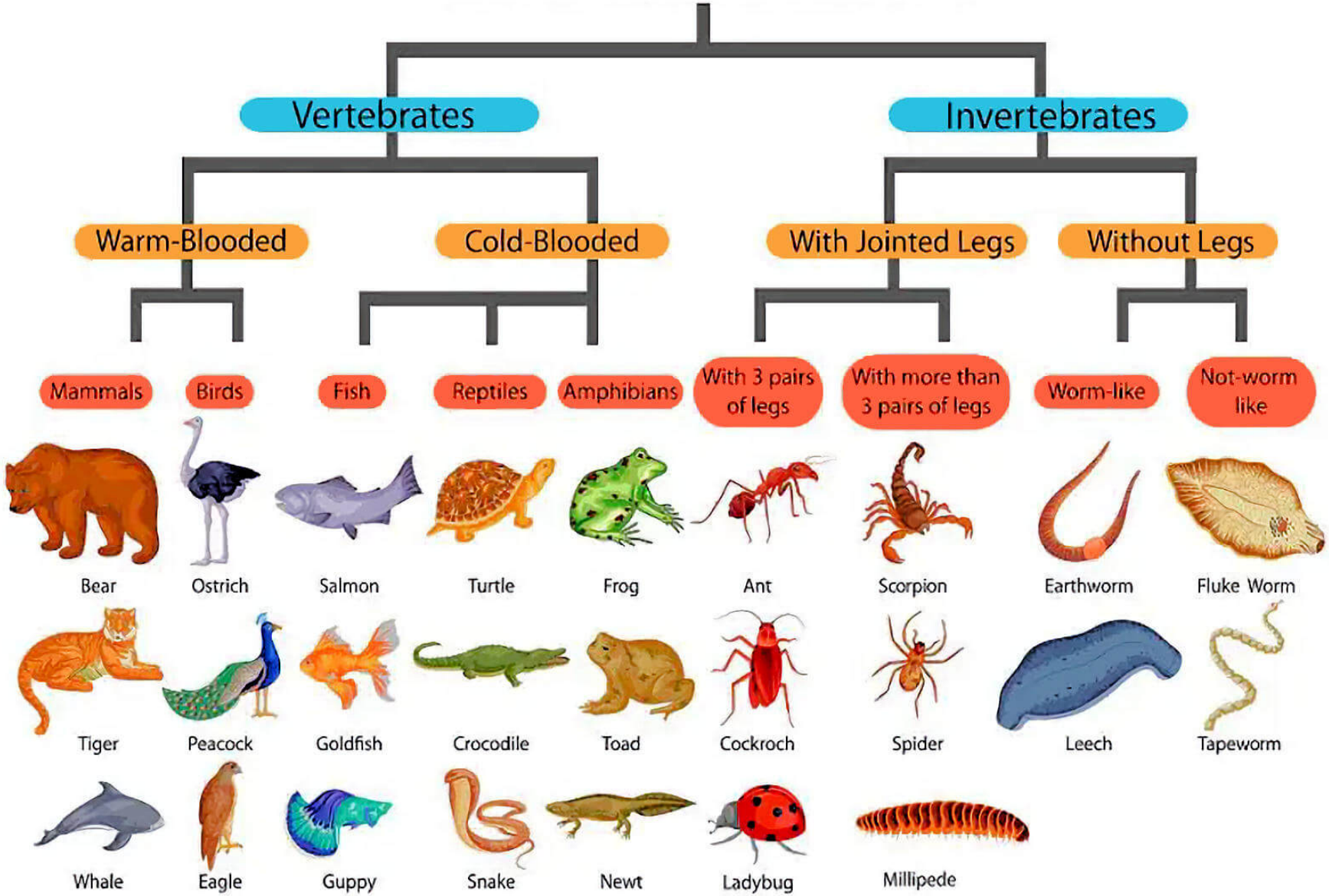
Plan

Introduction to learning

Supervised and unsupervised learning

Introduction to classification

classification animale



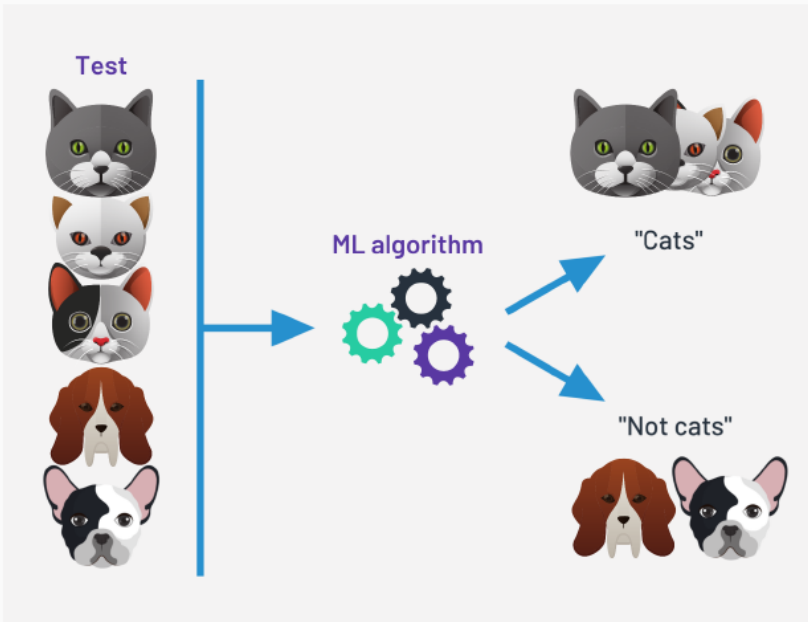
Introduction to classification

The aim of classification is to group (partition, segment) n observations into a number of homogeneous groups or classes. There are two main types of classification :

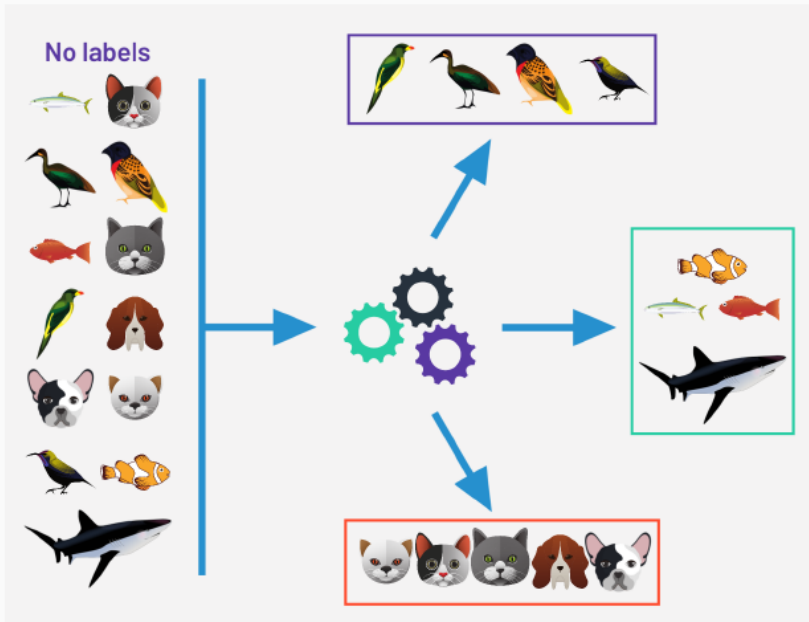
- supervised classification, often referred to simply as classification ;
- unsupervised classification, sometimes called partitionning, segmentation, or clustering.

Introduction to classification

Images from www.g2.com/articles/supervised-vs-unsupervised-learning



Supervised **classification**



Unsupervised **classification**

Introduction to classification

In supervised classification,

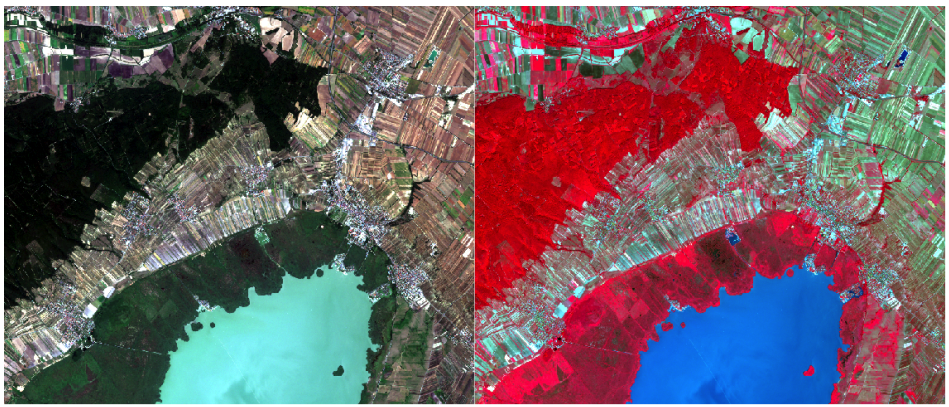
- we already know how many groups exist in the population ;
- we know the group to which each observation in the population belongs ;
- we want to classify the observations in the right groups based on different variables.

We can then use a classification rule to predict the groups to which new observations belong.

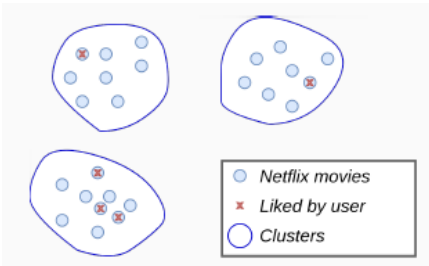
Introduction to classification

Classic examples of applications are :

- identify whether a bank transaction is fraudulent or not ;
- recognizing handwritten numbers ;
- image segmentation ;



- identify the type of cancer a patient has ;
- recommendation systems.



Introduction to classification

There are several families of supervised classification methods. The most common are :

- nearest neighbor method ;
- discriminant factor analysis ;
- classification trees ;
- logistic regression ;
- naive Bayesian ;
- neural networks ;
- vector support machines.

Introduction to classification

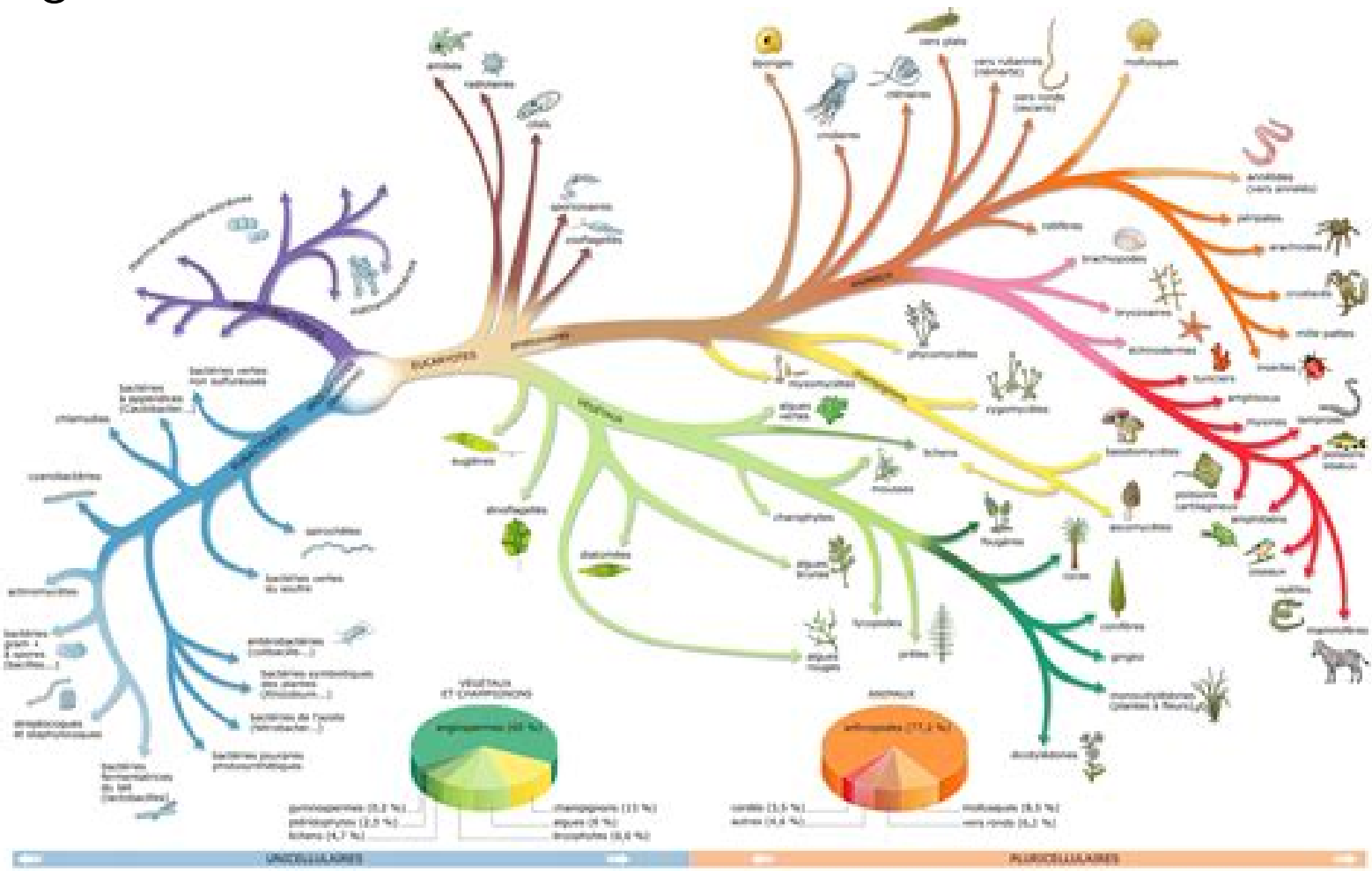
In unsupervised classification,

- in general, we don't know how many groups exist in the population ;
- we don't know the group to which each observation in the population belongs ;
- we want to classify observations into homogeneous groups based on different variables.

Introduction to classification

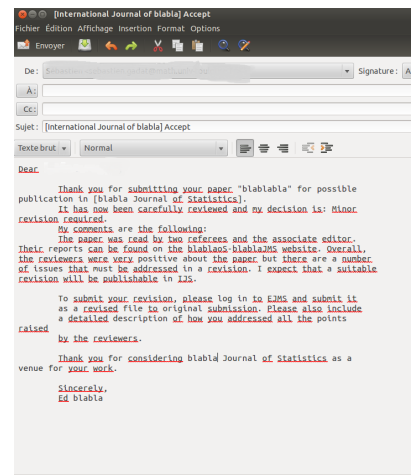
Typical applications are numerous. For example :

- in biology : the development of the taxonomy of living organisms ;



Introduction to classification

- in psychology : the determination of personality types present in a group of individuals ;
- en text mining : partitioning e-mails or texts according to subject.



Introduction to classification

There are several families of unsupervised classification methods. The most common are :

- hierarchical classification ;
- non-hierarchical classification, such as the k-means method ;
- density-based classification ;
- classification based on statistical/probabilistic models, e.g. a mixture of normal distributions.

Some sources of this lecture

- Marie Chavent's lectures
  `https://marie-chavent.perso.math.cnrs.fr/`
- Juliette Chevallier's lectures
  `https://juliette-chevallier.pages.math.cnrs.fr/`
- Example ROC curve
  `https://datatab.fr/tutorial/roc-curve`