Université de Strasbourg

Introduction to data mining

Maja Temerinac-Ott
temerinacott@unistra.fr

ICUBE

Slides adopted from Germain Forestier

Why data mining?

- Growing availability of huge amounts of data
  - Technology is available
  - Data collection: barcode, scanners, satellites, logs servers, etc.
  - To help store: database, data warehouses, digital libraries, www
- Why should we extract knowledge?
  - Economic necessity
  - E-commerce
  - High degree of competition
  - Customization, customer loyalty, market segmentation
- Customer data
- Digitalization of text, images, video, voice, etc.
- Internet and online catalogs

Why data mining?

- The data quantity is too big to be treated manually or by classical algorithms:
  - The number of entries is millions to billions
  - Multi-dimensional data
  - Heterogeneous sources of data
- The user is full of data, but does not know how to understand it:
  - “The greatest problem of today is how to teach people to ignore the irrelevant, how to refuse to know things, before they are suffocated. For too many facts are as bad as none at all." (W.H. Auden)
- What do we need?
  - Extract interesting and useful knowledge from the data : rules, regularity, irregularities, patterns, constraints

Data mining

- Extraction of implicit original (non-trivial) information, previously unknown and potentially useful from databases:
  - **Not trivial**: otherwise knowledge is not useful
  - **Implicit**: hidden knowledge is difficult to observe
  - **Unknown until now**: obvious!
  - **Potentially useful**: usable, understandable
- Whole process of discovery and interpretation of regularity in data
- Other names:
  - Knowledge Discovery in Databases (KDD)
  - Knowledge extraction
  - Data/pattern analysis
  - Data Analytics
  - Big Data

Group Task 1

- Form groups of 4-5 people
- You will get a list of steps in the data mining process

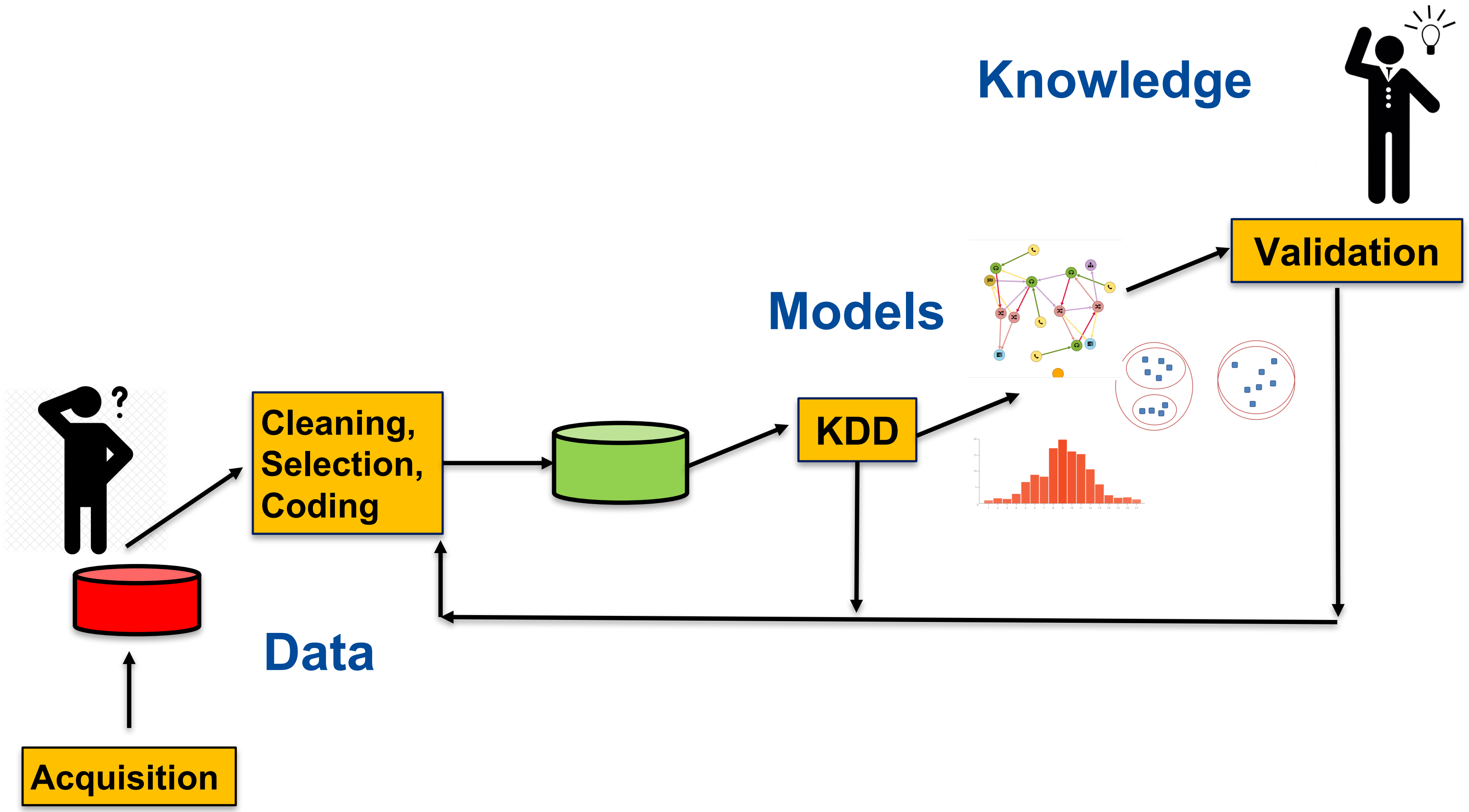
- Please order the steps in chronological order!

- You have 5 min for the task.

Data mining process

1. Identify the problem
2. Find data
3. Clean the data
4. Coding of data, actions on variables
5. Search for a model, knowledge, etc.
6. Validation and interpretation of the result, with possible return to the results in previous steps
7. Integration of new knowledge

Data mining process



Preparation of the data

- Existing or needed data:
  - Files: information contained in one or more independant files
  - Relational Database: information contained in several files by a common `key ‘
  - Database transaction
- Data cleaning:
  - duplicates, input errors, outliers, missing information (ignored observations, average values, mean values on class, regression, etc.)

Preparation of the data

- Data warehouses: collection of data collected from multiple heterogeneous sources
- Data is recorded, cleaned, transformed and integrated
- Usually modeled by a multidimensional data structure (cube):
  - data is structured along several lines of analysis (dimensions of the cube) such as time, location, etc.
  - A cell is the intersection of different dimensions
  - The calculation of each cell is carried out at loading
  - The response time is thus stable whatever is requested

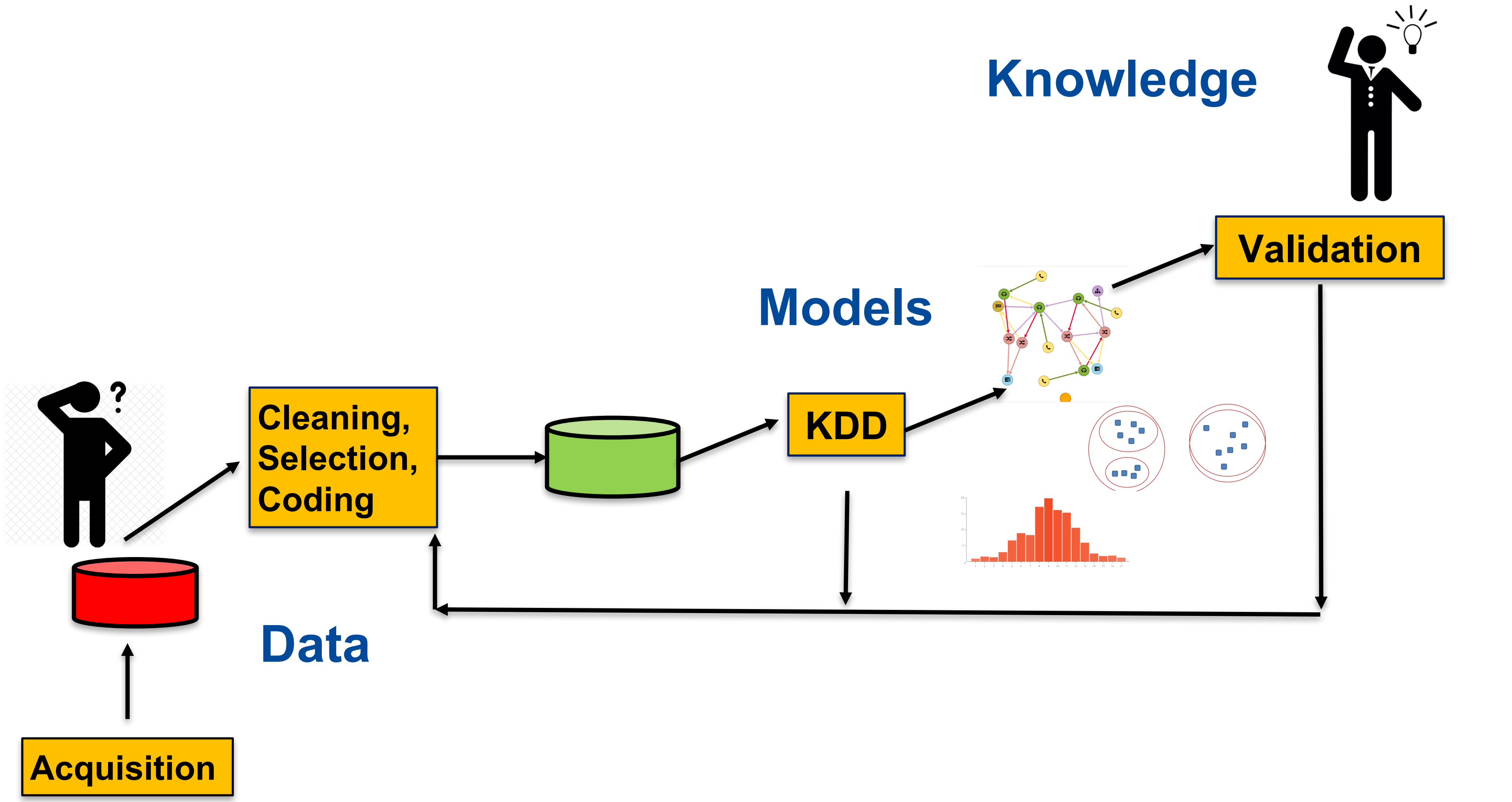
Preparation of the data

- The cubes are well suited for quick searches and analysis of data: On-Line Analytical Processing (OLAP)
  - What is the number of pairs of shoes sold by the store "OnVendDesChaussuresIci" in May 2016 AND Compare the sales with the same month of 2015 and 2014?
  - What are the components of production machines that have more unpredictable incidents during the period 2015-2016?
- Responses to OLAP requests may take a few seconds to several minutes.

Preparation of the data

- Selection of the data
  - Data sampling
  - Selection of sources
- Dimensionality reduction:
  - Selection or transformation of attributes
  - Weighting
- Coding:
  - Aggregation (sum, average), discretization, coding of attributes discrete, unification of benchmarks and standardization

Data mining process



Data mining

- Goal: Learn something new!
  - **Concepts**: regrouping of data based on shared characteristics
  - **Associations**: correlations between attributes or data
- Principles:
  - Getting the highest level of abstraction possible
  - Rules or truths that are the basis for other truths

Data mining

- Different approaches:
  1. **Estimation**:
     - create a model that best describes a prediction variable for real data
  2. **Classification**:
     - create a function that classifies an element among several pre-existing classes
  3. **Clustering** (regrouping):
     - search to identify a finite set of categories or groups which can describe the data
  4. **Dependency modelling**:
     - find a model that describes significant dependencies between variables

Data mining - learning

- **Supervised learning**:
  - Inductive model where the learner considers a set of grouped examples representative for the learning task (class of belonging, ownership, etc.): the examples are labeled beforehand
- **Predictive data mining**:
  - Divide / group instances into special classes for future predictions
  - Predicting unknown or missing values
- **Algorithms**:
  - Decision trees, classifications, genetic algorithms, regression (linear and non-linear)

Data mining - learning

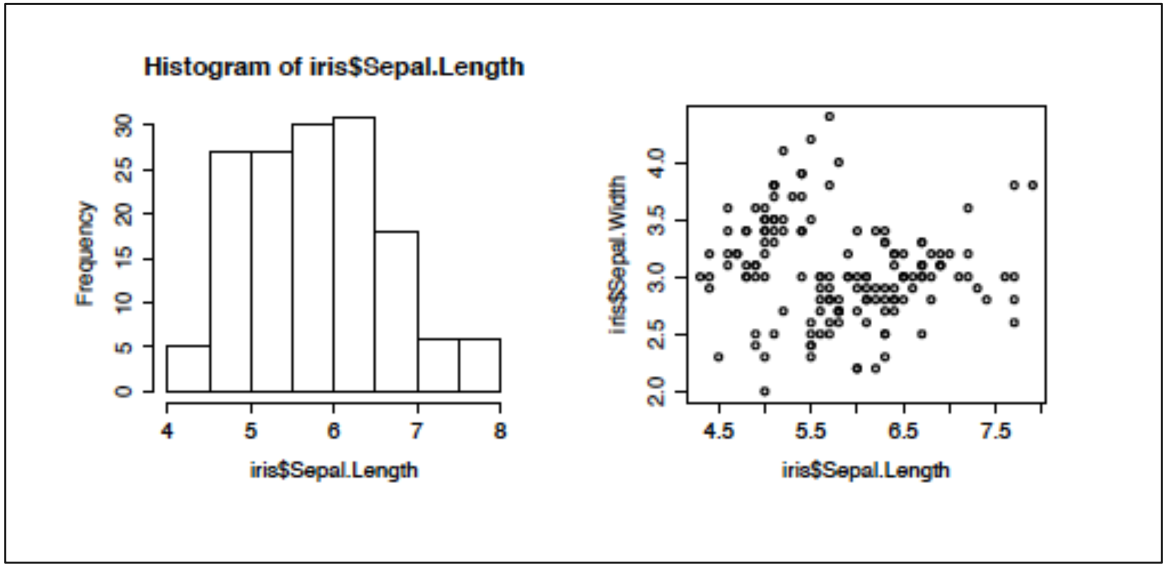
- **Induction**:
  - It is a commonly used technique
  - Generalization of an observation or reasoning established from singular cases.
  - It consists in drawing conclusions from a series of facts
- **Example**:
  - Induction:
    - water, oil and milk freeze when they are cooled, we will infer that all liquids must freeze, provided the cold is rather intense
  - Deduction:
    - all liquids are likely to freeze; so, if mercury is a liquid, it can freeze

Data mining - learning

- **Non-supervised learning**:
  - Construction of a model and discovery of relations is given without reference to other data
  - There is no prior information on the data
- **Explanation**:
  - Grouping instances into special classes based on their resemblance or the sharing of properties. The classes are unknown and are therefore created: they are used to explain "or summarize the data
- **Algorithms:**
  - Segmentation, grouping, discovery of associations and rules

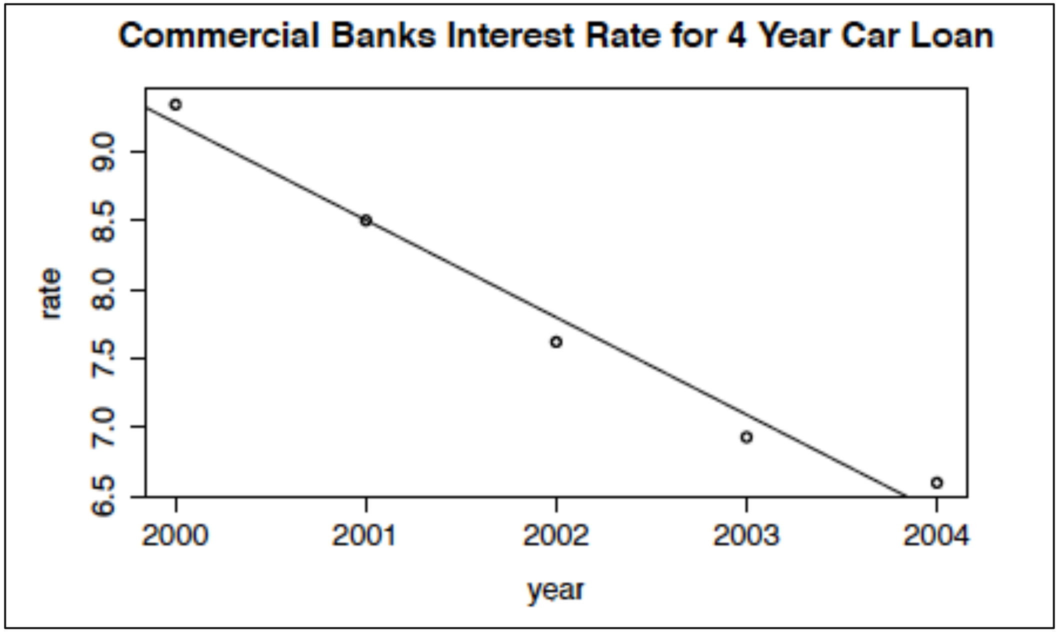
Data visualization

- Obtain a visual representation of the data
- Not always possible depending on the data type
- Not always possible depending on the amount of data



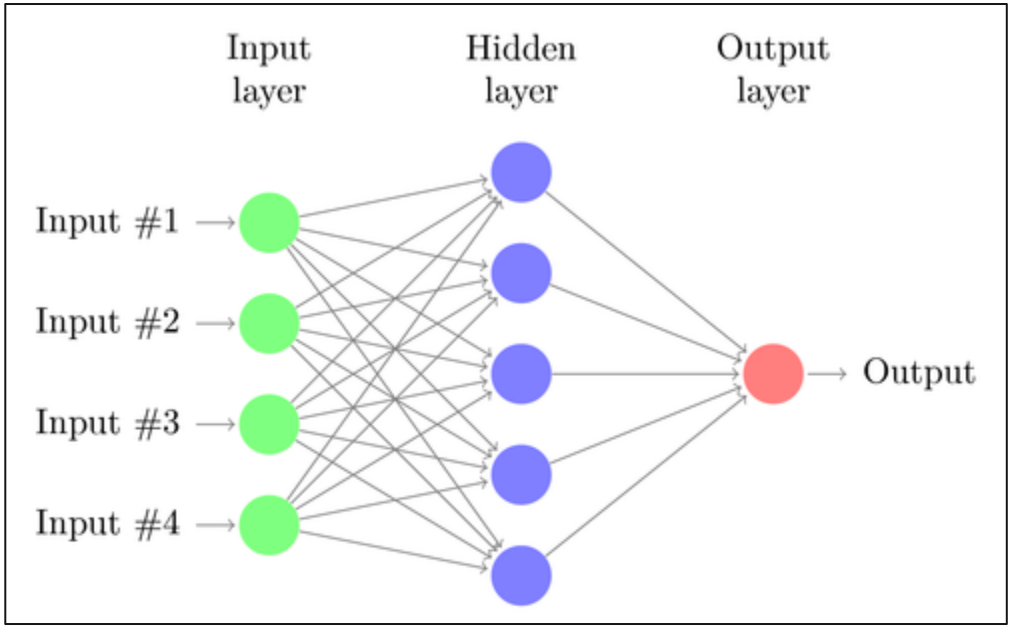
Approach 1: Estimation

- **Regression**:
  - Analyze the relationship of one variable vs. one or more other variables
  - Least-squared method



Approach 1: Estimation

- **Regression**:
  - Analyze the relationship of one variable vs. one or more other variables
  - Neural network

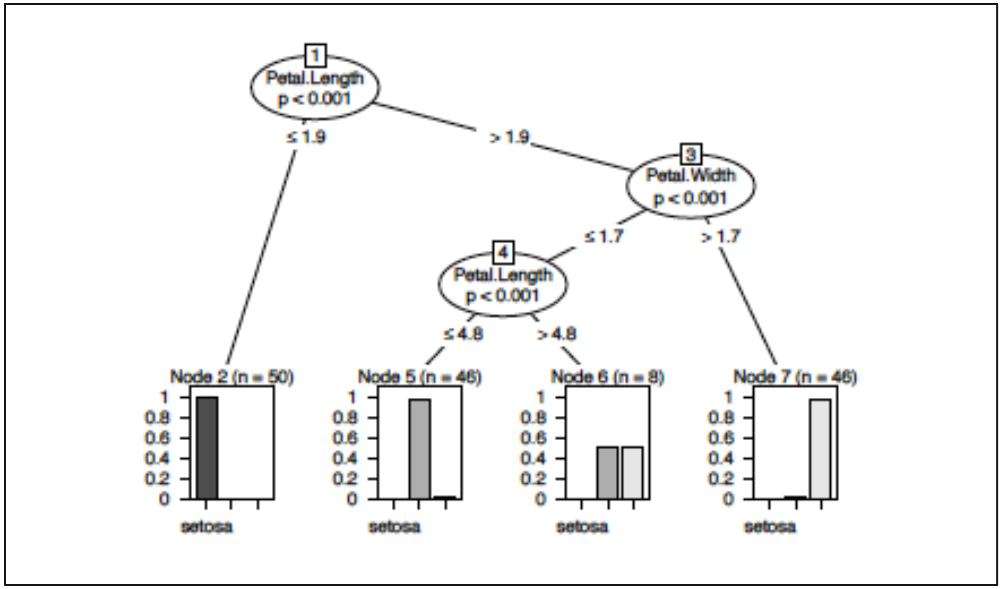


Approach 2: Classification

- Division of the data set into disjoint classes
- **Goal**:
  - search for a set of predicates characterizing a class of objects and which can be applied to unknown objects in order to identify their class of belonging.
- **Principales techniques**:
  - Decision trees
  - Bayesian classifier
  - k-nearest neighbor
  - Neuronal network
  - SVM
  - Genetic algorithm

Approach 2: Classification

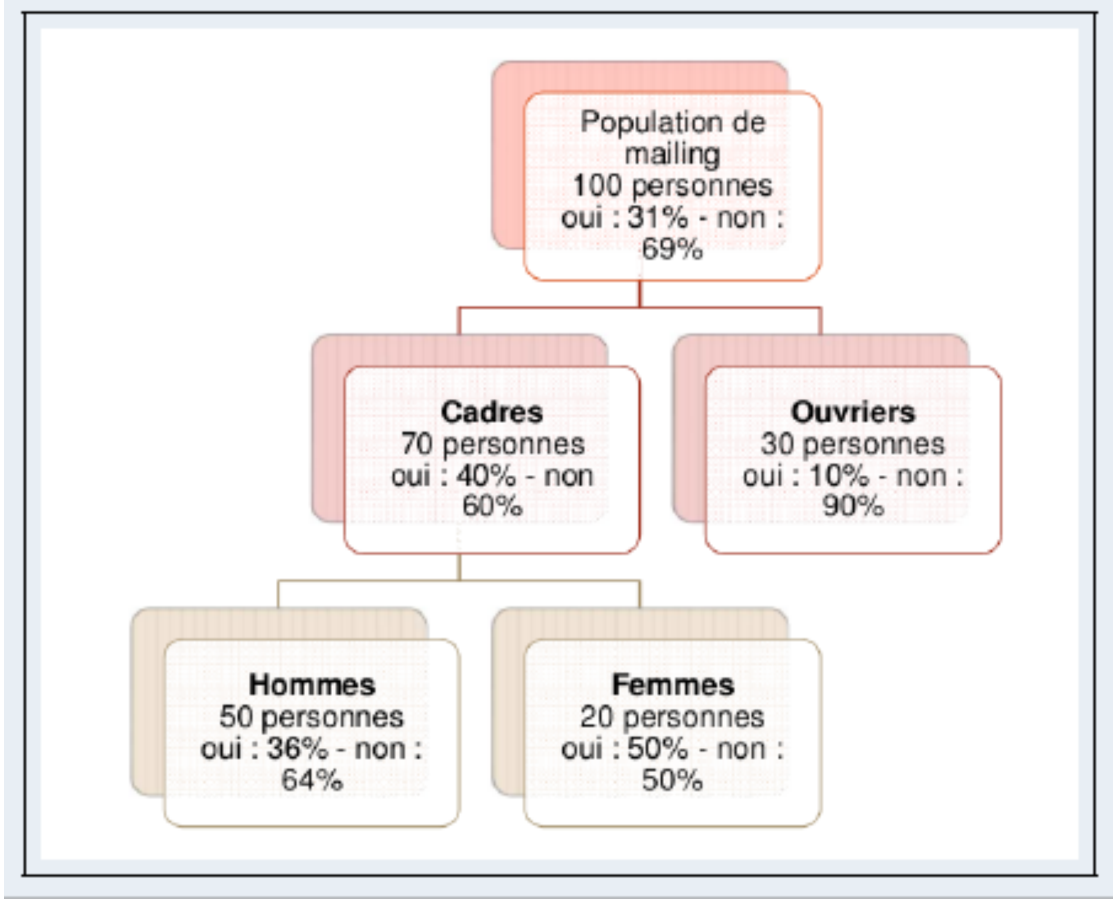
- **Decision trees**:
  - Classify objects into subclasses by hierarchical divisions
  - Automatic construction from a sample
  - There are several techniques to build the tree



- **Decision Tree** Example:
  - You are working in a company who sends gifts to potential customers, who than can respond to place an order.
  - If the customer does not respond to the gift the company pays 50 euro; otherwise it earns100 euro.
  - If you forget to send a gift to a responsive customer, you loose 100 euro
  - Given data of past customers in a given table, decide which group of people you should target in the future

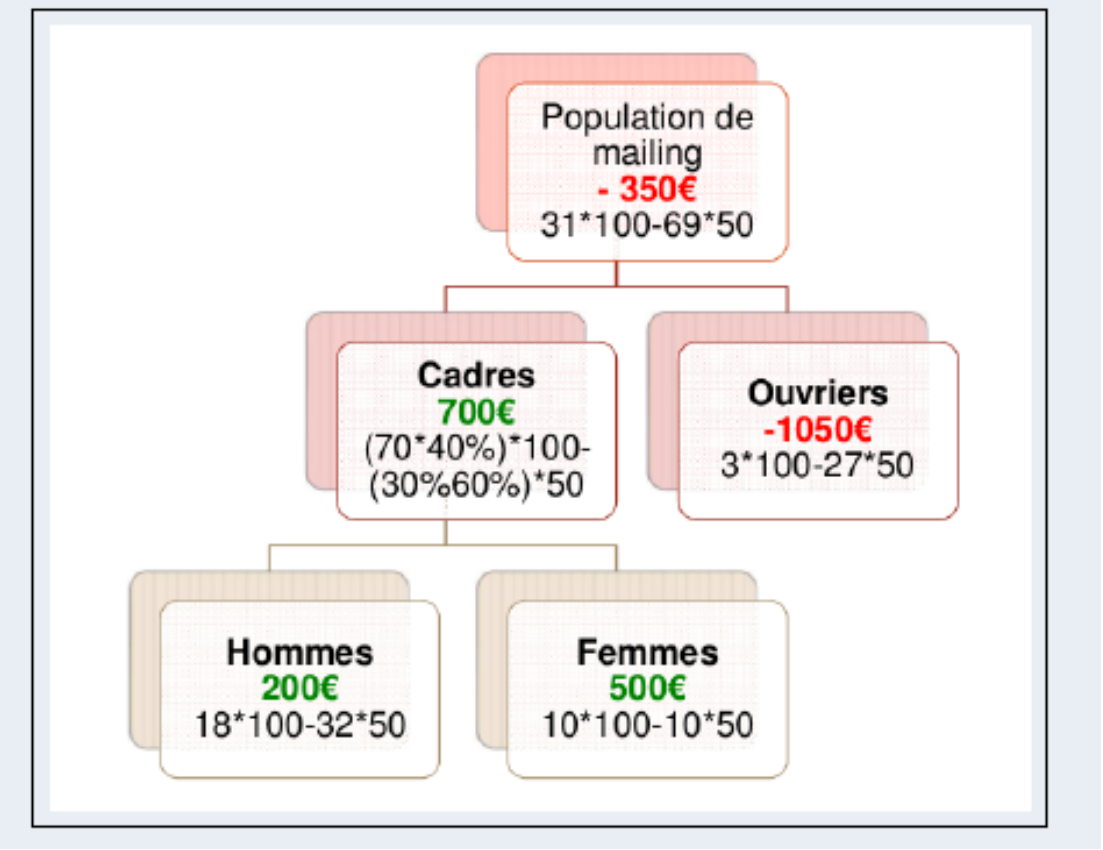
Nom	Prénom	Sexe	Profession	Réponse
Martin	Jeanne	F	Cadre	ok
Berluchette	Huguette	F	Ouvrière	ok
Sarkau	Sy	M	Ouvrier	non
Vil	Dominique	M	Cadre	non
Maitre	Kanter	M	Cadre	ok

- **Decision Tree** Example:



Approach 2: Classification

- **Decision Tree** Example:



Mail only to executives or only female executives

Approach 2: Classification

- **Bayesian classifier**:

$$\underbrace{P(c\,|\,x)}_{\text{Posterior Probability}} = \frac{\overbrace{P(x\,|\,c)}^{\text{Likelihood}}\,\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

$$P(c\,|\,X) = P(x_1\,|\,c)\times P(x_2\,|\,c)\times\cdots\times P(x_n\,|\,c)\times P(c)$$

  - *P*(*c*|*x*) is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
  - *P*(*c*) is the prior probability of *class*.
  - *P*(*x*|*c*) is the likelihood which is the probability of *predictor* given *class*.
  - *P*(*x*) is the prior probability of *predictor*.

Approach 2: Classification

Université de Strasbourg

- **Bayesian classifier** example: Male or female?

Training set:

Sex	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

Approach 2: Classification

- **Bayesian classifier** example: Male or female?

Test case:

Sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

- **Bayesian classifier** example: Male or female?

Sex	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033e-02	176.25	1.2292e+02	11.25	9.1667e-01
female	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667e+00

Gaussian naive bayes:

$$p(x = v \mid c) = \frac{1}{\sqrt{2\pi\sigma_c^2}}\, e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

$$P(\text{male}) = 0.5$$

$$p(\text{height} \mid \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789,$$

Approach 2: Classification

Université de Strasbourg

- **Bayesian classifier** example: Male or female?

$$P(\text{male}) = 0.5$$

$$p(\text{height} \mid \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789,$$

$$p(\text{weight} \mid \text{male}) = 5.9881 \cdot 10^{-6}$$

$$p(\text{foot size} \mid \text{male}) = 1.3112 \cdot 10^{-3}$$

$$\text{posterior numerator (male)} = \text{their product} = 6.1984 \cdot 10^{-9}$$

$$P(\text{female}) = 0.5$$

$$p(\text{height} \mid \text{female}) = 2.2346 \cdot 10^{-1}$$

$$p(\text{weight} \mid \text{female}) = 1.6789 \cdot 10^{-2}$$

$$p(\text{foot size} \mid \text{female}) = 2.8669 \cdot 10^{-1}$$

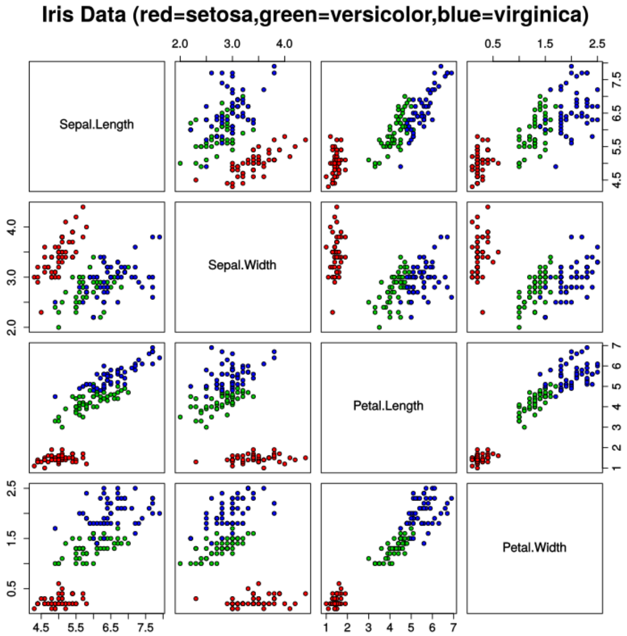
$$\text{posterior numerator (female)} = \text{their product} = 5.3778 \cdot 10^{-4}$$

Approach 2: Classification

- **Bayesian classifier** example:
  - Two classes:
    - $c1 = 01100, 11001, 10110, 10101, 10010$
    - $c2 = 01010, 11111, 11010, 11101, 10101$
  - Classify $x = 00111$

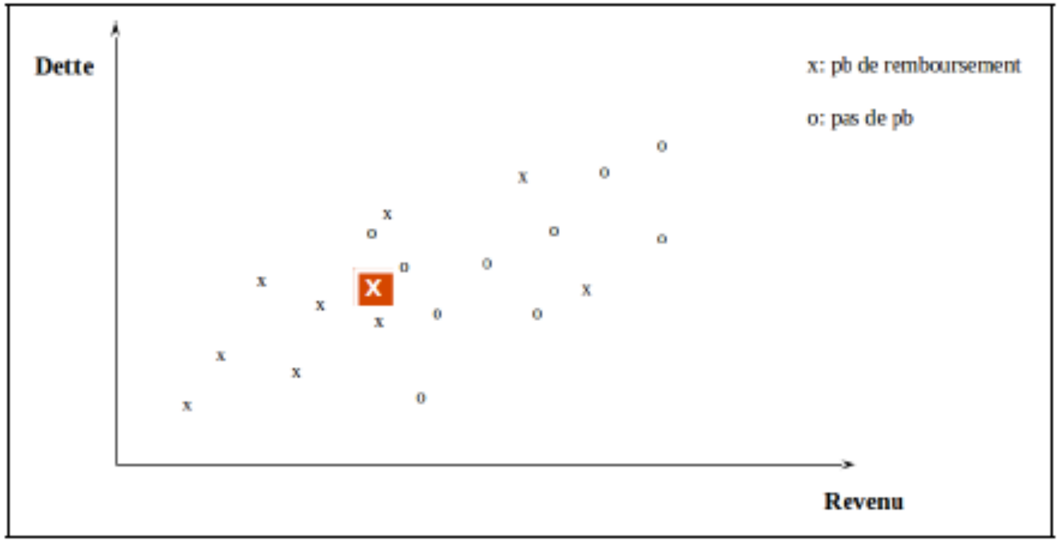
Approach 2: Classification

- **Bayesian classifier**:
  - Often used in text classification (i.e. spam)
  - Works with little data, updates are possible



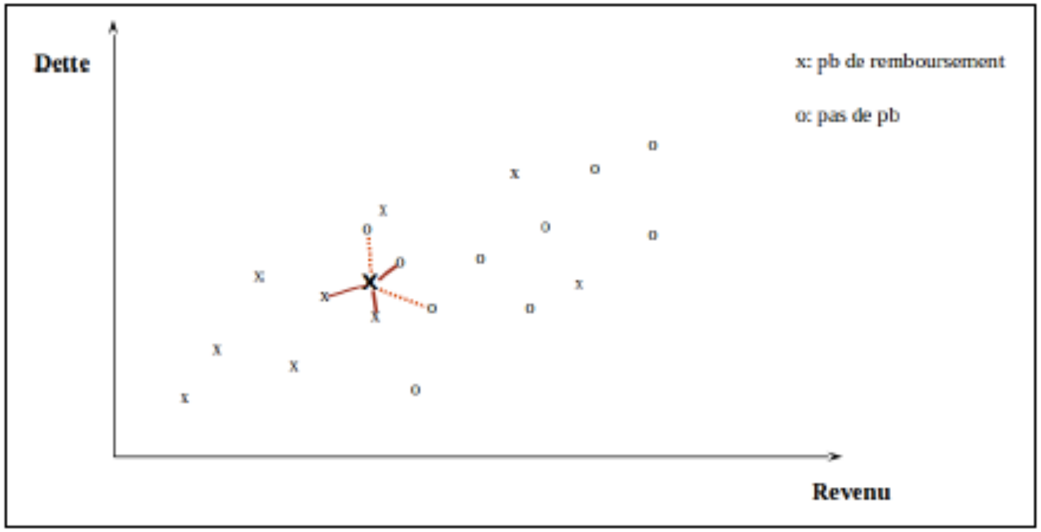
Use the four properties to predict the type
Of iris

- **K-nearest neighbor**:
  - All distances between the point X to be classified and all labeled points is computed
  - We keep the K closest labeled points. The Majority class in this set is attributed to X.



Approach 2: Classification

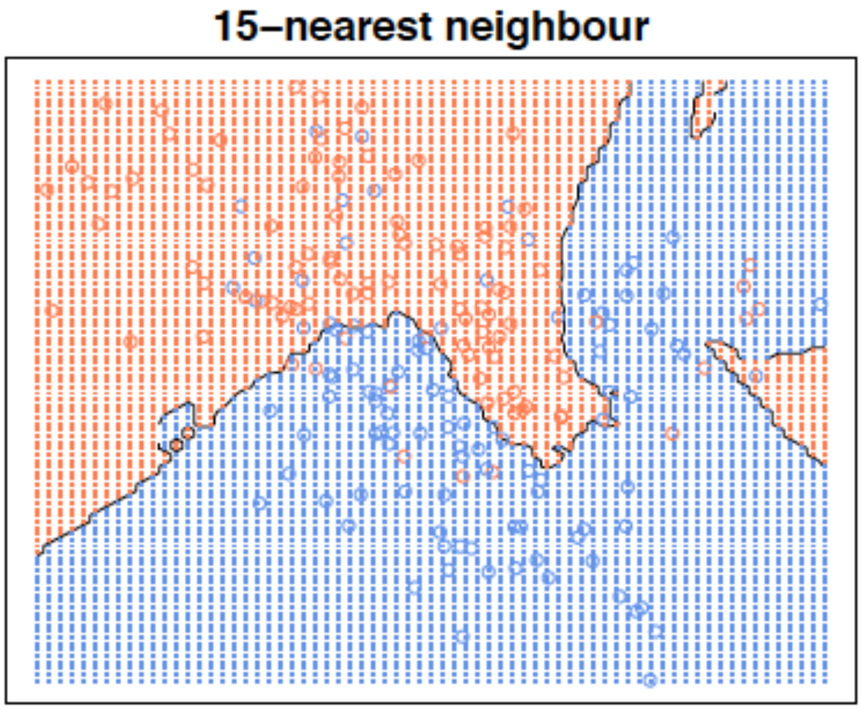
- **K-nearest neighbor**:
  - All distances between the point X to be classified and all labeled points is computed
  - We keep the K closest labeled points. The Majority class in this set is attributed to X.



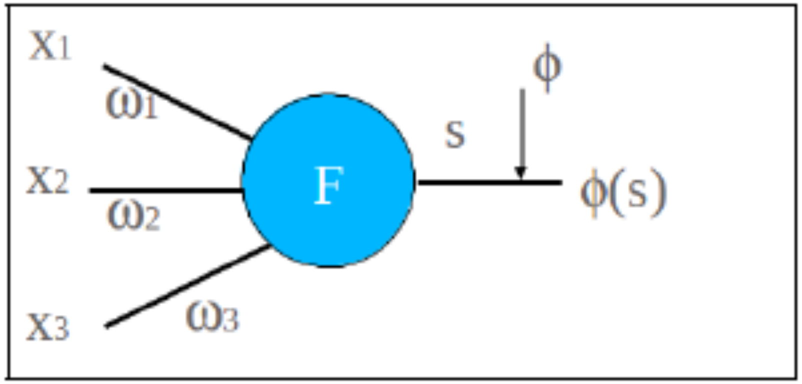
$$K = 3 \rightarrow x,\ K = 5 \rightarrow o$$

Approach 2: Classification

- **K-nearest neighbor**:
  - No hypothesis on the distribution of classes
  - Complexity increasing with the size of the training base



- **Neural networks**:
  - Inspired by the structure of the nervous system
  - A large number of connected neurons that process information
  - The response of the neuron depends on its state and the weights of the connections
  - The weights (or forces) are developed by experiment



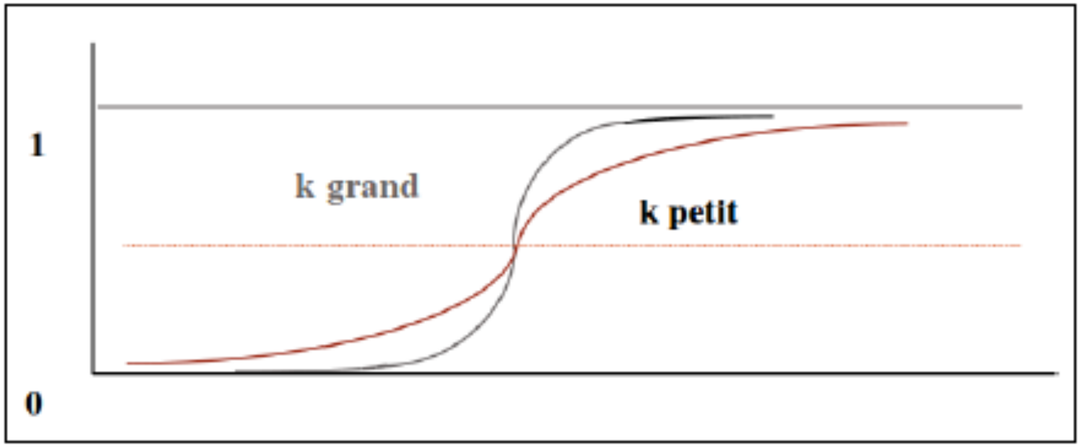
Approach 2: Classification

- **Neural networks** principle:
  - Construction of a network of simple computational units (neurons) linked by connections
  - Learning network parameters (weight of connections) using a set of examples
- Components of a neuron unite:
  - **inputs** (incoming connections or input variables)
  - **weights** on incoming connections
  - a **function F** which computes an output as a function of the inputs and the weight on the inlets
  - **activation function** which modifies the amplitude of the output of the node

Université de Strasbourg

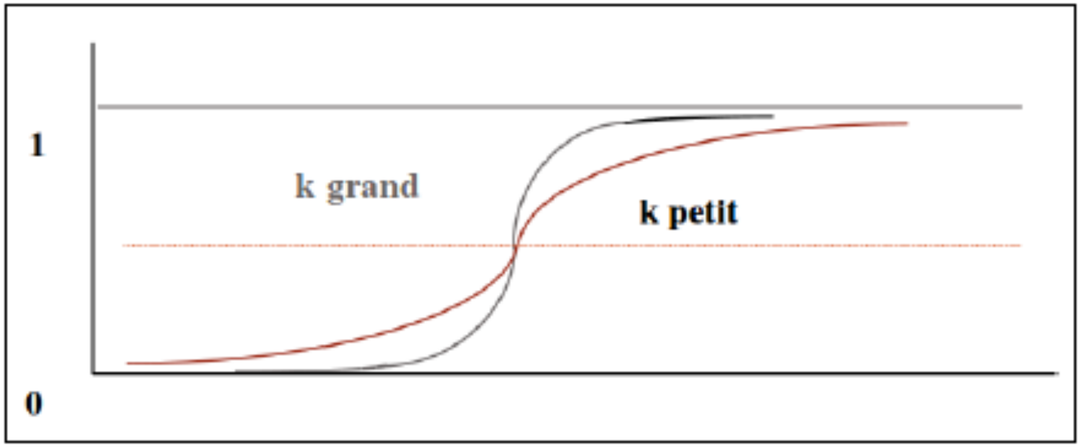
- **Activation function:**

  - $\Phi(s) = \text{linéaire}$
  - $\Phi(s) = \text{seuil}$
    - $\Phi(s) = 0 \quad \text{si} \quad s \leq a$
    - $\Phi(s) = 1 \quad \text{si} \quad s \geq a$
  - $\Phi(s) = 1/(1 + e^{-ks})$



Approach 2: Classification

- **Activation function:** $\Phi(s) = 1/(1 + e^{-ks})$
  - If the coefficient k is large, then the output is almost always close to 0 or 1: relatively symbolic neural network
  - If the coefficient k of $1/(1 + e^{-ks})$ is small, then the strength of each cell is well distributed between 0 and 1
  - Another implicit parameter is the center of the sigmoid function



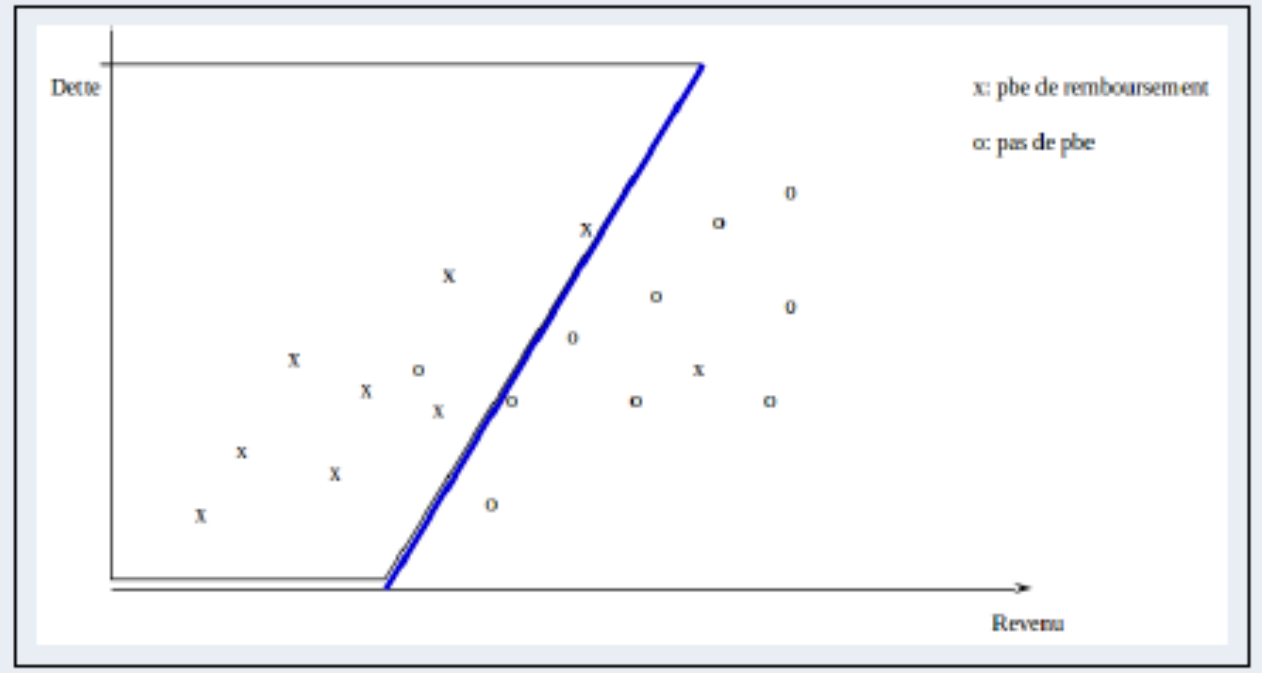
Approach 2: Classification

Université de Strasbourg

- **Simplest neural network:**
  - Single neuron
  - F = weighted sum of inputs
  - Activation function = thresholding
  - $s = 1$ if $w_1x_1 + w_2x_2 + \ldots > a$
  - $s = 0$ if $w_1x_1 + w_2x_2 + \ldots <= a$
  - Equation of a hyperplane!

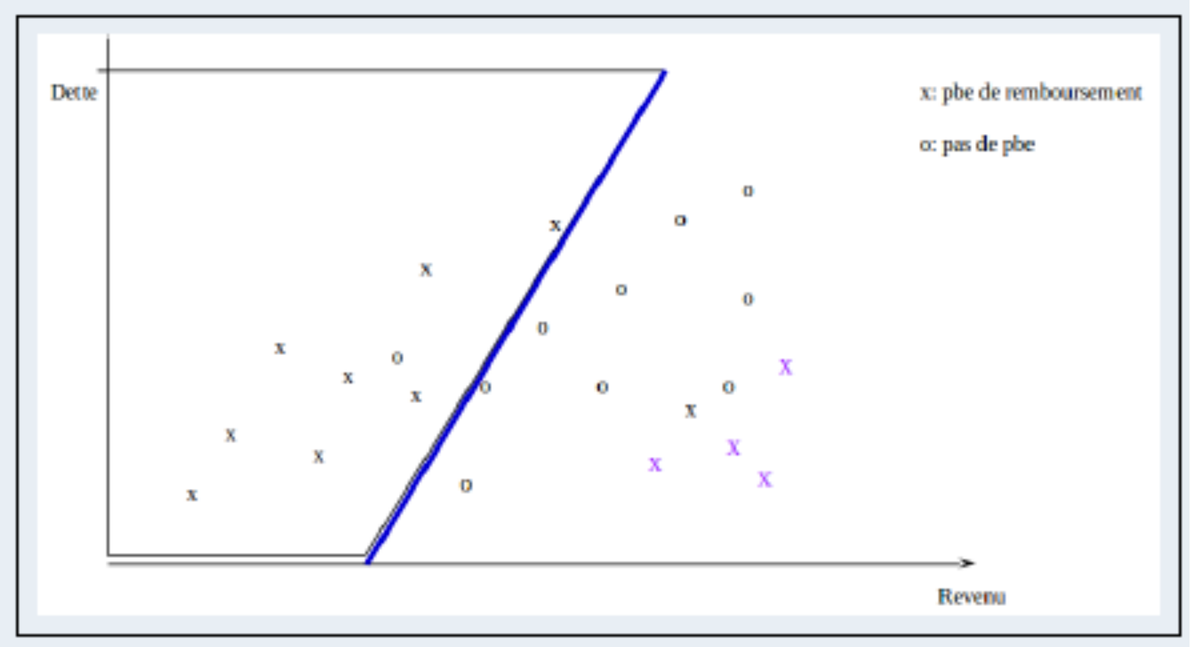
Approach 2: Classification

- **Simplest neural network:**
  - Linear separation



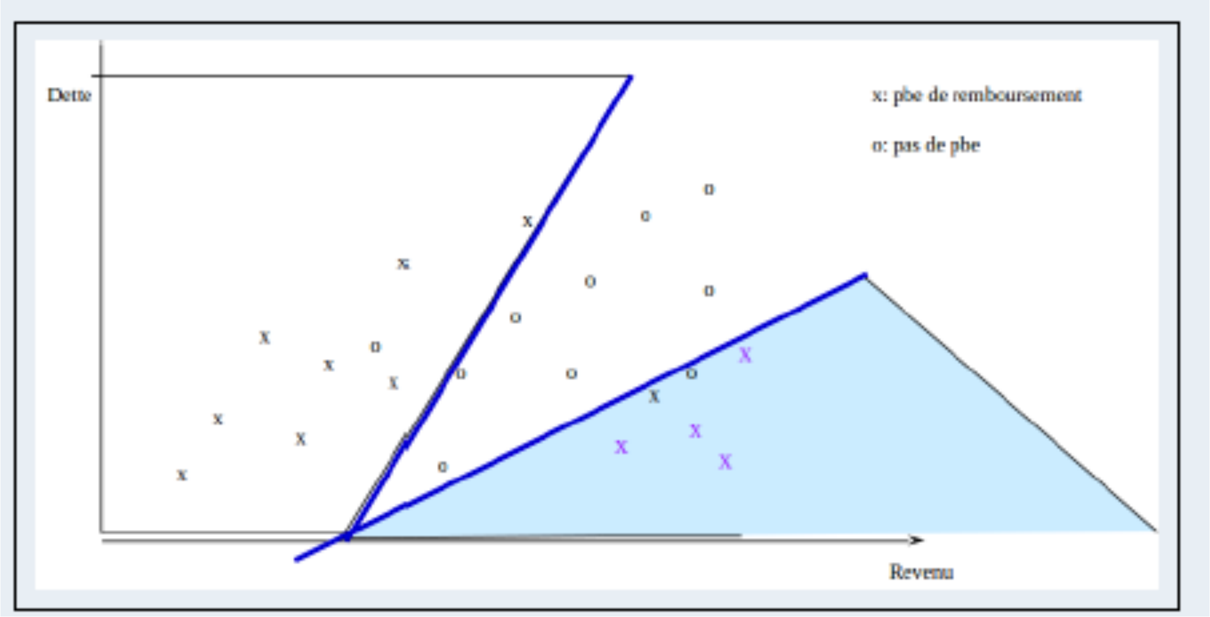
Approach 2: Classification

- **Simplest neural network:**
  - Additional examples
  - How to find a neuronal network discriminating between the two classes?



Approach 2: Classification

- **Simplest neural network:**
  - Additional examples



Approach 2: Classification

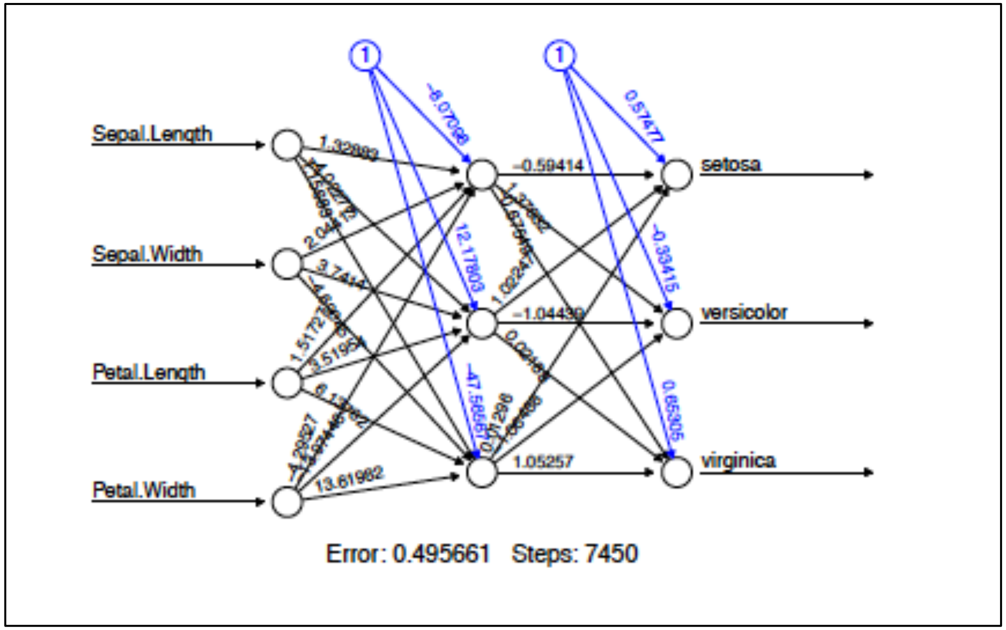
- NN In practice:
  - Choose a **calculation function** and an **activation function**
  - Choose an **architecture**:
    - Number of inputs
    - Number of outputs
    - Number of internal layers
    - Number of neurons of each of the internal layers
  - Select a **cost function**
  - Decide **when to stop** training

Approach 2: Classification

- **Advantages** of neural networks:
  - Robust to noise
  - Classification or estimation is quick after the training is completed
  - Available in all data mining software
- **Disadvantages**:
  - Black box: difficult to interpret the obtained model
  - Significant learning time
  - Selection of parameters is difficult

Approach 2: Classification

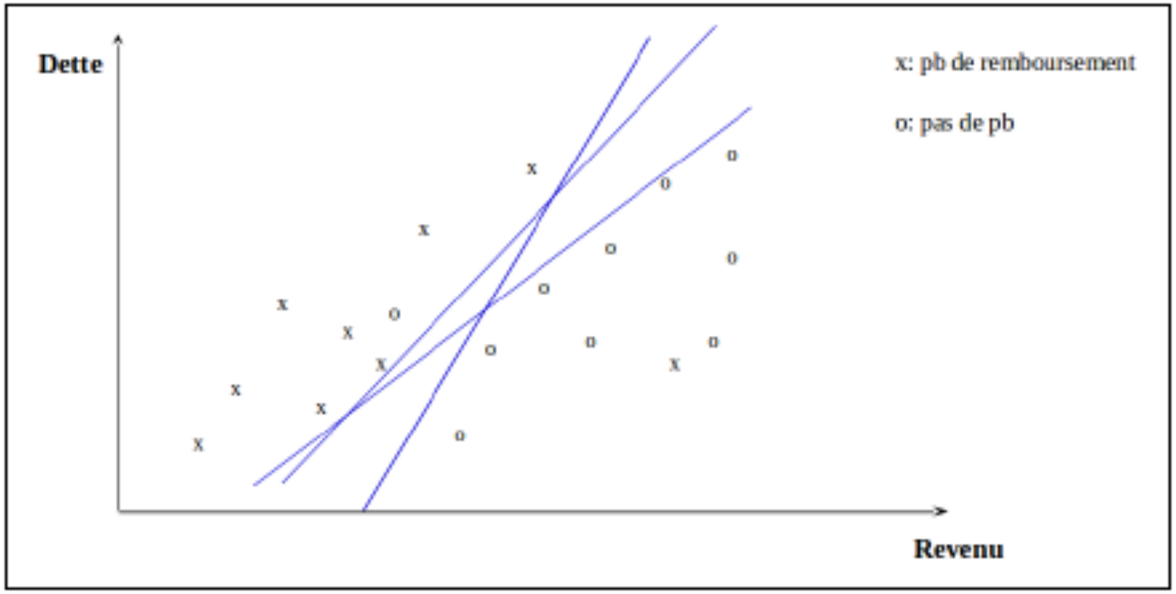
- Neural networks:
  - Renewed popularity with the arrival of Deep Learning
  - Frequently used in image analysis (Convolution Neural Network)



12.09.17

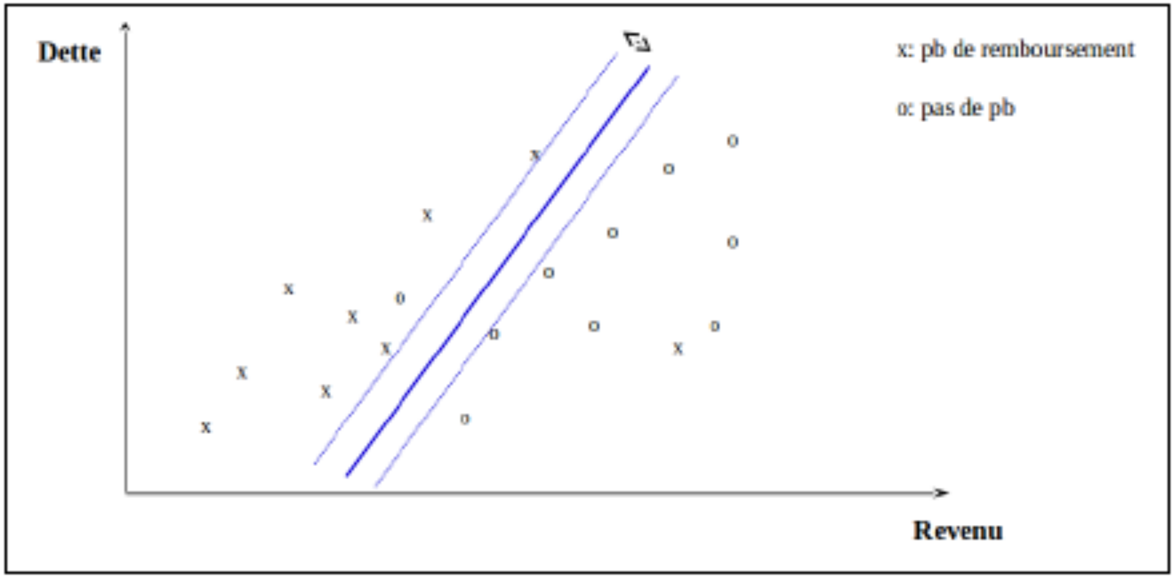
Approach 2: Classification

- **Support Vector Machine (SVM)**:
  - Divide the data into two classes using a hyperplane
  - Maximize the gap between this hyperplane and the data
  - Find the line that maximizes the gap among several possible

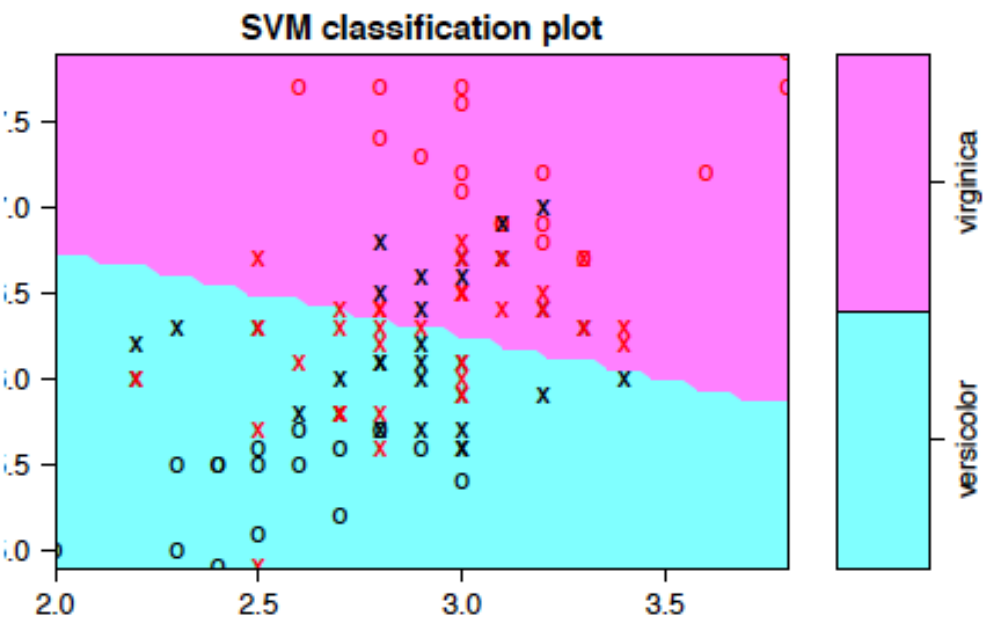


Approach 2: Classification

- **Support Vector Machine (SVM):**
  - Divide the data into two classes using a hyperplane
  - Maximize the gap between this hyperplane and the data
  - Find the line that maximizes the gap among several possible



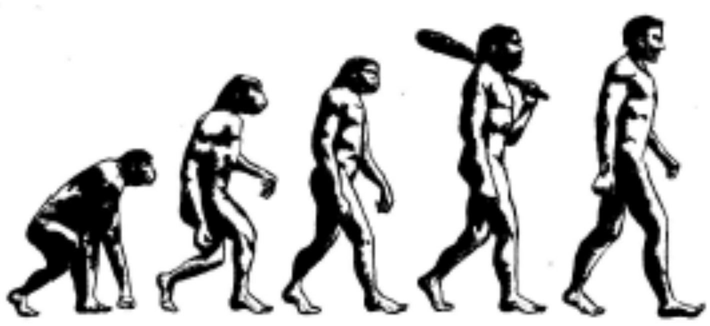
- **Support Vector Machine (SVM):**
  - Need to find a kernel suited to transform the data
  - Example of the Gaussian kernel: $K(x,y) = exp(-\frac{\|x-y\|}{2\sigma^2})$



Approach 2: Classification

Genetic algorithms:

- Inspired by the theories of the evolution of Darwin, Lamarck or Baldwin
- General optimization method
- Can be used in classification or estimation



Approach 2: Classification

- **Genetic algorithms:**
  - General scheme:
    - We define the “parameters” to be optimized: range of values, thresholds, etc. The corresponding genotype (chromosomes) is defined.
    - We define the function of computation of the phenotype and the function of an individual
    - The mechanisms and rates of crossing and mutation are defined
    - The function of electing survivors is defined

Approach 2: Classification

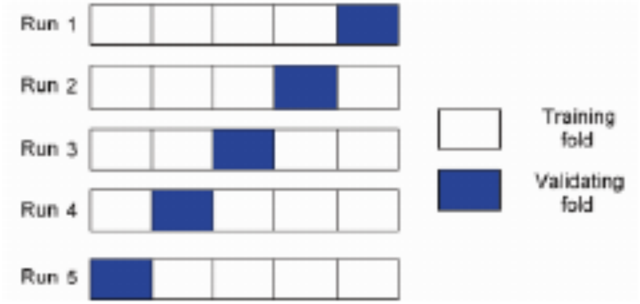
- **Genetic algorithms:**
  - General scheme:
    1. Initialize the population
    2. Calculate the degree of adaptation f(x) of each individual
    3. As long as not finished or no convergence:
       a) Reproduction of parents
          - select 2 individuals at a time
          - apply genetic operators

       b) calculate the degree of adaptation f(x) of each child

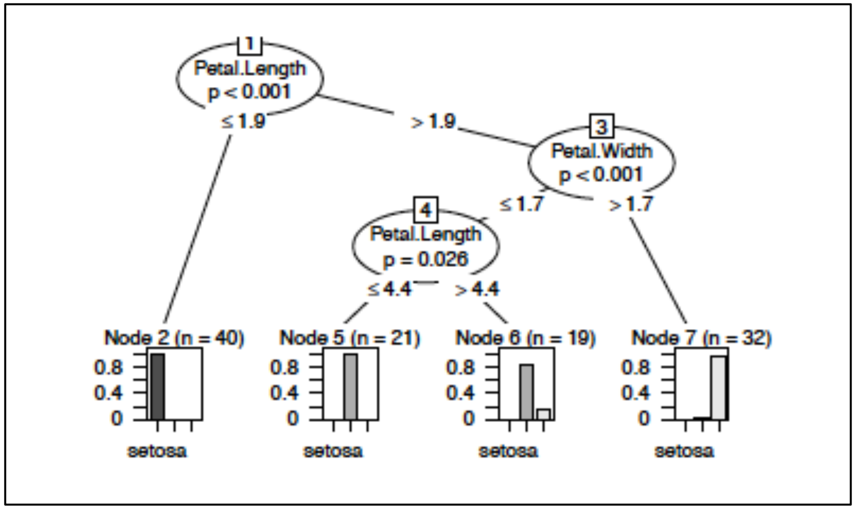
       c) select survivors from parents and children

Approach 2: Classification

- **Validation:**
  - Validation by test:
    - Data = learning set + test set
    - Construction of a model on the learning set and test model on the test set for which the results are known
    - Cross-validation



- **Validation:**
  - Validation by test:
    - **Split test / train data**, in general more data for learning
    - The number of cross-validation folds depends on the volume of the data

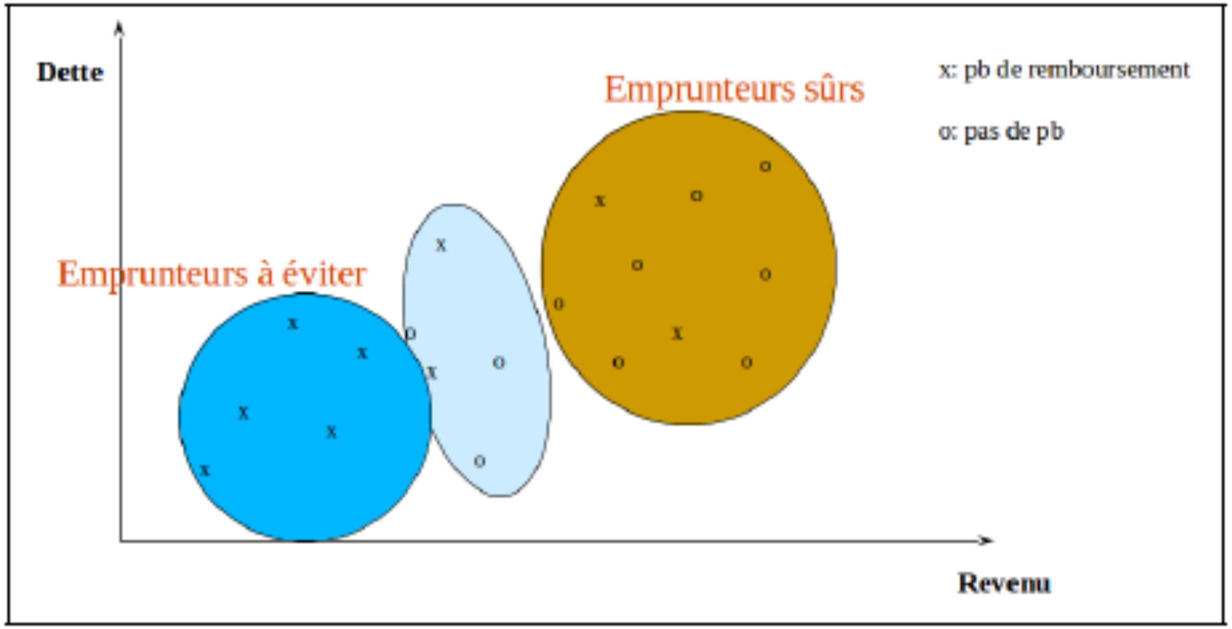


Approach 3: Clustering

- **Clustering:**
  - Aim of the clustering: obtain a simplified representation (structuring) of the initial data
  - Organization of a set of objects into a set of homogeneous and / or natural groupings

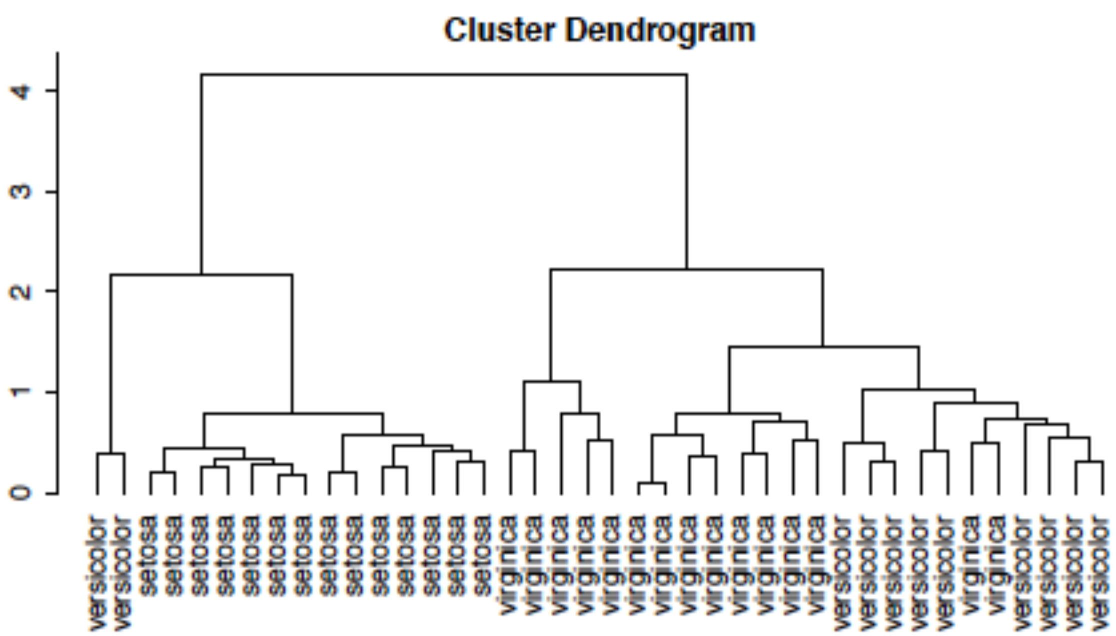
Approach 3: Clustering

- **Clustering:**
  - Automatic partitioning from data
  - No a priori semantic interpretation



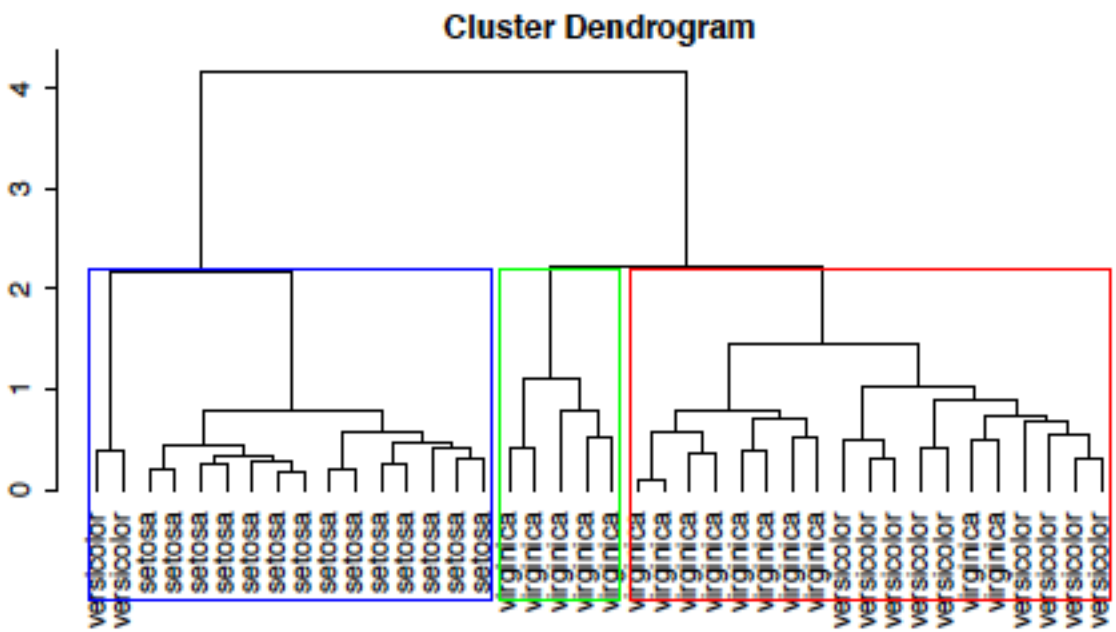
Approach 3: Clustering

- **Clustering:**
  - Hierarchical clustering



Approach 3: Clustering

- **Clustering:**
  - Hierarchical clustering



Approach 4:
Dependency modelling

Associations:

- Association rules: analysis of the shopping basket
- “On fridays, customers often buy beer packs and at the same time diapers”
- Are there any causal links between the purchase of a product P0 and a other product P1?

Approach 4: Dependency modelling

- **Associations:**
  - Rules of association: premise -> conclusion
  - Questions:
    - beurre -> pain ?
    - poisson, viande -> lait ?
    - fromage, pates -> vin ?

Identifiant	Transaction
1	beurre fruits lait pain
2	fruits lait pain
3	beurre fromage pain pâtes viande vin
4	fromage fruits lait légumes pain pâtes poisson
5	beurre fruits lait légumes pain pâtes poisson viande
6	beurre fromage légumes pain pâtes viande vin
7	beurre fromage lait légumes pain pâtes viande vin
8	fruits légumes poisson
9	beurre fromage lait pain pâtes viande vin
10	beurre fromage fruits lait légumes pain poisson viande

Approach 4: Dependency modelling

- **Associations:**
  - Formally:
    - Given a set of transactions D, find all the association rules X-> Y having support and confidence above the minimum thresholds predicted by the user
    - A transaction is a set of attributes (butter, fruit, milk, bread)
    - **Support**: % of transactions in D that contain X and Y
    - **Confidence**: % of transactions that contain X which also contain Y

Approach 4: Dependency modelling

- **Associations:**
  - Interpretation:
    - R : X -> Y (A%, B%)
    - A% of all transactions show that X and Y have been purchased at the same time (support of the rule) and B% of clients who purchased X have also purchased Y (confidence in the rule).

Approach 4: Dependency modelling

- **Associations:**
  - Two sub problems:
    - FIS: Find all frequent ensembles (**item sets**) that have support greater than or equal to a minimum value "**minusup**"
    - Generate all the association rules having confidence greater or equal to "**minconf**"

$$\text{support}_{A \Longrightarrow B} = \frac{|\{t\colon A \cup B \subseteq t\}|}{|T|} \qquad \text{confidence}_{A \Longrightarrow B} = \frac{|\{t\colon A \cup B \subseteq t\}|}{|\{t\colon A \subseteq t\}|}$$

Approach 4: Dependency modelling

Université de Strasbourg

- **Associations:**

Tickets	Produits achetés
1	beurre fruits lait pain
2	fruits lait pain
3	beurre fromage pain pâtes viande vin
4	fromage fruits lait légumes pain pâtes poisson
5	beurre fruits lait légumes pain pâtes poisson viande
6	beurre fromage légumes pain pâtes viande vin
7	beurre fromage lait légumes pain pâtes viande vin
8	fruits légumes poisson
9	beurre fromage lait pain pâtes viande vin
10	beurre fromage fruits lait légumes pain poisson viande

$$\text{Support} = \frac{\square \text{ et } \square}{\text{Tous tickets}}$$

$$\text{Confiance} = \frac{\square \text{ et } \square}{\square + \bigcirc}$$

	Support	Confiance
beurre $\rightarrow$ pain	70%	100%
poisson viande $\rightarrow$ lait	20%	100%
fromage pâtes $\rightarrow$ vin	40%	80%

Approach 4: Dependency modelling

Associations:

Tickets	Produits achetés
1	beurre fruits lait pain
2	fruits lait pain
3	beurre fromage pain pâtes viande vin
4	fromage fruits lait légumes pain pâtes poisson ~~vin~~
5	beurre fruits lait légumes pain pâtes poisson viande
6	beurre fromage légumes pain pâtes viande vin
7	beurre fromage lait légumes pain pâtes viande vin
8	fruits légumes poisson
9	beurre fromage lait pain pâtes viande vin
10	beurre fromage fruits lait légumes pain poisson viande

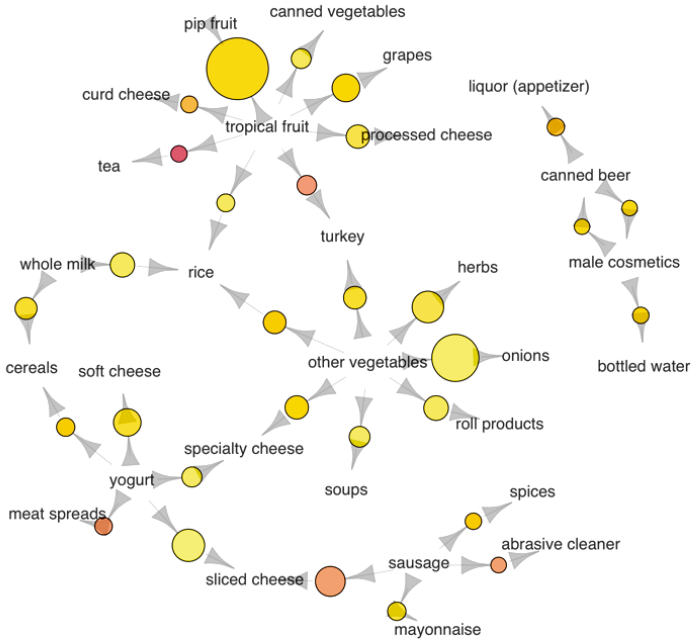
$$\text{Support} = \frac{\square \text{ et } \square}{\text{Tous tickets}}$$

$$\text{Confiance} = \frac{\square \text{ et } \square}{\square + \bigcirc}$$

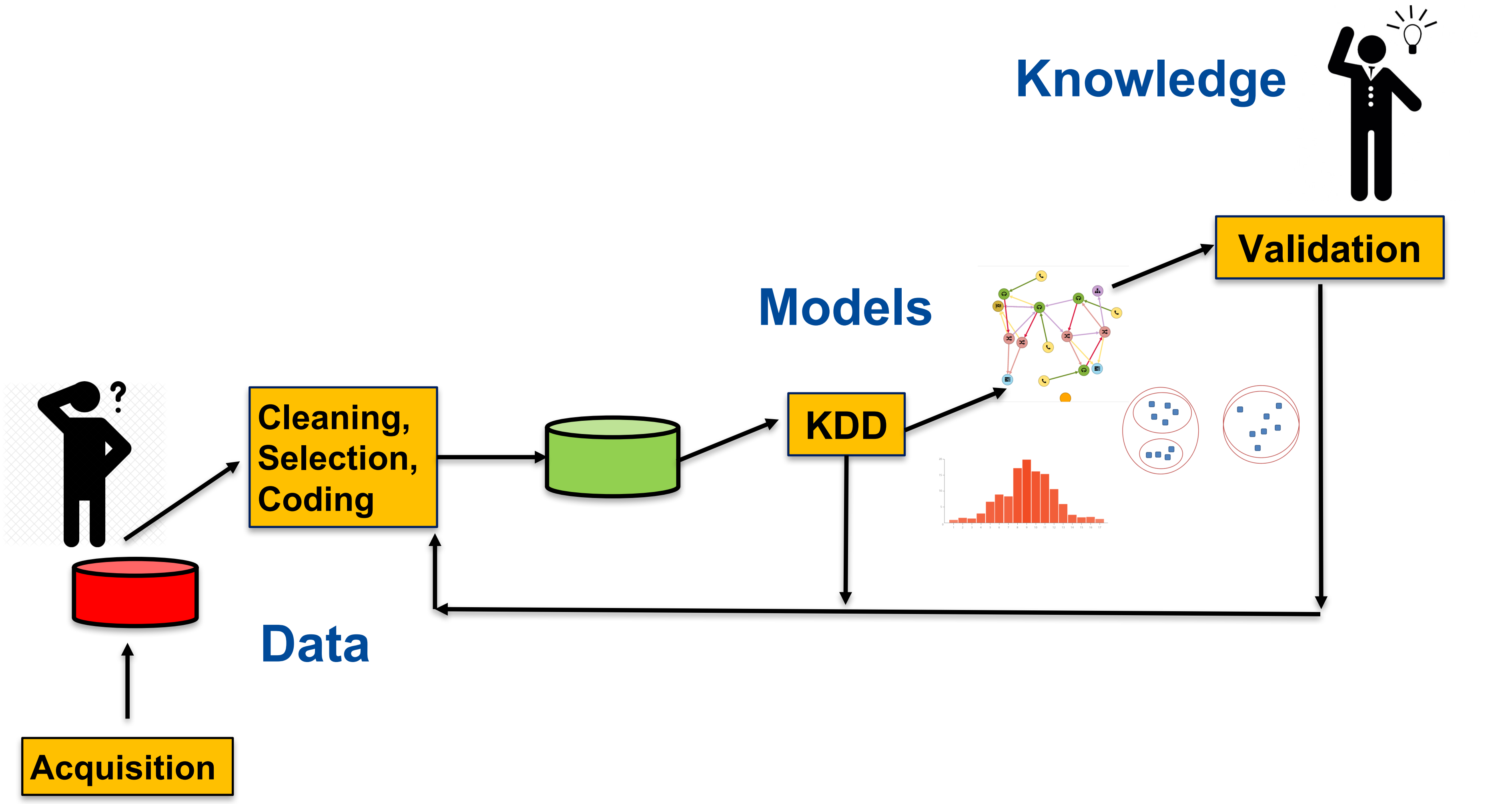
			Support	Confiance
beurre	→	pain	70%	100%
poisson viande	→	lait	20%	100%
fromage pâtes	→	vin	40%	80%

Université de Strasbourg

- **Associations:**
  - Association rules:
    - Numerous criteria for evaluating the interest of a rule
    - Difficult to scale to large volumes of data



Data mining process



Validation

- Generation of a large number of models
- Is the generated model interesting?
- How to measure the interest of a model:
  - Novelty
  - Easy to understand
  - Can be validated by new data (with a certain confidence)
  - Usefulness
  - Confirms (or rejects) the hypothesis of an expert

Validation

- Evaluation of a model
  - subjective: expert
  - objective: statistics and structure of the model
- Can we find all models? (completeness)
- Can we only generate the interesting models? (optimization):
  - Generating all the models and iterating according to certain measures and characteristics: nonrealistic
  - Generate only the models that satisfy a particular condition

Group Task 2

- Form groups of 4-5 people
- You will get a statement about data mining.

- What do you think? Is this statement correct or wrong? Why?

- You have 5 min for the task.

Conclusions

- **"Data mining methods are more inductive then methods based on hypothesis because there is no a priori knowledge on the data"**

- **False**: condition of application of the methods, choice of data, coding data, choice of explanatory variables, order of input of variables in the algorithm,..

Conclusions

- **"All data available must be used to generate the model"**

- **False**: coding of the data, order of input of the variables in the algorithm, irregular effects, outliers, influence of redundancies, correlations, the computer data model, saturation, instability ...

Conclusions

- **"With all these methods we will always find out something great!"**

- **False**: common sense solutions need to be found (specialists, experts). In fact, we need to find the best solution (among n) for difficult data

Conclusions

- **“Data mining is revolutionary!”**

- **False**: traditional data analysis methods + more specific methods (neural networks). Optimization of existing techniques because of thelarge amount of data.

Conclusions

- **Question**:
  - "Why so many algorithms?"
- **Answer**:
  - Because none is optimal in all cases

- As they are in practice complementary to one another, combining them intelligently (by constructing models) it is possible to achieve very significant performance gains