Data mining: une nouvelle façon de faire de la statistique?

Gilbert Saporta
Chaire de Statistique Appliquée
Conservatoire National des Arts et Métiers
292 rue Saint Martin, 75003 Paris
saporta@cnam.fr
http://cedric.cnam.fr/~saporta

Introduction

- L'information est la ressource du XXI ème siècle et la statistique un des métiers essentiels de son traitement.
- Le Data Mining (synonymes: *Fouille de données, extraction de connaissances ou KDD)* en est un avatar: nouveau champ d'application à l'interface de la statistique et des technologies de l'information (bases de données, intelligence artificielle, apprentissage etc.).

Introduction (suite)

- L’objectif: découvrir des structures et des « patterns » dans des grandes bases de données.
- Relations entre Data Mining et Statistique

Plan de la présentation

- **1.Quelques définitions du Data Mining**
- **2. Objectifs et outils**
- **3.Nouvelles mines : textes, web, données symboliques…**
- **4.Data Mining et statistique officielle**
- **5.Le Data Mining est-il de la statistique?**
- **6.Conclusions et perspectives**

1.Qu'est-ce que le Data Mining ?

- U.M.Fayyad, G.Piatetski-Shapiro : *" Data Mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data "*
- D.J.Hand : *" I shall define Data Mining as the discovery of interesting, unexpected, or valuable structures in large data sets"*

- La métaphore du Data Mining signifie qu'il y a des trésors ou pépites cachés sous des montagnes de données que l'on peut découvrir avec des outils spécialisés.
- Le Data Mining analyse des données recueillies à d'autres fins: c'est *une analyse secondaire* de bases de données, souvent conçues pour la gestion de données individuelles (Kardaun, T.Alanko,1998)
- Le Data Mining ne se préoccupe donc pas de collecter des données de manière efficace (sondages, plans d'expériences) (Hand, 2000)

Est-ce nouveau? Est-ce une révolution ?

- L'idée de découvrir des faits à partir des données est aussi vieille que la statistique *"Statistics is the science of learning from data. Statistics is essential for the proper running of government, central to decision making in industry,and a core component of modern educational curricula at all levels "* (J.Kettenring, 1997, ancien président de l 'ASA).

- Dans les années 60: Analyse Exploratoire (Tukey, Benzecri..)« *L'analyse des données est un outil pour dégager de la gangue des données le pur diamant de la véridique nature.* » (J.P.Benzécri 1973)

le Data Mining est né de :

- L'évolution des SGBD vers l'informatique décisionnelle avec les entrepôts de données (Data Warehouse).
- La constitution de giga bases de données : transactions de cartes de crédit, appels téléphoniques, factures de supermarchés: terabytes de données recueillies automatiquement.
- Développement de la Gestion de la Relation Client (CRM)
  - Marketing client au lieu de marketing produit
  - Attrition, satisfaction, etc.
- Recherches en Intelligence artificielle, apprentissage, extraction de connaissances
- Mais aussi une entreprise commerciale...

2.Objectifs et outils

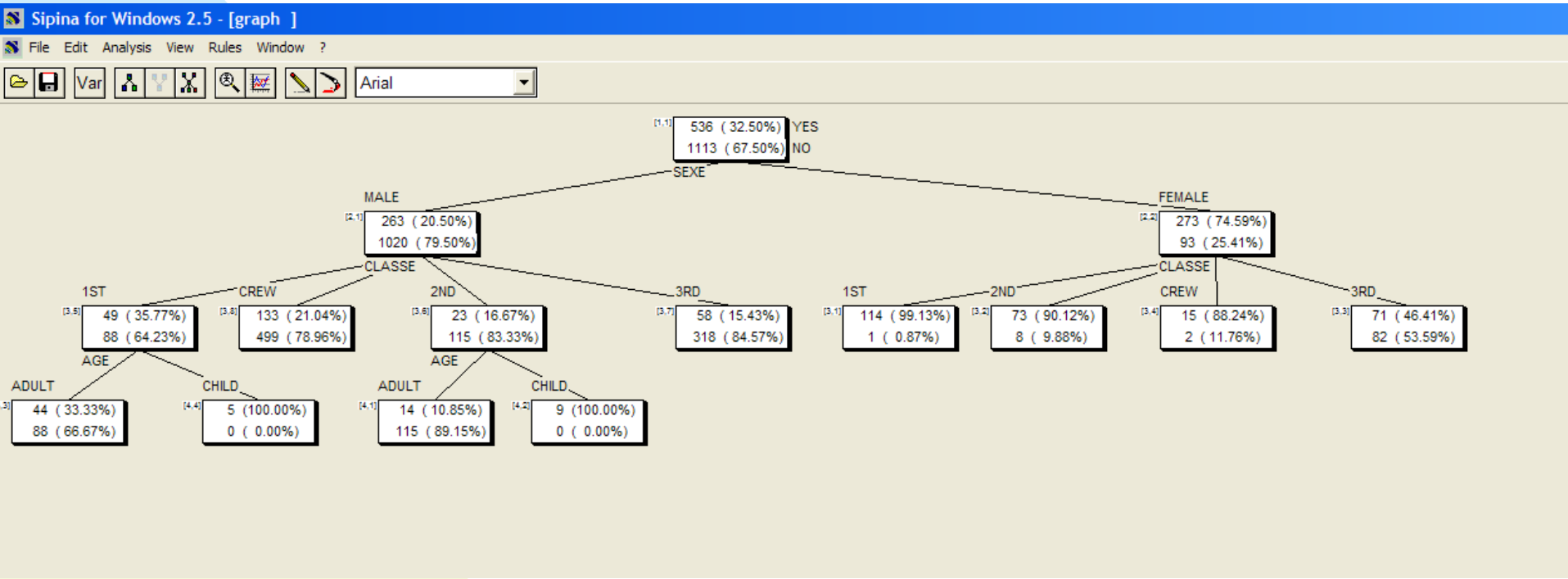
- Découvrir des structures dans les données.
- Deux types: modèles et « patterns » (ou comportements) *(D.Hand)*
- Autre distinction: prédictif (supervisé) ou exploratoire (non supervisé)

2.1 Modèles

Construire des modèles a toujours été une activité des statisticiens.Un modèle est un résumé global des relations entre variables, permettant de comprendre des phénomènes, et d'émettre des prévisions. « *Tous les modèles sont faux, certains sont utiles* » *(G.Box)*

- Le DM ne traite pas d'estimation et de tests de modèles préspécifiés, mais de la découverte de modèles à l'aide d'un processus de recherche algorithmique d'exploration de modèles:
  - linéaires ou non,
  - explicites ou implicites: réseaux de neurones, arbres de décision, SVM, régression logistique, réseaux bayesiens….
- Les modèles ne sont pas issus d'une théorie mais de l'exploration des données.

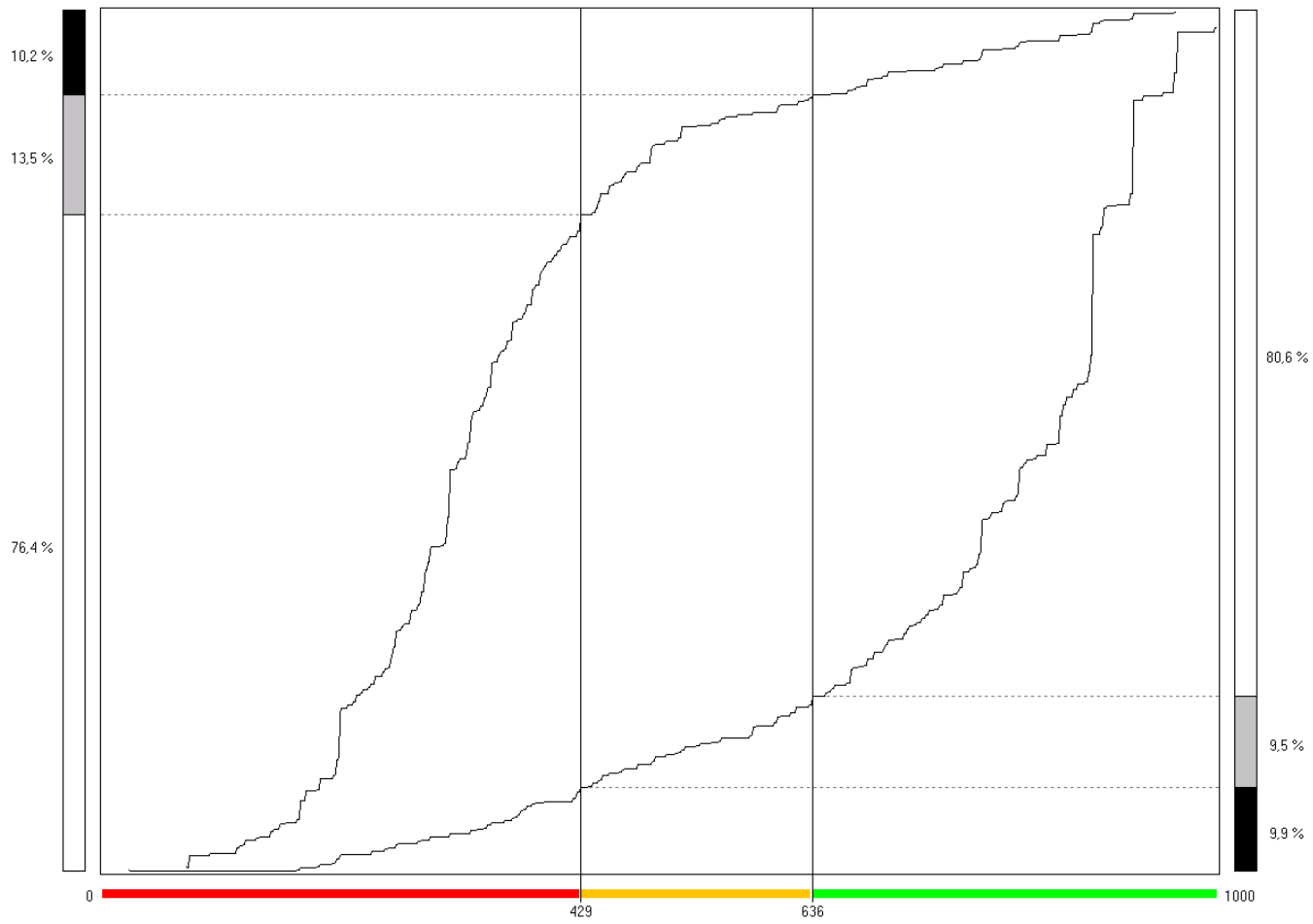
2.1.1 Arbres de décision



2.1.2 Scores

Ex: risque automobile en Belgique

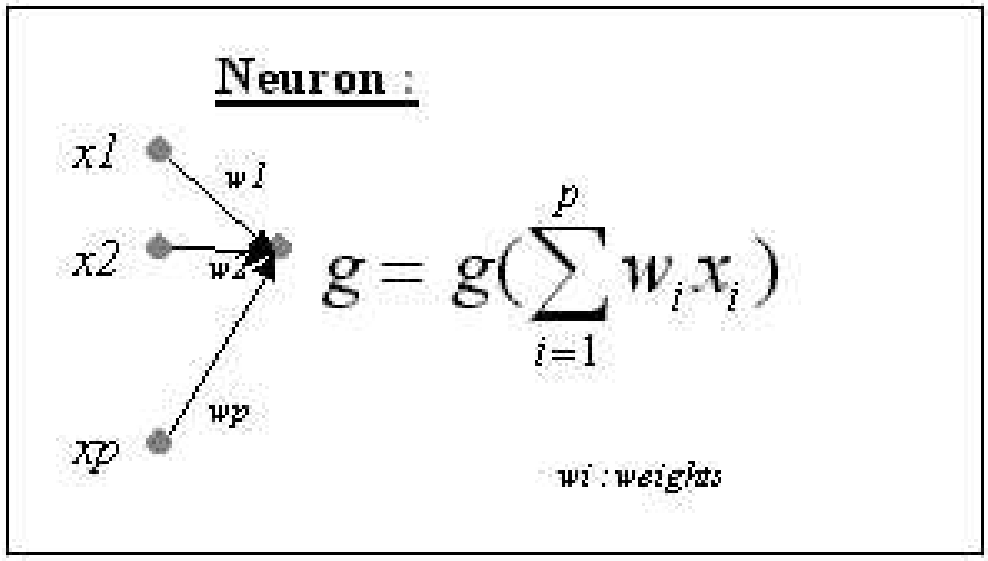
CATEGORIES	COEFFICIENTS DISCRIMINANT FUNCTION	TRANSFORMED COEFFICIENTS (SCORE)
2 . Use type		
USE1 - Profess.	-4.577	0.00
USE2 - private	0.919	53.93
4 . Gender		
MALE - male	0.220	24.10
FEMA - female	-0.065	21.30
OTHE - companies	-2.236	0.00
5 . Language		
FREN – French	-0.955	0.00
FLEM - flemish	2.789	36.73
24 . Birth date		
BD1 - 1890-1949 BD	0.285	116.78
BD2 - 1950-1973 BD	-11.616	0.00
BD? - ???BD	7.064	183.30
25 . Region		
REG1 - Brussels	-6.785	0.00
REG2 – Other regions	3.369	99.64
26 . Level of bonus-malus		
BM01 - B-M 1 (-1)	17.522	341.41
BM02 - Others B-M (-1)	-17.271	0.00
27 . Duration of contract		
C<86 - <86 contracts	2.209	50.27
C>87 - others contracts	-2.913	0.00
28 . Horsepower		
HP1 - 10-39 HP	6.211	75.83
HP2 - >40 HP	-1.516	0.00
29 . year of vehicle construction		
YVC1 - 1933-1989 YVC	3.515	134.80
YVC2 - 1990-1991 YVC	-10.222	0.00


10,2 %
13,5 %
76,4 %
80,6 %
9,5 %
9,9 %
0
429
636
1000

Scores (2)

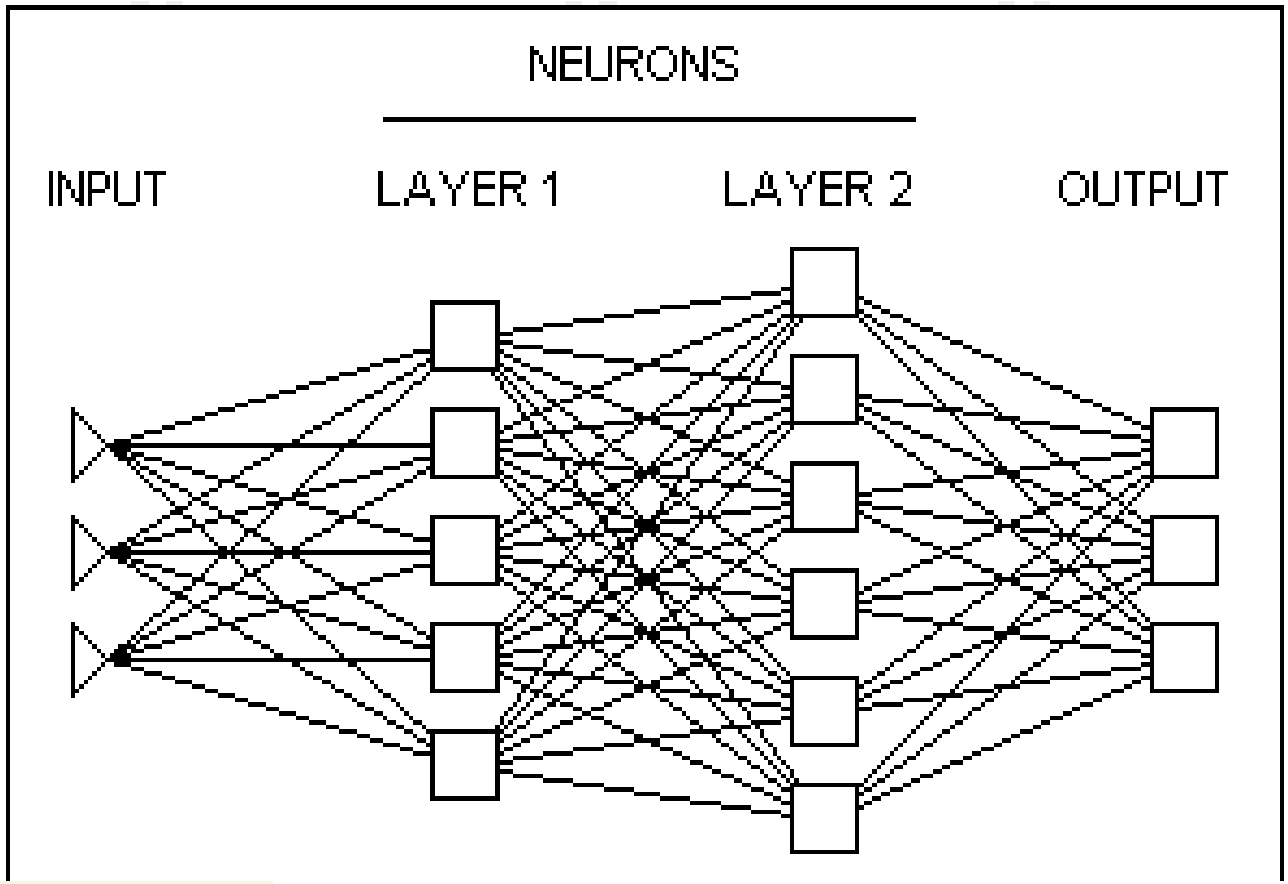
- Scores linéaires: analyse discriminante (Fisher) ou régression logistique
- Mais aussi: notation ou classement direct
- Une probabilité est un score!

2.1.3 Réseau de Neurones

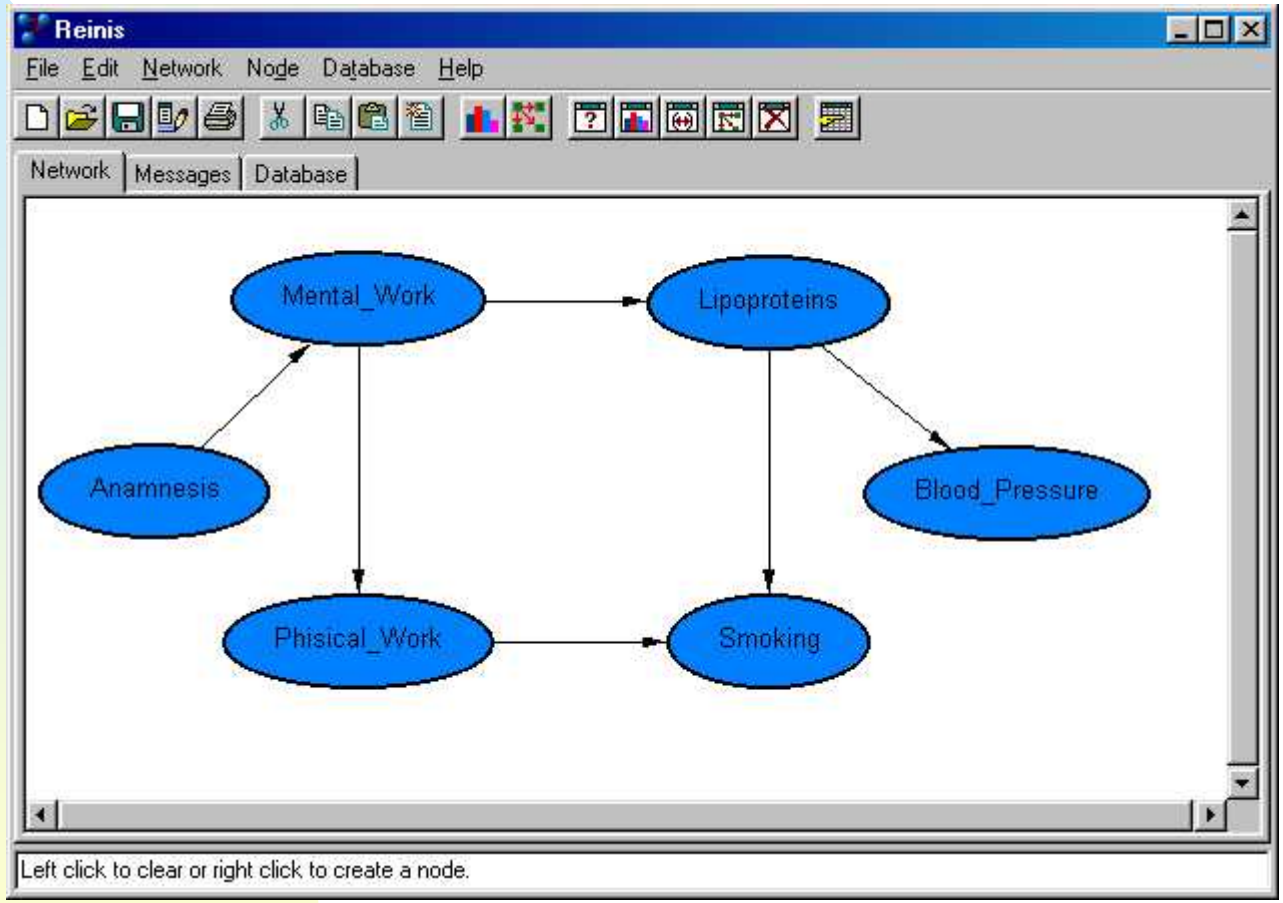


g transfer function)

$$g(v) = \frac{1}{1+e^{-kv}}$$


NEURONS
INPUT
LAYER 1
LAYER 2
OUTPUT

2.1.4 Réseaux bayesiens

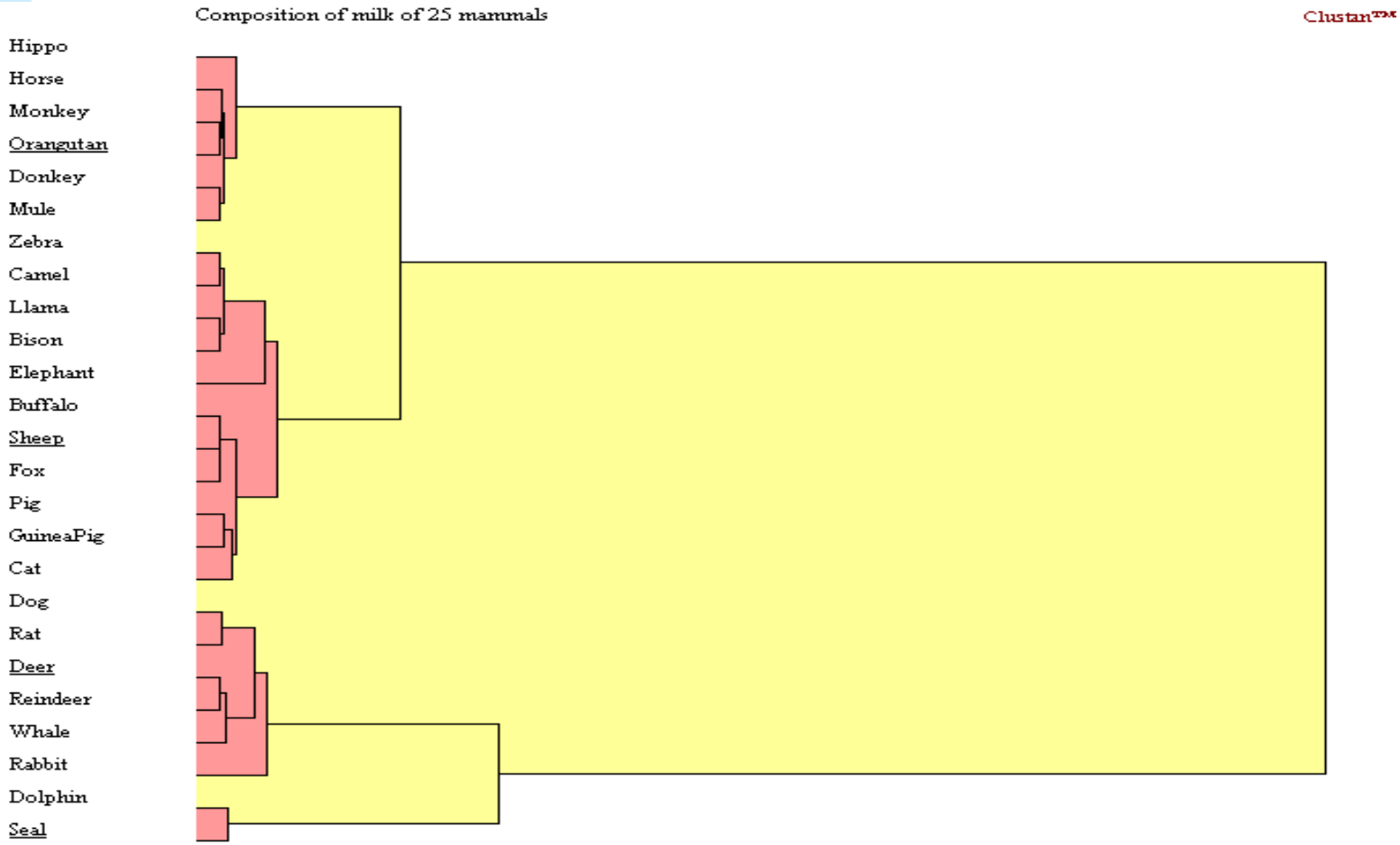


2.2 « Patterns » ou comportements

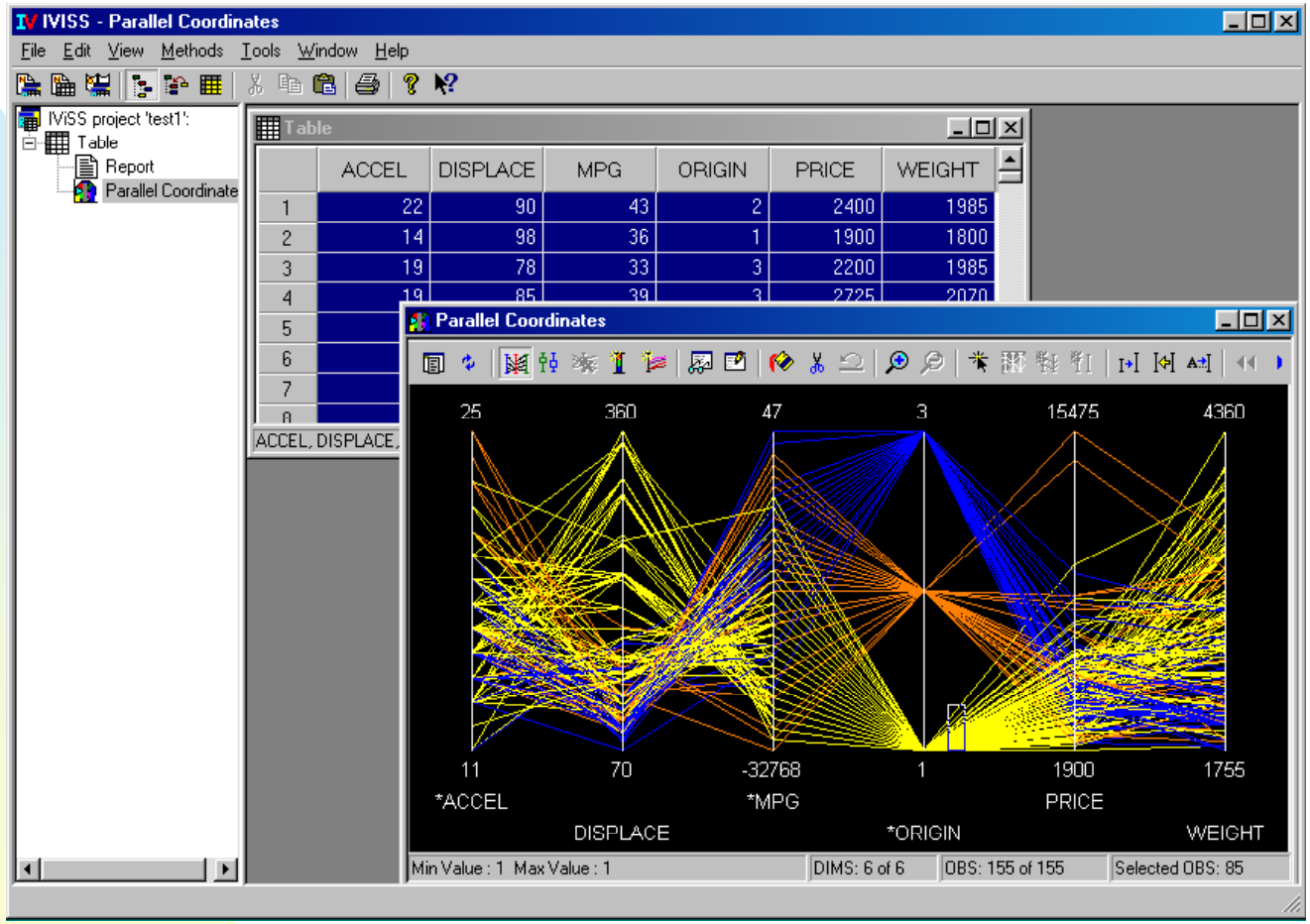
- une structure caractéristique présentée par peu d'observations:
  une « niche » de consommateurs à forte valeur, ou au contraire à haut risque..

- Outils: classification, cartes de Kohonen, visualisation...

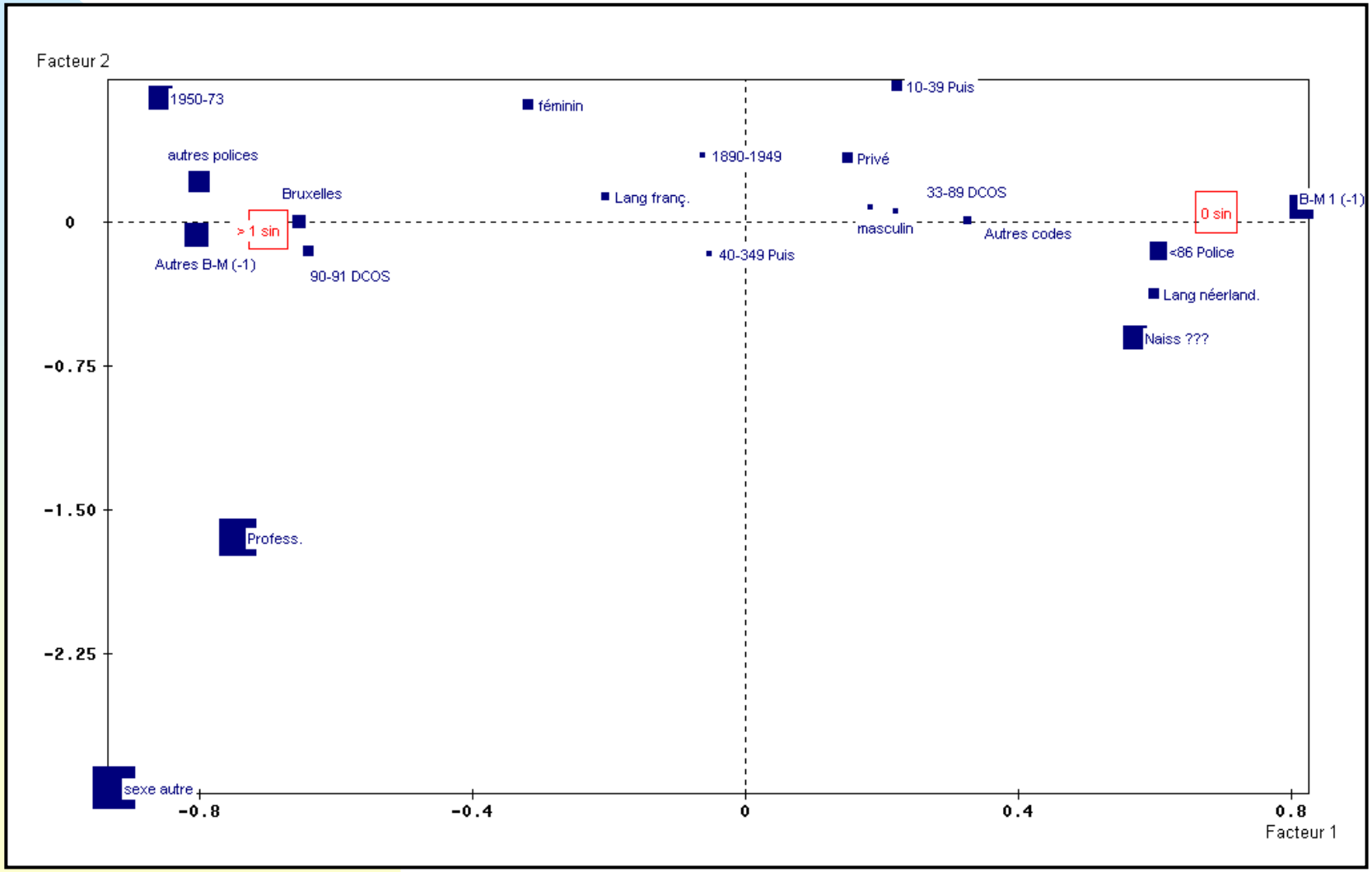
2.2.1Classification hiérarchique



2.2.2 Techniques de Visualisation; eg coordonnées parallèles

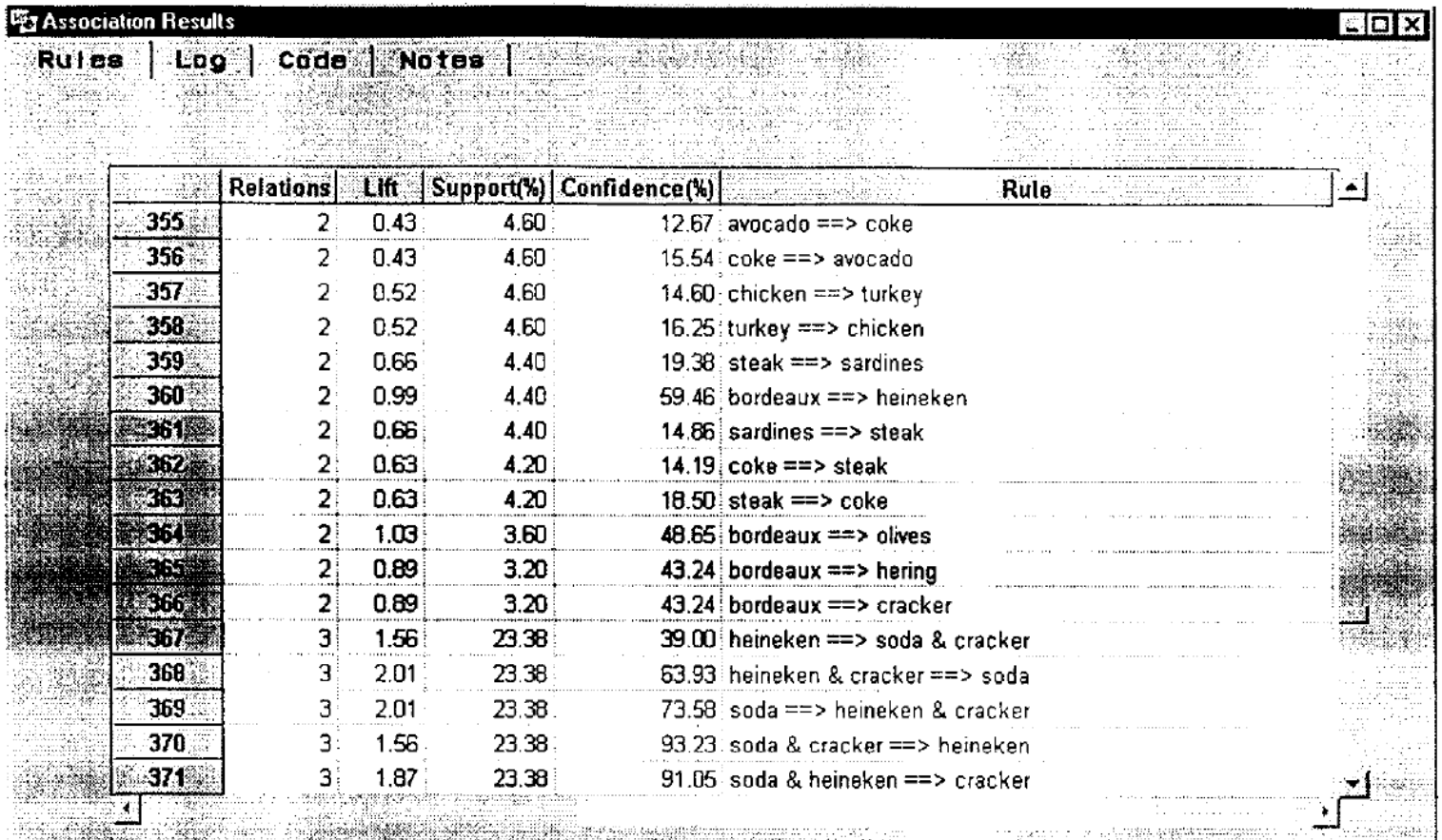


Visualisation suite: analyses factorielles: ACP,AFC,ACM



2.2.3 Découverte de règles d'association, l'analyse des tickets de caisse (*market basket analysis*):

- Ensembles d'articles achetés simultanément: « Couches et biéres! »



Association Results

Rules | Log | Code | Notes

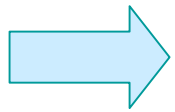
	Relations	Lift	Support(%)	Confidence(%)	Rule
355	2	0.43	4.60	12.67	avocado ==> coke
356	2	0.43	4.60	15.54	coke ==> avocado
357	2	0.52	4.60	14.60	chicken ==> turkey
358	2	0.52	4.60	16.25	turkey ==> chicken
359	2	0.66	4.40	19.38	steak ==> sardines
360	2	0.99	4.40	59.46	bordeaux ==> heineken
361	2	0.66	4.40	14.86	sardines ==> steak
362	2	0.63	4.20	14.19	coke ==> steak
363	2	0.63	4.20	18.50	steak ==> coke
364	2	1.03	3.60	48.65	bordeaux ==> olives
365	2	0.89	3.20	43.24	bordeaux ==> hering
366	2	0.89	3.20	43.24	bordeaux ==> cracker
367	3	1.56	23.38	39.00	heineken ==> soda & cracker
368	3	2.01	23.38	63.93	heineken & cracker ==> soda
369	3	2.01	23.38	73.58	soda ==> heineken & cracker
370	3	1.56	23.38	93.23	soda & cracker ==> heineken
371	3	1.87	23.38	91.05	soda & heineken ==> cracker

règles d'association (2)

- Règle $X \rightarrow Y$
- **Support**: $P(X \cap Y)$
- **Confiance**: $P(Y/X)$
- Logiciels: seuils s0 et c0
- Mais intéressant seulement si $P(Y/X)$ est très supérieur à $P(Y)$.
- **Lift**: $\frac{P(Y/X)}{P(Y)} = \frac{P(Y \cap X)}{P(X)P(Y)}$

2.3.Une Démarche ou des Outils?

- Certains vendeurs insistent sur des algorithmes spécifiques pour se différencier des outils statistiques classiques.
- Tendance actuelle: présenter une collection d'outils dans un logiciel unique et facile d'accès pour comparer plusieurs techniques sur les mêmes données: eg SAS Entreprise Miner.
- Le Data Mining est une démarche :

Données  Information  Connaissance

- 5ème PCRD de l'Union Européenne:
  - 2ème objectif : « développer une société de l'information conviviale »
- 6ème PCRD :
  - Knowledge society

3. De nouveaux gisements: textes, web, données symboliques...

- **3.1Text mining**

  Extraire de l'information de textes.

  Une part croissante de l'information se présente sous forme digitalisée: documents électroniques, nouvelles, brevets,réclamations, e-mails etc.

  Des techniques spéciales de classification supervisée ou non sont développées.

3.2 Webmining

Analyse de la fréquentation de sites web et du comportement des utilisateurs

Applications :

- fidélisation
- mesures d'efficacité de campagnes de promotion
- click analysis : optimisation des sites

3.3 données symboliques

- Sortir du cadre du tableau rectangulaire
- données floues ou intervalles, avec ou sans probabilités.
  - Eurostat a financé le projet SODAS (symbolic data analysis for official statistics), un consortium de 17 équipes.

4.DM et statistique officielle

- Une situation paradoxale?
- Les Instituts Nationaux de Statistique possédent des gisements de données avec les recensements et enquêtes. Cependant peu exploités par les techniques du Data Mining qui reste largement inconnu.

DM et Statistique Officielle : un changement culturel?

- Opposition entre statisticiens officiels « experts » et utilisateurs qui ne connaissent pas leurs données .
- Explorer une base de données sans a priori pour y trouver des structures inconnues n'est pas une idée familière pour les statisticiens publics.
- Cependant Eurostat finance des recherches en DM spécialement conçues pour les INS comme:

  SODAS, KESO (Knowledge Extraction for Statistical Offices),SPIN! (Spatial Mining for Data of Public Interest).

- Quelques domaines potentiels pour le DM :
  - statistique d'entreprise, en particulier innovation, santé financière
  - equipement et épargne des ménages
  - santé publique: détection de facteurs de risque
  - text mining pour les metadonnées
  - amélioration des sites web

- Explorer les bases officielles implique une redéfinition des rôles des INS :
- leur tâche essentielle pour la plupart est la production de données, l'analyse étant faite par d'autres instituts.
- En raison des impératifs de confidentialité, les INS ne peuvent laisser à d'autres certaines explorations de fichiers: nouvelles missions et nouveaux recrutements.
- Renforcement du service auprés des acteurs économiques (citoyens, entreprises).

5. Le Data Mining est-il de la statistique?

- Le DM utilise des techniques statistiques, mais tous ne le savent pas… ou préferent l'ignorer.

  Brad Efron : Statistics has been the most successful information science. Those who ignore Statistics are condemned to reinvent it.
- D'un côté: tentative pour marginaliser la statistique; comme autrefois les systèmes experts, les réseaux de neurones etc.
- De l'autre, nombre de statisticiens et économetres ignorent ou méprisent le phénomène en disant : *"Ce n'est pas de la statistique"*

- Mais qu'est ce que la statistique?

  Jerome Friedman (1997):

  "If being data related is not a sufficient reason for a topic to be considered part of our discipline, then what other qualifications are required? The answer so far seems to be that Statistics is being defined in terms of a set of tools, namely those currently being taught in our graduate programs. A few examples are: Probability theory, Real analysis, Measure theory, Asymptotics, Decision theory, Markov chains, Martingales, Ergodic theory, etc...

  The field of Statistics seems to be defined as the set of problems that can be successfully addressed with these and related tools"

One view recognizes that while the amount of data (and related applications) continue to grow exponentially, the number of statisticians is not growing that fast.Therefore our field should concentrate that small part of information science that we do best, namely probabilistic inference based on mathematics.

This is a highly defensible point of view that may well turn out to be the best strategy for our field. However, if adopted, we should become resigned to the fact that the roll of Statistics as a player in the information revolution will steadily diminish over time. This strategy has the strong advantage that it requires relatively little change to our current practice and academic programs.

- Another point of view, advocated as early as 1962 by John Tukey holds that Statistics ought to be concerned with data analysis. The field should be defined in terms of a set of problems (as are most fields) rather than a set of tools, namely those problems that pertain to data.

Un nouveau paradigme pour l'inférence

- L'inférence classique ne fonctionne plus pour les très grands ensembles de données: toute hypothèse nulle H_0 est rejetée quand n est grand:
  - Une corrélation de 0.002 est significativement différente de zéro avec un million d'individus.
- Il faut remplacer les test de signification par de la validation croisée, du rééchantillonnage…
- On testera si une structure reste valable dans une autre partie des données que celle qui a été explorée.

Dangers et illusions (1/3):

- Les structures trouvées sont-elles valides?
- Il est inévitable de trouver des comportements, en raison d'une recherche combinatoire. Existent-ils vraiment ?
  - "False discovery rate" (Benjamini & Hochberg,1995)
- Le traitement exhaustif n'est sans doute pas la meilleure idée: un bon échantillonnage est souvent plus sur.

Dangers et illusions: (2/3)

- Il faut vérifier l'utilité de ce que l'on « découvre»: corrélation n'est pas causalité et promouvoir B n'entrainera pas forcément des meilleures ventes de A!
- Acceptabilité des méthodes:
  - prédire et comprendre peuvent ne pas aller de pair
  - réticences aux boîtes noires
- Qualité des données, un enjeu majeur.
  - Robustesse aux outliers
  - Données manquantes, fusion de fichiers..

Dangers et illusions (3/3)

- Découvrir des structures « inattendues » est une idée trompeuse: on a d'autant plus de chances de trouver quelque chose d'intéressant que l'on connaît mieux ses données
- Une démarche complètement automatique est aussi une idée fallacieuse. L'expertise et l'intervention du spécialiste sera toujours nécessaire.

L'apport de la théorie de l'apprentissage(1/3)

- V.Vapnik: une nouvelle vision de la modélisation prédictive.

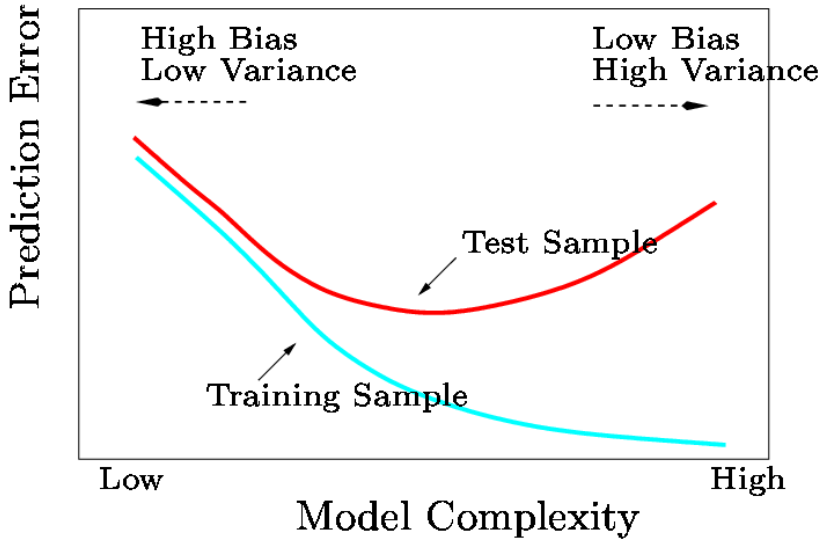


Figure 2.11: *Test and training error as a function of model complexity.*

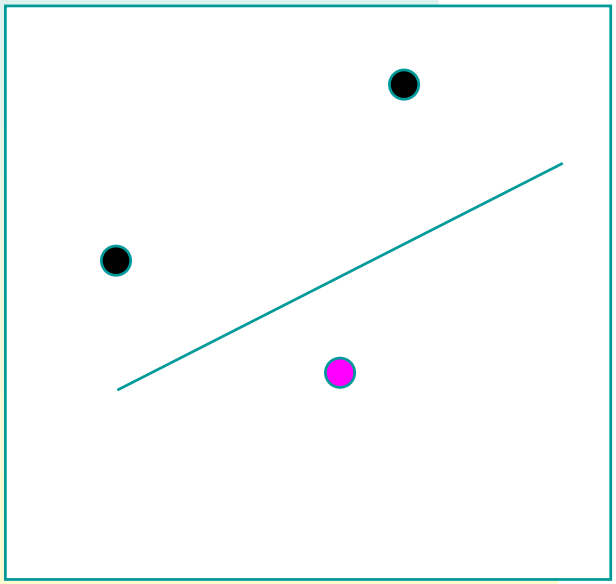
L'apport de la théorie de l'apprentissage (2/3)

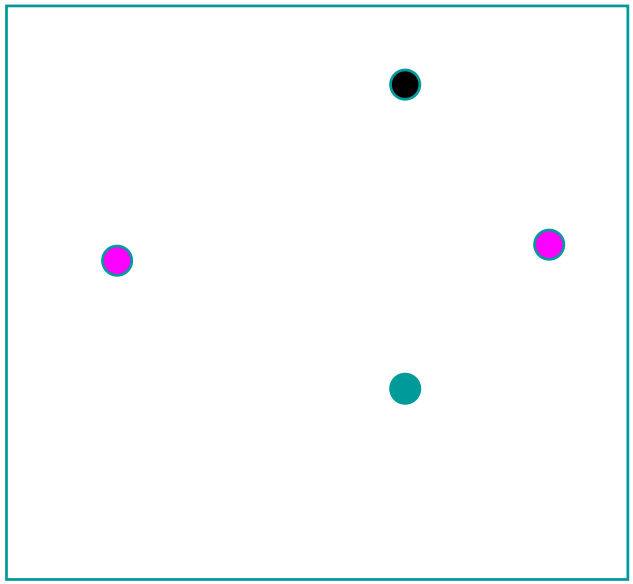
- La complexité d'un modèle n'est pas le nombre de parametres: VC dimension
- Possibilité de travailler dans des espaces de dimension infinie!

$$R < R_{\text{emp}} + \sqrt{\frac{h\left(\ln\left(2n/h\right)+1\right)-\ln q/4}{n}}$$

L’apport de la théorie de l’apprentissage (3/3)

- En 2-D, les fonctions linéaires (droites) peuvent “hacher” 3 points, mais pas 4





Aucune ligne droite ne peut séparer les points noirs des points roses

6.Conclusions et perspectives

- Le Data Mining est une discipline née en dehors de la statistique, dans la communauté des bases de données et de l'IA dans le but de valoriser les bases de données.
- Le Data Mining offre des perspectives nouvelles pour la statistique et répond au défi du traitement des gigabases de données.
- Le Data Mining est la branche de la statistique exploratoire qui cherche à découvrir des structures inconnues et utiles.

- Les statisticiens peuvent y trouver de nouveaux problèmes de recherche et de nouveaux débouchés, et devraient être les mieux placés en tant que spécialistes du risque et de l'incertain.
- S'ils s'en désintéressent, d'autres prendront leur place en réinventant la statistique, qui risque de voir son rôle diminuer .
- Cela implique une redéfinition du rôle de la statistique dans les sciences de l'information, et une réflexion sur les cursus universitaires.

- **Conséquences sur l'enseignement:**
- ne pas se limiter à l'utilisation des packages, mais introduire de l'informatique fondamentale et appliquée: Bases de données, optimisation combinatoire, algorithmique, etc.

- " We may have to moderate our romance with mathematics. Mathematics (like computing) is a tool, a very powerful one to be sure, but not the only one that can be used to validate statistical methodology. Mathematics is not equivalent to theory, nor vice versa. Theories are intended to create understanding and mathematics, is not the only way to do this" J.Friedman.

Le point de vue de l'Académie des Sciences (juillet 2000)

- Constat 6: La recherche dans le domaine de la "gestion des connaissances d'affaires" (business intelligence) représente une enjeu majeur, tant national qu'international, ayant mobilisé des investissements mondiaux dépassant 30 milliards $ en 1999

- Recommandation 6: une mise à niveau des laboratoires des universités et des centres de recherche en statistique informatique. En particulier les laboratoires de statistique doivent pouvoir recruter des enseignants et chercheurs en informatique, et être dotés des moyens de calcul adéquats, pour mener des recherches coordonnées en statistique et informatique, orientées par exemple vers le "Data Mining" et les domaines connexes

Références

- Académie des Sciences, Rapport sur la science et la technologie n°8, *La statistique*, 2000, Paris
- Fayyad U.M., Piatetsky-Shapiro G., Smyth P., and Uthurusamy R. (eds.) *Advances in Knowledge Discovery and Data Mining*. Menlo Park, California: AAAI Press ,1996.
- J.Friedman : *Data Mining and statistics, what's the connection?* 1997 *http://www-stat.stanford.edu/~jhf/ftp/dm-stat.ps*
- J.Friedman: *The role of Statistics in Data Revolution*, ISI,Helsinki, 1999, *http://www.stat.fi/isi99/index.html*
- D.Hand: *Why data mining is more than statistics write large*, ISI,Helsinki, 1999, *http://www.stat.fi/isi99/index.html*
- D.Hand: *Methodological issues in data mining*, in Compstat 2000, Physica-Verlag, 77-85, 2000
- *n°spécial du Journal de la SFdS (2001)*