De l'analyse exploratoire à la modélisation prédictive:

le chemin de la science des données

Gilbert Saporta
CEDRIC- CNAM

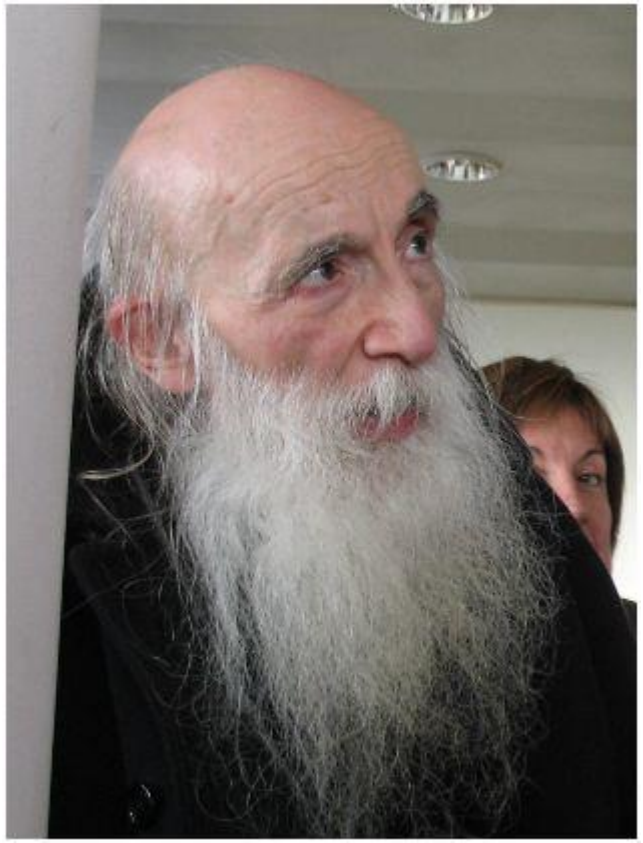
1. Analyse des données *vs* statistique mathématique

- Années 60: un mouvement international en réaction aux abus de la formalisation
- ***Laisser parler les données***
- Statistique et ordinateurs

John Wilder Tukey (1915-2000)

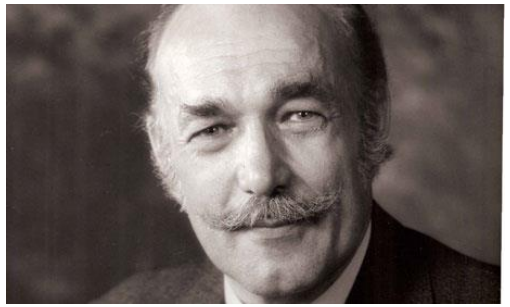
- The Future of Data Analysis (1962)
- « *He (Tukey) seems to identify statistics with the grotesque phenomenon generally known as mathematical statistics and find it necessary to replace statistics by data analysis* » (Anscombe, 1967).

1. Introduction. For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. And when I have pondered about why such techniques as the spectrum analysis of time series have proved so useful, it has become clear that their "dealing with fluctuations" aspects are, in many circumstances, of lesser importance than the aspects that would already have been required to deal effectively with the simpler case of very extensive data, where fluctuations would no longer be a problem. All in all, I have come to feel that my central interest is in *data analysis*, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.



Jean-Paul Benzécri (1932-)

- « *Statistique n'est pas probabilité. Sous le nom de statistique mathématique, des auteurs (qui, je vous le dis en français, n'écrivent guère dans notre langue...) ont édifié une pompeuse discipline, riche en hypothèses qui ne sont jamais satisfaites dans la pratique* » (1972)
- « *L'analyse des données est un outil pour dégager de la gangue des données le pur diamant de la véridique nature* » (1973)

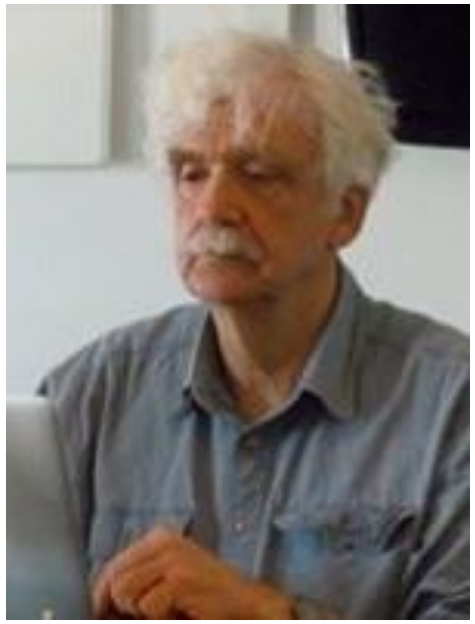


John Nelder (1924-2010)

- "Statistics is intimately connected with science and technology, and few mathematicians have experience or understand of methods of either. This I believe is what lies behind the grotesque emphasis on significance tests in statistics courses of all kinds; a mathematical apparatus has been erected with the notions of power, uniformly most powerful tests, uniformly most powerful unbiased tests, etc. etc., and this is taught to people, who, if they come away with no other notion, will remember that statistics is about significant differences (…). The apparatus on which their statistics course has been constructed is often worse than irrelevant---it is misleading about what is important in examining data and making inferences." (J.A. Nelder, in discussion of Chatfield 1985)

Japon, Pays-Bas, Canada, Italie…

Chikio Hayashi (1918-2002)



Jan de Leeuw (1945-)

Shizuiko Nishisato (1935-)

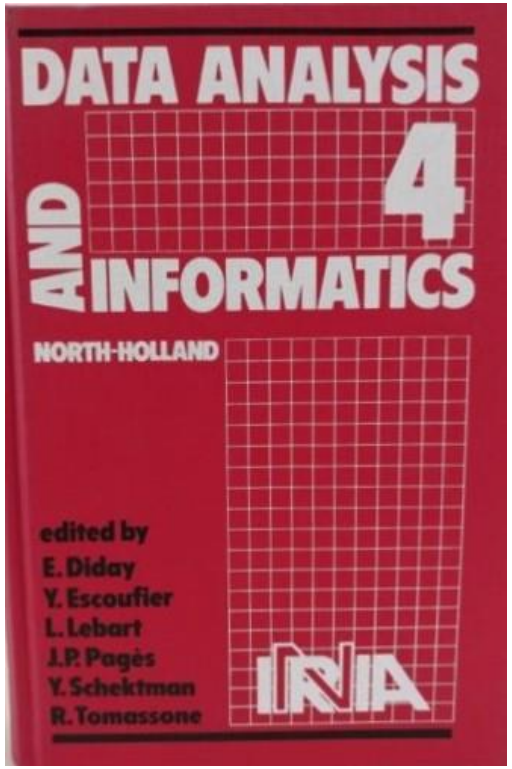
Carlo Lauro (1943-)

Conférences

- Data Analysis and Informatics de 1977 à 1991

Edwin Diday

Ludovic Lebart



ANALISI DEI DATI E INFORMATICA
Incontri con la Scuola Francese

DATA ANALYSIS AND INFORMATICS
Meeting the French School

ANALYSE DE DONNÉES ET INFORMATIQU
Rencontre avec l'École Française

NAPOLI (Italia) - 30 giugno ... 5 luglio 1980
NAPLES (Italy) - June 30th ... July 5th, 1980
NAPLES (Italie) - 30 juin ... 5 juillet 1980

Organizzati dall' / *Organized by* / Organisé par

I. S. D. U. N. ISTITUTO DI STATISTICA E DEMOGRAFIA DELL' UNIVERSITÀ DI NAPOLI

I. N. R. I. A. INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Direttori / *Course Directors* / Directeurs

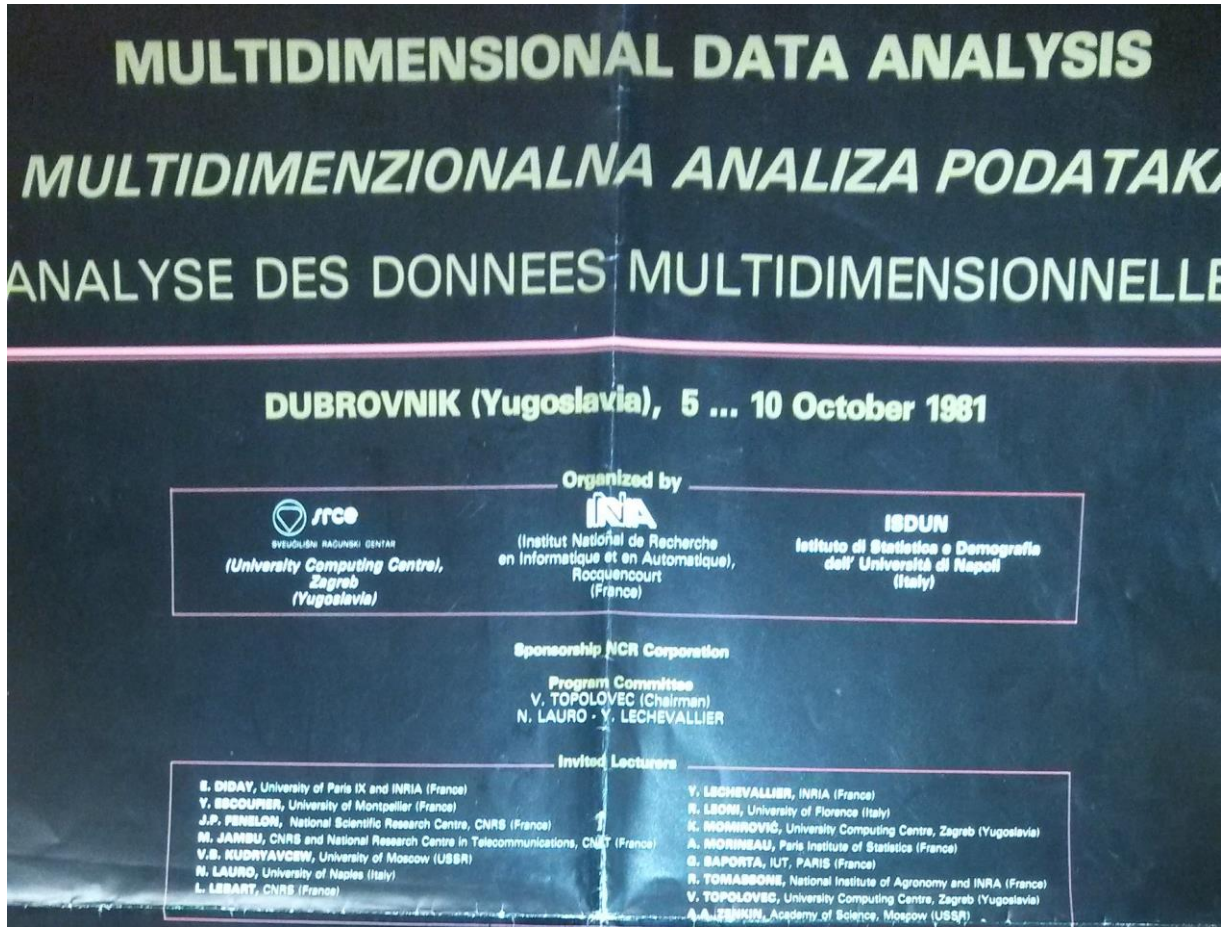
N. LAURO & E. DIDAY

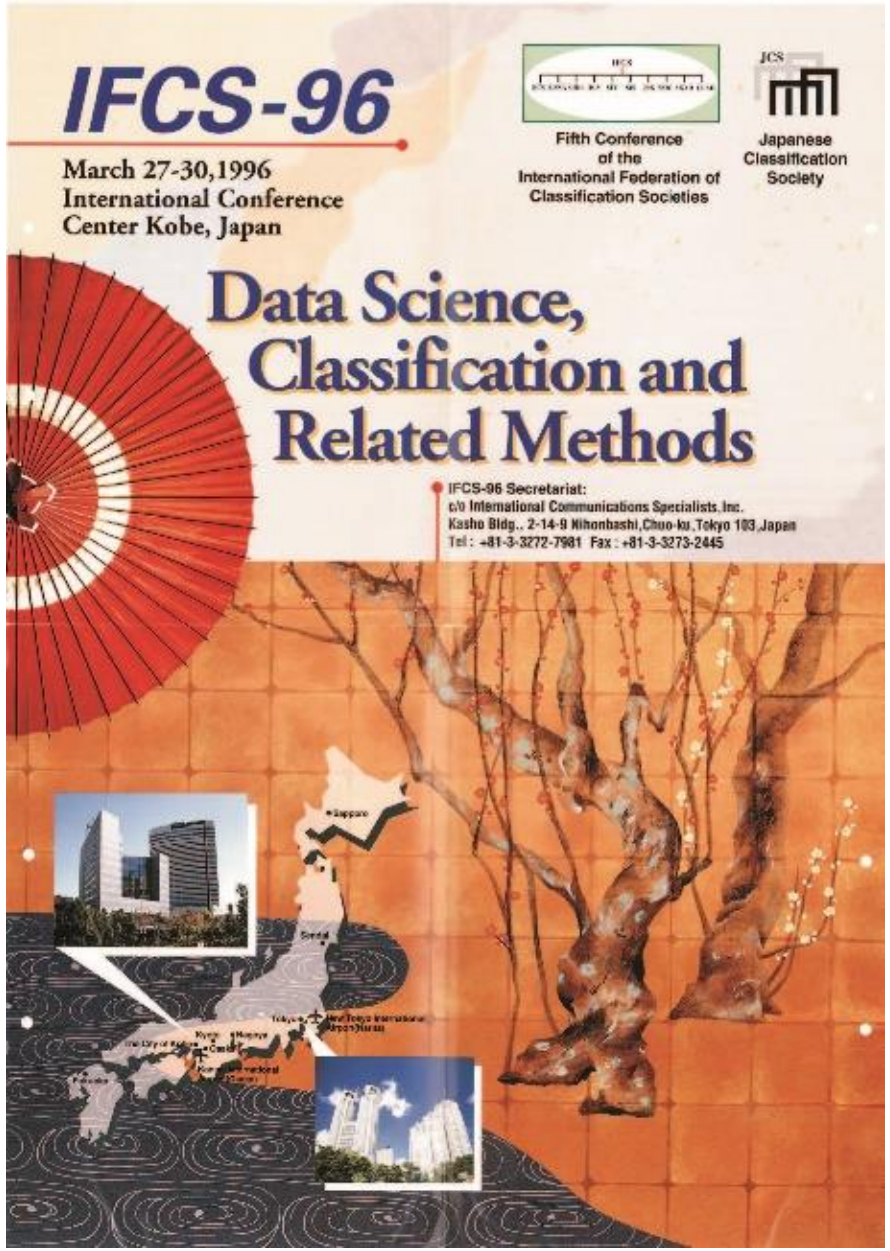
Conferenzieri / *Lecturers* / Conférenciers

E. DIDAY	Université de Paris IX et INRIA
Y. ESCOUFIER	Université de Montpellier
J. M. FENELON	Centre National de la Recherche Scientifique
M. JAMBU	Centre National de la Recherche Scientifique
N. LAURO	Università di Napoli
L. LEBART	Centre National de la Recherche Scientifique
Y. LECHEVALLIER	Institut National de Recherche en Informatique et en Automatique
A. MORINEAU	Institut de Statistiques de Paris
G. SAPORTA	Institut Universitaire de Technologie de Paris
R. TOMASSONE	Institut National de Recherche Agronomique

TEMI / *TOPICS* / THÈMES

Recenti sviluppi sull'analisi fattoriale e l'analisi delle corrispondenze
Recent developments in factor analysis and correspondence analysis
Récents développements en analyse factorielle et analyse des correspond




IFCS-96
March 27-30,1996
International Conference
Center Kobe, Japan
Fifth Conference
of the
International Federation of
Classification Societies
Japanese
Classification
Society
Data Science,
Classification and
Related Methods
IFCS-96 Secretariat:
c/o International Communications Specialists, Inc.
Kasho Bldg., 2-14-9 Nihonbashi, Chuo-ku, Tokyo 103, Japan
Tel : +81-3-3272-7981 Fax : +81-3-3273-2445

2. L’analyse exploratoire
(ou non supervisée)

Une collection de méthodes de réduction de dimension

- Analyses « factorielles »
  - ACP
  - AFC, ACM
- Classification
  - Partitionnement: k-means, nuées dynamiques
  - Hiérarchies

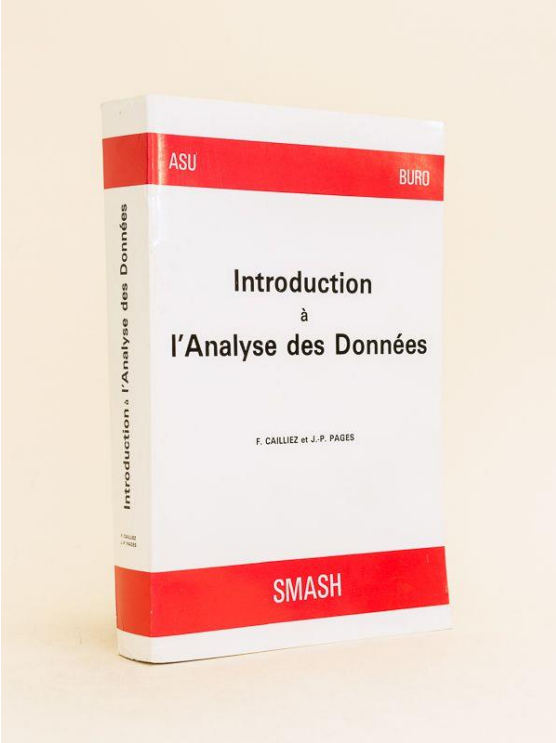
Puis vint le temps des synthéses:

- Toutes les méthodes (factorielles) sont des cas particuliers de:

l'ACP

l'analyse canonique

1976



J.Douglas Carroll (1939 -2011)

Encore plus de méthodes sont des cas particuliers du principe d'association maximale :

$$Max_Y \sum_{j=1}^{p} \Phi(Y, X_j)$$

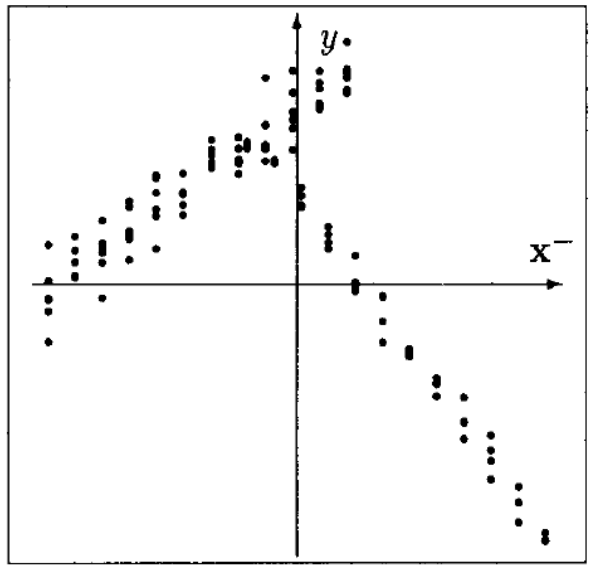
M.Tenenhaus (1977), J.F. Marcotorchino (1986), G.S. (1988)

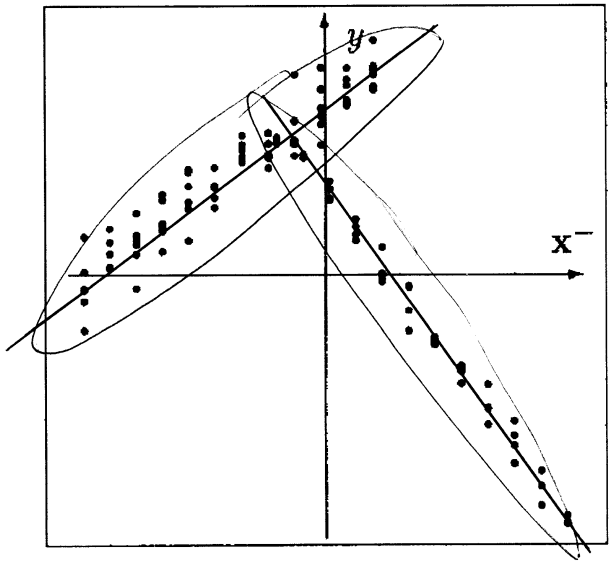
- Quelques cas

critère	analyse
$\max \sum_{j=1}^{p} r^2(c, x_j)$ avec x_j numériques	ACP
$\max \sum_{j=1}^{p} \eta^2(c, x_j)$ avec x_j catégorielles	ACM
$\max \sum_{j=1}^{p} R^2(c, \mathbf{X}_j)$ avec $\mathbf{X}_j$ bloc	ACG (Carroll)
$\max \sum_{j=1}^{p} Rand(Y, x_j)$ avec Y et x_j catégorielles	Partition centrale
$\max \sum_{j=1}^{p} \tau(y, x_j)$ avec y et x classements	Règle de Condorcet

…et le temps des méthodes typologiques ou clusterwise :

- Recherche simultanée d'une partition et de k modèles locaux





Diday, 1974
Charles, 1977
Späth, 1979
DeSarbo & Cron, 1988

Preda & S., 2005
Bougeard, Niang, & S. 2017

Adapté de Hennig, 2000

Et pendant ce temps, à Montpellier…

- 1970: Opérateur d'Escoufier **W**
  - Processus

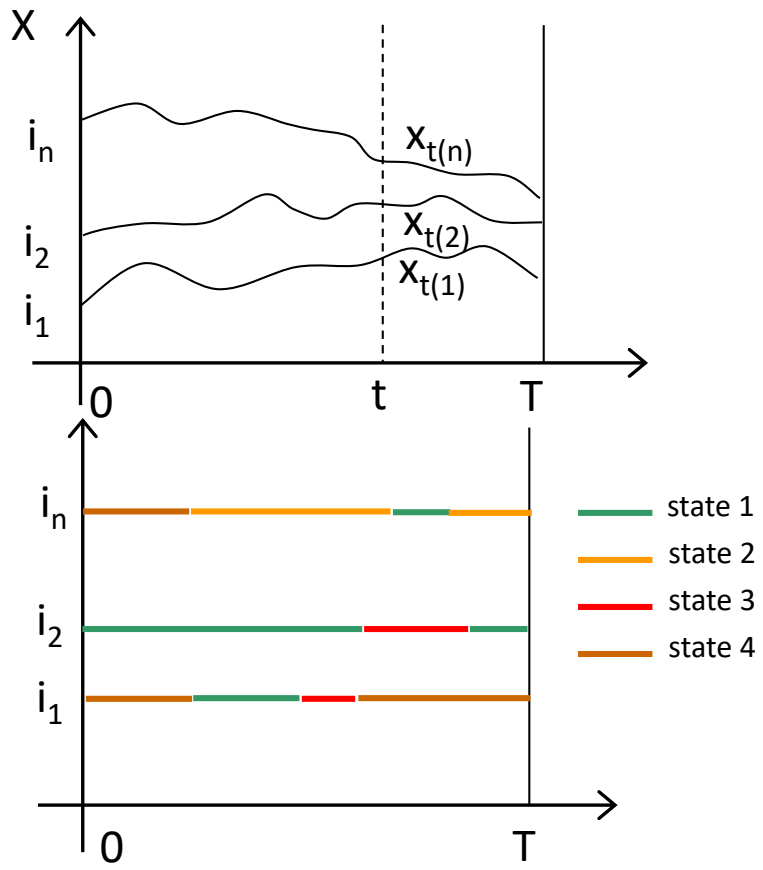
$$\mathbf{W}(Y) = \int_0^T X_t E(X_t Y) dt$$

  - Tableau **W=XX'**
- 1973: coefficient RV
  - Méthodes multiblocs, approche STATIS.

H. L'Hermier des Plantes, P.Robert, C.Lavit, F.Glaçon, R.Sabatier, M.Vivien, X.Bry…

Puis les extensions à de nouveaux types de données

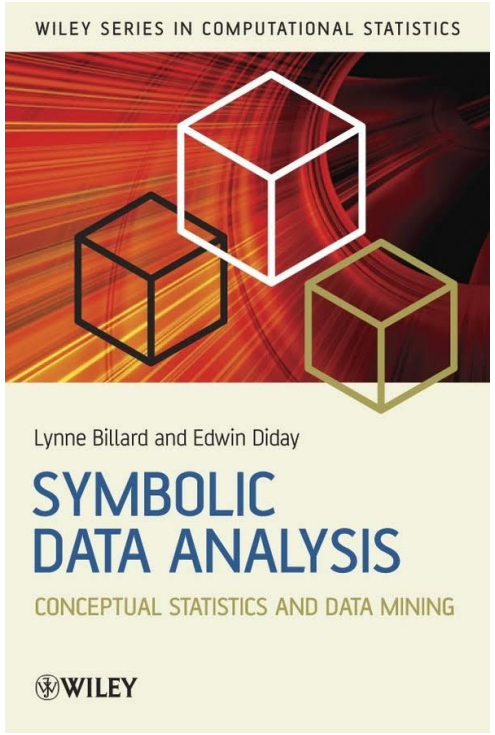
- Données fonctionnelles



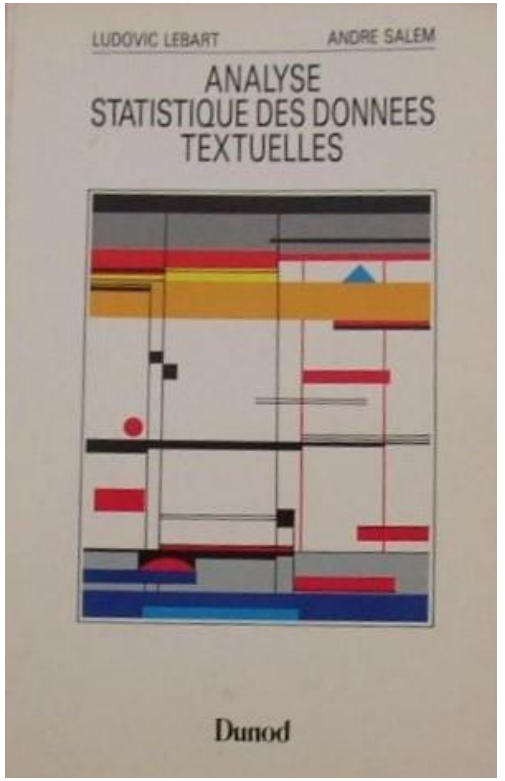
Jean-Claude Deville
1974

Jim Ramsay
1982

- Données symboliques



- Données textuelles



Vers l'analyse non-linéaire

- ACP semi-linéaire (Dauxois & Pousse 1976, Gifi 1990)

$$\arg\max V\left(\sum_{j=1}^{p}\Phi_j\left(x^j\right)\right) \quad \text{au lieu de} \quad \arg\max V\left(\sum_{j=1}^{p}a_j x^j\right)$$

- Kernel PCA (Schölkopf, B., Smola, A., Müller, K.L., 1998)
  - MDS dans l'espace étendu RKHS

$k(\mathbf{x},\mathbf{y})$ fonction simple de $<\mathbf{x},\mathbf{y}>$

Le temps des méthodes sparse

- Inspirées du Lasso.
- Utile en grande dimension p>>n:
  - Problèmes d'interprétation
  - Résultats instables
- ACP sparse

SCoTLASS Simplified Component Technique–Lasso, Jolliffe & al. (2003) contraintes L_1 pour obtenir des coefficients nuls

$$\max \mathbf{u'Vu} \quad \text{sous} \quad \|\mathbf{u}\|^2 = 1 \text{ et } \sum_{j=1}^{p} |u_j| \leq t$$

3. Modélisation prédictive
(ou supervisée)

Les deux cultures

1928-2005

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

- The generative modelling culture
  - seeks to develop stochastic models which fits the data, and then make inferences about the data-generating mechanism based on the structure of those models. Implicit (…) is the notion that there is a true model generating the data, and often a truly “best” way to analyze the data.
- The predictive modelling culture
  - is silent about the underlying mechanism generating the data, and allows for many different predictive algorithms, preferring to discuss only accuracy of prediction made by different algorithm on various datasets. **Machine Learning** is identified by Breiman as the epicenter of the Predictive Modeling culture.

From Donoho, 2015

Machine Learning

- Esprit similaire à celui de l'analyse des données
  - « Les modèles doivent suivre les données, pas l'inverse » *JPB principe n°2*
  - George Box (1919-2013): « All models are wrong, some are useful », 1987
- Pas ou peu d'hypothèses de distribution
- Privilégie l'efficacité sur l'interprétabilité
  - SVM, réseaux de neurones et deep learning, forêts aléatoires, boosting et méthodes d'ensemble, etc: boîtes noires

Validation empirique

- Le Machine Learning insiste sur :
  - Différence entre ajustement et prévision
  - Usage systématique d'ensembles d'apprentissage et de validation

Une démarche avec 3 échantillons pour choisir entre plusieurs modèles:

- Apprentissage: pour estimer les paramètres des modèles
- Test : pour choisir le meilleur modèle
- Validation : pour estimer la performance sur des données futures
  - **Estimer les paramètres ≠ estimer la performance**

- Précurseurs:
  - Paul Horst (1903-1999)

    « *the usefulness of a prediction procedure is not established when it is found to predict adequately on the original sample; the necessary next step must be its application to at least a second group. Only if it predicts adequately on subsequent samples can the value of the procedure be regarded as established*» 1941
  - Leave one out : Lachenbruch et Mickey, 1968
  - Validation croisée: Stone, 1974

- **Elémentaire?**
  - Pas si sur…
  - Voir publications en économétrie, épidémiologie, .. prédictions rarement validées sur des données « hold-out » (sauf en prévision de séries temporelles)

Paradoxes

- Comprendre sans prédire
  - Un modèle qui s'ajuste bien peut fournir des prévisions médiocres au niveau individuel (eg épidémiologie)
- Prédire sans comprendre
  - Des modèles ininterprétables (eg *deep learning*) peuvent donner de bonnes prévisions
  - *Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms* (Vapnik, 2006).

Vladimir Vapnik (1936-)

4. Et maintenant:
Big Data, Data Science etc.

What steam was to the 19th century, and oil has been to the 20th, data is to the 21st. It's the driver of prosperity, the revolutionary resource that is transforming the nature of economic activity, the capability that differentiates successful from unsuccessful societies.

The Data Manifesto, Royal Statistical Society, 2014

Big Data: faux espoirs et défis

Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

Comprendre pour mieux prédire

- Ouvrir les boîtes noires
  - « Model agnostic methods » pour mesurer l'importance des variables
    - une variable est importante si sa suppression (ou une permutation aléatoire) affecte sérieusement la qualité des prévisions
- Corrélation *versus* causalité et le retour de l'expérimentation
  - As Box et al. put it, "To find out what happens when you change something, it is necessary to change it." … the best way to answer causal questions is usually to run an experiment. (Varian, 2016)
  - Big Data et inférence causale: A/B testing, publicité sur le web (Bottou,2013)

Arthur M. Sackler COLLOQUIA

Drawing Causal Inference from Big Data

This meeting was held March 26-27, 2015 at the National Academy of Sciences 2101Constitution Ave. NW in Washington, D.C.
Organized by Richard M. Shiffrin (Indiana University), Susan Dumais (Microsoft Corporation), Mike Hawrylycz (Allen Institute), Jennifer Hill (New York University), Michael Jordan (University of California, Berkeley), Bernhard Schölkopf (Max Planck Institute) and Jasjeet Sekhon (University of California, Berkeley)
Graduate Student / Postdoctoral Researcher travel awards sponsored by the National Science Foundation and the Ford Foundation.

July 5, 2016
vol. 113, n° 27

Conclusions (?)

- Esprit de l’analyse des données toujours actuel
- Vers plus d’interprétabilité
- Science et conscience

Quelques souvenirs

Dubrovnik 1981

Japon 1987

ICOTS IV, Marrakech, 1994

Merci pour votre attention