Présentation des concepts de l'Analyse des données symboliques

E. Diday
Paris–Dauphine University

OUTLINE

- **INTRODUCTION**
- PART 1: BUILDING SYMBOLIC DATA FROM STANDARD OR COMPLEX DATA
- PART 2: THE SYMBOLIC DATA ANALYSIS PARADIGM
- PART 3: ILLUSTRATIVE EXAMPLES
- PART 4: ANALYSIS PROCESS
- PART 5: FUTURE and CONCLUSION

PART 1

BUILDING SYMBOLIC DATA FROM STANDARD OR COMPLEX DATA

FROM DATA BASES INTENDED FOR MANAGEMENT TO DATA SCIENCE TOOLS

New tools are needed to transform big and complex

- data bases intended for management
- to data bases usable for Data Science tools.

➢ Symbolic Data are among these tools.

STANDARD UNITS

Standard units are described by single-valued data given by:

- **Numerical variables (as age, weight,..) or**
- **Categorical variables express groups (Nationality, team name,…)!**

Units						
Players	age	height	weight	Nationality	Club	Team
Player 1						
Messi						Barça
Ronaldo						Real Madrid

From units to groups of units

UNITS	**GROUPS**
Players	Teams
Words	Documents
Inhabitant	Regions
Cells	Tumors
Patients	Treatment
Pixels	Images
Specimen	Species

GROUPS AS NEW UNITS

Such groups allow a summary of the population and often

- Represent the real units of interest.

- They cannot be described by single valued data as there is variability between the units contained in each group.

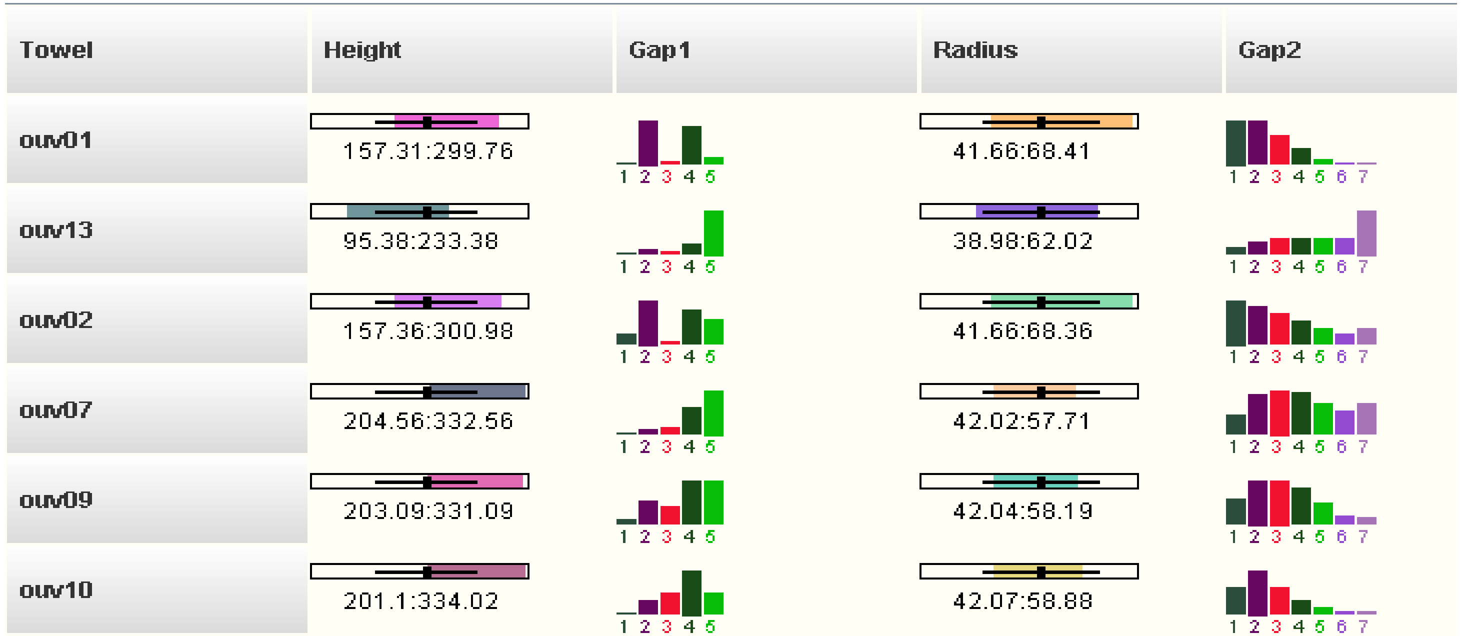
SYMBOLIC DATA

- **The SDA domain** is born by considering classes (i.e. groups) of a given population to be units instead of standard statistical units.
- It is an answer to the challenge of

➢ Complex Data

➢ Big Data

An example of groups described by symbols



Symbolic variables are random variables of random variable value

Towel	Height	Gap1	Radius	Gap2
ouv01	157.31:299.76	1 2 3 4 5	41.66:68.41	1 2 3 4 5 6 7
ouv13	95.38:233.38	1 2 3 4 5	38.98:62.02	1 2 3 4 5 6 7
ouv02	157.36:300.98	1 2 3 4 5	41.66:68.36	1 2 3 4 5 6 7
ouv07	204.56:332.56	1 2 3 4 5	42.02:57.71	1 2 3 4 5 6 7
ouv09	203.09:331.09	1 2 3 4 5	42.04:58.19	1 2 3 4 5 6 7
ouv10	201.1:334.02	1 2 3 4 5	42.07:58.88	1 2 3 4 5 6 7

SYMBOLIC DATA EXPRESS VARIABILITY INSIDE CLASSES OF INDIVIDUALS

TEAM OF THE MONDIAL	WEIGHT	NATIONALITY	NB OF GOALS
BARSA	[75 , 89]	{French}	{0.8 (0), 0.2 (1)}
MANCHESTER	[80, 95]	{Fr, Alg, Arg }	{0.1 (0), 0.3 (1), …}
PARIS-ST G.	[76, 95]	{Fr, Tun }	{0.4 (0), 0.2 (1), …}
DORTMUND	[70, 85]	{Fr, Engl, Arg }	{0.2 (0), 0.5 (1), …}

Here the variation (of weight, nationality, …) concerns the players of each team.

Therefore each cell can contain:

An interval, a sequence of categorical values, a sequence of weighted values as a barchart, a distribution, …or numbers.

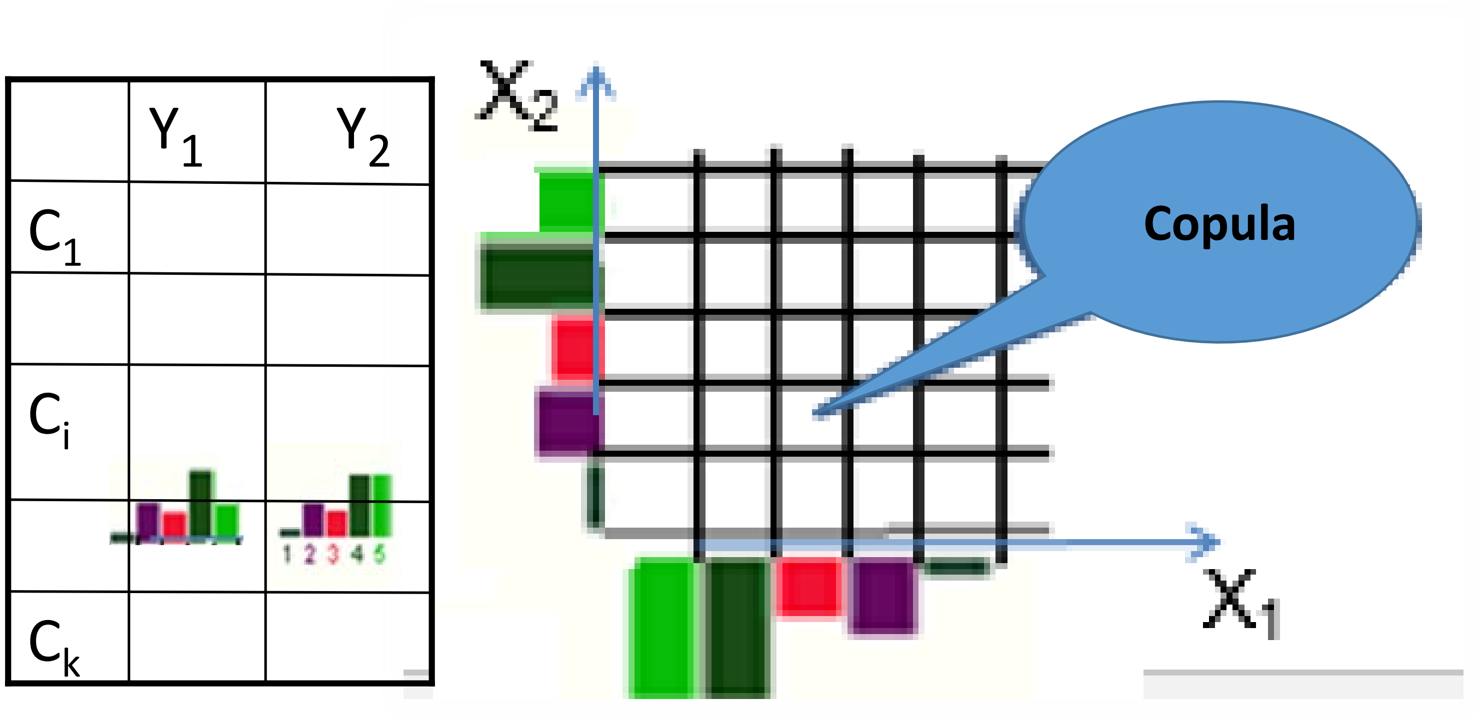
THIS NEW KIND OF VARIABLES ARE CALLED « SYMBOLIC » BECAUSE THEY ARE NOT PURELY NUMERICAL IN ORDER TO EXPRESS THE INTERNAL VARIATION INSIDE EACH CLASS.

NEEDED TOOLS TO DESCRIBE GROUPS

- Aggregation of the units contained in the groups are needed
- They leads to new statistical units described by:
- symbols (intervals, distributions, list of words or categories, etc.)
- single-valued data are not suitable because they cannot incorporate the additional information on data structure (ie unpaired variables) and internal variability available in symbolic data.

Bi-plot of histogram variables

- The joint probability can be inferred by a copula model



In case of independency the probability of the joint is the product which is a case of copula which allows many other models as Franck, etc.
IN case of BIG DATA it is a very economical way to get the joint.

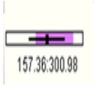
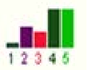
From lower level of individual observation to higher level observation of classes: higher level models are needed

Table 1

Individual	X_1		X_j	
ind_1				
Messi			$\mathbf{X_{ij}}$	
ind_n				

A number
(age of Messi)

Table 2

Team	X'_1		X'_j	
C_1				
C_i				
C_k				

A symbolic data
(age of Messi team)

X_j is a standard random numerical variable

X'_j is a random variable with histogram value

- Question: if the law of Xj is given what is the law of X'_j ? (Dirichlet models useful).

Some SDA principle

Four principles guide this paper in conformity with the Data Science framework.

- **First,** new tools are needed to transform huge data bases intended for management to data bases usable for Data Science tools. This transformation leads to the construction of new statistical units described by aggregated data in term of symbols as single-valued data are not suitable because they cannot incorporate the additional information on data structure available in symbolic data.
- **Second,** we work on the symbolic data as they are given in data bases and not as we wish that they be given. For example, if the data contains intervals we work on them even if the within interval uniformity is statistically not satisfactory. Moreover, by considering Min Max intervals we can obtain useful knowledge, complementary to the one given without the uniformity assumption. Hence considering that the Min Max or interquartile and the like intervals are false hypothesis has no sense in modern Data Science where the aim is to extract useful knowledge from the data and not only to infer models (even if inferring models like in standard statistics, can for sure give complementary knowledge).
- **Third,** by using marginal description of classes by vectors of univariate symbols rather than joint symbolic description by multivariate symbols as 99% of the users would say that a joint distribution describing a class leads often to sparse data with too much low or 0 values and so has a poor explanatory power in comparison with marginal distributions describing the same class. Nevertheless, a compromise can be obtained by considering joints instead of marginal between the more dependent variables.
- **Fourth,** we say that the description of a class is much more explanatory when it is described by symbolic variables (closer from the natural language of the users), than by its usual analytical multidimensional description. This principle leads to a way to compare clustering methods by the explanatory power of the clusters that they produce.

A Basic SDA formalism in case of categorical variables

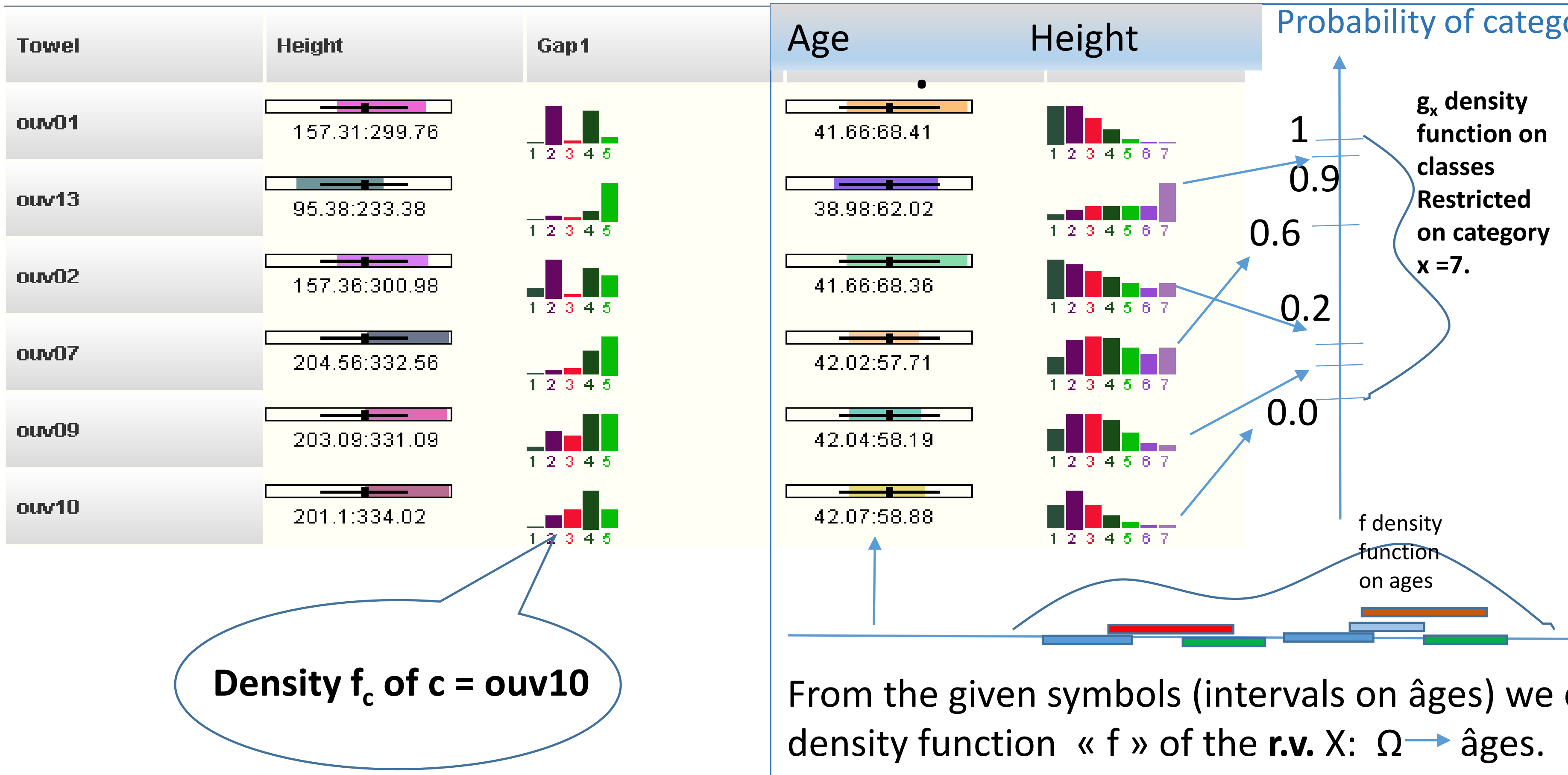
Three random variables C, X, A defined on the ground population Ω.

C a class variable:
Ω → P such that C(w) = c where c is a class of a given partition P.
X a categorical value variable:
Ω → M such that X(w) = x Є M a set of categories M.
A an aggregation function which associates to a class c a symbol

s = A(c)

Examples:
s = [min, max] , s = interquartile interval
s = cumulative distribution, s = a barchart etc.

The three basic density function as symbols: f, f_c, g_x



From the symbols (barcharts) we can Induce a density function « g » of the variable « Height » from the set of classes to the set of histograms

From the given symbols (intervals on âges) we can induce a density function « f » of the **r.v.** X: $\Omega \rightarrow$ âges.
Also, from the given histograms we can find a density « f » of the rv X' on heights.

From complex data to symbolic data

What are Complex Data?

Complex data are any data set which cannot be considered as a standard data table.

This case happen when variables are unpaired as they are not defined on the same unit.

Example 1 of complex data plants in towers of nuclear power plants

- Towers of nuclear power plants are described by
- **Table 1) Observations: Cracks .**

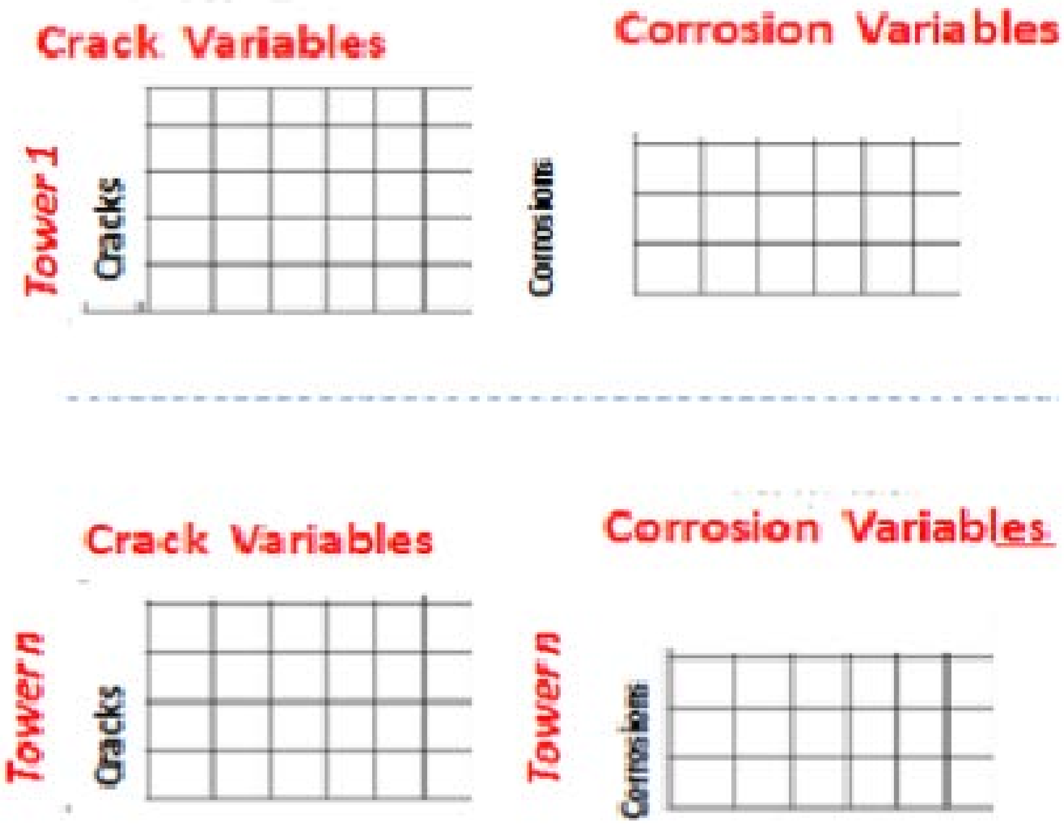
  Variables: Cracks description.
- **Table 2) Observations: corrosions.**

  Variables: corrosion description .
- **Table 3) Observations: vertices of a grid.**

  Variables: Gap depression from the ground.

Example of Complex Data

Craks



Example 2 of complex data in Official Statistics

The objects are regions described by

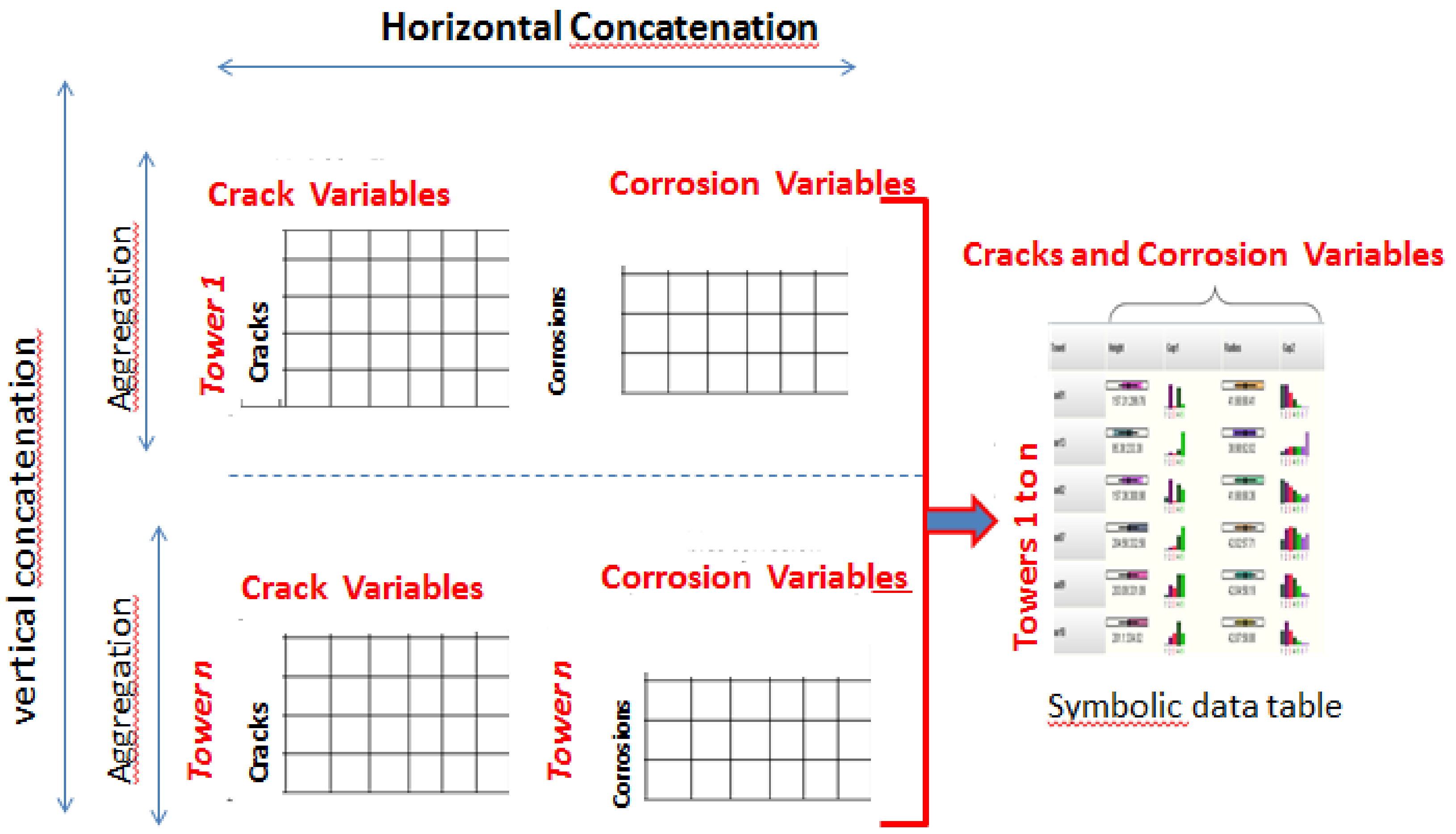
- **Table 1) Observations: hospitals .**

  Variables: size, patients number, ..
- **Table 2) Observations: schools.**

  Variables: schools description.
- **Table 3) Observations: inhabitants.**

  Variables: Socio demographic ,..

FROM COMPLEX DATA TO SYMBOLIC DATA by AGREGATION AND CONCATANATION PROCESS



SDA: The Three major questions

Building Symbolic data Table

From Complex Data

- **Agregation,**
- **Concatanation**

Managing Symbolic data table

- Sorting rows
- Sorting variables
- **Extending SQL queries** to Symbolic Data

Analysing Symbolic data tables

•Extending to symbolic data:

- Standard Statistics
- Standard Data Mining,
- Standard Learning Machine.

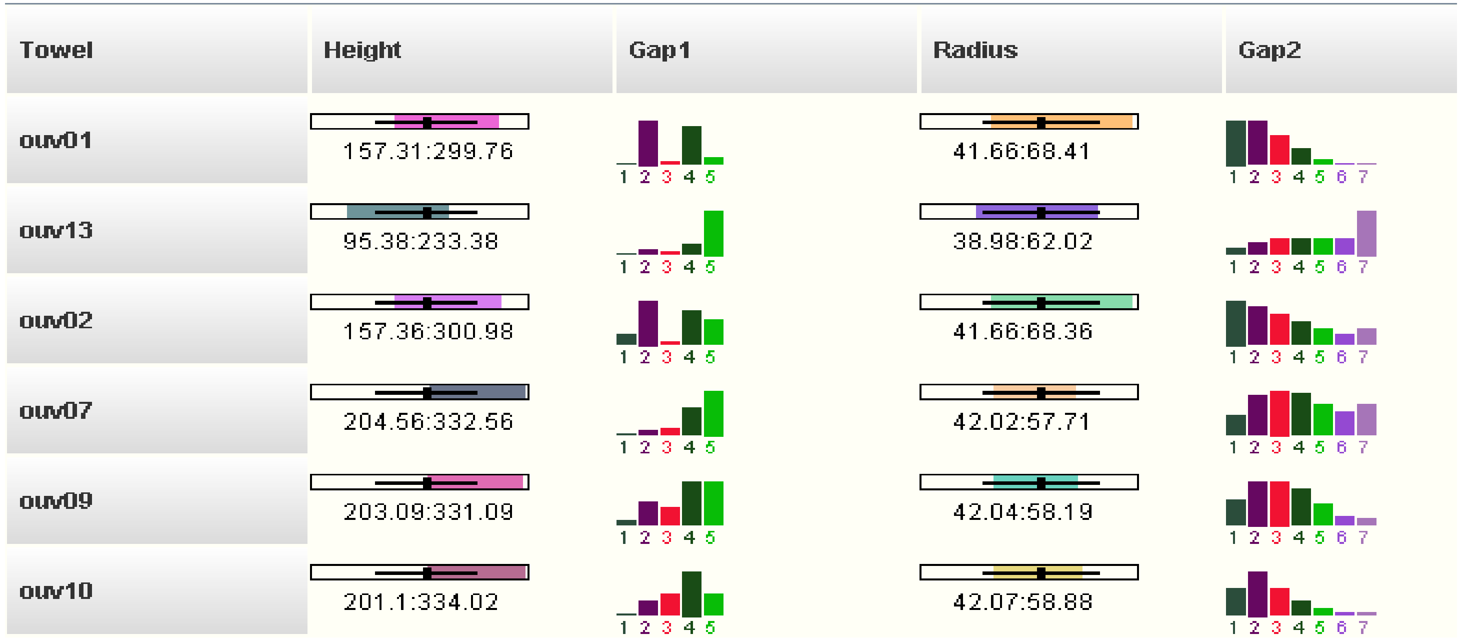
Agregation tools in SDA

- How to categorize the ground variables in order to maximize the explanatory power of the Symbolic Data Table?

Find the discretisation which:

- **First:** discriminates as well as possible these classes.
- **Second**: Maximizes the correlation between the bins.
- **Third** : Minimizes the entropy of the rows.

Coding in order to improve the explanatory power of a the classes



Towel	Height	Gap1	Radius	Gap2
ouv01	157.31:299.76	1 2 3 4 5	41.66:68.41	1 2 3 4 5 6 7
ouv13	95.38:233.38	1 2 3 4 5	38.98:62.02	1 2 3 4 5 6 7
ouv02	157.36:300.98	1 2 3 4 5	41.66:68.36	1 2 3 4 5 6 7
ouv07	204.56:332.56	1 2 3 4 5	42.02:57.71	1 2 3 4 5 6 7
ouv09	203.09:331.09	1 2 3 4 5	42.04:58.19	1 2 3 4 5 6 7
ouv10	201.1:334.02	1 2 3 4 5	42.07:58.88	1 2 3 4 5 6 7

- Maximize: distances between rows,
- Maximize: correlations between column
- Minimize: entropy in each cell

Disavantage of the aggregation process

- Lost of correlation between the ground variables

SDA answer:

- adding correlations is possible at the class level.
- In case of unpaired variables correlation has no meaning as units are different.
- The Joint instead of marginal lead to sparse data tables
- The marginal have better explanatory power.

SOME ADVANTAGES of SYMBOLIC DATA

- Work at the needed level of generality without loosing variability.
- Reduce simple or Complex and/or BIG DATA.
- Reduce number of observations and number of variables.
- Reduce missing data.
- Ability to extract explanatory knowledge and decision from Complex /Big data in opposition with black box decision methods.
- Solve confidentiality (classes are not confidential as individuals).
- Facilitate interpretation of results: decision trees, factorial analysis new graphic kinds.
- Increase explanatory power of the methods by remaining with the user language of the initialy given variables.

PART 2

SYMBOLIC DATA ANALYSIS

What is Symbolic Data Analysis?

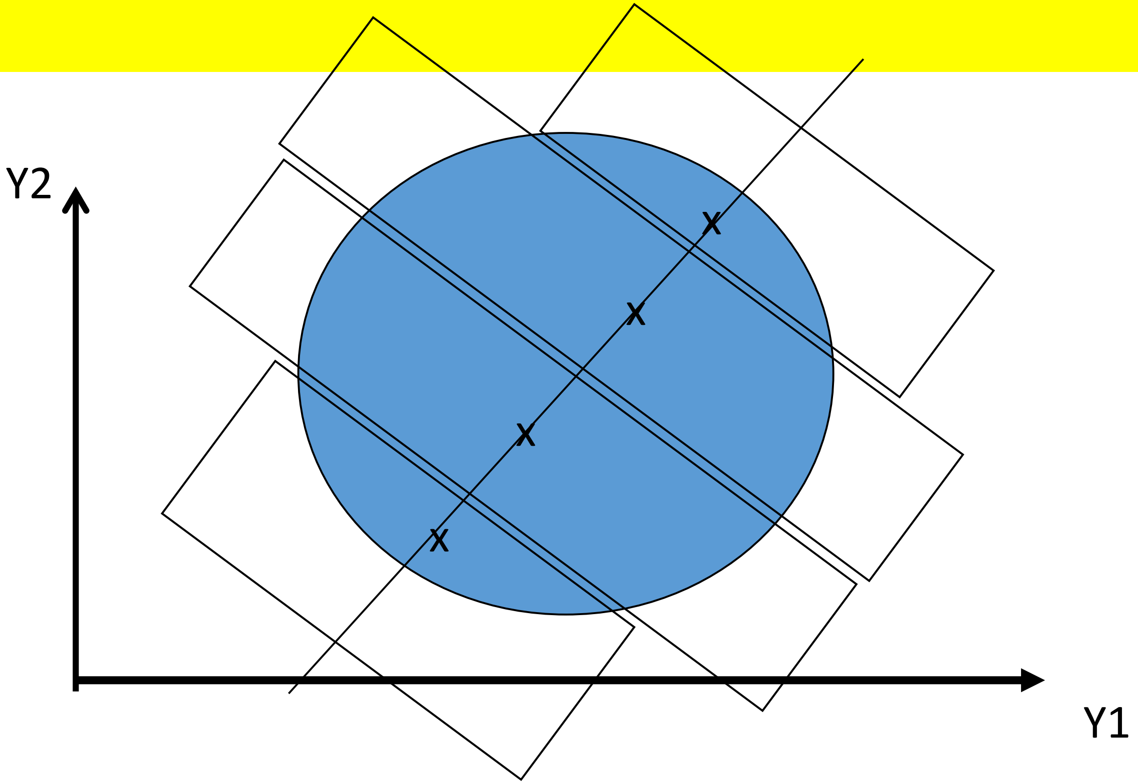
It is an emerging area of Data Science based on :

- ➢aggregating individual level data into group-based summarized by symbols.
- ➢developing complementary Data Science tools enhancing our understanding of the data.
- ➢ increasing the explanatory power of machine learning in the case of:
- ➢ standard, large and complex datasets.

SYMBOLIC DATA ANALYSIS TOOLS HAVE BEEN DEVELOPPED

- Graphical visualisation of Symbolic Data
- Correlation, Mean, Mean Square, distribution of a symbolic variables.
- Dissimilarities between symbolic descriptions
- Clustering of symbolic descriptions
- S-Kohonen Mappings
- S-Decision Trees
- S-Principal Component Analysis
- S-Discriminant Factorial Analysis
- S-Regression
- Etc... Much remains to be done

From standard observations to classes, the correlation is not the same!



- **Observations data are uniformly distributed in the circle:**
- **no correlation between Y1 and Y2 for intial observations data.**
- **A correlation appears between the two variables for the centers of a given partition in 4 classes.**

WHY SYMBOLIC DATA CANNOT BE REDUCED TO A CLASSICAL STANDARD DATA TABLE?

Symbolic Data Table

Players category	Weight	Size	Nationality
Very good	[80, 95]	[1.70, 1.95]	{0.7 Eur, 0.3 Afr}

Transformation in classical data

Players category	Weight Min	Weight Max	Size Min	Size Max	Eur	Afr
Very good	80	95	1.70	1.95	0. 7	0.3

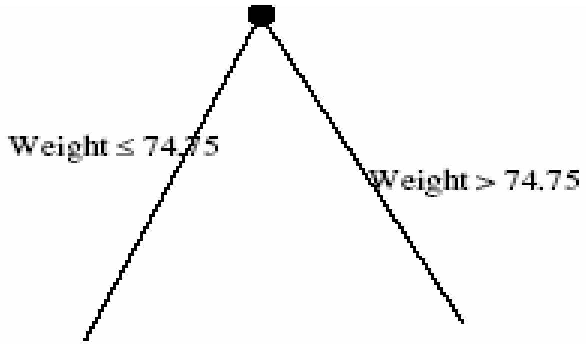
Concern:

The initial variables are lost and the variabilité is lost!

Divisive Clustering   or Decision tree

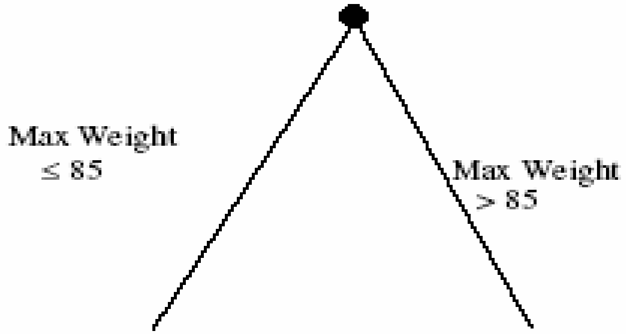
Symbolic Analysis

Weight



Classical Analysis

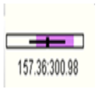
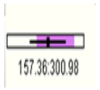
Max Weight



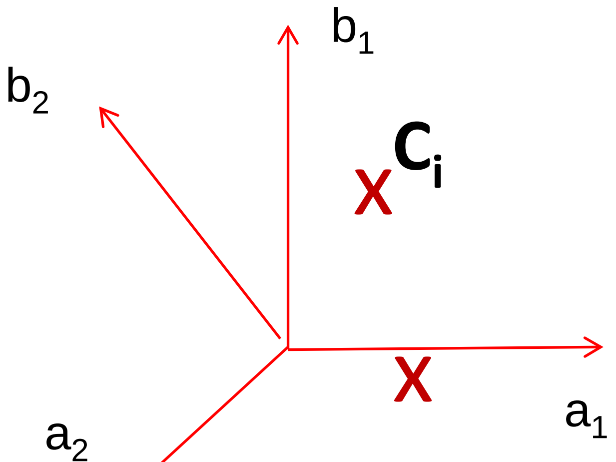
VIZUALIZATION OF SYMBOLIC DATA

	a_1	b_1	a_2	b_2
C_1				
C_i	a_{1i}	b_{1i}	a_{2i}	b_{2i}
C_k				

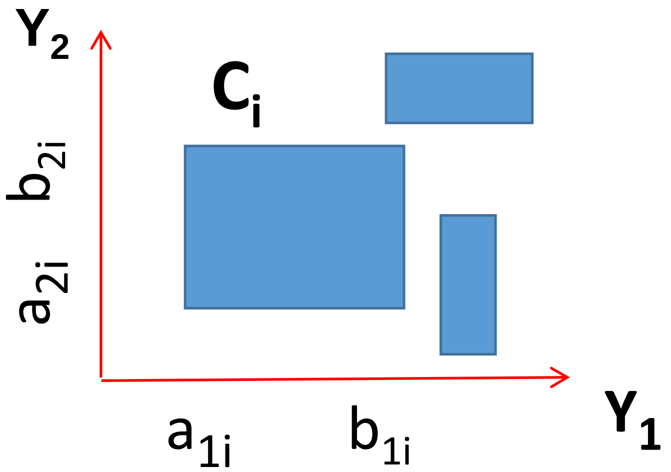
$(Y_1(C_i), Y_2(C_i)) = ([a_{1i}, b_{1i}], ([a_{2i}, b_{2i}])$

	Y_1	Y_2
C_1		
C_i		
C_k		

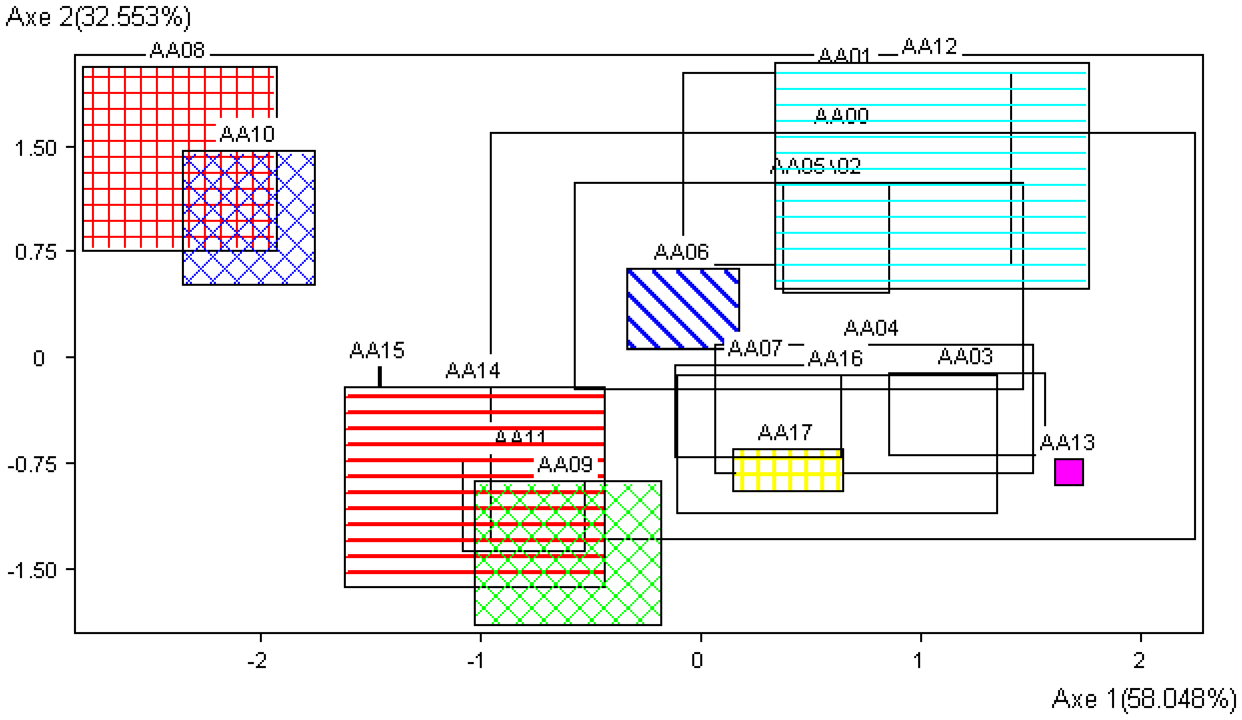
Numerical representation of interval variables



Bi-plot of interval variables

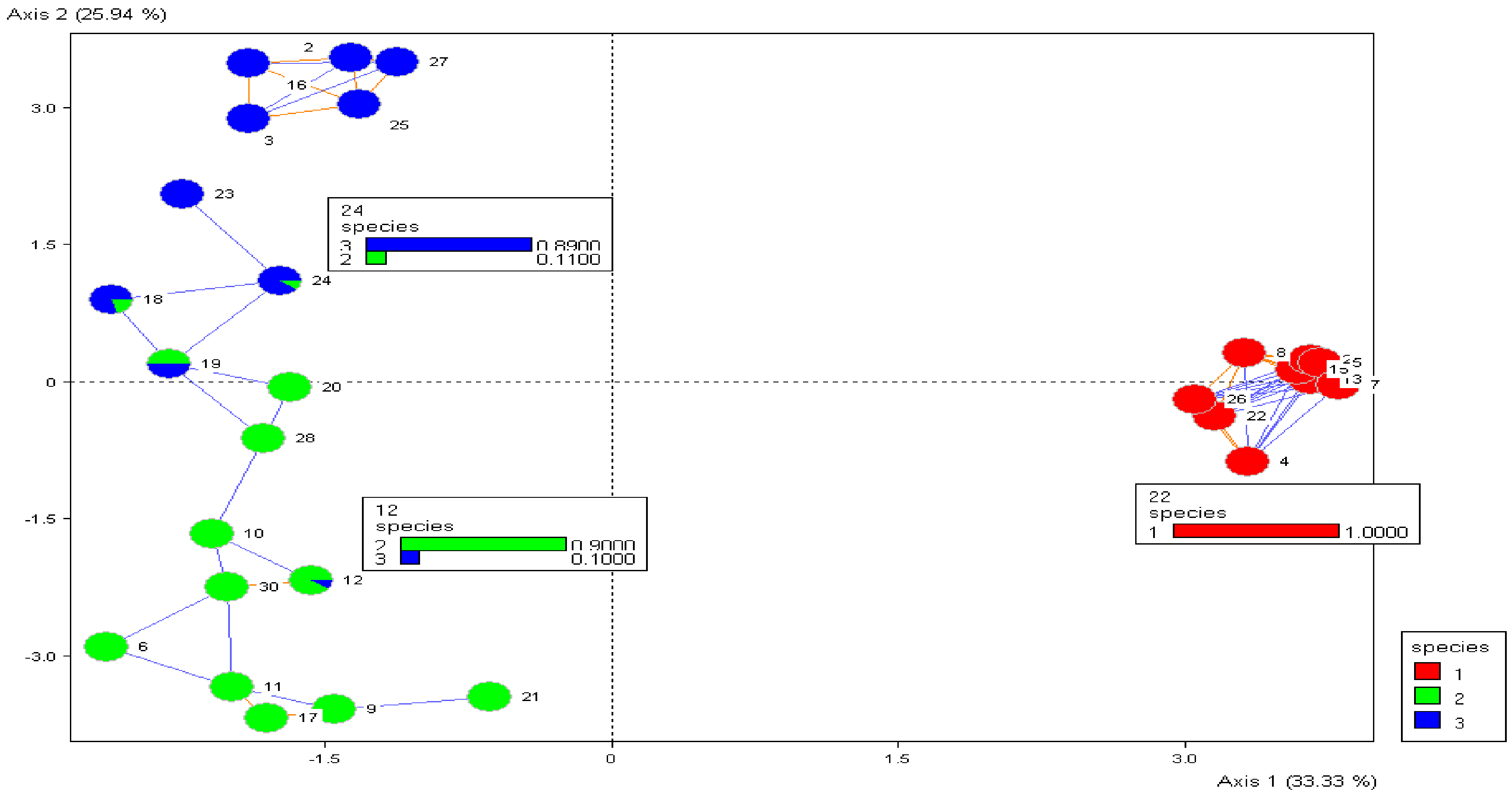


PCA OF INTERVAL DATA



Each class is represented by a rectangle which express its variability
A standard PCA would represent each class by a point.

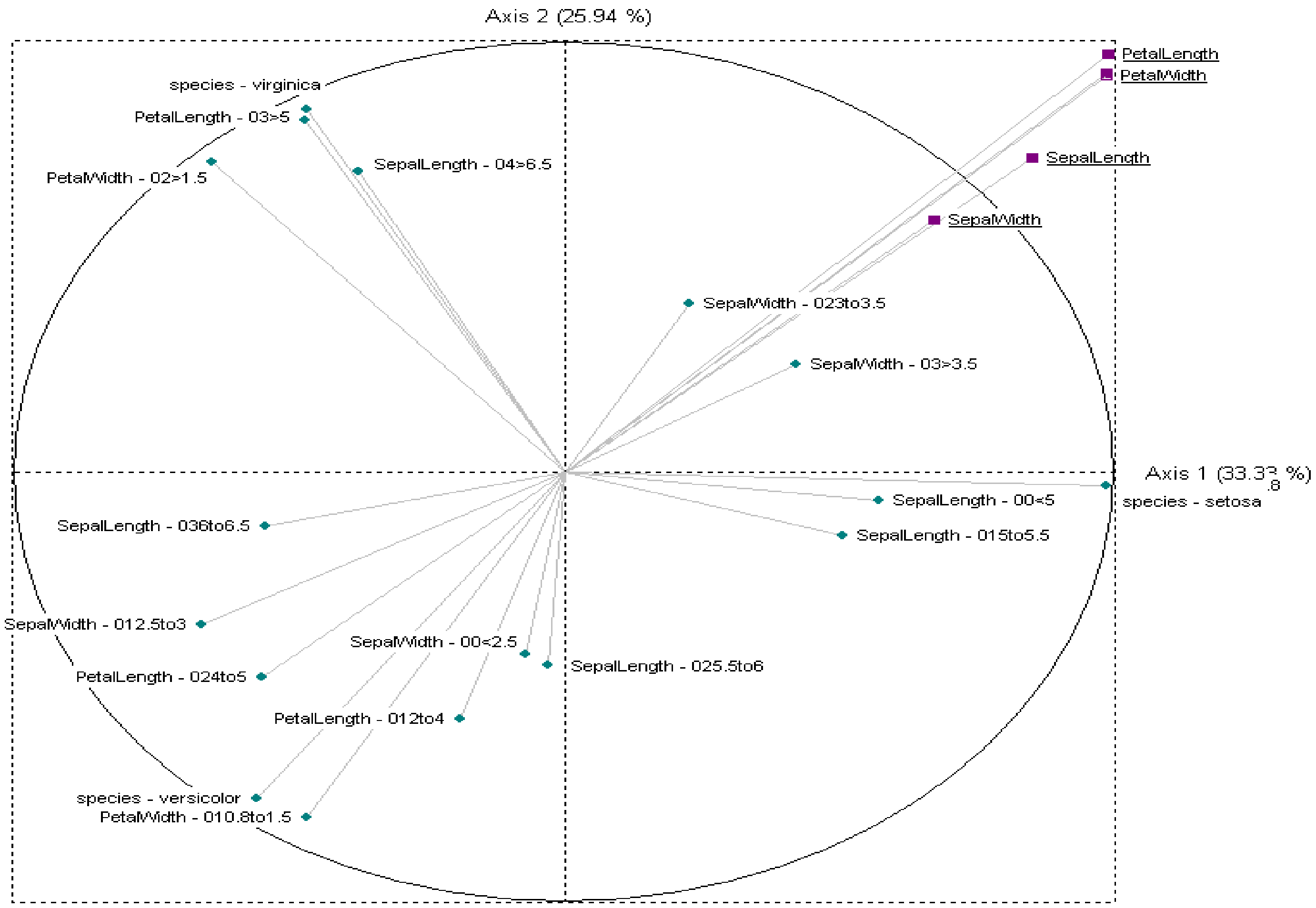
PCA and NETWORK OF BAR CHART DATA of 30 Iris Fisher Data Clusters*



Any symbolic variable (set of bins variables) can be projected. Here the species variable.

* SYROKKO Company afonso@syrokko.com

The Symbolic Variables contributions are inside the smallest hyper cube containing the correlation sphere of the bins



PART 3: Illustrative examples

Objects and Symbolic Data

Any object can be described by symbolic data when it varies in time and /or position or among its parts:

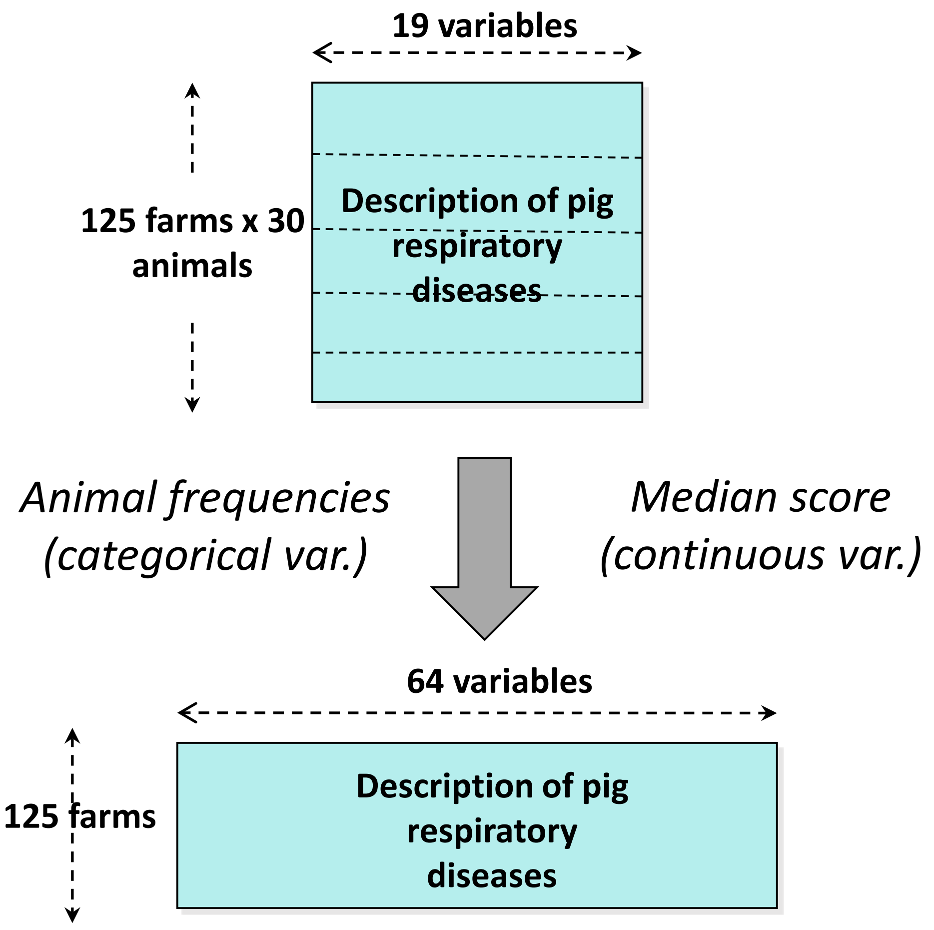
- **power point cooling towers**,
- **boats** of risk in a harbor,
- financial **stocks** behavior,
- **text** section of a book
- **web intrusions** in a company,
- **Cells** in an image or **images**

Application Domain

“Concordance” or “Discordance” “specificity”, “typicality” , Ranking of:

- **power point cooling towers**,
- **boats** of risk in a harbor,
- financial **stocks** behavior,
- **text** section of a book
- **web intrusions** in a company,
- **Cells** in an image or **images**

HIERARCHICAL DATA*



Symbolic procedure

From numerical description of pigs to symbolic description of Farms

- Numerical variables

and

- Categorical variables

are transformed in Bar Chart of the frequencies based on 30 animals,

Or in interval value variables

*C. Fablet, S. Bougeard (AFSSA)

Step 1: Symbolic Description of Farms*

* SYROKKO Company afonso@syrokko.com

Telephone calls text mining in order to discover “themes” without using semantic

INITIAL DATA: 2 814 446 rows

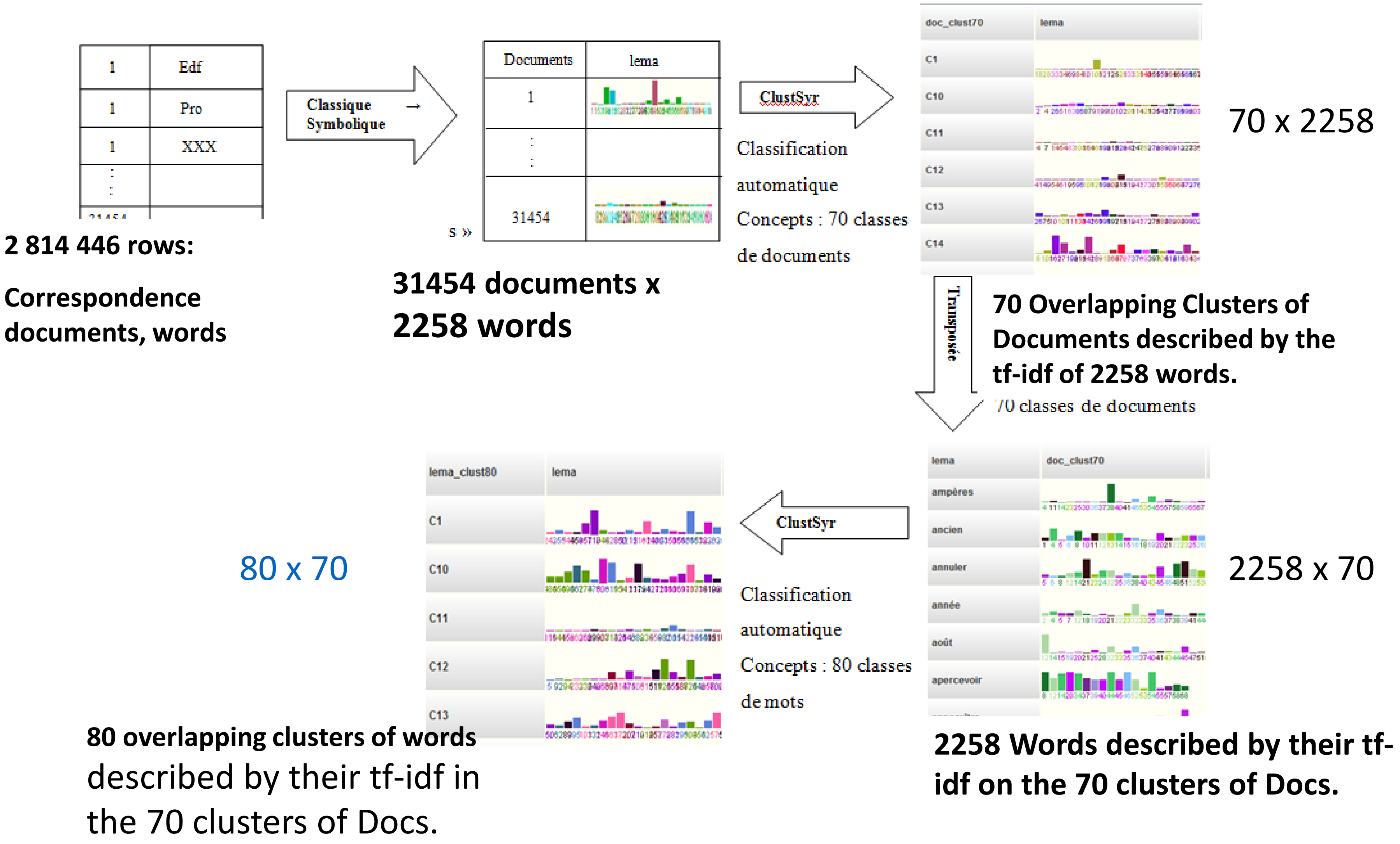
Documents	Words
Doc1	bonjour
Doc1	oui
Doc1	monsieur
………	
Doc2	panne
……	

Correspondence between documents and words.

Each calling session is called a document.
We start after lemmatisation with a table of

- 31454 documents
- 2258 words

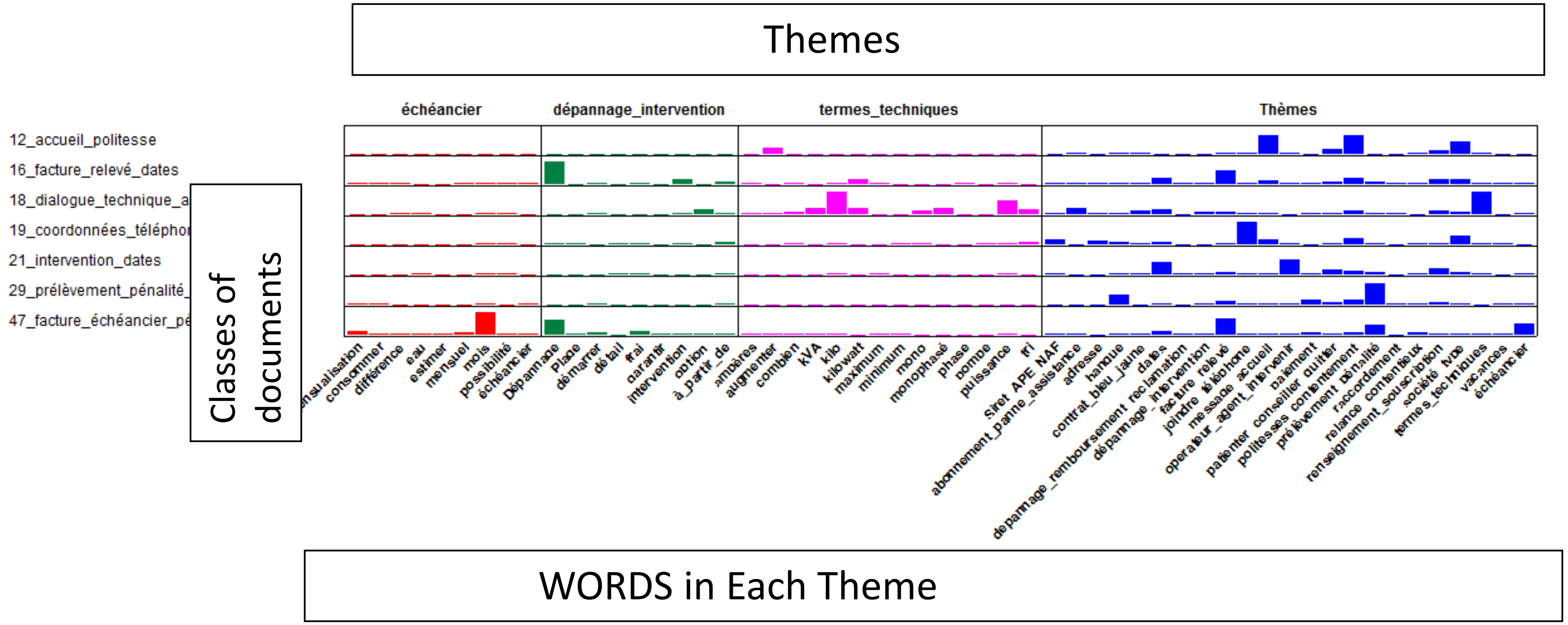
First Steps:building overlapping clusters of documents and words: CLUSTSYR



Next step:

STATSYR

Each cluster of documents is described by the 80 clusters of words called “themes”



GRAPHICAL REPRESENTATION by NETSYR from SYR software

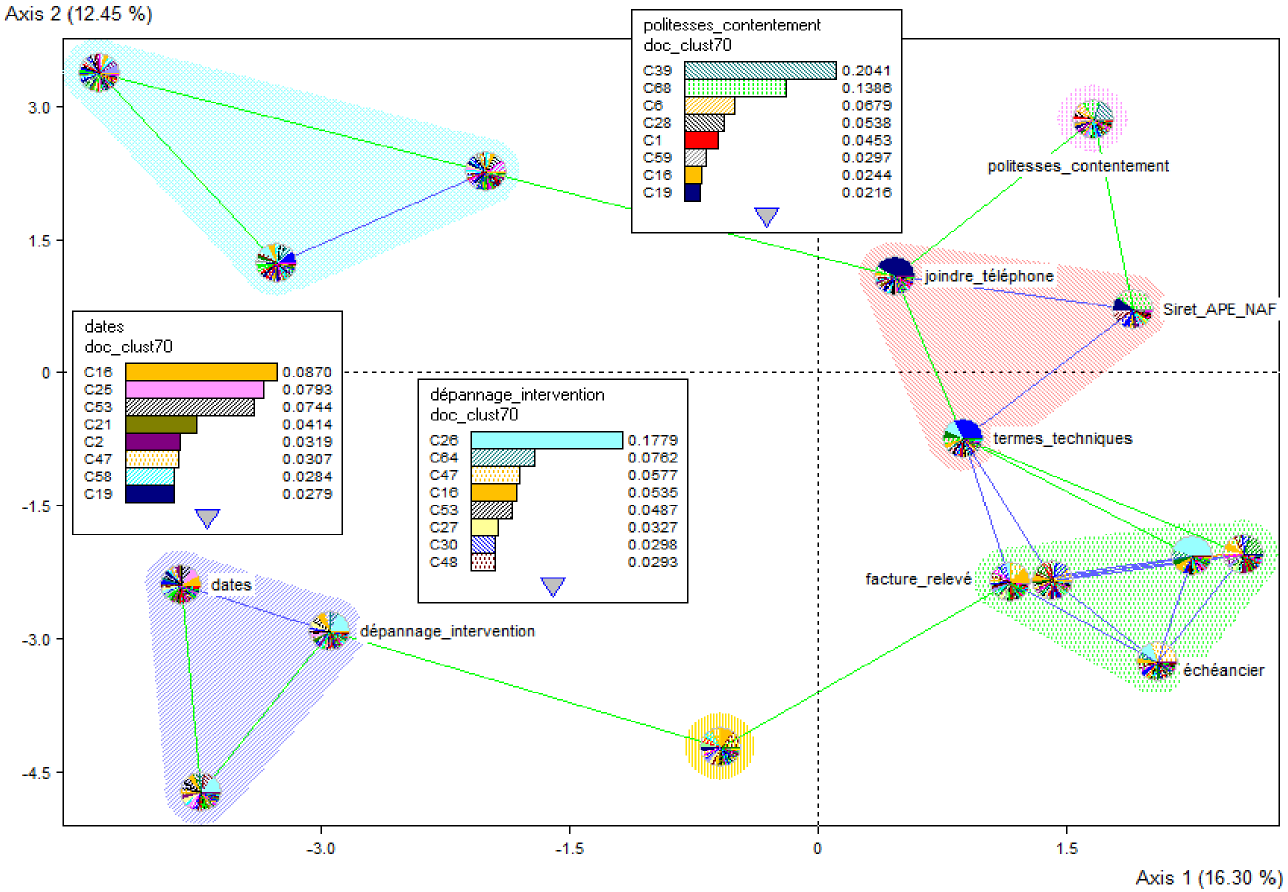
GRAPHICAL REPRESENTATION of themes ,
document classes, by Pie Charts
And their Bar chart description.

Overlapping Clusters

SOCIAL NEWORK Based on dissimilarities

ANNOTATION :
of Themes and Document classes

Moving, Zooming…



We obtain finally a clear representation of the main themes , their classes and their links : “failures”, “budget”,”addresses”, “vacation” etc..

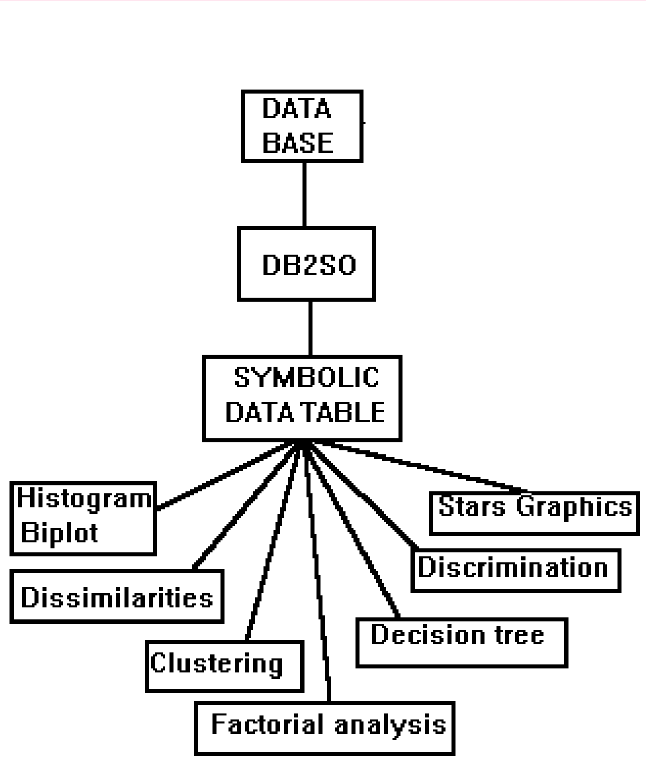
PART 4

ANALYSIS PROCESS

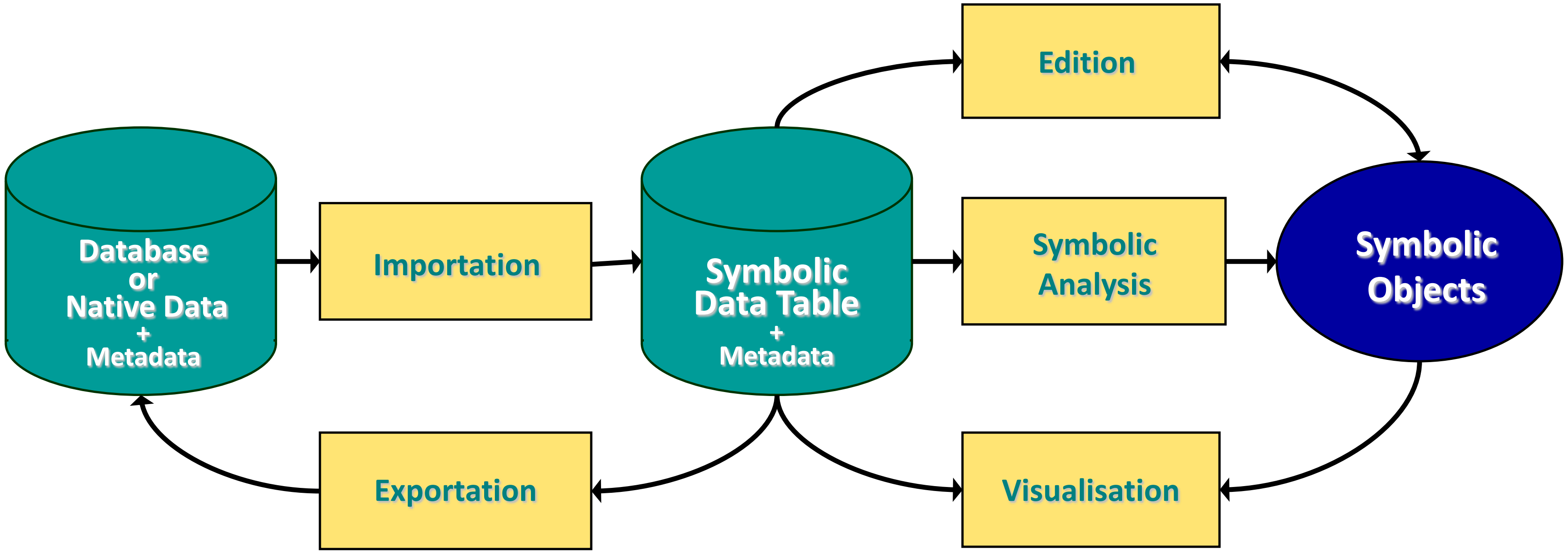
The most popular Softwares

- SODAS
- SYR
- RSDA

AN OVERVIEW ON THE SODAS SOFTWARE



SODAS Architecture



SODAS
Symbolic Data Analysis Software

- To build symbolic data from standard or complex data and analyze symbolic data

- SODAS: **academic free** package, through registration required and a code needed for installation

- Also the software can be download with explanation of the methods, user manuals and much Symbolic Data Bases at : http://www.ceremade.dauphine.fr/SODAS/

SODAS Web Site

SODAS

SODAS in english

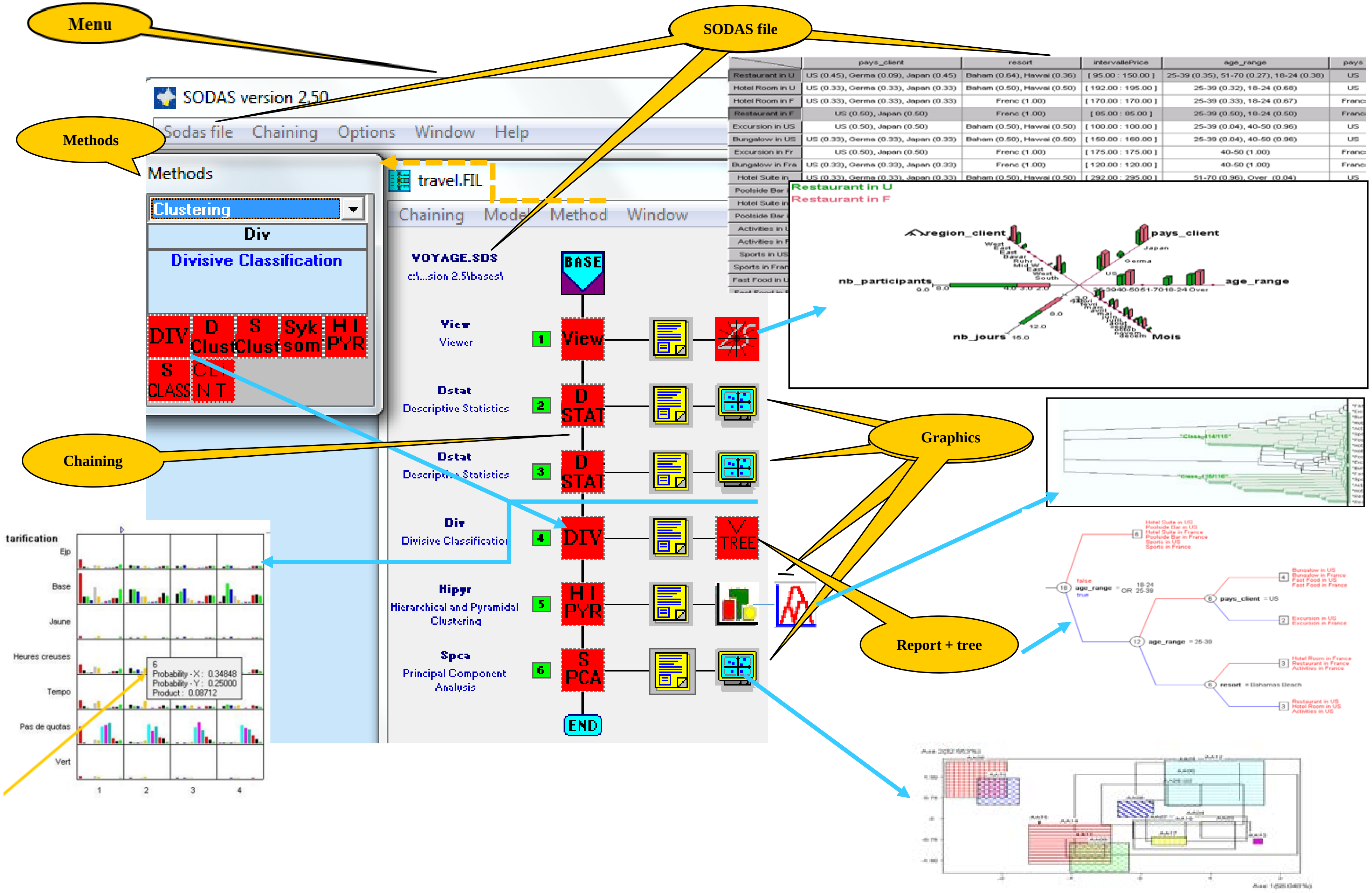
présenté par le laboratoire "LISE-CEREMADE"

UN LOGICIEL D'ANALYSE de DONNEES SYMBOLIQUES

Un nouvel outil pour le "DATA WAREHOUSE" et le "DATA MINING"

Manuel utilisateur	Présentation du projet et du logiciel ECOLES SODAS	Données et exemples de traitements avec SODAS
Téléchargement du logiciel SODAS Version 1.2 Version 2.5		Présentation des méthodes
Publications SODAS	Programmes MAJ et compatibles SODAS	Participants SODAS

SODAS Chaining

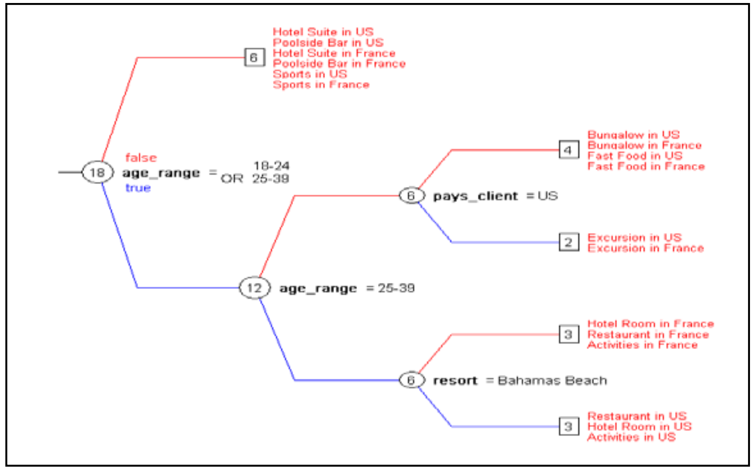


Some examples of SODAS methods

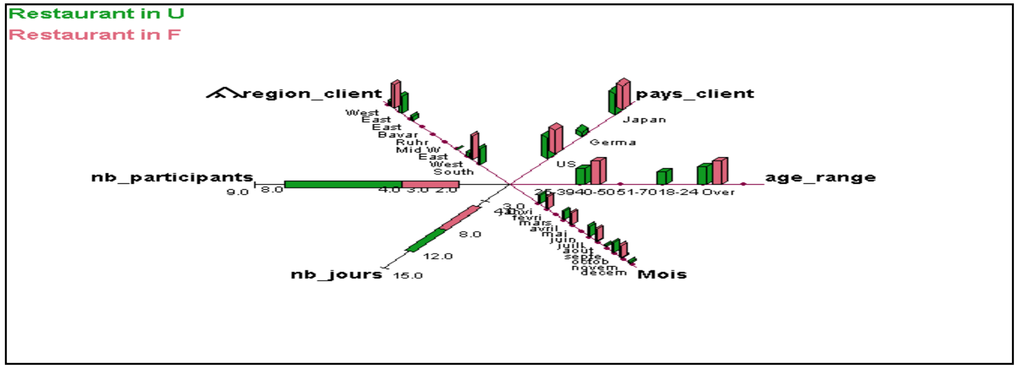
STAT Method: Descriptive Statistics for Symbolic Data

FDA Method: Factorial Discriminant Analysis for Symbolic Data

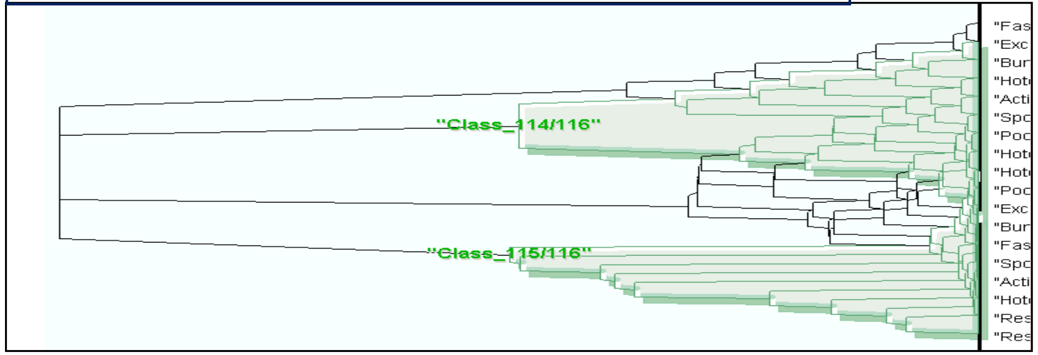
DIV Method: Divisive Clustering for Symbolic Data



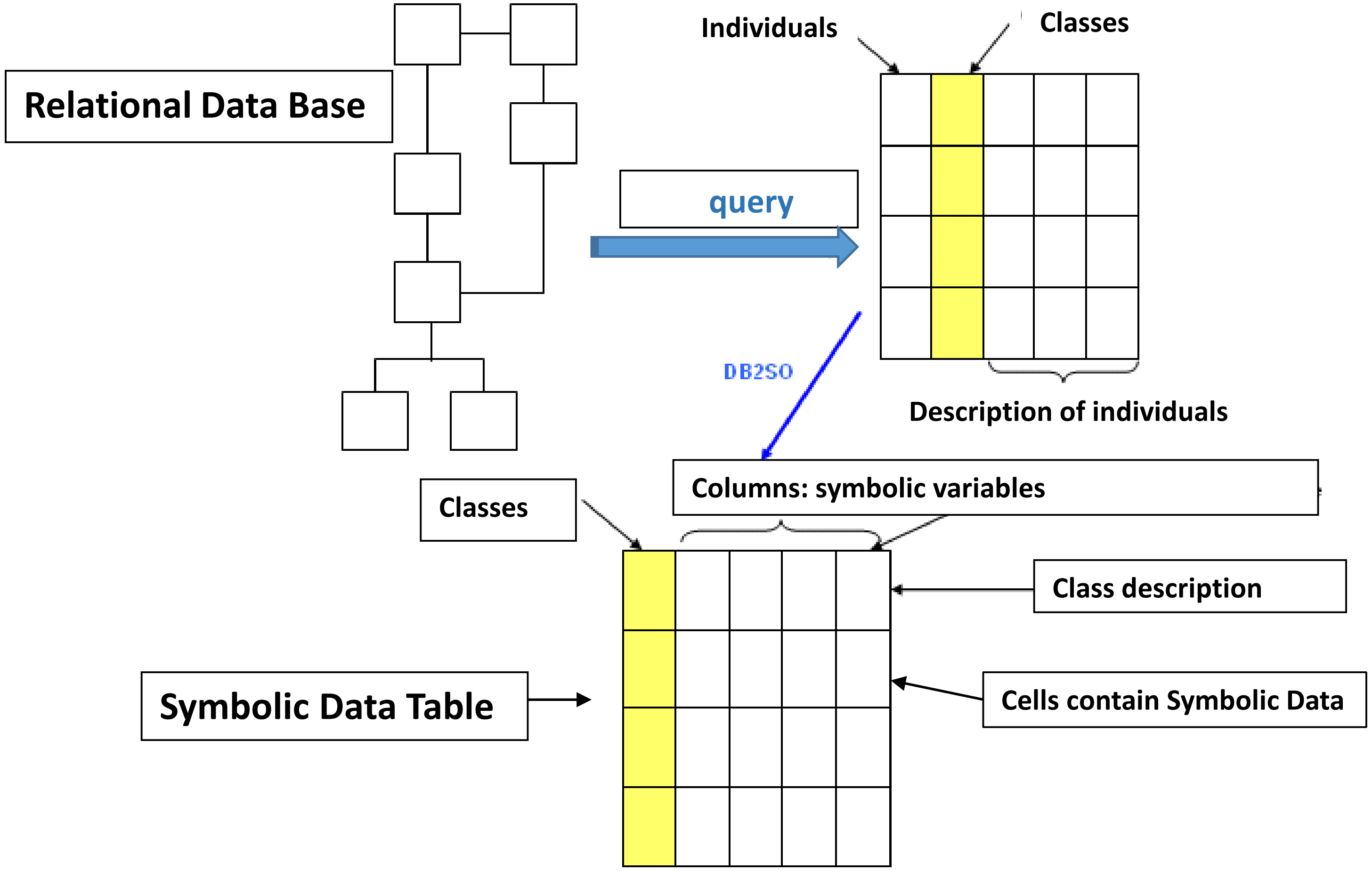
SOE-VIEW Method: Symbolic Objects Editor



HIPYR Method: Pyramidal Clustering on Symbolic Objects



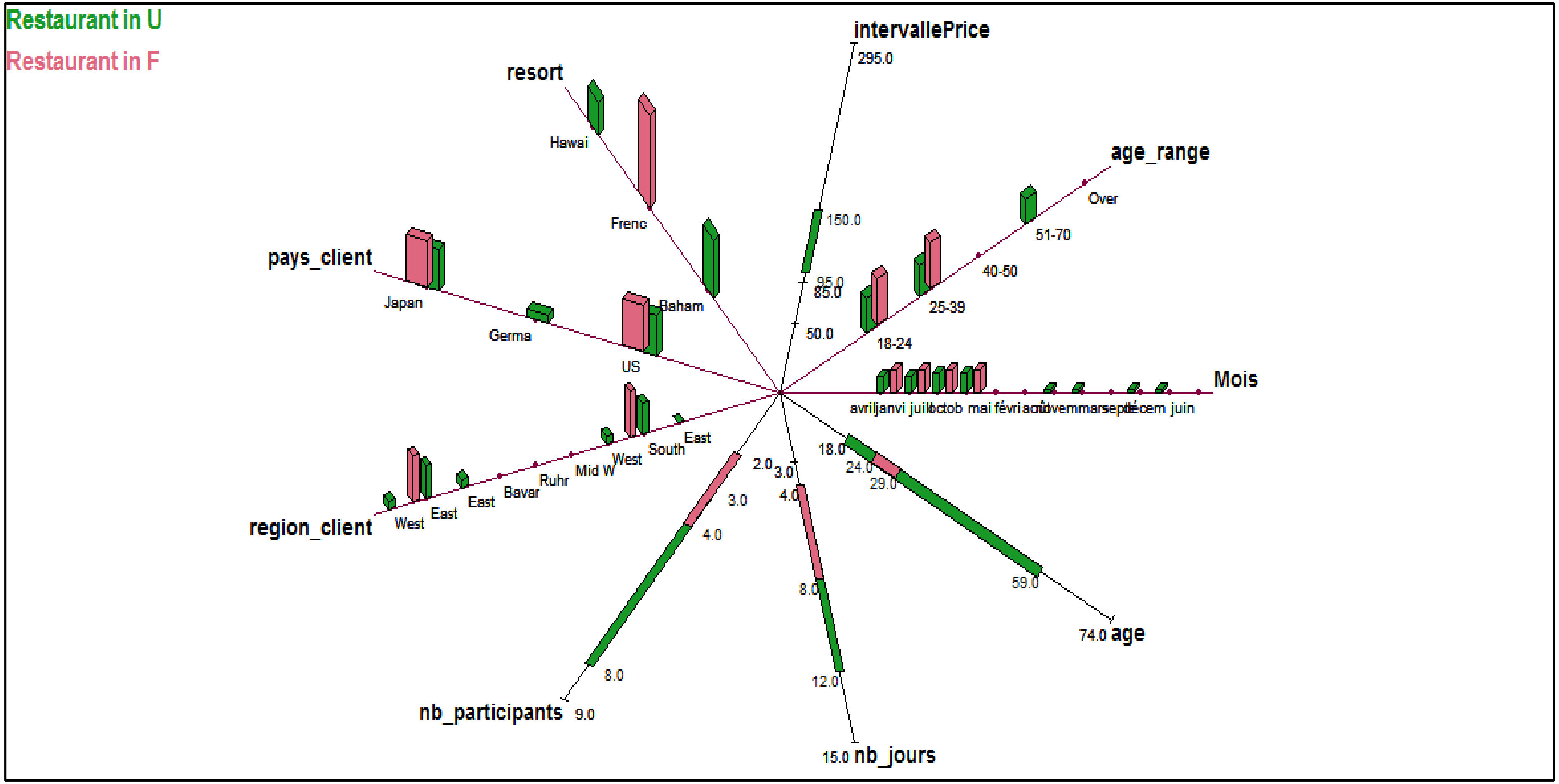
FROM DATA BASE TO SYMBOLIC DATA IN SODAS



SOE (Symbolic Object Editor) Method

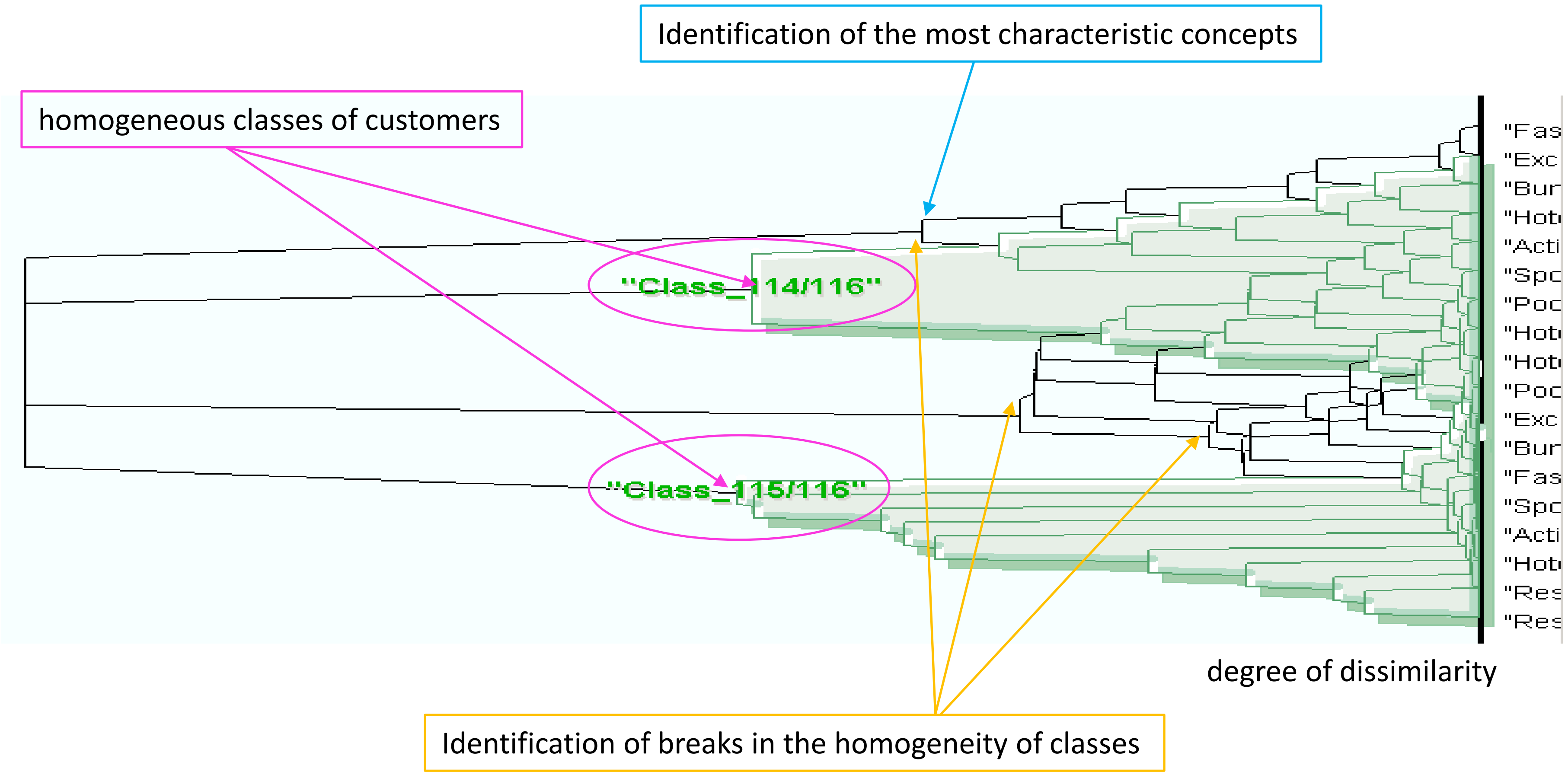
Zoom Star Representation 3D of SO

Comparison of several SO



HIPYR Method

Hierarchical and Pyramidal Clustering on Symbolic Objects

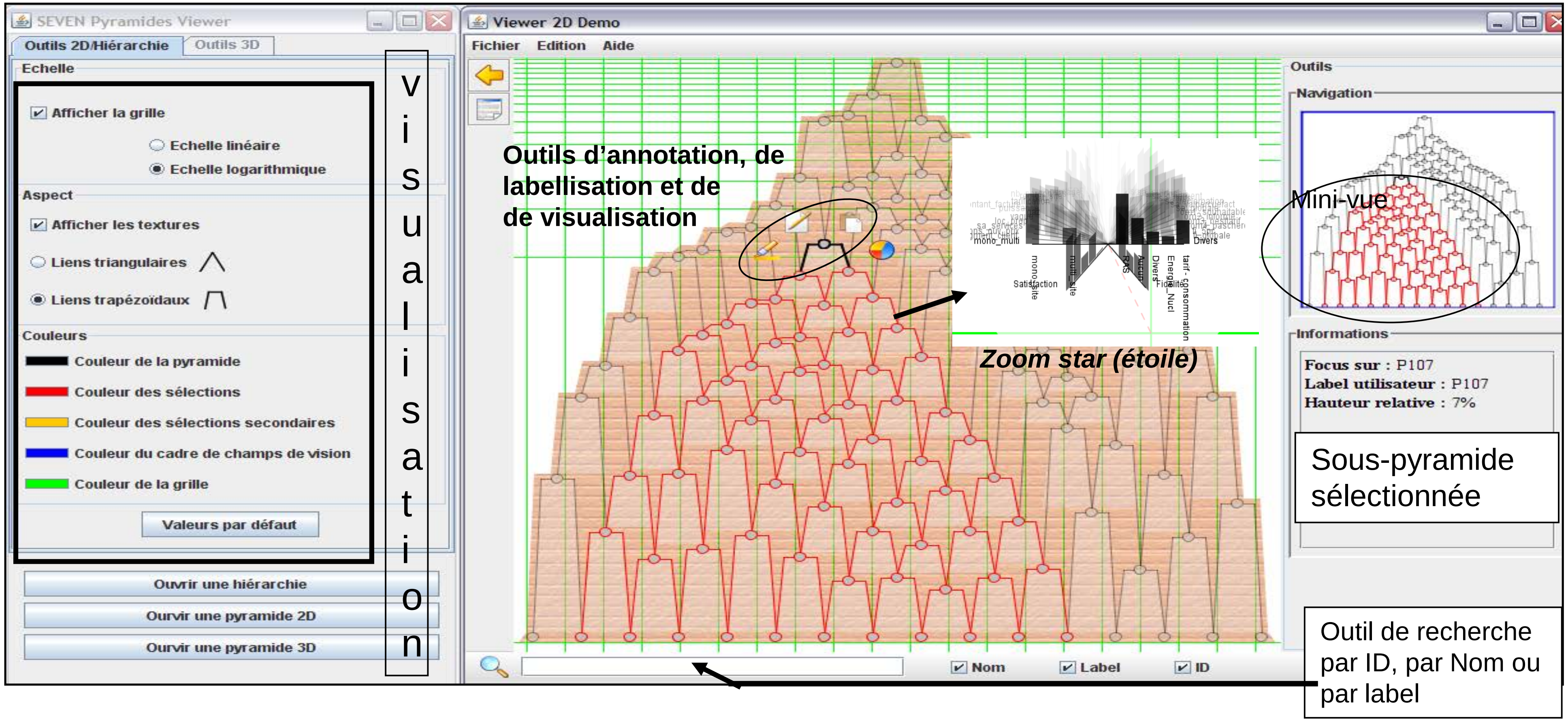


DAUPHINE UNIVERSITÉ PARIS | CEREMADE UMR CNRS 7534

HIPYR Method

Hierarchical and Pyramidal Clustering on Symbolic Objects

PYR2D: *New intuitive visualization interface*



SCLUST method

Symbolic Dynamic Clustering on Symbolic Objects

EDITION OPTIMAL PARTITION

Classe: 1, Cardinal: 4

Restaurant in US, Hotel Room in US, Bungalow in US, Activities in US

Classe: 2, Cardinal: 5

Excursion in US, Excursion in France, Bungalow in France, Food in US, Fast Food in France

Classe: 3, Cardinal: 3

Hotel Room in France, Restaurant in France, Activities in France

Classe: 4, Cardinal: 2

Sports in US, Sports in France

Classe: 5, Cardinal: 4

Hotel Suite in US, Poolside Bar in US, Hotel Suite in France, Poolside Bar in France

In summary

- **Symbolic data Tables generalize Standard Data Tables.**
- **SDA is a tool for extending standard methods of Statistics, Data Mining, Learning machine etc. to Complex and Big Data.**
- **Any Classical Data Analysis can be enhanced by a complementary Symbolic Data Analysis where the units are classes.**

PART 5

FUTURE and CONCLUSION

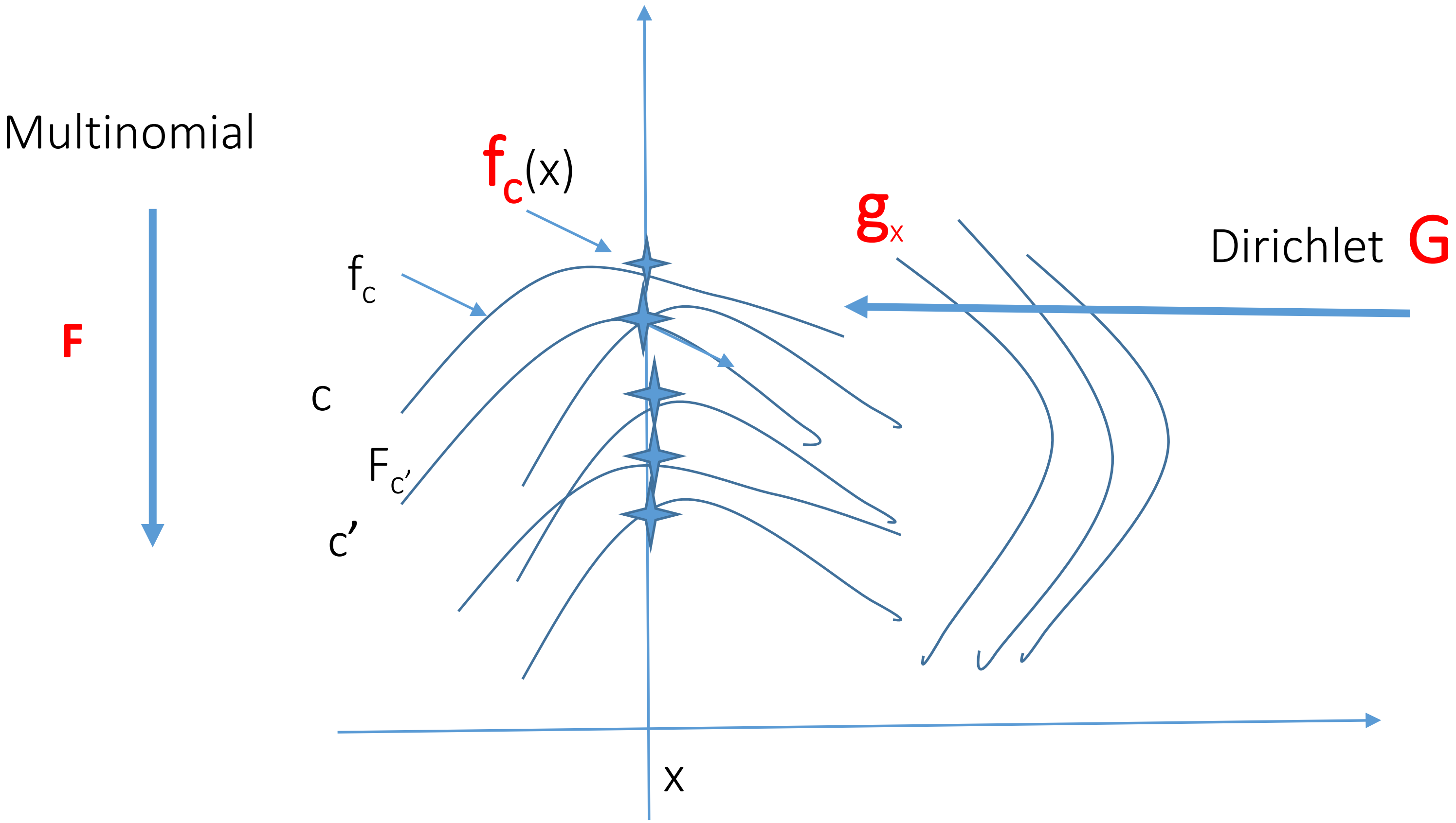
CONCLUSION

- SDA can enhance standard Data Mining and Statistics by complementary results.
- Symbolic data have to be build from given standard or complex data.
- Symbolic data cannot be reduced to standard data.
- Complex data can be simplified in symbolic data.
- Big Data bases can be reduced in symbolic data
- Standard software as EXCEL or Data Bases queries has to be extended to symbolic data tables

FUTURE

- \- To continue the symbolic extension of standard methods of statistic, machine learning and data science, data mining. Which become a case.
- Several explanatory criteria are on the way to be defined from which individuals, classes, symbolic variables and symbolic data tables can be placed in order from the more towards to the less characteristic.
- There are based on four random variables
- They lead to a symbolic extension of the Tf-Idf, the LDA (Latent Dirichlet allocation), the standard likelihood.
- Much remains to be done in order to compare and improve the different criteria and to extend them into the parametric and numerical cases in order to improve the explanatory power of machine learning .
- These tools have potential applications in many domains.

Four basic Random Variables



SUMMARY AND FUTURE

- ❑ **We have introduced new kinds of data: symbolic data.**
- ❑ **We have introduced new kinds of units: classes**
- ❑ **We have extended standard data analysis to COMPLEX and BIG Data**
- ❑ **CLASSES ARE THE UNITS OF THE FUTURE.**
- ❑ **SYMBOLIC DATA ARE THE NUMBERS OF THE FUTURE!!!**

Actuellement d’après google scholar en moyenne 1 article par jour apparaît dans le monde

Références récentes

- **Applications:**
- G. Nuemi, F. Afonso, A. Roussot, L. Billard, J. Cottenet, E. Combier, E. Diday, C. Quantin (2013): classification of hospital pathways in the management of cancer: application to lung cancer in the region of burgundy, Cancer Epidemiology journal, Elsevier. 2013 Oct; 37(5):688-96. Epub 2013 Jul 10. doi: 10.1016/j.canep.2013.06.007.
- C. Guinot, D. Malvy, J-F. Schemann, F. Afonso, R. Haddad, E. Diday (2015): Strategies evaluation in environmental conditions by symbolic data analysis: application in medicine and epidemiology to trachoma. ADAC (Advances in Data Analysis and Classification). March 2015, Volume 9, Issue 1, pp 107-119. DOI: 10.1007/s11634-015-0201-2.
- M. Ochs, E. Diday, F. Afonso, (2016) "From the Symbolic Analysis of Virtual Faces to a Smiles Machine," IEEE Trans Cybern. Volume: 46 Issue:2. doi: 10.1109/TCYB.2015.2411432
- **Overview:**
- E. Diday (2016) “Thinking by classes in Data Science: the symbolic data analysis paradigm”. WIREs Comput Stat 2016, 8:172–205. Doi: 10.1002/wics.1384.

RECENT BOOKS

- **F. Afonso, E. Diday, C. Toque (2018) “Data Science par Analyse des Données Symboliques”. Book (448 pages). TECHNIP editor.**
- **L. Billard E. Diday (2019) “Clustering Methodology for Symbolic Data”. 2020 John Wiley & Sons Ltdt. Print ISBN:9780470713938 |Online ISBN:9781119010401 |DOI:10.1002/9781119010401).**
- **Diday E., Rong G., Saporta G., Wang H., (editors and co-authors) (2020) Advances in Data Science (Symbolic, Complex and Network Data). ISTE WILEY Science Publishing Ltd.**
- **Last one contains:**
- **Emilion R., Diday E. (2020) " Likelihood in the Symbolic Context" Chapter 2 in Advances in Data Sciences, edited by Diday E., Rong G., Saporta G., Wang H. , (2020), Publisher: ISTE WILEY Science Publishing Ltd). . http://www.iste.co.uk/book.php?id=1597.**
- **Diday E. (2020) Explanatory Tools for Machine Learning in the Symbolic Data Analysis Framework. Chap 1 in Advances in Data Sciences, edited by Diday E., Rong G., Saporta G., Wang H., Publisher: ISTE WILEY Science Publishing Ltd). . http://www.iste.co.uk/book.php?id=1597**

- Trois livres parus en 2018, 2019 et 2020.