# FROM DATA MINING TO KNOWLEDGE MINING: SYMBOLIC DATA ANALYSIS AND THE SODAS SOFTWARE

E. Diday
University of Paris IX Dauphine
and INRIA

## AIM

**FROM HUDGE DATA IN AN ECONOMIC WAY**

**-Extract new knowledge**

**-Summarize**

**-Concatenate**

**-Solve confidentiality**

**-Explain correlation**

**HOW? By working on HIGHER LEVEL UNITS called CONCEPTS necessary described by more complex data extending Data Mining to Knowledge Mining.**

## OUTLINE

1) THE MAIN IDEA:
FIRST AND SECOND ORDER OBJECTS.

2)THE INPUT OF A SYMBOLIC DATA ANALYSIS:

SYMBOLIC DATA TABLE.

3) SOURCE OF SYMBOLIC DATA: FROM DATA

BASES, FROM CATEGORICAL VARIABLES.

4) MAIN OUTPUT OF SYMBOLIC DATA ANALYSIS

ALGORITHMS: SYMBOLIC DESCRIPTIONS AND

SYMBOLIC OBJECTS.

5) THE MAIN STEPS OF A SDA.

6)TOOLS OF SYMBOLIC DATA ANALYSIS

7) SYNTHETICAL VIEW OF THE SODAS PROJECT

## THE MAIN IDEA:
## FIRST AND SECOND ORDER OBJECTS

THE ARISTOTLE ORGANON (IV B.C.) CLEARLY DISTINGUISHES "FIRST ORDER OBJECTS" (AS THIS HORSE OR THIS MAN) CONSIDERED AS A UNIT DESCRIBING AN INDIVIDUAL OF THE WORLD , FROM "SECOND ORDER OBJECTS" (AS A HORSE OR A MAN) ALSO TAKEN AS A UNIT DESCRIBING A CLASS OF INDIVIDUALS.

# FOUR IDEAS

1) **ONLY TWO LEVELS OF UNITS:**
   First level: Individuals
   Second level: concepts

2) **THE CONCEPTS CAN BE CONSIDERED AS NEW INDIVIDUALS OF HIGHER LEVEL.**
3) **A CONCEPT IS DESCRIBED BY USING THE DESCRIPTION OF A CLASS OF INDIVIDUALS OF ITS EXTENT.**
4) **THE DESCRIPTION OF A CONCEPT MUST EXPRESS THE VARIATION OF THE INDIVIDUALS OF ITS EXTENT**

**FROM FIRST ORDER OBJECTS TO SECOND ORDER OBJECTS IN OFFICIAL STATISTICS**

| Units | Classes | Descr. Var. of the Units | | |
|---|---|---|---|---|
| Case n° | Region | Bedroom | Dining-Living | Socio-Econ Group |
| 11401 | Northern-Metropolitan | 2 | 1 | 1 |
| 11402 | Northern-Metropolitan | 2 | 1 | 3 |
| 11403 | Northern-Metropolitan | 1 | 3 | 3 |
| | | | | |
| 12315 | East-Anglia | 1 | 3 | 3 |
| 12316 | East-Anglia | 2 | 2 | 1 |
| 14524 | Greater-London | 1 | 2 | 3 |

## FROM FIRST ORDER OBJECTS TO SECOND ORDER OBJECTS IN OFFICIAL STATISTICS

| Classes | Descriptive variable of the units | | |
|---|---|---|---|
| Region | Bedroom | Dining-Liv | Socio-Ec gr |
| Northern-Metropolitan | 2 | 1 | 1 |
| Northern-Metropolitan | 2 | 1 | 3 |
| Northern-Metropolitan | 1 | 3 | 3 |
| | | | |
| East-anglia | 1 | 3 | 3 |
| East-anglia | 2 | 2 | 1 |
| East-anglia | 1 | 2 | 3 |
| | | | |

| Classes | Descriptive variables of the classes | | |
|---|---|---|---|
| Region | Bedroom | Dining-Liv | Socio-Ec gr |
| Northern-Metropolitan | (2\3) 2, (1\3) 1 | (2\3) 1, (1\3) 3 | (1\3) 1, (2\3) 3 |
| | | | |
| East-anglia | (2\3) 1, (1\3) 2 | (2\3) 2, (1\3) 3 | (2\3) 3, (1\3) 1 |
| | | | |

Schweitzer (1984): "Distributions are the numbers of the future"

**MORE GENERALY, WHAT IS THE INPUT OF A SYMBOLIC DATA ANALYSIS?**
**SYMBOLIC DATA TABLE**
**+ BACKGROUND KNOWLEDGE**

**SYMBOLIC DATA TABLE :**
**THE CELLS CAN CONTAIN:**
**- SEVERAL QUALITATIVE OR QUANTITATIVE WEIGHTED VALUES**
**- INTERVALS**
**- HISTOGRAMS**

## EXAMPLE OF SYMBOLIC DATA TABLE

| PRODUCT | WEIGHT | TOWN | COLOUR |
|---|---|---|---|
| PRODUCT 1 | 3.5 | Londres | {red, white, yellow} |
| PRODUCT 2 | [ 3 , 8 ] | {Paris, Londres } | |
| PRODUCT 3 | {3.1 , 4.6, 7.2} | | { 0.3 red, 0.7 green} |
| PRODUCT 4 | [(0.4) [2,3[ , (0.6) [3, 8]] | | |
| | | | |

**THE CELLS CAN CONTAIN:**

**-SEVERAL QUALITATIVE OR QUANTITATIVE WEIGHTED VALUES**

**-INTERVALS**

**- HISTOGRAMS**

**SOME BACKGROUND KNOWLEDGE CAN BE GIVEN**

**VARIABLES CAN BE:**

**-TAXONOMIC:** « A SOCIO-ECONOMIC GROUP IS CONSIDERED TO BE "SELF-EMPLOYED" IF IT IS "PROFESSIONAL SELF-EMPLOYED" OR "OWN ACCOUNT NON-PROFESSIONAL".

**-HIERARCHICALLY DEPENDENT :**

THE VARIABLE: “DOES THE COMPANY HAS COMPUTERS? “ AND THE VARIABLE: “ KIND OF COMPUTER” ARE HIERARCHICALLY LINKED.

**- WITH LOGICAL DEPENDENCIES:**

« IF AGE(W) IS LESS THAN 2 MONTHS THEN HEIGHT(W) IS LESS THAN 10 ».

## SOURCE OF SYMBOLIC DATA

**.FROM CATEGORICAL VARIABLES:**
- GIVEN. (AS « TYPE OF EMPLOYMENT »)
- OBTAINED BY CLUSTERING.

**.FROM DATA BASES:** QUERY CREATING A NEW CATEGORICAL VARIABLE: cartesian prod

**.FROM EXPERT:** NATIVE SYMBOLIC DATA:

Scenario of road accidents, species of insects

**.FROM CONFIDENTIAL DATA**
IN ORDER TO HIDE THE INITIAL DATA BY LESS ACCURACY

**.FROM STOCHASTIC DATA TABLE:**
THE PROBABILITY DISTRIBUTION , THE HISTOGRAM THE PERCENTILES OR THE RANGE OF ANY RANDOM VARIABLE ASSOCIATED TO EACH CELL OF A DATA TABLE

### EXAMPLE

| | Mathematics | Physics | Litterature | |
|---|---|---|---|---|
| | | | | |
| Tom | $X_M$ | $X_P$ | $X_L$ | |
| | | | | |
| | | | | |
| Paul | | | | |
| | | | | |

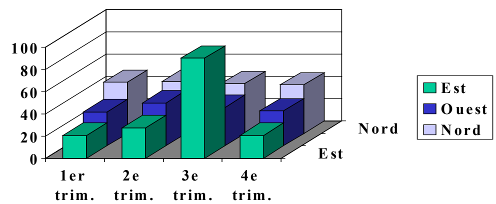
$X_M$ is the random variable which associates to each exam of TOM his mark in mathematics.

From $X_M$ several kinds of symbolic objects can be defined by using in each cell:
- Probability distr.
- Histograms
- Inter-quartile intervals

.FROM TIME SERIES

- IN DESCRIBING INTERVALS OF TIME:
( the variation of the values each week)

- IN DESCRIBING A TIME SERIES BY THE HISTOGRAM OF ITS VALUES.




# FROM RULES
## Example:Districts description

| Districts | *male-full-time-employee %* | *male-part-time-employee %* | *male-self-employed %* | Z |
|---|---|---|---|---|
| D1 | 8% | 5% | 20% | Z1% |
| D2 | 12% | 9% | 15% | Z2% |
| | | | | |
| Dn | | | | |

# Example of Rule defining class

- ***R1**: male-full-time-employee%(X,low) ∧ male-part-time-employee%(X,low) ∧ neighbor(X,Y) ∧ comm-activities(Y,high)*

→ ***male-self-employed%(X,high)***

- 70 districts X satisfy the rule: the low percentage of full-time and part-time male employees in district X adjacent neighbor of *Y*, with many commercial activities, implies a high percentage of self-employed males in *X*.

# Districts description of rules

| Descript of rules extent | *male-full-time-employee%* | *male-part-time-employee %* | *male-self-employed%* | Z |
|---|---|---|---|---|
| R1 | [8%,12%] | [5%, 9%] | [20%, 15%] | [Z1%,Z2] |
| R2 | [2%, 6%] | [4%,8%] | [18%,14%] | ….. |
| | | | | |
| Rn | | | | |

## MAIN OUTPUT OF SYMBOLIC DATA ANALYSIS ALGORITHMS:

SYMBOLIC DESCRIPTIONS

SYMBOLIC OBJECTS.

## SYMBOLIC DESCRIPTIONS

| Description | AGE | SPC |
|---|---|---|
| D1 | {12,20,28} | {employee,worker} |
| D2 | [5, 33] | {teacher,countryman} |

CONCEPTS ARE MODELED BY SYMBOLIC OBJECTS

WHATS A CONCEPT?

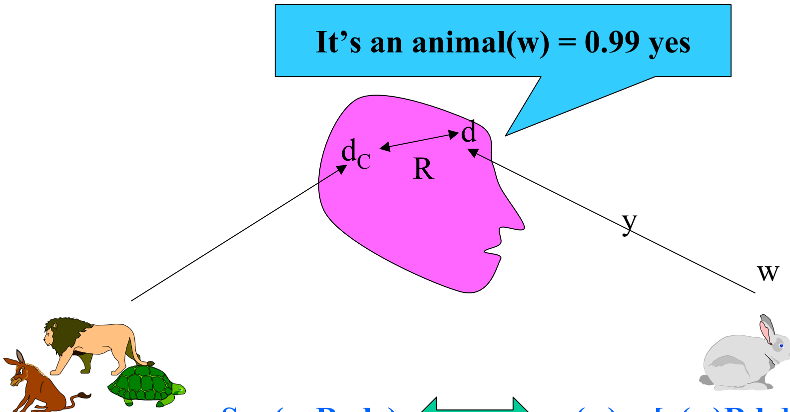
A CONCEPT IS DEFINED BY AN

* INTENT : ITS CHARACTERISTIC PROPERTIES

* EXTENT:THE SET OF INDIVIDUALS WHICH SATISFY THESE PROPERTIES

LIKE OUR MIND, SYMBOLIC OBJECTS MODEL CONCEPTS BY AN INTENT AND A WAY OF COMPUTING THE EXTENT

# SYMBOLIC OBJECT




**$S = (a, R, d_C)$** $\Longleftrightarrow$ **$a(w) = [y(w)Rd_C]$**

## TWO KINDS OF SYMBOLIC OBJECTS

### BOOLEAN SYMBOLIC OBJECTS

**S = (a, R, d1)**

**d1= {12, 20 ,28} x {employee, worker}]**
**R = ($\subseteq$ , $\subseteq$ ),**

**a(w) = [age(w) $\subseteq$ {12, 20 ,28}] $\wedge$ [SPC(w) $\subseteq$ {employee, worker}]**

**a(w) $\in$ {TRUE, FALSE}.**

## THE MEMBERSHIP FUNCTION« a » MODAL CASE

**S = (a, R, d):**

**a(w) = [age(w) $R_1$ {(0.2)12, (0.8) [20 ,28]}] $\wedge$ [SPC(w) $R_2$ {(0.4)employee, (0.6)worker}]**

**a(w) $\in$ [0,1].**

**First approach: simple or flexible matching**

**R= ($R_1$, $R_2$ ): $r \, R_i \, q = \sum_{j=1,k} r_j \, q_j \, e^{(r_j - \min(r_j, q_j))}$ .**

**Second approach:**

**Probabilistic: if dependencies, copulas, derivation of the joint distribution, transforming the joint density in [0,1].**

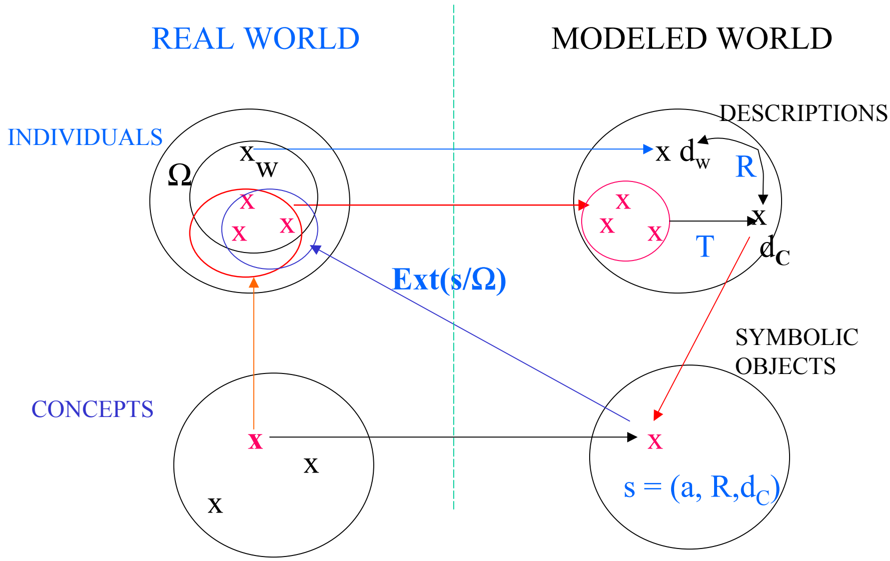
## EXTENT OF A SYMBOLIC OBJECT S:

**BOOLEAN CASE:**

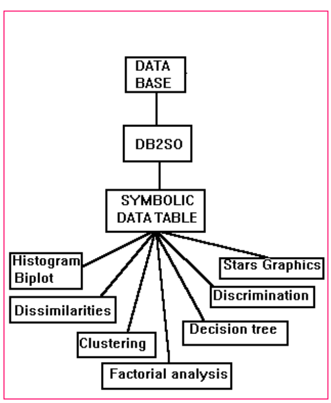
**EXT(s) = {W $\in$ $\Omega$ / a(W) = TRUE}.**

**MODAL CASE**

**$EXT_\alpha$ (S)= $EXTENT_\alpha$ (a) = {W $\in$ $\Omega$ / a( W ) $\geq$ $\alpha$}.**




## AN OVERVIEW ON THE SODAS SOFTWARE



## THE MAIN STEPS FOR A SYMBOLIC DATA ANALYSIS IN SODAS

**. PUT THE DATA IN A RELATIONAL DATA BASE (Oracle, Acces, ...)**

**.DEFINE A CONTEXT BY GIVING**

*** The Units (Individuals, Households,...)**

*** The Classes (Regions, Socio-economics groups,...)**

*** The descriptive variables of the units**

**. BUILD A SYMBOLIC DATA TABLE WHERE**

*** The units are the preceding classes**

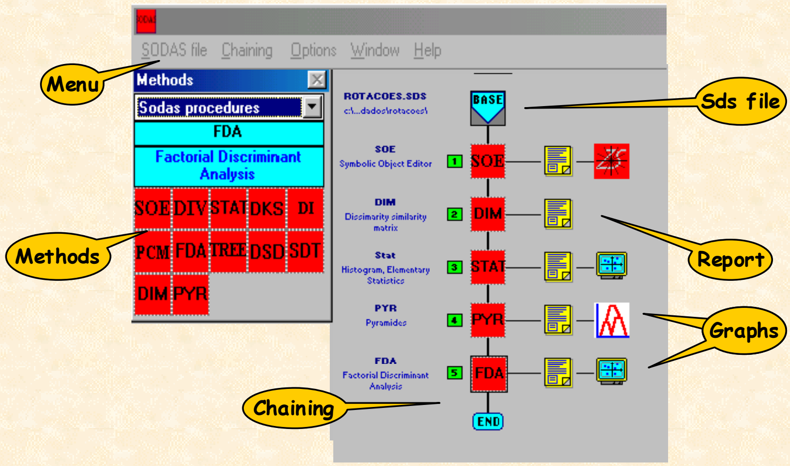
*** The descriptions of each class is obtained by a generalization process applied to its members**

**APPLY**

**SYMBOLIC DATA ANALYSIS TOOLS**

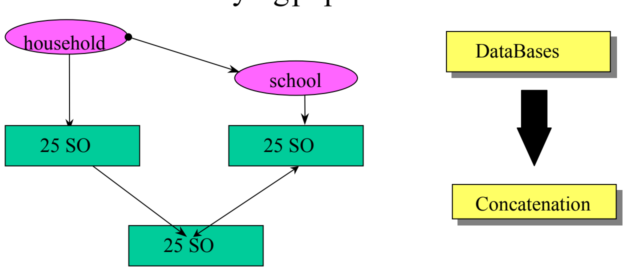
- **Correlation, Mean, Mean Square**
- **Histogram of a symbolic variable**
- **Dissimilarities between symbolic descriptions**
- **Clustering of symbolic descriptions**
- **Principal component Analysis**
- **Decision Tree**
- **Graphical visualisation of Symbolic Objects**

# SODAS Software




## Concatenation of summarized data from several populations

Join two or more sets of SO based on different underlying populations



# DE CARVALHO'S DISSIMILARITY MEASURES

A straightforward extension of similarity measures for classical data matrices with nominal variables.

| | ***Agreement*** | ***Disagreement*** | ***Total*** |
|---|---|---|---|
| ***Agreement*** | $\alpha=\mu(A_j \cap B_j)$ | $\beta=\mu(A_j \cap c(B_j))$ | $\mu(A_j)$ |
| ***Disagreement*** | $\chi=\mu(c(A_j) \cap B_j)$ | $\delta=\mu(c(A_j) \cap c(B_j))$ | $\mu(c(A_j))$ |
| ***Total*** | $\mu(B_j)$ | $\mu(c(B_j))$ | $\mu(Y_j)$ |

where $\mu(V_j)$ is either the cardinality of the set $V_j$ (if $Y_j$ is a *nominal* variable) or the length of the interval $V_j$ (if $Y_j$ is a *continuous* variable).

# DE CARVALHO'S DISSIMILARITY MEASURES

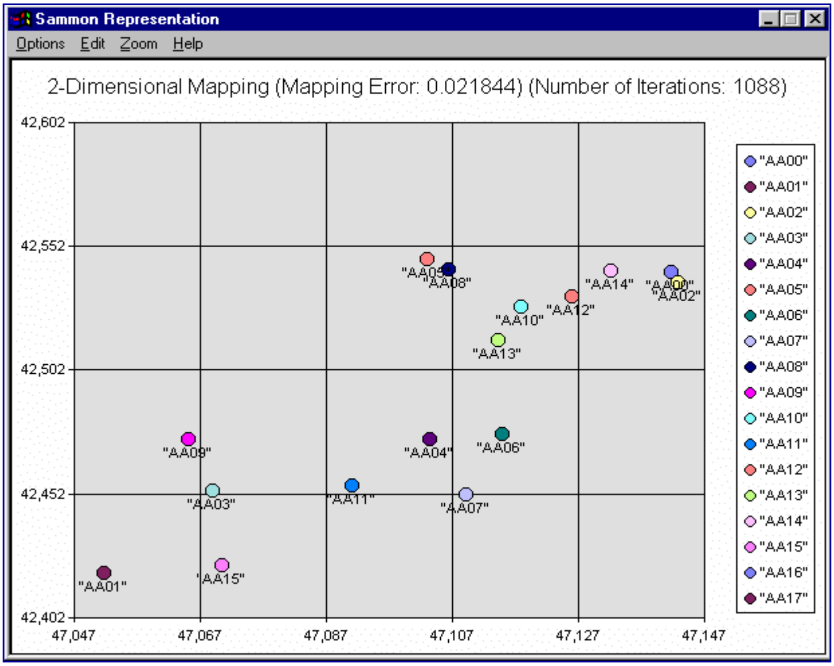
Five different similarity measures $s_i$, $i = 1, ..., 5$, are defined:

| $s_i$ | ***Comparison Function*** | ***Range*** | ***Property*** |
|---|---|---|---|
| $s_1$ | $\alpha/(\alpha+\beta+\chi)$ | [0,1] | metric |
| $s_2$ | $2\alpha/(2\alpha+\beta+\chi)$ | [0,1] | semi metric |
| $s_3$ | $\alpha/(\alpha+2\beta+2\chi)$ | [0,1] | metric |
| $s_4$ | $\frac{1}{2}[\alpha/(\alpha+\beta)+\alpha/(\alpha+\chi)]$ | [0,1] | semi metric |
| $S_5$ | $\alpha/[(\alpha+\beta)(\alpha+\chi)]^{1/2}$ | [0,1] | semi metric |

The corresponding dissimilarities are $d_i = 1 - s_i$.
The $d_i$ are aggregated on p variables by the generalised Minkowski metric, thus obtaining:

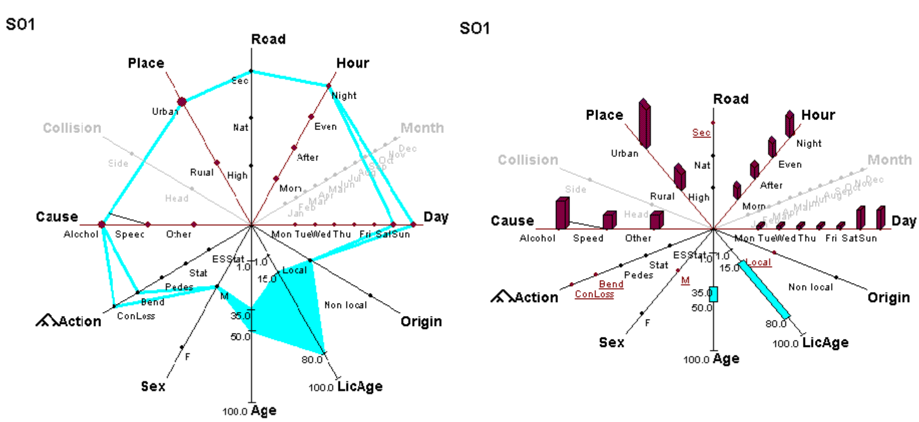
SO_1

$$d_i(a,b)=\sqrt[q]{\sum_{j=1}^{p}\left[w_j d_i(A_j,B_j)\right]^q} \qquad 1 \le i \le 5$$




# SOE Software

## Operations available in a Graphic Window



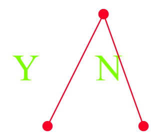

<u>Figure21: Examples of *2D* and *3D Zoom Stars*</u>

The object name and the variables names can be moved

# Method:
# DIV algorithm

Divisive and symbolic algorithm

- THE METHOD
  - divisive — recursive, descendant
  - binary — binary question
  - symbolic — input: symbolic data




output: symbolic interpretation

clusters: assertion object

# (within-cluster inertia)

Additive criterion

$$W(P) = \sum_{C_k \in P} Q(C_k)$$

**Q** measures the quality of a cluster

$P=(C_1,C_2,\ldots,C_k)$

$$Q(C) = \frac{1}{2n_k} \sum_{\omega_i \in C_k} \sum_{\omega_j \in C_k} d^2(\omega_i, \omega_j)$$

$n_k$ = number of individuals in $C_k$

d = distance or dissimilarity between symbolic objects (Haussdorf, KHI2)

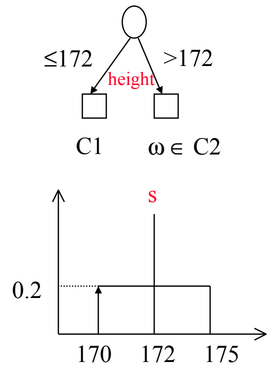
Normalization: inverse of dispersion (symbolic variance)

inverse of maximum deviation

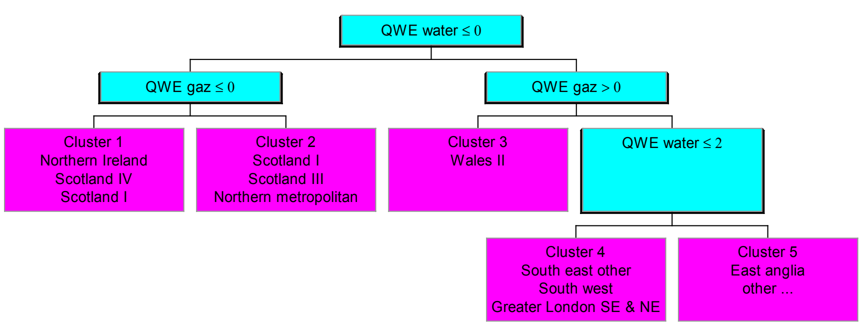
# Choice of the cut value: S

- numerical or ordinal
  – order the value of the variables
  – choice s in the middle of 2 consecutive values
- interval
  – reduce the interval in a point: the center
  – choice s: idem numerical method
- probabilistic
  – on probabilistic distribution
  – choice s = mediane

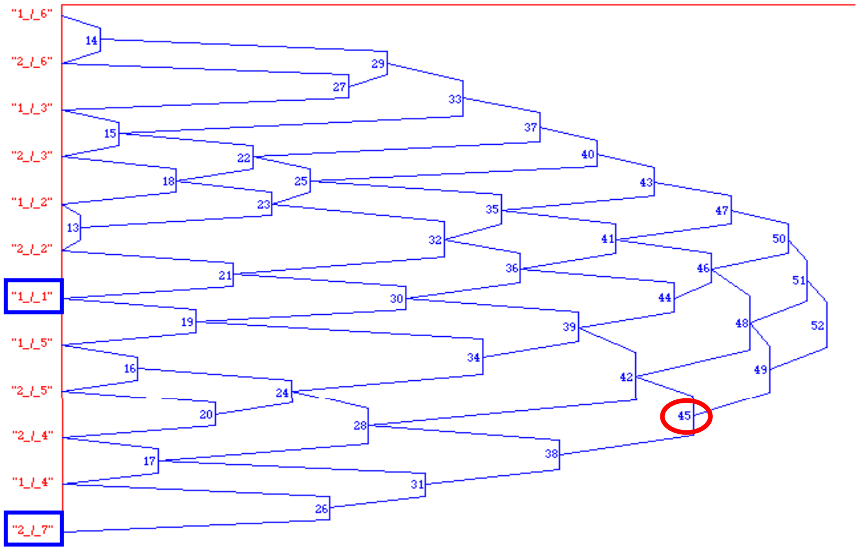
Individual ω de C:
height(ω)= [170,175]



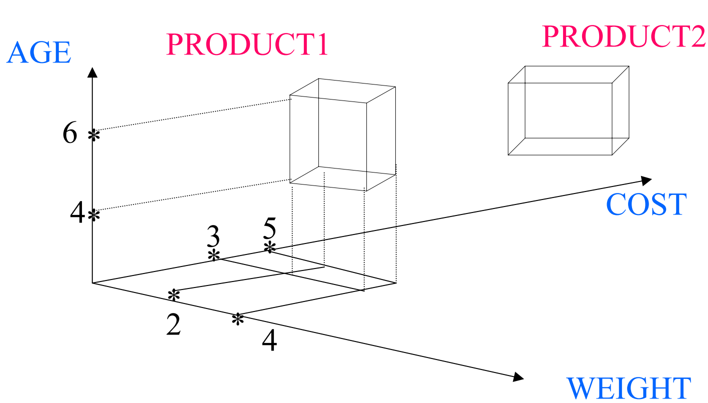

# Output results
# Clustering tree

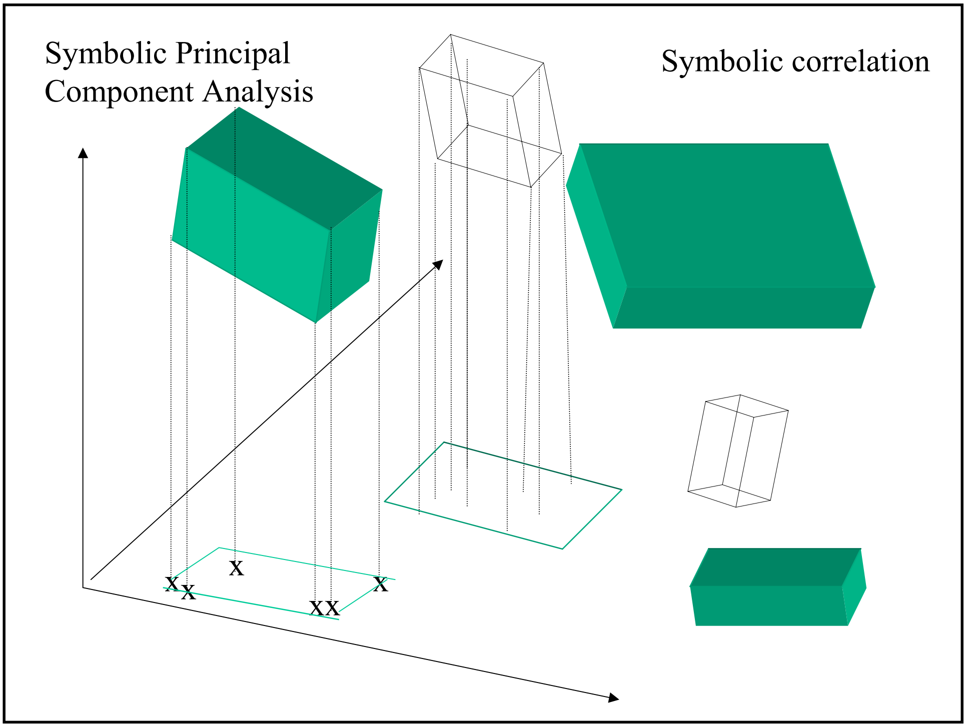


# Pyramid

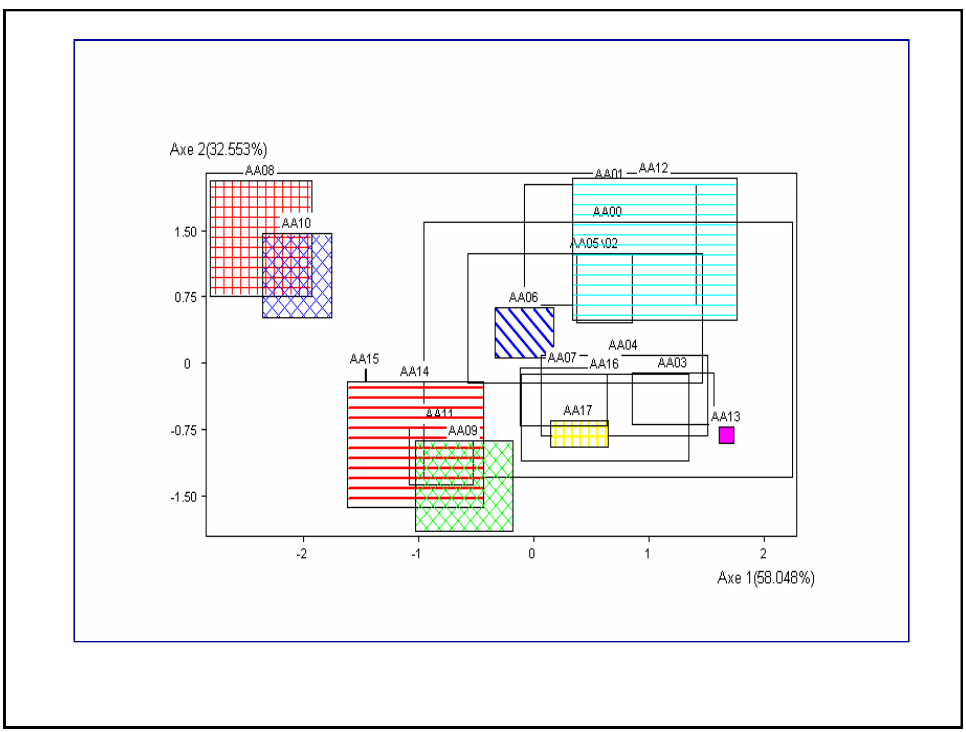



| PRODUCTS | WEIGHT | COST | AGE | PROFIT |
|---|---|---|---|---|
| PRODUCT1 | [2,4] | [3,5] | [4,6] | [0,3] |
| PRODUCT2 | [4,5] | [3,4] | [1,6] | [2,7] |
| PRODUCT3 | [1,6] | [2,7] | [5,8] | [6,9] |



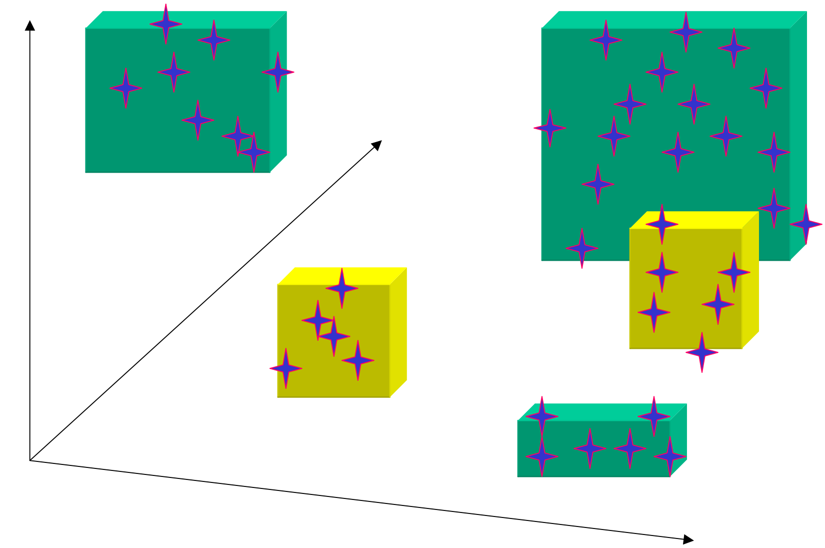


Symbolic Principal
Component Analysis
Symbolic correlation
x
x
x
xx
x

Axe 2(32.553%)
Axe 1(58.048%)
AA08
AA10
AA01
AA12
AA00
AA05
AA02
AA06
AA04
AA07
AA16
AA03
AA15
AA14
AA11
AA09
AA17
AA13
1.50
0.75
0
-0.75
-1.50
-2
-1
0
1
2

## SYMBOLIC DESCRIPTION OF CLASSES



# Decisional *Symbolic* Data Analysis

- Discrimination

Factorial Discriminant Analysis
(Lauro, Verde, Palumbo, 2000)

Geometrical Visualisation of Symbolic description well separated from each other on a factorial discr plane.

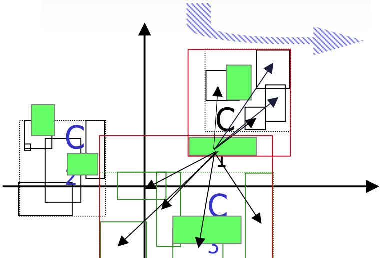
- Classification rules

*Based on the proximity between objects on factorial planes*

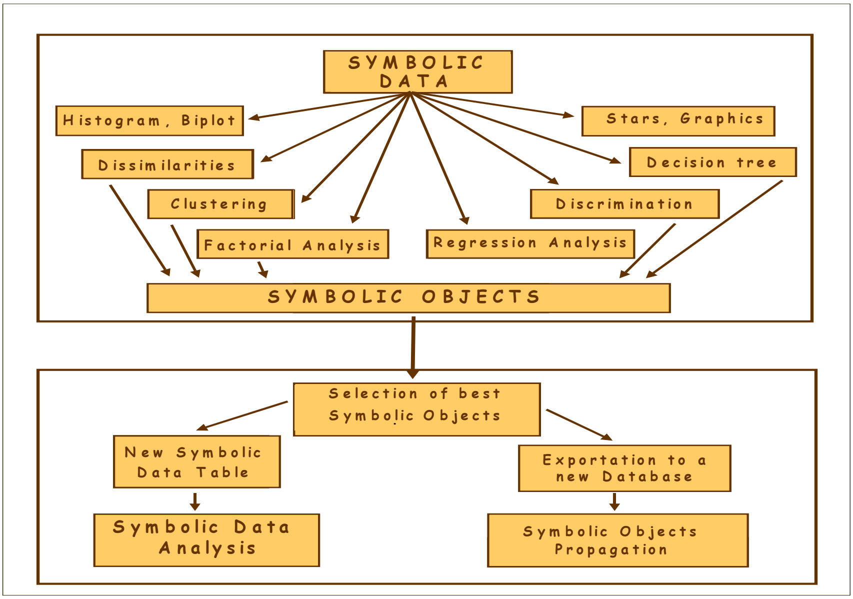
Geometrical rules:

*Minimum SO class volume increasing*

*De Carvalho's dissimilarity measure*



## THE SODAS 2 SOFTWARE FROM ASSO




**NEW PROBLEMS APPEAR**
.QUALITY, ROBUSTNESS RELIABILITY OF THE APPROXIMATION OF A CONCEPT BY A SYMBOLIC OBJECT,
.THE SYMBOLIC DESCRIPTION OF A CLASS,
.THE CONSENSUS BETWEEN SYMBOLIC DESCRIPTIONS ETC..

**MUCH HAS TO BE DONE:**
SYMBOLIC
.REGRESSION, FACTORIAL ANAL.
.MULTIDIMENSIONAL SCALING,
.MIXTURE DECOMPOSITION,
.NEURAL NETWORK, KOHONEN MAP
.CONCEPT PROPAGATION…...

## Some recent advances:

- Mixture decomposition of Distributions of distributions (by Copulas, Dirichlet and Kraft stochastic process)

- Stochastic Symbolic Conceptual lattices using capacity theory

- Symbolic class description

-Symbolic Regression

-NEXT FUTUR

-- Spatial symbolic clustering by pyramids

- Symbolic time series.

- Consensus between different description of the same set of units

# AIM ATTAINED

FROM HUDGE DATA BASES

IN AN ECONOMIC WAY

WE ARE ABLE TO: -Extract new knowledge

-Summarize

-Concatenate

-Solve confidentiality

-Explain Correlation

HOW? By working on HIGHER LEVEL UNITS

extending Data Mining to Knowledge Mining.

## CONCLUSION

**Symbolic Data Analysis is an extension of standard data analysis therefore**

**First principle: any Symbolic Data Mining method must have as a special case method of Data Mining on standard data.**

**Second principle : the output must be a symbolic description or symbolic object**

**New problems appear as the quality, robustness and reliability of the approximation of a concept by a symbolic object, the symbolic description of a class, the consensus between symbolic descriptions etc..**

**Due to the intensive development of the information technology the great chapters of the standard statistics will have to be think in these new terms.**

# References

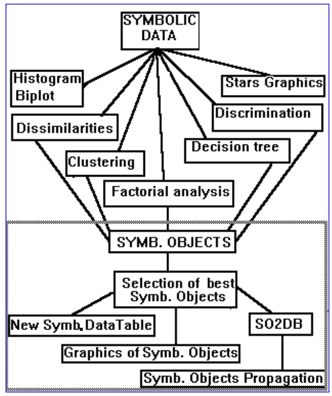

SPRINGER, 2000 :
"Analysis of Symbolic Data"
H.H., Bock, E. Diday, Editors . 450 pages.

**JASA (Journal of the American Statistical Association)**
**"From the Statistic of Data to the Statistic of Knowledge:**
**Symbolic Data Analysis" Billard, Diday June, 2003 .**

**Electronic Journal of S. D. A.: ESDA**
**E. Diday, R. Verde, Y. Lechevallier**


Download SODAS and SODAS information :
**www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm**

# THE SODAS 2 SOFTWARE FROM ASSO

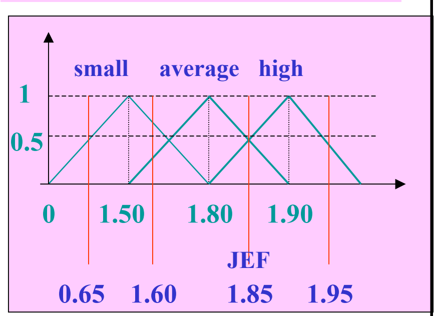



# FROM FUZZY DATA TO SYMBOLIC DATA

| | height | weight | hair |
|---|---|---|---|
| Paul | 1.60 | 45 | yellow |
| Jef | 1.85 | 80 | yellow |
| Jim | 0.65 | 30 | black |
| Bill | 1.95 | 90 | black |

Initial Data

From Numerical to Fuzzy Data




| | height | | | weight | hair |
|---|---|---|---|---|---|
| | small | average | high | | |
| Paul | 0.70 | 0.30 | 0 | 45 | yellow |
| Jef | 0 | 0.50 | 0.50 | 80 | yellow |
| Jim | 0.50 | 0 | 0 | 30 | black |
| Bill | 0 | 0 | 0.48 | 90 | black |

Fuzzy Data

| | height | | | weight | hair |
|---|---|---|---|---|---|
| | small | average | high | | |
| {Paul, Jef } | [0, 0.70] | [0.30, 0.50] | [0, 0.50] | [45, 80] | yellow |
| {Jim, Bill} | [0, 0.50] | 0 | [0, 0.48] | [30, 90] | black |

Symbolic Data