25 ans de statistique décisionnelle

Gilbert Saporta
Conservatoire National des Arts et Métiers

gilbert.saporta@cnam.fr
http://cedric.cnam.fr/~saporta

Bienvenue au CNAM

Créé en 1794 par l'abbé Grégoire pour perfectionner l'industrie nationale

Formation tout au long de la vie
pour près de 90 000 auditeurs par an.
Recherche et diffusion de la culture scientifique et technique

Grégoire

Licence de maths.applis option **statistique**
Master de statistique
17 cours à la carte
Master d'**actuariat**

Souvenirs de Vannes

- Deux congrès mémorables
  - 1977
  - 1993
- Deux anniversaires:
  - 20 ans en mars 1991
  - 30 ans en octobre 2001
- La CPN (1984-2005)
  - De SEETQG à STID
- Une figure tutélaire

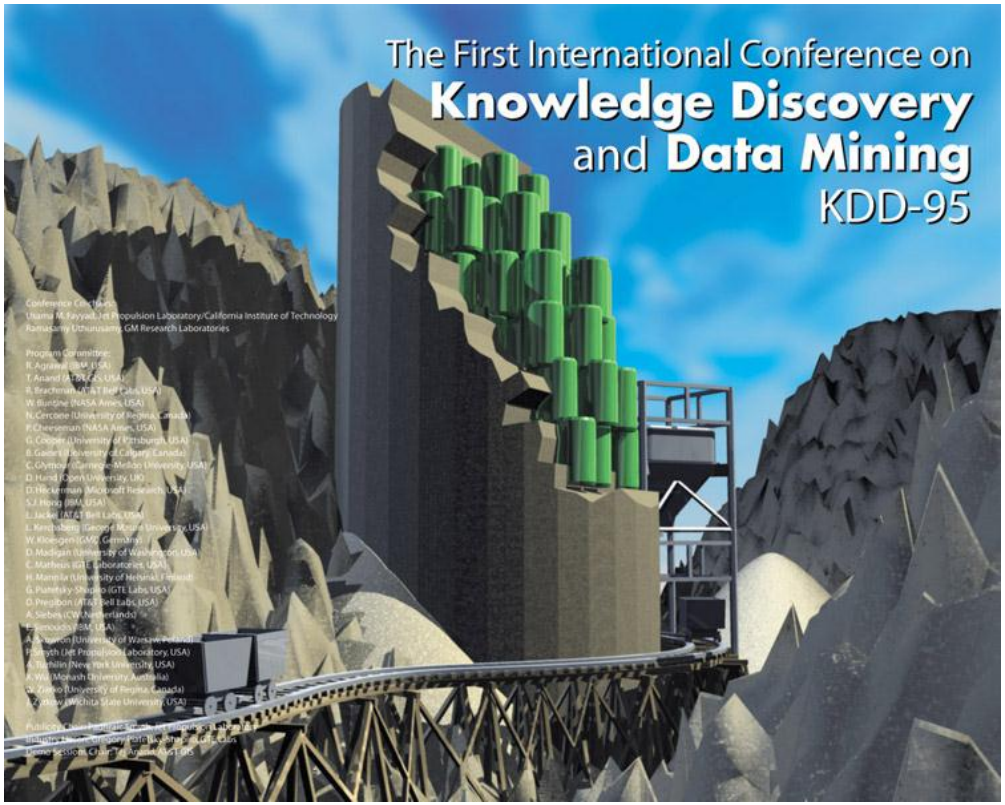
et quelques personnalités

1996

- Statistique décisionnelle
- Data Mining (fouille de données)
- Modélisation prédictive
- KDD (extraction de connaissances)

Termes inconnus en 1984!

- Decision support systems (informatique décisionnelle) dès les années 70 mais sans statistique. Business intelligence après 90


The First International Conference on
Knowledge Discovery
and Data Mining
KDD-95

Data mining et KDD

- « Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms »

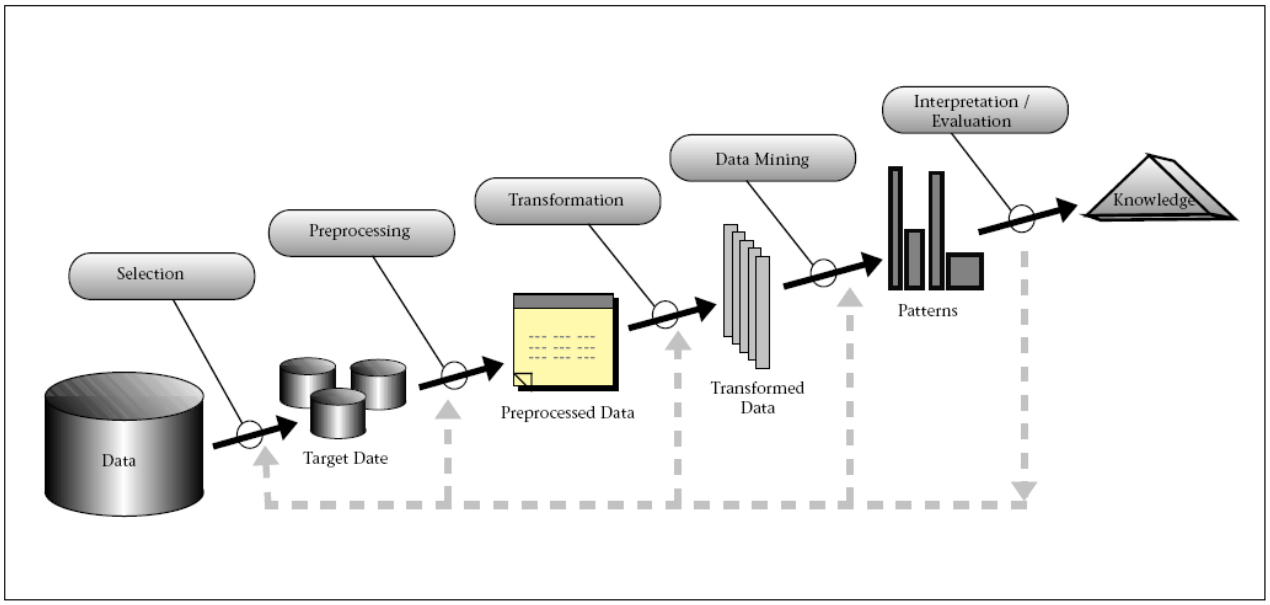


Figure 1. An Overview of the Steps That Compose the KDD Process.

Data Mining and Knowledge Discovery in Databases: An overview, Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, Communications of ACM, 39:11, November 1996

4 générations

- 1ère génération: statistique classique (tests)
- 2ème génération (70): statistique multidimensionnelle (discriminante, logistique)
- **3ème génération (80) machine learning (arbres de décisions, réseaux de neurones)**
- **4ème génération (90) théorie de l'apprentissage (SVM)**

Une nouvelle manière de concevoir des modèles

- Modèles pour comprendre ou modèles pour prévoir?
  - Compréhension des données et de leur mécanisme générateur à travers une représentation simple (parcimonieuse)
  - Prédire de nouvelles observations avec une bonne précision

- Paradoxe n° 1
  - Un « bon » modèle statistique ne donne pas nécessairement des prédictions précises au niveau individuel. Exemple facteurs de risque en épidémiologie
- Paradoxe n°2
  - On peut prévoir sans comprendre:
    - pas besoin d'une théorie du consommateur pour faire du ciblage
    - un modèle n'est qu'un algorithme

- En data mining, un bon modèle est celui qui donne de bonnes prévisions
  - capacité prédictive sur de nouvelles observations (« généralisation »)
  - différent de l'ajustement aux données (prédire le passé)
    - Un modèle trop précis sur les données se comporte de manière instable sur de nouvelles données : phénomène de surapprentissage
    - Un modèle trop robuste (rigide) ne donnera pas un bon ajustement sur les données
  - modéles issus des données

Théorie statistique de l'apprentissage

- V.Vapnik a fournir le cadre conceptuel pour obtenir des modèles robustes et précis
- Apports principaux:
  - complexité h $\neq$ nombre de paramètres
  - compromis entre ajustement et généralisation
  - erreur de généralisation fonction de h/n:la complexité peut augmenter avec la taille des données

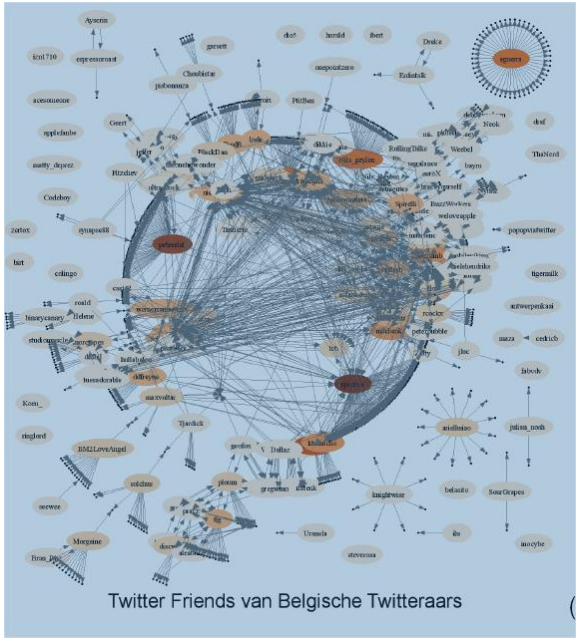
Quelques domaines d'application

- CRM:
  - profilage: associations et recommandations
  - fidélité (churn)
- Détection de fraude
- Notation des risques (scoring)

- Santé, bioinformatique
- Grande distribution

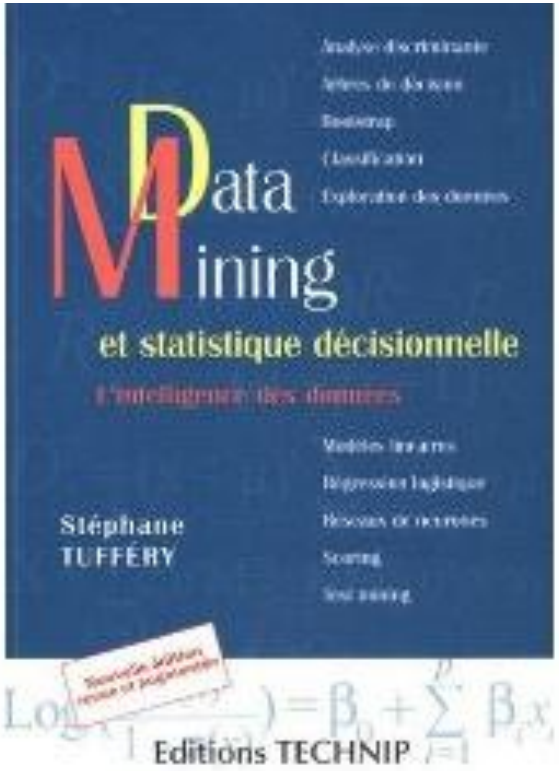
Enjeux de la prochaine décennie

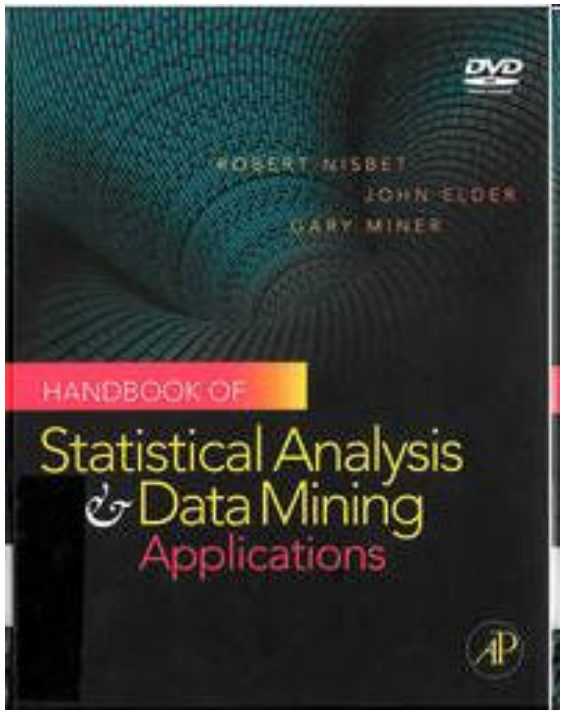
- Des données toujours plus nombreuses:
  - puces RFID
  - Réseaux sociaux
- Flux de données



- Tirer le meilleur des deux approches: data driven et hypothesis driven en les combinant
- Plus que jamais un métier d'avenir pour ceux qui savent combiner les compétences statistiques et informatique
- éthique et traitements de données personnelles

Lectures recommandées:





Merci
pour votre attention

SECOND ANNOUNCEMENT
CALL FOR **PAPERS**

The 19th International Conference on Computational Statistics
Paris - France, **August** 22nd-27th, **2010**

Information : http://www.compstat2010.fr - info@compstat2010.fr

Conservatoire National des Arts et des Métiers
292 rue Saint-Martin - 75003 PARIS - France

Organized by:

Cnam iasc INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE INRIA