# Data Mining

## Spécificités, Applications et Outils

Ricco Rakotomalala
Université Lumière Lyon 2
Laboratoire ERIC

Ricco Rakotomalala

- ricco.rakotomalala@univ-lyon2.fr

- http://chirouble.univ-lyon2.fr/~ricco/data-mining
Ressources, documentation, logiciels, données…

# Plan

1. Un exemple introductif
2. Spécificités du Data Mining
3. Le « scoring » marketing
4. Les logiciels libres (gratuits) de Data Mining

# Data Mining ?

Une démarche plus qu'une théorie !

# Exemple introductif : demande de crédit bancaire

L'expert se fonde sur son « expérience » pour prendre la bonne décision

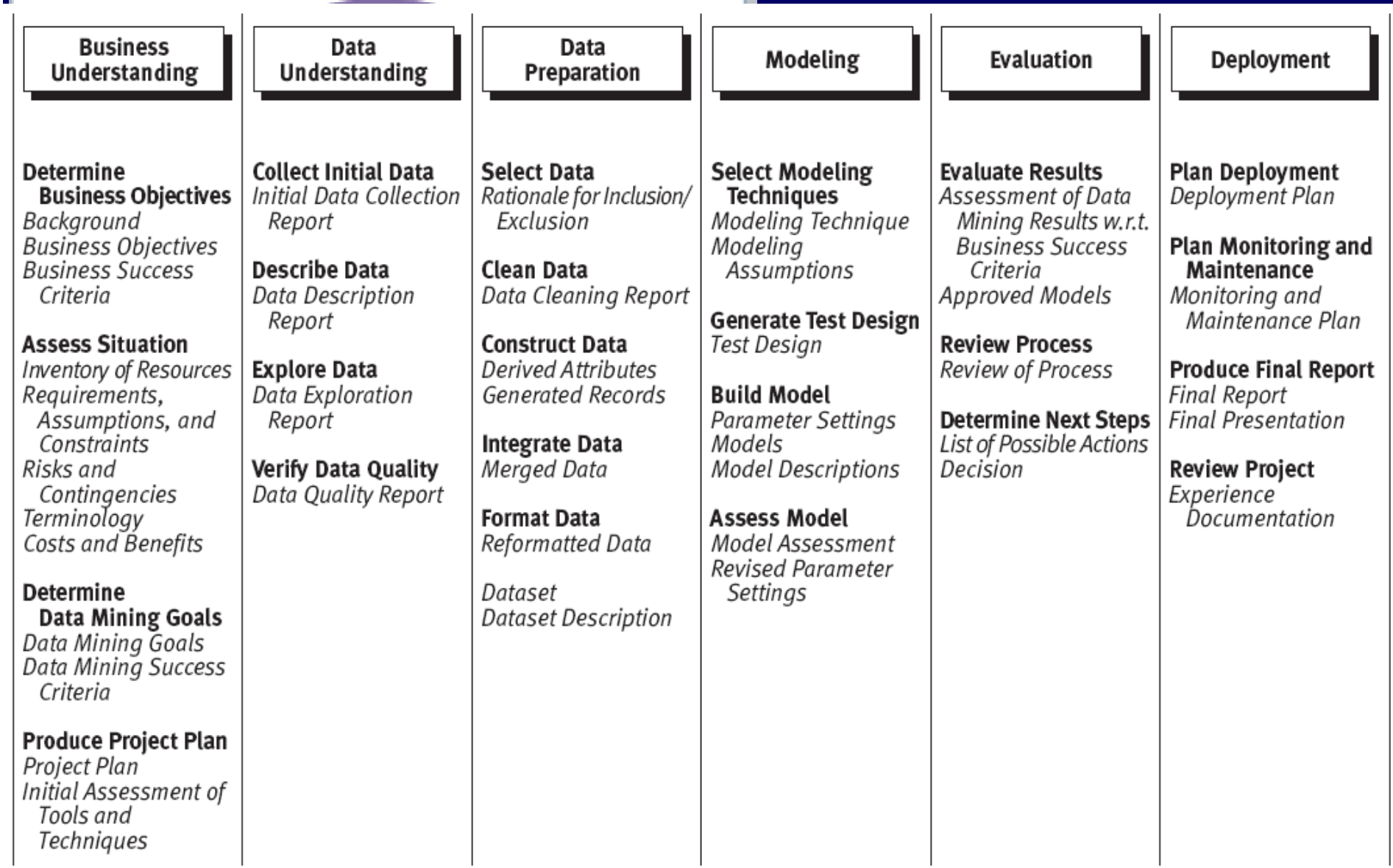
- divorcé
- 5 enfants à charge
- chômeur en fin de droit
- compte à découvert

# Expérience de l'entreprise : ses clients et leur comportement

- coûteuse en stockage
- inexploitée

Comment et à quelles fins utiliser cette expérience accumulée

# CRISP-DM 1.0, Step-by-step Data Mining Guide, SPSS Publication
# Le processus ECD (Extraction de Connaissances à partir de Données)
# Knwoledge Discovery in Databases (KDD)



## Est-ce vraiment nouveau ?

Définition :
Processus non-trivial d 'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données (Fayyad, 1996)

Data Mining : Une nouvelle façon de faire de la statistique ?
http://cedric.cnam.fr/~saporta/DM.pdf

*L'analyse des données est un outil pour dégager de la gangue des données le pur diamant de la véridique nature.*» (J.P.Benzécri1973)

The basic steps for developing an effective process model ?
http://www.itl.nist.gov/div898/handbook/pmd/section4/pmd41.htm

1. Model selection
2. Model fitting
3. Model validation

# Spécificités du Data Mining ?

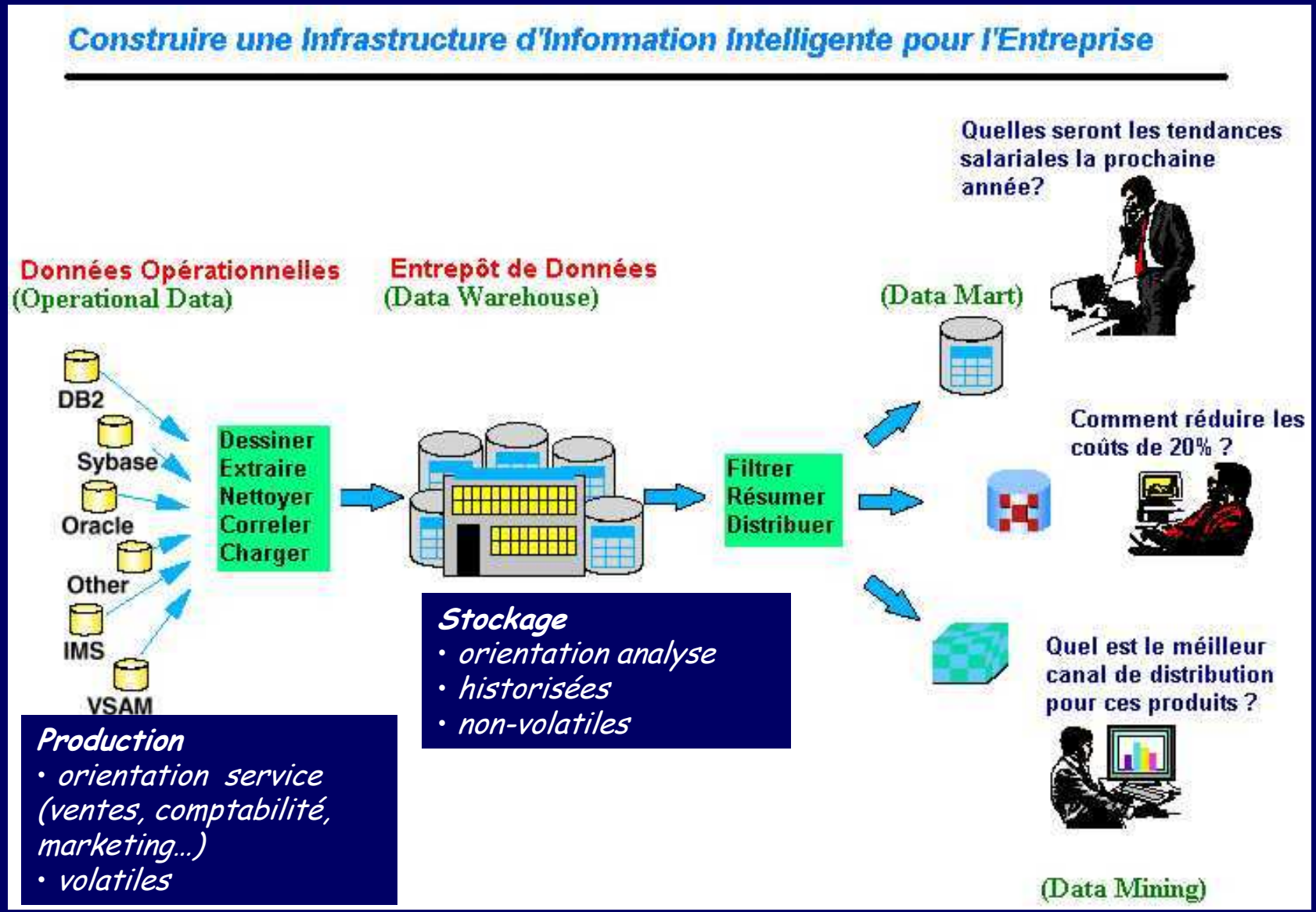
(1) Sources de données
(2) Techniques utilisées
(3) Multiplicité des supports

# Spécificités du Data Mining

## Sources de données

- valoriser les fichiers de l'entreprise
- construire des entrepôts
- modifier le système d'information de l'entreprise

# Les sources de données



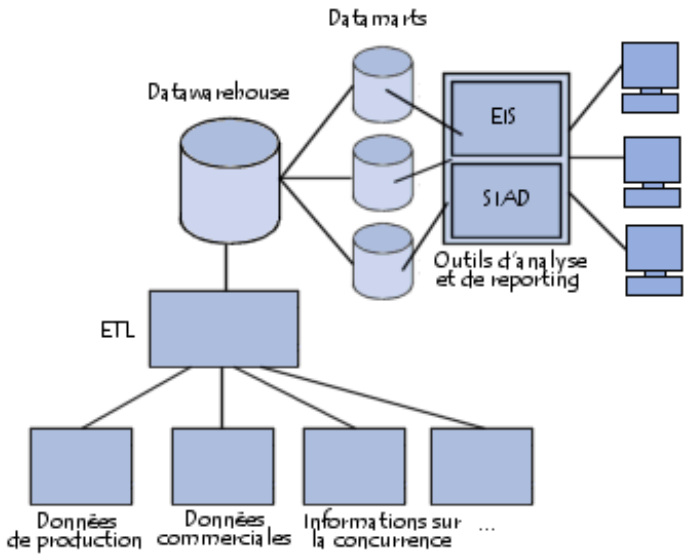
# B.D. de gestion vs. B.D. décisionnelles

| | Systèmes de gestion (opérationnel) | Systèmes décisionnels (analyse) |
|---|---|---|
| Objectif | dédié au métier et à la production<br>ex: facturation, stock, personnel | dédié au management de l'entreprise (pilotage et prise de décision) |
| Volatilité (perennité) | données volatiles<br>ex: le prix d'un proiduit évolue dans le temps | données historisées<br>ex: garder la trace des évolutions des prix, introduction d'une information daté |
| Optimisation | pour les opérations associées<br>ex: passage en caisse (lecture de code barre) | pour l'analyse et la récapitulation<br>ex: quels les produits achetés ensembles |
| Granularité des données | totale, on accède directement aux informations atomiques | agrégats, niveau de synthèse selon les besoins de l'analyse |

Entrepôts/Datamarts : Sources de données pour l'analyse
Conséquence : la volumétrie devient un élément important

# Data Mining vs. Informatique Décisionnelle (Business Intelligence)

L'**informatique décisionnelle** (… BI pour *Business Intelligence*) désigne les moyens, les outils et les méthodes qui permettent de collecter, consolider, modéliser et restituer les données d'une entreprise en vue d'offrir une aide à la décision et de permettre aux responsables de la stratégie d'une entreprise d'avoir une vue d'ensemble de l'activité traitée.
*(http://fr.wikipedia.org/wiki/Informatique_d%C3%A9cisionnelle)*




http://www.commentcamarche.net/entreprise/business-intelligence.php3

- Sélectionner les données (par rapport à un sujet et/ou une période)
- Trier, regrouper ou répartir ces données selon certains critères
- Élaborer des calculs récapitulatifs « simples » (totaux, moyennes conditionnelles, etc.)
- Présenter les résultats de manière synthétique (graphique et/ou tableaux de bord) → REPORTING

La notion de modélisation « statistique » (apprentissage, exploration de données) est mise de côté !

# Spécificités du Data Mining

Techniques d'exploration de données :

- Des techniques d'origines diverses, issues de cultures différentes
- …mais qui traitent des problèmes similaires
- et qui partent toujours d'un tableau de données

# Des cultures et des communautés différentes…

## Statistiques
Théorie de l'estimation, tests
Économétrie

*Maximum de vraisemblance et moindres carrés*
*Régression logistique, …*

## Analyse de données (Statistique exploratoire)
Description factorielle
Discrimination
Clustering

Méthodes géométriques, probabilités
ACP, ACM, Analyse discriminante, CAH, …

| | var 1 | var 2 | … | var J |
|---|---|---|---|---|
| individu 1 | | | | |
| individu 2 | | valeurs | | |
| … | | | | |
| individu n | | | | |

## Informatique « Machine Learning »
Apprentissage symbolique
Reconnaissance de formes

Une étape de l'intelligence artificielle
Réseaux de neurones, algorithmes génétiques…

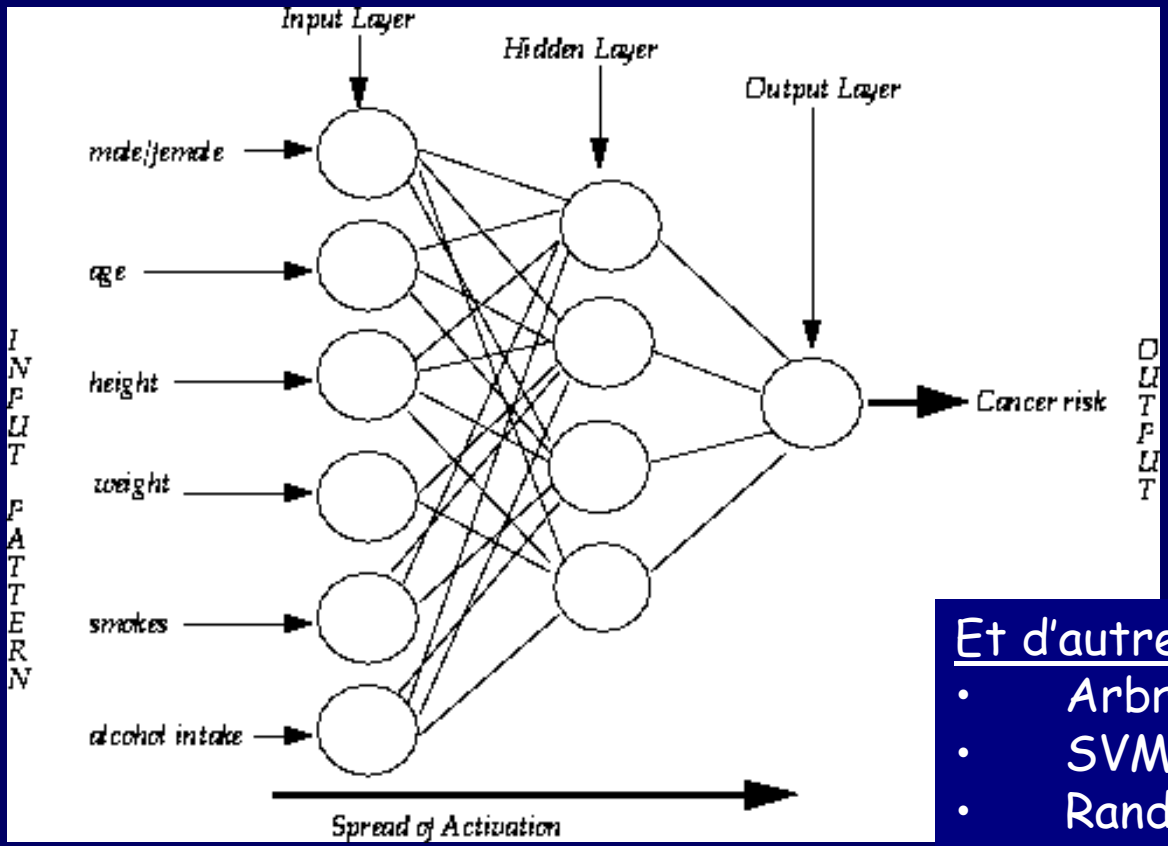
## Informatique (Base de données)
Exploration des bases de données

Volumétrie
Règles d'association, motifs fréquents, …

Très souvent, ces méthodes reviennent à optimiser les mêmes critères, mais avec des approches / formulations différentes

# Techniques issues de l'Intelligence Artificielle « Machine Learning »

## Les réseaux de neurones artificiels




Et d'autres encore…

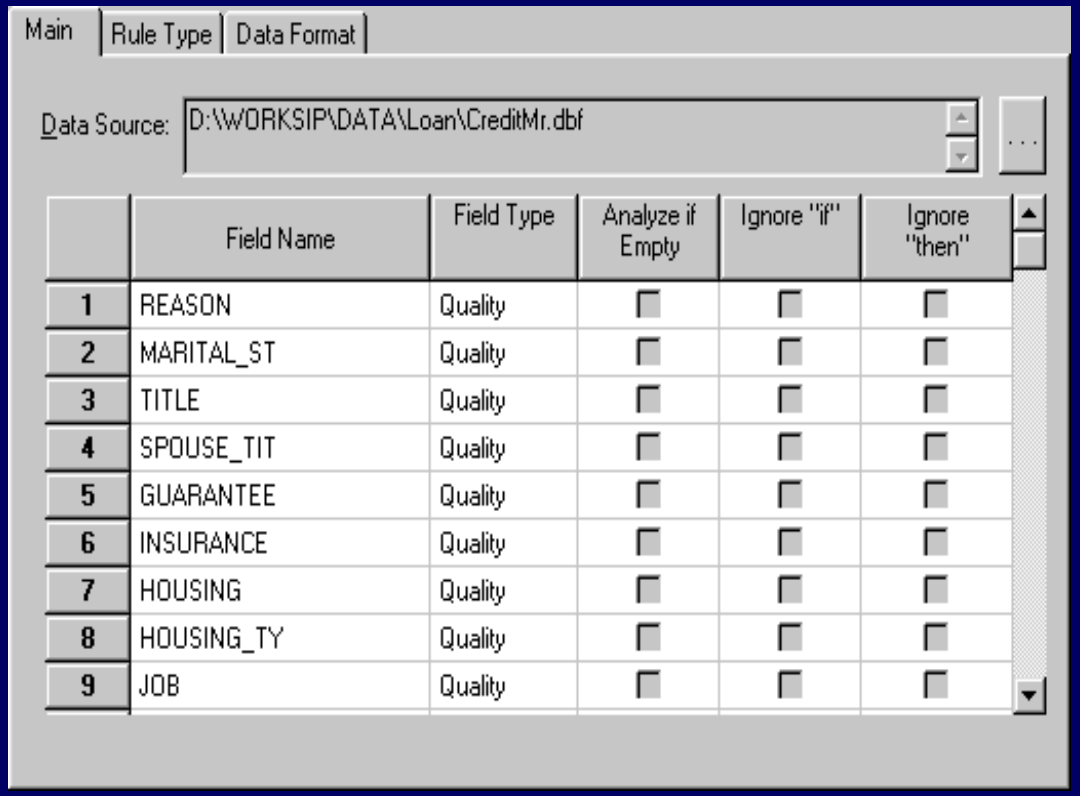
- Arbres de décision
- SVM (Support vector machine)
- Random Forest
- Etc.

Une lecture possible

1. Biais de représentation
2. Biais d'apprentissage

# Techniques en provenance des BD

## Les règles d'association




| | Field Name | Field Type | Analyze if Empty | Ignore "if" | Ignore "then" |
|---|---|---|---|---|---|
| 1 | REASON | Quality | ☐ | ☐ | ☐ |
| 2 | MARITAL_ST | Quality | ☐ | ☐ | ☐ |
| 3 | TITLE | Quality | ☐ | ☐ | ☐ |
| 4 | SPOUSE_TIT | Quality | ☐ | ☐ | ☐ |
| 5 | GUARANTEE | Quality | ☐ | ☐ | ☐ |
| 6 | INSURANCE | Quality | ☐ | ☐ | ☐ |
| 7 | HOUSING | Quality | ☐ | ☐ | ☐ |
| 8 | HOUSING_TY | Quality | ☐ | ☐ | ☐ |
| 9 | JOB | Quality | ☐ | ☐ | ☐ |

*If* MARITAL_ST *is* Divorced
*Then*
SPOUSE_TIT *is* None
*Rule's probability:* 0.952
*The rule exists in* 40 *records.*

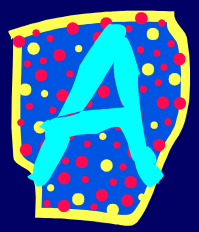
*If* MARITAL_ST *is* Divorced
and LOAN_LENGT = 4.00
*Then*
GUARANTEE *is* No
*Rule's probability:* 0.966
*The rule exists in* 28 *records.*

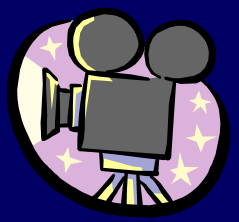
# Spécificités du Data Mining

Élargissement des supports
- Text mining
- Image mining
- …autres…

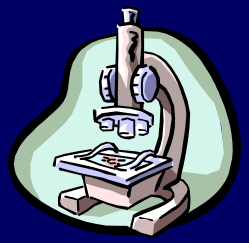
L'appréhension des sources multiples

# Élargir les supports – Les données non structurées





Rôle fondamental de la préparation des données !



| | var 1 | var 2 | ... | var J |
|---|---|---|---|---|
| individu 1 | | | | |
| individu 2 | | valeurs | | |
| ... | | | | |
| individu n | | | | |

Prédiction
Structuration
Description
Association

Les applications

Filtrage automatique des e-mails (spams,...)
Reconnaissance de la langue à une centrale téléphonique
Détection des images à problèmes sur le web
Analyse des mammographies
Etc.

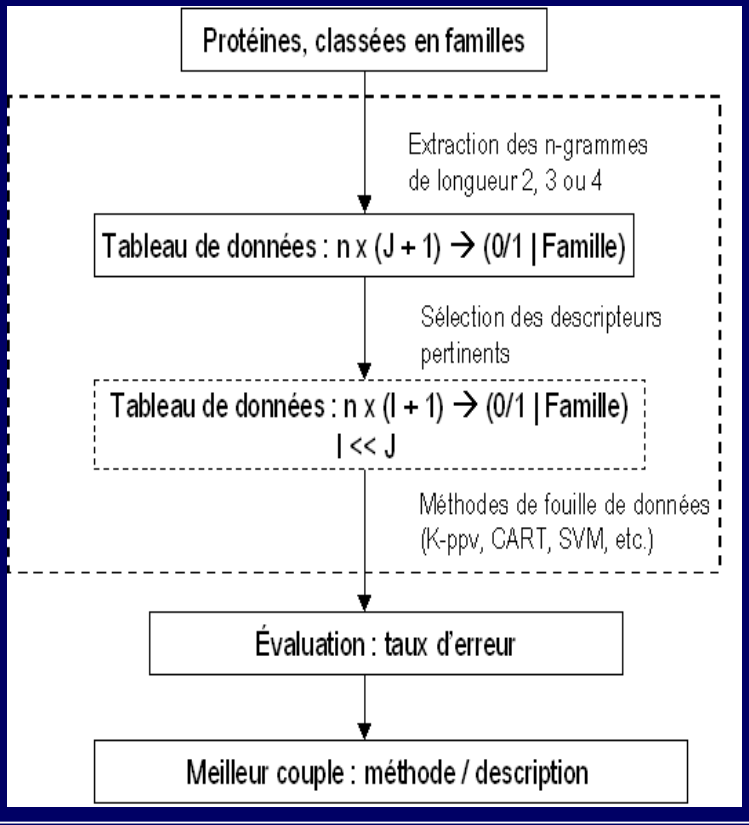
# Classement de protéines

Les protéines sont décrites par une suite d'acides aminés
Il y a 20 types d'acides aminés (alphabet de 20 caractères)

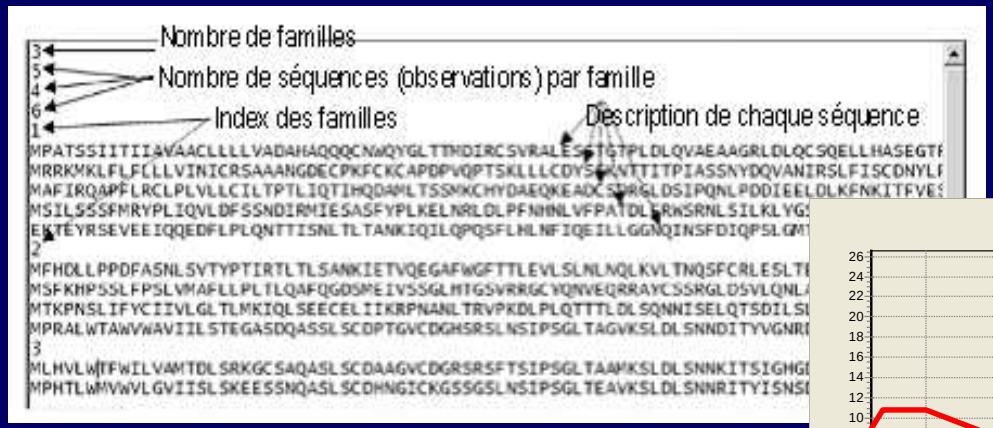
Les protéines sont regroupés selon leur fonction
→ Les protéines proposant les mêmes fonctions sont « similaires »

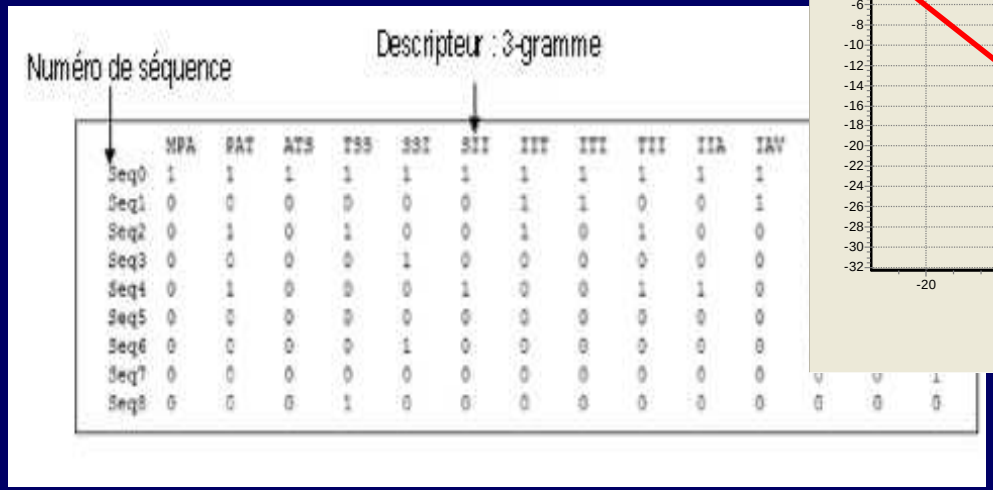
Peut-on construire une fonction de classement permettant d'associer une protéine à une fonction à partir de sa description ?

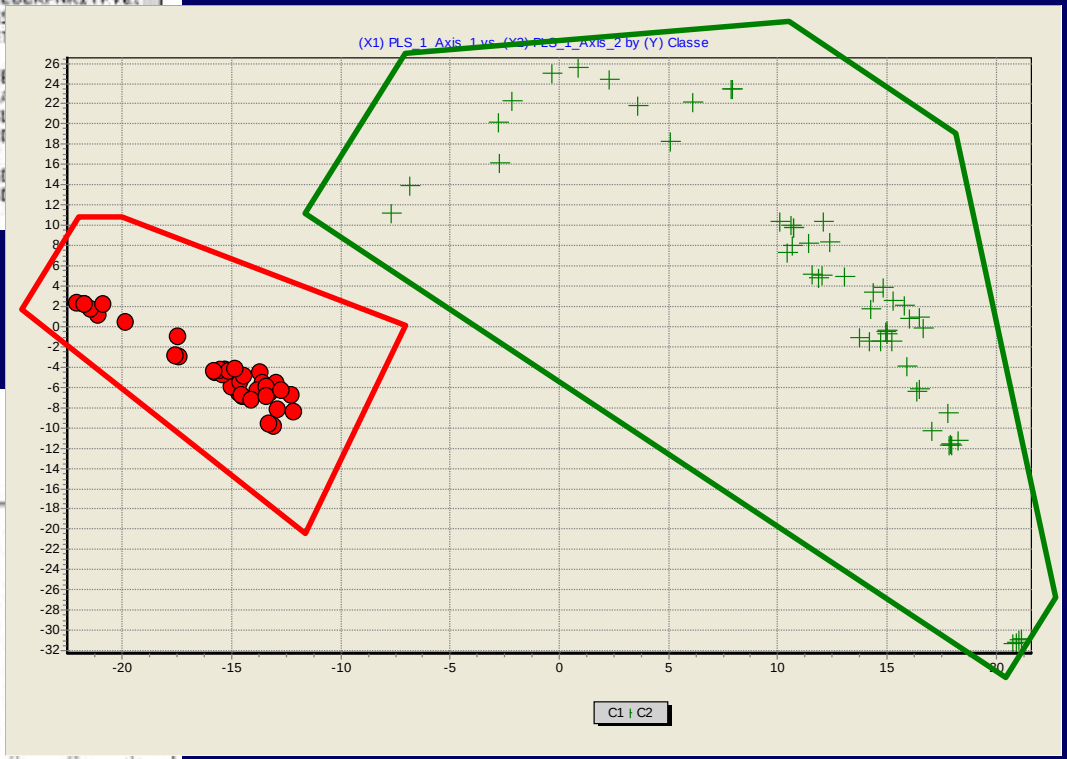
Démarche inspirée de la catégorisation de textes



# Classement de protéines




Discrimination de 2 familles
(Axes PLS)

Danger : Créer ses propres problèmes

# Scoring

Le « ciblage clientèle »

L'application reine du Data Mining

# Le publipostage pour promouvoir un produit

Objectif : promouvoir un produit

Rôle du ciblage : solliciter les clients les plus réceptifs
- optimiser un budget limité
- ne pas agacer les clients « hostiles »
- se donner une idée des performances du ciblage

Outils :
- base de données clientèle
- une variable supplémentaire : clients appétents (+) et non-appétents (-) [*totalement inconnue au départ*]
- construire un « score » pour trier la base selon l'appétence du client
- envoyer le courrier en priorité aux clients réceptifs

Remarque : la démarche peut être reproduite dans d'autres domaines (campagne de dépistage,…)

# La démarche du ciblage

*Schéma général*

2000 clients sollicités au hasard
100 clients ont répondu positivement = 100/2000 → 5%

| Title | Insuran | Childı | Wages |
|---|---|---|---|
| Mrs | No | 2 | 1408 |
| Mr | No | 2 | 1294 |
| Mrs | No | 1 | 1810 |
| Mrs | Yes | 0 | 1800 |
| Mr | No | 5 | 1770 |
| Mr | No | 1 | 1550 |
| Mrs | Yes | 2 | 1561 |
| Mrs | Yes | 2 | 1561 |
| Mrs | No | 1 | 1660 |
| Mrs | No | 2 | 1408 |
| Mrs | Yes | 1 | 1402 |
| Mrs | No | 0 | 862 |
| Mr | Yes | 1 | 1914 |
| Mrs | No | 2 | 2324 |
| Mrs | No | 2 | 862 |
| Mrs | No | 0 | 892 |
| Mr | No | 1 | 2214 |
| Mrs | No | 1 | 2021 |
| Mr | No | 1 | 1425 |
| Mrs | No | 0 | 1863 |
| Mrs | No | 0 | 1318 |
| Mr | Yes | 1 | 1800 |
| Mrs | No | 1 | 981 |
| Mrs | No | 2 | 2900 |
| Mr | No | 0 | 5400 |

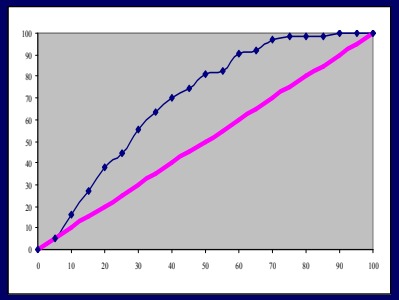
Base de données clientèle
(202.000 clients)

| Title | Insuran | Childı | Wages | Retour |
|---|---|---|---|---|
| Mrs | No | 2 | 1408 | + |
| Mr | No | 2 | 1294 | + |
| Mrs | No | 1 | 1810 | - |
| Mrs | Yes | 0 | 1800 | + |
| Mr | No | 5 | 1770 | + |
| Mr | No | 1 | 1550 | - |
| Mrs | Yes | 2 | 1561 | + |

1000 test

1000 apprentissage

Courbe LIFT
Évaluer la performance du ciblage



$S(R) = \Phi(X)$

Fonction score : permet de classer les clients selon l'appétence

200 000 clients

| Title | Insuran | Child | Wages | SCORE |
|---|---|---|---|---|
| Mr | No | 0 | 2185 | 0.9997 |
| Mrs | No | 1 | 900 | 0.9992 |
| Mrs | No | 2 | 3000 | 0.9987 |
| Mr | No | 1 | 1410 | 0.9976 |
| Mrs | No | 2 | 1600 | 0.9956 |
| Mrs | No | 0 | 1520 | 0.9931 |
| Mr | No | 0 | 5400 | 0.9898 |
| Mrs | No | 2 | 2400 | 0.9888 |
| Mrs | Yes | 3 | 1237 | 0.987 |
| Mr | No | 2 | 1572 | 0.9863 |
| Mrs | No | 1 | 2621 | 0.9861 |
| Mrs | No | 2 | 1782 | 0.9855 |
| Mr | No | 0 | 2400 | 0.9841 |
| Mrs | No | 2 | 1020 | 0.9836 |
| Mrs | No | 0 | 1812 | 0.9828 |
| Mrs | No | 0 | 1470 | 0.9821 |
| Mrs | No | 2 | 1320 | 0.9799 |
| Mrs | No | 1 | 1080 | 0.9788 |

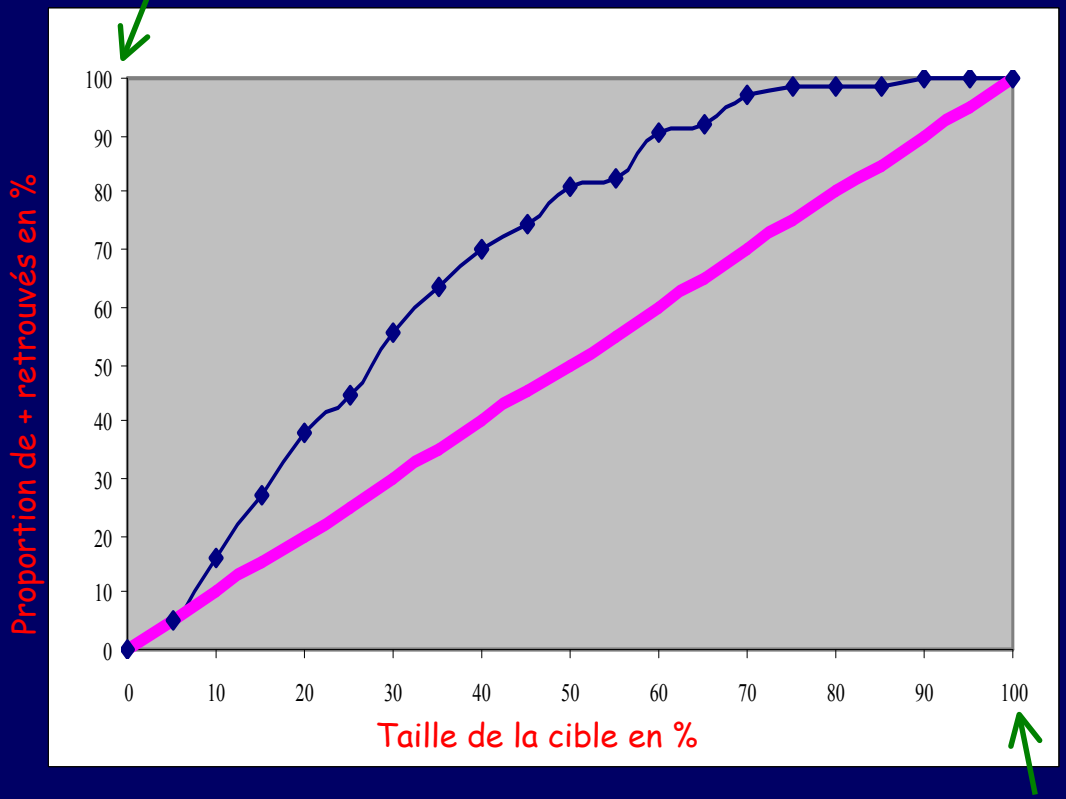
## Etapes

1. Base de données
2. Constitution des données étiquetées
3. Apprentissage d'un modèle de prédiction
4. Évaluation des performances
5. Déploiement

Potentiel de + : 5% de 200 000 = 10 000 clients +

# Intérêt de la courbe LIFT [GAIN CHART] par rapport à une évaluation « classique » (matrice de confusion + taux d'erreur de classement)

100 % des + dans le fichier test = 50 individus
100 % du potentiel dans la base = 10.000 individus




100 % de la cible dans le fichier test = 1.000 individus
100 % de la cible dans la base = 200.000 individus

## Avantages

1. « Transposabilité » des résultats en travaillant sur les pourcentages

2. Gradation de la prise de décision

   → Budgets fixés

   → Parts de marchés fixés

# Le scoring dans la pratique
# (ou la pratique du scoring)

Utilisation directe des bases de l'entreprise, déjà étiquetées

\+ Réduction des coûts
Rapidité d'action

\- Les « positifs » sont ceux qui ont déjà pris le produit
Comment ont été sollicités les « positifs » ?
Parmi les négatifs, certains ont été sollicités, d'autres non…

Est-ce que le « scoring » dans ce contexte permet de retrouver les clients les plus réactifs par rapport à un type de sollicitation ?

# Outils logiciels

Pour les enseignements …

# Logiciels de DATA MINING – Fonctionnalités

## Accès et préparation des données
Accéder à un fichier / une BD
Rassembler des sources différentes

## Méthodes de Fouille de données
Lancer les calculs avec différents algorithmes
Bibliothèque de méthodes

## Enchaîner les traitements
Faire coopérer les méthodes sans programmer

## Évaluer les connaissances
Validation croisée, etc.

## Exploiter les sorties
Rapports, visualisation interactive, etc.

## Appliquer/exploiter les modèles
Modèles en XML (PMML), code C, DLL compilées
Prédiction directe sur de nouveaux fichiers

Logiciels commerciaux
Prototypes de recherche

# Organisation des logiciels

## Logiciels pilotés par menu

(+) Organisation de type « tableur »
(+) Rapidité de prise en main
(-) Enchaînement « à la main » des traitements
(-) Pas de trace des opérations effectuées
(-) Et donc reproductibilité difficile des traitements

STATISTICA, XL-MINER, …

## Ligne de commande

(+) Souplesse et puissance de la programmation
(+) Sauvegarde des traitements, reproductilibité
(-) Apprentissage d'un langage

SAS, S-PLUS, R,…

## Filière (diagramme de traitements)

(+) Programmation « visuelle » - Pas d'apprentissage
(+) Enchaînement des traitements
(+) Sauvegarde des traitements, reproductilibité
(-) Pas la puissance d'un « vrai » langage de programmation

(1)

SPAD
SAS → Enterprise Miner
SPSS → Clementine
S-PLUS → Insightful Miner
STATISTICA → Data Miner
…

*Et les outils libres (gratuits, open source…)* → (2)

ORANGE
TANAGRA
WEKA

# Logiciels gratuits pour les enseignements ?

1. S'attacher au fond (*méthode , savoir*) et non pas à la forme (*savoir-faire*)

*Qu'importe le logiciel et son mode opératoire ?*

2. Former des étudiants qui iront sur le marché du travail

*Il existe des « standards » selon les domaines, les logiciels commerciaux les respectent, il faut familiariser les étudiants avec ces standards*

4. Notoriété ?

*Former des étudiants sur les logiciels « inconnus »…*

3. L'apprentissage d'un logiciel ne doit pas requérir des compétences spécifiques (hors domaine)

*Accès à des formats de fichiers reconnus, langage de script spécifiques*

# Cahier des charges d'un logiciel pour les enseignements

1. Gratuité sans restrictions – Au moins dans le cadre des enseignements
2. Installation simplifiée – Pas de serveurs lourds à installer
3. Gestion simplifiée des données - Format texte / tableur
4. Fonctionnement par diagramme de traitements – La fameuse notion de « filière »
5. Évaluation des méthodes (supervisées), outils de scoring et comparaisons
6. Résultats lisibles, possibilité de les reprendre dans un traitement de texte

Pouvoir définir des traitements, les comparer, les évaluer sur des jeux de données sans avoir à passer par un apprentissage compliqué d'un logiciel spécifique

## WEKA – ORANGE – TANAGRA pour les enseignements

Des outils riches mais des objectifs, des cultures et des approches très différentes

**WEKA**

Privilégié pour les expérimentations
et la comparaison de performances en supervisé

**ORANGE**

Très marqué « machine learning », souple
avec des efforts pour la convivialité et la simplicité
Exploration graphique excellente

**TANAGRA**

Un outil pluri-culturel tourné vers les études
de statistique exploratoire et de fouille de données

**Alors ?**

OUI pour les aspects méthodologiques

Ces outils permettent d'assurer des enseignements de Data Mining dans le sens comprendre et mettre en oeuvre les méthodes, interpréter les résultats… sur des fichiers plats de taille modérée (quelques dizaines de milliers d'obsv…)

NON pour les aspects opérationnels

Accès aux entrepôts, le reporting dynamique, le déploiement…

# Bibliographie : Culture générale Data Mining

« Le Data mining », R. Lefebure et G. Venturi, ed. Eyrolles, 2001.

« Data Mining et Statistique décisionnelle », S. Tufféry, ed. Technip, 2005.
*Plutôt guide pratique : repères pour les projets, opportunités, rapide et très peu technique*

« Extraction de connaissances à partir de données », D. Zighed et R. Rakotomalala, in Techniques de l'Ingénieur, H3-744, 2003.

http://morgon.univ-lyon2.fr/Introduction_au_datamining_cours.htm