DE LA STATISTIQUE DES DONNÉES

À LA

STATISTIQUE DES CONNAISSANCES:

L'ANALYSE DES DONNEES SYMBOLIQUES

E. Diday

Université Paris IX Dauphine

PLAN

- Données, connaissances et notre objectif
- Individus, catégories, concepts
- Des concepts aux données symboliques
- Pourquoi on ne code pas les données symboliques sous forme de données classiques ?
- Stratégie classique versus symbolique
- Des données complexes aux données symboliques
- Le logiciel SODAS issu de deux projets européens
- L'extension des méthodes classiques
- Les méthodes et problèmes spécifiques
- Conclusion: perspectives et moralité
- Références
- Diffusion

DONNEES , CONNAISSANCES

Définition des données:

Ce sont des grandeurs ou des qualités décrivant des entités du monde appelés individus.

Définition des connaissances:

Ce sont des informations d'ordre intensionnel (donc non réduites à des grandeurs ou des qualités) qui portent sur des entités du monde, appelées concepts et qui sont munies d'une extension.

Selon quel principe sont-elles construites?

« En s'arrachant hors de l'objet (qu'il soit individu ou concept) et hors de soi » (J.P. Sartre)

L’ OBJECTIF

Notre objectif est d’extraire des informations nouvelles sur des individus et des concepts au travers des données et connaissances qui les modélisent en un point de départ.

Bergson (La pensée et le mouvant (1934))

« Prendre des concepts déjà faits, les doser et les combiner ensemble jusqu’à ce qu’on obtienne un équivalent pratique du réel »

Un peu de vocabulaire

- Sujet: être libre, conscient, responsable
- Objet: chose du monde réel ou virtuel, le contraire du sujet
- Catégorie: modalité d'une variable qualitative
- Classes: partie d'un ensemble
- Concepts: entité munie d'une intension et d'une extension
- Réifier: c'est rendre « objet » ou « chose »
- Décrire: c'est ce qu'on dit d'un objet (ce n'est pas l'objet!)
- « décrire c'est détruire »
- « je est un autre » (Rimbaud).

Individus, variables

- Individus : unités statistiques de l'analyse des données
- Les Individus de premier ordre: ils réifient les sujets qui deviennent "objets" d'étude.
- Les individus de second ordre: ils réifient les catégories, classes et concepts
- Le variables: se sont les descripteurs des individus
- PAS de symétrie objets versus variables!!!
- Aristote (Traité des catégories): "un objet a une couleur mais une couleur n'a pas un objet!"

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), Arnault 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 GSM

Symbolique : des concepts

Espèces d'oiseaux

IRIS, Régions d'habitation

Joueurs d'une équipes (Marseille

Types d'image (marines,..

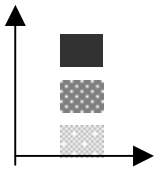
Magasins d'une chaîne

Usagers

Trajectoire dans les services et Hôpitaux

Bénéficiaires

Niveaux de consommation



STATISTIQUE INDIVIDUELLE VERSUS STATISTIQUE CONCEPTUELLE

La description d'un individu de la base de données:

La description d'un oiseau, d'un champignon, d'un habitant, d'une feuille de maladie se fait à l'aide de variables à valeur unique qualitative ou quantitative:

Taille = 20, Couleur = Rouge

La description d'un concept :

une espèce d'oiseau ou de champignon , un IRIS (INSEE), une trajectoire de patients, un bénéficiaire d'assurance, un type d'image…

doit tenir compte de la VARIATION des individus de son extension: les oiseaux de l'espèce, les habitants de l'IRIS, les patients ayant suivi la même trajectoire, les feuilles de maladies d'un bénéficiaire dans une période donnée….

Taille = [20, 30], Couleur = {0.3Rouge, 0.7 vert}+ règles, Taxonomies

LA REIFICATION DES CATEGORIES, CLASSES ET CONCEPTS EN INDIVIDUS D'ANALYSE DE DONNEES

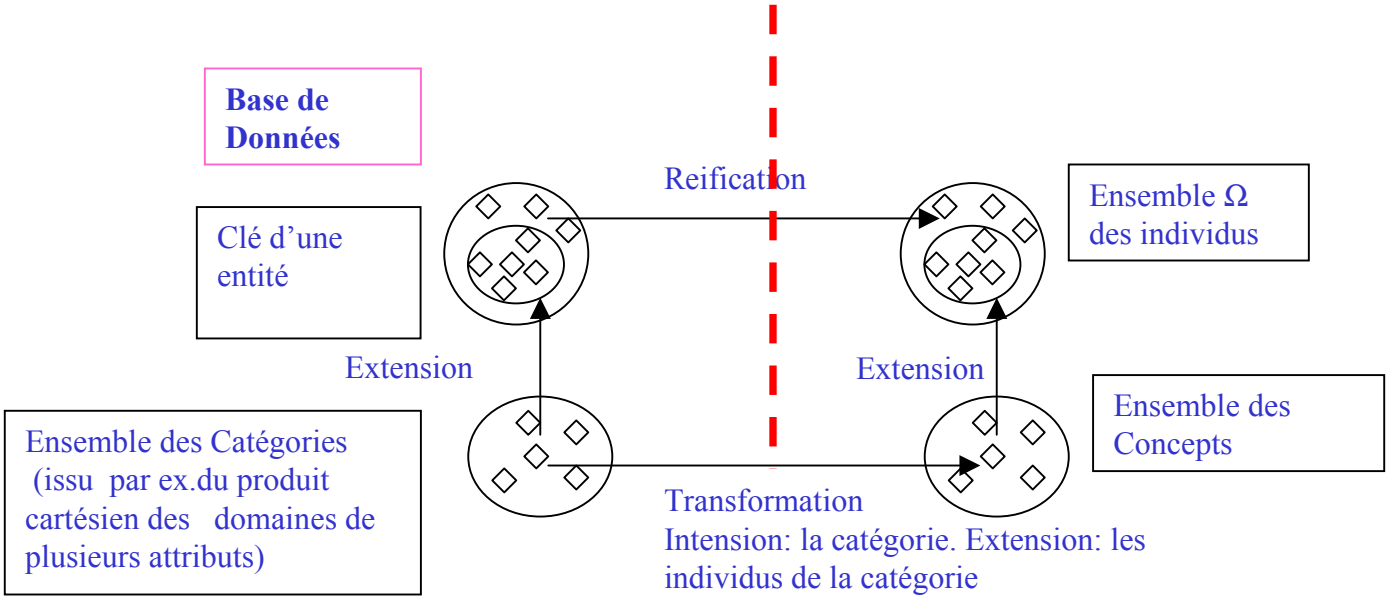
LES CATEGORIES SONT DEFINIES DE FACON EXHAUSTIVE PAR LES MODALITES D'UNE VARIABLE QUALITATIVE OU UN PRODUIT CARTESIEN DE TELLES VARIABLES.

LES CLASSES SONT DEFINIES PAR UNE CLASSIFICATION AUTOMATIQUE OU UNE CATEGORIE

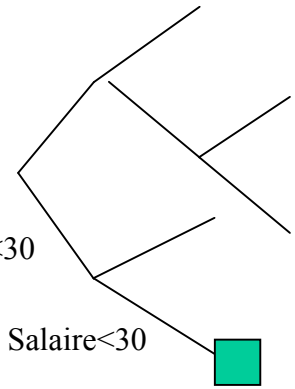
UN CONCEPT PEUT ETRE DEFINIS A PARTIR D'UNE CATEGORIE OU D'UNE CLASSE

LES **INDIVIDUS** QUI LES REIFIENT SONT CONSIDERES COMME DES OBJETS DONT LA DESCRIPTION N'EST JAMAIS EXHAUSTIVE.

LA TRANSFORMATION DE CATEGORIES EN CONCEPTS



- Exemple : les branches d'un arbre de décision constituent des catégories définies par des conjonctions de propriétés transformables en concepts qui peuvent être décrits par d'autres variables.



MODELISATION DES BENEFICIAIRES de L'ASSURANCE MALADIE SUR UNE PERIODE DONNEE

Individus Catégories

Occurrences	Bénéficiaire	AnnéeRembour	Prise Charge (type nominal)
111111	236	1996	21
111112	236	1996	31
111113	236	2002	31
111114	362	1995	1
111115	362	1996	21
111116	235	1994	1
111117	235	2000	31

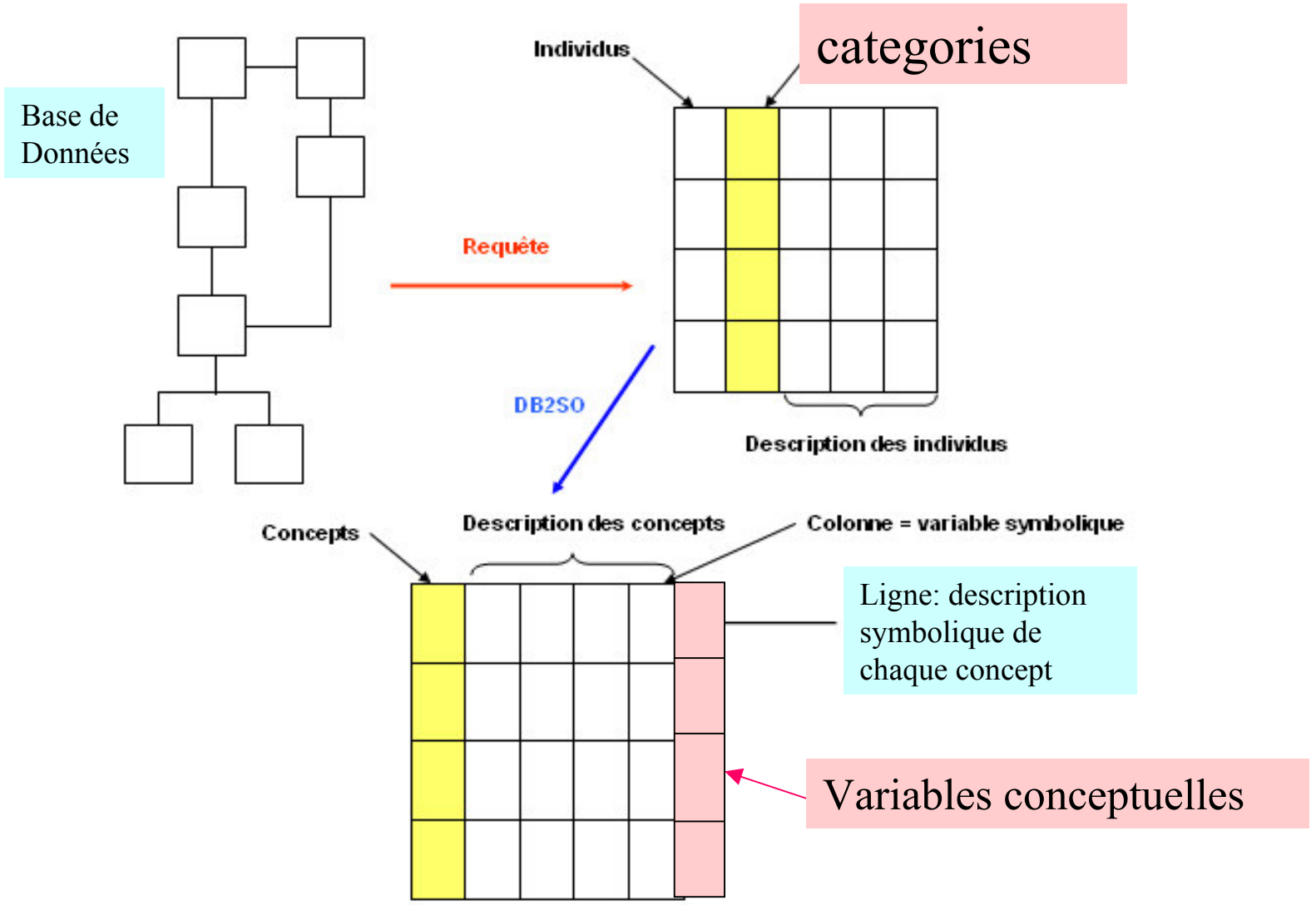
Généralisation

Concepts

Bénéficiaire	Année Rembour (intervalle)	Prise en Charge (diagramme)	Age
236	[1996,2002]	21(33.3), 31(66 ,6)	72
362	[1995,1996]	1(50%), 21(50%)	85
235	[1994,2000]	1(50%), 31(50%)	65

Age est une variable ajoutée liée aux concepts

DE LA BASE DE DONNEES AUX CONCEPTS



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

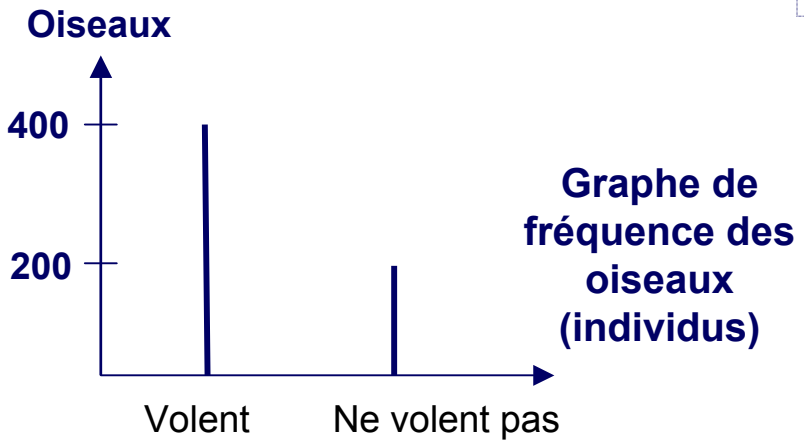
Espèce	Vole	Couleur	Taille	Migre
Hirondelle	Oui	0.3n,0.7gris	[60, 85]	Oui
Autruche	Non	0.1noir,0.9g	[85, 160]	Non
Pingouin	Non	0.5n,0.5gris	[70, 95]	Oui

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

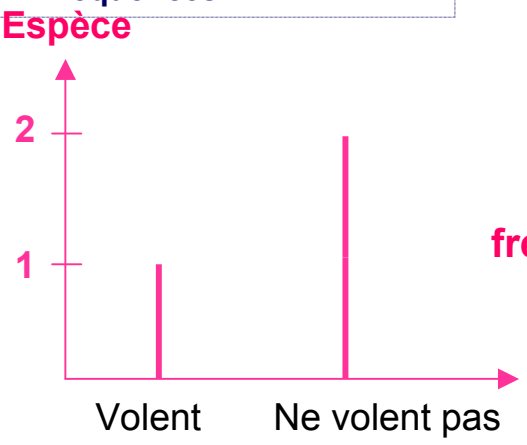
« L'espèce » est un concept qui devient la nouvelle unité statistique; on ne s'intéresse plus aux 600 individus en tant que tels

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



Graphe de fréquence des oiseaux (individus)



Graphe de fréquence des espèces (concepts)

COMPARAISON ENTRE LA STATISTIQUE DES INDIVIDUS DÉCRITS PAR DES DONNÉES CLASSIQUE ET LA STATISTIQUE DES CONCEPTS DÉCRITS PAR DES DONNÉES SYMBOLIQUES.

La statistique des oiseaux n'est pas la statistique des espèces d'oiseaux

La statistique des feuilles de maladies n'est pas la statistique des assurés

Les données classiques qui décrivent les individus de base ne sont pas des données symboliques qui décrivent les concepts.

CONLUSION: Les deux approches sont différentes et complémentaires

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

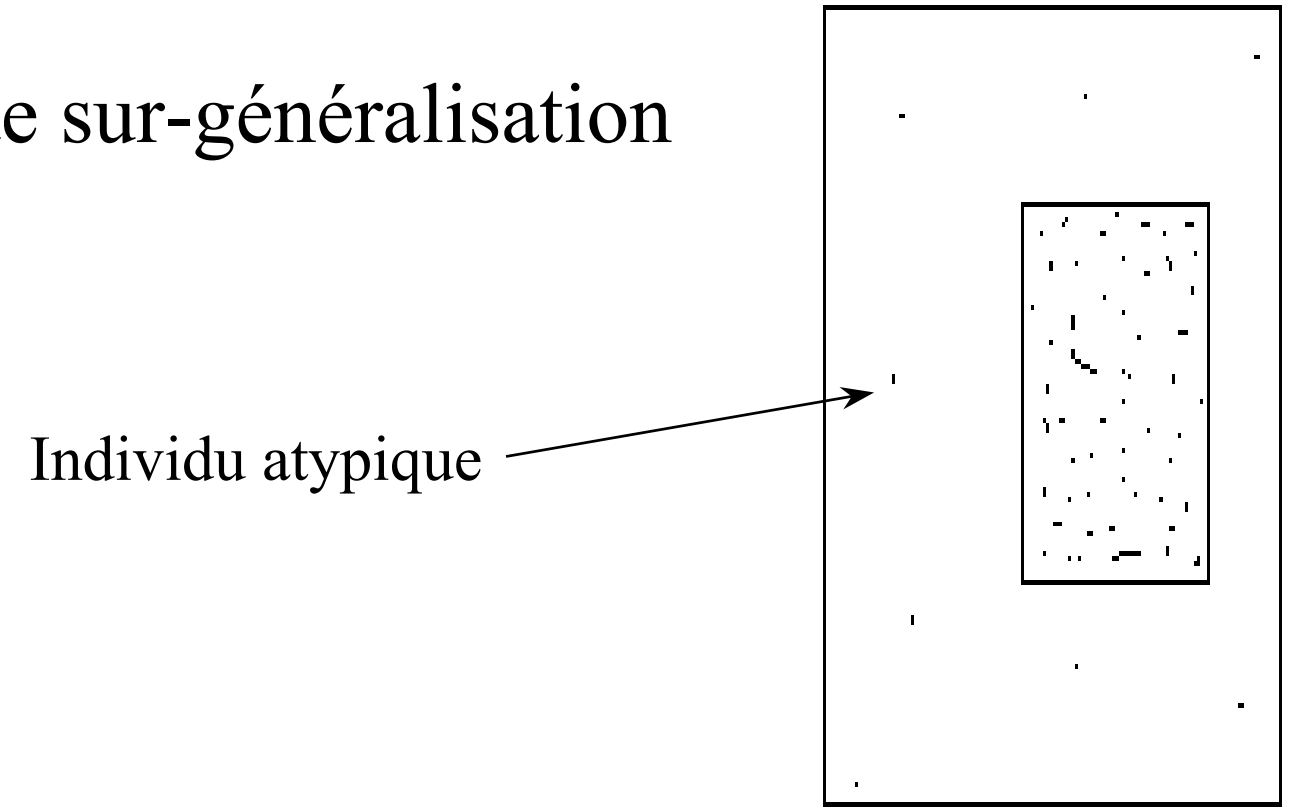
POUR EXPRIMER LA VARIATION INTERNE DES CONCEPTS

Chaque cellule peut contenir:

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

Réduction de la description d'OS

Problème de sur-généralisation



COMMENT CONSERVER DES LIENS PERDUS PAR GENERALISATION?

	Y1	Y2
C1	a	2
C1	a	2
C1	a	2
C1	b	1
C1	b	1
C1	c	2
C2	b	1
C2	b	3
C2	a	2

	Y1	Y2
C1	{a, b, c }	{1, 2}
C2	{a, b }	{1, 2, 3}

En ajoutant des connaissances supplémentaires:

EXEMPLE: ici on garde deux règles

[Y1 = a] ⟶ **[Y2 = 2]**

[Y2 = 1] ⟶ **[Y1 = b]**

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**

- **DEPENDANCES LOGIQUES**

Approche Classique Versus Symbolique: les données d'entrée et les méthodes de traitement

Données d'entrée

Classique

• Quantitatives: Points de **R** (nbres réels)
• Qualitatives Ordinales : Points de **N** (nbres naturels)
• Qualitatif non ordonné : Valeur nominale

Dans chaque case

Symbolique

- Diagrammes, Histogrammes ou Distributions
- Suite de valeurs
- Suite de valeurs pondérées
- Valeurs munies de règles (hiérarchie, variables mère-fille, « si….alors…. »…)
- Taxonomie (ex. : St Denis est inclus dans région parisienne)
- Fonctions
- graphes
- Séquences

Méthodes d'analyse

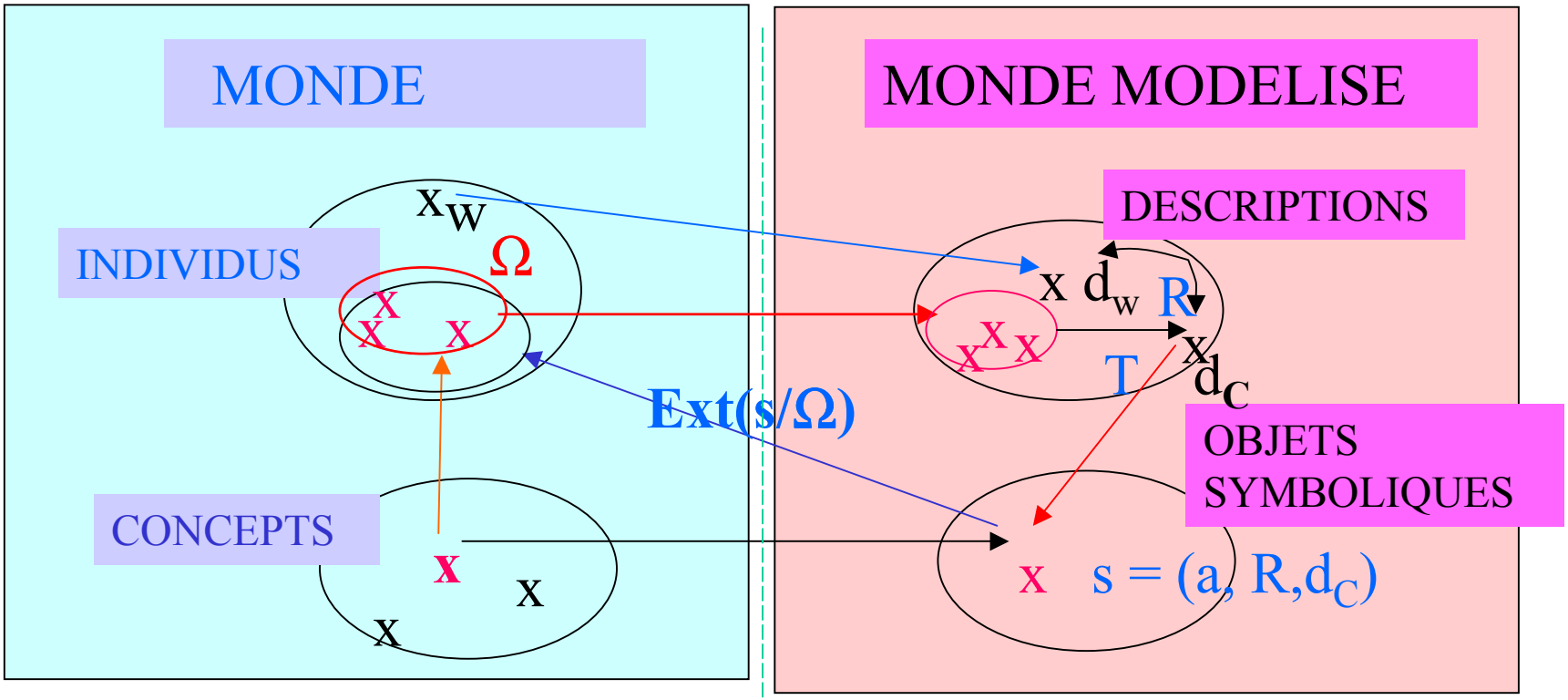
• Stat descriptive (Histos, Corrélations , biplots)
•Typologie (hiérarchies, pyramides, K-means, Nuées dynamiques, Cartes de Kohonen ,…)
•Décomposition de mélange de lois
•Arbres de Décision, boosting, baging, …
•Calcul et Représentation de dissimilarités
•Inférence de règles ou d'arbres de causalités
•Méthodes de visualisation (points)
•Analyse factorielle (ACP, AFC, …)
•Régression classique, PLS
•Réseaux neuronaux, VSM (Vector Support Machine), Etc.
•Treillis de Galois (données binaires)

Toutes les méthodes classiques se généralisent sur des concepts modélisés par des données symboliques :

✚ (PLUS)

+ Méthodes propres à l'analyse symbolique
-Indicateurs et fonctions de décision symboliques basés sur des concepts
-Dissimilarités (Hausdorff, …)
-Description symbolique de classes en sous-classes homogènes, discriminantes et séparantes.
- Explication symbolique de corrélations. Etc

MODÉLISATION ET APPRENTISSAGE DES CONCEPTS PAR QUATRE ESPACES



Exemple: Ω : 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 Ω 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

$$CAP(C, A) = Prob([X1 = A] \cup [X2 = A]) = p1+p2-p1p2$$

→ **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] x {employés, paysans}] ⟶ **par généralisation**

R = (⊆ , ⊆), ⟶ **appariement**

a(w) = [age(w) ⊆ [18, 52] ∧ [CSP (w) ⊆ {employés, paysans}]

agrégation

a(w) ∈ {VRAI, FAUX} ⟶ **fonction de reconnaissance**

OBJETS SYMBOLIQUES MODAUX

S = (a, R, d):

a(w) = [age(w) R_1 [(0.2)[12, 20]], (0.8) [20 ,28]]] $\wedge^*$

[SPC(w) R_2 [(0.4) employee, (0.6) worker]]

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

=> R$\rightarrow$ Appariement ,

Exemple: Paul Lévy, Hellinger, Kullback…

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

$\wedge^* \rightarrow$ Agrégation, copules

EXTENSION D'UN OBJET SYMBOLIQUE

CAS BOOLEEN :

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

CAS MODAL

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

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

→ RÉUTILISER LE CONCEPT SUR UNE AUTRE BASE,

→ IDENTIFIER UN INDIVIDU DE SON EXTENSION,

→AMELIORER PAR APPRENTISSAGE SA MODELISATION,

RÉDUIRE LES DONNÉES

→MASSIVES,

→MANQUANTES

→ LA CONFIDENTIALITÉ

EN SE DONNANT LA POSSIBILITE DE LES RETROUVER.

EN QUOI L'ADS EST INNOVANTE PAR RAPPORT AUX APPROCHES CLASSIQUES EN STAT, AD, DATA MINING?

La démarche classique: on dispose d'un tableau de données classique comportant une valeur unique par case (quantitative ou qualitative) .

La démarche symbolique:

On dispose d'une Base de Donnée,

→une requête fournit un tableau de données classiques muni d'une variable privilégiée dont les modalités sont des catégories.

→on construit par généralisation un nouveau tableau dont les unités sont des concepts (réifiant les catégories précédents) décrits par des données symboliques munies de connaissances supplémentaires.

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 EXPLICATIVES CAR S'EXPRIMANT DANS LE LANGAGE DE L'UTILISATEUR.

CINQ PRINCIPES

1) A CHAQUE ETAPE, SEULEMENT DEUX NIVEAUX:

Premier niveau: les individus

Second niveau: les concepts

2) LES CONCEPTS PEUVENT EUX-MÊME ÊTRE CONSIDÉRÉS COMME DES UNITÉS ET REIFIES AU MÊME TITRE QUE LES INDIVIDUS

3) UN CONCEPT PEUT ÊTRE DÉCRIT EN UTILISANT UNE CLASSE D'INDIVIDUS DE SON EXTENSION

4) LA DESCRIPTION D'UN CONCEPT DOIT EXPRIMER LA VARIATION DES INDIVIDUS DE SON EXTENSION

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

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.

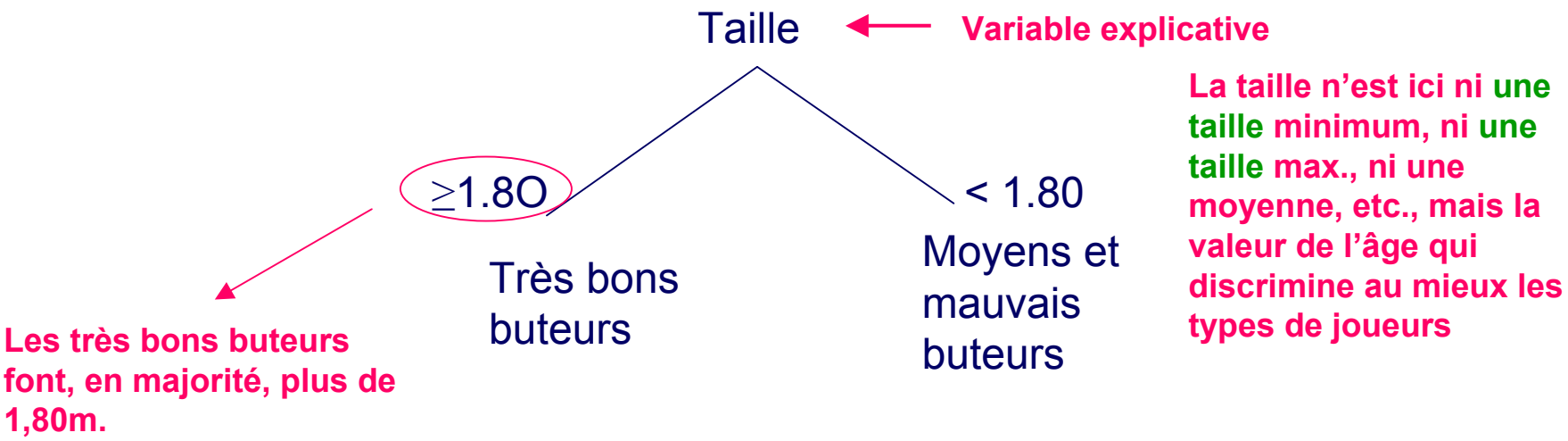
Exemple 1: Perte d'information du codage classique de données symboliques

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

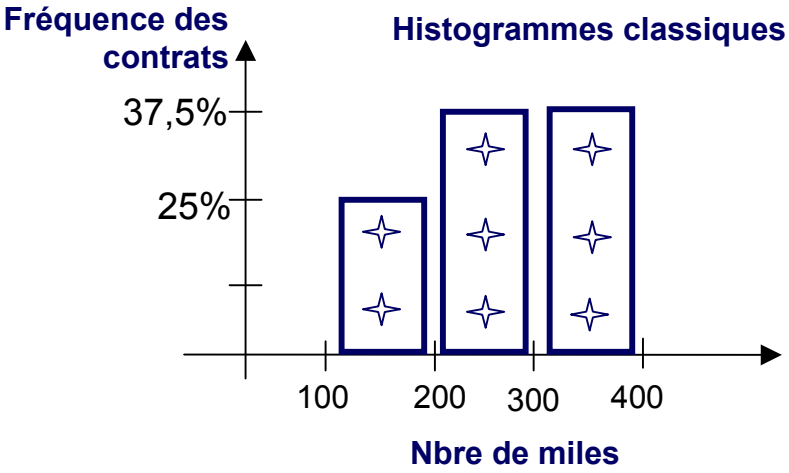


Perte d'information du codage classique de données symboliques

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

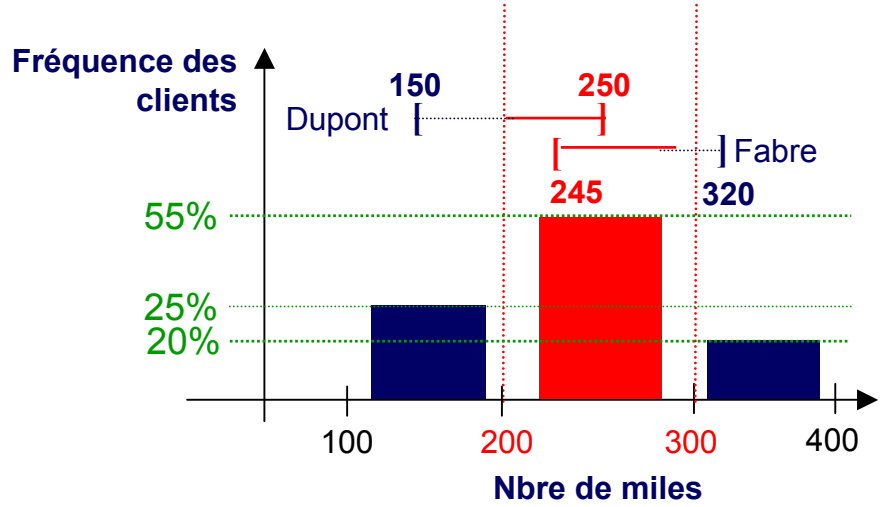
Consommation	Client concerné	Nbre de miles
Achat 1	M. Dupont	150
Achat 2	M. Dupont	180
Achat 3	M. Dupont	250
Achat 1	Mme Fabre	270
Achat 2	Mme Fabre	245
Achat 3	Mme Fabre	310
Achat 4	Mme Fabre	320
Achat 5	Mme Fabre	315

Histogrammes classiques



Client	Nbre de miles
M. Dupont	[150;250]
Mme Fabre	[270;320]

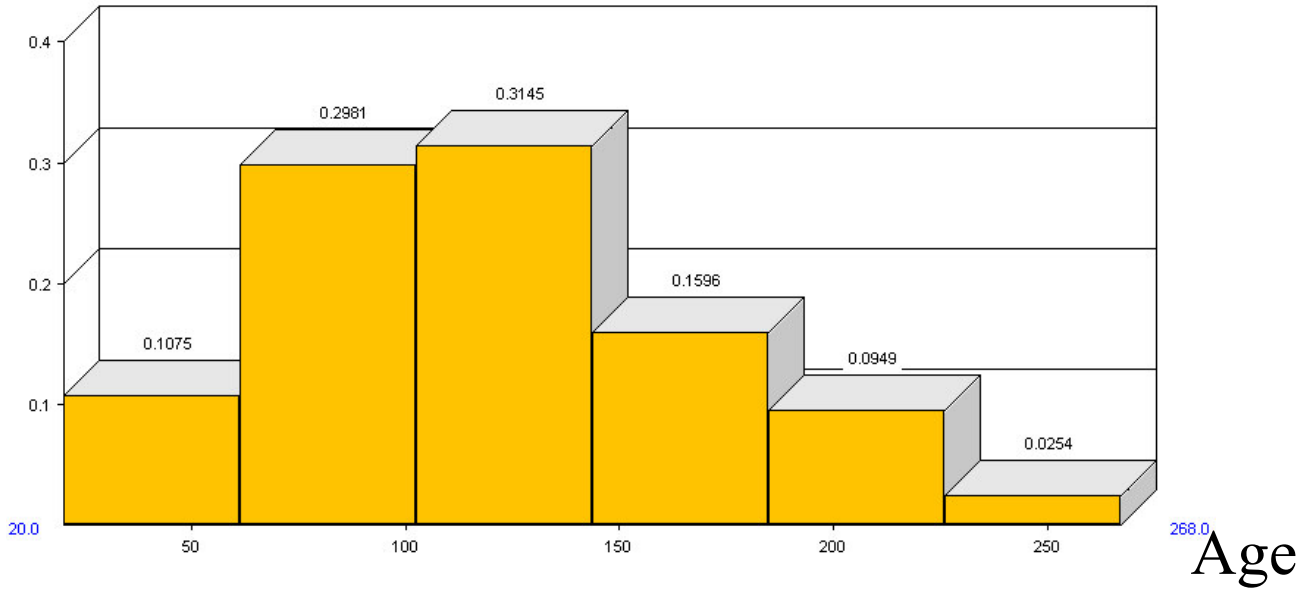
Histogrammes symboliques sur le concept « client » :



L'histogramme [200, 300] représente la somme des portions d'intervalles entrant dans cet écart . 50% de la conso de M. Dupont et 60% de la conso de Mme Fabre sont pris en compte pour former cet histogramme.

L'approche symbolique permet de construire un histogramme d'une variable à valeur intervalle et de conserver ainsi les min ET les max

Exemple 2: 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



L'approche symbolique permet de construire un histogramme d'une variable à valeur intervalle ou histogrammes.

Application: Détection de profils symboliques rares (outliers)

Perte d'information du codage classique de données symboliques

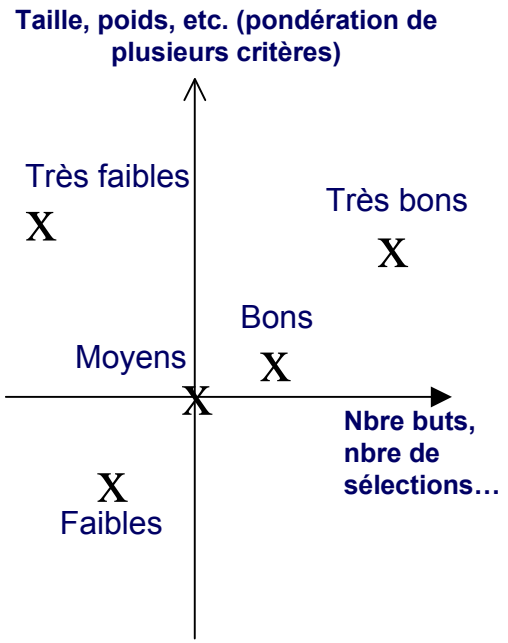
EXEMPLE 3 : Analyse en composantes principales

En codage classique, chaque concept est représenté par un point

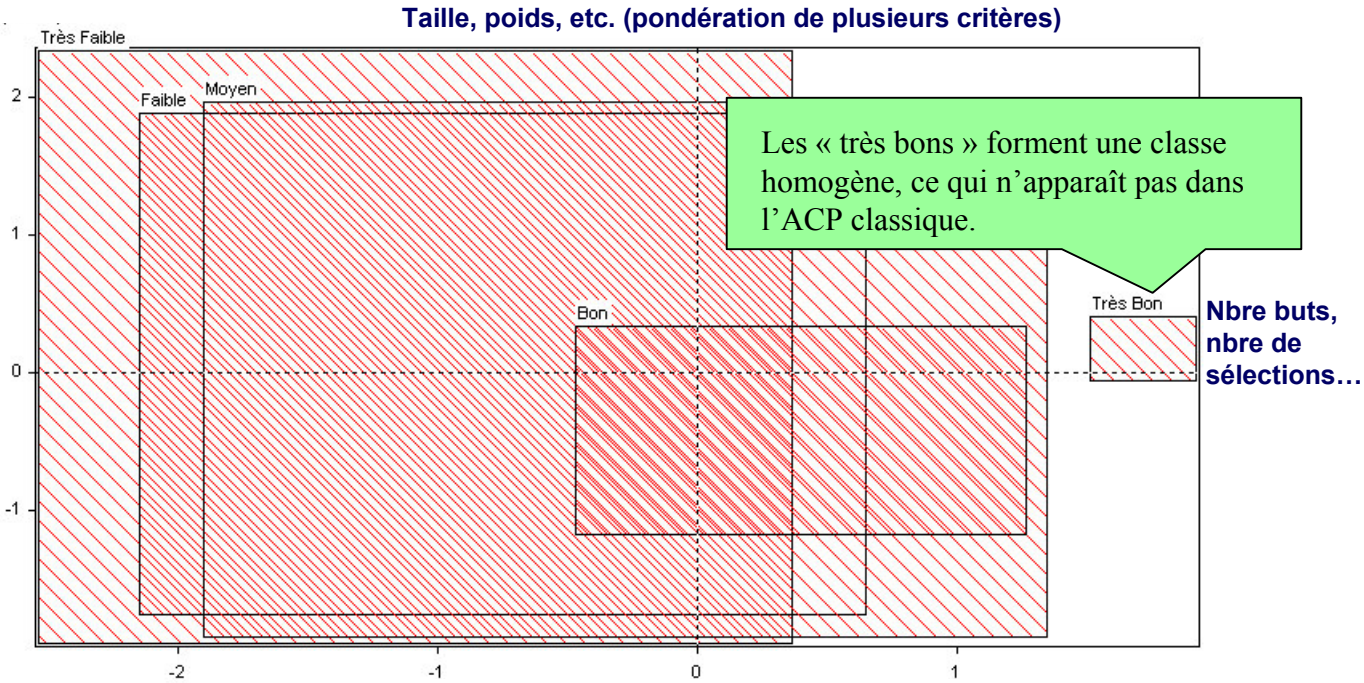
En codage symbolique :

→ 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 poids, etc. / le nbre de buts marqués, le nbre de sélections, etc..



ACP Classique



ACP symbolique (A. Chouakria et al.)

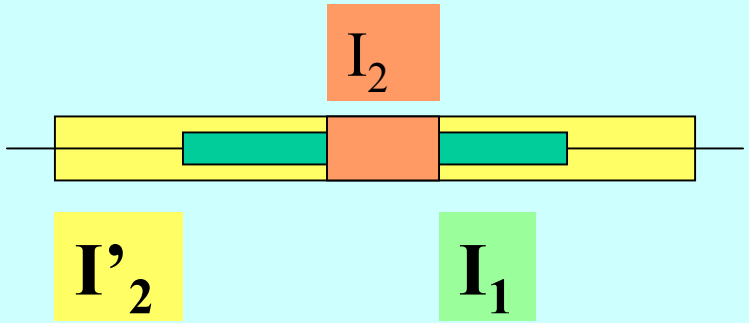
EXEMPLE 4: DISSIMILARITES Classique versus symbolique

Pour calculer une distance entre deux intervalles, l'utilisation de l'écart des min et de l'écart des max

$$d(I_1, I_2) = \left| \text{Min}(I_1) - \text{Min}(I_2) \right| + \left| \text{Max}(I_1) - \text{Max}(I_2) \right|$$

est erroné

Exemple:



$d(I_1, I_2) = d(I_1, I'_2)$ alors que intuitivement I_1 et I_2 sont plus proches puisque leurs bornes sont plus proches.

La raison:

l'écart $\left| \text{Min}(I_k) - \text{Max}(I_j) \right|$ n'est pas pris en compte comme par ex avec la dissimilarité de Haussdorf.

Stratégie Classique versus Symbolique: Les trajectoires

Trajectoires classiques : ce sont celles des unités statistiques de base décrites par des données classiques.

Trajectoires symboliques: ce sont celles des concepts décrits par des données symboliques.

Par exemple: extension des méthodes classiques pour la prédiction à partir d'une série temporelle d'intervalles.

Stratégie Classique versus Symbolique

Classique:

→les unités statistiques de base sont l'objet de l'analyse.

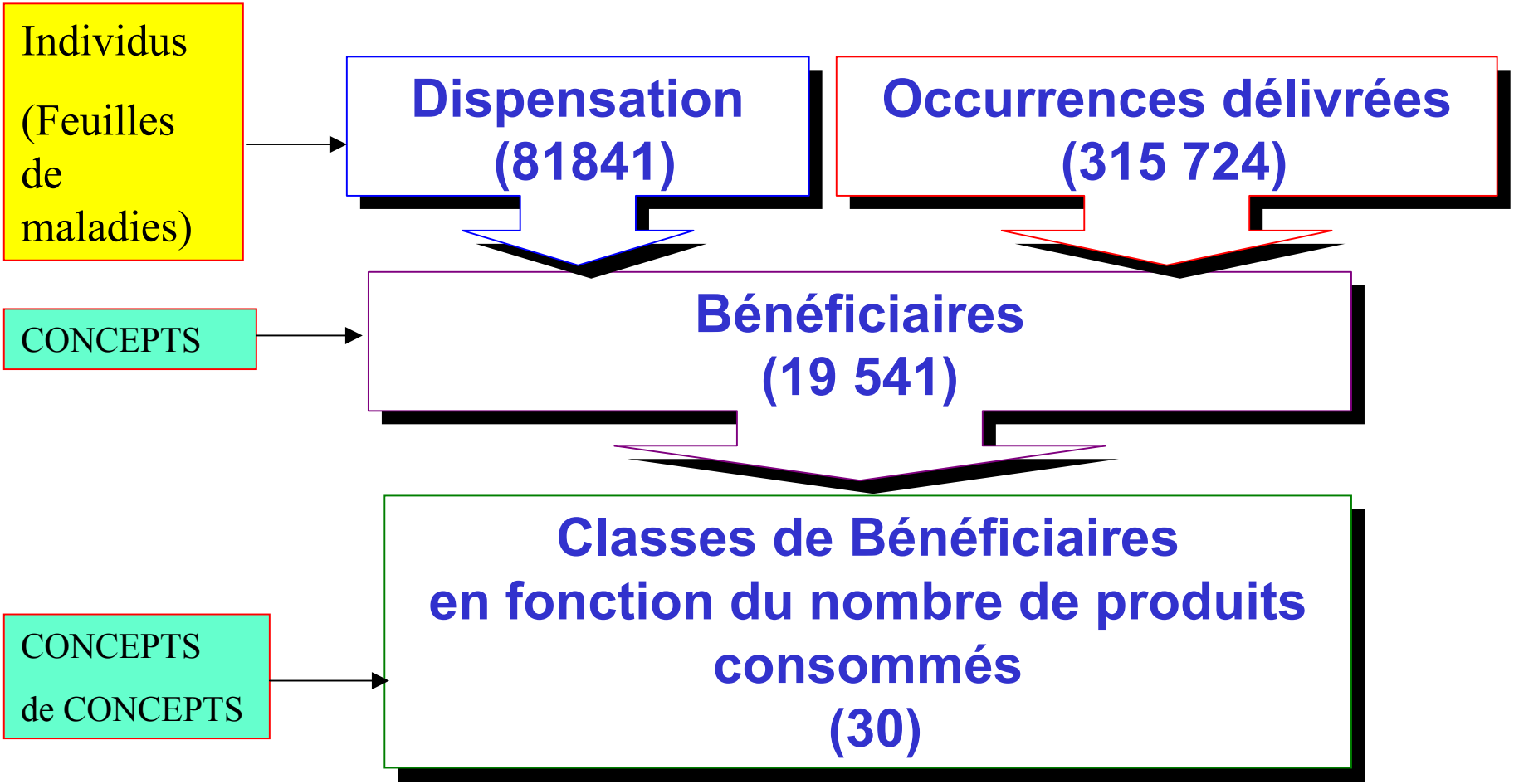
→ Elles sont décrites par des variables classiques parfois munies de variables cibles à expliquer.

Symbolique:

→Les concepts (souvent construits à partir des variables cibles de premier niveau) constituent l'objet de l'analyse.

→ Ils sont décrits par des variables symboliques parfois munies de variables « conceptuelles » cibles à expliquer.

DES INDIVIDUS AUX CONCEPTS



Quatre Niveaux d'Unités Statistiques

Niveau 1: Les produits achetés

Niveau 2: Les clients

Niveau 3: Trente catégories de consommation

Niveau 4: Trois catégories de consommation

(Grand, Moyen, Petit)

Niveau 1: données classiques,

Niveaux 2, 3, 4: données symboliques

Stratégie Classique versus Symbolique

Les individus (Unités Statistiques de la base):

→Les feuilles de maladies exprimant la consommation d'assurés sociaux.

→Variables descriptives: Taux de remboursement, médicament générique, date, type de médecin …

→Variables cibles: coût de la consommation sur une période.

→Les concepts:

→niveau de consommation décrit par les mêmes variables.
→Mêmes variables descriptives mais symboliques.

→Variable cible à expliquer petite, moyenne, grande consommation.

Données complexes

versus données symboliques

Données incomplètes classiques deux types:

→Ont un sens mais sont absentes: le passage aux concepts les réduit voire les fait disparaître.

→N'ont pas de sens: type de camion d'une entreprise qui n'en a pas: le passage aux concepts tient compte de variables hiérarchique dites « mères-filles ».

Exemple: régressions symbolique puis régressions fille etc.

Données complexes

versus données symboliques

Données spatio-temporelles:

→en passant des villes aux régions on obtient des concepts « régions » définis par des données symboliques exprimant la variation.

→Les séries temporelles associées aux concepts « régions » peuvent être représentées par la variation de la probabilité p_i ou de « l'information » $p_i \, \mathrm{Log} \, p_i$ de séquences à 1, 2, …, k éléments.

Trajectoires Classiques versus Trajectoires Symboliques

Exemple de Trajectoire classique: évolution de la vente d'un article précis (définit par un code de transaction), portable X sur une période donnée pour un individu.

Exemple de Trajectoire symbolique: évolution de la consommation GSM d'un segment de population.

Nomadisme: ceux qui ont modifié l'abonnement de 1 à 2 fois, 2 à 4, 4 et + sur trois mois → Trois concepts décrits par des var. symboliques (age, sexe, CSP, résidence, coût d'abonnement….)

Persistance: ceux qui sont restés consommateurs 1, 2, 3, 4 mois entre Octobre et Décembre 2003 → Quatre concepts décrits par des var. symboliques (age, sexe, CSP, résidence, coût d'abonnement ….).

Classique versus Symbolique : les données spatio-temporelles

Pour chaque pathologie et pour chaque hôpital, il existe en général 4 à 5 trajectoires de patient possibles

Trajectoire A
Patient 1urologie, 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…

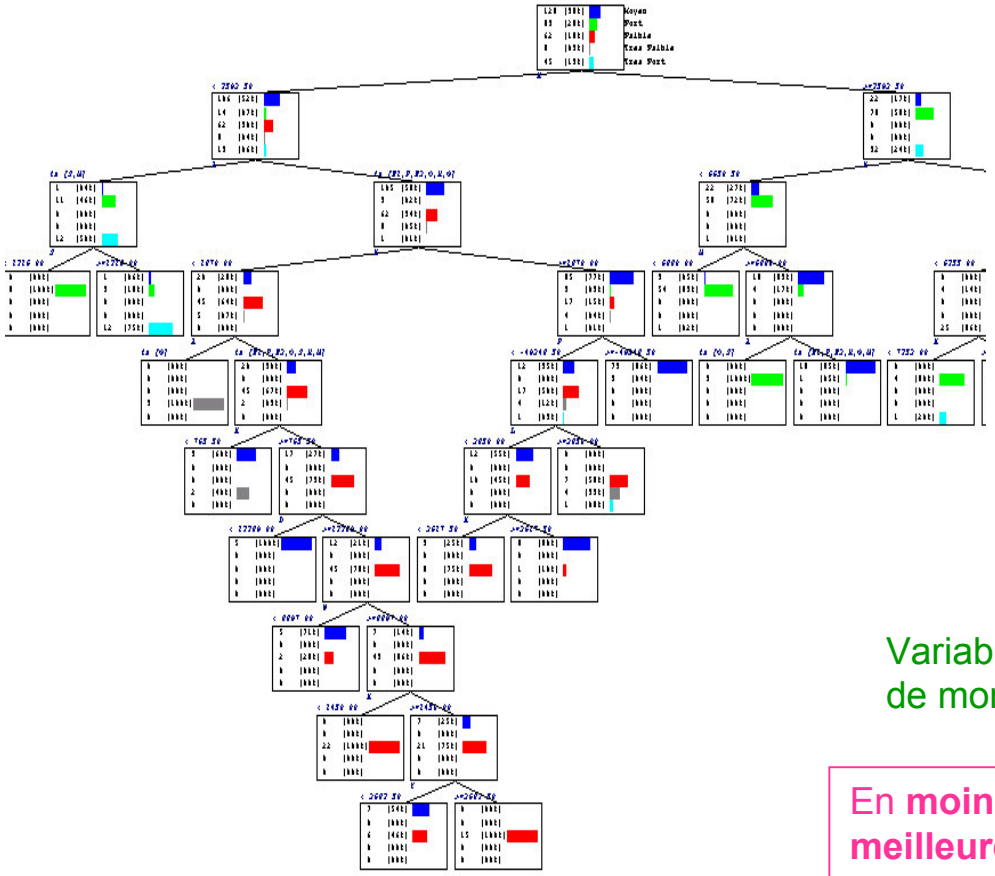
On peut aussi construire, à partir de milliers de trajectoires, des classes de trajectoires (rassemblant chacune des centaines de trajectoires). Ces classes seront les nouveaux concepts sur lesquels est effectuée l'analyse statistique.

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

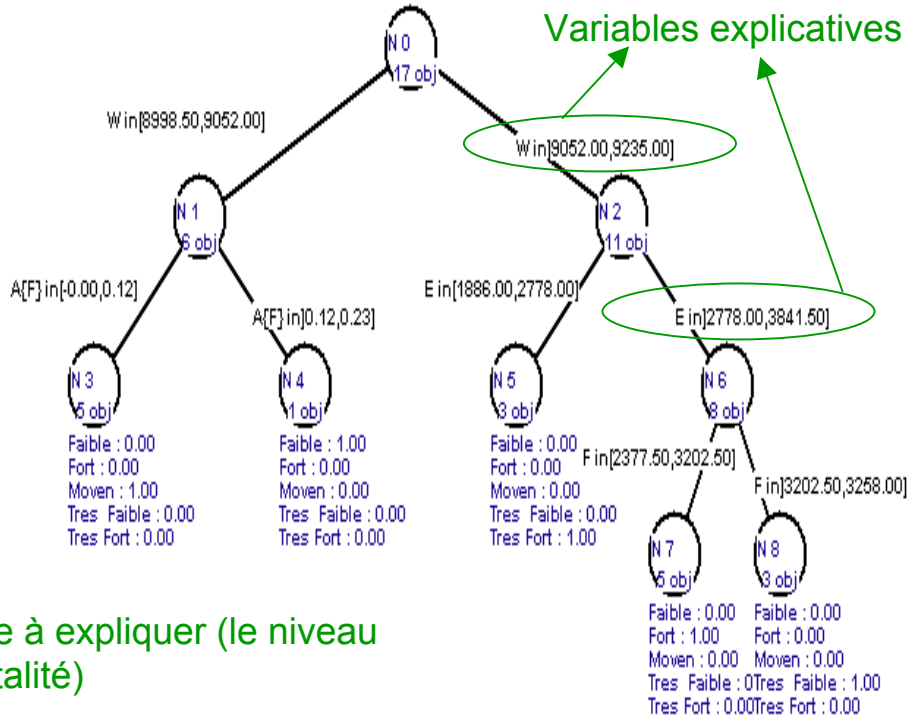
Classique / symbolique : une comparaison

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

Arbre « classique » sur les patients



Arbre « symbolique » sur les trajectoires



Variable à expliquer (le niveau de mortalité)

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 données symboliques

Au départ chaque case contient un objet complexe

	Catégor	Image	Texte	Sequ.
i1	Cj		doc1	agbdc
--	------	--------	--------	
in	Ck		docn	dgabh

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

	Image	Texte	Sequ
C1	{image}1	{doc}1	{gba}1
---	--------	--------	
Ck	{image}k	{doc}k	{ahd}k

Description d'objets complexes

Exemple: C_i = images maritimes

	Catég	Image	Texte	Sequ.
i1				
---		--------	--------	-------
in				

Généralisation

	Image	Texte	Sequ
C1			
------	--------	--------	
Ck			

Données Classiques

Données Symboliques

Données complexes

versus données symboliques

Données imprécises:

→ Le passage aux concepts exprime la variation de ces données.

Exemple: Taille (Jean) = 1. 50 +/- 0.1,

Taille (Paul) = 1. 60 +/- 0.2

Si Paul et Jean sont blonds, le concept « blond » est décrit par: Taille (Blond) = [1.49, 1.62]

Données complexes

versus données symboliques

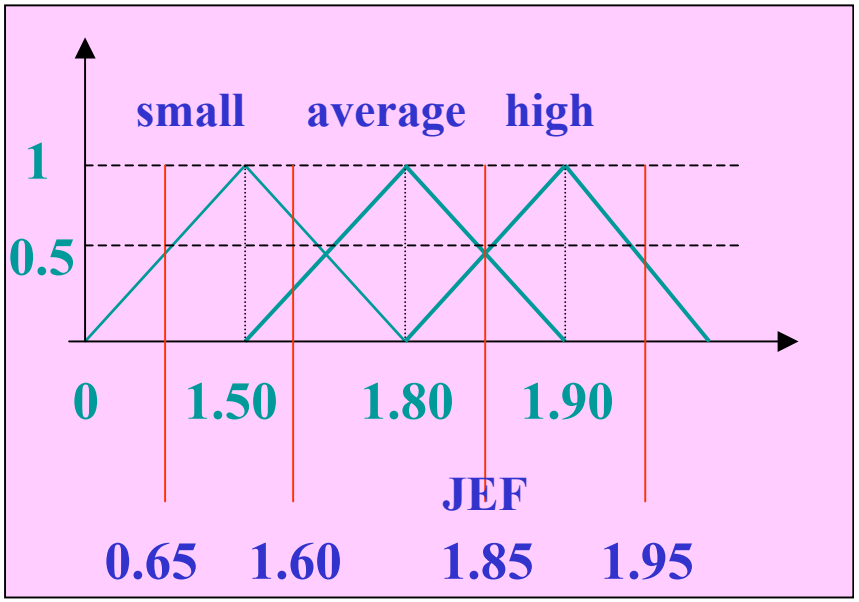
Donnée floues:

Le passage aux concepts exprime la variation des données floues.

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

From Numerical to Fuzzy Data



Initial 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

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] ∧ [caractéristiques des écoles]

Données classiques versus Données symboliques

Le cas des DONNEES CONFIDENTIELLES

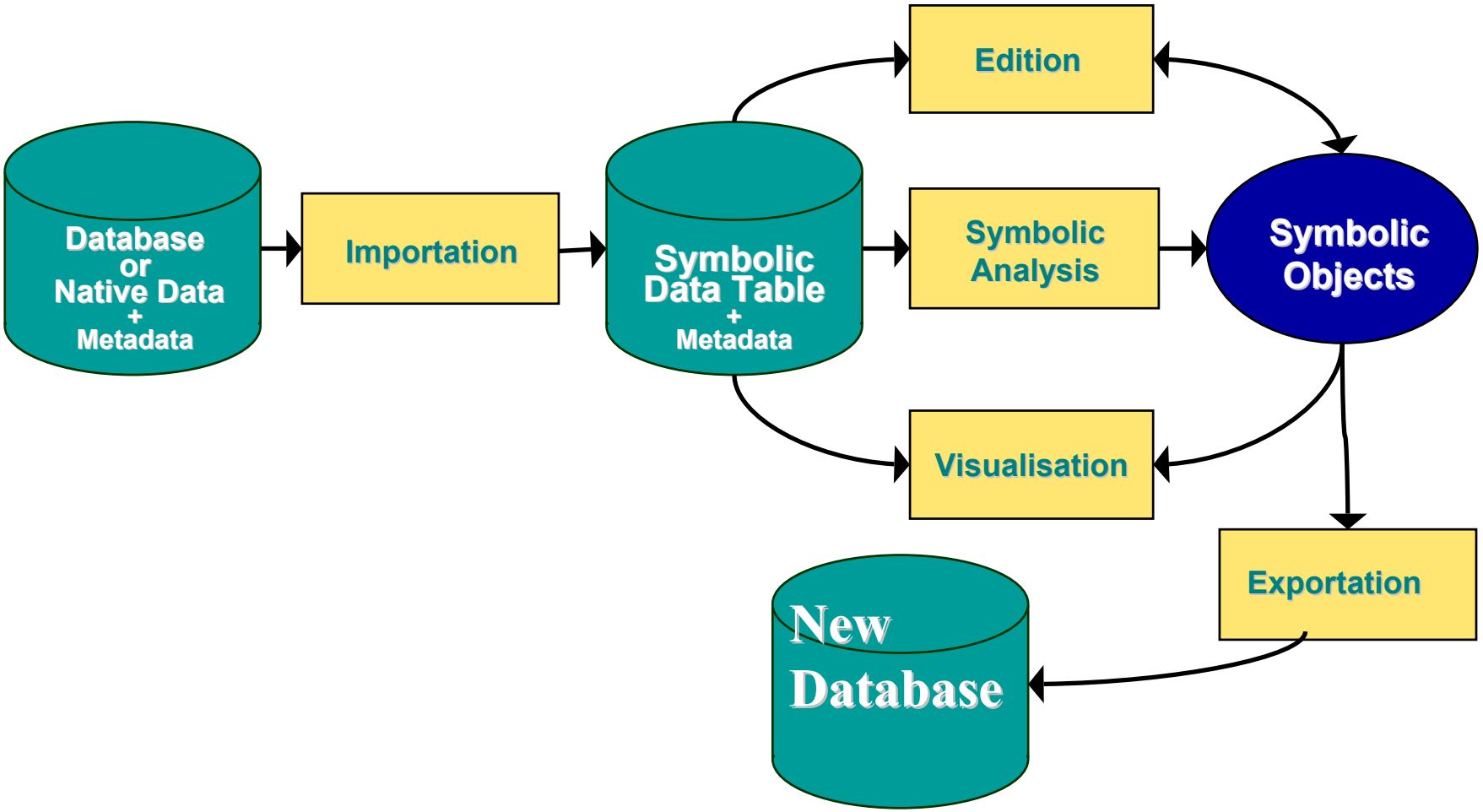
Approche Classique: les individus sont décrits par des données confidentielles

Approche Symbolique: les concepts sont décrits par des données symboliques qui ne sont plus confidentielles puisque les individus n'apparaissent plus.

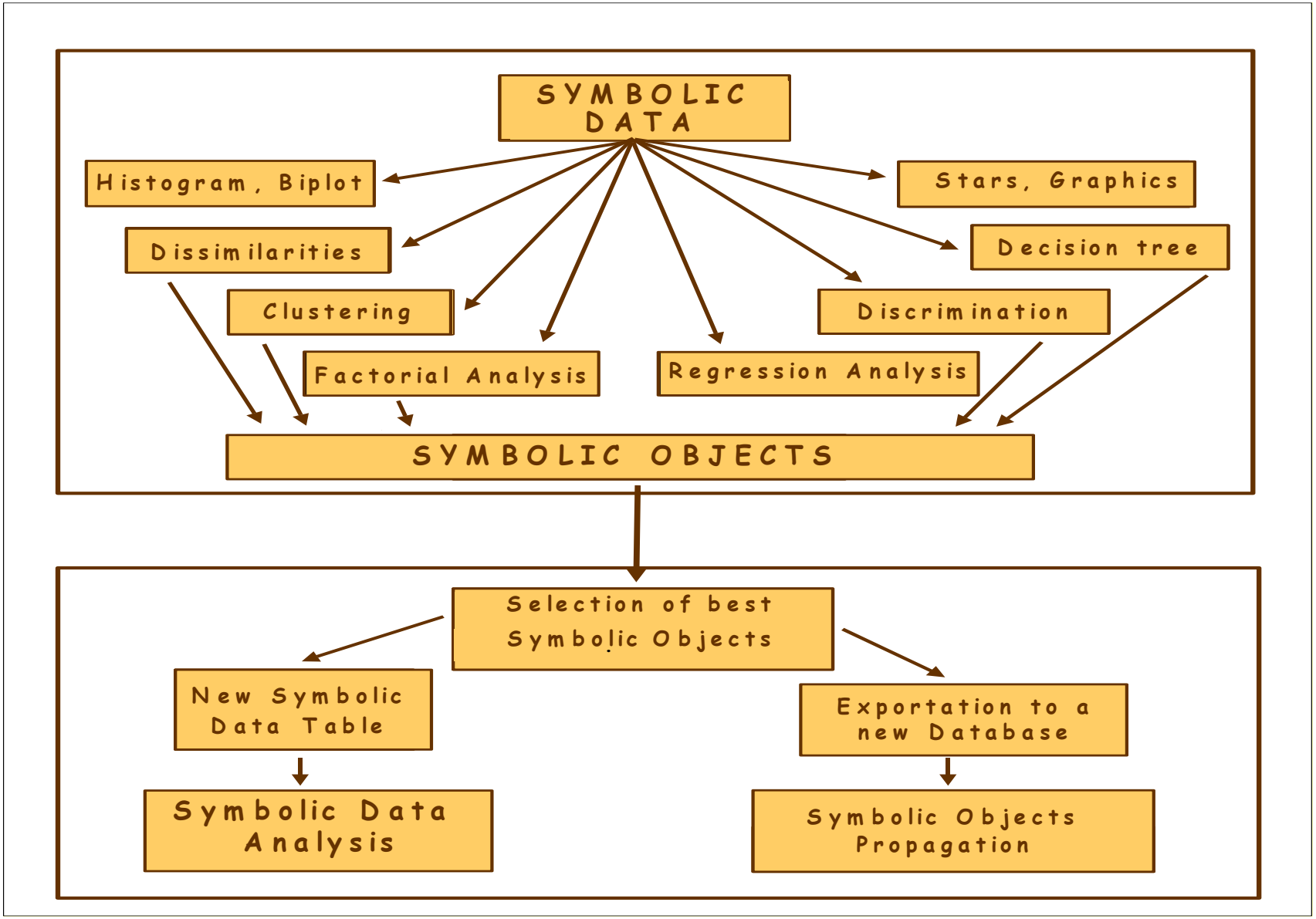
ASSO-SODAS Consortium

- **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), Madrid (UCM).
- **Centre de Recherche:** INRIA (F),
- **Compagnies:** CISIA (F), TES (L), EDF, THOMSON (F)

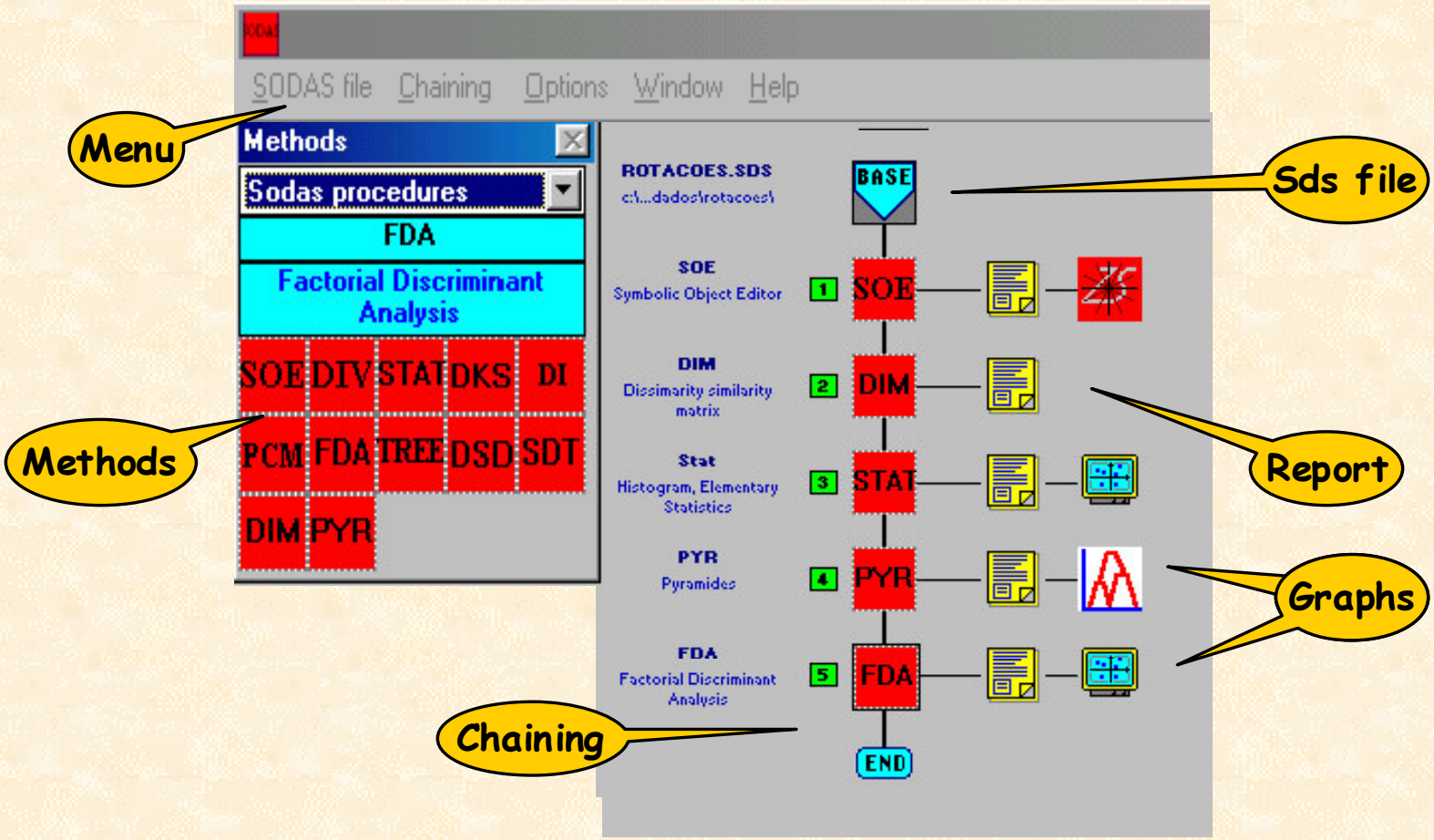
ASSO Architecture



THE SODAS 2 SOFTWARE FROM ASSO



SODAS Software

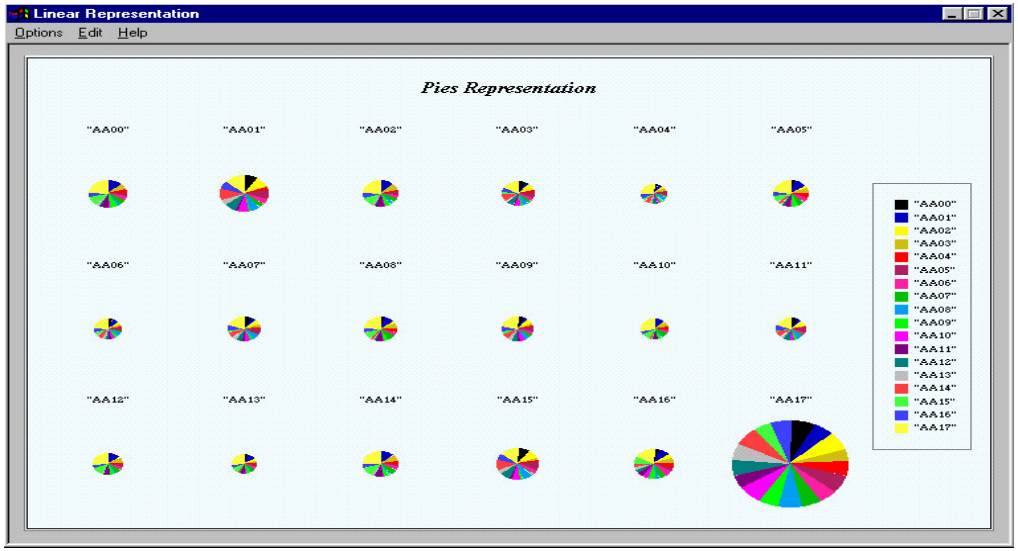


Extension de Méthodes classiques aux concepts

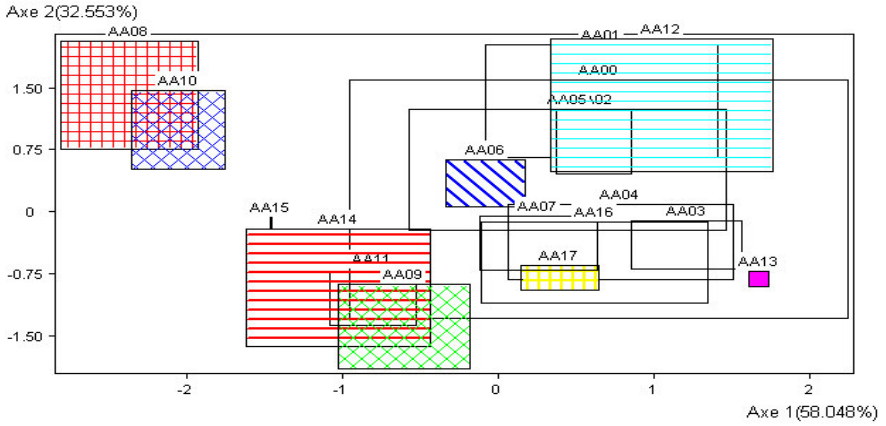
- Histos, correlation
- Visualisation en Etoiles
- Biplots
- ACP, AFC
- Décomposition de mélanges
- Multidimensional Scaling
- Typologie (Nuées Dynamiques , Pyramides)
- Régression
- Réseaux neuronaux
- Arbres de Décision
- Extraction de Règles
- Treillis de Galois

Autres exemples de méthodes de SODAS

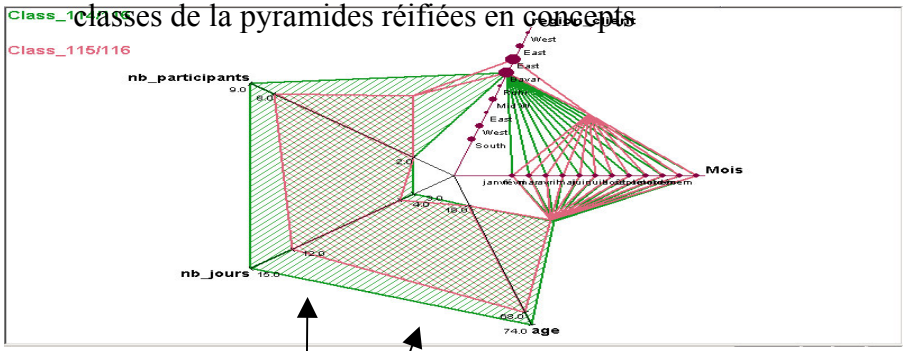
CARTE DE KOHONEN DE CONCEPTS



ANALYSE FACTORIELLE: ACP



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

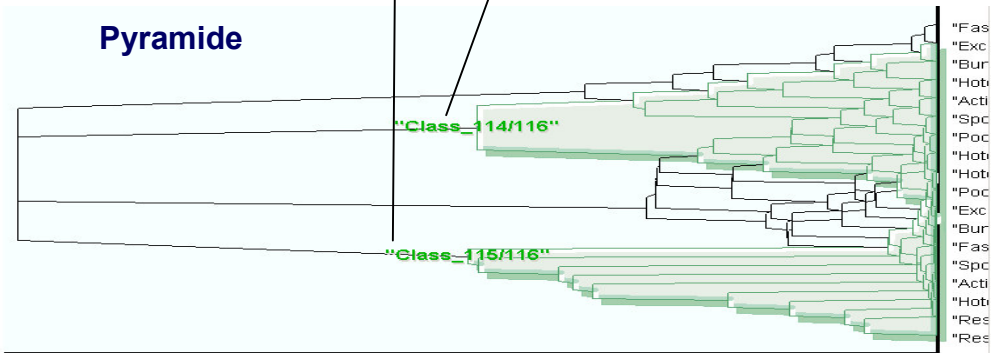


Méthode DIV (division en classes de concepts homogènes et description symbolique de ces classes réifiées en concepts)

```
   - 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)
```

Pyramide



La représentation cartésienne

- Cas de variables à valeur intervalle
- Cas de variables à valeur histogramme (utilisation des copules)

MODELISATION PROBABILISTE

CAS STANDARD: Les variables sont des variables aléatoires à valeur quantitative ou qualitative.

- CAS SYMBOLIQUE:

Les variables sont 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	Bénéficiare	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
Bénéficiare	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

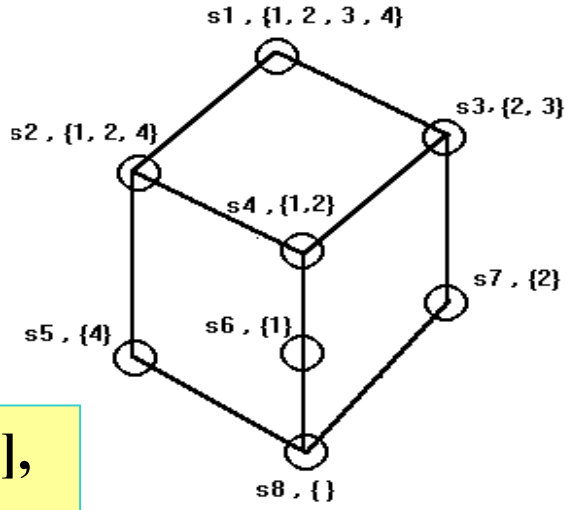
Treillis de Galois

- C’est la structure naturelle des objets symb car ils représentent de façon cohérente les intension et extension des objkets symboliques dits complets
- Objets symb complets: l’intension de l’extension est la même intension

Tableau de données symboliques

	Y1	Y2	Y3
W1	{a, b}	∅	{g}
W2	∅	∅	{g, h}
W3	{c}	{e, f}	{g, h, i}
W4	{a, b, c}	{e}	{h}

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



Objets symboliques induits du Treillis de concepts de concepts

s_2 : $a_2(w) = [y_2(w) \subseteq \{e\}] \wedge [y_3(w) \subseteq \{g,h\}]$,
$Ext(s_2) = \{1, 2, 4\}$

s_3 : $a_3(w) = [y_1(w) \subseteq \{c\}]$,
$Ext(s_3) = \{2, 3\}$

s_4 : $a_4(w) = [y_1(w) \subseteq \{a,b\}] \wedge [y_2(w) = \varnothing] \wedge [y_3(w) \subseteq \{g,h\}]$,
$Ext(s_4) = \{1, 2\}$

s_5 : $a_5(w) = [y_2(w) \subseteq \{e\}] \wedge [y_3(w) \subseteq \{h\}]$,
$Ext(s_5) = \{4\}$

PARTITIONNEMENT D'UN ENSEMBLE DE CONCEPTS

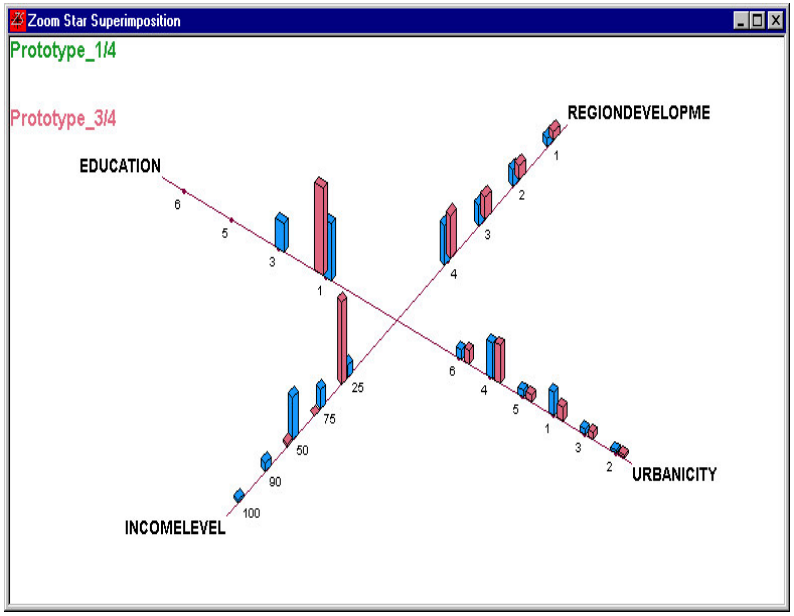
Y. Lechevallier, F. De Carvalho, R. Verde

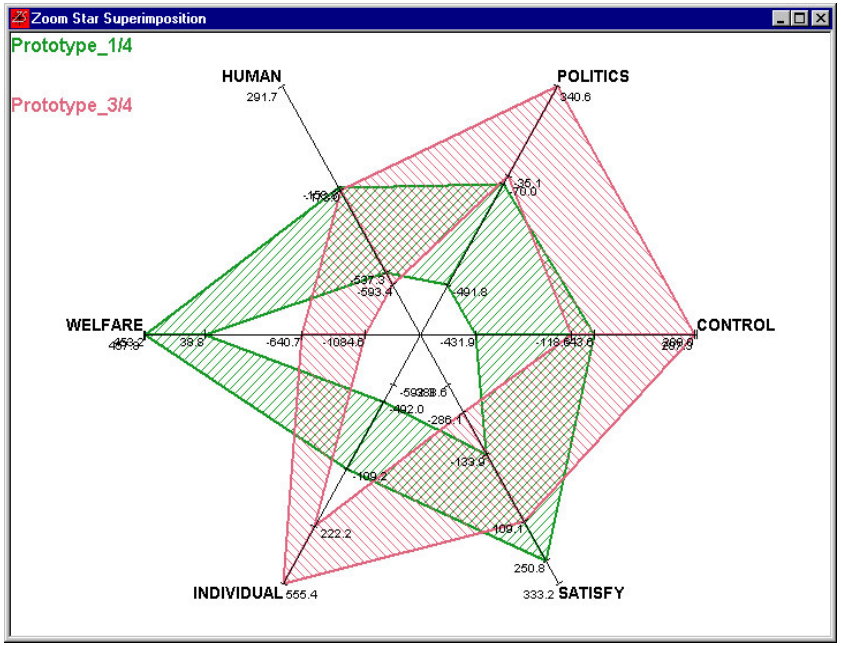
CLASSIQUE		SYMBOLIQUE
Moyennes	→	Prototypes (= Obj Symb)
K-means	→	Nuées Dynamiques
Dissimilarité standard	→	Dissimilarité symbolique
(Euclidienne, KHI2…)		(Haussdorf, Ichino, De Carvalho,…)

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: 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.

Règles d'association classiques versus symboliques

- Règles d'association symboliques
  - **Ex: $1/5<P_v\leq 2/5 \wedge [70, 100] \subseteq X \rightarrow 0<P_P\leq 1/5$**
  - **Si pour un client donné,**
  - la fréquence d'achat de viande est comprise entre 1/5 et 2/5 de ses achats totaux
  - **Et si** le client a dépensé entre 70 et 100
  - **Alors** le client a également acheté du poisson (pour moins de 1/5 de ses achats totaux)

- Règles d'association classiques et symboliques obtenues à partir des matrices classiques et symboliques précédentes

N	Règles Classiques	Confiance %
1	c→p	60
2	p→v	58
3	p→c	50
4	v→p	47

N	Règles Symboliques	Confiance %
1	$[70,100]\subseteq X \wedge 1/3<P_v\leq 2/3 \rightarrow 1/3<P_p\leq 2/3$	98%
2	$[70,100]\subseteq X \wedge 2/3<P_p\leq 1 \rightarrow 1/3<P_v\leq 2/3$	94
3	$1/3<P_v\leq 2/3 \rightarrow 0<P_p\leq 1/3$	79
4	$0<P_c\leq 1/3 \rightarrow 1/3<P_p\leq 2/3$	74

Extraction de règles d'association avec l'algorithme Apriori

Agrawal et al. (1993), Agrawal et Srikant (1994)…

- Extraction de règles d'association.
  - Exemple du panier de la ménagère:
    - » {lait, oeufs}$\rightarrow${beurre}

- Input de l'algorithme Apriori
  - n items I = $\{i_1,\ldots, i_n\}$ (lait, oeufs, beurre…)
  - m transactions T=$\{t_1, \ldots, t_m\}$ $t_i \in P(I)$-Ø (tickets de caisse)

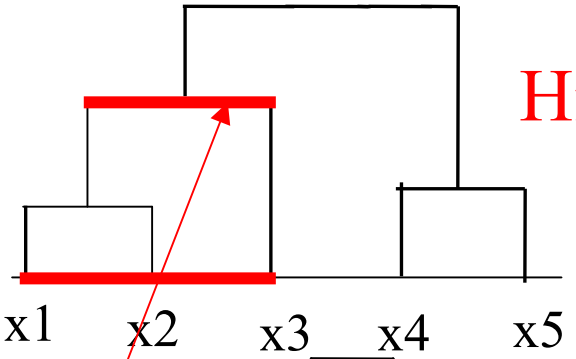
- Output: Règles d'association: $X \rightarrow Y$, $X \subset I$, $Y \subset I$, $X \cap Y = \emptyset$
  - avec un support et une confiance supérieurs à deux seuils minimum *minsup* et *minconf*

- $sup(X \rightarrow Y) = card(t \in T / X \cup Y \subseteq t) / card(T)$

  = Proportion des transactions contenant à la fois X et Y
- $conf(X \rightarrow Y) = sup(X \rightarrow Y) / sup(X)$

  = Proportion des transactions contenant Y parmi celles contenant X

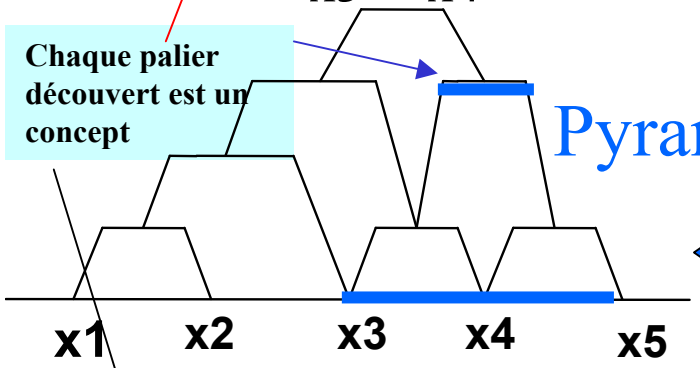
QUALITE DE LA REPRESENTATION SPATIALE



Hierarchies

Ultrametric dissimilarities = U

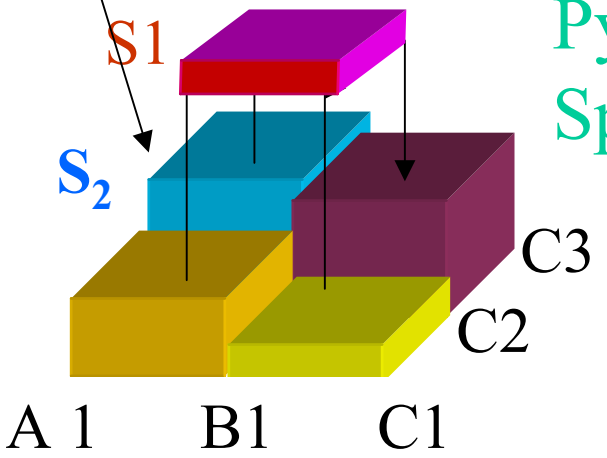
$W = |d - U|$



Pyramides

Robinsonian dissimilarities = R

$W = |d - R|$



Pyramides Spatiales

Yadidean dissimilarities = Y

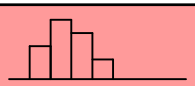
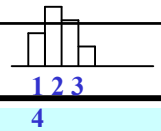
$W = |d - Y|$

Problèmes et Méthodes spécifiques

- Indicateurs conceptuels
- Codage par partition optimale (pas Fisher !)
- Ordre de variables symboliques
- Description symbolique de classes
- Explication symbolique des corrélations

COMMENT CONSERVER LA CORRELATION ET L'EXPLIQUER?

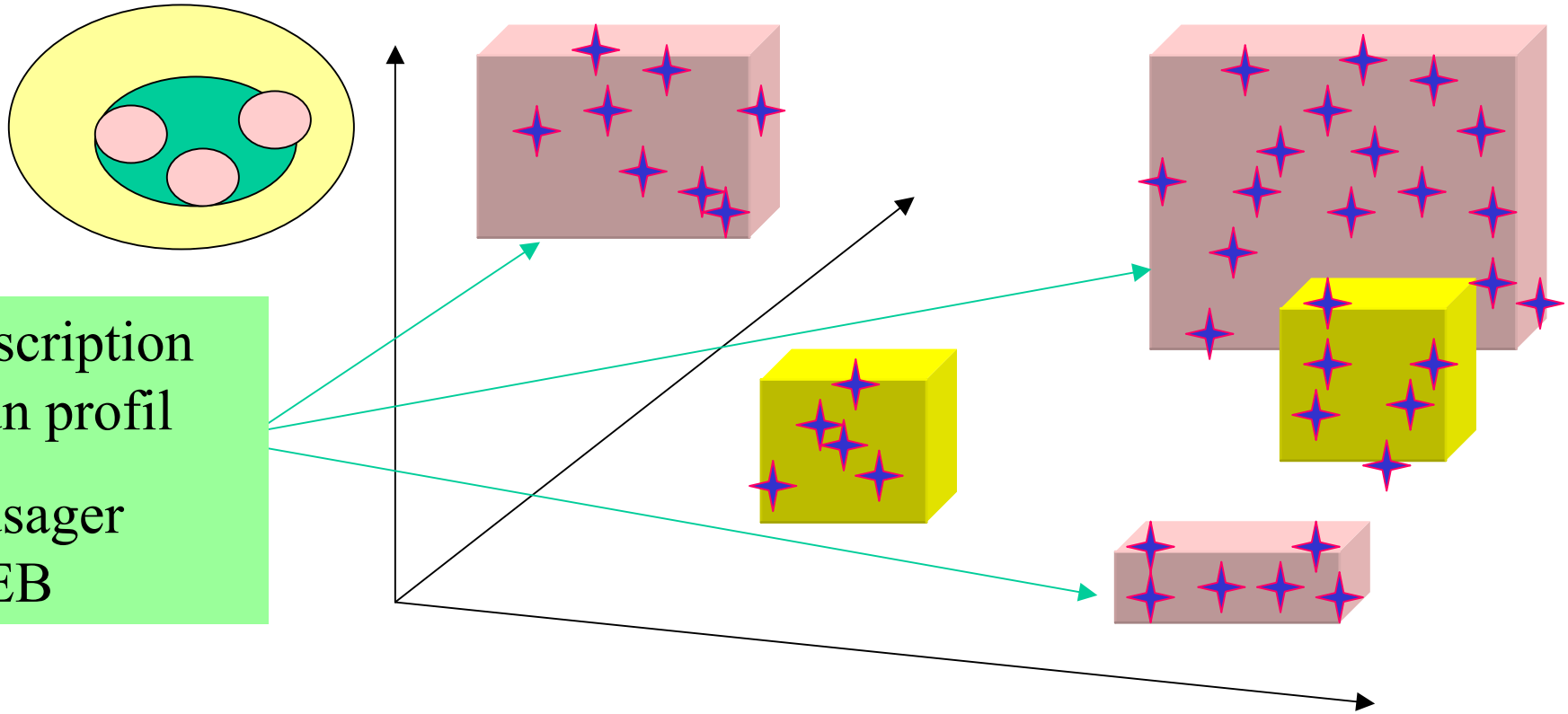
Indiv	Concept	opht	dent	lieu	période
i1	C1	12.5	3	Lyon1	
i2	C1	9.6	2	Paris 3	
i3	C1	11.4	4	Paris 3	
i4	C2	3.2	1		
i5	C2	7.1	4		

Concept	opht	dent	lieu	période	Cor(opht, phar)
C1	[9.6, 12.5]		{Lyon1, Paris 3}		Cor_{C1} (opht, dent)
C2	[3.2, 7.1]		Paris 3		Cor_{C2} (opht, dent)
C3	[4.7, 8.1]				Cor_{C3} (opht, dent)
C4	[5, 16]		Pau4		Cor_{C4} (opht, dent)

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

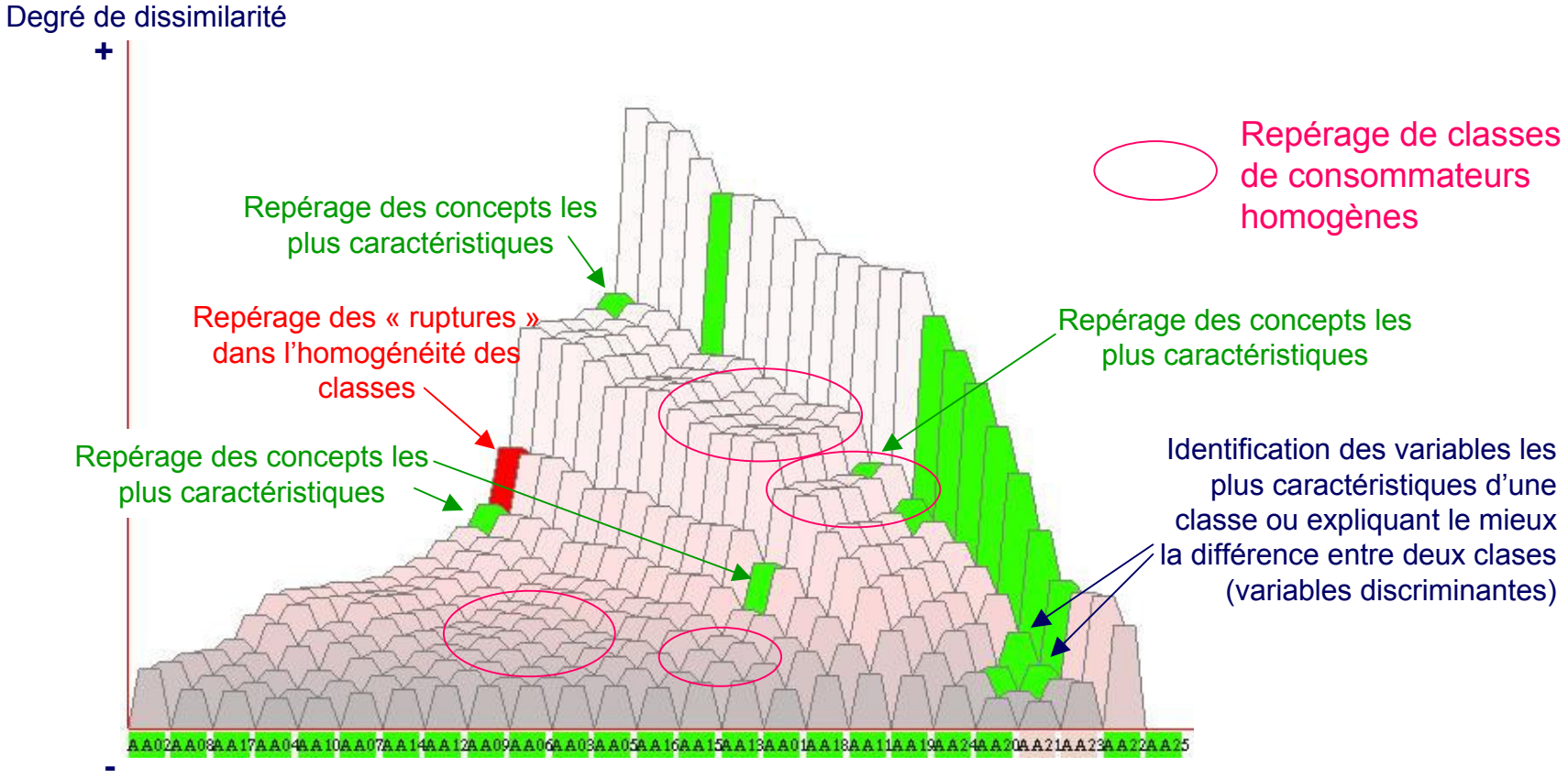
DESCRIPTION SYMBOLIQUE D'UNE CLASSE:
Homogènes, séparante et discriminante
S. Winsberg, L. Mehdi

Exemple:Trouver une description d'un profil de traces d'usagers web, en sous classes homogènes séparantes des autres profils et expliquant la durée du cheminement.



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

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



- **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.**

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!

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.

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.

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.

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

WILEY, 2007 (à paraître)

"Symbolic Data Analysis and the SODAS software."

E. Diday, M. Noirhomme , (Editors).

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 Toulouser. 442 pages.**

DIFFUSION DE L'ANALYSE DES DONNEES SYMBOLIQUES

EUROPE: 18 équipes de 9 pays européens ont réalisés SODAS (EUROSTAT)

ETATS UNIS: Un contrat de coopération avec la NSF + JASA

UNE REVUE INTERNATIONALE D'Analyse de Données Symboliques:

Electronical Journal of SDA (JSDA) at

www.jsda.unina2.it/newjsda/volumes/index.htm

UNIVERSITE DAUPHINE

ECOLES D'ANALYSE DE DONNEES SYMBOLIQUES

SITE www.ceremade.dauphine.fr/%7Etouati/sodas-pagegarde.htm

CREATION D'UNE ENTREPRISE: SYROKKO (un vent nouveau …) pour valoriser SODAS et l'ADS dans l'industrie et la fonction publique.

SITE www.syrokko.com

EPILOGUE

LA REIFICATION: Le terme réification vient du mot latin *res* qui veut dire « chose ». « Réifier » veut donc dire « chosifier ». « Un être humain adulte doit faire un effort considérable pour s'abstenir de découper le monde qui l'environne en « corps », en choses ou en objets physiques distincts et séparés les uns des autres. Un objet individuel possède ou exemplifie des propriétés grâce auxquelles on peut le percevoir, le reconnaître, le catégoriser et le conceptualiser ». (Pierre Jacob De l'intention(2004))

Ce qui est réifié ne peut être décrit complètement

De chaque objet nous observe l'infini….

Nous voulons regarder, le doute nous punit,

le doute, morne oiseau, nous bat de son aile

et l'infini s'enfuit d'une fuite éternelle…Rimbaud:

Quelques Références

- Afonso F., Billard L., E. Diday (2004) : Régression linéaire symbolique avec variables taxonomiques, Revue RNTI, Extraction et Gestion des Connaissances (EGC 2004),G.Hébrail et al. Eds, Vol. 1, p. 205-210, Cépadues, 2004.
- Afonso F., Diday E. (2005) : Extension de l'algorithme Apriori et des règles d'association aux cas des données symboliques diagrammes et intervalles, Revue RNTI, Extraction et Gestion des Connaissances (EGC 2005), Vol. 1, pp 205-210, Cépadues, 2005.
- Aristotle (IV BC): Organon Vol. I Catégories, II De l'interprétation. J. Vrin edit. (Paris) (1994).
- Arnault A., Nicole P. (1662) : La logique ou l'art de penser, Froman, Stutgart (1965).
- Appice A., D'Amato C., Esposito F., Malerba D. (2006): Classification of Symbolic Objects: A Lazy Learning Approach. Intelligent Data Analysis, 10 (4), 301 – 324
- .Bezerra B. L. D., De Carvalho F.A.T. (2004): A symbolic approach for content-based information filtering. Information Processing Letters, 92 (1), 45-52.
- Billard L. (2004): Dependencies in bivariate interval-valued symbolic data.. In: Classification, Clustering and New Data Problems . Proc. IFCS'2004. Chicago. Ed. D. Banks. Springer Verlag, 319-354.
- Billard L., Diday E. (2006): Symbolic Data Analysis: Conceptual Statistics and Data Mining. To be published by Wiley.

Billard L., Diday E. (2005): Histograms in symbolic data analysis 2005. Intern Stat. Inst. 55.
Billard L., Diday E. (2006) “ Descriptive Statistics for Interval-valued Observations in the Presence of Rules”. Computational Statistics 21;2.Pages 187-210.
Bravo Llatas M.C. (2004): Análisis de Segmentación en el Análisis de Datos Simbólicos. Ed. Universidad Complutense de Madrid. Servicio de Publicaciones. ISBN:8466917918. (http://www.ucm.es/BUCM/tesis/mat/ucm-t25329.pdf)
Brito, P. (2005) : Polaillon, G., Structuring Probabilistic Data by Galois Mathématiques et Sciences Humaines, 43ème année, nº 169, (1), pp. 77-104.
Brito, P. (2002): Hierarchical and Pyramidal Clustering for Symbolic Data, Journal of the Japanese Society of Computational Statistics, Vol. 15, Number 2, pp. 231-244.
Caruso C., Malerba D., Papagni D. (2005). Learning the daily model of network traffic. In M.S. Hacid, N.V. Murray, Z.W. Ras, S. Tsumoto (Eds.) Foundations of Intelligent Systems, 15th International Symposium, ISMIS'2005, Lecture Notes in Artificial Intelligence, 3488, 131-141, Springer, Berlin,
Germania. Cazes, P., Chouakria, A., Diday, E. Schektman, Y. (1997) Extension de l’analyse en composantes principales à des données de type intervalle, Revue de Statistique Appliquée XIV(3), 5–24.
Ciampi A., Diday E., Lebbe J., Perinel E., R. Vignes (2000): Growing a tree classifier with imprecise data. Pattern. Recognition letters 21. pp 787-803.

- De Carvalho F.A.T., Eufrasio de A. Lima Neto, Camilo P.Tenerio (2004): A new method to fit a linear regression model for interval-valued data. In: Advances in Artificial Intelligence: Proceedings of the Twenty Seventh German Conference on Articial Intelligence (eds. S. Biundo, T. Fruchrirth, and G. Palm). Springer-Verlag, Berlin, 295-306.De Carvalho F.A.T., De Souza R., Chavent M., Y. Lechevallier (2006): Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. Pattern Recognition Letters, 27 (3), 167-179
- De Carvalho F.A.T., Brito P., Bock H. H. (2006), Dynamic Clustering for Interval Data Based on L_2 Distance, Computational Statistics, accepted for publication.
- De Carvalho, F. A. T. (1995): Histograms In Symbolic Data Analysis. Annals of Operations Research,Volume 55, Issue 2, 229-322.
- De Souza, R. M. C. R. and De Carvalho, F. A. T. (2004): Clustering of Interval Data based on City-Block Distances. Pattern Recognition Letters, Volume 25, Issue 3, 353-365.
- Diday E. (1987 a): The symbolic aproach in clustering and related methods of Data Analysis. In "Classification and Related Methods of Data Analysis", Proc. IFCS, Aachen, Germany. H. Bock ed.North-Holland.
- Diday E. (1987 b): Introduction à l'approche symbolique en Analyse des Données. Première Journées Symbolique-Numérique. Université Paris IX Dauphine. Décembre 1987.

- Diday E. (1989): Introduction à l'Analyse des Données Symboliques. Rapport de Recherche INRIA N° 1074 (August 1989). INRIA Rocquencourt 78150. France.
- Diday E. (1991) : Des objets de l'Analyse des Données à ceux de l'Analyse des Connaissances. In « Induction Symbolique et Numérique à partir de données ». Y. Kodratoff, Diday E. Editors. CEPADUES-EDITION.ISBN 2.85428.282 5.
- Diday E. (2000): L'Analyse des Données Symboliques : un cadre théorique et des outils pour le Data Mining. In : E. Diday, Y. Kodratoff, P. Brito, M. Moulet "Induction symbolique numérique à partir de données". Cépadues. 31100 Toulouse. www.editions-cepadues.fr. 442 pages.
- Diday E. (2002): An introduction to Symbolic Data Analysis and the Sodas software. Journal of Symbolic Data Analysis. Vol. 1, n° 1. International Electronic Journal. www.jsda.unina2.it/JSDA.htm.
- Diday E., Esposito F. (2003): An introduction to Symbolic Data Analysis and the Sodas Software IDA. International Journal on Intelligent Data Analysis". Volume 7, issue 6. (Decembre).
- Diday E., Emilion R. (2003): Maximal and stochastic Galois Lattices. Journal of Discrete Applied Mathematics, Vol. 127, pp. 271-284.
- Diday E. (2004): Spatial Pyramidal Clustering Based on a Tessellation. Proceedings IFCS'2004, In Banks andal. (Eds.): Data Analysis, Classification and Clustering Methods Heidelberg, Springer-Verlag.

- Diday E., Vrac M. (2005): Mixture decomposition of distributions by Copulas in the symbolic data analysis framework. Discrete Applied Mathematics (DAM). Volume 147, Issue 1, 1 April, Pages 27-41.
- 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
- Diday E.(1995): Probabilist, possibilist and belief objects for knowledge analysis. Annals of Operations Research. 55, pp. 227-276.
- Diday E., Murty N. (2005): Symbolic Data Clustering. In Encyclopedia of Data Warehousing and Mining . John Wong editor . Idea Group Reference Publisher.
- Duarte Silva, A. P., Brito, P. (2006): Linear Discriminant Analysis for Interval Data,Computational Statistics, accepted for publication.
- Gioia, F. and Lauro, N.C. (2005) Basic Statistical Methods for Interval Data, Statistica applicata, 1.
- Gioia, F. and Lauro, N.C. (2006): Principal Component Analysis on Interval Data, Computational statistics, In press.
- Hardy, A. and Lallemand, P. (2002): Determination of the number of clusters for symbolic objects described by interval variables, In Studies in Classification, Data Analysis, and Knowledge Organization, Proceedings of the IFCS'02 Conference, 311-318.

- Hardy, A, Lallemand, P. and Lechevallier, Y. (2002) : La détermination du nombre de classes pour la méthode de classification symbolique SCLUST, Actes des Huitièmes Rencontres de la Société Francophone de Classification, 27-31
- Hardy, A. and Lallemand, P. (2004): Clustering of symbolic objects described by multi-valued and modal variables, In Studies in Classification, Data Analysis, and Knowledge Organization, Proceedings of the IFCS'04 Conference, 325-332
- Hardy, A. (2004): Les méthodes de classification et de determination du nombre de classes: du classique au symbolique, In M. Chavent, O. Dordan, C. Lacomblez, M. Langlais, B. Patouille (Eds), Comptes rendus des Onzièmes Rencontres de la Société Francophone de Classification, 48-55
- Hardy, A. (2005): Validation in unsupervised symbolic classification, Proceedings of the Meeting "Applied Stochastic Models and Data Analysis " (ASMDA 2005), 379-386
- Irpino, A. (2006): Spaghetti PCA analysis: An extension of principal components analysis to time dependent interval data. Pattern Recognition Letters, Volume 27, Issue 5, 504-513.
- Irpino, A., Verde, R. and Lauro N. C. (2003): Visualizing symbolic data by closed shapes, Between Data Science and Applied Data Analysis, Shader-Gaul-Vichi eds., Springer, Berlin, pp. 244-251.

- Lauro, N.C., Verde, R. and Palumbo, F. (2000): Factorial Data Analysis on Symbolic Objects under cohesion constrains In: Data Analysis, Classification and related methods, Springer-Verlag, Heidelberg
- M. Limam, E. Diday, S. Winsberg (2004): Symbolic Class Description with Interval Data. Journal of Symbolic Data Analysis, 2004, Vol 1
- D. Malerba, F. Esposito, M. Monopoli (2002): Comparing dissimilarity measures for probabilistic symbolic objects.In A. Zanasi, C. A. Brebbia, N.F.F. Ebecken, P. Melli (Eds.)Data Mining III, Series Management Information Systems, Vol 6, 31-40, WIT Press, Southampton, UK. Mballo C., Asseraf M., E. Diday (2004): Binary tree for interval and taxonomic variables. A Statistical Journal for Graduates Students"Volume 5, Number 1, April 2004.
- Milligan , G.W., Cooper M.C. (1985): An examination of procedures for determining the number of clusters in a data set. Psychometrica 50, 159-179.
- MenesesE., Rodríguez-Rojas O. (2006): Using symbolic objects to cluster web documents. WWW 2006: 967-968.

- Noirhomme-Fraiture, M. (2002): Visualization of Large Data Sets : the Zoom Star Solution, Journal of Symbolic Data Analysis, vol. 1, July.
- <http://www.jsda.unina2.it/>http://www.jsda.unina2.it
- Prudêncio R. B. C., Ludermir T., F. de A. T. De Carvalho (2004): A Modal Symbolic Classifier for selecting time series models.Pattern Recognition Letters, 25 (8), 911-921.
- Rodriguez O. (2000): "Classification et modèles linéaires en Analyse des Données Symboliques". Thèse de doctorat, University Paris 9 Dauphine.
- Schweizer B. (1985) "Distributions are the numbers of the futur" . Proc. sec. Napoli Meeting on "The mathematics of fuzzy systems". Instituto di Mathematica delle Faculta di Mathematica delle Faculta di Achitectura, Universita degli studi di Napoli. p. 137-149.
- Schweizer B. , Sklar A. (2005): Probabilist metric spaces . Dover Publications INC. Mineola, New-York.Soule A., K. Salamatian, N. Taft, R. Emilion (2004): “Flow classfication by histograms” ACM SIGMETRICS, New York. http://rp.lip6.fr/~soule/SiteWeb/Publication.php
- Stéphan V. (1998): "Construction d'objets symboliques par synthèse des résultats de requêtes". (1998). Thesis. Paris IX Dauphine University.
- Vrac M, Diday E., Chédin A. (2004) : Décomposition de mélange de distributions et application à des données climatiques. Revue de Statistique Appliquée, 2004, LII (1), 67-96.
- Vrac M, Diday E., Chédin A. (2004) : Décomposition de mélange de distributions et application à des données climatiques.Revue de Statistique Appliquée, 2004, LII (1), 67-96.