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 »

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**

DES INDIVIDUS AUX CONCEPTS: POURQUOI?

Parce que souvent c'est le concept qui est l'entité que l'on veut étudier!

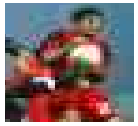
- EXEMPLE 1: pas les tickets de caisse mais les clients.
- EXEMPLE 2: Pas les déclarations d'accidents mais les assurés
- EXEMPLE 3: pas des traces web mais les entreprises qu'elles représentent

Approche Classique Versus Symbolique: les unités de l'étude

Classique : des individus

Oiseaux 

Habitants 

Joueurs de foot (Zidane,…) 

Images 

Articles vendus 

Traces d'usager WEB

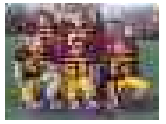
Feuilles de maladies

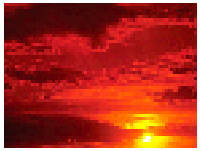
Abonnés GSM 

Symbolique : des concepts

Espèces d'oiseaux 

Régions d'habitation 

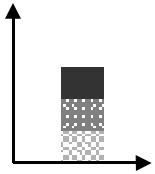
Joueurs d'Équipes (Lyon, …) 

Types d'image (marines,…) 

Magasins d'une chaîne 

Usagers

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'une image, d'un produit vendu, 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, les images maritimes, les produits vendus dans un des magasins d'une chaîne, un bénéficiaire d'assurance…

doit tenir compte de la variation des individus de son extension: les oiseaux de l'espèce, les images maritimes de la base, les produit vendu du magasin, les feuilles de maladies d'un bénéficiaire dans une période donnée….

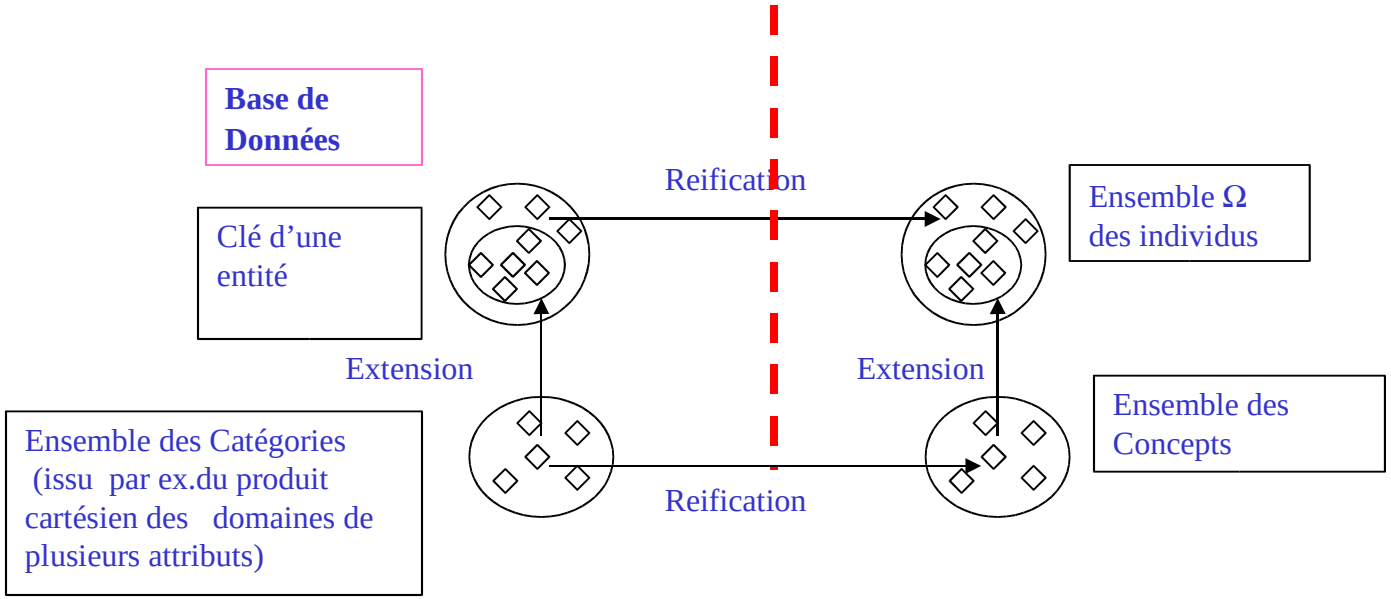
Taille = [20, 30], Couleur = {0.3Rouge, 0.7 vert}+ règles, Taxonomies

LA REIFICATION DES INDIVIDUS ET DES CONCEPTS

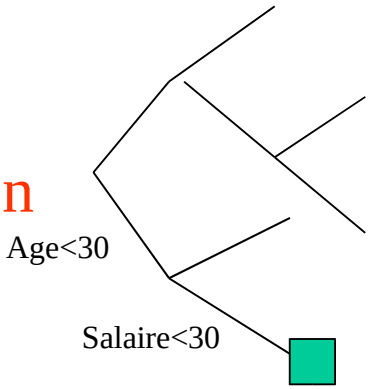
LES CATEGORIES SONT DEFINIES DE FACON EXHAUSTIVE PAR LES MODALITES D'UNE VARIABLE QUALITATIVE OU UN PRODUIT CARTESIEN DE TELLES VARIABLES.

LES **INDIVIDUS** ET LES **CONCEPTS** SONT CONSIDERES COMME DES ENTITES DU MONDE LEUR DESCRIPTION N'EST JAMAIS EXHAUSTIVE.

LA REIFICATION DES CATEGORIES EN CONCEPTS



- Exemple : les branches d'un arbre de décision constituent des catégories définies par des conjonctions de propriétés réifiables 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

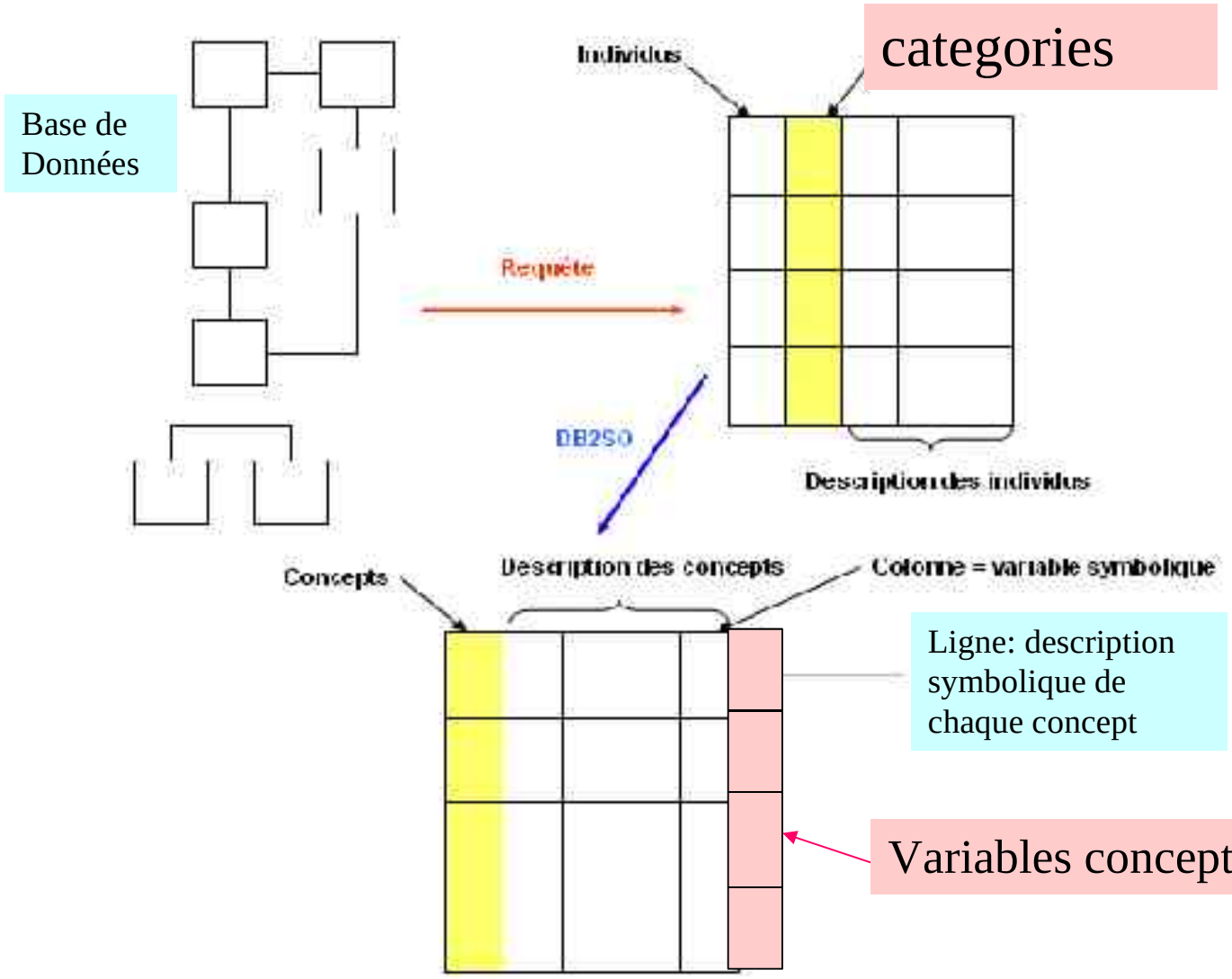
Généralisation

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

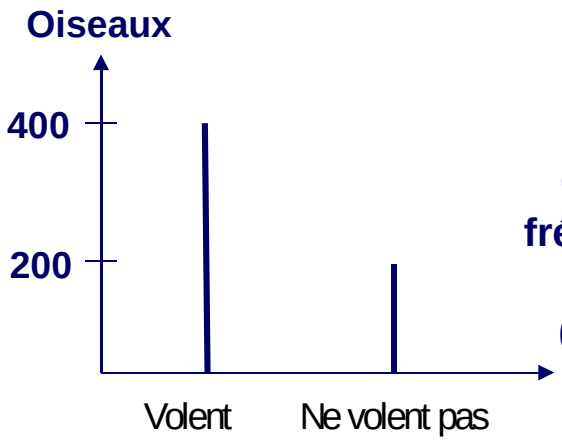
Espèce	Vole	Couleur	Taille	Migre
Hirondelle	Oui	0.3n,0.7gris	[60, 85]	Oui
Autruche	Non	0.1noir,0.9g	[85, 160]	Non
Pingouin	Non	0.5n,0.5gris	[70, 95]	Oui

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

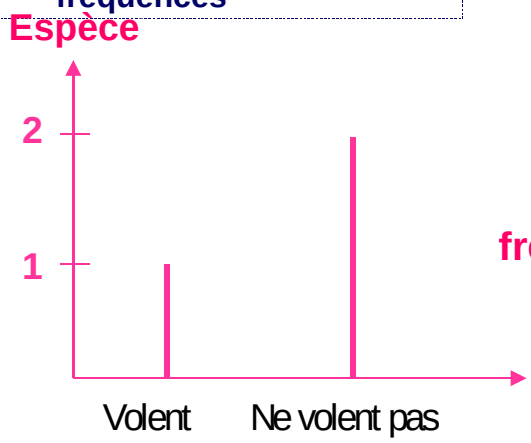
« L'espèce » est un concept qui devient la nouvelle unité statistique; on ne s'intéresse plus aux 600 individus en tant que tels

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:

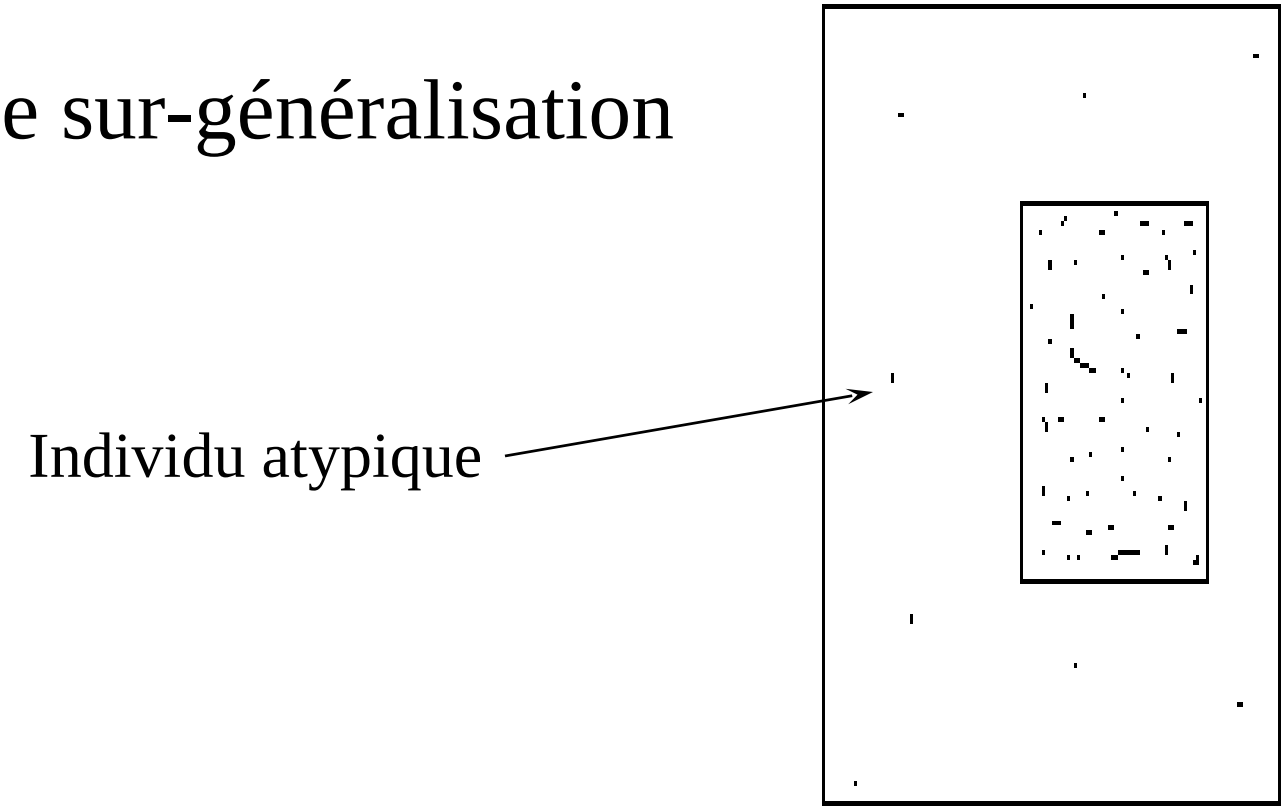
- une ou plusieurs valeurs qualitatives ou quantitatives

- 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
Dans chaque case

Classique

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

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

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

OBJECTIF

→ RÉUTILISER LE CONCEPT SUR UNE AUTRE BASE,

→ IDENTIFIER UN INDIVIDU DE SON EXTENSION,

→AMELIORER PAR APPRENTISSAGE SA MODELISATION,

RÉDUIRE LES DONNÉES

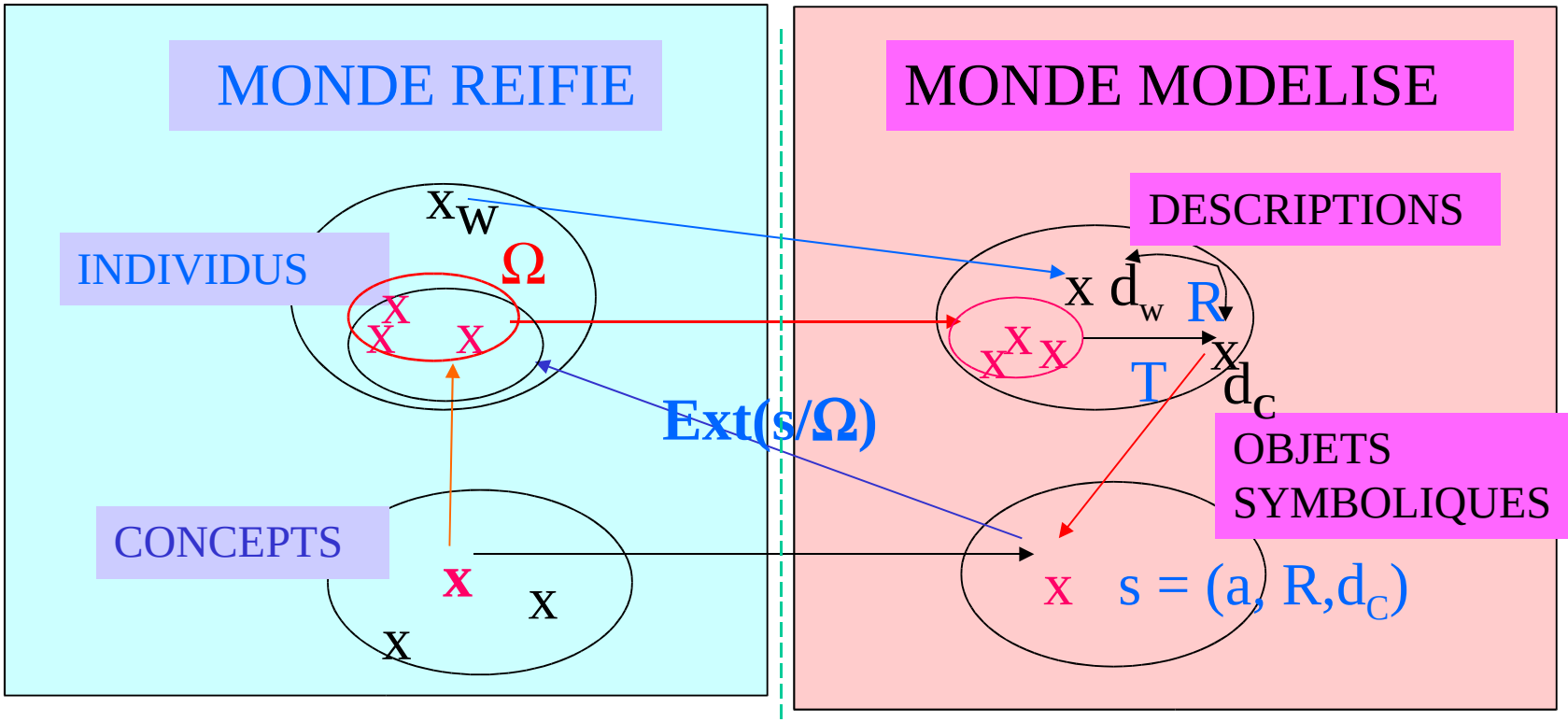
→MASSIVES,

→MANQUANTES

→ LA CONFIDENTIALITÉ

EN SE DONNANT LA POSSIBILITE DE LES RETROUVER.

MODÉLISATION ET APPRENTISSAGE DES CONCEPTS PAR QUATRE ESPACES



Exemple: concept= les autruches, Ω : base de données décrivant des oiseaux, contenant 3 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

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

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

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]	[1.80, 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	1.80	1.95	0. 7	0.3

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

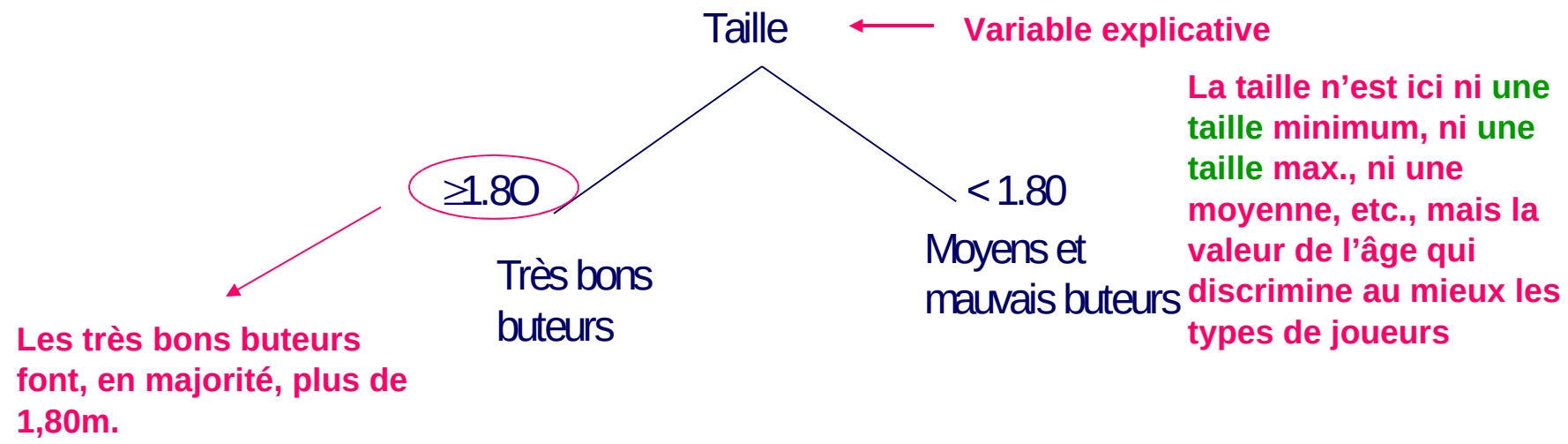
Exemple 1: Perte d'information du codage classique de données symboliques

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 concepts décrits par SODAS

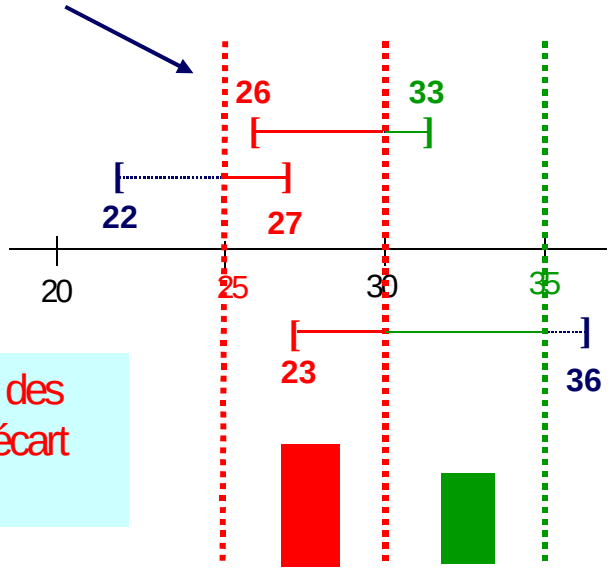


EXEMPLE 2: Perte d'information du codage classique de données symboliques

HISTOGRAMME . L'approche classique perd la notion de variable symbolique et ne permet de construire un histogramme que sur les min les max

Articles vendus dans la classe A de magasins	**[22,27]**
Articles vendus dans la classe B de magasins	**[26,33]**
Articles vendus dans la classe C de magasins	**[23, 36]**

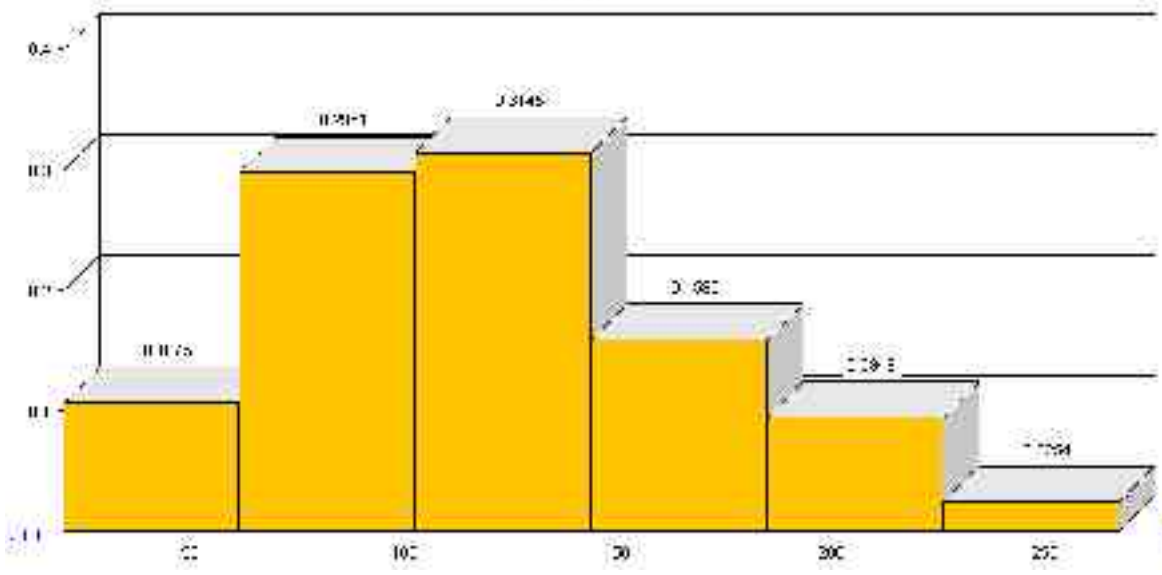
Construction d'un histogramme symbolique



La barre [25, 30] représente la somme des portions d'intervalles entrant dans cet écart

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

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



Age

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

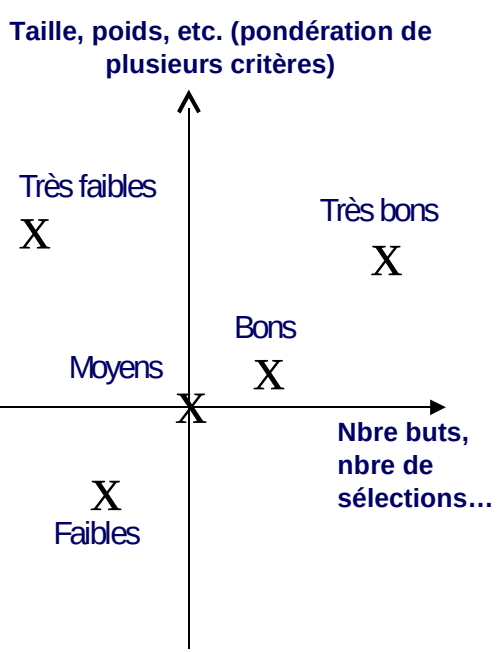
EXEMPLE 3 : Analyse en composantes principales

En codage classique, chaque concept est représenté par un point

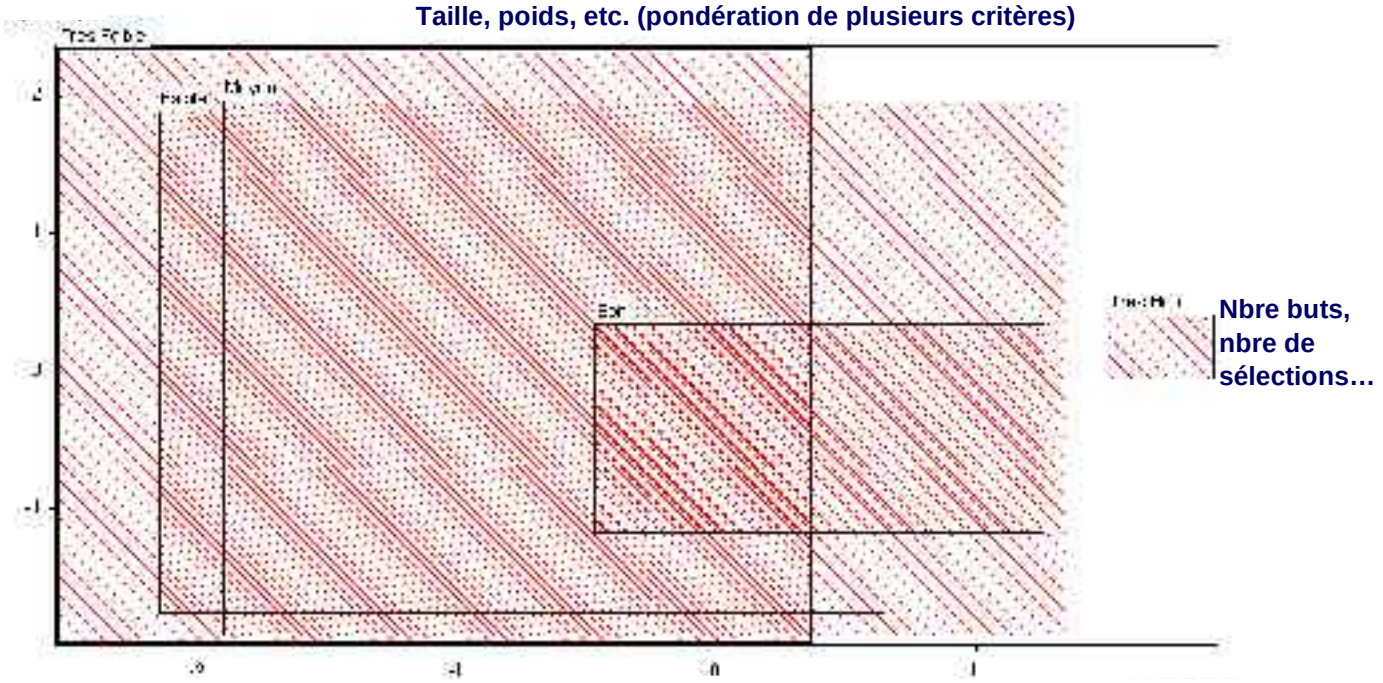
En codage symbolique :

→ 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

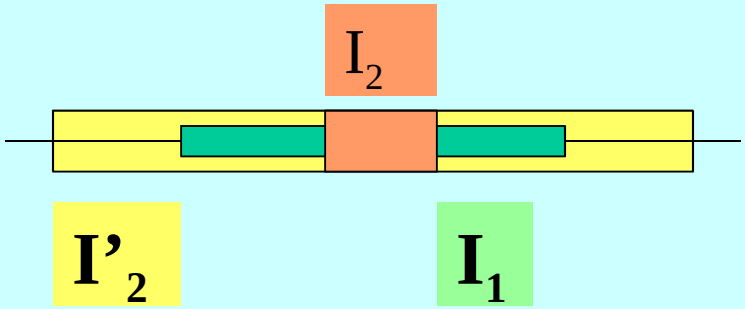
EXEMPLE 4: DISSIMILARITES Classique versus symbolique

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.

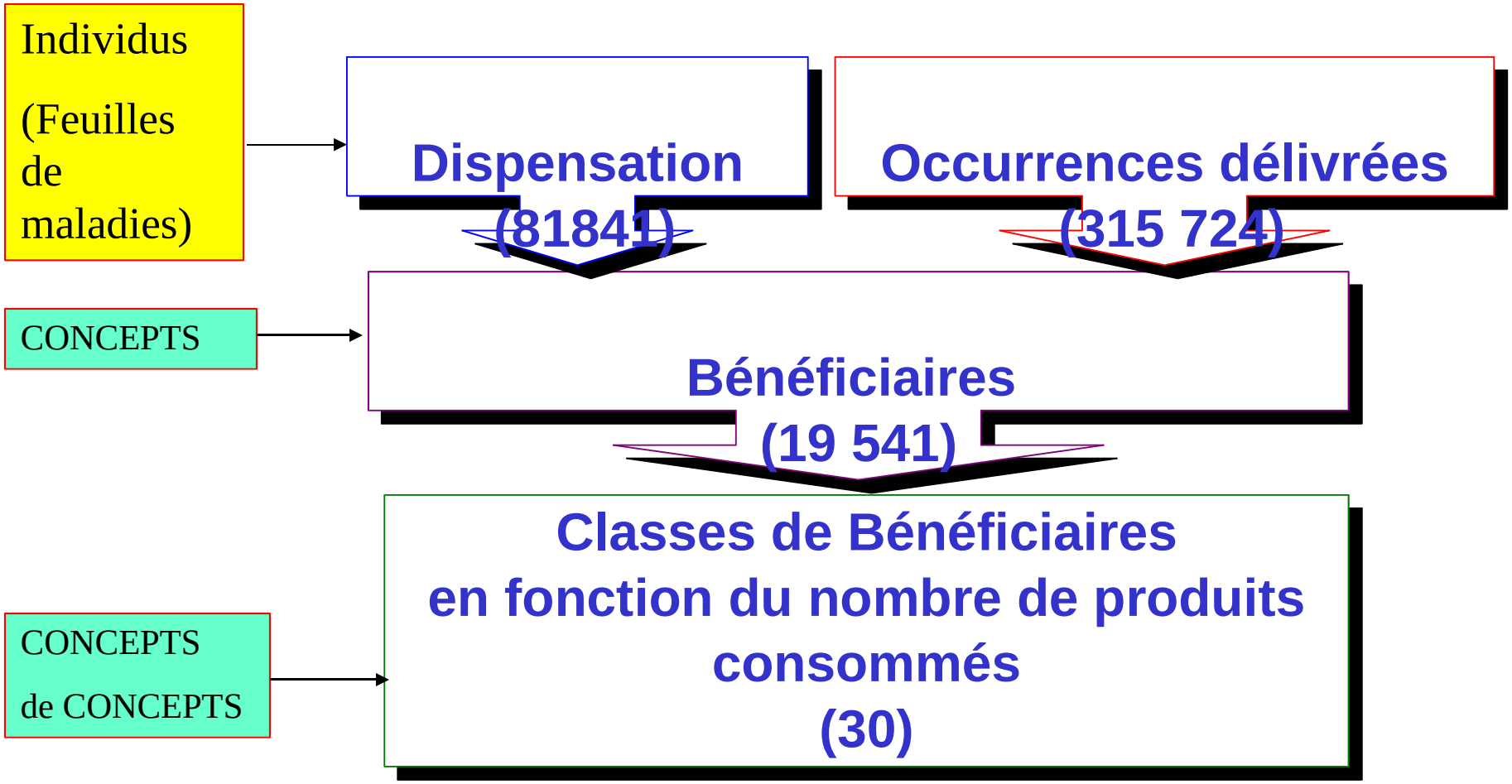
→ Elles sont décrites par des variables classiques parfois munies de variables cibles à expliquer.

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.

DONNEES COMPLEXES

- Données Incomplètes
- Données spatio-temporelles
- Trajectoires
- Images
- Textes
- Séquences
- Imprécises
- Floues, possibilistes, crédibiliste,…
- Structurées
- Fonctionnelles
- ----------

Dans tous ces cas le passage des individus aux concepts fait apparaître de la variation exprimable par des données symboliques.

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 ….).

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

	Catégor	Image	Texte	Séqu.
i1	C_j		doc1	agbdc
--	------	--------	--------	
in	C_k		doc$_n$	dgabh

	Image	Texte	Séqu
C_1	{image}$_1$	{doc}$_1$	{gba}$_1$
---	--------	--------	
C_k	{image}$_k$	{doc}$_k$	{ahd}$_k$

Description classique d'objets complexes

Exemple: C_i = images maritimes

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

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

Généralisation

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:

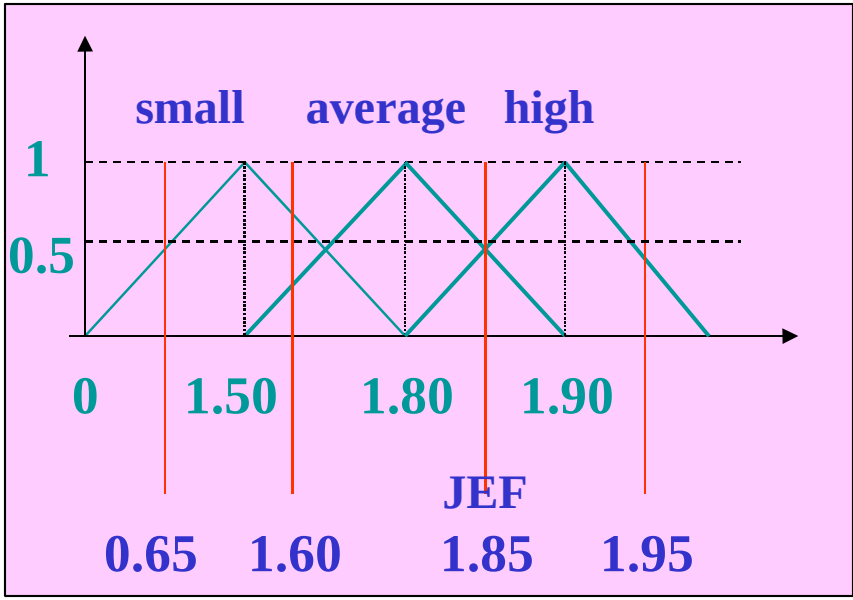
Le passage aux concepts exprime la variation des données floues.

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

Initial Data

From Numerical to Fuzzy Data



	height			weight	hair
	small	average	high		
Paul	0.70	0.30	0	45	yellow
Jef	0	0.50	0.50	80	yellow
Jim	0.50	0	0	30	black

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, 0.40]	[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 symbolique des foyers] ∧ [caractéristiques symbolique des écoles]

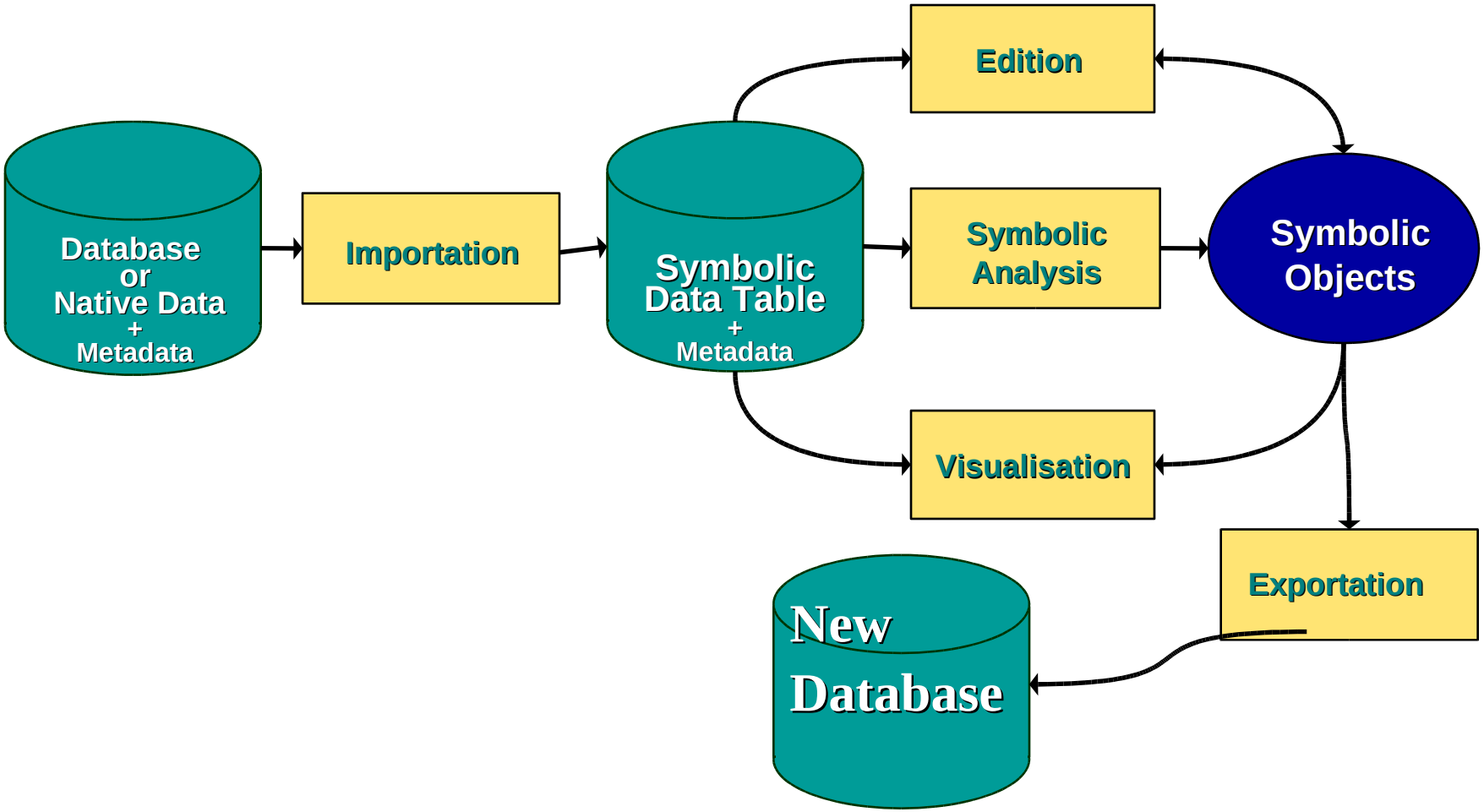
Données classiques versus Données symboliques

Le cas des DONNEES CONFIDENTIELLES

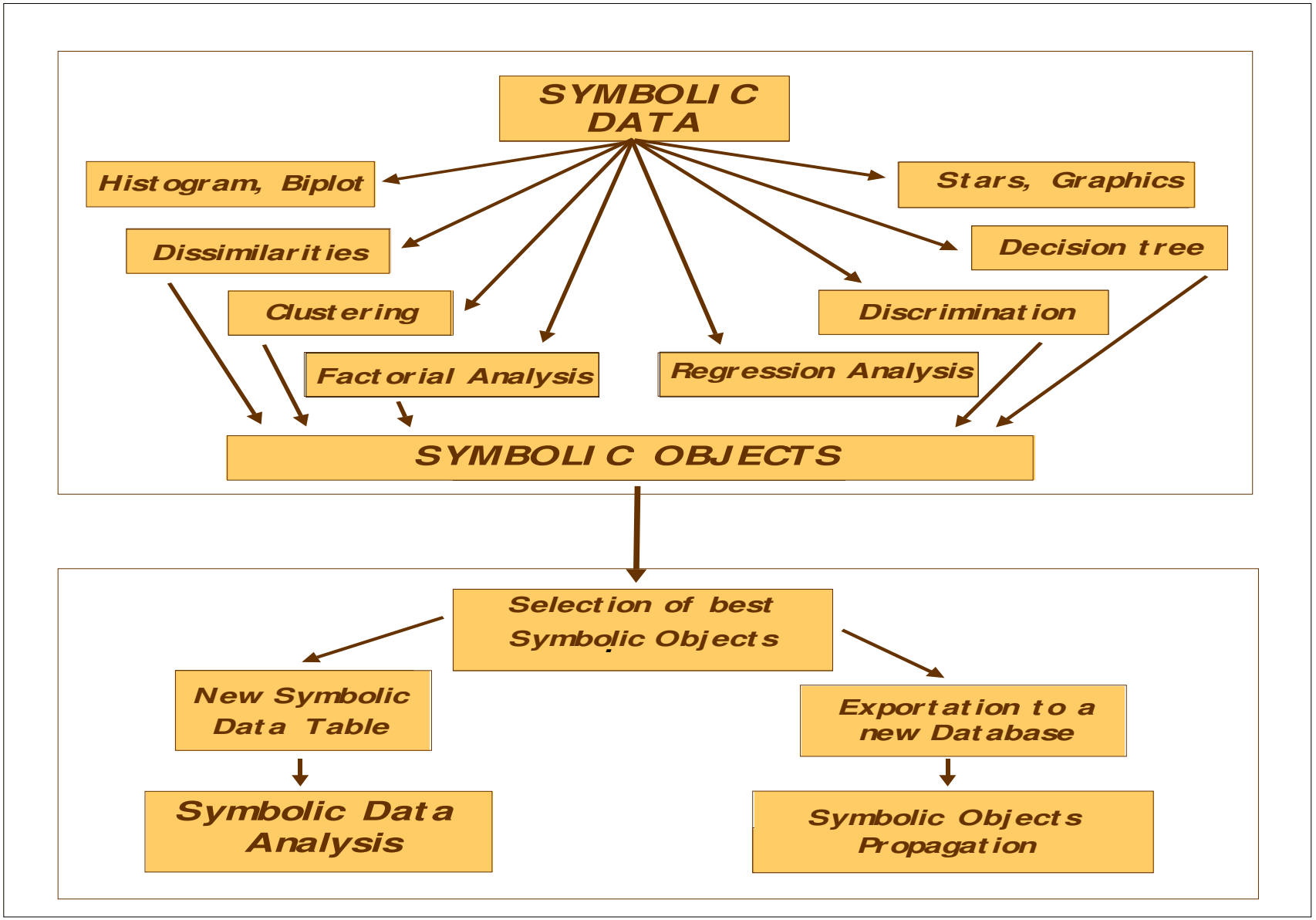
Approche Classique: les individus sont décrits par des données confidentielles

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

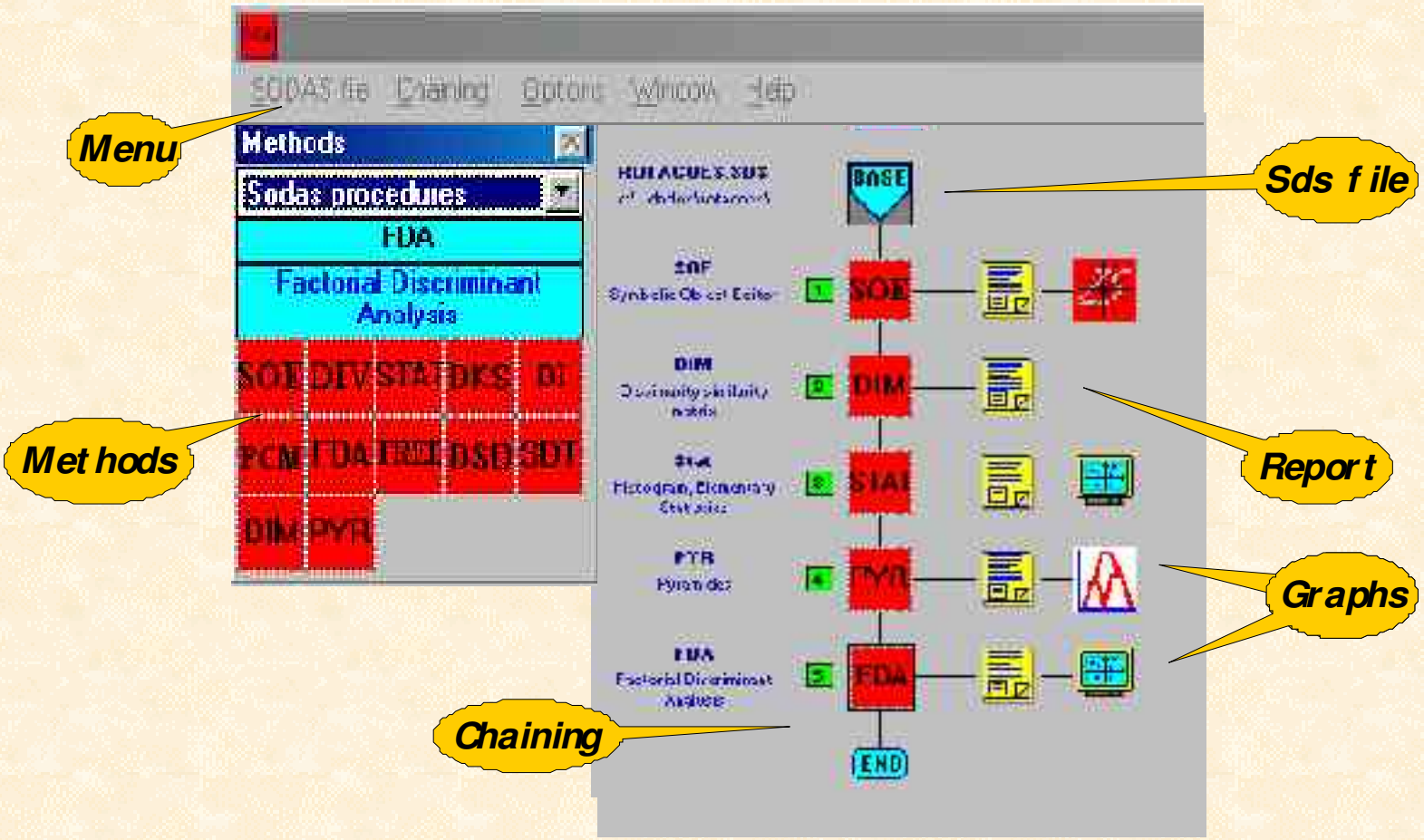
ASSO Architecture



THE SODAS 2 SOFTWARE FROM ASSO



SODAS Software

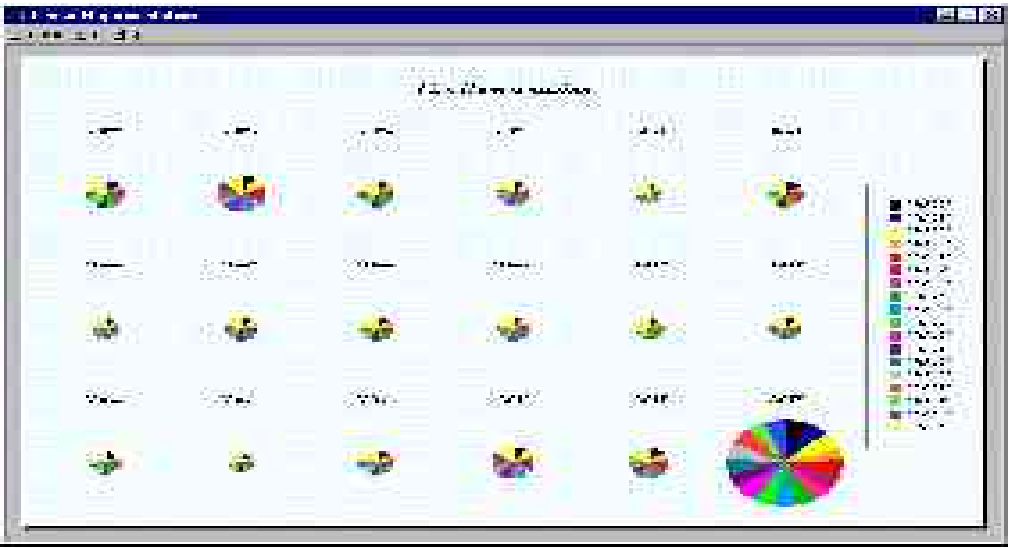


Extension de Méthodes classiques aux concepts

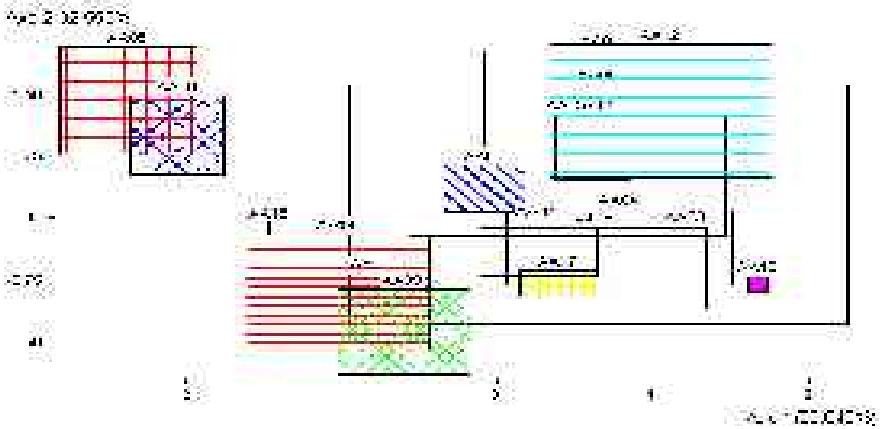
- Histos, correlation
- Visualisation en Etoiles
- Biplots
- ACP, AFC
- Décomposition de mélanges
- Multidimensional Scaling
- Typologie (Nuées Dynamiques , Pyramides, hiérarchies)
- Régression
- Réseaux neuronaux
- Arbres de Décision
- Treillis de Galois etc.

Autres exemples de méthodes de SODAS

CARTE DE KOHONEN DE CONCEPTS

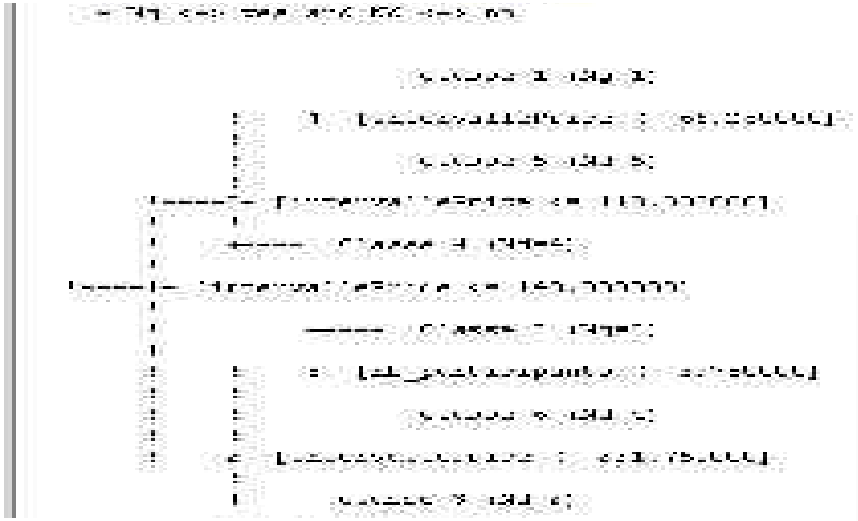


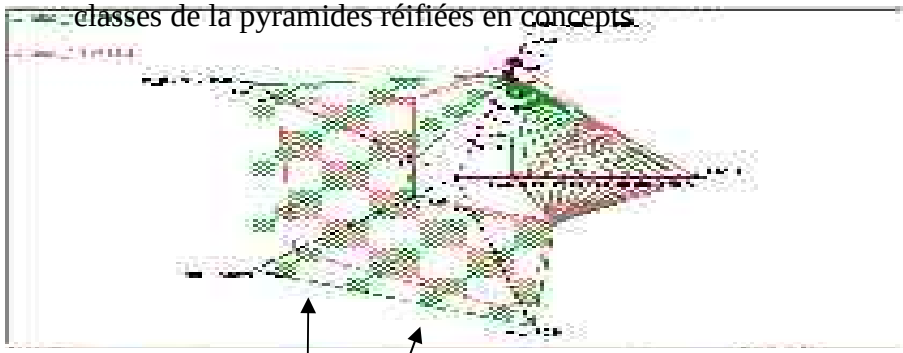
ANALYSE FACTORIELLE: ACP



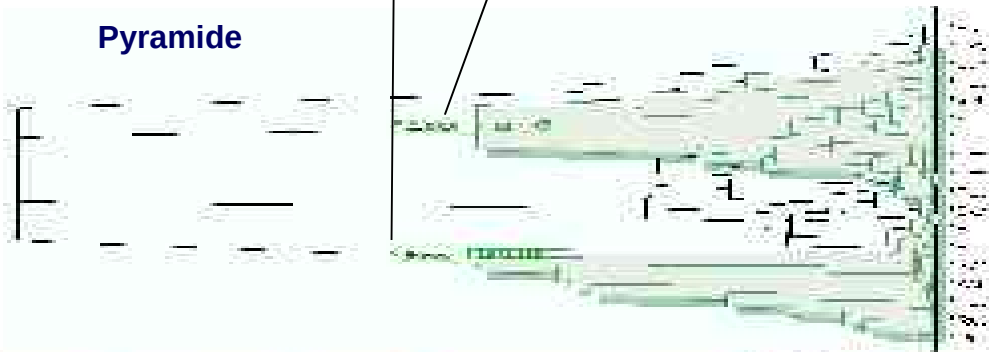
Superposition de deux deux étoîles associées à deux classes de la pyramides réifiées en concepts

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





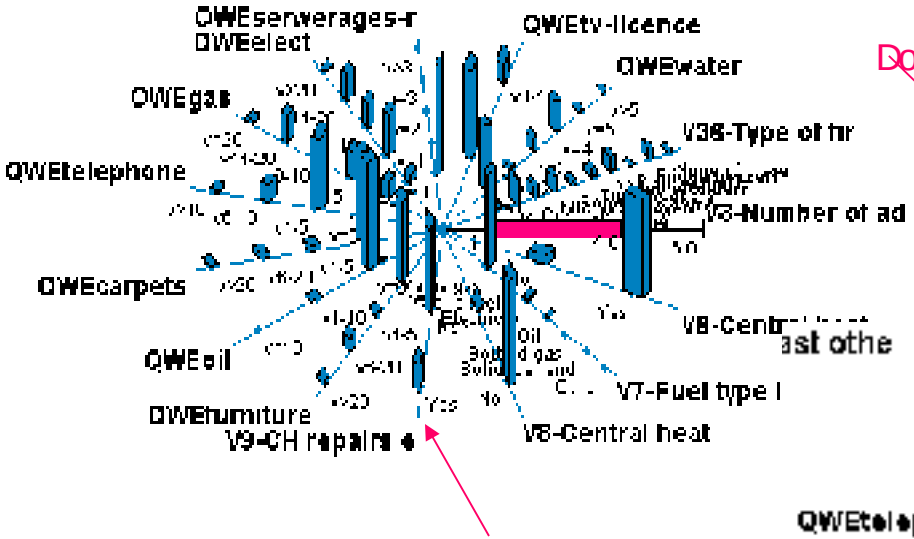
Pyramide



Description symbolique d'un concept : l'étoile

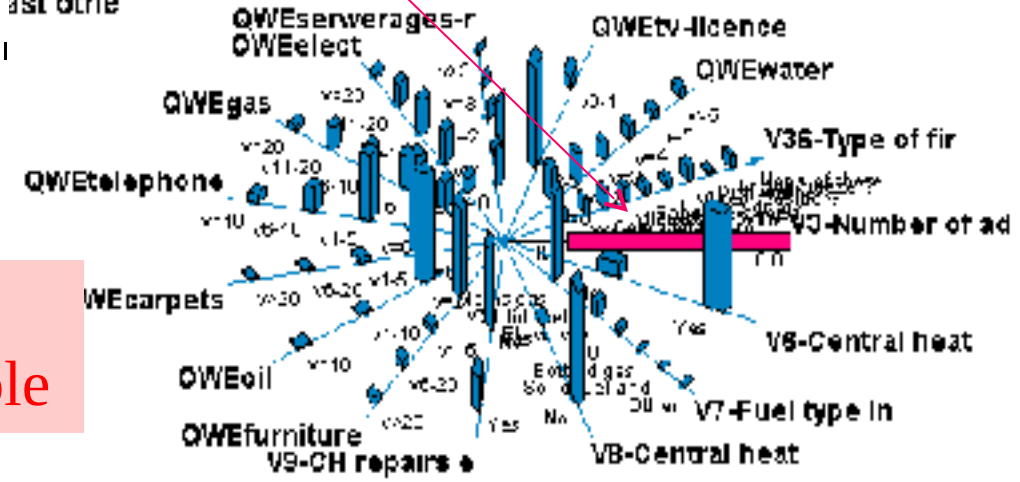
Ex. : La consommation d'énergie (électricité,gaz, pétrole) de foyers en Angleterre, par région

Northern metropolitan



Données quantitatives

South East



Données qualitatives

Chaque étoîle est associée à un concept . Chaque axe à une variable

Une telle représentation, de concepts, est un point de départ ; toutes les méthodes de STAT, Data Mining, AD, peuvent ensuite être étendues à des ensembles d'étoiles munies de connaissances supplémentaires :

- Analyse factorielle (ACP)
- Classification (hiérarchies, Pyramides)
- Nuées Dynamiques, Etc.
- Corrélation
- Discrimination
- Segmentation (arbres de décision)
- Cartes de Kohonen
-Réseaux de neuronnes

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

ASSURANCES SOCIALES (CCSMA)

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

Bénéficiaire	Spéc. 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

Les variables sont à valeur « variable aléatoire »

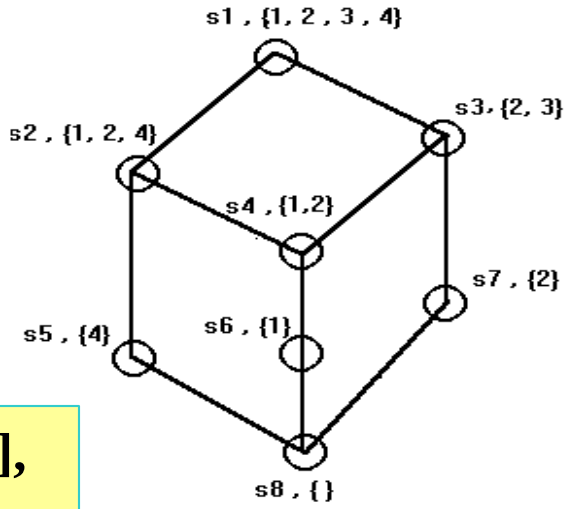
Théories utiles pour l'ADS

- Capacités de Choquet
- Copules de Sklar
- Topologie de Hausdorff
- Algèbre des intervalles
- Ensembles aléatoires
- Dissimilarités entre distributions

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

CLASSIQUE

Moyennes

K-means

Dissimilarité standard

(Euclidienne, KHI2…)

SYMBOLIQUE

Prototypes (= Obj Symb)

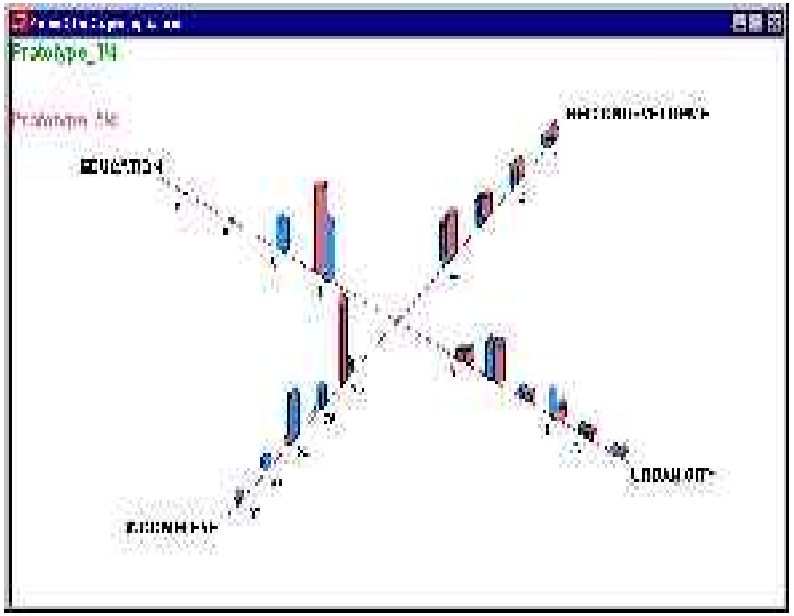
Nuées Dynamiques

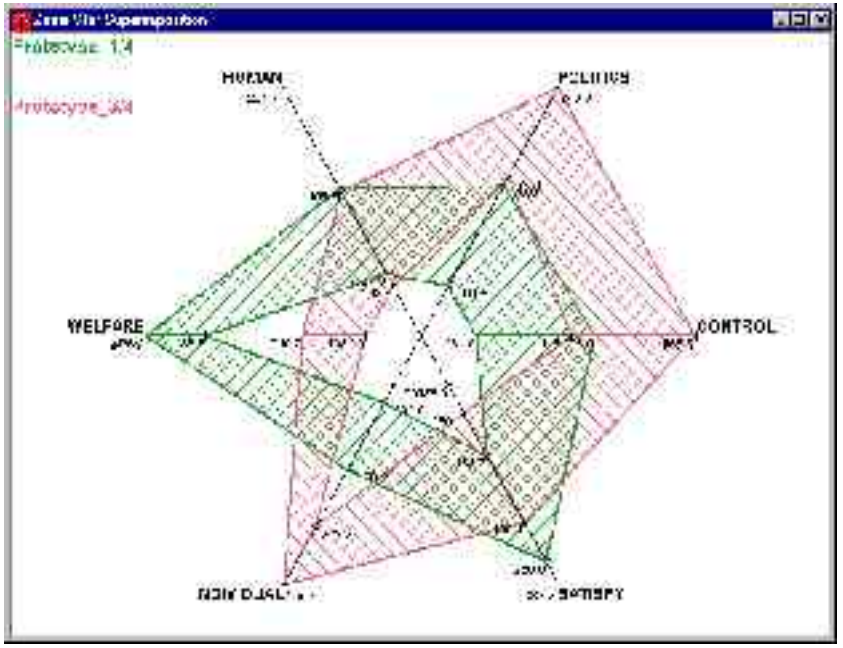
Dissimilarité symbolique

(Haussdorf, Ichino, De Carvalho,…)

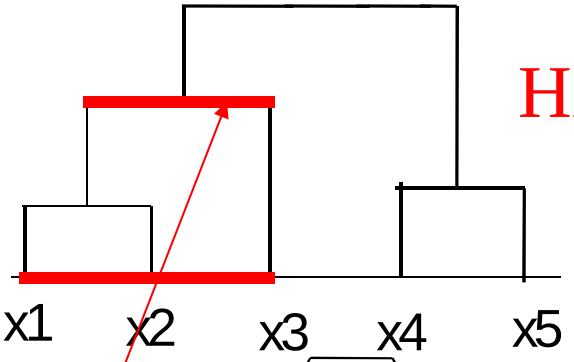
Représentation graphique des prototypes

Comparison entre les classes 1 et 3





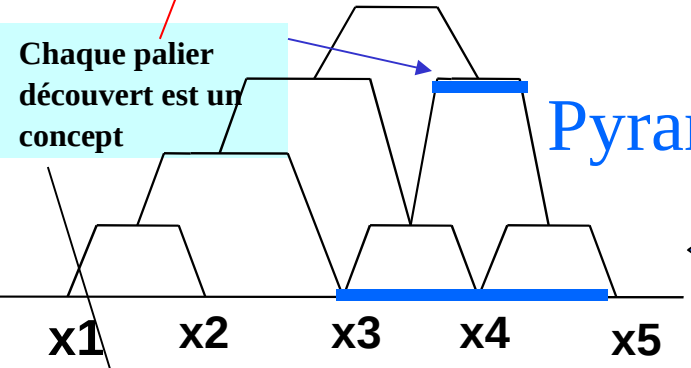
QUALITE DE LA REPRESENTATION SPATIALE



Hierarchies

Ultrametric dissimilarities = U

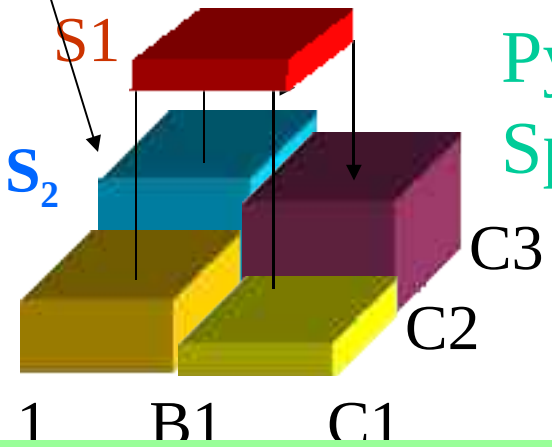
$$W = |d - U|$$



Pyramides

Robinsonian dissimilarities = R

$$W = |d - R|$$



Pyramides Spatiales

Yadidean dissimilarities = Y

$$W = |d - Y|$$

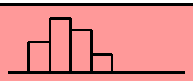
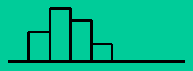
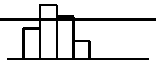
Meilleure qual. avec moins de paliers

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?

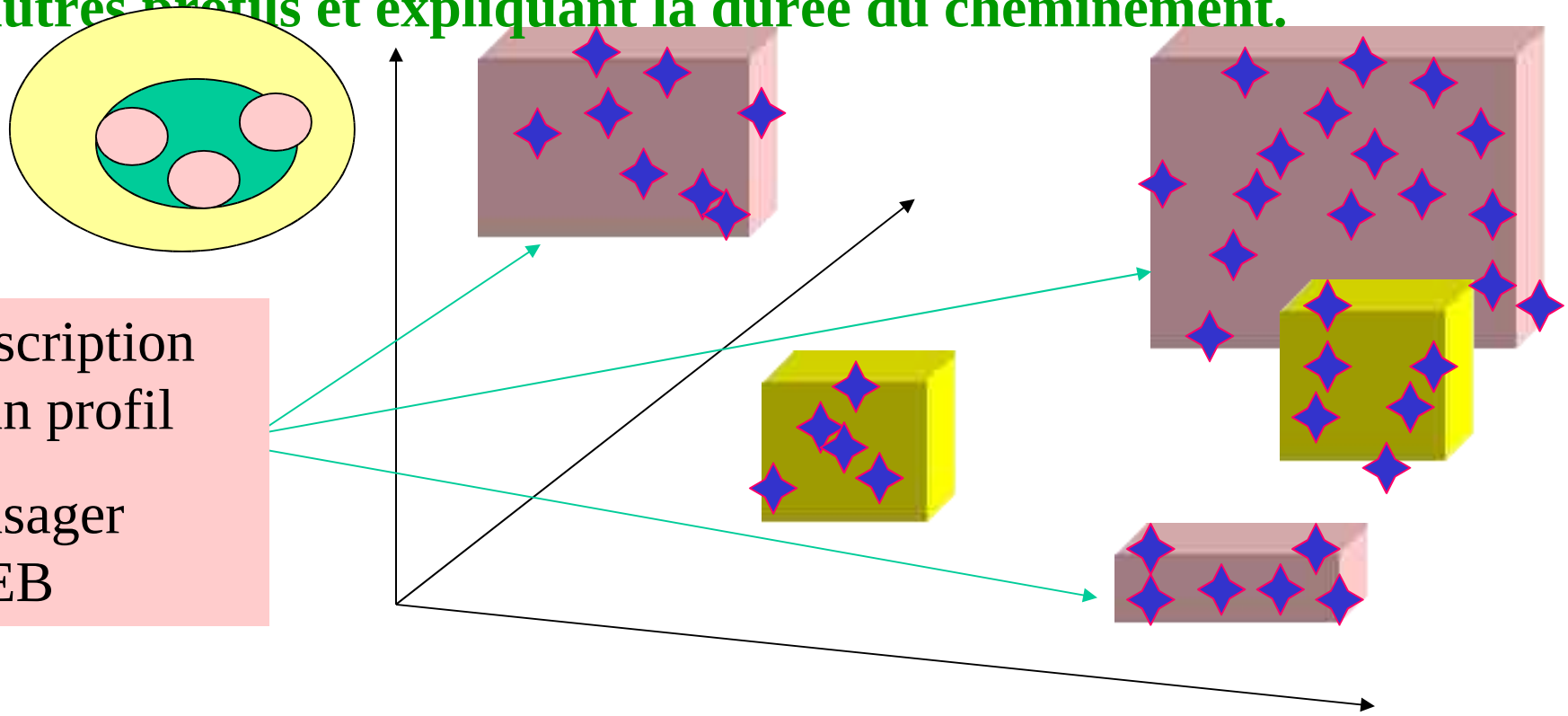
Indiv	Concept	opht	dent	lieu	période
i1	C1	12.5	3	Lyon1	
i2	C1	9.6	2	Paris 3	
i3	C1	11.4	4	Paris 3	
i4	C2	3.2	1		
i5	C2	7.1	4		

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

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:
En sous-classes Homogènes, séparante et discriminante

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



Description d'un profil

d'usager WEB

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 plus ou moins complexes décrits par des données classiques: on manque de bras!

MORALITÉ

Ne plus penser seulement en termes de variables et d'individus classiques.

Penser aussi en termes de catégories, concepts et variables symboliques.

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 aussi des catégories d'individus (induits par des variables qualitatives intéressantes ou une typologie) à étudier en tant que concepts réifiés.

- 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.

References

SPRINGER, 2000 :

“Analysis of Symbolic Data”

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

JASA (Journal of the American Statistical Association)

“From the Statistic of Data to the Statistic of Knowledge: Symbolic Data Analysis” L. Billard, E. Diday June, 2003 .

Electronic Journal of S. D. A.: JSDA

E. Diday, R. Verde, Y. Lechevallier

Download SODAS and SODAS information :

www.ceremade.dauphine.fr/˜touati/sodas-pagegarde.htm

E. Diday (1995) "Probabilist, possibilist and belief objects for knowledge analysis"Annals of Operations Research. 55,pp. 227-276.

E. Diday (2000) "Analyse des données symboliques: théorie et outil pour la fouille de connaissances" TSI (Technique et Science Informatiques). Vol 19, n°1-2-3 , Janvier

Diday E., Kodratoff Y., Brito P., Moulet M. (2000): "Induction symbolique numérique à partir de données". Cépadues. 31100 Toulouse. www.editions-cepadues.fr. 442 pages.

Stephan V., G. Hébrail, Y. Lechevallier (2000) "Generation of symbolic objects from Relational Databases" Chapt. 5 In Bock, Diday, Springer Verlag.

E. Diday, R. Emilion (1997) "Treillis de Galois Maximaux et Capacités de Choquet" . C.R. Acad. Sc. t.325, Série 1, p 261-266. Présenté par G. Choquet en Analyse Mathematique.

Diday E., Emilion R. (2003) "Maximal stochastic Galois Lattice". DAM. Journal of Discrete Applied Mathematics . Volume 127, issue 2 , 15 April 2003, Pages 271-284.

Diday E. (2004) " Spatial Clustering " Proc IFCS'2005. Chicago. 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

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 …)

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 (2004))

Ce qui est réifié ne peut être décrit complètement

Rimbaud:

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…