

# Pattern Mining

## Part I



**Fabrice Guillet**

# General outline

- Introduction: Generalities on data mining and pattern mining
- Mining transactional patterns: itemsets and association rules
- Quality and Interestingness measures

# Pattern Mining

## Part I

### Introduction: Generalities on data mining and pattern mining

Ecole Polytechnique de l'université de Nantes

F. Guillet

# Outline

## Introduction: Generalities on data mining and pattern mining

- Generalities on Business Intelligence
- Generalities on data mining
- KDD Process and Method
- KDD models

# Introduction

Generalities on « Business Intelligence »

# Introduction to “Business Intelligence”

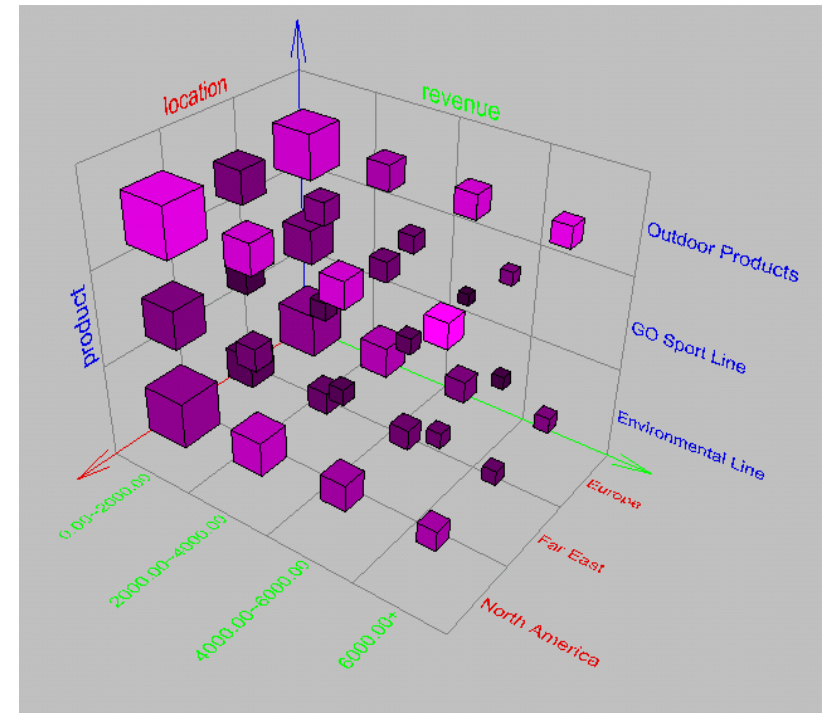
- **« Classical » Computer science :**  
**Data** oriented (IS + applications)
  - Databases DB, relational tables
  - Applications
- **Business intelligence Evolution toward :**  
**Computer science for “Decision-making”**  
**(informatique décisionnelle)**  
a **Decision Maker**
  - Data Wharehouse (BI)
  - **Knowledge Discovery in Data (KDD, Data Mining)**
  - Knowledge Management (KM) and Engineering (KE)

# Data Warehouse

- Computer science for “Decision” : Data Warehouse

Data -> Decision maker

- Data selection, history, analysis for decision maker
- DB -> Multidimensional data (time, location, product, income)
- Data requests -> OLAP (on-line analytical processing)
- Reporting



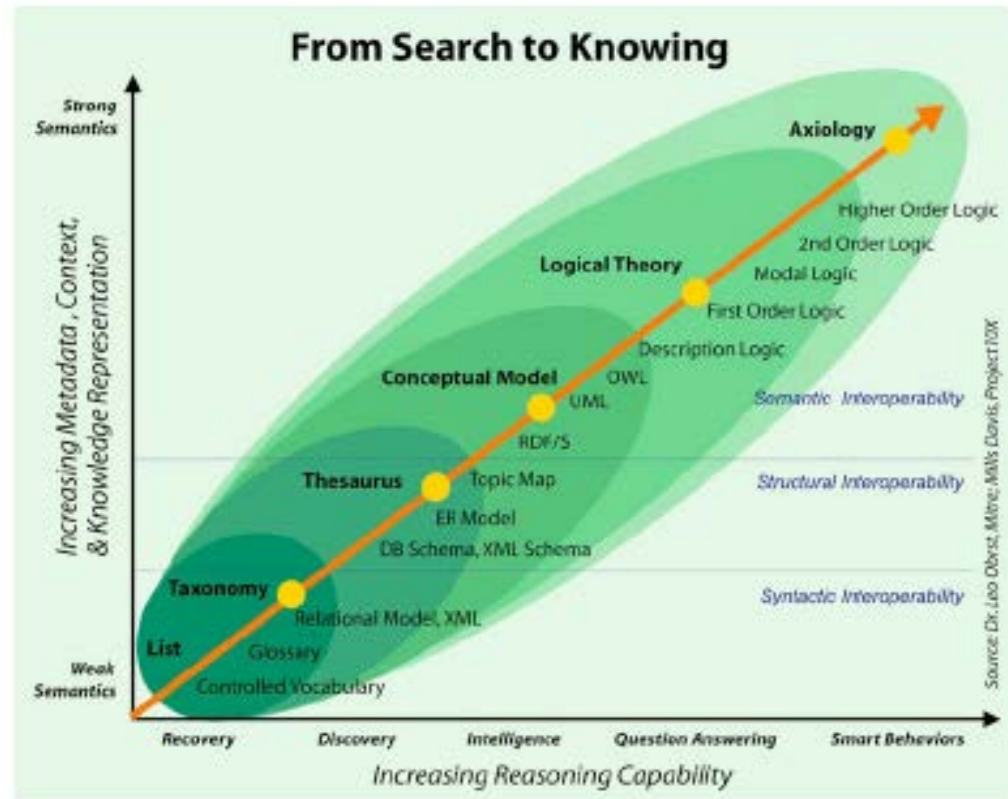
# Knowledge Management and Engineering

- Computer science for “Decision” : **Knowledge Management and Engineering**

Decision maker -> “Knowledge”  
Data

- Man -> computer
- interviews, formalization, storing, capitalization
- DB -> KB, semantic networks
- Linked with Semantic web technology (Ontology, Knowledge Engineering)

Figure-10 Semantics for information, knowledge, and reasoning





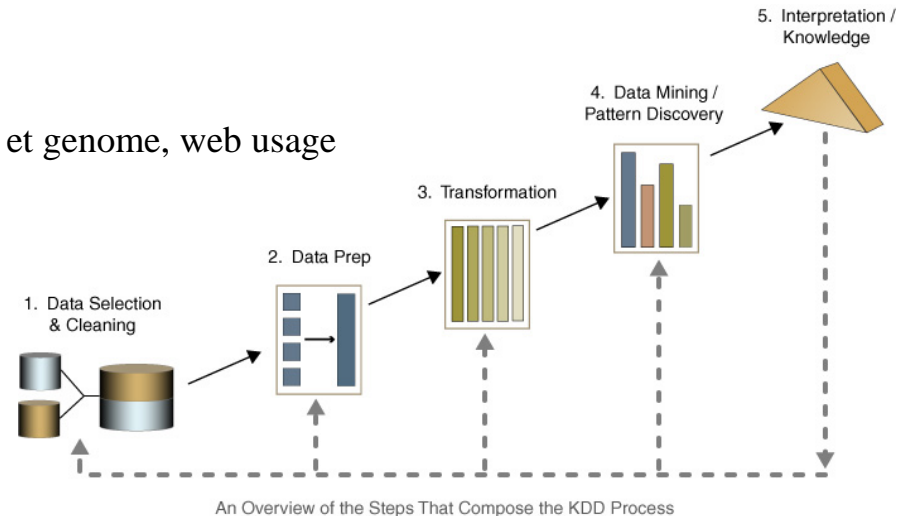
# Data Mining

- **Computer science for “Decision” : Data Mining**

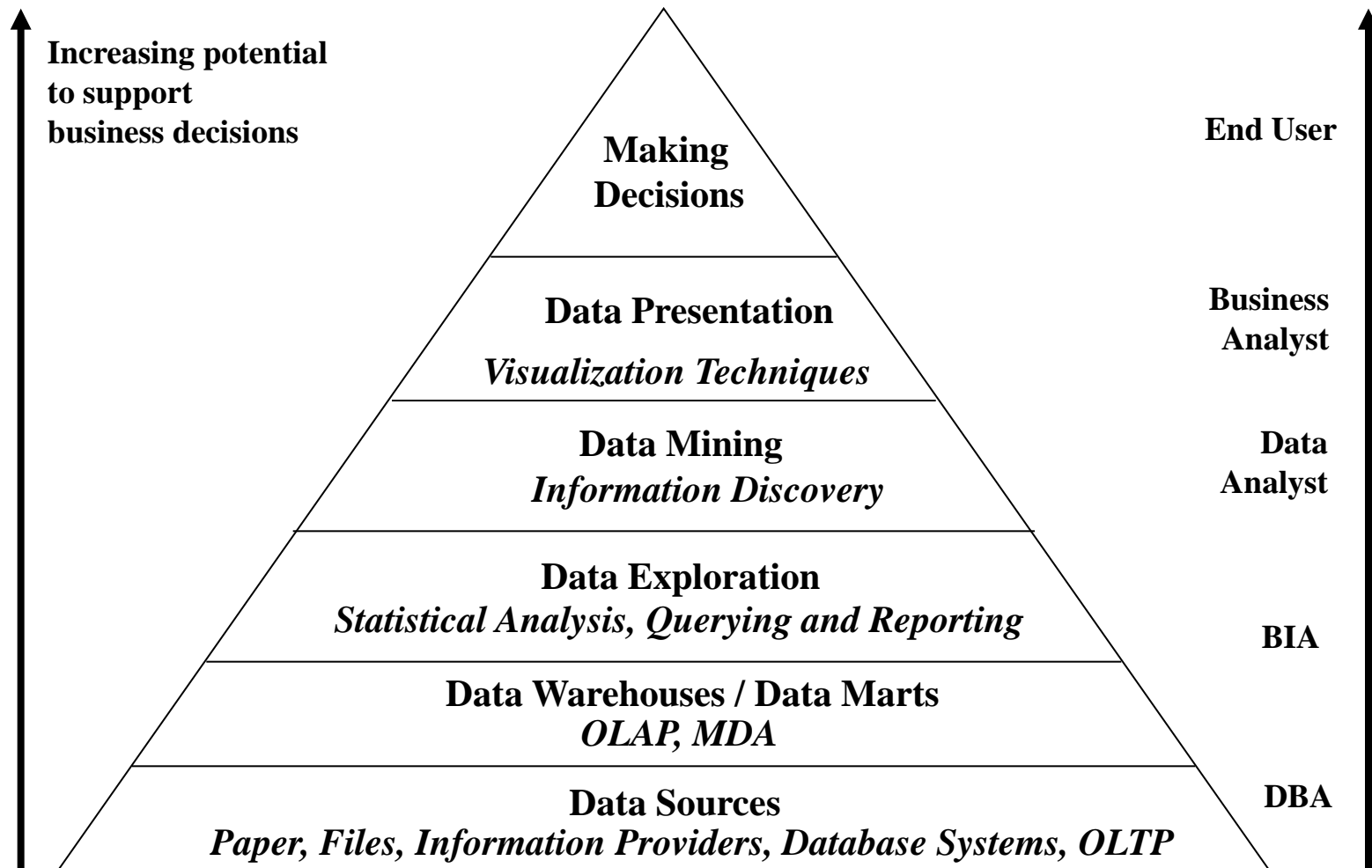
Knowledge Discovery in Data (KDD)

Data -> Prediction Model -> “Knowledge” -> Decision maker

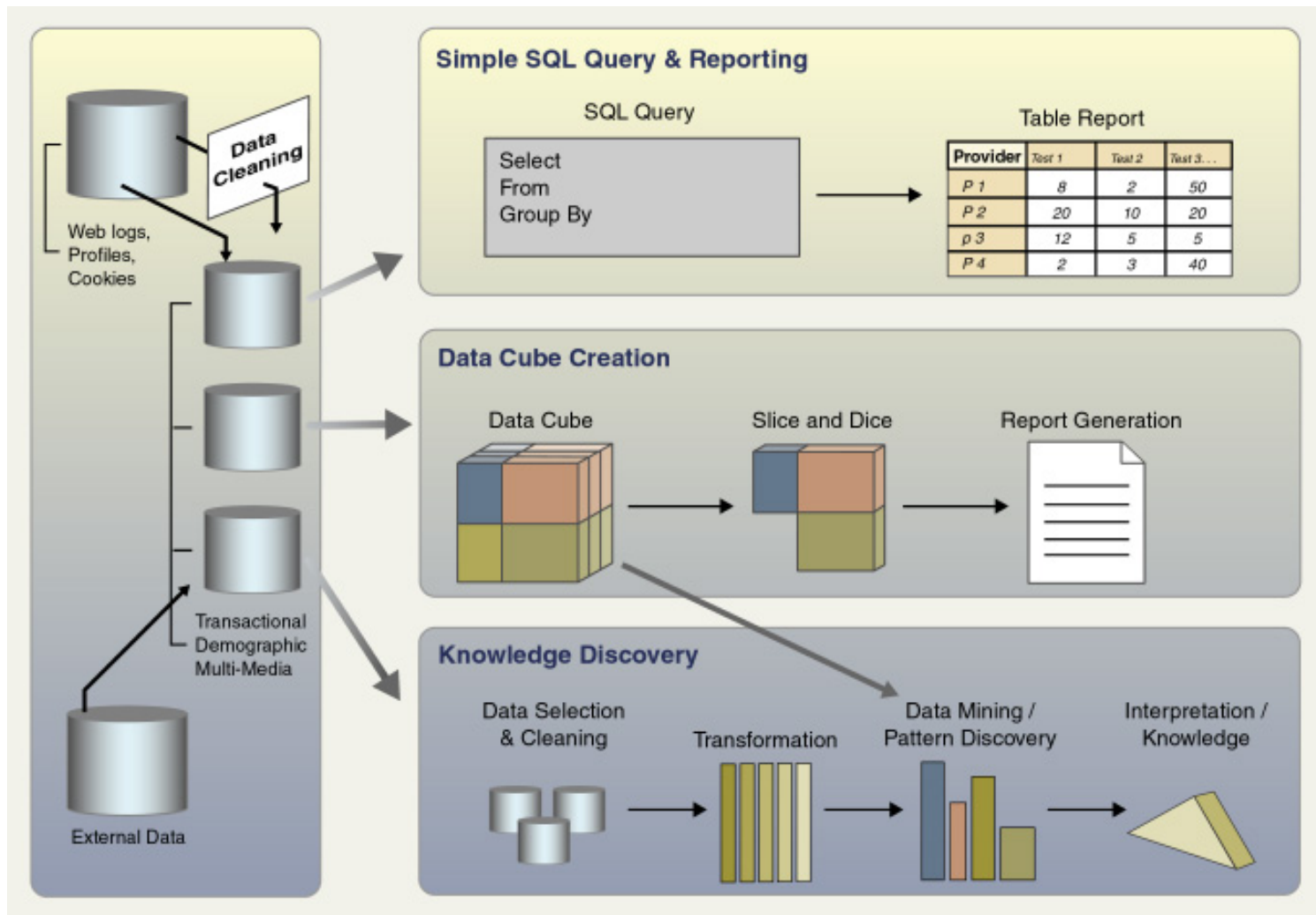
- Structured data :
  - Transactional (DB) : Data Mining
  - Sequences : Sequence Mining (bio-info et genome, web usage (logs), ...)
  - Spatiotemporal : ST Mining (GIS)
- Semi-structured data (XML based) :
  - web : Web Mining
  - Multimedia : Multimedia Mining
  - Graphs : Graph Mining
  - Social networks: Social mining
  - Knowledge : Knowledge Mining (sem Nets, Conceptual graphs, OWL ontology, ...)
- Weakly structured data :
  - Texts : Text Mining



# Evolution



# More : from SQL, DataWarehouse, to KDD

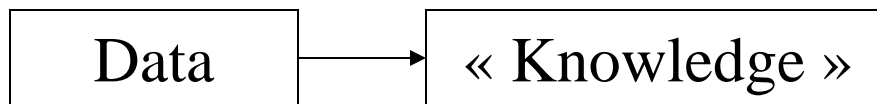


# Introduction

## Generalities on Data Mining

# Data-mining : recall of definitions

- « Extraction of relevant information, previously unknown, and potentially usefull from data" (Frawley et Piatetski-Shapiro)
- « Discovery of new correlations, tendencies, and models by mining a large amount of data" (John Page)
- « Decision-aid process where users seek for model of interpretation for their data" (Kamran Parsaye)
- "Torturer l'information possible jusqu'à ce qu'elle avoue" (Dimitris Chorafas)



Data-mining :

The stage of knowledge extraction in the global process of  
"Knowledge Discovery in database (KDD)"

# Application Areas

## **Industry**

Finance

Insurance

Telecommunication

Transport

Consumer goods

Data Service providers

Utilities

## **Application**

Credit Card Analysis

Claims, Fraud Analysis

Call record analysis

Logistics management

promotion analysis

Value added data

Power usage analysis

# Applications

- Banking: loan/credit card approval
  - predict good customers based on old customers
- Customer relationship management:
  - identify those who are likely to leave for a competitor.
- Targeted marketing:
  - identify likely responders to promotions
- Fraud detection: telecommunications, financial transactions
  - from an online stream of event identify fraudulent events
- Manufacturing and production:
  - automatically adjust knobs when process parameter changes
- Medicine: disease outcome, effectiveness of treatments
  - analyze patient disease history: find relationship between diseases
- Molecular/Pharmaceutical:
  - identify new drugs
- Scientific data analysis:
  - identify new galaxies by searching for sub clusters
- Web site/store design and promotion:
  - find affinity of visitor to pages and modify layout

# Inputted Data / outputted Knowledge

- **Data** : accurate information referring to real world
  - transaction, example, item, case, instance, object, ...
  - Attributes, variables + domain of values (symbolic/numeric)
  - DB transaction (data-mining), text (text-mining), internet (web-mining), knowledge (K-mining).
- **Knowledge** : global information, generalized, synthetical
  - class, abstract cluster of data (good, bad) :
    - Extension : list of transactions (rows)
    - Intension : common attributes (columns), boolean expression
    - Statistics : proba, frequency, value distribution...
  - links between classes (general/specific, distance, similarity, implication, deviations, ...)
  - Interestingness measure (quality) : interest, relevance, confidence, surprise...
  - => Pattern



# Inputted Data / outputted Knowledge

**Data** = table

Row = transactions

Column = attribute value

**Class/cluster** = sub table

Extension : list of rows

Intension : common attributes

(reduced description = abstraction = learning)

ID	Outlook	Temperature	Humidity	Windy	Class
1	sunny	hot	high	false	N
2	sunny	hot	high	true	N
3	overcast	hot	high	false	P
4	rain	mild	high	false	P
5	rain	cool	normal	false	P
6	rain	cool	normal	true	N
7	overcast	cool	normal	true	P
8	sunny	mild	high	false	N
9	sunny	cool	normal	false	P
10	rain	mild	normal	false	P
11	sunny	mild	normal	true	P
12	overcast	mild	high	true	P
13	overcast	hot	normal	false	P
14	rain	mild	high	true	N

ID	Outlook	Temperature	Humidity	Windy	Class
1	sunny	hot	high	false	N
2	sunny	hot	high	true	N
8	sunny	mild	high	false	N
9	sunny	cool	normal	false	P
11	sunny	mild	normal	true	P

Extension : { 1,2,8,9,11 }

Intension : Outlook=sunny

ID	Outlook	Temperature	Humidity	Windy	Class
1	sunny	hot	high	false	N
2	sunny	hot	high	true	N

Extension : { 1,2 }

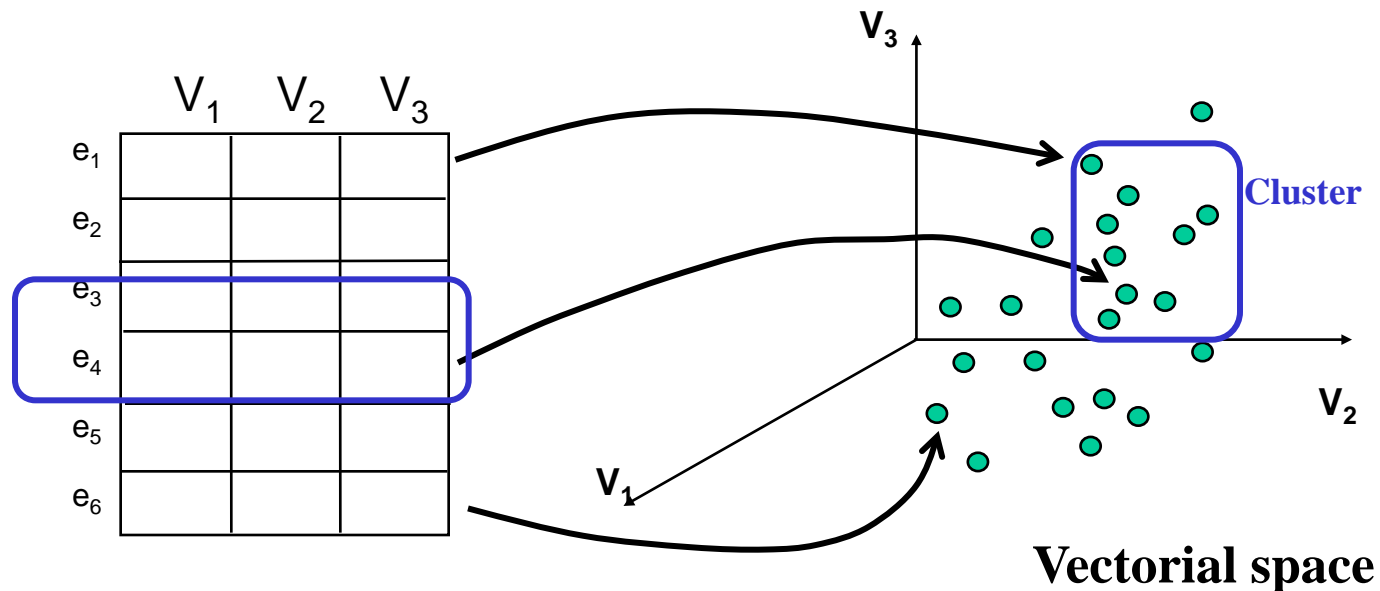
Intension : Outlook=sunny  
and Temperature = hot

*Frequent itemset*: Temperature=hot and Outlook=sunny ?

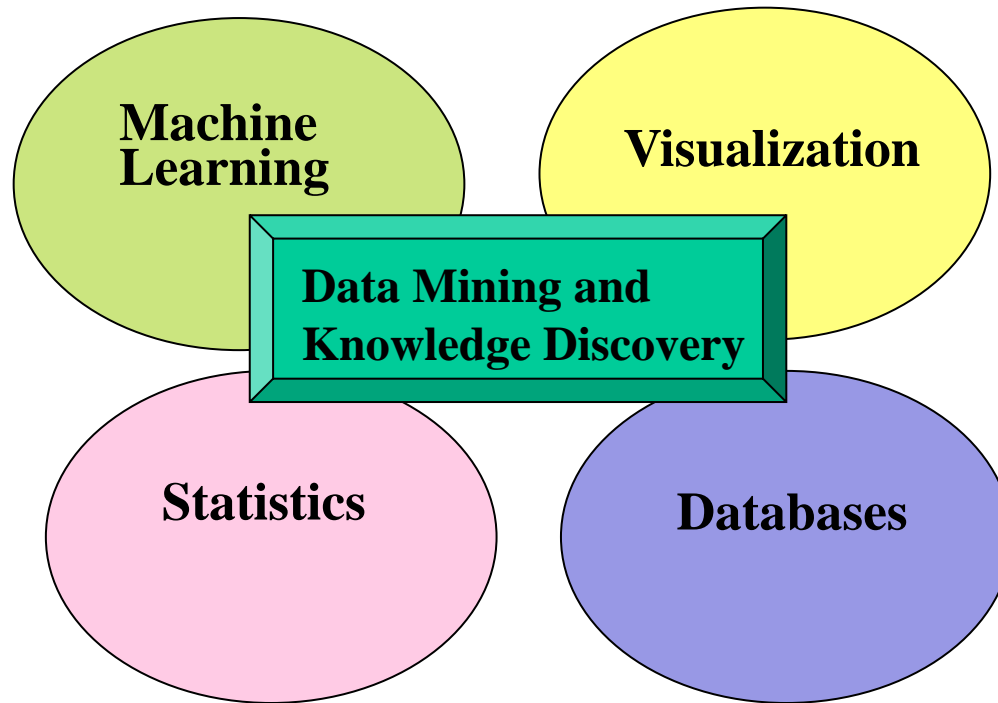
*Rule*: Temperature=hot --> Outlook=sunny ?

# Inputted Data / outputted Knowledge

- A set  $E$  of  $n$  transactions (examples, instances, objects, rows) :  $E = \{e_1, e_2, \dots, e_n\}$
- A set  $V$  of  $p$  variables (attributes, items, descriptors, columns) :  $V_1, V_2, \dots, V_p$ 
  - Each variable  $V_i$  is associated to a domain of values  $D_i$  ( $V_i \in D_i$ )



# Related Fields

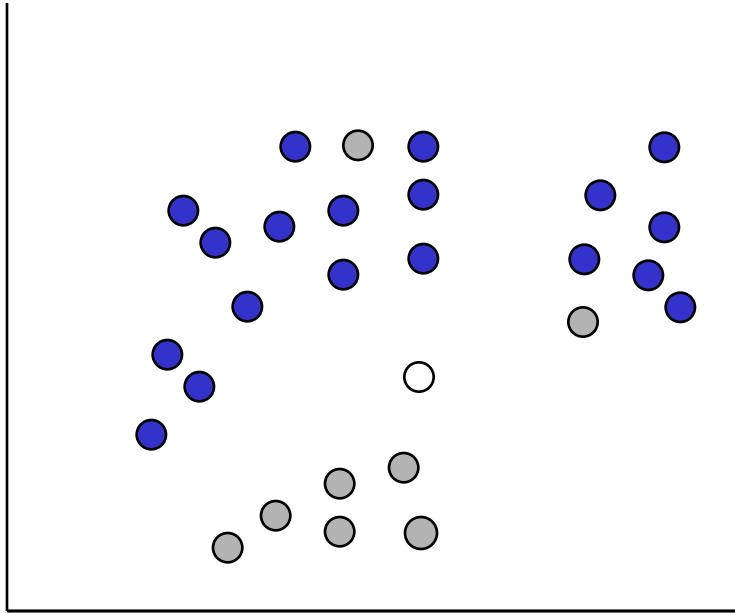


# Major Data Mining Tasks

- **Classification:** predicting an item class
- **Clustering:** finding clusters in data
- **Associations:** e.g.  $A \& B \rightarrow C$
- **Visualization:** to facilitate human discovery
- **Summarization:** describing a group
- **Deviation Detection:** finding changes
- Estimation: predicting a continuous value
- Link Analysis: finding relationships
- ...

# Data Mining Tasks: Classification

**Learn a method for predicting the instance class from pre-labeled (classified) instances**

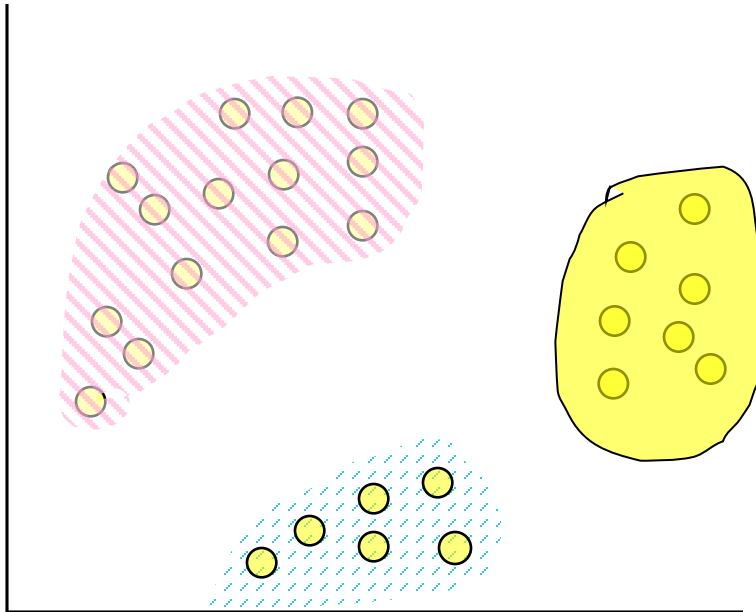


Many approaches, Classifiers :  
Statistics,  
Decision Trees,  
Neural Networks,  
kNN,  
Linear/logistic Regression,  
SVM  
...

Given a set of points from classes ● ●  
what is the class of new point ○?

# Data Mining Tasks: Clustering

**Find “natural” grouping of instances given un-labeled data**



## Unsupervised learning

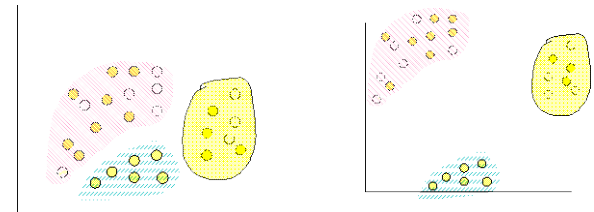
Many approaches :

K-means, EM

Kohonen maps (SOM),

PCA, AHC, MDS

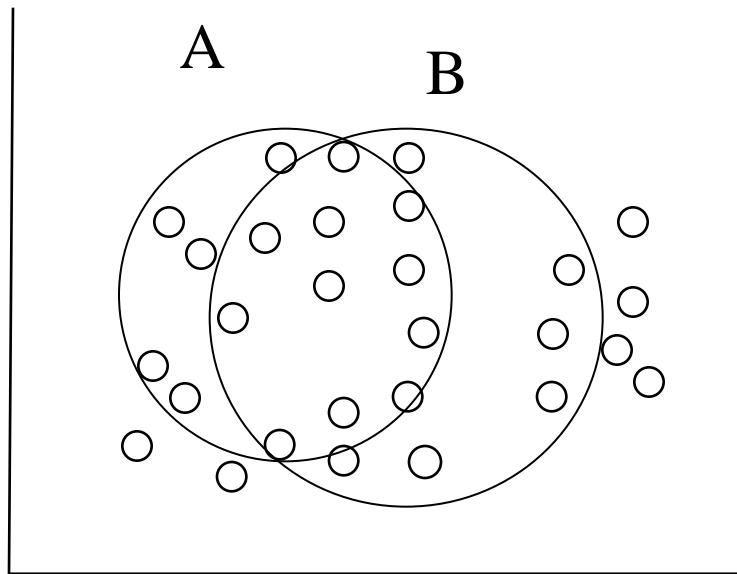
...



# Data Mining Tasks: Association rules

**Implication tendencies between variables :  $A \rightarrow B$**

**Not clustering, not classification**



Apriori algorithm

# Introduction

## KDD Process and Method

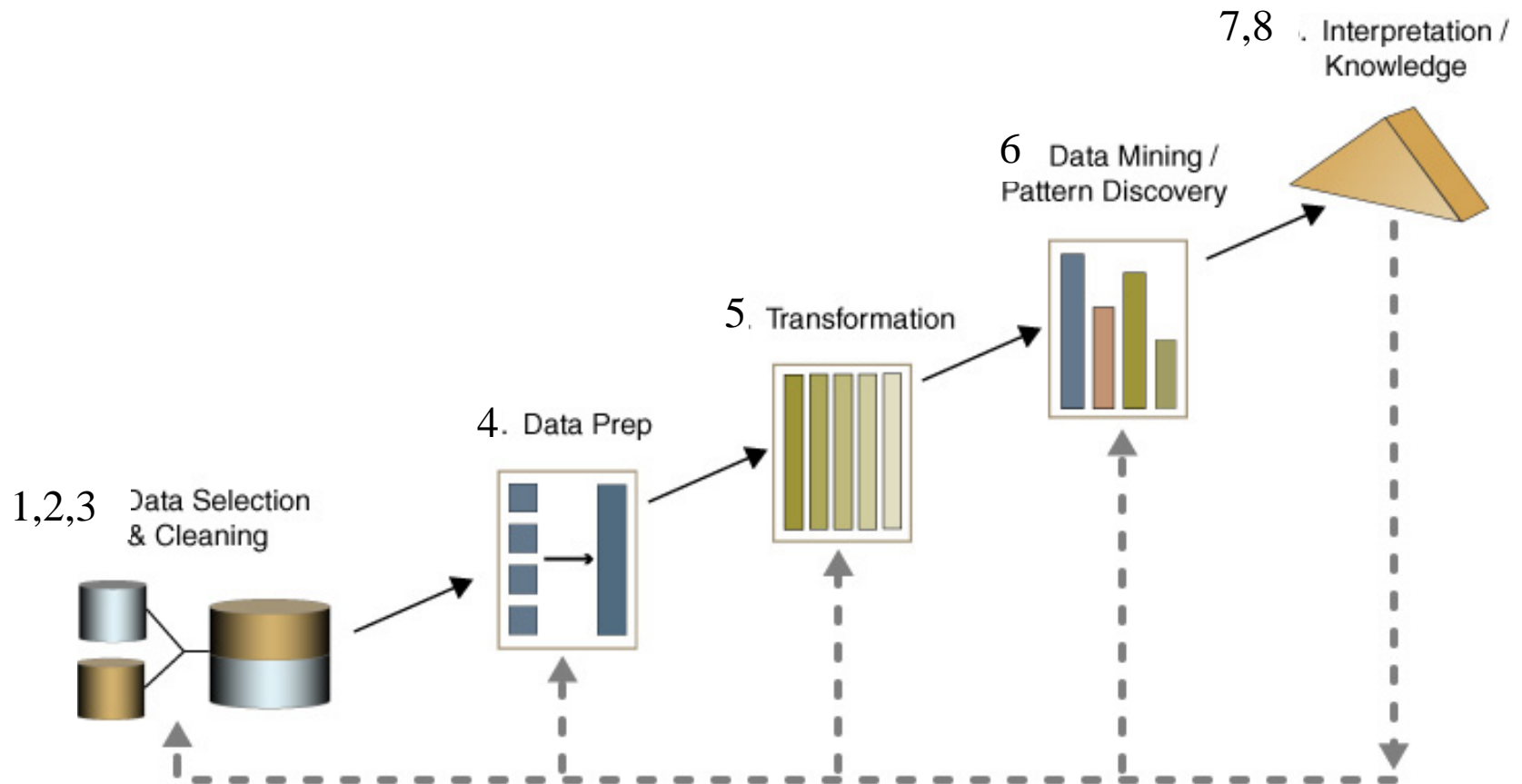


# Phases of the KDD Process

A process + a method

- Step 1 : Problem formulation - definition and goals
- Step 2 : Data collection - localization
- Step 3 : Data selection – sampling and feature selection
- Step 4 : Data cleaning
- Step 5 : Transformations on variables
- Step 6 : DM Model selection
- Step 7 : Interpretation, Evaluation
- Step 8 : Knowledge deployment, storing and management

# Knowledge Discovery



An Overview of the Steps That Compose the KDD Process

# KDD process :

## 1 Problem formulation

- Problem, goal, evaluation of the result (quality measure?)
- 2 types of problems :
  - Supervised (classification) : find an explanation of a precise phenomena (i.e. a given class variable). Ex: churn of customers
  - Unsupervised (clustering) : discovery of new clusters, with a given criterion (homogeneity)
- Form of the result : (readable for a decision maker...)

# KDD Process :

## 2 Data collection

Structure générale des données (forme de l'ensemble d'apprentissage)

- Data Sources : Information System, BD, Data warehouse
- Relevant variables (potential of explanation, reduction of dimensions)
- Ratio : number of variables / number of examples
  - + /+ : time consuming
  - +/- : too few examples => false generalizations
  - -/+ : optimal zone
  - -/- : instability

# KDD Process :

## 3 Data selection

Learning on a sample

- Sampling strategy (sample/whole set)
  - Advantage : accuracy evaluation + reduction of the computing time.
  - Inconvenient : sampling strategy ? (random, quotas, stratified),
  - Pareto law : 20/80%

# KDD Process :

## 4 Data cleaning

- Data quality and DB
- Processing of null values, missing values, and bad values
- Hard and time consuming step
- Manual strategy on reduced DB
- Few automated strategy on large DB (ex: distribution test)

# KDD Process :

## 5 Feature selection

Adaptation to next stage (data maining),  
or transformations for the decision maker

- Monovariabe transformations
  - Change the unit of a measure, normalization of values, ...
  - Change date/duration, place -> coordinate (geographic data), ...
  - Domain transformation :
    - symbolic -> numeric,
    - discretization (continue -> discrete)
    - binarization : n symbolic variables -> n binary variables
- Multivariable transformations
  - ratios of 2 variables
  - combinaisons of variables (linear ot not)

# KDD Process :

## 6 model selection

the kernel step : Selection of the DM model

- Learning strategy with sampling
  - Learning sample + test sample
- Model : formal representation/pattern, algorithm, parameters
- 3 kinds of models :
  - Numeric
    - Numerical formula (combination of variables) :
    - Linear regression, clustering, EM, PCA, AHC...
  - Logic
    - Logic formula (conjunction of variables, rules)
    - Decision trees, Bayesian networks, association rules, FCA ...
  - Exploratory
    - Data Projections => graphical/visual evaluation (clouds of points)
    - Factorial analysis (PCA,..), Kohonen maps, data visualization, ...



# KDD Process :

## 7 Interpretation, Evaluation

- Evaluation by the decision maker (subjective : new, usefull, actionable)
- Validation with specific quality measure depending on the model
  - Error on test sample
  - Confidence interval with a bootstrap sampling
  - Entropy of decision tree, support and confidence of rules, ...
- Readability of the results
  - ex : 200 000 association rules ?

=> Need for visual representations

# KDD Process :

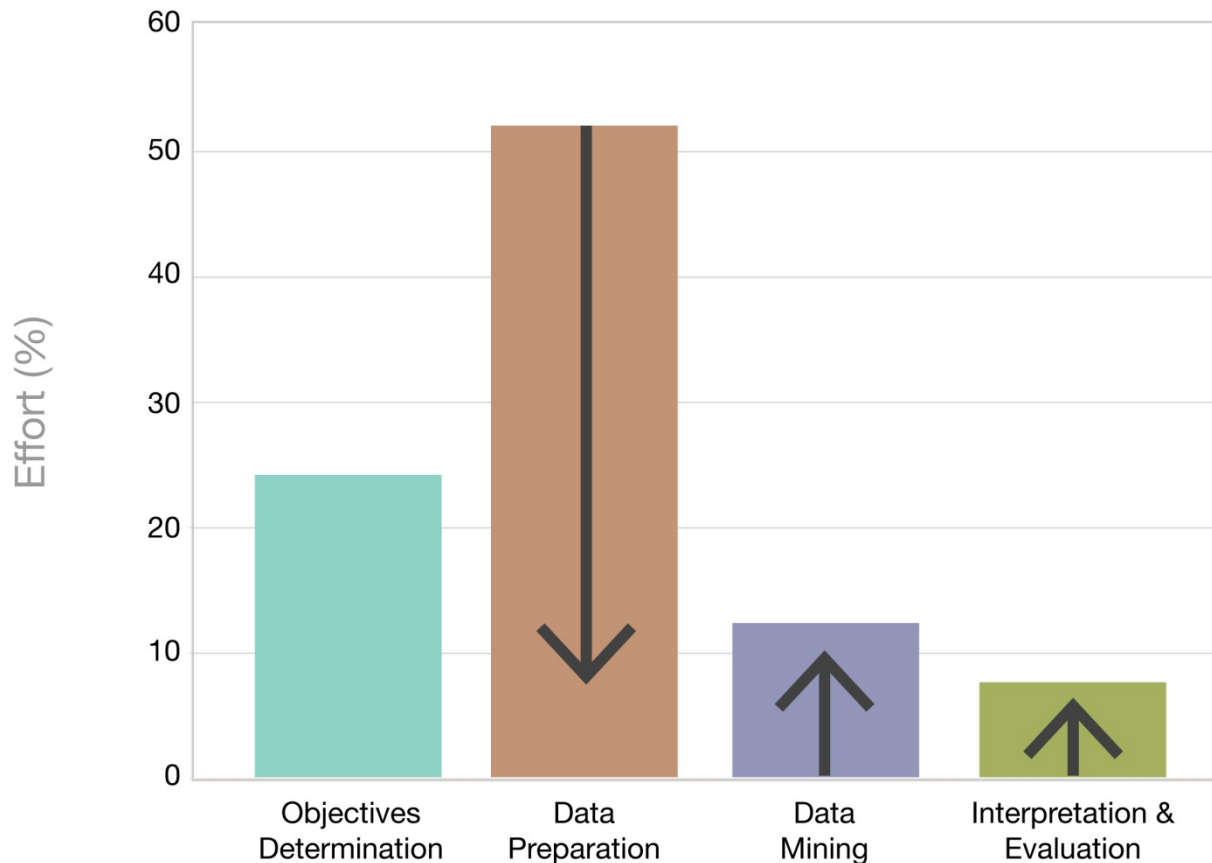
## 8 Knowledge deployment

Consequence of the discovered knowledge : decision -> action

- Effect of the decision (marketing strategy, ...)
- A report ?
- The knowledge is formalized + stored in the IS (data warehouse)
- Toward knowledge management (capitalization)?

# Required effort for each KDD Step

- Arrows indicate the direction we hope the effort should go.



# KDD Process :

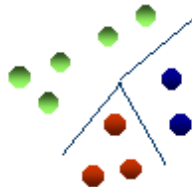
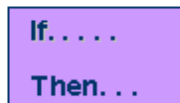
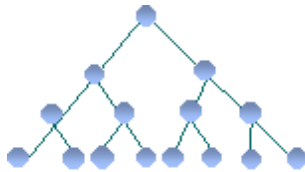
## Conclusion

- A decision making oriented system (not an automated system)
- A set of tools and models (Jack knife)
- Oriented to decision makers (not data analysts)
- Need for graphical interfaces, simplicity, readability, automated setting of parameters
- Supporting the scale factors (large data)
- Many progresses keep to make...

# Introduction

## KDD models

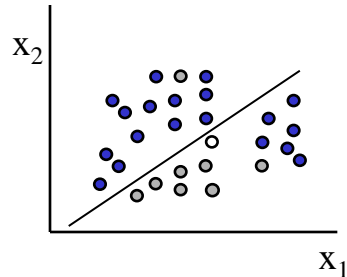
Models are...



- Decision Trees
- Nearest Neighbor Classification
- Neural Networks
- Rule Induction
- K-means Clustering
- Patterns

# Classification (supervised learning)

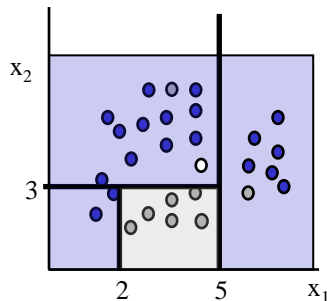
Classifiers: Linear/logistic regression, Decision tree, Perceptron /Neural networks, K-nearest-neighbors, ...



**Linear Regression** :  $\beta_0 + \beta_1 x_1 + \beta_2 x_2 \geq 0 \quad (\sum_i \beta_i x_i)$

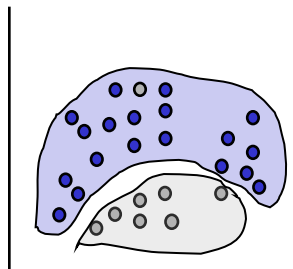
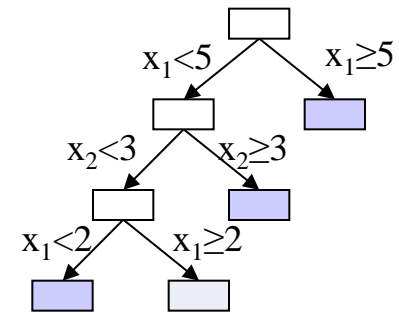
Regression computes  $\beta_i$  from data to minimize squared error to 'fit' the data

Not flexible enough



**Decision Trees** (ID3, C4.5, CART, ...):

if  $X > 5$  then blue  
else if  $Y > 3$  then blue  
else if  $X > 2$  then green  
else blue



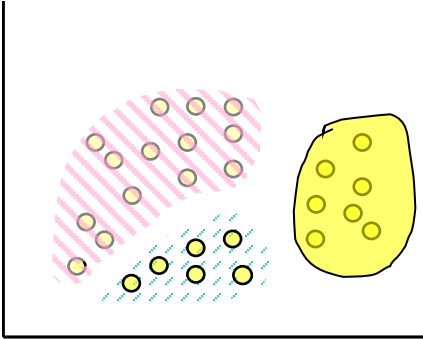
**Neural Nets** :

Can select more complex regions

Can be more accurate

Also can overfit the data – find patterns in random noise

# Clustering (non supervised learning)



## Non hierarchical:

**K-means, PAM, SOM, DBscan, SVM, mixtures models,  
topic modeling (LSA, LDA, ), ...**

## Representation learning (PCA, MDS, TSNE, word2vec, ), autoencoders...

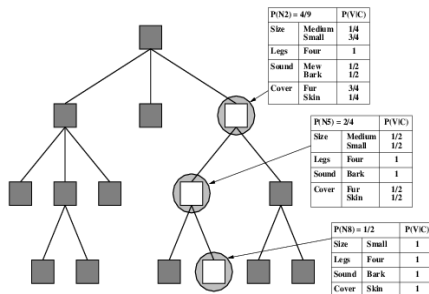


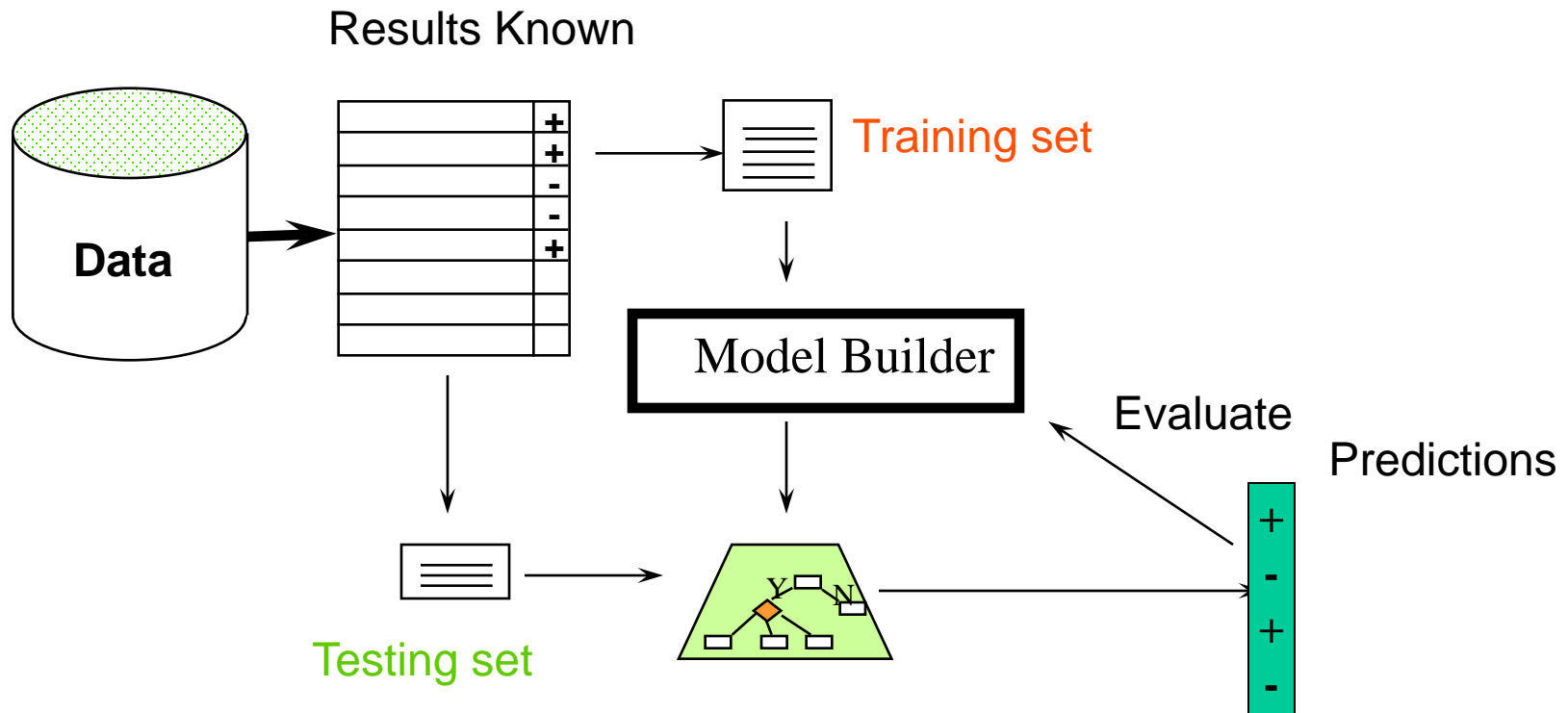
Figure 1. A small hierarchy of probabilistic concepts for the domain of household pets, illustrating COBWEB's representation and organization of knowledge.

## Hierarchical :

## HAC, FCA, Optics, COBWEB, ...



# Learning and Classification



# Map of learning models (scikit)

## **Supervised learning** (regression/classification, kernel, probabilistic)

- Linear Models : linear regression, ridge, lasso, bayesian regression, logistic regression, Linear/Quadratic Discriminant Analysis, Kernel ridge regression, Isotonic regression, Cross decomposition (PLS, CCA) ...
- Support Vector Machines (regression, classification, kernel)
- Stochastic Gradient Descent
- Nearest Neighbors (regression, classification)
- Gaussian Processes (regression, classification, kernel)
- Naïve Bayes
- Decision Trees (classification, regression, ID3, C4.5, C5.0, CART)
- Pattern based classification : contrast sets, associative classification
- Neural network models (supervised)
- Ensemble methods : Bagging methods, AdaBoost, gradient, voting, Random Forests, ...
- Multiclass and multilabel
- Semi-Supervised
- (Feature selection, Probability calibration)

+ model selection (cross validation) + hyperparameter tuning + evaluation/validation metrics ( precision, Jaccard, F-measure, mse, r2, ...)

## **Unsupervised learning** (numeric, kernel, probabilistic)

- Clustering : K-Means, Affinity Propagation, Spectral clustering, Hierarchical clustering, DbScan, Optics, Birch
- Graph clustering :
- Binary/symbolic clustering : FCA, Freq Patterns
- Gaussian mixture models
- Neural network models (unsupervised) : SOM, Boltzman, autoencoders, word2vec, ...
- Linear dimensionality reduction (matrix factorization problems) : PCA, SVD, kernel PCA, ICA, NMF, LDA
- Manifold learning (non-linear dimensionality reduction) : isomap, Hessian / Laplacian Eigenmaps, MDS T-SNE, ...
- Biclustering
- Novelty and Outlier Detection
- Covariance estimation, Density Estimation

+ model selection (bootstrap) + hyperparameter tuning + evaluation/validation metrics (silhouette, Dunn, Calinsky, ...)

# KDD models : Decision trees

# KDD models : Decision trees

Supervised learning, Logic Model, discrete variables (binary or symbolic)

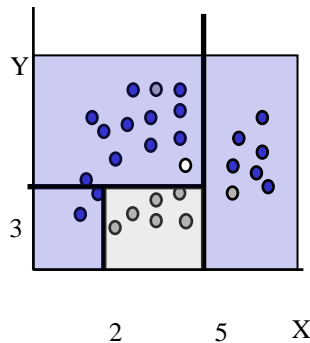
Partitioning among variable values

A tree (hierarchy of partitions)

Recursive algorithm

Classification rules : a path root-leaf (+ measure)

goal : 1 class of object in a leaf (homogeneity)



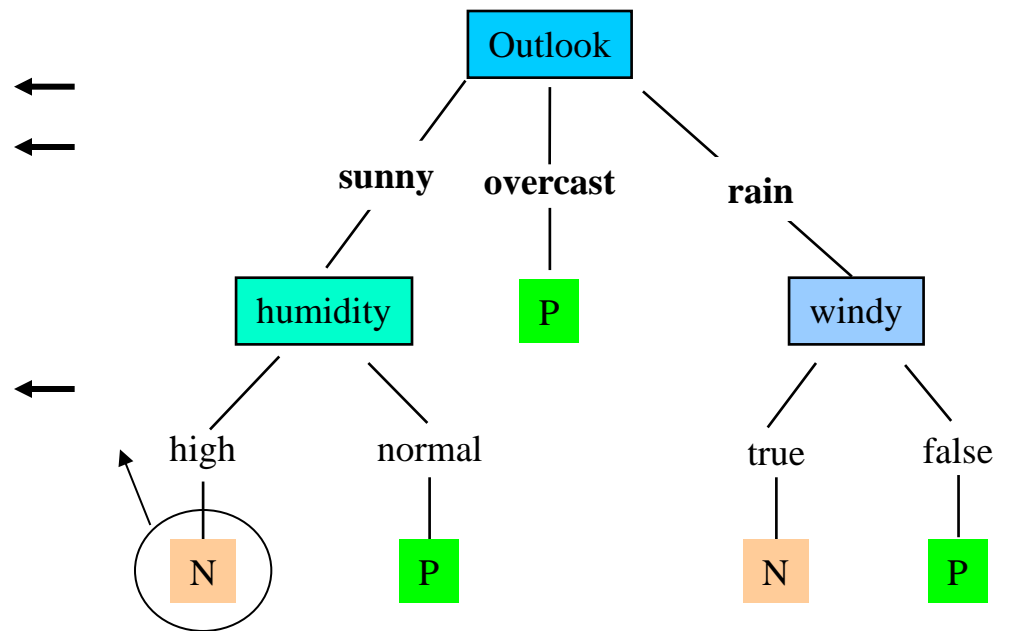
**Decision Trees** (ID3, C4.5, CART, ...):

if  $X > 5$  then blue  
else if  $Y > 3$  then blue  
else if  $X > 2$  then green  
else blue

# KDD models : Decision trees

## Exemple 1

Outlook	Tempreature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N



# KDD models : Decision trees

## criteria

- Hierarchical Partitioning :
  - Selection the best variable (selection criteria).
  - Partition the data among the values of the variable.
  - Recursively apply on each new part or stop (stop criteria).
- Examples of criteria :
  - entropy :  $-\sum p_i \log_2(p_i)$  ,  $p_i = \text{proba}(C=C_i)$  probability of the class  $C_i$  in the part
  - Gini  $1 - \sum_{i=1}^n p_i^2$
  - Chi 2
  - ...

# KDD models : Decision trees

## Exercise and learn&test

- Mushroom dataset

	Odorant	Anneau1	Chapeau bombé	Pied large	Taches	Comestible
champignon 1	1		1	1		1
champignon 2	1		1	1	1	1
champignon 3			1	1	1	1
champignon 4	1			1	1	
champignon 5			1	1		1
champignon 6	1		1	1		1
champignon 7			1	1		1
champignon 8	1	1		1		1
champignon 9			1	1		
champignon 10	1	1	1	1		1
champignon 11			1		1	1
champignon 12		1	1	1		
champignon 13	1	1	1	1		
champignon 14	1		1	1		1
champignon 15				1	1	1
champignon 16			1	1		1
champignon 17		1			1	
champignon 18	1		1	1	1	1
champignon 19	1		1	1		1
champignon 20			1	1		1
champignon 21	1	1	1		1	1
champignon 22	1		1	1	1	1
champignon 23			1	1	1	1
champignon 24			1			

# KDD models : Decision trees

## Exercise and learn&test

- Mushroom dataset

### Contengency Table

	C=0	C=1	
V=0	n00	n10	n.0
V=1	n01	n11	n.1
	n0.	n1.	n

### Frequencies

$p00 = \text{proba}(C=0|V=0) = n00 / n.0$   
 $p10 = \text{proba}(C=1|V=0) = 1 - p00$   
 $p01 = \text{proba}(C=0|V=1) = n01 / n.1$   
 $p11 = \text{proba}(C=1|V=1) = 1 - p01$

$p0. = n0. / n$   
 $p1. = n1. / n = 1 - p0.$

**Entropy**  $E(p0, p1, p2, \dots) = -p0 \log_2(p0) - p1 \log_2(p1) - p2 \log_2(p2) - \dots$

Global Entropy of the table (before splitting with variable V)

$E(C) = E(p0., p1.) = -p0. \log_2(p0.) - p1. \log_2(p1.)$

Conditional entropy of the 2 subtables (after splitting with variable V)

$E(C|V=0) = E(p00, p10)$

$E(C|V=1) = E(p01, p11)$

$E(C|V) = p.0 E(C|V=0) + p.1 E(C|V=1)$

Difference of entropy

$EE(V) = E(C) - E(C|V)$

Entropy Gain

$GE(V) = EE(V) / E(C)$



# KDD models : Decision trees

## Exercise and learn&test

- Mushroom dataset

First level of the tree :

Contengency Tables

	Positive Exemples (1)	Negative Exemples (0)	Total
Comestible	18	6	24

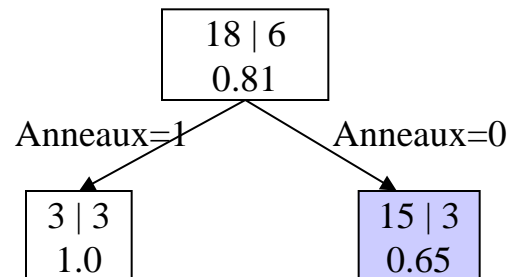
Odorant	10	2	12
non Odorant	8	4	12
Anneaux	3	3	6
non Anneaux	15	3	18
Bombé	16	4	20
non Bombé	2	2	4
Large	16	4	20
non Large	2	2	4
Taches	8	2	10
non Taches	10	4	14

Conditional Entropy	Entropy of partition	Entropy gain
	<b>0,8112781</b>	

0,650022422	0,7841591	3,34%
0,918295834		
1	0,7375168	9,09%
0,650022422		
0,721928095	0,7682734	5,30%
1		
0,721928095	0,7682734	5,30%
1		
0,721928095	0,8042904	0,86%
0,863120569		

Conditional Gini	Gini of partition	Gini gain
	<b>0,375</b>	

0,27777778	0,361111	3,70%
0,44444444		
0,5	0,333333	11,11%
0,27777778		
0,32	0,35	6,67%
0,5		
0,32	0,35	6,67%
0,5		
0,32	0,371429	0,95%
0,40816327		



# KDD models : Decision trees

## Exercise and learn&test

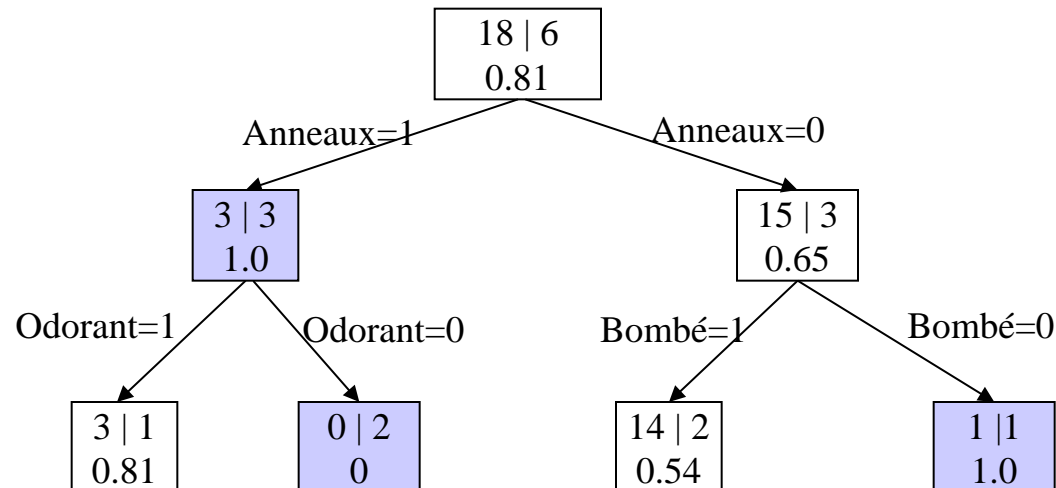
- Mushroom dataset

	Positive Exemples (1)	Negative Exemples (0)	Total
Comestible	3	3	6
Odorant	3	1	4
non Odorant	0	2	2
Bombé	2	2	4
non Bombé	1	1	2
Large	2	2	4
non Large	1	1	2
Taches	1	1	2
non Taches	2	2	4

Conditional Entropy	Entropy of partition	Entropy gain
	1	
0,811278124	0,5408521	45,91%
0		
1	1	0,00%
1		
1	1	0,00%
1		
1	1	0,00%
1		
1	1	0,00%
1		

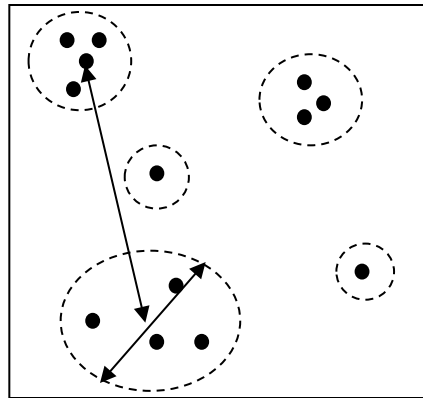
	Positive Exemples (1)	Negative Exemples (0)	Total
Comestible	15	3	18
Odorant	7	1	8
non Odorant	8	2	10
Bombé	14	2	16
non Bombé	1	1	2
Large	14	2	16
non Large	1	1	2
Taches	6	2	8
non Taches	8	2	10

Conditional Entropy	Entropy of partition	Entropy gain
	0,6500224	
0,543564443	0,6426554	1,13%
0,721928095		
0,543564443	0,5942795	8,58%
1		
0,543564443	0,5942795	8,58%
1		
0,811278124	0,7616392	-17,17%
0,721928095		



# KDD Models : K-means

## KDD Models : K-means clustering



- Principle of k-means [macQueen 1967]
  - Clustering (non supervised)
  - create  $k$  disjointed classes with 2 criteria
    - (1) Small clusters (small diameter)
    - (2) Not close clusters (large distance between center of clusters)

# KDD Models : K-means clustering

The basic steps to follow given a raw dataset

## 1. Initialization

- Define the number of clusters ( $k$ ).
- Designate a cluster centre (centroid) for each cluster.

2. Assign each data point to the closest cluster centre.

3. Calculate the new cluster centre

4. Repeat 2 until stability of centroids.

# KDD Models : K-means

## Algorithm

### **Entrée :**

- k : number of clusters
- X : les éléments à classer

Choisir k noyaux  $N_i$  initiaux

Initialiser les k classes  $C_i$  avec leur noyau  $N_i$ :  $C_i = \{N_i\}$

### **Répéter**

#### **Pour tout x de X faire**

$C(x) = C_i$  la classe du noyau le plus proche/similaire de x

$C_i = C_i \cup \{x\}$  ajouter x à la classe

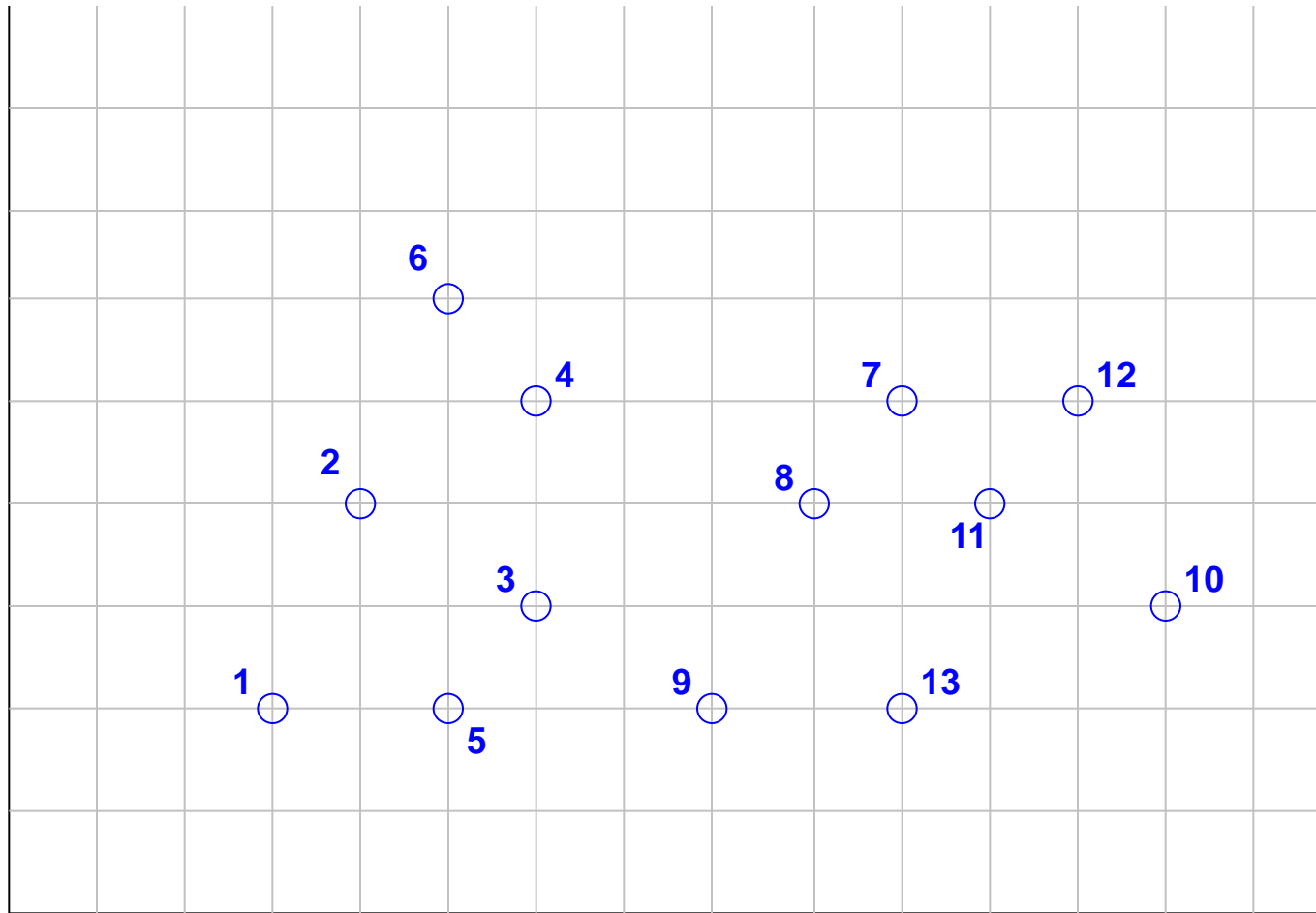
$N_i = B(C_i)$  nouveau noyau (barycentre de la classe)

#### **Fait**

**Tant que** noyaux non stables

## KDD Models : K-means

### Example



## KDD Models : K-means

Example : dependance to initial centroids

$K = 2$

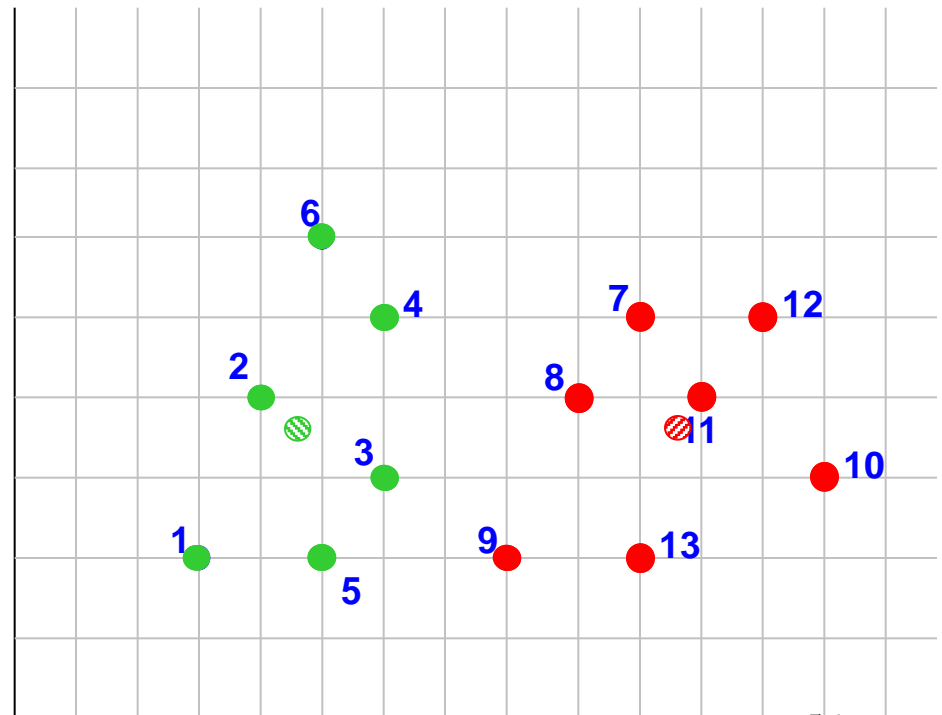
1st solution

**N1** = (6,3) = point 3

**N2** = (8,2) = point 9

$B(\mathbf{C1}) = (4.8, 3.7)$

$B(\mathbf{C2}) = (10.4, 3.6)$





# K-Means

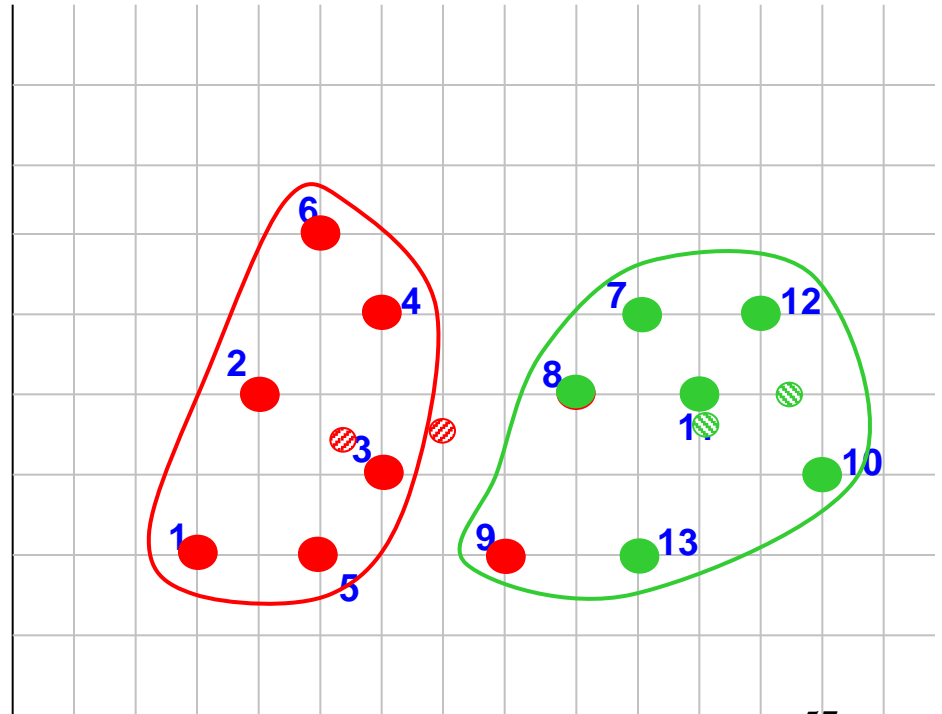
Example : dependance to initial centroids

$K = 2$

2nd solution

**N1** = (13,3) = point 10

**N2** = (11,2) = point 13



# K-means

## Clustering Quality : Validity criterion

- Pour trouver les meilleurs classes, il faut
  - Minimize the distance intra-cluster (intra)
  - Maximize the distance inter-clusters (inter)

$$intra = \frac{1}{N} \sum_{i=1}^k \sum_{x \in C_j} ||x - z_j||^2$$

$$inter = \min(||z_i - z_j||^2)$$

$N$ : number of points $z_i$ : centre du groupe $i$
--

- Maximize the validity  $V$  :

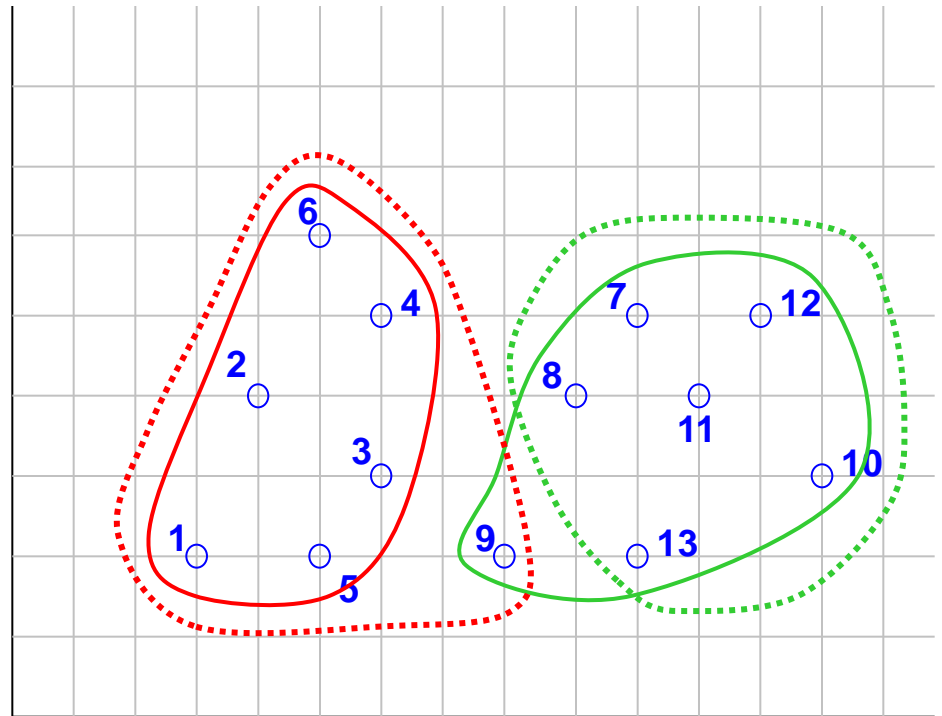
$$V = \frac{inter}{intra}$$

# k-means

## Example of validity

Intra = 3.66 Inter = 31.3  
V = 8.55

Intra = 3.75 Inter = 30.9  
V = 8.24



# K-Means

## Optimization of the number k of cluster

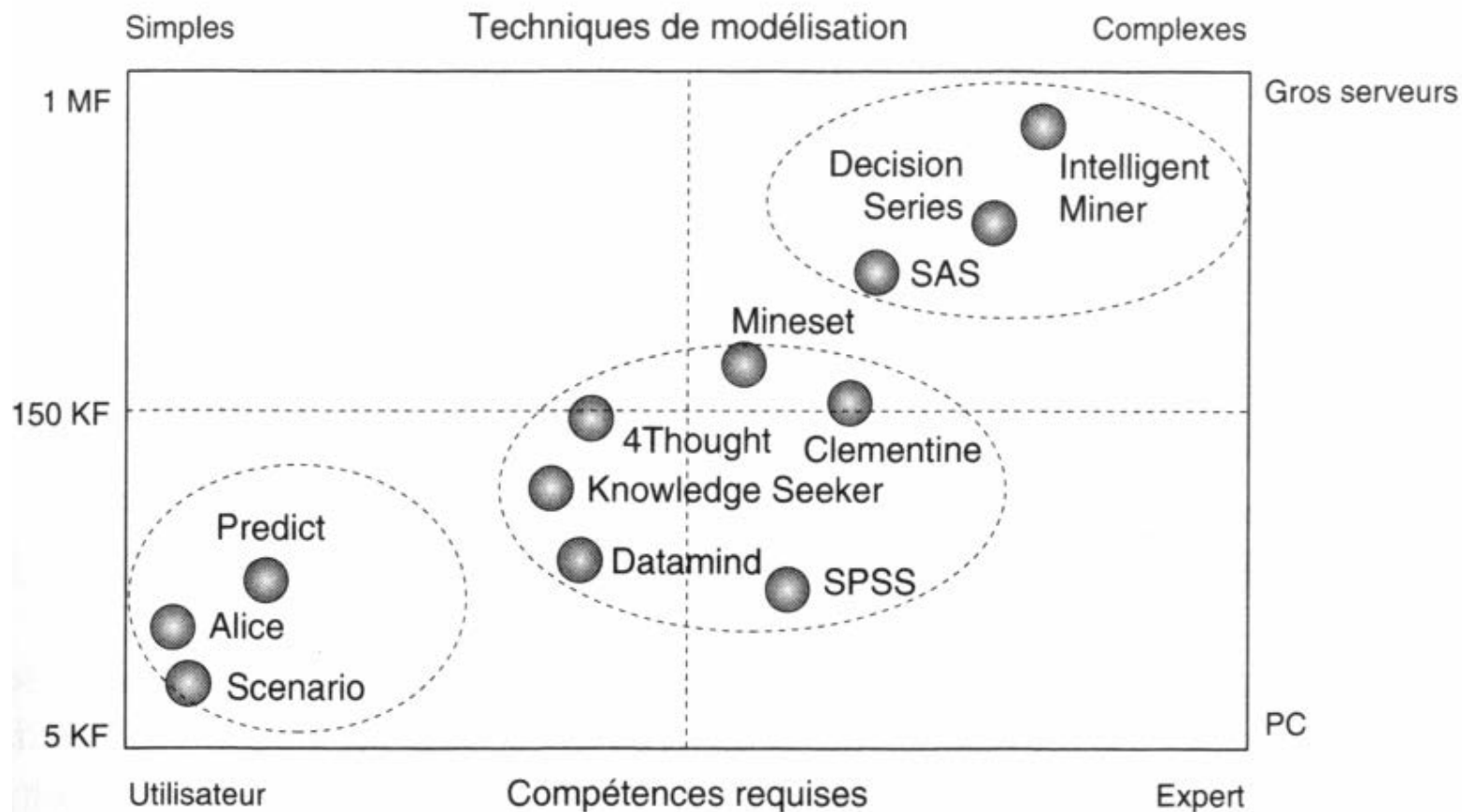
- k : given by the user
- Increase k from 2 until  $V(k)$  is maximal

$$V(k-1) < V(k)$$

$$V(k+1) < V(k)$$

Some KDD softwares

# Quelques logiciels de Data Mining

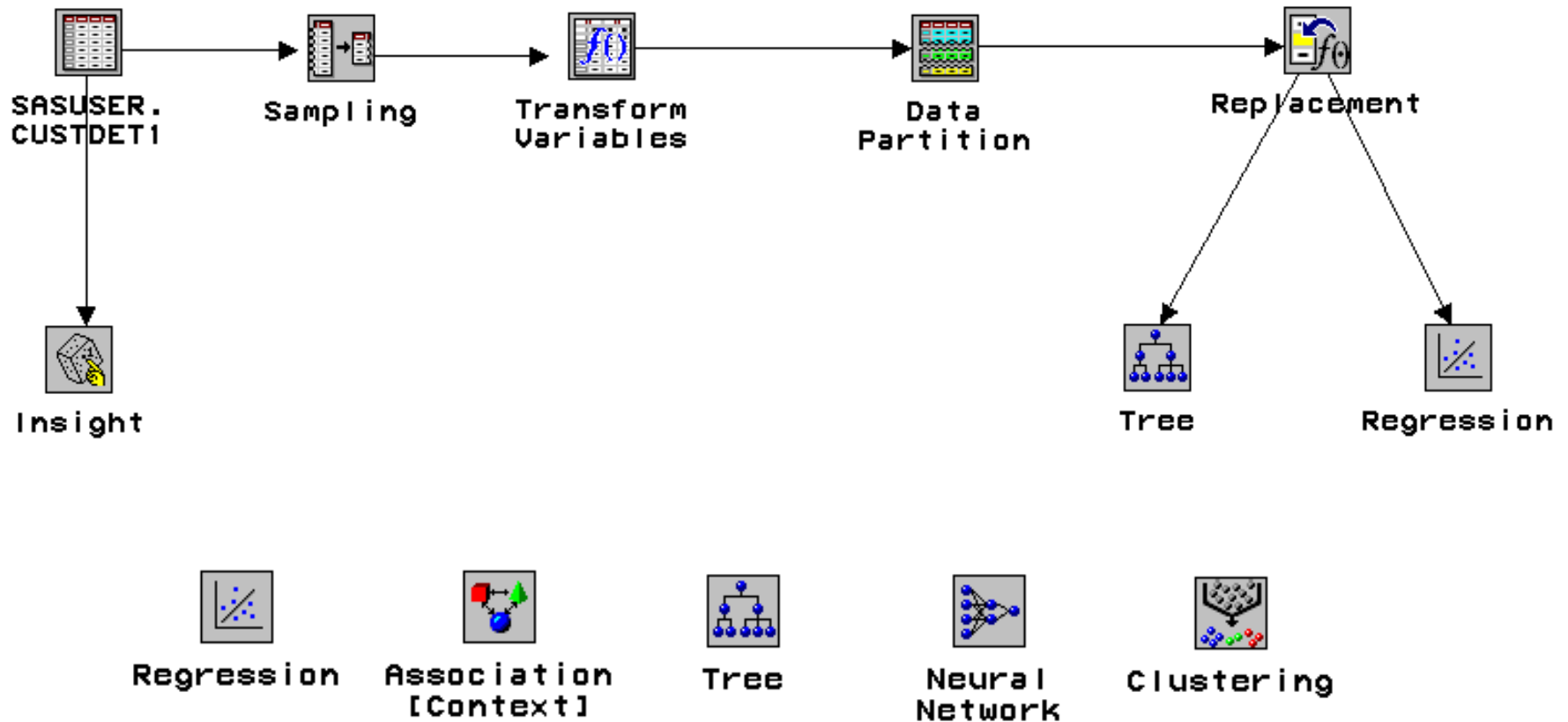


## Some Data Mining softwares

- Commercial :
  - SAS – Enterprise Miner
  - IBM – Intelligent Miner, SPSS
  - Oracle – ODM
  - KXEN
  - SPADE
  - SISENSE...
- Non commercial:
  - Statistiques : R, Tanagra
  - Orange (Python), Weka (java), Knime, rapidMiner, Lavasoft, ...
  - Python ecosystem : scipy, matplotlib, pandas, scikit-learn, skimage, tensorflow-keras
  - R+Python : anaconda, jupyter/jupyterlab
- Others
  - Big Data, map-reduce, hadoop, amazon AWS, apache SPARK-Mlib (Scala, Java, Python, R)

# Data Mining softwares

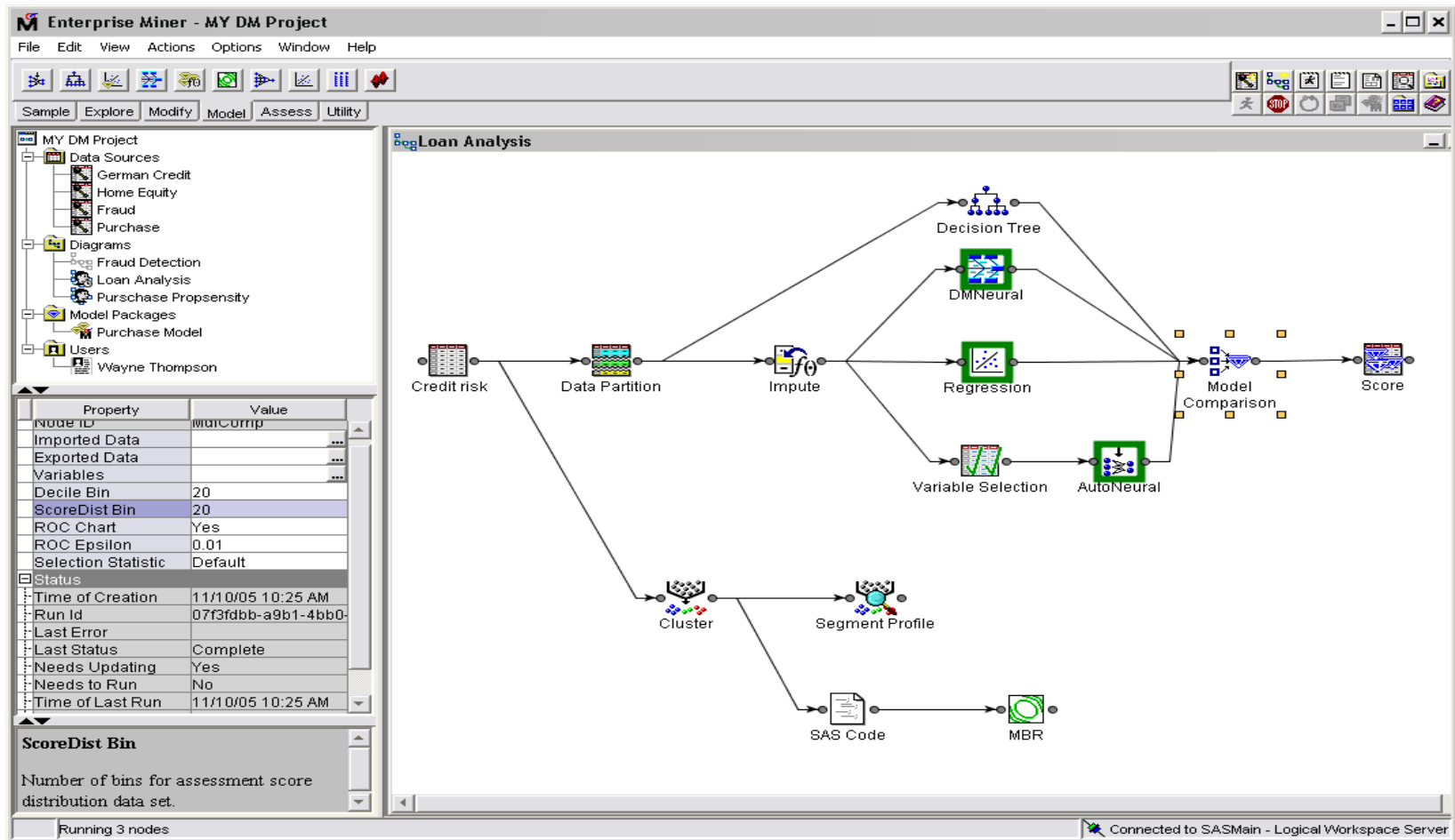
## SAS Enterprise Miner





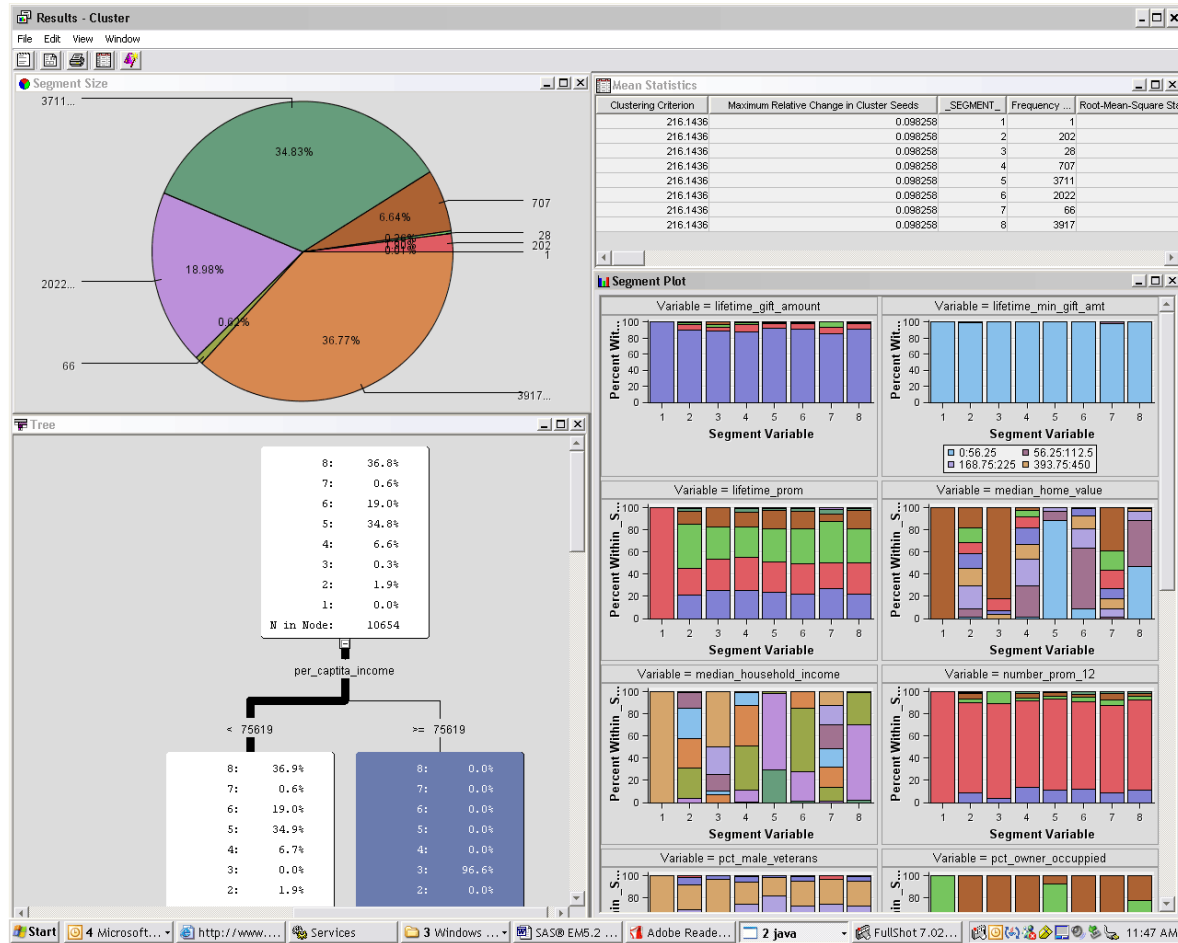
# Data Mining softwares

## SAS Enterprise Miner



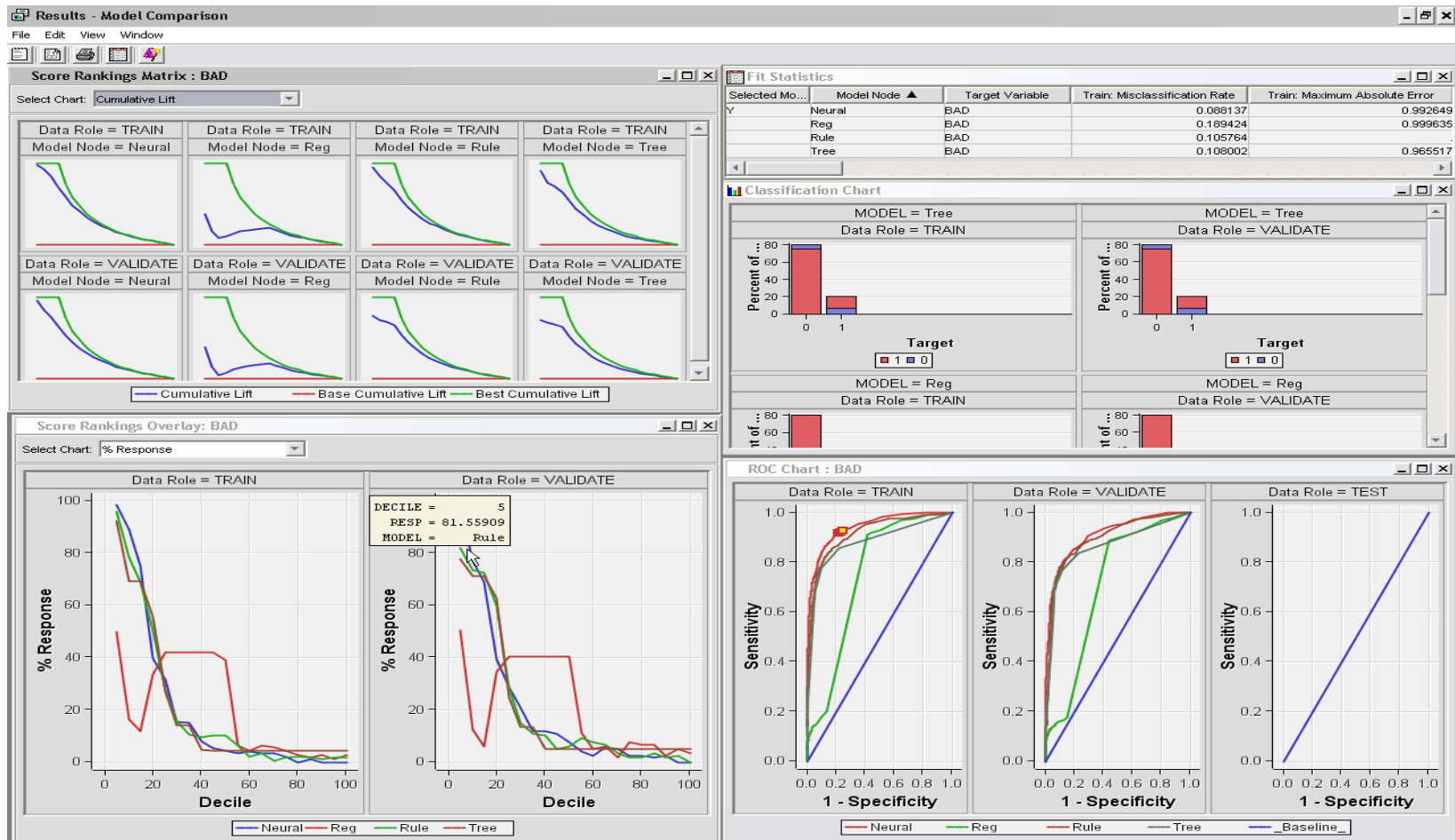
# Data Mining softwares

## SAS Enterprise Miner



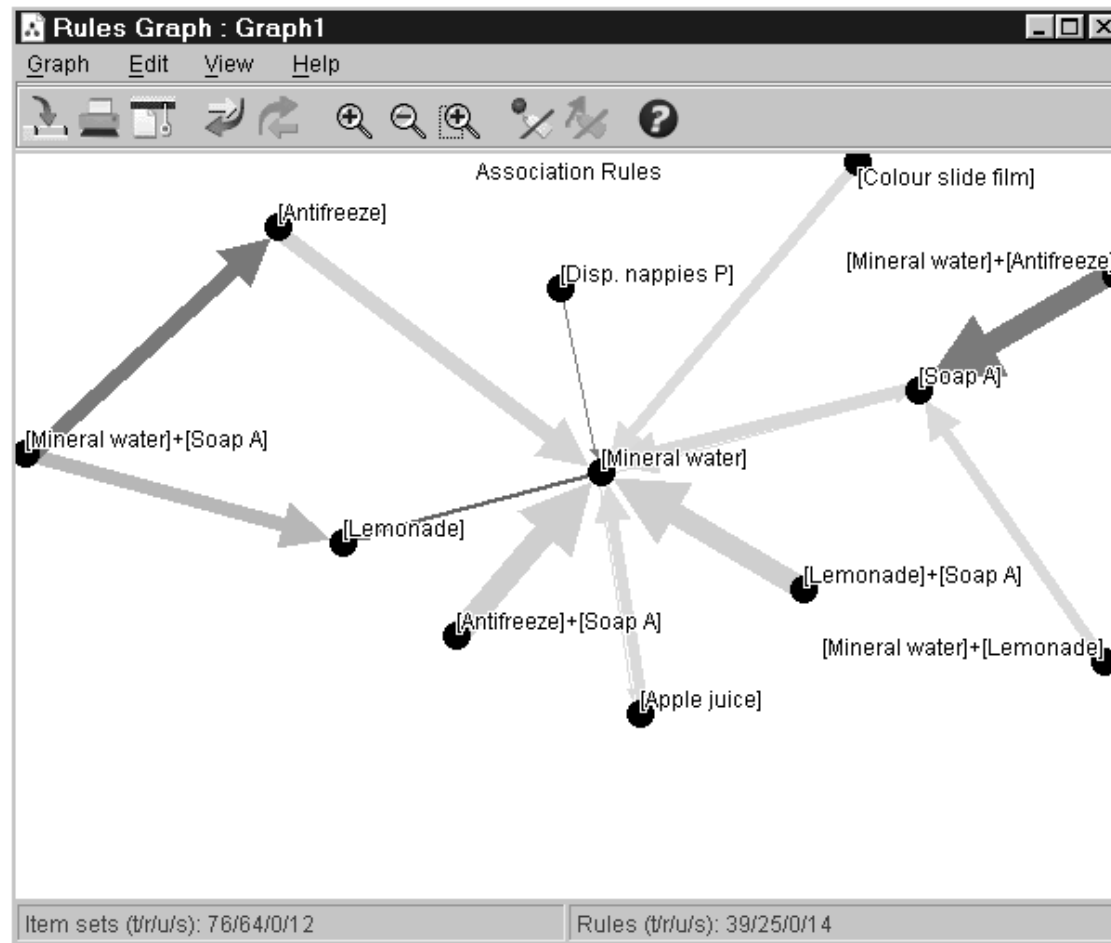
# Data Mining softwares

## SAS Enterprise Miner



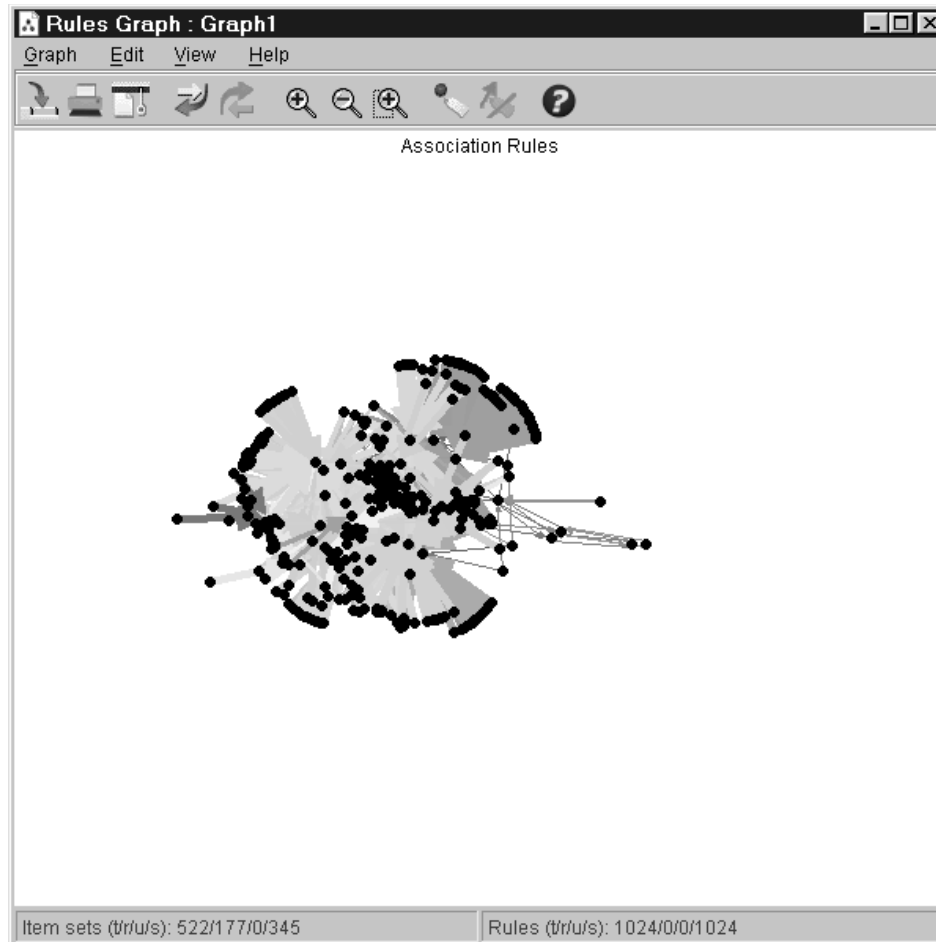
# Data Mining softwares

## Intelligent Miner (IBM)

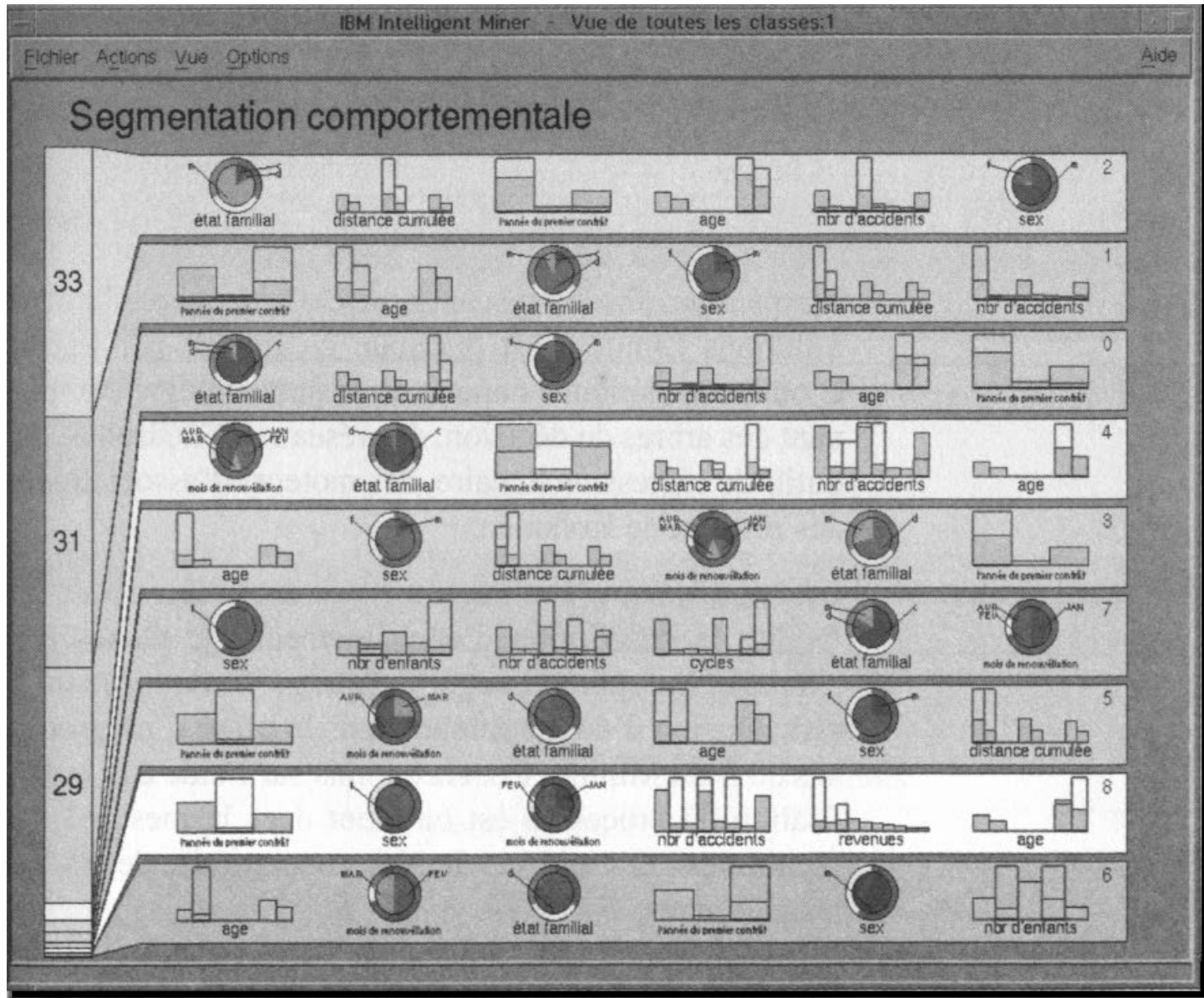


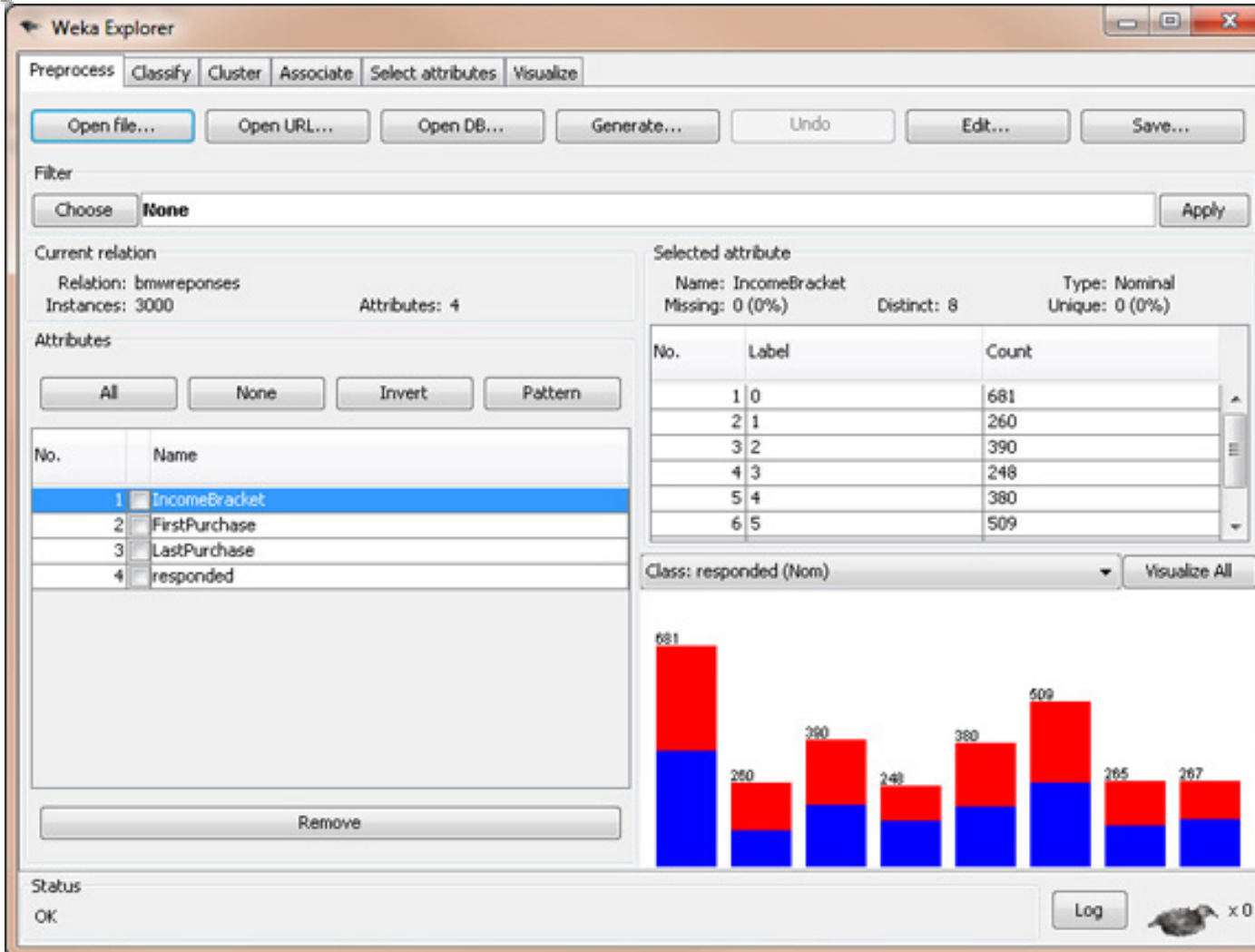
# Data Mining softwares

## Intelligent Miner (IBM)



# Intelligent Miner (IBM)





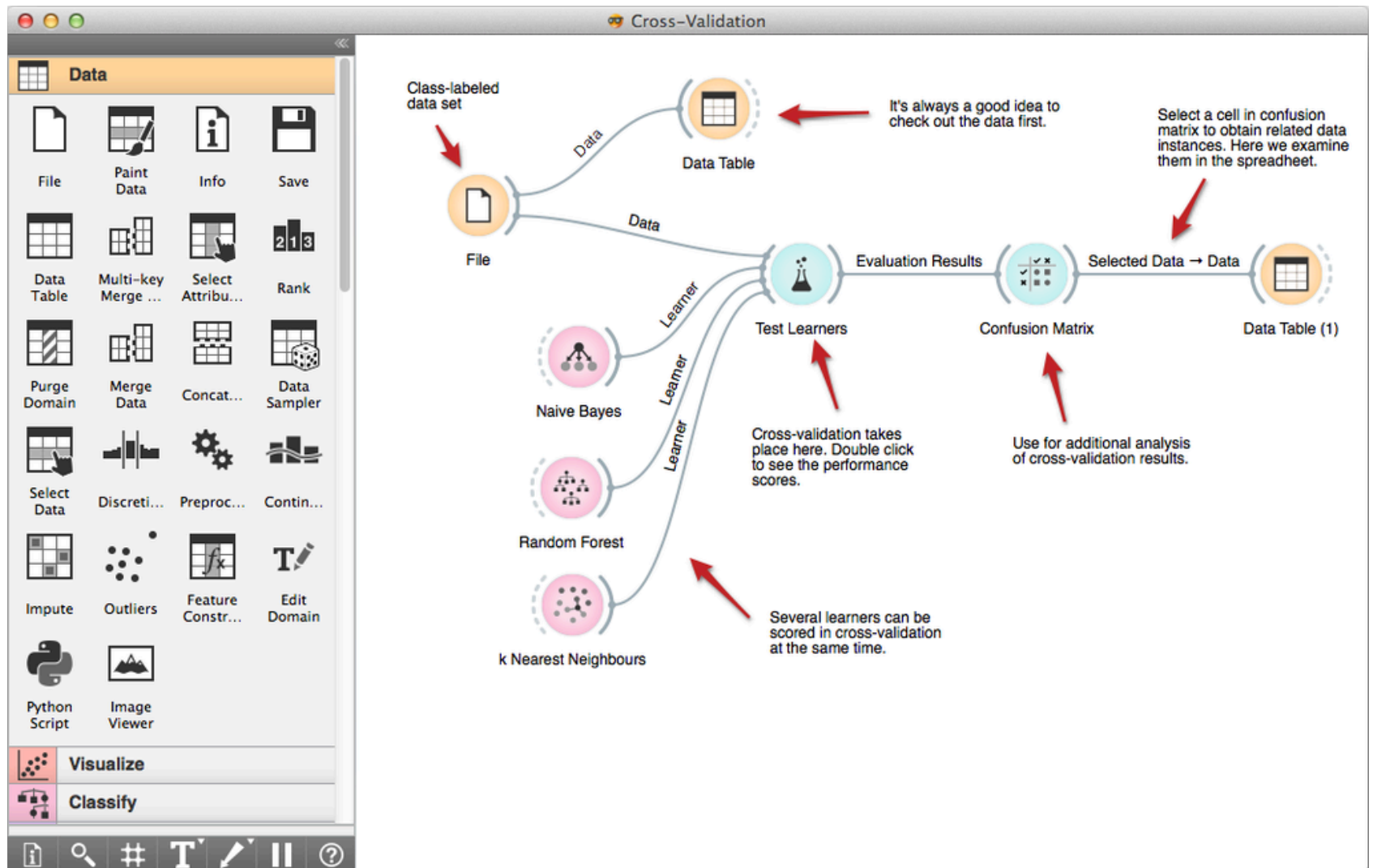
The screenshot displays the Weka Explorer interface. The 'Preprocess' tab is active, showing the 'IncomeBracket' attribute selected. The 'Current relation' is 'bmwreponses' with 3000 instances and 4 attributes. The 'Attributes' list includes 'IncomeBracket', 'FirstPurchase', 'LastPurchase', and 'responded'. The 'Selected attribute' section shows 'IncomeBracket' with 8 distinct values and 0 missing values. A bar chart visualizes the distribution of 'IncomeBracket' values, with the y-axis representing the count of instances for each value.

No.	Label	Count
1	0	681
2	1	260
3	2	390
4	3	248
5	4	380
6	5	509

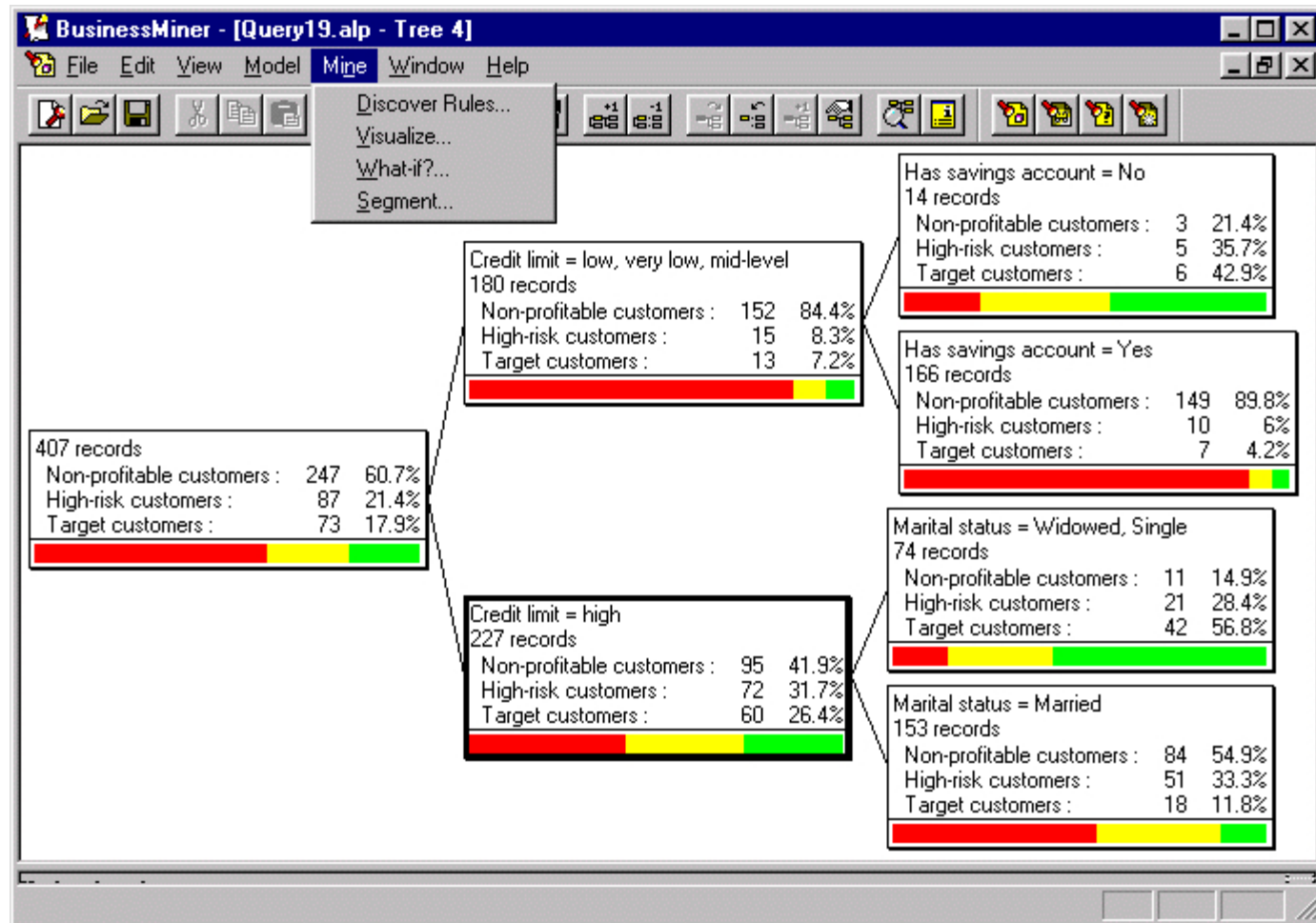
The bar chart shows the distribution of 'IncomeBracket' values. The y-axis represents the count of instances for each value. The bars are colored red and blue, with the red portion representing the count of instances for each value. The counts are: 681, 260, 390, 248, 380, 509, 265, and 267.

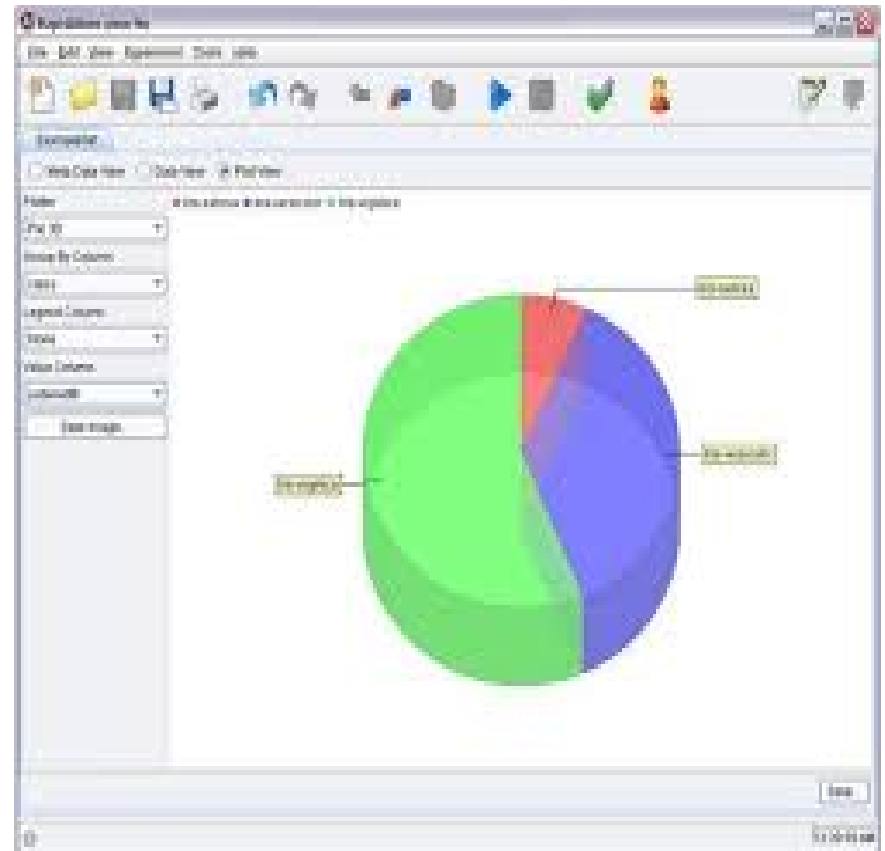
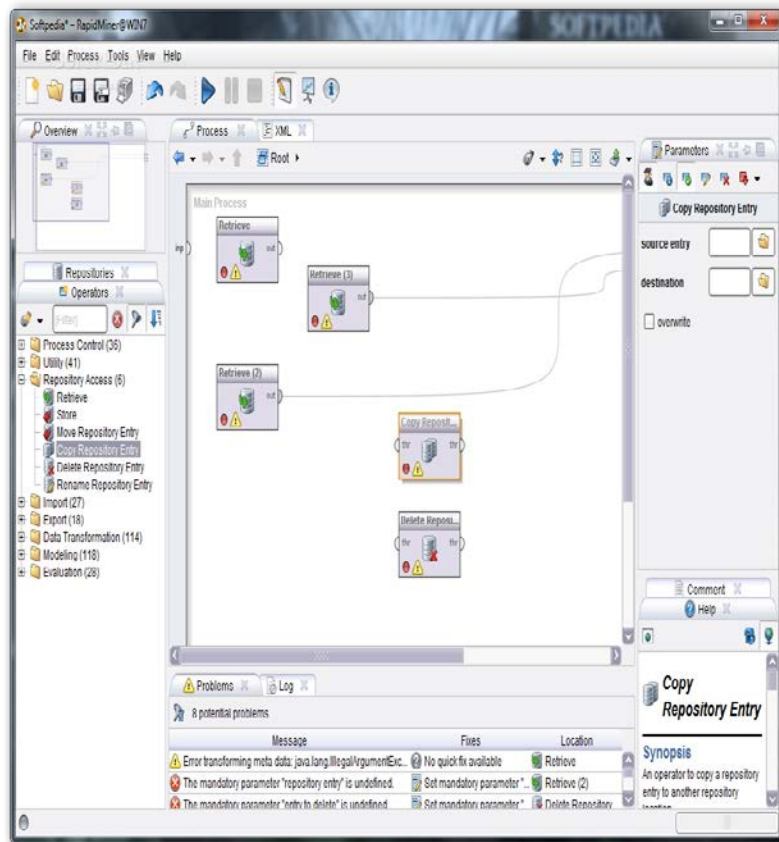
Weka GUI Choo...  
Waikato Environment for Knowledge Analysis  
(c) 1999 - 2003  
University of Waikato  
New Zealand

GUI  
Simple CLI Explorer  
Experimenter KnowledgeFlow

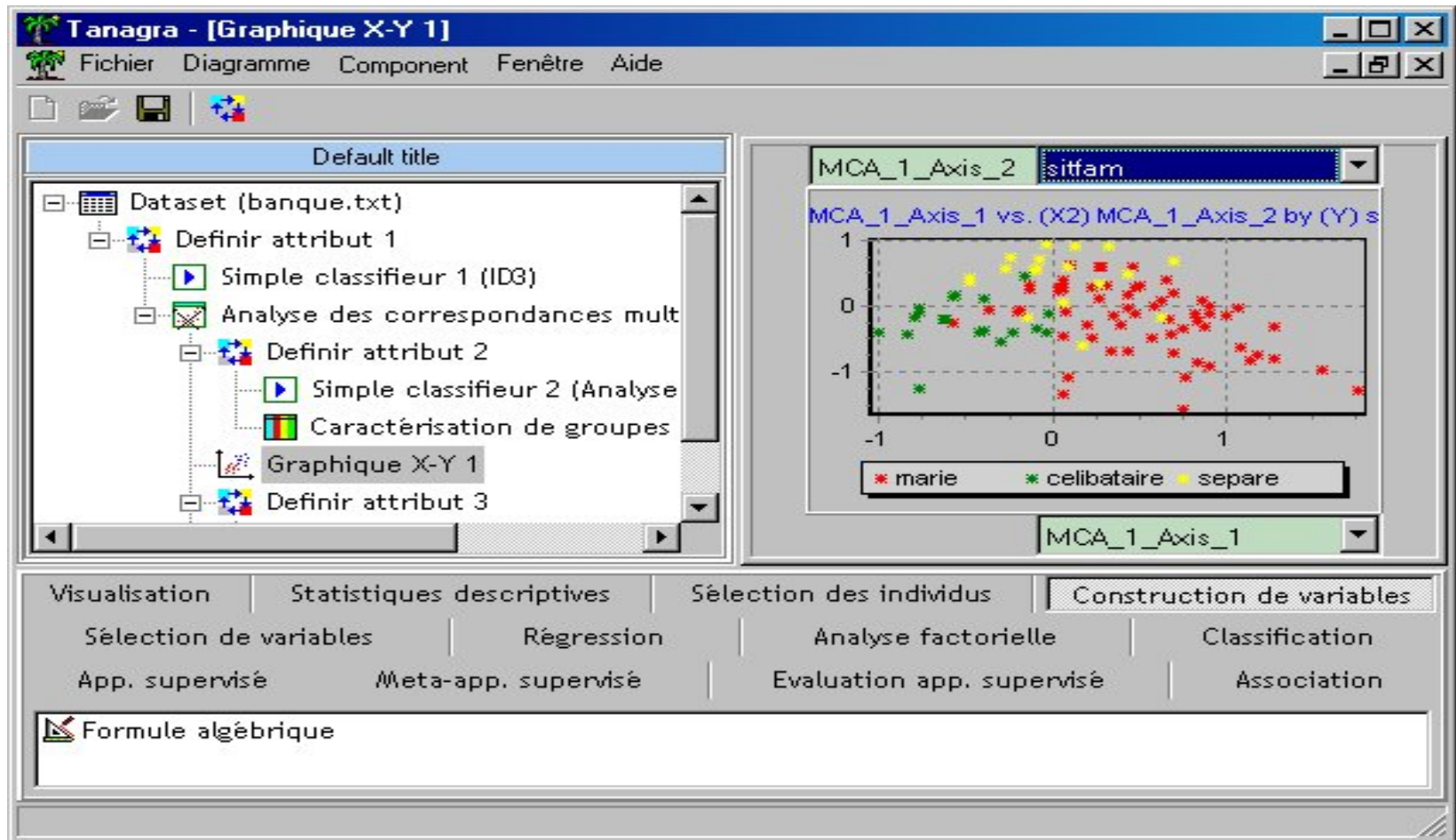








# Tanagra



<http://eric.univ-lyon2.fr/~ricco/tanagra>

# A Comparison



Procedure	R-Programming	RapidMiner	Weka	Orange
Partitioning of dataset into training and testing sets.	Pass (but limited partitioning methods)	Pass (but limited partitioning methods)	Pass (but limited partitioning methods)	Pass (but limited partitioning methods)
Descriptor scaling	Pass	Pass	Fail (cannot save parameters for scaling to apply to future datasets)	Fail (no scaling methods)
Descriptor selection	Fail (no wrapper methods)	Pass	Pass (but is not part of KnowledgeFlow)	Fail (no wrapper methods)
Parameter optimization of machine learning/statistical methods	Fail (not automatic)	Pass	Fail (not automatic)	Fail (not automatic)
Model validation using cross-validation and/or independent validation set	Pass (but limited error measurement methods)	Pass	Pass (but cannot save model so have to rebuild model for every future dataset)	Pass (but cannot save model so have to rebuild model for every future dataset)

## Some tools

- Professional
- A set of tools and models (Jack knife)
- Oriented to decision makers (not data analysts)
- Need for graphical interfaces, simplicity, readability, automated setting of parameters
- Supporting the scale factors (large data)
- Many progresses keep to make...

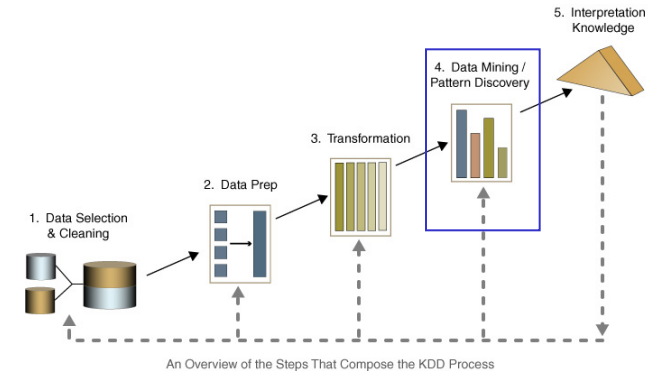
# Some examples

# Relational Pattern Mining Generalities

# Data Mining

Knowledge Discovery in Data (KDD)

Data -> Model -> “Knowledge” -> Decision maker



- Structured data :
  - Transactional: itemsets, association rules (AR), generalized AR
  - Sequences : Sequences, rule sequences
  - Spatiotemporal : spatial patterns, spatiotemporal patterns, trajectories, rules
- Semi-structured data (XML based) :
  - Multimedia : Multimedia patterns (zones, scenes, transitions, repetitions, ...)
  - Graphs : frequent subgraphs, dynamics (social networks)
  - Knowledge : Knowledge Mining (sem Nets, Conceptual graphs, OWL ontology, ...)
- Weakly structured data :
  - Texts : textual patterns (concepts), rules (Text Mining)

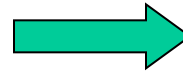


# Transactional Patterns :

## Itemsets, association rules (AR), generalized AR

Transactions

TID	Produce
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CEREAL
4	MILK, BREAD, SUGAR
5	MILK, CEREAL
6	BREAD, CEREAL
7	MILK, CEREAL
8	MILK, BREAD, CEREAL, EGGS
9	MILK, BREAD, CEREAL



Frequent Itemsets:

Milk, Bread (4)  
Bread, Cereal (3)  
Milk, Bread, Cereal (2)  
...



Rules:

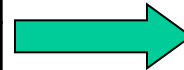
Milk => Bread (66%)

...

# Sequential Patterns

- Sequential pattern mining for **multiple** data sequences (sequence mining)

Sequence ID	Purchase data record
1	<bread, cheese>
2	<(wheat, milk), bread, (berry, sausage)>
3	<(bread, pumpkin, sausage)>
4	<bread, cheese, sausage>
5	<cheese>



Frequent subsequences:

Bread, Cheese (2)  
Bread, Sausage (3)  
...

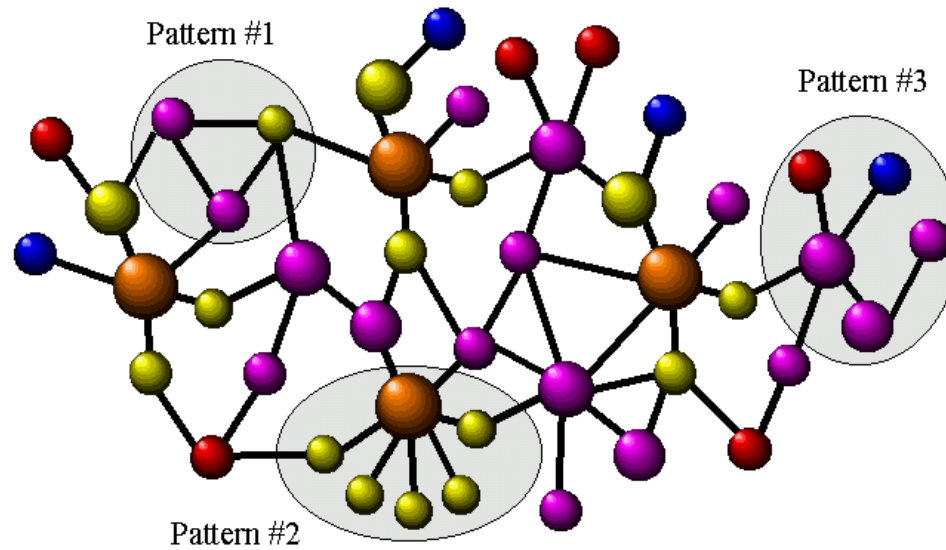
- Sequential pattern mining for a **single** data sequence (episode mining)

Data sequence
<S1 S2 S3 S4 S5 S6 S7 ... ... Sn>



Sequential Rules:  
Bread > Cheese (66%)

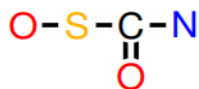
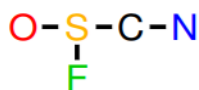
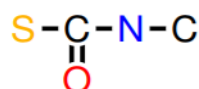
# Graph Patterns



# Graph Patterns

## Frequent subgraph mining in molecules

example molecules  
(graph database)



frequent molecular fragments ( $s_{\min} = 2$ )

\* (empty graph)

3

S

3

O

3

C

3

N

3

O-S

2

S-C

3

C=O

2

C-N

3

O-S-C

2

S-C-N

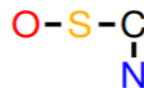
3

S-C=O

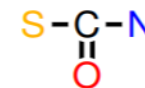
2

N-C=O

2



2



2

The numbers  
below the subgraphs  
state their support.

From Christian Borgelt - <http://www.borgelt.net/>

# Références

- **KD Nuggets:** [www.kdnuggets.com](http://www.kdnuggets.com)
- **The Data Mine:** [www.the-data-mine.com](http://www.the-data-mine.com)

## *Data mining :*

- Han J. and Kamber M., “Data Mining – Concepts and Techniques”, Morgan Kaufmann, 2001.
- Pang-Ning Tan. "Introduction to data mining", Pearson Addison Wesley, 2005, Isbn: 0321321367
- M. J. Zaki and W. Meira, Jr. "Data Mining and Analysis: Fundamental Concepts and Algorithms", Cambridge University Press, 2013. (version électronique pdf disponible)
- I. H. Witten et E. Frank "Data Mining: Practical Machine Learning Tools and Techniques", Elsevier, 2005. ISBN (version électronique pdf disponible)

## *Machine learning / statistics :*

- Friedman, J., Hastie, T., & Tibshirani, R. (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Springer Series in Statistics.
- Bishop, C. M., & Nasrabadi, N. M. (2006). "Pattern recognition and machine learning" (Vol. 1, p. 740). New York: springer.
- Murphy, K. P. (2012). Machine learning: a probabilistic perspective. The MIT Press.
- Saporta G., "Probabilités, analyse des données et statistique", Technip, 1990.

## *Others :*

- Dunja Mladenic, Nada Lavrac, Marko Bohanec, Steve Moyle. Data Mining and Decision Support. Kluwer. 2003 ISBN 1-4020-7388-7
- Willi Klösgen, Jan Zytkow. Handbook of Data Mining and Knowledge Discovery. Oxford University Press. 2002 ISBN: 0-19-511831-6
- Fayyad U.M., Piatetsky-Shapiro G. et Smyth P. (eds) : Advances in knowledge discovery and data mining. AAAI Press. 1996.