

**DE LA STATISTIQUE DES DONNÉES  
À LA STATISTIQUE DES CONCEPTS:  
L'ANALYSE DES DONNEES  
SYMBOLIQUES**

E. Diday

Université Paris IX Dauphine

# PLAN

- Des individus aux concepts: atomes de connaissances
- Des concepts aux données symboliques exprimant leur variation interne
- Pourquoi on ne code pas les données symboliques sous forme de données classiques ?
- Stratégie classique versus symbolique
- Le logiciel SODAS issu de deux projets européens
- L'extension des méthodes classiques
- Aspects théoriques: modèle probabiliste, treillis de Galois stochastiques
- Une application industrielle
- perspectives et moralité
- Références
- Diffusion
- Conclusion

## DES INDIVIDUS AUX CONCEPTS

Dans l'Organon (IV AJC), Aristote distingue clairement les **unités de premier ordre** (comme cet homme ou ce cheval), des **unités de second ordre** (comme l'homme, le cheval ou l'animal).

**Unités de premier ordre → INDIVIDUS**

**Unités de second ordre → CONCEPTS**

## CONCEPTS: Intension, extension

Dans "la logique ou l'art de penser" (1662), Arnauld et Nicole

UN CONCEPT EST DEFINI PAR UNE

\* **INTENSION** : SES PROPRIETES CARACTERISTIQUES.

\* **EXTENSION**: L'ENSEMBLE DES INDIVIDUS QUI  
SATISFONT CES PROPRIETES

# Approche Classique Versus Symbolique: les unités de l'étude

## Classique : des individus

Oiseaux



Habitant, logement



Joueur de foot (Zidane,...)



Image



Articles vendus



Traces d'usager WEB

Patients victime d'infarctus

Feuilles de maladies

Abonnés Mobiles



## Symbolique : des concepts

Espèces d'oiseaux



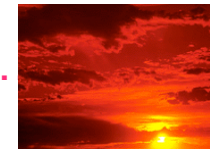
IRIS, Régions d'habitation



Joueurs d'une équipes (Marseille)



Types d'image ( marines,..)



Magasins d'une chaîne



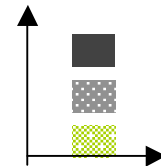
Usagers

Trajectoire dans les services et Hôpitaux

Bénéficiaires



Niveaux de consommation



# Des unités statistiques classiques aux concepts, la statistique n'est pas la même!

Sur une île se trouvent 400 hirondelles, 100 autruches, 100 pingouins :

Tableau de données classiques

Oiseau	Catégorie	Vole	Taille (cm)
1	Pingouin	Non	80
2	Hirondelle	Oui	70
600	Autruche	Non	125

Tableau de données symboliques

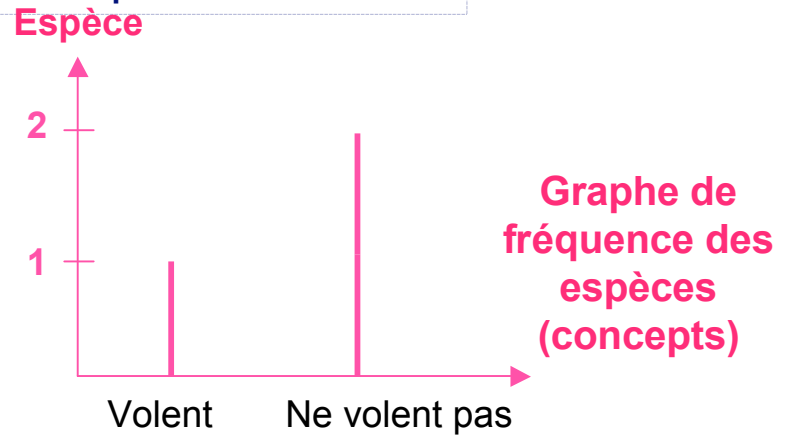
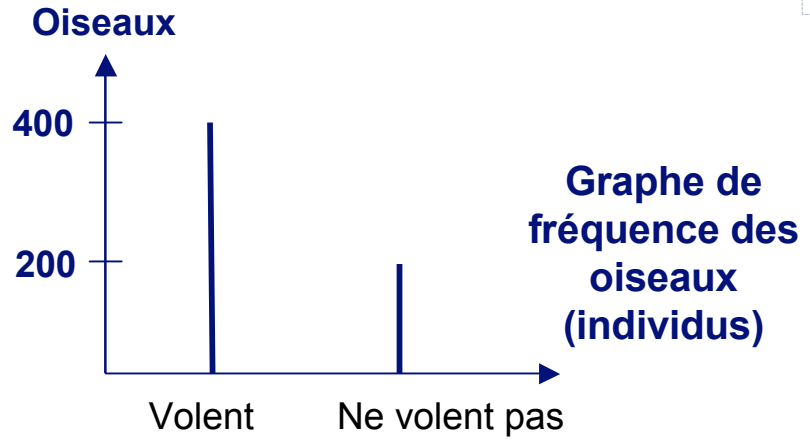
Espèce	Vole	Couleur	Taille	Migre
Hirondelle	Oui	0.3n,0.7gris	[60, 85]	Oui
Autruche	Non	0.1noir,0.9g	[85, 160]	Non
Pingouin	Non	0.5n,0.5gris	[70, 95]	Oui

« Pingouin », « hirondelle » et « autruche » sont les concepts construits à partir de la variable Catégorie

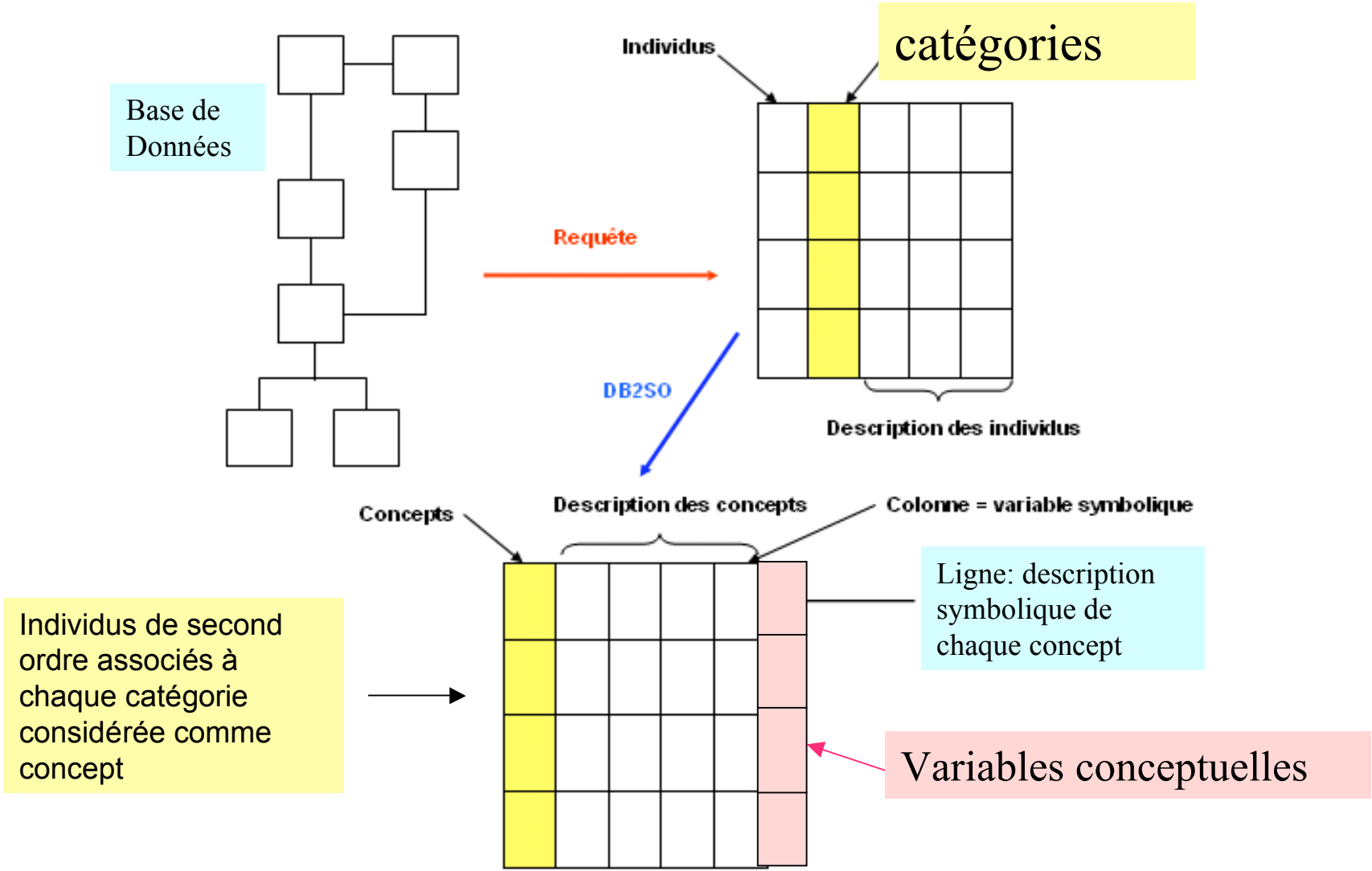
« L'espèce » est un concept qui devient la nouvelle unité statistique; on ne s'intéresse plus aux 600 individus en tant que tels

Les variations dues aux individus inclus dans l'extension de chaque concept sont conservées sous forme d'intervalle ou de diagramme de fréquences

Ajout d'une variable « conceptuelle » : elle s'applique au concept



# DE LA BASE DE DONNEES AUX CONCEPTS



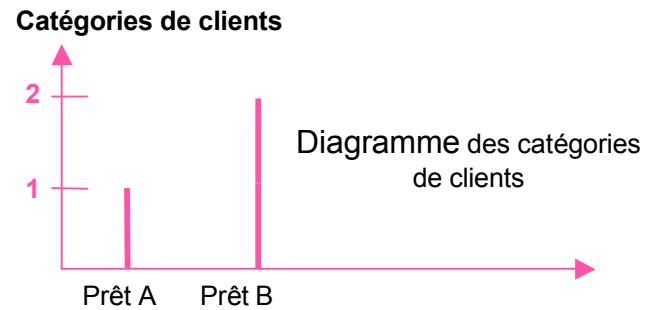
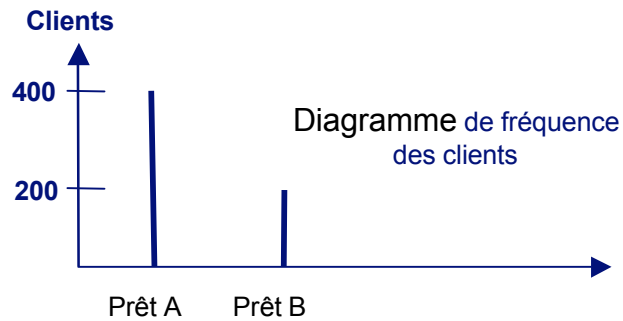
**APPLICATION AUX PRETS BANQUAIRES: caractériser les profils sociaux démographiques des types de prêts selon le lieu (50.000 IRIS x Types de prêts).**

Client	Catégorie de client	Iris	Prêt	Habitat	Revenu
1	Prêt B iris 2	2	B	Pavillon	80
2	Prêt A iris 1	1	A	HLM	70
600	Prêt B iris 3	3	B	HLM	125

Catégorie de client	Iris	Prêt	Habitat	Revenu	Nombre d'habitants de l'iris
Prêt A iris 1	1	A	0.3Pav,0.7 HLM	[60, 85]	10000
Prêt B iris 2	2	B	0.7Pav,0.3HLM	[70, 120]	4000
Prêt B iris 3	3	B	0.9Pav,0.1HLM	[78, 200]	2000

Tableau des données classiques initiales.

Tableau des descriptions symboliques



(Revolving)

Les diagrammes de fréquence sur les individus (clients) et les concepts catégories de clients sont inversés.

**L'APPROCHE SYMBOLIQUE N'EST PAS MEILLEURE  
QUE L'APPROCHE CLASSIQUE!!!**

Elle est **DIFFERENTE** et **COMPLEMENTAIRE**.

EXEMPLE:

FAIRE LA STATISQUE DES ESPECES D'OISEAUX N'EST  
PAS MEILLEUR QUE FAIRE LA STATISTIQUE DES  
OISEAUX: C'EST **DIFFERENT** ET **COMPLEMENTAIRE**.

Si on peut dire que l'Analyse des données a rendu les individus  
à la statistique, alors on peut dire aussi que l'Analyse des  
Données Symboliques lui rend les concepts.

## DONNEES SYMBOLIQUES

EQUIPE	POIDS	NATIONALITE	NOMBRE DE BUTS
DIJON	80.5	{Française}	12
LYON	[ 75 , 89 ]	{Fr, Brés, Arg }	
PARIS-ST G.	{83.1 , 84.6, 87.2, ...}		{0.3 (0), 0.4 (1), ...}
NANTES	[(0.4) [70,80[, (0.6)[80, 90]		

LES VARIABLES SONT DITES SYMBOLIQUES

CAR A VALEUR NON PUREMENT NUMERIQUES indispensable

POUR EXPRIMER LA VARIATION INTERNE DES CONCEPTS

Chaque cellule peut contenir:

- une ou plusieurs valeurs qualitatives ou quantitatives
- un intervalle
- un diagramme, histogramme, une f. de répartition, une courbe...

# ANALYSE DES DONNEES SYMBOLIQUES: 3 ETAPES

**PREMIERE ETAPE:** DES INDIVIDUS AUX CATEGORIES.

**DEUXIEME ETAPE:** DES CATEGORIES AUX CONCEPTS DECRITS PAR DES VARIABLES SYMBOLIQUES et AUGMENTATION DE LA DIMENSION PAR DES VARIABLES CONCEPTUELLES et des CONNAISSANCES SUPPLEMENTAIRES.

**TROISIEME ETAPE:** EXTRACTION DE NOUVELLES CONNAISSANCES PAR EXTENSION (au moins) DES OUTILS STANDARDS DE LA STATISTIQUE, DE L'AD ET DU DATA MINING AUX CONCEPTS DECRITS PAR DES DONNEES SYMBOLIQUES

# CONNAISSANCES SUPPLEMENTAIRES

EN PLUS DU TABLEAU DE DONNEES SYMBOLIQUES  
POSSIBILITE D'AJOUT EN ENTREE DE :

- VARIABLES DECRIVANT SPECIALEMENT  
LES CONCEPTS (i.e. PAS LES INDIVIDUS)

- VARIABLES TAXONOMIQUES



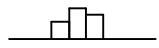
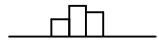
- DEPENDANCES HIERARCHIQUES

- CORRELATIONS

- DEPENDANCES LOGIQUES

## COMMENT CONSERVER LA CORRELATION ET L'EXPLIQUER?

Indiv	Zone	opht	dentaire	lieu	assureur
i1	C1	12.5	3,5	Lyon1	Dupont
i2	C1	9.6	2,1	Paris3	Durand
i3	C1	11.4	6.5	Lyon1	Dupont
i4	C2	3.2	1,6	Paris1	Charles
i5	C2	7.1	4,8	Lyon2	Cartier

Concept	opht	dentaire	lieu	assureur	Cor(opht, dent)
C1	[9.6, 12.5]	[2.1, 6.5]	{Lyon1, Paris 3}		Cor <sub>C1</sub> (opht, dent)
C2	[3.2, 7.1]	[1.6, 4.8]	Paris 3		Cor <sub>C2</sub> (opht, dent)
C3	[9.2, 10.1]	[6.2, 8.1]	Pau 1		Cor <sub>C3</sub> (opht, dent)
C4	[5, 8.4]	[7.3, 9.4]	Pau 4		Cor <sub>C4</sub> (opht, dent)

**Ensuite: expliquer la corrélation par régression ou arbre de décision symbolique. Résultat: la période et le lieu expliquent la corrélation des coûts opht et dent = un vendeur d'assurances.**

# 3 Avantages de l'ADS

## **Etudier les bonnes unités statistiques:**

Les assurés plutôt que les feuilles de maladies.

## **Réduire le nombre d'individus**

En passant des individus aux classes d'individus ou concepts:

- Réduction des données Massives,
- Réduction des Données Manquantes
- Réduction de la Confidentialité

## **Réduire le nombre de variables**

- Une variable à valeur intervalle ce n'est pas deux variables Min, Max
- Une variable à valeur histogramme à 20 classes ce n'est pas 20 variables numériques.

# CINQ PRINCIPES

**1) A CHAQUE ETAPE, SEULEMENT DEUX NIVEAUX:**

Premier niveau: les individus réifiés en individu de premier ordre

Second niveau: les concepts réifiés en individu du second ordre

**2) LES CONCEPTS PEUVENT EUX-MÊME ÊTRE CONSIDÉRÉS  
COMME DES UNITÉS ET REIFIÉS EN INDIVIDUS DE  
SECOND ORDRE (ISO)**

**3) UN ISO PEUT ÊTRE DÉCRIT EN UTILISANT UNE CLASSE  
D'INDIVIDUS DE L' EXTENSION DU CONCEPT QU'IL  
REPRESENTE.**

**4) LA DESCRIPTION D'UN ISO DOIT EXPRIMER LA  
VARIATION DES INDIVIDUS DE CETTE EXTENSION**

**5) POUR ANALYSER CES ISO IL FAUT TENIR COMPTE DE  
CETTE VARIATION ET LA REPRÉSENTER**

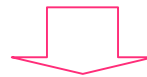
**6) LES DESCRIPTIONS DOIVENT ÊTRE EXPLICATIVES EN  
S'EXPRIMANT DANS LE LANGAGE DE L'UTILISATEUR  
SOUS FORME DE CONJONCTIONS DE PROPRIÉTÉS  
RELATIVES AUX VARIABLES INITIALES QU'IL A FOURNI.**

# Comparaison données classiques / données symboliques au niveau du codage

Pourquoi on ne code pas les données symboliques sous forme de données classiques ?

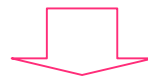
Tableau symbolique

Cat. de buteurs	Poids	Taille	Nationalité
Très Bons	[80, 95]	[1.70, 1.95]	{0.7 Eur, 0.3 Afr}



Codage en données classiques

Catégorie de buteurs	Poids Min	Poids Max	Taille Min	Taille Max	Eur	Afr
Très Bons	80	95	1.70	1.95	0.7	0.3

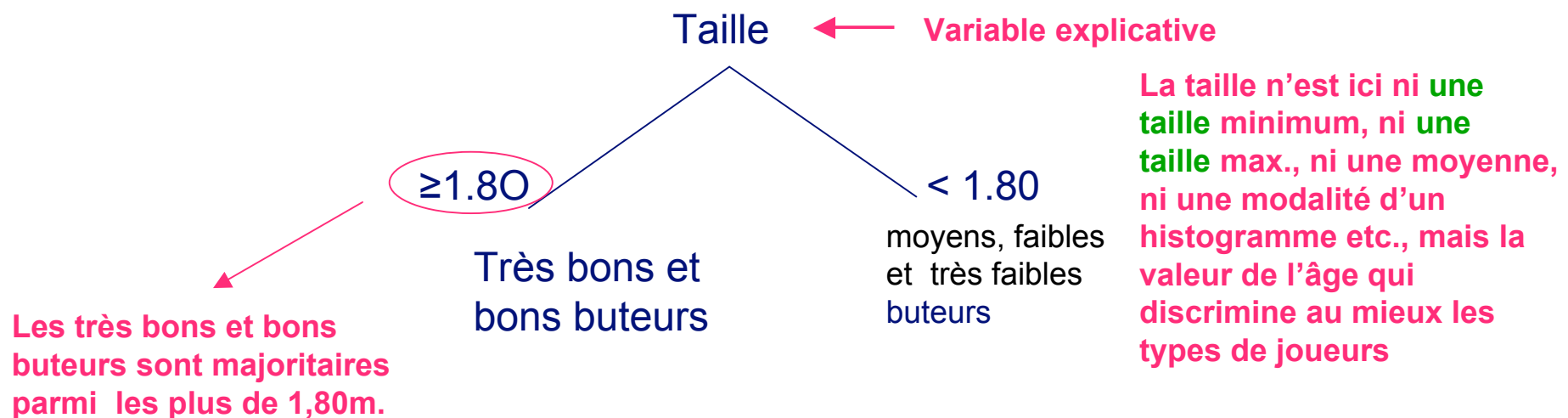


**Codage classique : on perd les variables initiales ,  
on les démultiplie, on perd la variation.**

# Exemple 1: Perte d'information du codage classique de données symboliques

## Les arbres de décision

- En codage classique la variable « Taille » n'existe plus car seules « Taille Min » et « Taille max » demeurent.
- Le codage symbolique fournit l'arbre suivant qui discrimine les classes de buteurs et que le codage classique ne peut fournir:
- Les catégories obtenues peuvent être réifiées en individus décrits par SODAS



# Perte d'information du codage classique de données symboliques

L'approche classique perd la notion de variable à valeur intervalle et ne permet de construire un histogramme que sur les min OU les max

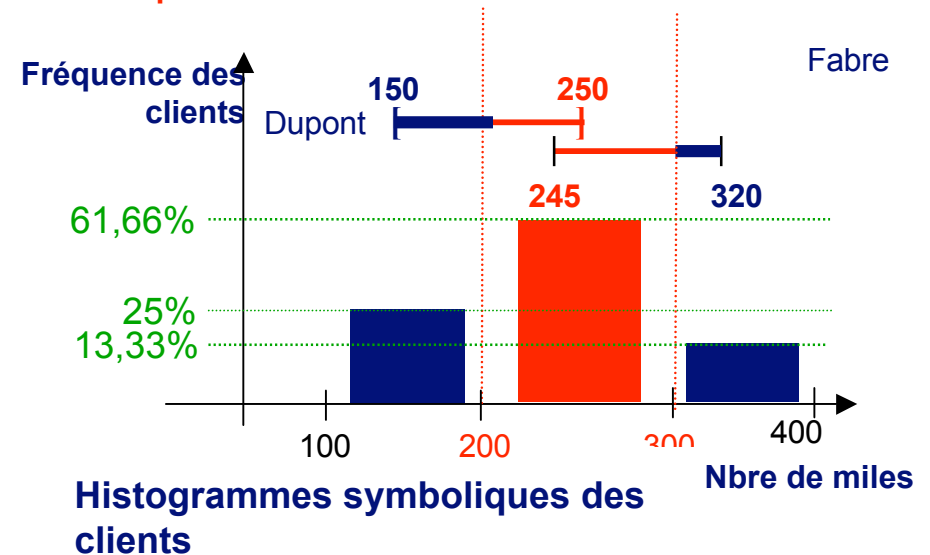
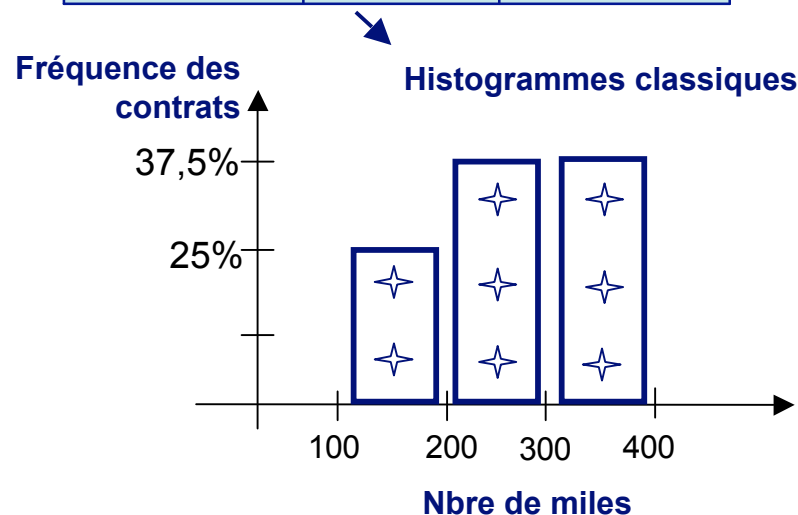
Consommation	Client concerné	Nbre de miles
Achat 1	M. Dupont	150
Achat 2	M. Dupont	180
Achat 3	M. Dupont	250
Achat 1	Mme Fabre	270
Achat 2	Mme Fabre	245
Achat 3	Mme Fabre	310
Achat 4	Mme Fabre	320
Achat 5	Mme Fabre	315

Client	Nbre de miles
M. Dupont	[150;250]
Mme Fabre	[245;320]

Client	Min	Max
M. Dupont	150	250
Mme Fabre	245	320

Données symboliques sur le concept « client »

codage classique



L'histogramme [200, 300] représente la somme des portions d'intervalles entrant dans chaque classe de l'histogramme. Il exprime le fait que pour M Dupont et Mme Fabre l'intervalle de consommation [ 200, 300] apparaît dans une plus grande proportion de leur consommation.

L'approche symbolique permet de construire un histogramme d'une variable à valeur intervalle et de conserver ainsi la variation.

# Perte d'information du codage classique de données symboliques

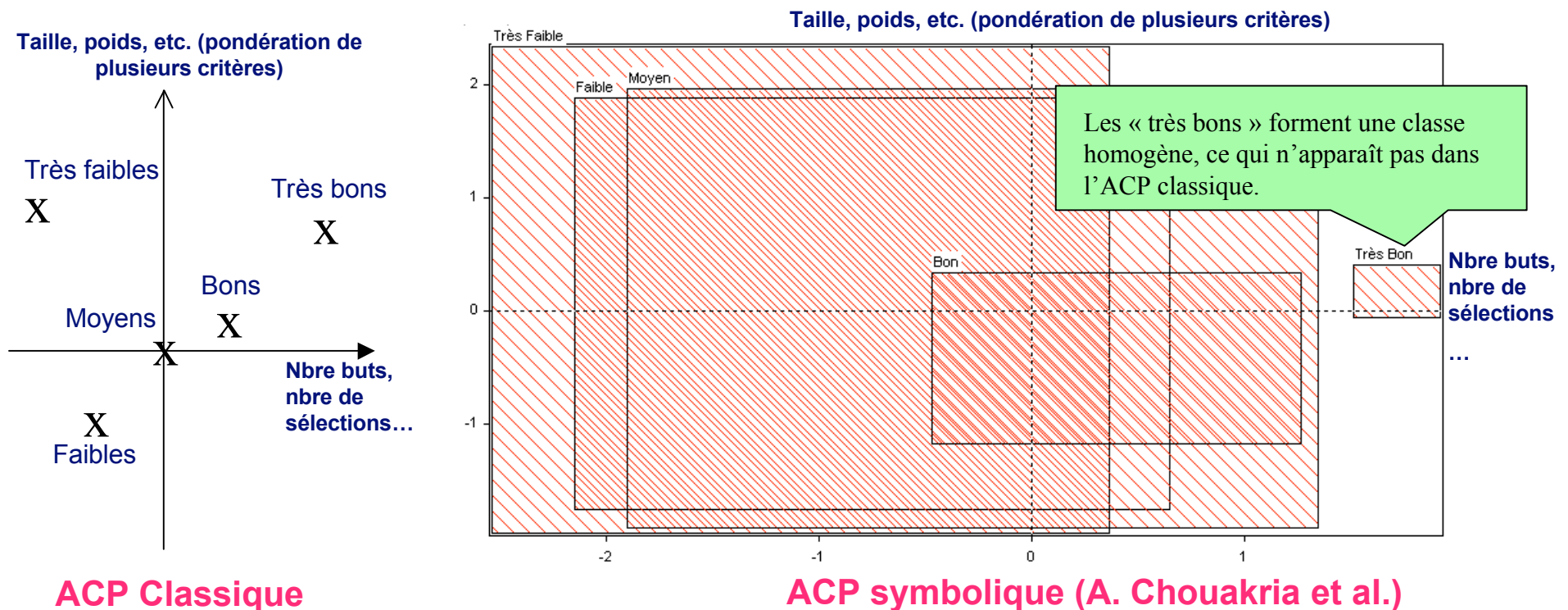
## EXEMPLE 3 : Analyse en composantes principales

En codage classique, chaque concept est représenté par un point

En codage symbolique :

→ chaque concept est représenté par une surface, ici un rectangle exprimant la variation du concept (de la valeur min. à la valeur max. prise par les individus inclus dans le concept).

→ Chaque concept peut être encore décrit par une conjonction de propriétés réduite aux axes factoriels retenus; ici : la taille, le p...



# Classique versus Symbolique : les données spatio-temporelles

Pour chaque pathologie et pour chaque hôpital, il existe en général plusieurs trajectoires de patient possibles

**Trajectoire A**  
Patient 1 urologie, hôpital Saint Louis

1. Visite au service d'urologie
2. Séjour en salle de repos
3. Séjour en salle d'opération
4. Séjour en salle de réanimation

**Trajectoire B**  
Patient 2 urologie, hôpital Saint Louis

- Visite au service d'urologie
- Visite en salle de radiographie
- Séjour en salle d'opération
- Séjour en salle de repos

**Trajectoire C...**  
Patient 3 urologie, hôpital Saint Louis



Chaque trajectoire peut être traitée comme un concept sur lequel sera réalisée l'analyse statistique. A chaque concept (trajectoire) peuvent être associées des variables : traitement suivi par les patients, nombre de jours d'hospitalisation, nbre de patients concernés...

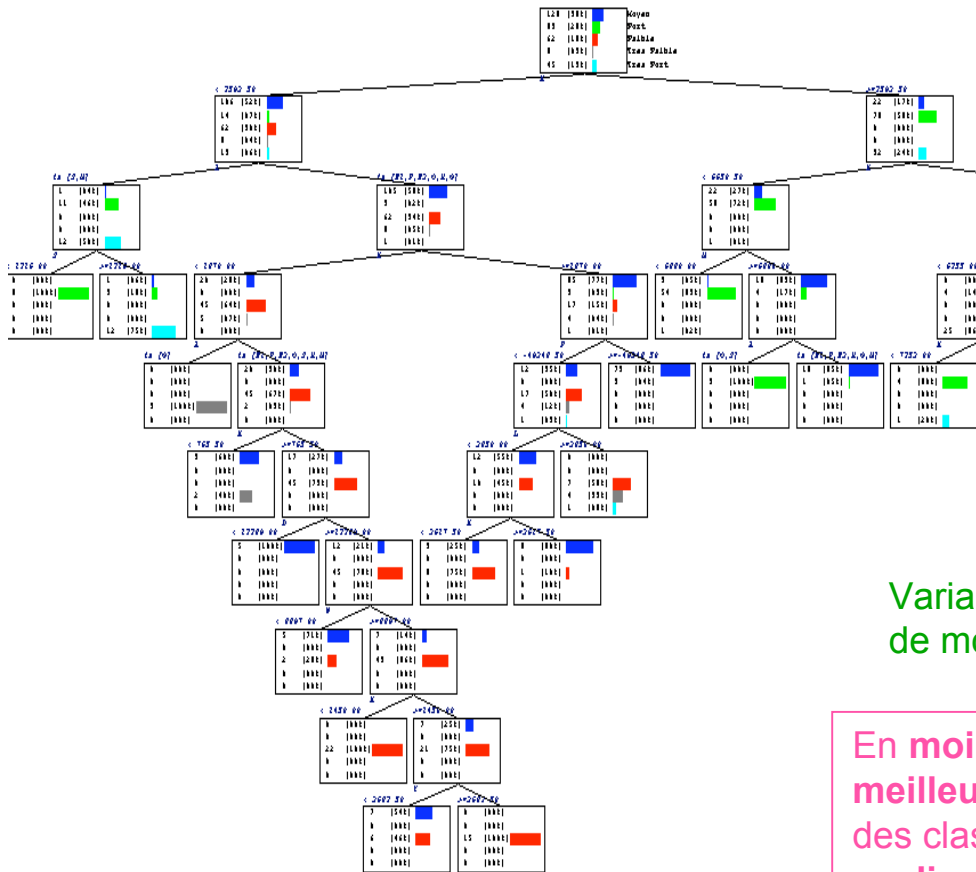
On peut aussi construire, à partir de milliers de trajectoires, des classes de trajectoires (rassemblant chacune des centaines de trajectoires). Ces classes seront les nouveaux concepts sur lesquels est effectuée l'analyse statistique.

**L'historique d'un patient/client/produit... est décrit par des propriétés qui généralisent tout ce qui s'est produit dans cette période.**

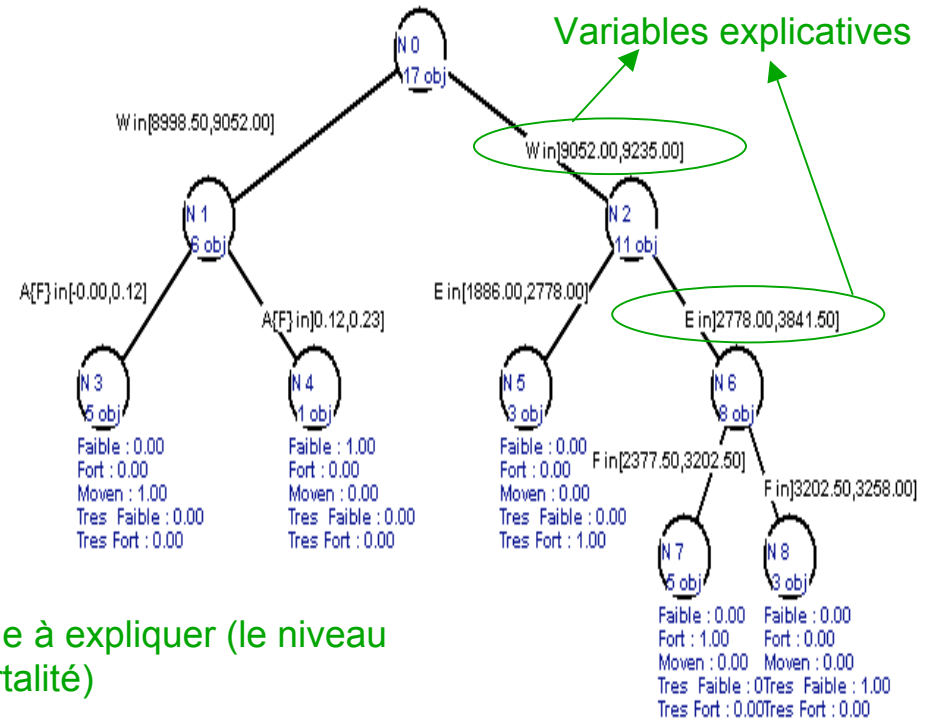
# Classique / symbolique : une comparaison

Arbres de décision établis sur 1000 données initiales (patients) que l'on veut regrouper en classes homogènes suivant une même trajectoire d'hospitalisation.  
 Variable à expliquer (ex. la mortalité) et des variables explicatives cliniques-biologiques.

Arbre « classique » sur les patients



Arbre « symbolique » sur les trajectoires



En moins de branches, moins de nœuds et avec une meilleure discrimination, l'arbre symbolique permet d'obtenir des classes de patients plus homogènes et clairement expliquées vis-à-vis de la variable « mortalité ».

# Données complexes versus symboliques: données structurées

Tableau classique

Foyer	Ville	Taille foyer	Localisation	CSP
Jones	Londres	2	Picadilly	3
Tom	Paris	5	Bercy	1
Bulle	Paris	3	La Défense	2

Description symbolique de Londres par les foyers


Ville	Taille foyers	Localisation	CSP
Londres	[1;8]	Picadilly(43%);....	

Tableau classique

École	Ville	Statut
Sherry	Londres	Privé
Laplace	Paris	Public
Welcome	Londres	Public

Description symbolique de Londres par les écoles

Ville	Statut	Spécialisation	
Londres	{{(privé, 37%);(public, 63%)}}	{{(oui, 17%); (non, 83%)}}	


## Concaténation

Londres = [caractéristiques des foyers]  $\wedge$  [caractéristiques des écoles]

# Données complexes versus données symboliques

Au départ chaque case contient un objet complexe

Des catégories aux concepts: chaque case contient un ensemble d'objets complexes

	Catégo	Image	Texte	Séqu.
i1	$C_j$		doc1	agbd
---	-----	-----	-----	c
in	$C_k$		docn	dgab

h

	Image	Texte	Séqu
$C_1$	{image}1	{doc}1	{gba}1
---	-----	-----	
$C_k$	{image}k	{doc}k	{ahd}k

Description d'objets complexes

Exemple:  $C_i = \text{images maritimes}$

	Catég	Image	Texte	Séqu.
i1				
---		-----	-----	-----
in				

	Image	Texte	Séqu
$C_1$			
---	-----	-----	
$C_k$			

Généralisation

Données Classiques

Données Symboliques

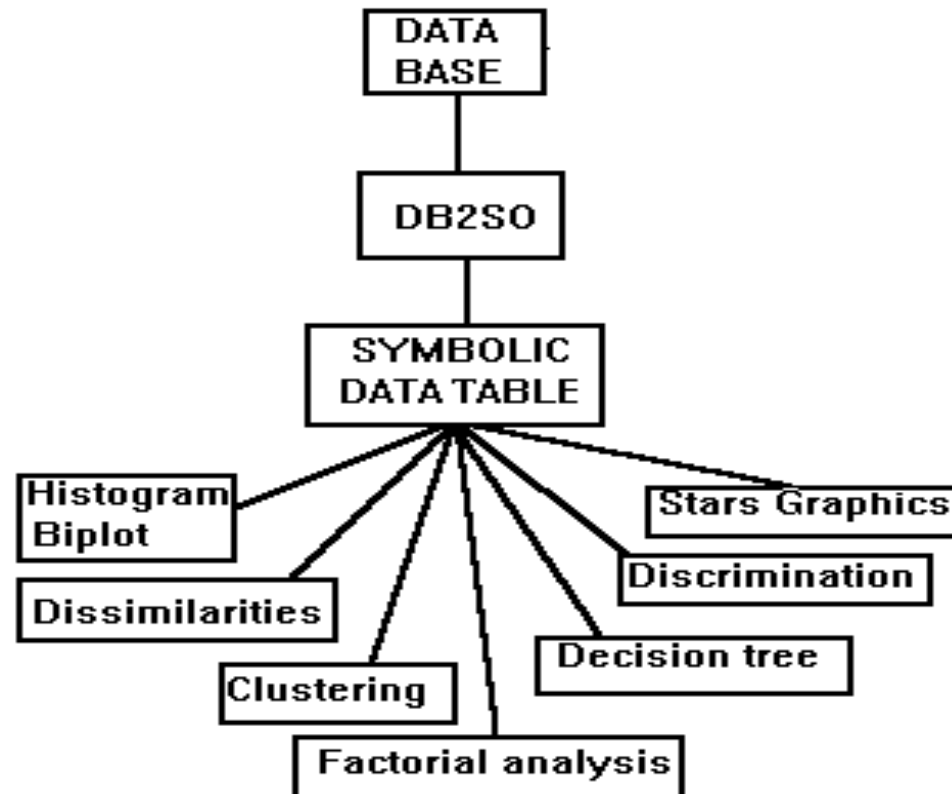
# ASSO-SODAS Consortium

Instituts Nationaux de Statistique:

INE (P), STATFI (FIN), EUSTAT (E) , ONS (UK)

- Universités: Namur(FUNDP-B), Napoli(DMS-I), Paris(DAUPHINE-F), Aachen (RWTH-D), Porto(FEP-P), Bari(DIB-I), Athens(UOA-GR), Recife (UFEP-BR)
- Centre de Recherche: INRIA (F),
- Companies: CISIA (F), TES (L), EDF, THOMSON (F)

# SODAS: de la BDR à l'ADS



# SODAS Software

The screenshot displays the SODAS software interface. On the left, a 'Methods' menu is open, showing a list of procedures: 'Sodas procedures', 'FDA', 'Factorial Discriminant Analysis', and a grid of other methods including SOE, DIV, STAT, DKS, DI, PCM, FDA, TREE, DSD, SDT, DIM, and PYR. A yellow callout bubble labeled 'Menu' points to the menu header, and another labeled 'Methods' points to the grid of procedure names.

The main window shows a 'chaining' diagram for a file named 'ROTACOE.SDS'. The diagram is a vertical flowchart starting with a blue 'BASE' box, followed by five red boxes numbered 1 to 5: 'SOE' (Symbolic Object Editor), 'DIM' (Dissimilarity similarity matrix), 'STAT' (Histogram, Elementary Statistics), 'PYR' (Pyramides), and 'FDA' (Factorial Discriminant Analysis). Each step includes a document icon and a monitor icon. The flow ends with a blue 'END' box. A yellow callout bubble labeled 'Sds file' points to the 'BASE' box, 'Report' points to the document icons, and 'Graphs' points to the monitor icons.

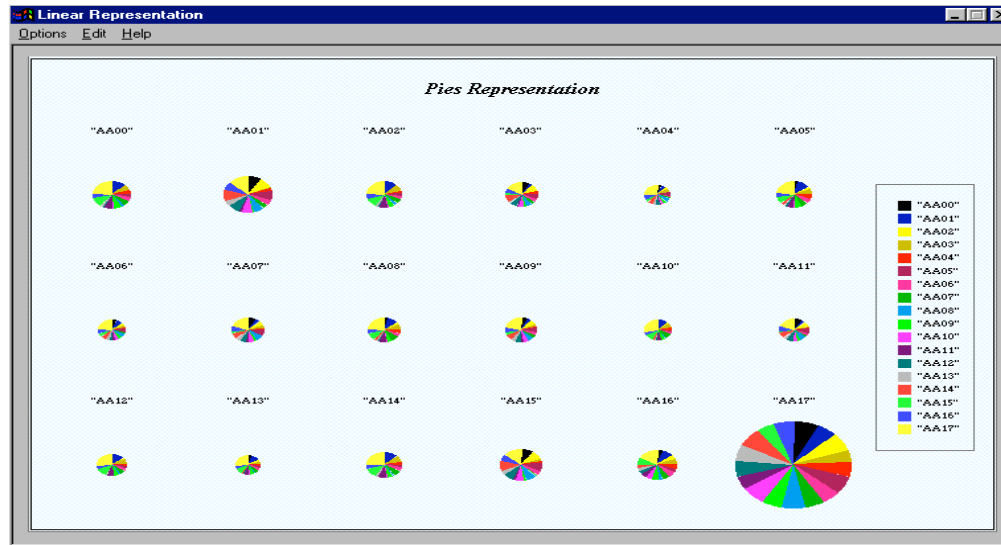
A yellow callout bubble labeled 'Chaining' points to the vertical flow of the diagram.

# **Extension de Méthodes classiques aux concepts: 3 livres 1 Springer, 2 Wiley (2008)**

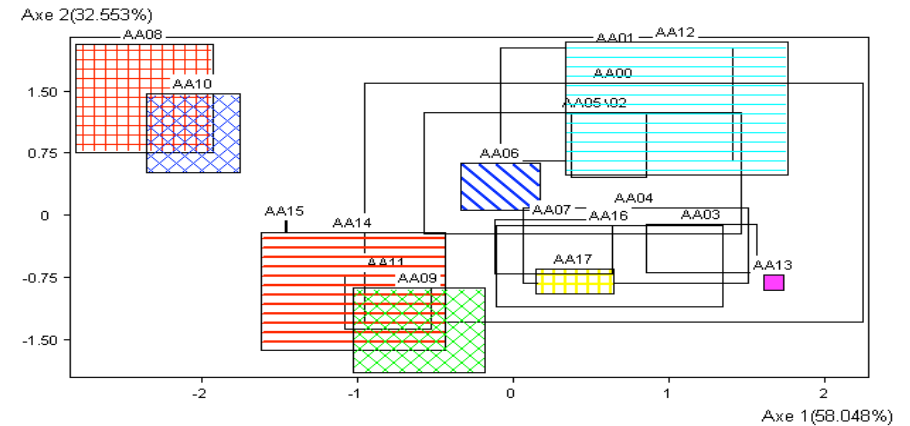
- Histos, correlation
- Visualisation en Etoiles
- Biplots
- ACP, AFC
- Décomposition de mélanges
- Multidimensional Scaling
- Typologie (Nuées Dynamiques , Pyramides)
- Régression
- Réseaux neuronaux, Cartes de Kohonen
- Arbres de Décision
- Extraction de Règles
- Treillis de Galois
- etc.

# Autres exemples de méthodes de SODAS

## CARTE DE KOHONEN DE CONCEPTS



## ANALYSE FACTORIELLE: ACP



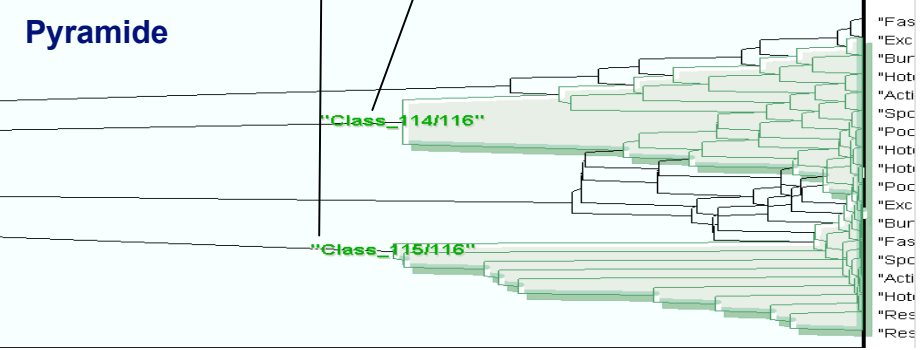
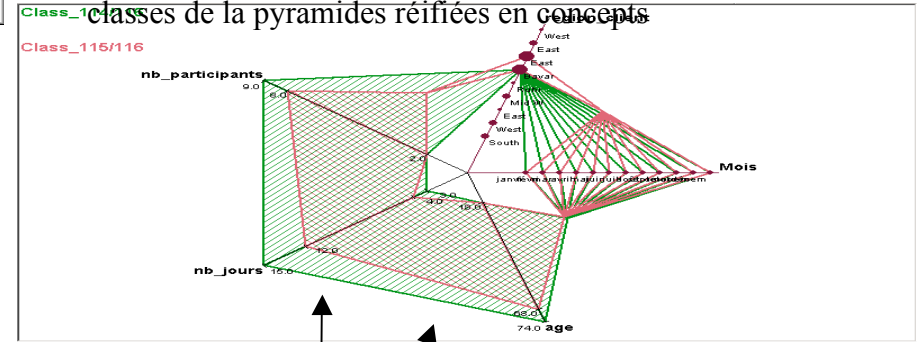
Superposition de deux deux étoiles associées à deux classes de la pyramides réifiées en concepts

## Méthode DIV (division en classes de concepts homogènes et description symbolique de ces classes réifiées en concepts)

```

- Ng <-> yes and Nd <-> no

+---- Classe 1 (Ng=1)
!----4- [intervallePrice <= 66.250000]
!
+---- Classe 5 (Nd=5)
!
!----3- [intervallePrice <= 110.000000]
!
+---- Classe 4 (Nd=4)
!
!----1- [intervallePrice <= 140.000000]
!
+---- Classe 2 (Ng=2)
!----5- [nb_participants <= 3.750000]
!
+---- Classe 6 (Nd=4)
!
!----2- [intervallePrice <= 231.750000]
!
+---- Classe 3 (Nd=2)
    
```



# MODELISATION PROBABILISTE

**Cas Classique :** Les variables sont des variables aléatoires à valeur quantitative ou qualitative.

## - CAS SYMBOLIQUE:

Les variables sont des Variables Aléatoires à valeur

- . Variable aléatoire
- . Loi de probabilité
- . Fonction de répartition
- . Diagramme
- . Intervalle inter-quartile
- . Suite de valeurs (ord,nom,quantitatives)

## ASSURANCES SOCIALES (MSA)

INDIVIDUS	CONCEPTS	$y$		$z$
Dispensation D	Bénéficiaire	Spéc. Médicale	Montant Remboursé	Taux de remb
D11	Ben1	6	1500	100
D12	Ben1	6	200	35
D13	Ben1	2	819	50
D21	Ben2	1	1800	10
D22	Ben2	5	300	25

CONCEPTS	$Y$		$Z$
Bénéficiaire	Spec. Médicale	Montant Remboursé	Taux de remb
Ben 1	$X^1_M$	$X^1_G$	$X^1_T$
Ben 2			
Ben n	$X^n_M$	$X^n_G$	$X^n_T$

# PARTITIONNEMENT D'UN ENSEMBLE DE CONCEPTS

Y. Lechevallier, F. De Carvalho, R.  
Verde

## CLASSIQUE

Moyennes

K-means

Dissimilarité standard

(Euclidienne,  
KHI2...)

## SYMBOLIQUE

Prototypes (= Obj Symb)

Nuées Dynamiques

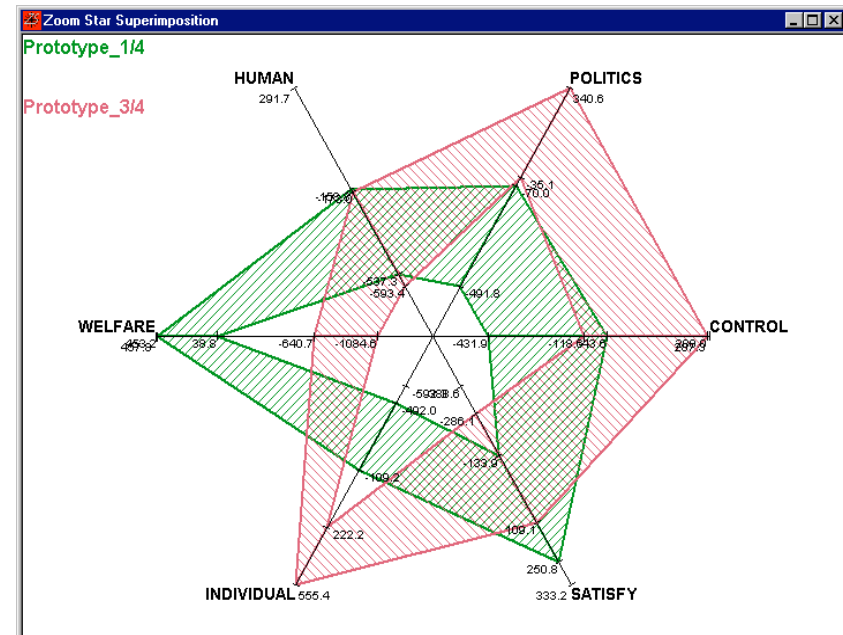
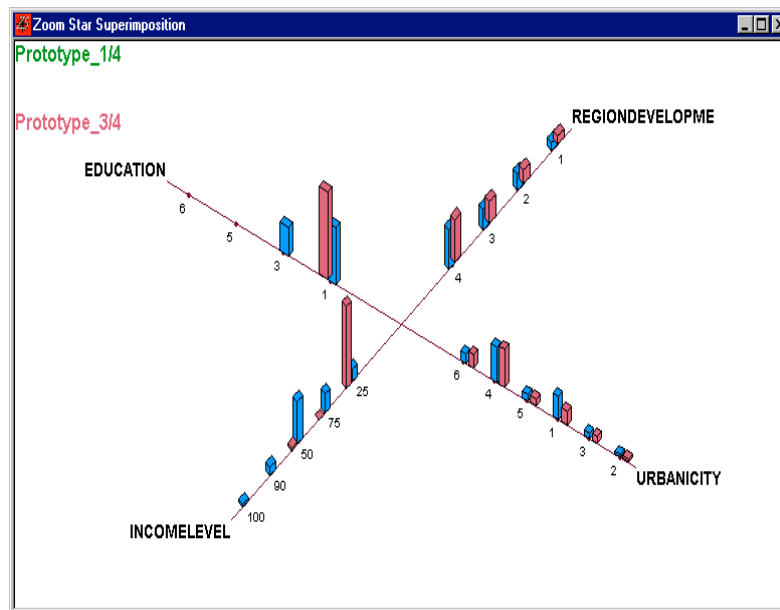
Dissimilarité symbolique

(Hausdorff, Ichino, De  
Carvalho,...)

# Représentation graphique des prototypes

M. Noirhomme et al.

## Comparison entre les classes 1 et 3



# Extension de l'algorithme Apriori au cas des données symboliques

Afonso et Diday (2004), Afonso (2005)...

- Découvrir des règles au niveau des concepts.
- Cas du panier de la ménagère: passer des bons d'achats au Concepts clients
- Exemple avec **3 items**: v: viande, p: poisson, c: céréales et une variable **montant** de la transaction

Transaction	Client	Item=Y	Montant=X
T1	c1	v	50
T2	c1	v,p,c	70
T3	c1	v,p,c	90
T4	c1	v	60
T5	c2	v,p	60
T6	c2	v,p,c	90
T7	c2	v	60
T8	c3	v,p	55
T9	c3	v	100
...	...	...	...

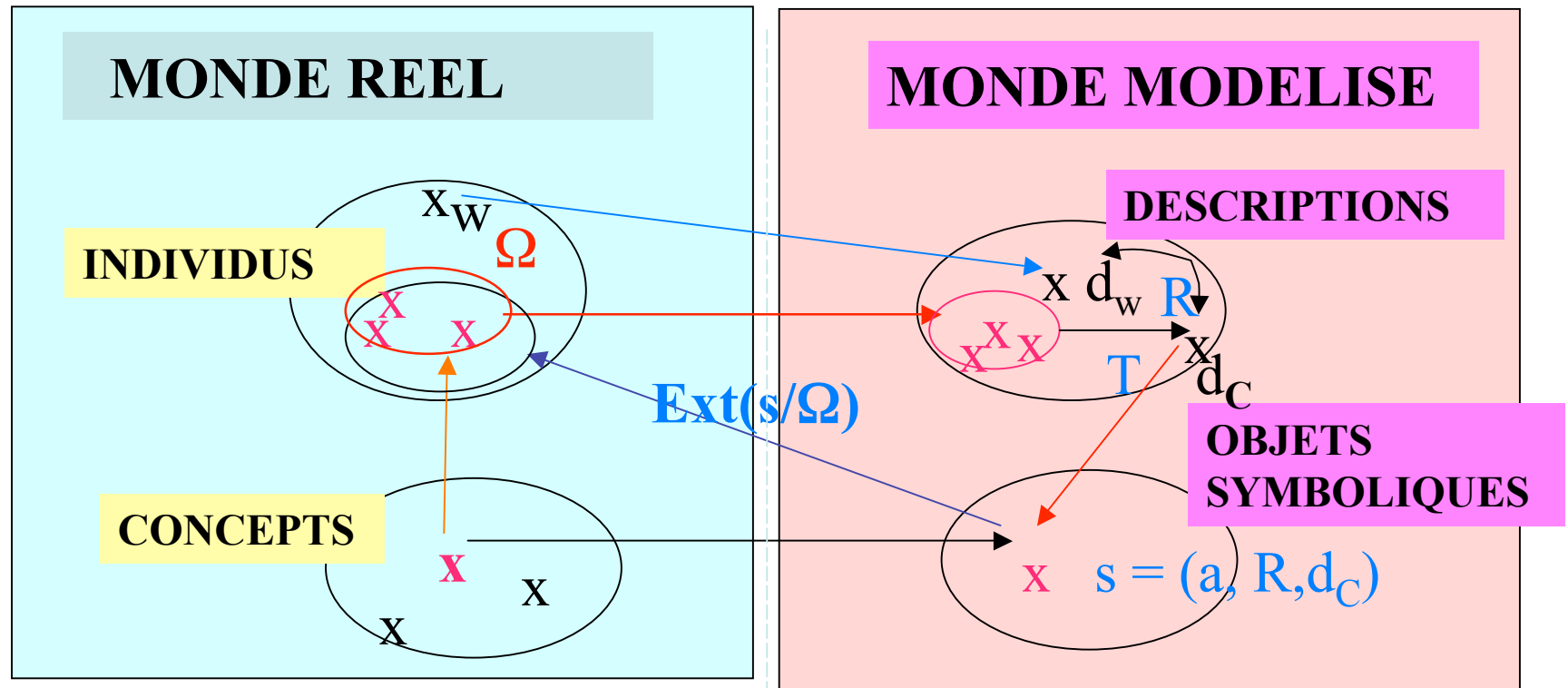
Matrice de données pour l'apriori classique



Clients	Items=Y	Montant=X
c1	1/2v,1/4p,1/4c	[50,90]
c2	2/5v,2/5p,1/5c	[60,90]
c3	3/4v,1/4p	[55,100]
...	...	...

Matrice symbolique où les concepts sont les clients décrits par une variable à valeurs diagrammes et une variable intervalle.

# MODÉLISATION ET APPRENTISSAGE DES CONCEPTS PAR QUATRE ESPACES



**Exemple:**  $\Omega$  : base de données décrivant des oiseaux, contenant 3 autruches. concept= les autruches,  $d_w$ : description d'un oiseau.  $d_c$ : description des trois autruches obtenue grâce à l'opérateur de généralisation  $T$ .  $R$ : relation binaire exprimant l'adéquation entre  $d_w$  et  $d_c$ .  $a$ : fonction d'appartenance d'un individu à un concept. L'extension de l'objet symb  $s$  dans  $\Omega$  entraîne 2 espèces d'erreurs.

Apprentissage des opérateurs par l'amélioration de la qualité de l'adéquation entre l'extension du concept et celle de l'objet symbolique qui le modélise.

# CONSTRUCTION D'UN OBJET SYMBOLIQUE POUR MODÉLISER UN CONCEPT

IL FAUT:

→ **un opérateur de généralisation T**

Exemple: T-norme, possibilités, capacités

La capacité de deux concepts  $C = (C_1, C_2)$  de satisfaire l'événement A

$$CAP(C, A) = \text{Prob}([X_1 = A] \cup [X_2 = A]) = p_1 + p_2 - p_1 p_2$$

→ **un opérateur de comparaison R entre la description d'un individu et celle d'une classe.**

Exemple: Inclusion, Appariement, Probabilité conditionnelle (qu'un concept soit satisfait par un individu donné connaissant la probabilité a priori qu'un individu satisfasse au concept)

→ **un opérateur d'agrégation:** pour agréger les résultats des comparaisons pour chaque variable.

→ Exemple: produit, copules...

## DEUX TYPES D'OBJETS SYMBOLIQUES

### OBJETS SYMBOLIQUES BOOLEENS

$S = (a, R, d1)$  modélise un concept  $C$  réifiant la catégorie employés x paysans.

$d1 = [18, 52] \times \{\text{employés, paysans}\} \longrightarrow$  par généralisation

$R = (\subseteq, \subseteq), \longrightarrow$  appariement

$a(w) = [\text{age}(w) \subseteq [18, 52] \wedge [\text{CSP}(w) \subseteq \{\text{employés, paysans}\}]]$

agrégation

$a(w) \in \{\text{VRAI, FAUX}\} \longrightarrow$  fonction de reconnaissance

## OBJETS SYMBOLIQUES MODAUX

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

$a(w) = [\text{age}(w) \mathbf{R}_1 [(0.2)[12, 20], (0.8) [20, 28]]] \wedge^*$

$[\text{SPC}(w) \mathbf{R}_2 [(0.4) \text{employee}, (0.6) \text{worker}]]$

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

$\Rightarrow R \rightarrow$  Appariement ,

Exemple: Paul Lévy, Hellinger, Kullback...

$R = (\mathbf{R}_1, \mathbf{R}_2) : r \mathbf{R}_i q = \sum_{j=1, k} r_j q_j e^{(r_j - \min(r_j, q_j))}$ .

$\wedge^* \rightarrow$  Agrégation, copules

# EXTENSION D'UN OBJET SYMBOLIQUE

## CAS BOOLEEN :

$$\mathbf{EXT(s) = \{w \in \Omega / a(w) = \mathbf{VRAI}\}.$$

## CAS MODAL

$$\mathbf{EXT}_{\alpha}(S) = \mathbf{EXTENT}_{\alpha}(a) = \{w \in \Omega / a(w) \geq \alpha\}.$$

## INTERÊT DE LA MODÉLISATION D'UN CONCEPT PAR UN OBJET SYMBOLIQUE

- RÉUTILISER LE CONCEPT SUR UNE AUTRE BASE,
- IDENTIFIER UN INDIVIDU DE SON EXTENSION,
- AMÉLIORER PAR APPRENTISSAGE SA MODÉLISATION,

### RÉDUIRE LES DONNÉES

- MASSIVES,
- MANQUANTES
- LA CONFIDENTIALITÉ

EN SE DONNANT LA POSSIBILITÉ DE LES RETROUVER  
PAR CALCUL D'EXTENSION.

# Treillis de Galois Stochastiques

- C'est la structure naturelle des objets symb car ils représentent de façon cohérente les intension et extension des objets symboliques dits complets
- Objets symboliques complets: l'intension de l'extension est la même intension.
- Théorème: A mesure que la connaissance s'améliore le treillis stochastique de concepts se stabilise et converge

CRAS (1998) Diday Emilion (Choquet)

## Tableau de données symboliques

Y1	Y2	Y3	W1 {a, b}	{g}	W2	{g,
			$\emptyset$			
	$\emptyset$		$\emptyset$			

## Objets symboliques induits du Treillis de concepts de concepts

$$s_2 : a_2(w) = [y_2(w) \subseteq \{e\}] \wedge [y_3(w) \subseteq \{g, h\}],$$

$$\text{Ext}(s_2) = \{1, 2, 4\}$$

$$s_3 : a_3(w) = [y_1(w) \subseteq \{c\}],$$

$$\text{Ext}(s_3) = \{2, 3\}$$

$$s_4 : a_4(w) = [y_1(w) \subseteq \{a, b\}] \wedge [y_2(w) = \emptyset]$$

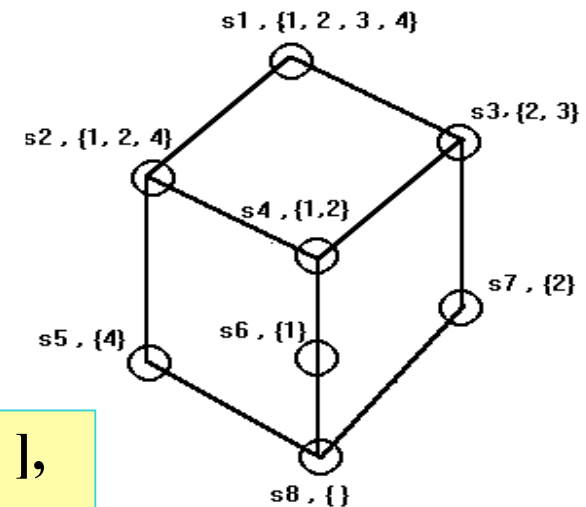
$$\wedge [y_3(w) \subseteq \{g, h\}],$$

$$\text{Ext}(s_4) = \{1, 2\}$$

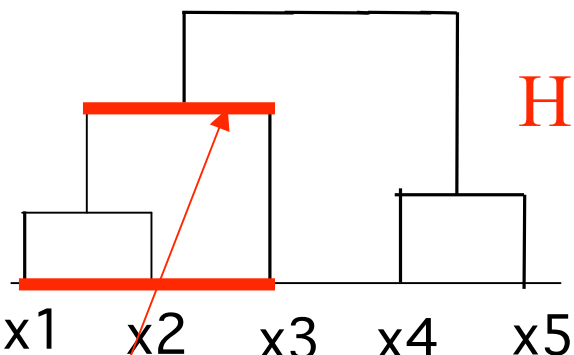
$$s_5 : a_5(w) = [y_2(w) \subseteq \{e\}] \wedge [y_3(w) \subseteq \{h\}],$$

$$\text{Ext}(s_5) = \{4\}$$

Treillis de Galois issu du tableau de données symboliques dont les unités sont des concepts



# QUALITE DE LA REPRESENTATION SPATIALE

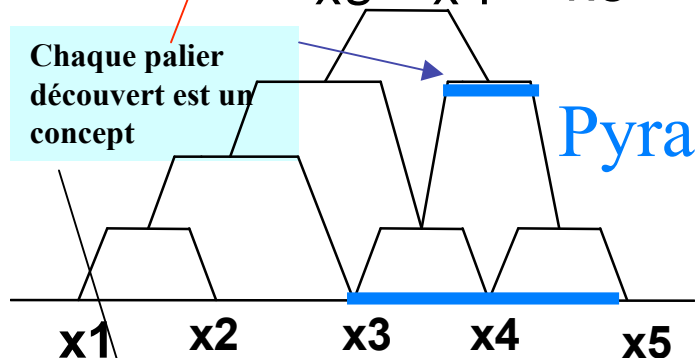


Hierarchies

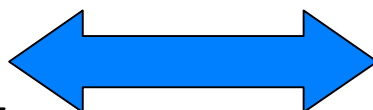


Ultrametric  
dissimilarities = U

Chaque palier  
découvert est un  
concept

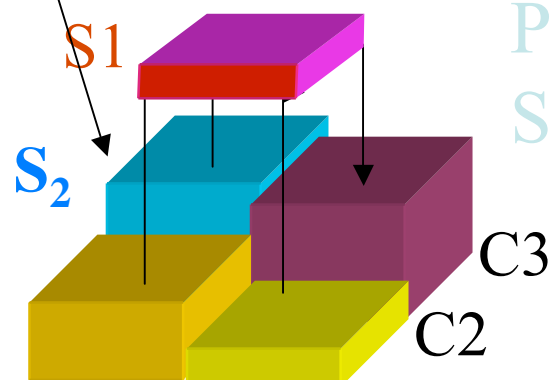


Pyramides

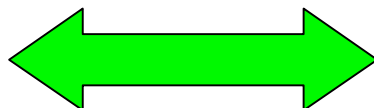


$$W = |d - U|$$

Robinsonian  
dissimilarities = R



Pyramides  
Spatiales



$$W = |d - R|$$

Yadidean  
dissimilarities = Y

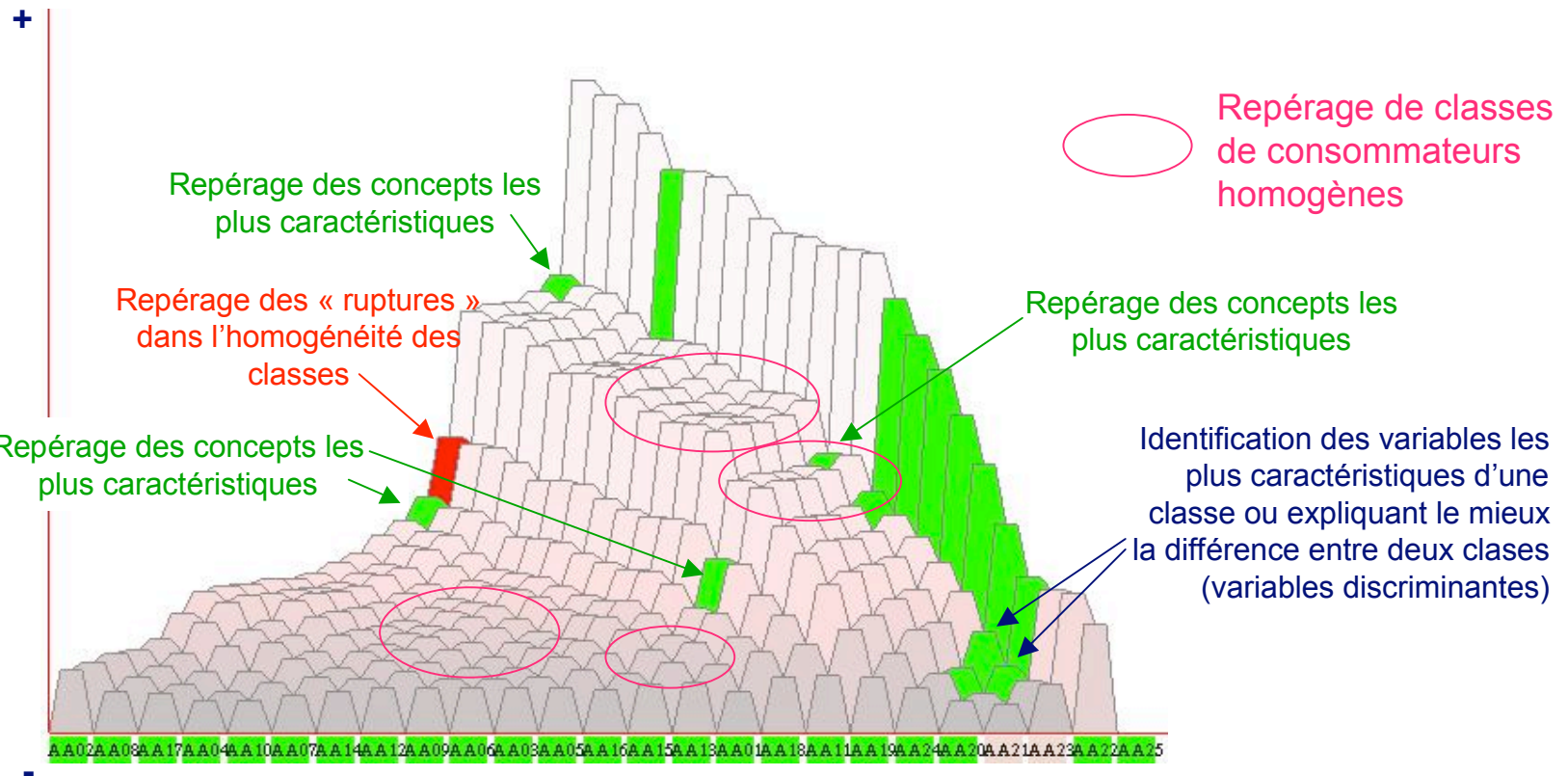
A 1 B 1 C 1

$$W = |d - Y|$$

## Exemple d'application de nos méthodes (K. PAK, M. Rahal) :

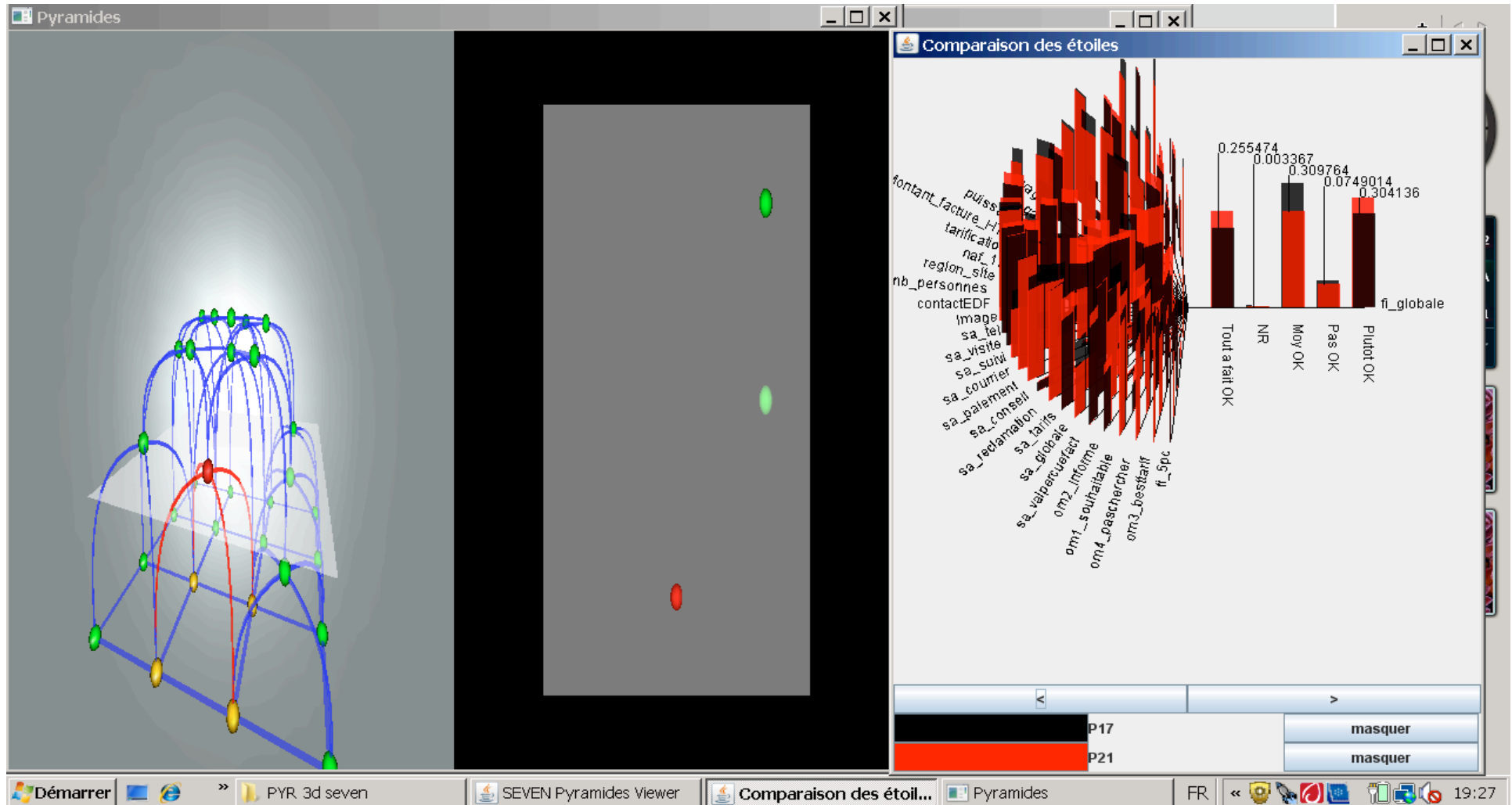
### Exemple 2 : Pyramide de la consommation d'énergie par les foyers en Angleterre

Degré de dissimilarité



- L'analyse va pouvoir se focaliser sur les groupes de classes de consommateurs les plus caractéristiques (application de toutes les autres méthodes d'analyse sur ces données, reconstitution d'une pyramide à partir des nouvelles classes sélectionnées...).
- De nouvelles segmentations de clients pourront être faites.
- Des ruptures seront repérées et devront être expliquées.
- Une nouvelle classe ajoutée a posteriori sera automatiquement placée auprès de celles qui lui ressemblent le plus.

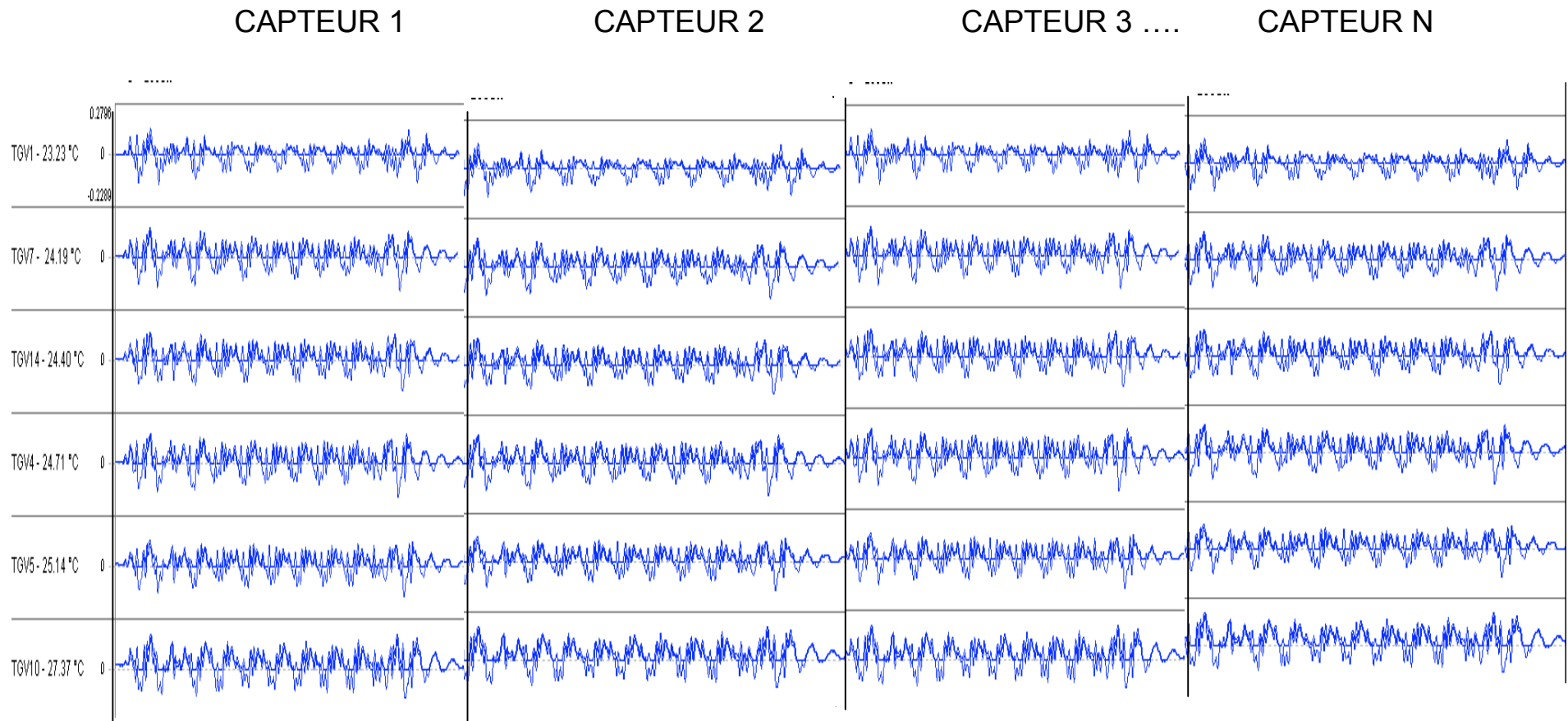
# Classification Spatiale



Réalisé dans le cadre de l'ANR SEVEN (EDF, LIMSI, Dauphine).

Théorie de la classification spatiale: E. Diday (2008) "Spatial classification". DAM (Discrete Applied Mathematics) Volume 156, Issue 8, Pages 1271-1294.

# Détection d'anomalies sur des ouvrages publics (LCPC) Laboratoire Central Des Ponts et Chaussées



**Chaque ligne représente un TGV passant sur un pont à une certaine température: Chaque case du tableau contient jusqu'à 800.000 valeurs**

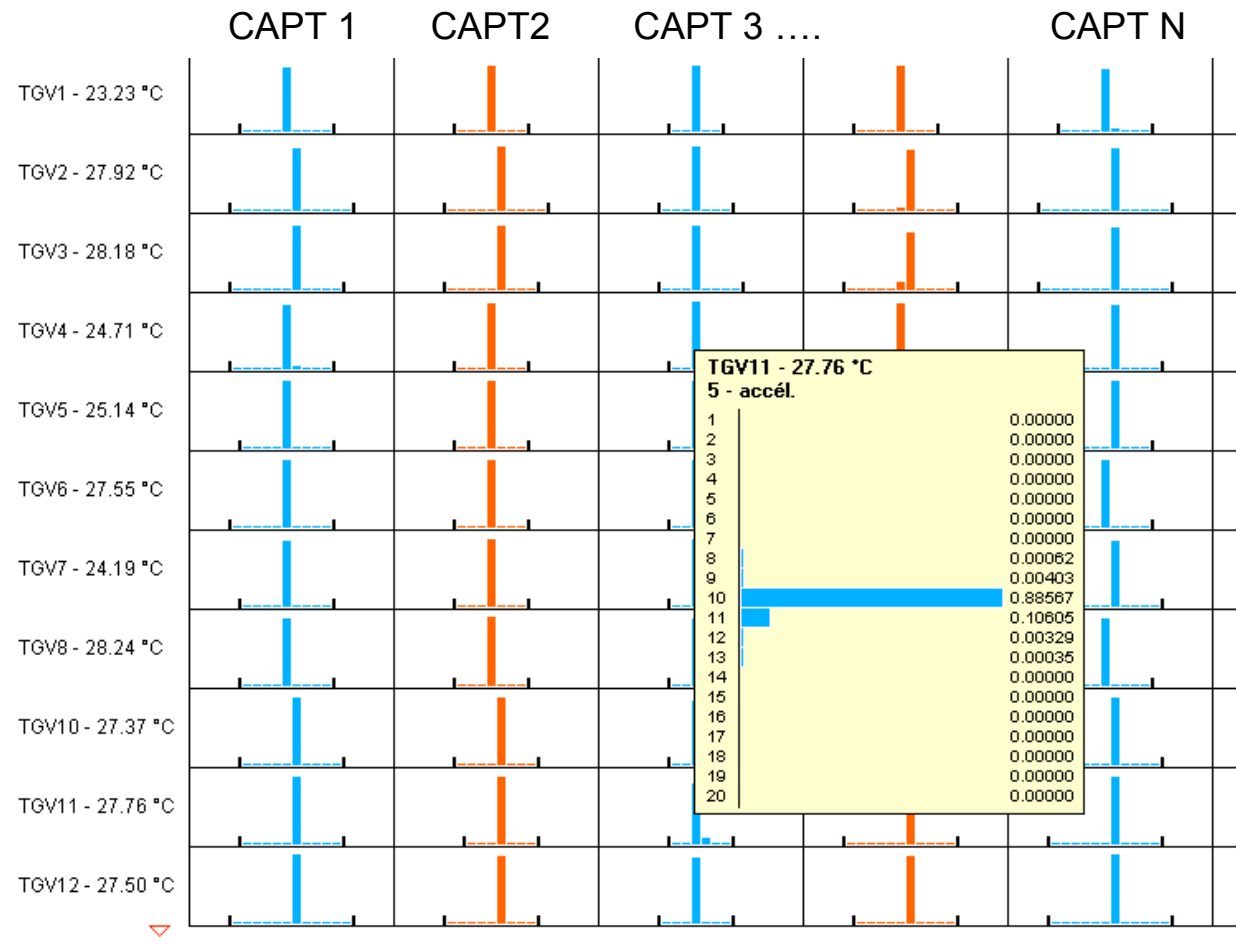
# Construction du Tableau de données symboliques

Transformation des Signaux en histogrammes

- Par Projection sur l'axe des ordonnées
- Par Ondelettes

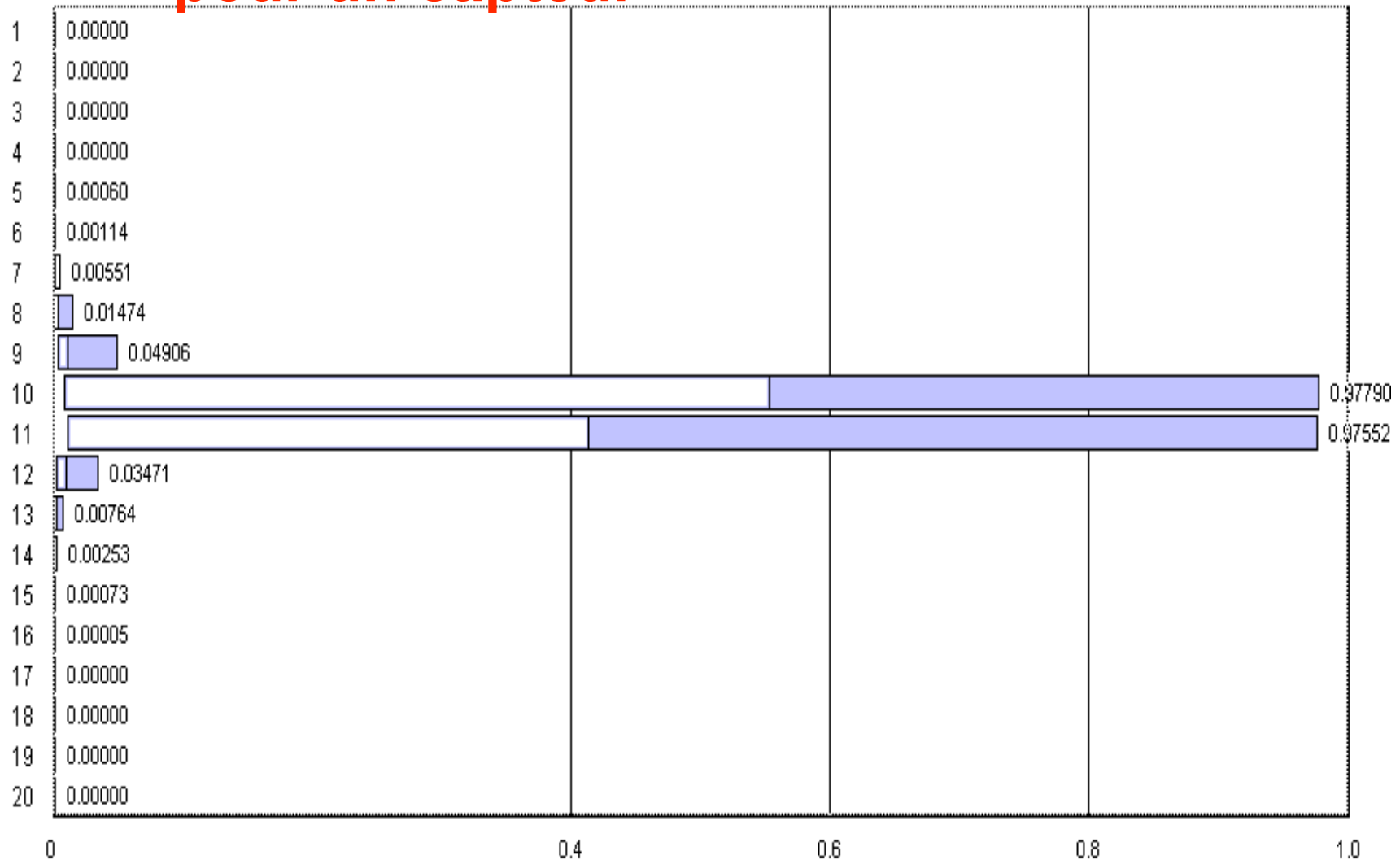
**Il en résulte un tableau de données dont chaque case contient un histogramme à 20 intervalles.**

# Construction du Tableau de données symboliques

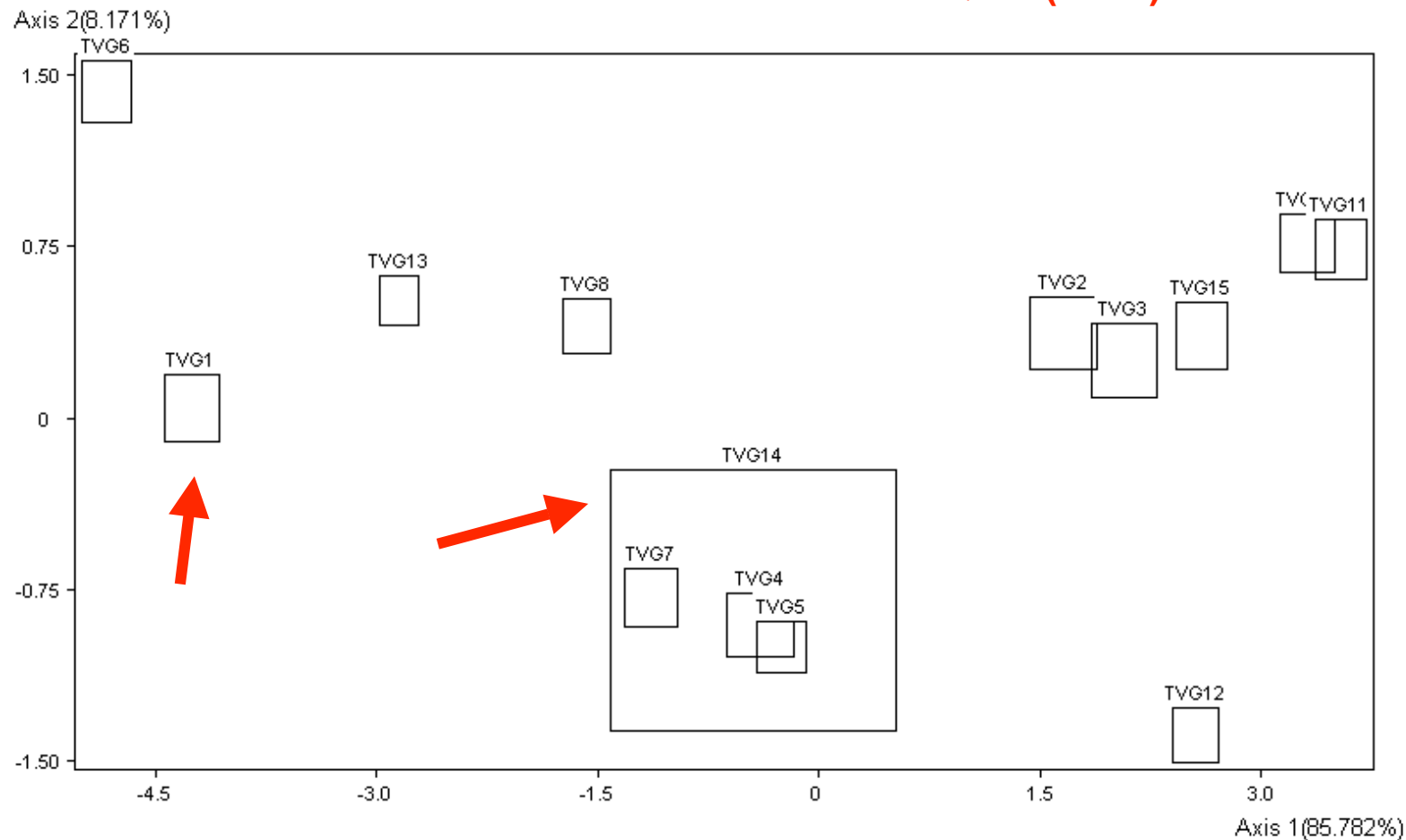


**Tableau de données symboliques dont chaque case contient un histogramme à 20 intervalles représentant chaque signal pour chaque capteur.**

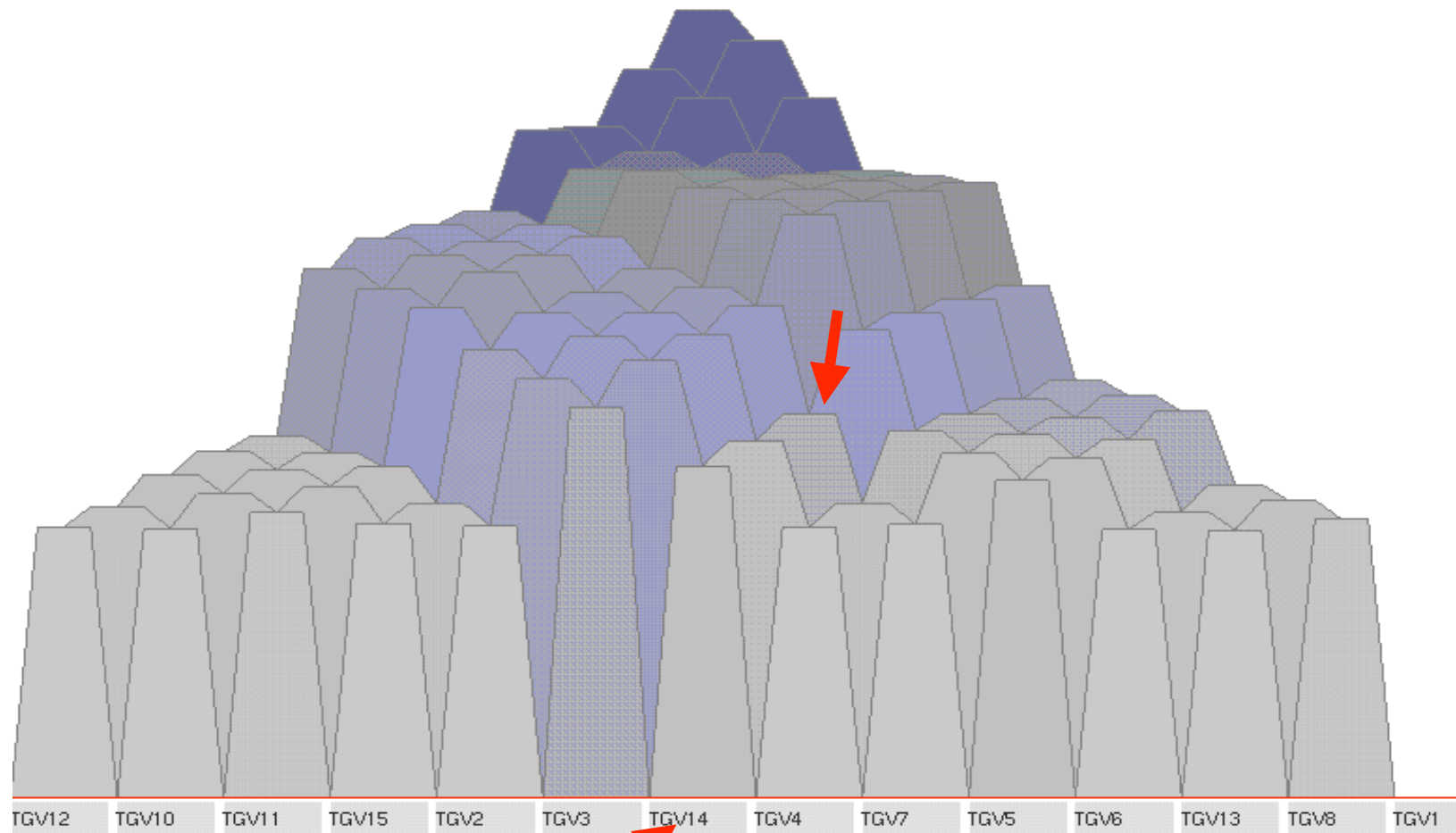
# Variation des fréquences des intervalles des histogrammes pour un capteur



## ANALYSE FACTORIELLE SYMBOLIQUE (ACP)

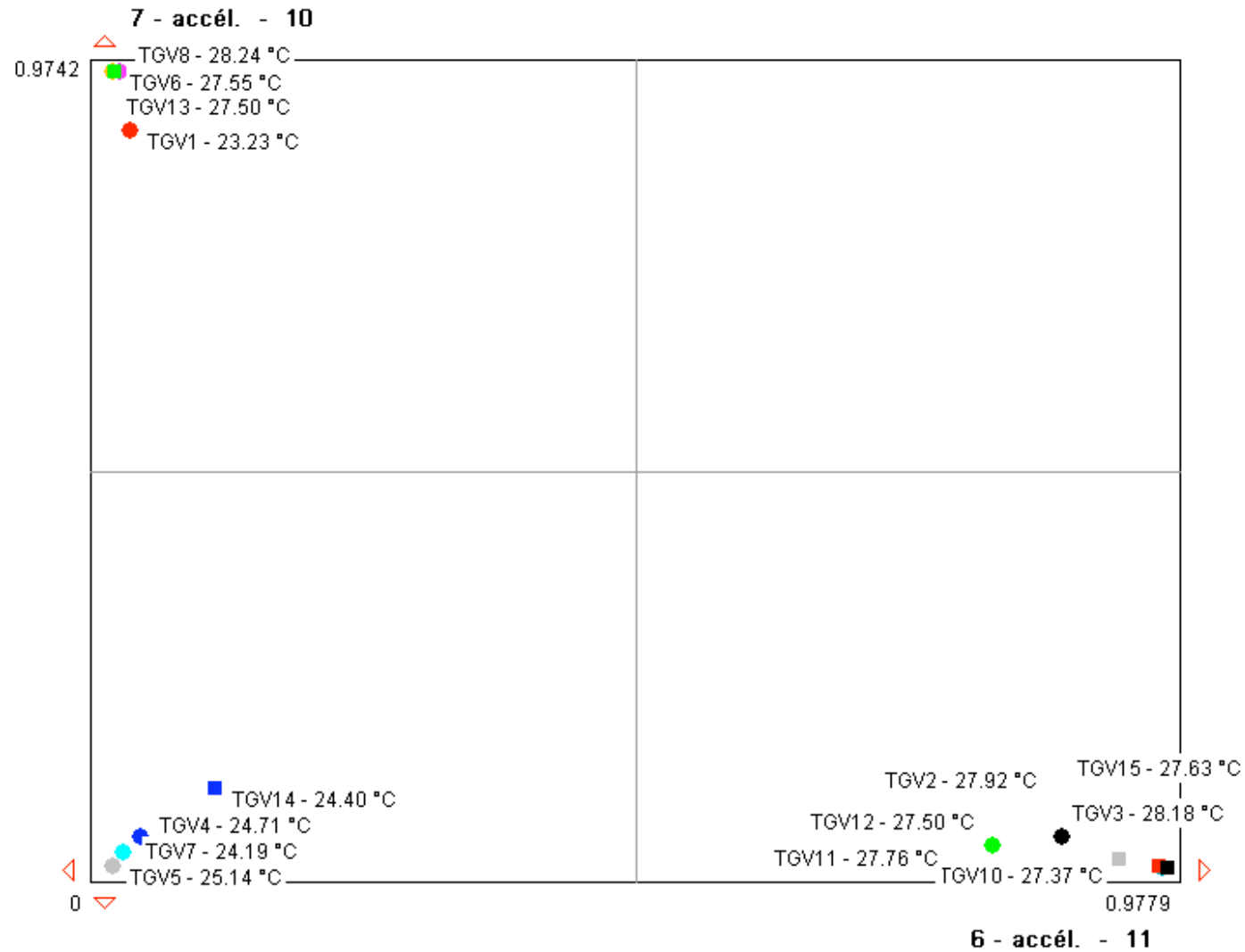


ACP Symbolique appliquée au tableau des intervalles interquartiles. Le TGV1 est en dehors de son groupe de température et le TGV 14 recouvre la classe des basses températures.



- La pyramide symbolique fait apparaître:
- 1) le TGV1 en dehors de son groupe de température
  - 2) Le TGV 14 couvre tous les TGV de sa température.

# Réduction du nombre de variables pour l'identification



# Conclusion

L'ADS a permis de découvrir des groupes caractéristiques de TGV faisant ressortir des anomalies et de montrer que peu de capteurs suffisent à les distinguer.

# LIVRES consacrés à l'ADS

**SPRINGER, 2000 :**

**“Analysis of Symbolic Data”**

**H.H., Bock, E. Diday, Editors . 450 pages.**

**WILEY, 2006**

**L. Billard , E. Diday “Symbolic Data Analysis, conceptual statistic and Data Mining”.[www.wiley.com](http://www.wiley.com)**

**WILEY, 2008**

**“Symbolic Data Analysis and the SODAS software.” 457 pages  
E. Diday, M. Noirhomme , ([www.wiley.com](http://www.wiley.com))**

# Articles de synthèse dans livres ou journaux

- E. Diday (2005) "Categorization in Symbolic Data Analysis". In « handbook of categorization in cognitive science ». Edited by H. Cohen and C. Lefebvre. Elsevier editor. <http://books.elsevier.com/elsevier/?isbn=0080446124>
- JASA (Journal of the American Statistical Association) "From the Statistic of Data to the Statistic of Knowledge: Symbolic Data Analysis". L. Billard, E. Diday June, 2003 .
- Diday E.(2000): "Un cadre théorique et des outils pour le Data Mining". Chapitre 1 de 90 pages, dans "Induction symbolique numérique à partir de données". Diday E., Kodratoff Y., Brito P., Moulet M. (eds) Cépadues. 31100 Toulouse. 442 pages.

# **DIFFUSION DE L'ANALYSE DES DONNEES SYMBOLIQUES**

## **UNIVERSITE DAUPHINE**

SITE SODAS CEREMADE PARIS Dauphine LISE: des mémoires d'étudiants expliquent et illustrent SODAS.

## **REVUE INTERNATIONALE D'Analyse de Données Symboliques:**

Electronical Journal of SDA (JSDA) at

[www.jsda.unina2.it/newjsda/volumes/index.htm](http://www.jsda.unina2.it/newjsda/volumes/index.htm)

**ENTREPRISE: « SYROKKO » (un vent nouveau ...)  
pour valoriser SODAS et l'ADS dans l'industrie**

[www.syrokko.com](http://www.syrokko.com)

**MORALITÉ:** dans votre travail vérifiez si vos unités d'étude sont des individus ou des concepts.

- Si ce sont des individus demandez-vous s'il n'y aurait pas des catégories d'individus (induits par des variables qualitatives intéressantes ou une typologie) à étudier en tant que concepts .

- Si ce sont des concepts pensez à prendre en compte leur variation interne (i.e. des individus de leur extension) pour les décrire par des variables symboliques munies de connaissances supplémentaires.

**PERSPECTIVES:** Le champs de recherche et d'application est immense puisqu'il faut tout reprendre en AD, STAT et Data Mining en pensant autrement, c'est à dire en termes de concepts et de données symboliques plutôt que d'individus décrits par des données classiques ou complexes: on manque de bras!

# CONCLUSION

Nous avons montré que la représentation des données et connaissances n'est pas seulement un domaine d'utilisation normal des outils standards de la Statistique, de la Fouille de Données (Data Mining) ou de l'Analyse des Données plus ou moins complexes, mais de plus, le fait de s'intéresser aux connaissances et aux concepts qui en forment les atomes en tant qu'unités d'étude remet totalement en cause ces outils et nécessite leur renouvellement complet aussi bien dans leur théorie que dans leur pratique et dans la façon de les penser.