# Symbolic data analysis of complex data

Edwin Diday,
CEREMADE,
University Paris-Dauphine, France

Beijing 2011

# OUTLINE

- What is the Symbolic Data Analysis (SDA) paradigm?
- Why SDA is a good tool for Complex Data Mining?
- The SYR software.

# What is a PARADIGM?

From *The Structure of Scientific Revolutions* (**Kuhn,** 1962)

We can define a **scientific paradigm** by

- What is the failure in the actual practice?
- What is the paradigm shift?
- What is to be observed and scrutinized?
- What kind of questions and how are they structured?
- What are the principles and the theoretical development?
- What is the applicability domain?

# What is the actual failure which has produced the SDA Paradigm?

The failure is that in the actual practice

- Only the “individual” kind of observations is considered.
- Therefore, these individual observations are only described by standard numerical and categorical variables.

# What is the SDA paradigm shift ?

It is the transition

- from “individual observations” described by standard variables of numerical and categorical values.
- To “higher level observations” described by variables of symbolic values taking care of their internal variation (intervals, probability distributions, sets of categories or numbers, random variables,…) which can not be treated as numbers.

# From lower level of individual observation to higher level observation variables

**Standard Data Table**

| | $X_1$ | | $X_j$ | |
|---|---|---|---|---|
| $ind_1$ | | | | |
| | | | | |
| $ind_i$ | | | $\mathbf{X_{ij}}$ | |
| | | | | |
| $ind_n$ | | | | |

A number
(age of Zidane)

**Symbolic Data Table**

| | $Y_1$ | | $Y_j$ | |
|---|---|---|---|---|
| $C_1$ | | | | |
| | | | | |
| $C_i$ |  | |  | |
| | | | | |
| $C_k$ | | | | |

A symbolic data (age of Zidane team)

$X_j$ is a Random variable of numerical or categorical value

$Y_j$ is a Random variable of value: a random variable represented by a distribution.

(Distributions are the number of the future! Schweitzer 1984)

# Space of representation of the symbolic Data

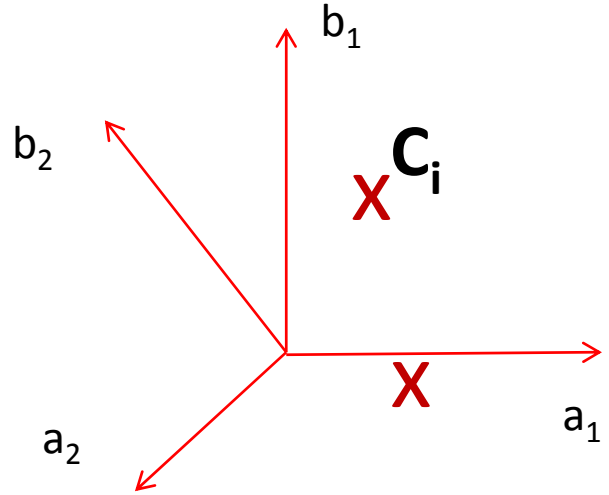
Numerical variables for symbolic data

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

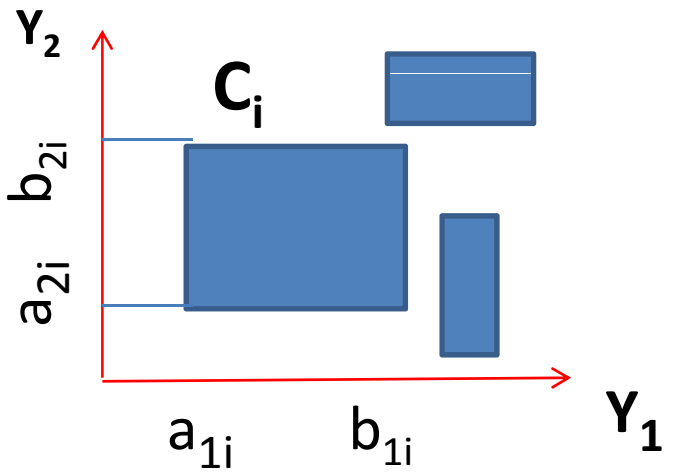
$(Y_1(C_i), Y_2(C_i)) = ([a_{1i}, b_{1i}], ([a_{2i}, b_{2i}])$

Symbolic variables

| | $Y_1$ | $Y_2$ |
|---|---|---|
| $C_1$ | | |
| | | |
| $C_i$ | | |
| | | |
| $C_k$ | | |

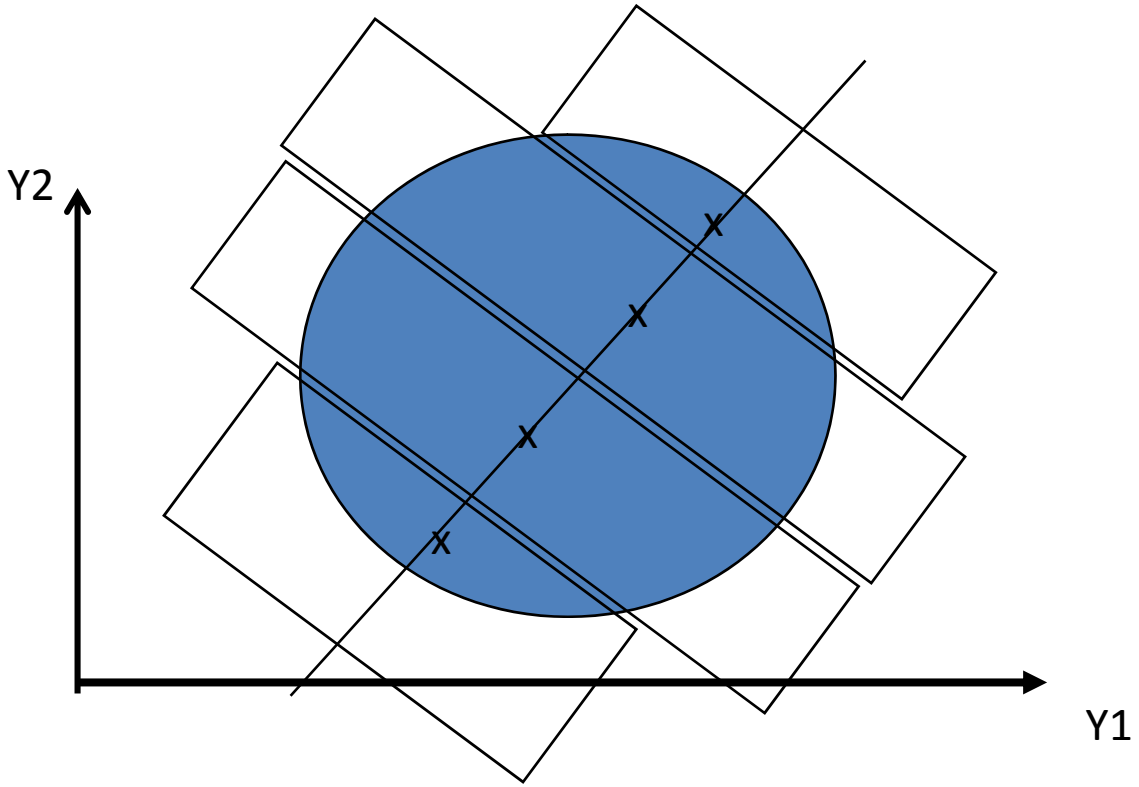



Numerical Space: 4 variables , no variation appears




Symbolic Space: 2 variables , variation appears

# From standard observations to higher level observations, The correlation is not the same!




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

# What is to be observed and scrutinized?

In SDA the observed and scrutinized are

**"higher level observations".**

In opposition to

**Individual observations:** a player, a fund, a stock,…

**Higher level observations are:**

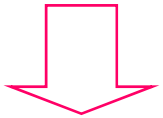
- **Classes:** a player subset, a subset of funds, of stocks, …
- **Categories:** American funds, European funds,…
- **Concepts:** an intent: volatile American funds.

  an extent: the volatile American funds of a given data base.

# 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**

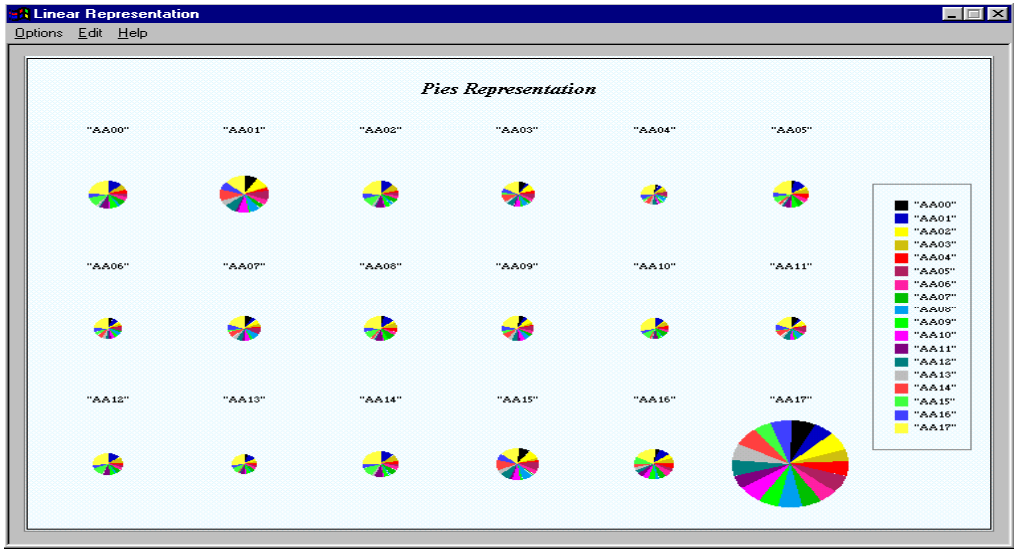
| Players category | Weight Min | Weight Max | Size Min | Size Max | Eur | Afr |
|---|---|---|---|---|---|---|
| Very good | 80 | 95 | 1.70 | 1.95 | 0. 7 | 0.3 |

**Concern:**

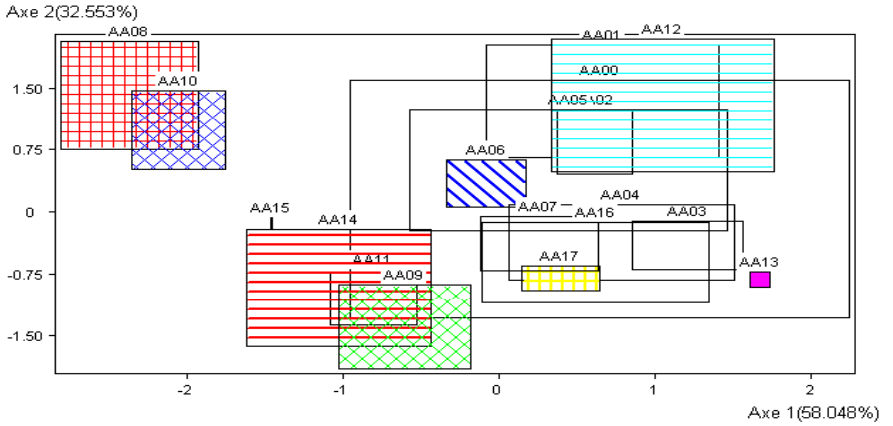
**The initial variables are lost and the variation is lost!**

# 1. Non parametric: Extending Data Mining to Symbolic Data

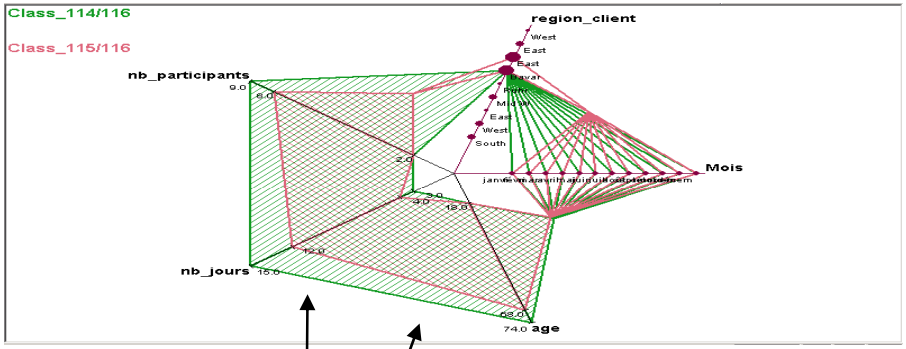
## Kohonen map



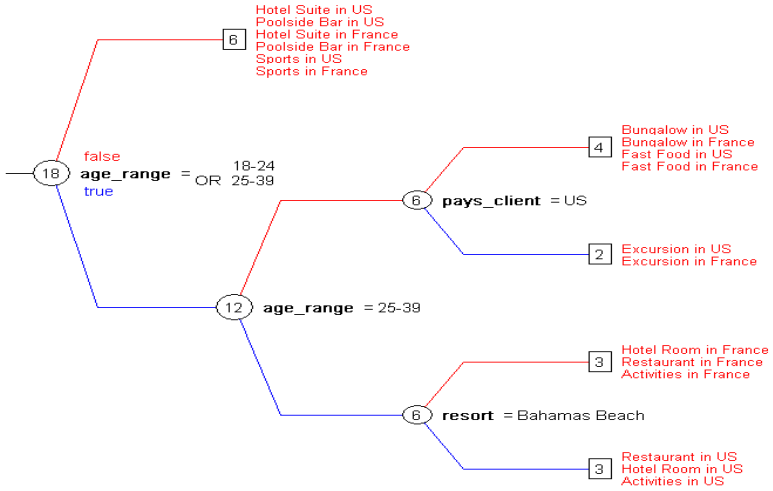

## Principal componnent



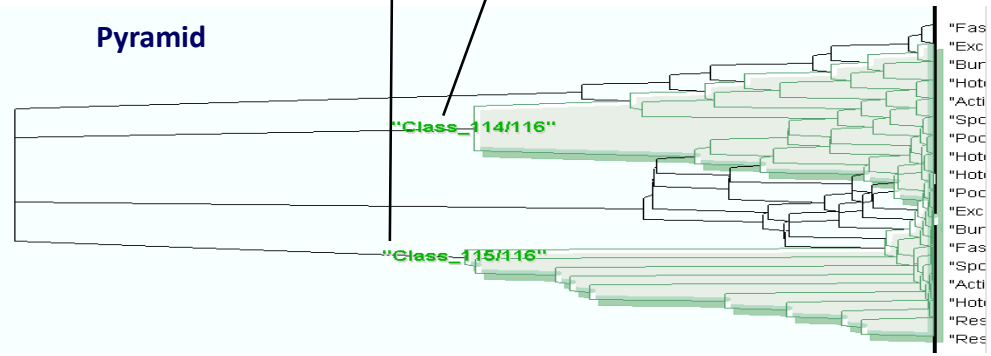

## Zoom stars overlapping




## Top down clustering tree or decision tree




## Pyramid



# From the observed to the variables

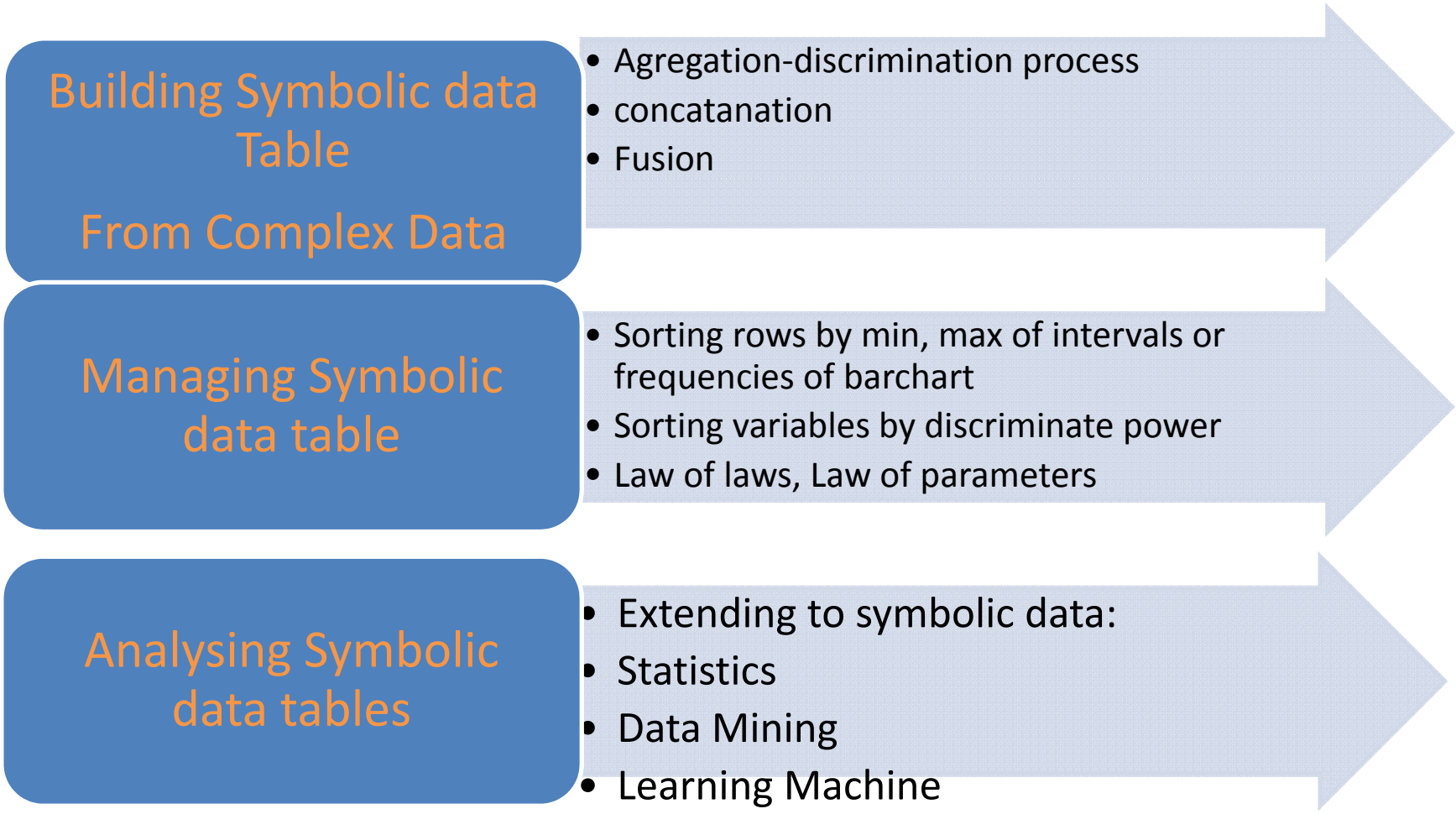
**World of observation**

- Individuals
- Classes
- Categories
- Concepts

**World of variables**

- Standard numerical and categorical
- Symbolic variables taking care on the internal variation
- A value of a standard categorical variable
- Intent defined by symbolic variables plus a way for calculating the extent.

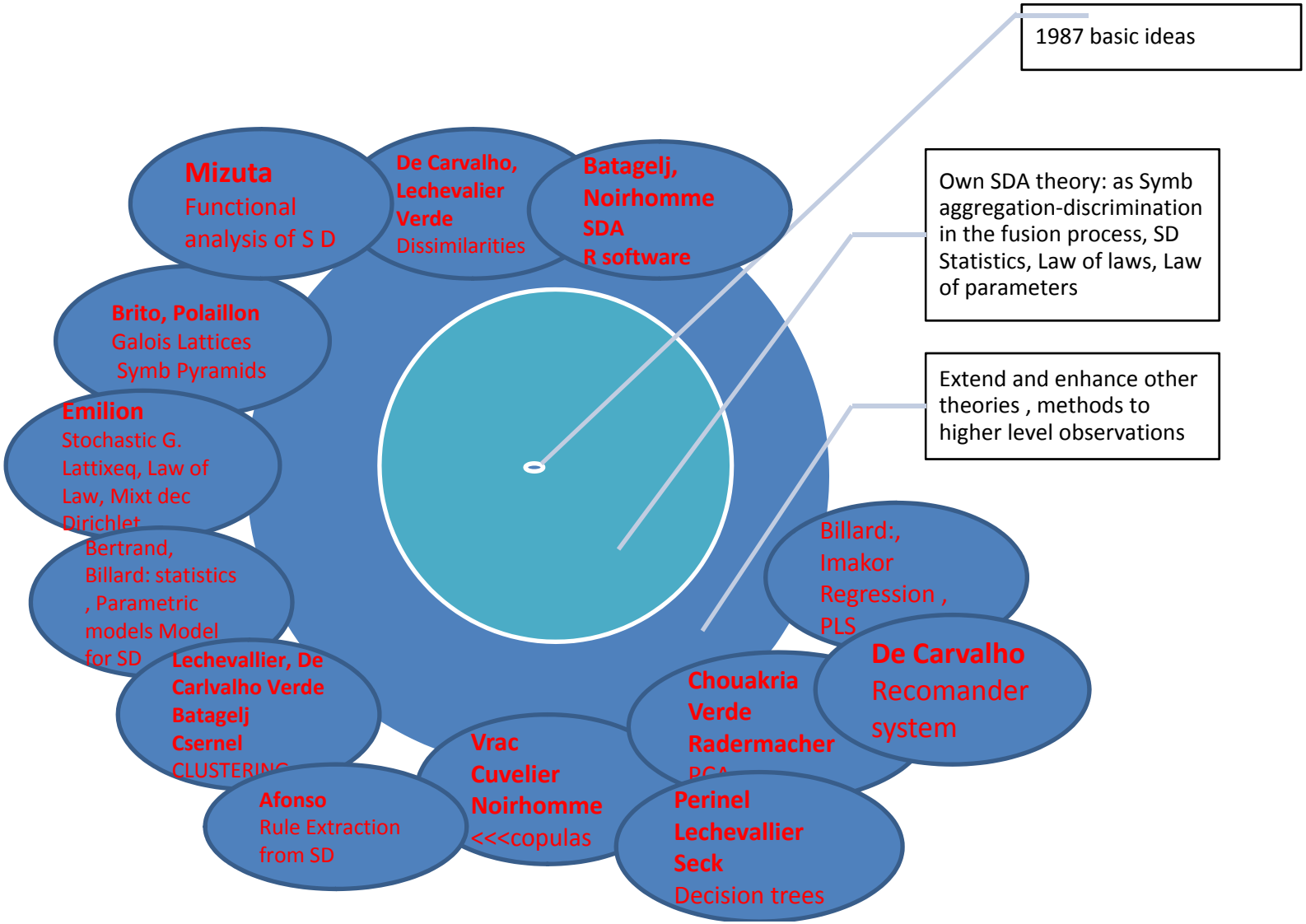
# What kind of questions and how are they structured?



# Some SDA Principles

- **The agregation process must discriminate SD between the higher level observations.**
- **Work in the Symbolic space as much as possible.**
- **Represent as much as possible in the output the internal variation of the higher level observations.**
- **Don’t reduce the output to be just numerical and try to remain in the same kind of SD than in input.**
- **Any extension of a standard theory or method to SD must contain this theory or method as a special case.**

# SDA Theory Expansion



# What is the SDA applicability domain?

- Standard data table: unique set of individuals

  Standard numerical and (or) categorical variables, induce categories which can be considered as “higher level observations” described by discriminate symbolic variables taking care of their internal variation.

- Native symbolic data table: no individuals
- Complex data: several sets of different individuals

# What are Complex Data?

**Any data which cannot be considered as a "standard observations x standard variables" data table.** Several data tables describing different kind of observations by different variables.

**Examples**

- Hierarchical Data
- Multi source Data
- Specific complex data: Textual Data, images Multimedia data (text and images, …).

# NUCLEAR POWER PLANT

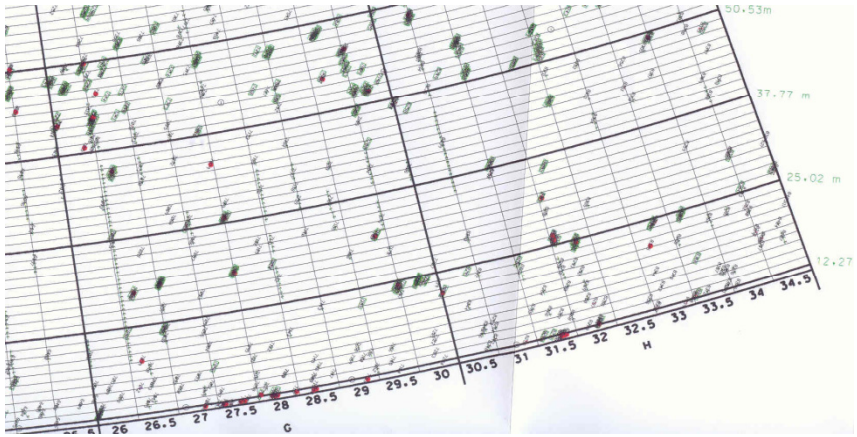
**Nuclear thermal power station**

## Inspection :

Inspection machine

Craks

Cartography of the towel by a grid




**PB: FIND CORRELATIONS BETWEEN 3 CLASSICAL DATA TABLES OF DIFFERENT UNITS AND VARIABLES:**

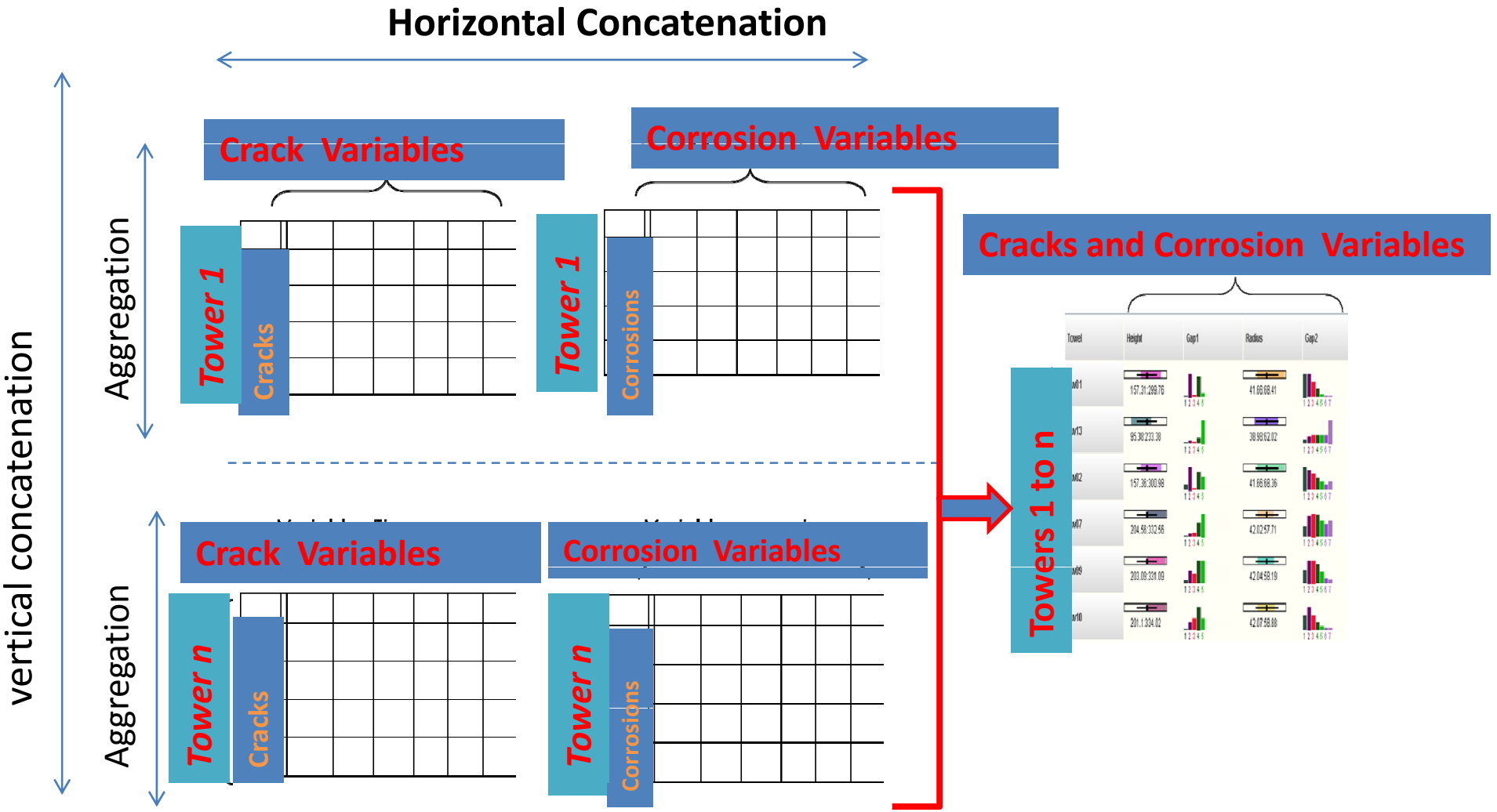
**Table 1) Observations: Cracks . Variables: Cracks description.**

**Table 2) Observations: vertices of a grid. Variables: Gap deviation at different periods compared to the initial model position.**

**Table 3) Observations: vertices of a grid. Variables: Gap depression from the ground.**

**ARE Transformed in ONE Symbolic Data Table where the concepts are interval of height or each concept is a tower. On this new table correlation between variables can be calculated.**

# From complex data to symbolic data table: The Fusion process

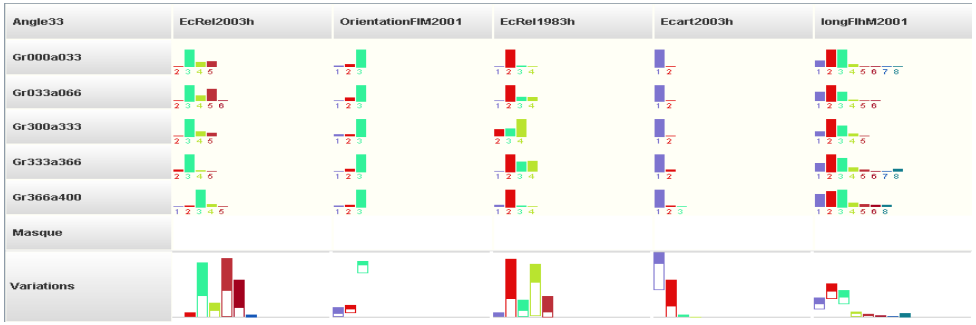


# Advantages of the SD Table obtained by complex data fusion

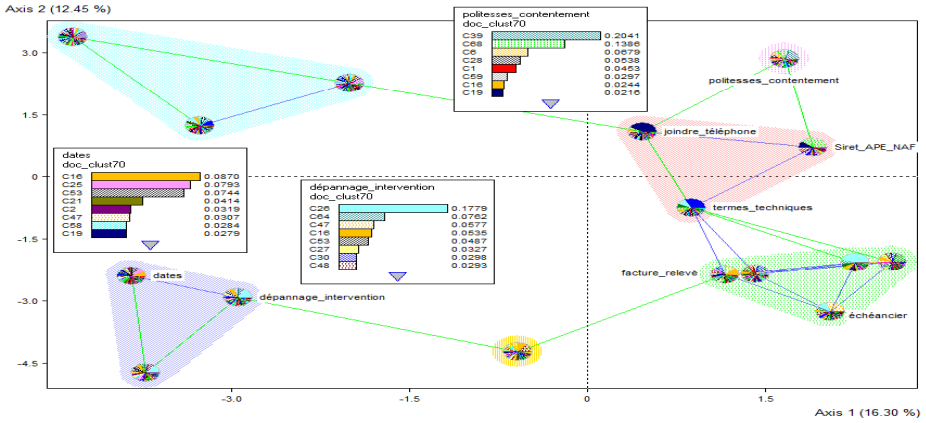
- Allows the synthesis of multiple and heterogeneous data in a unique SDT.
- Significative Correlations between heterogeneous variables can be obtained
- The individuals (the towers) can be considered as higher level observations in a more complet and fiable form.
- Allows the higher level observations positioning and their comparison in a much better way than considering the multiple d.ata sources separately

# SYR SOFTWARE

- Produce a Symbolic Data Table by fusioning complex data.

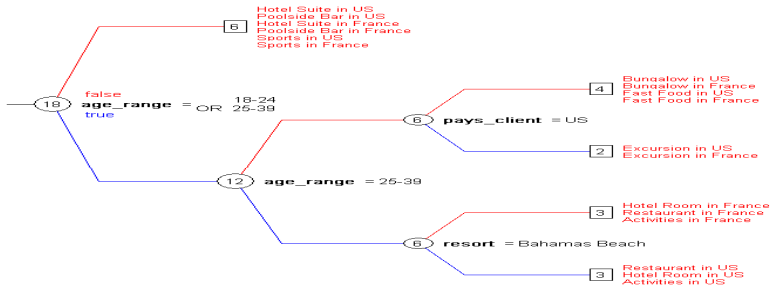



- Manage Symbolic Data Tables: sort rows and columns by discriminant power




- Analyse Symbolic data tables: SSTAT,SPCA,Sclustering,etc.

- Produce network, rules and decision trees.



# GRAPHICAL REPRESENTATION

## NETSYR

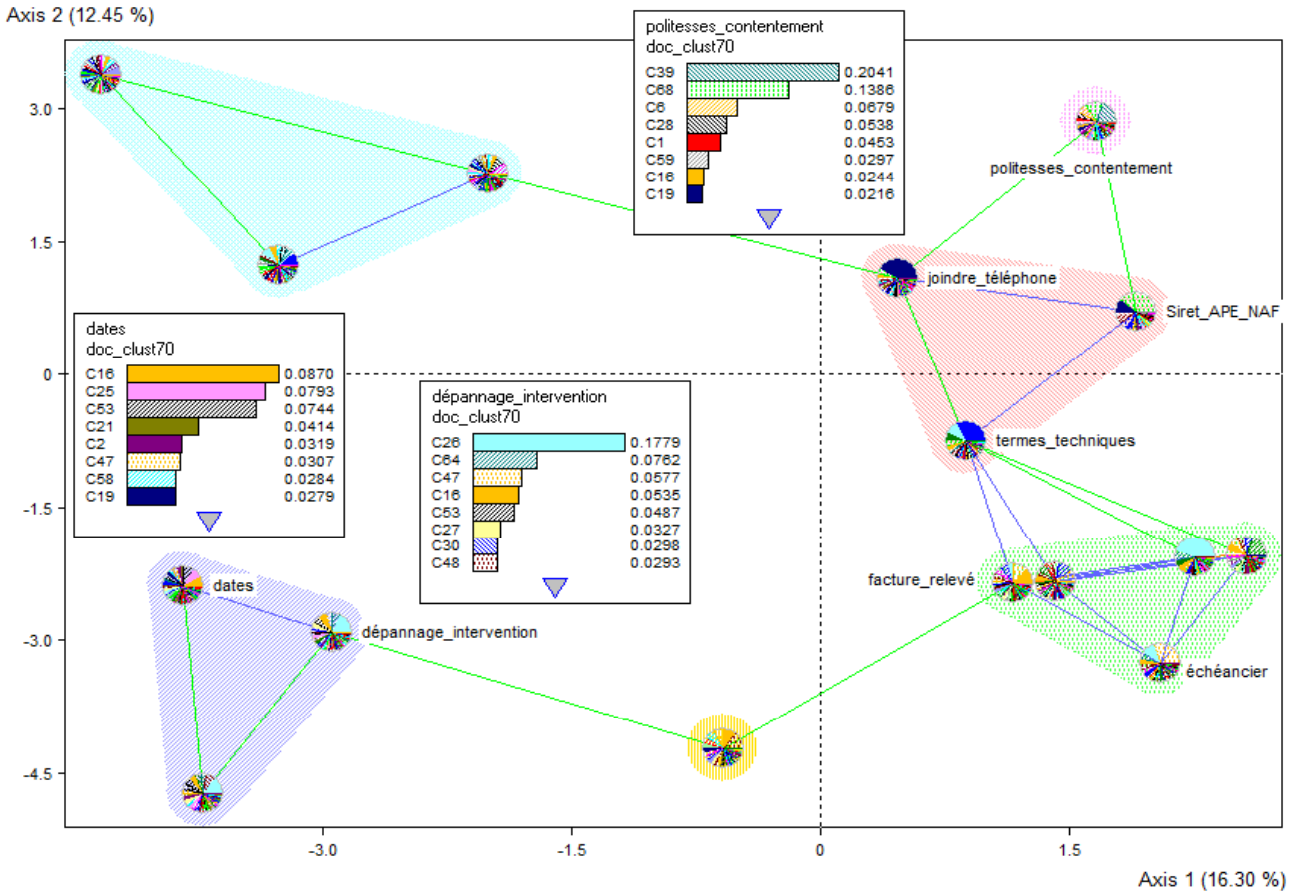
**GRAPHICAL REPRESENTATION** of higher level observations, by Pie Charts
And their Bar chart description.

**Overlapping Clusters**

**Induced SOCIAL NEWORK**
**Based on dissimilarities**

**ANNOTATION :**
of variables and higher level observation

Moving, Zooming…




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

# Conclusion

- We have shown that SDA is a new paradigme based on the transition from *standard individual observations* to *higher level observations* described by symbolic data.
- SDA is a useful tool for Complex Data Mining.
- Much remains to be done for improving the fusion process
- Much remains to be done for extending actual methods to SD, for example:

  the topology inside the symbolic space, summarizing graphs, social networks, for extending Factorial Analysis, Canonical Analysis, PLS etc., to *higher level observations.*

# Some References


- E. Diday, M. Noirhomme éditeurs et co-authors (2008) " Symbolic Data Analysis and the SODAS software"
  Book 457 pages. Wiley. ISBN 978—0-470-01883-5.

- L. Billard, E. Diday (2006) "Symbolic Data Analysis: conceptual statistics and data Mining". Book Wiley. 330 pages. ISBN 0-470-09016-2

- Afonso F., Diday E., Badez N., Genest Y., Orcesi A. (2010) Use of symbolic data analysis for structural health monitoring applications. *IALCCE 2010. Taiwan.*

# HOW SYMBOLIC DATA ARE BUILT?
# From Database to Concepts