

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

- Le Data Mining est un nouveau champ situé au croisement de la statistique et des technologies de l'information (bases de données, intelligence artificielle, apprentissage etc.) dont le but est de découvrir des structures dans de vastes ensembles de données.

- Deux types: modèles et « patterns » (ou comportements)  
(D.Hand)

## 1.1 Définitions:



- 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)



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

Conference Co-Chairs:  
Usama M. Fayyad, Jet Propulsion Laboratory/California Institute of Technology  
Ramakrishnan Uthurusamy, GM Research Laboratories

Program Committee:  
R. Agrawal (IBM, USA)  
T. Anand (AT&T, USA)  
R. Berchman (AT&T Bell Labs, USA)  
W. Buntine (RASA Ames, USA)  
N. Cercone (University of Regina, Canada)  
R. Chensu (NASA Ames, USA)  
G. Cooper (University of Pittsburgh, USA)  
B. Gao (University of Calgary, Canada)  
C. Glymour (Carnegie Mellon University, USA)  
D. Hand (Open University, UK)  
D. Heckerman (Microsoft Research, USA)  
S. J. Hong (IBM, USA)  
L. Jackel (AT&T Bell Labs, USA)  
J. Kerschberg (Georgia Institute of Technology, USA)  
W. Kloeckner (GMD, Germany)  
D. Madigan (University of Washington, USA)  
C. Mathieu (GTE Laboratories, USA)  
H. Mannila (University of Helsinki, Finland)  
G. Piatetsky-Shapiro (GTE Labs, USA)  
D. Pregibon (AT&T Bell Labs, USA)  
A. Siebes (CWI, Netherlands)  
E. Simoudis (IBM, USA)  
A. S. Wilson (University of Warwick, UK)  
P. Wright (Jet Propulsion Laboratory, USA)  
A. Yeh (New York University, USA)  
A. Yeh (Monash University, Australia)  
D. Zelenko (University of Regina, Canada)  
L. Zou (Wichita State University, USA)

Publication: International Journal of Knowledge Discovery and Data Mining  
Volume 1, Number 1, 1996  
Editor: Usama M. Fayyad

# 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, Benzécri) « *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)

## 1.2 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

# Le défi de l'explosion du volume de données (Michel Béra, 2009)

- In the 90s



Large in	
Neural Networks	Statistics
100,000 Weights	50 parameters
50,000 examples	200 cases

- Today

- **Web transactions** At Yahoo ! (Fayyad, KDD 2007)



**± 16 B events - day, 425 M visitors - month, 10 Tb data / day**

- **Radio-frequency identification** (Jiawei, Adma 2006)



A retailer with 3,000 stores, selling 10,000 items a day per store  
**300 million events per day** (after redundancy removal)

- **Social network** (Kleinberg, KDD 2007)



**4.4-million-node network** of declared friendships on blogging community  
**240-million-node network** of all IM communication over one month on  
Microsoft Instant Messenger

- **Cellular networks**



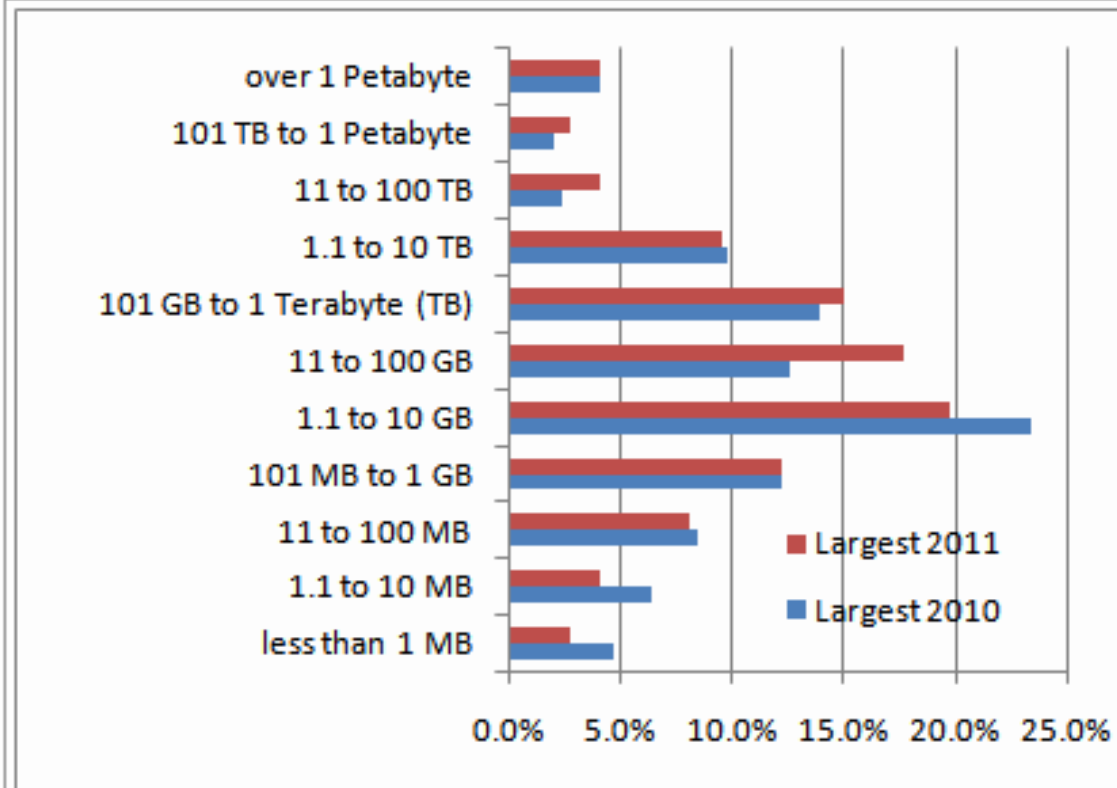
A telecom carrier generates **hundreds of millions of CDRs / day**  
The network generates technical data : **40 M events / day** in a large city



## Largest dataset analyzed / data mined

 Tweet 8

What was the largest database / dataset you analyzed? [148 votes]



<http://www.kdnuggets.com>



## Industries / Fields where you applied Analytics / Data Mining in 2011

### Industries / Fields where you applied Analytics / Data Mining in 2011?

[228 voters] 2011 % of voters 2010 % of voters

CRM/ consumer analytics (57)	25.0%	26.8%
Banking (43)	18.9%	19.2%
Health care/ HR (38)	16.7%	13.1%
Education (37)	16.2%	9.9%
Fraud Detection (32)	14.0%	12.7%
Science (31)	13.6%	10.3%
Social Networks (30)	13.2%	6.6%
Credit Scoring (29)	12.7%	8.0%
Direct Marketing/ Fundraising (28)	12.3%	11.3%
Insurance (28)	12.3%	10.3%
Finance (26)	11.4%	11.3%
Telecom / Cable (25)	11.0%	10.8%
Retail (24)	10.5%	8.0%

<http://www.kdnuggets.com>

## 1.3 Objectifs et outils

- Le Data Mining cherche des structures de deux types : **modèles** et **patterns**
- Patterns
  - une structure caractéristique possédée par un petit nombre d'observations: niche de clients à forte valeur, ou au contraire des clients à haut risque
  - Outils: classification, visualisation par réduction de dimension (ACP, AFC etc.), règles d'association.

# 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) \*

\* Box, G.E.P. and Draper, N.R.: Empirical Model-Building and Response Surfaces, p. 424, Wiley, 1987

# Modèles

- Le DM ne traite pas d'estimation et de tests de modèles préspecifiés, 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.

- 
- Autre distinction: **prédictif** (supervisé) ou **exploratoire** (non supervisé)

# Des outils ou un process?



- Le DM est souvent présenté comme un ensemble intégré d'outils permettant entre autres de comparer plusieurs techniques sur les mêmes données.
- Mais le DM est bien plus qu'une boîte à outils:



# Data mining et KDD

- « Le Data Mining est une étape dans le processus d'extraction des connaissances, qui consiste à appliquer des algorithmes d'analyse des données »

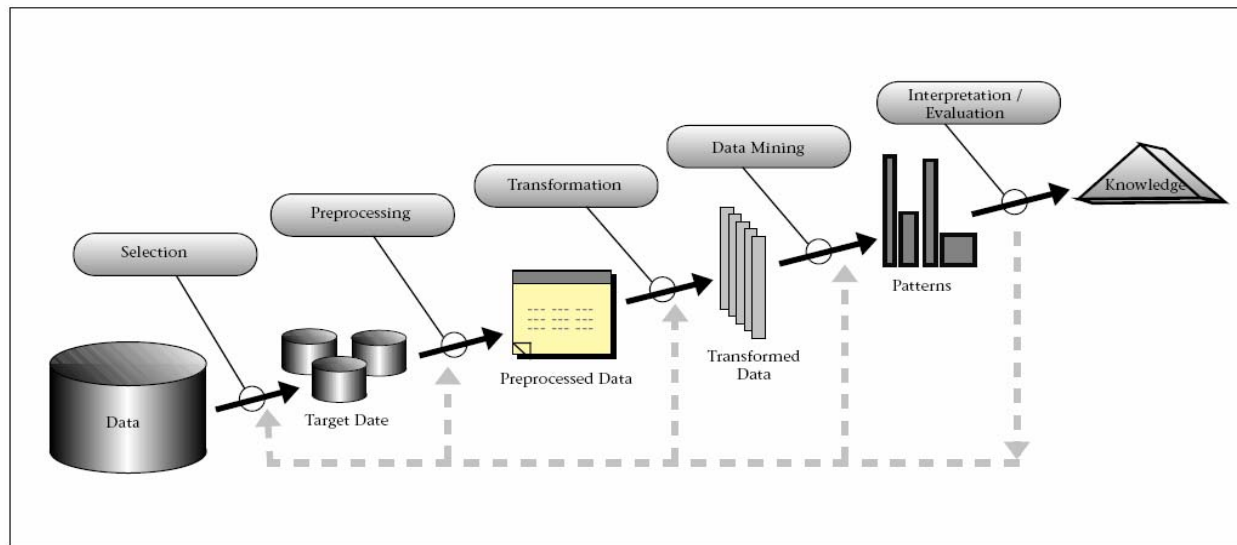


Figure 1. An Overview of the Steps That Compose the KDD Process.



## Algorithms for data analysis / data mining



Tweet




comments

Which methods/algorithms did you use for data analysis in 2011? [311 voters]

Decision Trees/Rules (186)		59.8 %
Regression (180)		57.9 %
Clustering (163)		52.4 %
Statistics (descriptive) (149)		47.9 %
Visualization (119)		38.3 %
Time series/Sequence analysis (92)		29.6 %
Support Vector (SVM) (89)		28.6 %
Association rules (89)		28.6 %
Ensemble methods (88)		28.3 %
Text Mining (86)		27.7 %
Neural Nets (84)		27.0 %
Boosting (73)		23.5 %
Bayesian (68)		21.9 %
Bagging (63)		20.3 %
Factor Analysis (58)		18.7 %
Anomaly/Deviation detection (51)		16.4 %
Social Network Analysis (44)		14.2 %
Survival Analysis (29)		9.32 %
Genetic algorithms (29)		9.32 %
Uplift modeling (15)		4.82 %

## 2. Trois techniques emblématiques du Data Mining



- Une méthode non supervisée:
  - Règles d'association
- Deux méthodes supervisées
  - Arbres de décision
  - Scores








## 2.1 La recherche de règles d'association ou l'analyse du panier de la ménagère

- Illustré avec un exemple industriel provenant de PSA Peugeot-Citroen .
- (Thèse CIFRE de Marie Plasse).

# PROBLEMATIQUE INDUSTRIELLE








## Les données

→ Plus de 80000 véhicules décrits par plus de 3000 attributs binaires

Véhicules	A1	A2	A3	A4	A5	...	Ap
	1	0	0	1	0		0
	0	0	1	1	0		0
	0	1	0	0	1		0
	1	0	0	0	1		0
	0	1	0	0	1		1
	0	1	0	0	1		0
	0	0	1	0	0		0

Matrice de données binaires

=

Véhicules	Attributs présents
	{A1, A4}
	{A3, A4}
	{A2, A5}
	{A1, A5}
	{A2, A5, Ap}
	{A2, A5}
	{A3}

Données de transaction

- Trouver des corrélations entre les attributs...
- ... grâce à la recherche de règles d'association

# LA RECHERCHE DE REGLES D'ASSOCIATION

## Rappel de la méthode

- ⇒ Origine marketing : analyser les ventes des supermarchés  
*"lorsqu'un client achète du pain et du beurre,  
 il achète 9 fois sur 10 du lait en même temps"*

- ⇒ Formalisation :  $A \rightarrow C$  où  $A \cap C = \emptyset$

- ⇒ Fiabilité : **Support** : % de transactions contenant A et C

$$\text{sup}(A \rightarrow C) = P(A \cap C) = P(C / A) \cdot P(A)$$

- ⇒ Précision : **Confiance** : % de transactions contenant C sachant qu'elles ont A

$$\text{conf}(A \rightarrow C) = P(C / A) = \frac{P(A \cap C)}{P(A)} = \frac{\text{sup}(A \rightarrow C)}{\text{sup}(A)}$$

- ⇒ Algorithmes :

- ⇒ Recherche des **sous-ensembles fréquents** (avec minsup)
- ⇒ Extraction des **règles d'association** (avec minconf)



$$s(A \rightarrow C) = 30 \%$$

⇒ 30% des transactions contiennent à la fois



$$c(A \rightarrow C) = 90 \%$$

⇒ 90% des transactions qui contiennent  
 contiennent aussi



- Apriori (Agrawal & Srikant, 1994)
- Partition (Saverese et al., 1995)
- Sampling (Brin & Motwani, 1997)
- Eclat (Zaki, 2000)
- FP-Growth (Han & Pei, 2003)