

# Analyse de données

Jamal Atif

[jamal.atif@dauphine.fr](mailto:jamal.atif@dauphine.fr)

Université Paris-Dauphine, Licence MIDO



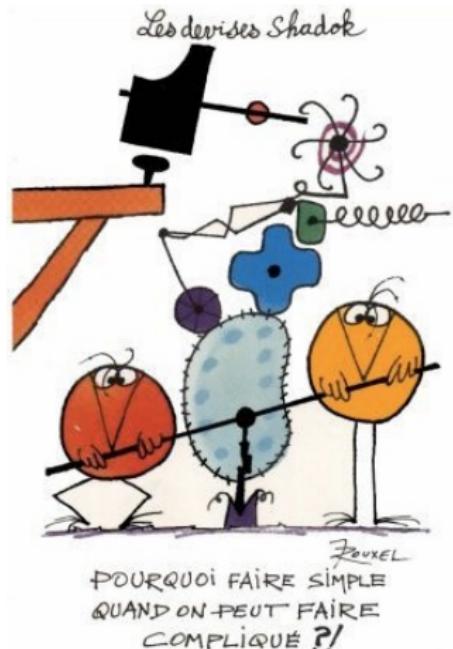
2014-2015

# Introduction

# L' analyse de données, qu'est-ce que c'est ?



# Question de vocabulaire... (1)



Attention :

- ▶ historiquement : plusieurs « point de départ »
- ▶ domaine récent dont le vocabulaire n'est pas fixé
- ▶ évolution rapide
- ▶ domaine applicatif *versus* domaine de recherche

## Question de vocabulaire... (2)

Conséquence : on ne va pas parler que d'analyse de données

- ▶ reconnaissance des formes  
*(pattern recognition)*
- ▶ Analytics
- ▶ apprentissage automatique  
*(machine learning)*
- ▶ fouille de données (*data mining*)
- ▶ intelligence artificielle
- ▶ statistique
- ▶ ...

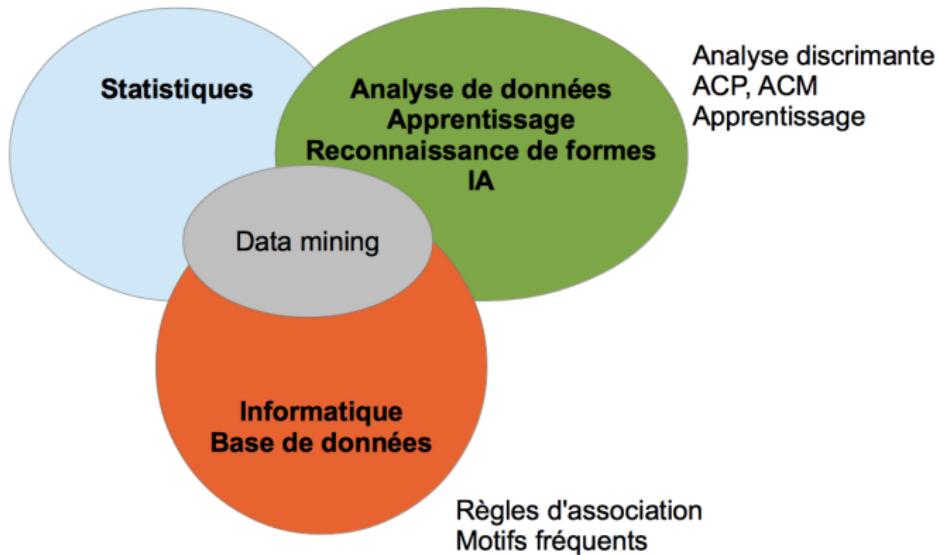
⇒ domaines différents avec des intersections plus ou moins grandes

## Data Sciences

Dans ce cours : Analyse et Fouille des Données  
(AFD)

# La rencontre de plusieurs disciplines

Régression  
Maximum de vraisemblance, moindres carrés



# Dans ce cours

## Définitions

- ▶ Extraction de connaissances à partir de données (KDD) :
  - ▶ Cycle de découverte d'information regroupant la conception des grandes bases de données ou les entrepôts de données (data warehouses).
  - ▶ Ensemble des traitements à effectuer pour extraire de l'information aux données.
  - ▶ L'**analyse et la fouille de données** est un des traitements.
- ▶ Analyse et fouille de données = data mining
  - ▶ Ensemble des techniques d'exploration de données permettant d'extraire des connaissances sous la forme de **modèles de description** afin de :
    - ▶ **Décrire** le comportement actuel des données.
    - ▶ Et/ou **Prédire** le comportement futur des données.

# Dans ce cours

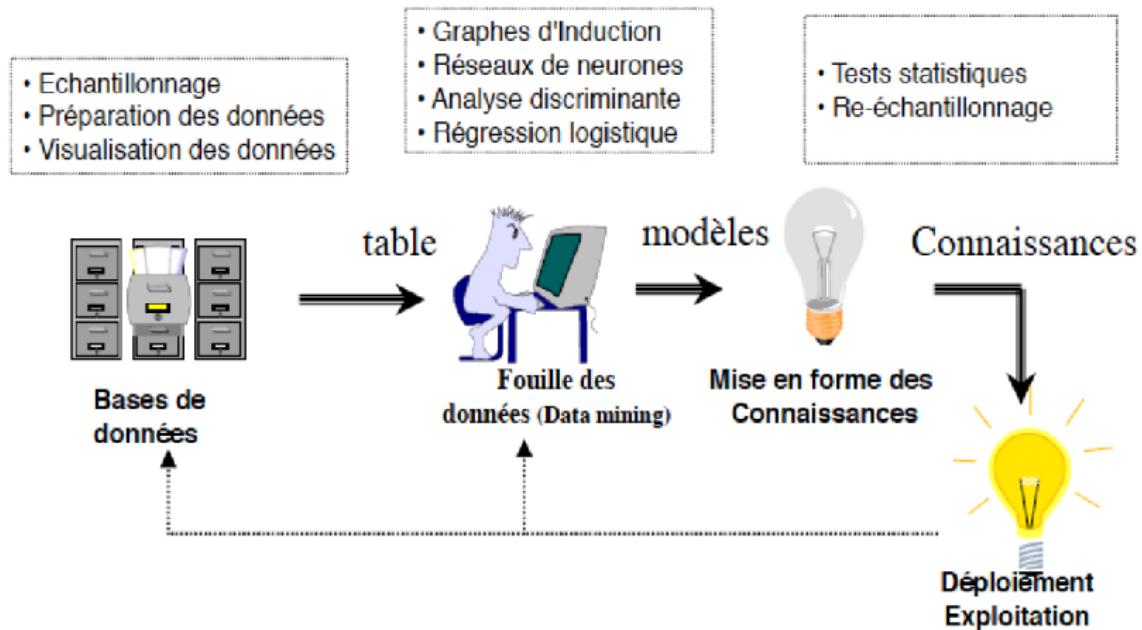
## Définition (S. Tuffery)

L'AFD est l'ensemble des :

- ▶ algorithmes et méthodes
  - ▶ ... destinés à l'exploration et l'analyse
  - ▶ ... de (souvent) grandes bases de données informatiques
  - ▶ ... en vue de détecter dans ses données des règles, des associations, des tendances inconnues (non fixées a priori), des structures particulières restituant de façon concise l'essentiel de l'information utile
  - ▶ ... pour l'aide à la décision.

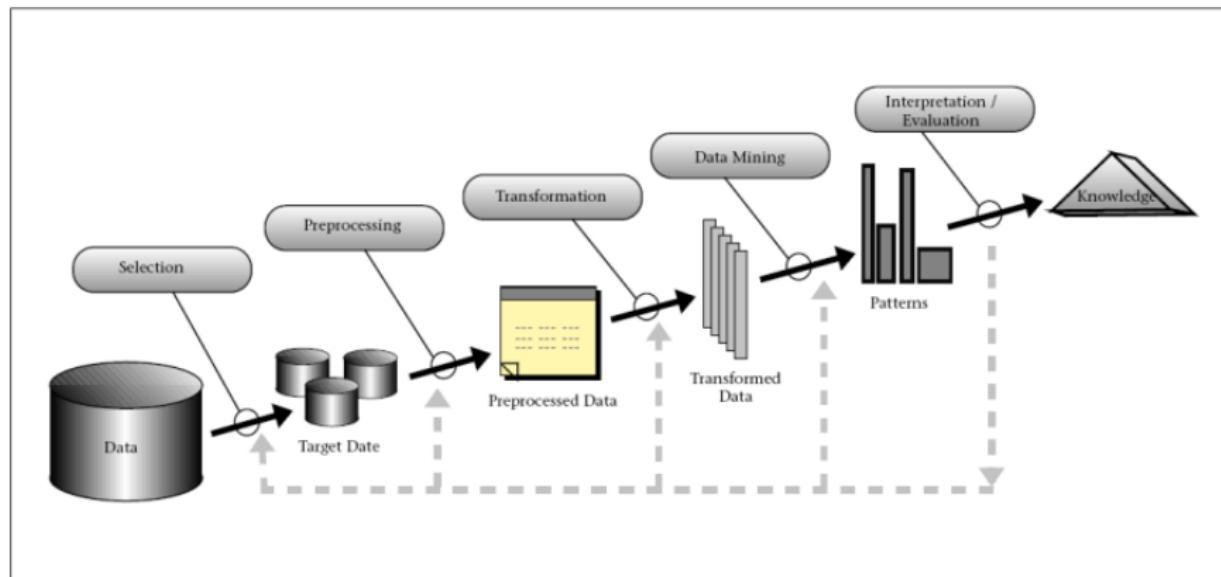
# AFD : une démarche plus qu'une théorie

Processus ECD (extraction de connaissances à partir de données) ou KDD (Knowledge Discovery in Databases)



# AFD : une démarche plus qu'une théorie

Processus ECD (extraction de connaissances à partir de données) ou KDD  
(Knowledge Discovery in Databases)



# Pourquoi l'AFD ?

Un sujet d'actualité...

L'exploitation des données est importante car c'est :

- ▶ **méthode scientifique** ⇒ nécessité de savoir exploiter des données
  - ▶ c'est la base de la méthode scientifique (observations → lois/règles)
  - ▶ les données (et leur exploitation) au cœur de beaucoup d'avancés récentes
- ▶ **source de revenus**
  - ▶ modèle économique des entreprises du web (Google, Facebook, Amazon, ...)
  - ▶ fournit un service gratuit
  - ▶ seule « valeur » : capacité à exploiter les données collectées
- ▶ **nouvelle « approche de programmation »**
  - ▶ « rêve » de l'intelligence artificielle : l'ordinateur qui apprend
  - ▶ il y a des algorithmes que l'on ne peut pas/sait pas formaliser

# Références

- ▶ [http://www.nytimes.com/2009/08/06/technology/06stats.html?\\_r=3](http://www.nytimes.com/2009/08/06/technology/06stats.html?_r=3)
  - ▶ <http://radar.oreilly.com/2010/06/what-is-data-science.html>
- ⇒ intérêt de savoir utiliser des méthodes statistiques pour exploiter de grandes masses de données aussi bien d'un point de vu économique (facebook, google, ...) que scientifique (CERN et autre).

## Facebook's Data Science Group

“... on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of our analyses to other members of the organization”

# Pourquoi l'AFD ?

Dans l'industrie

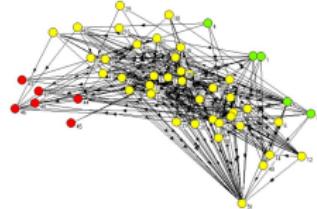


## Carte de crédit

- ▶ tous les achats sont enregistrés
- ▶ détection des fraudes/comportement à risque
- ▶ ciblage
- ▶ accord de prêt
- ▶ ...

## Navigation Web

- ▶ historique de la navigation
- ▶ ciblage/marketing
- ▶ optimisation des sites / du traffic
- ▶ ...

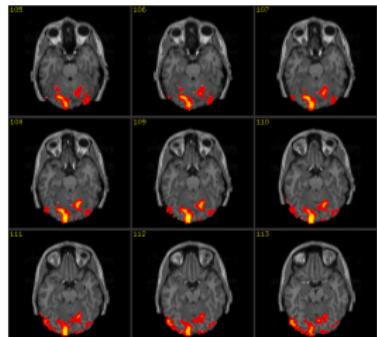


# Pourquoi l'AFD ?

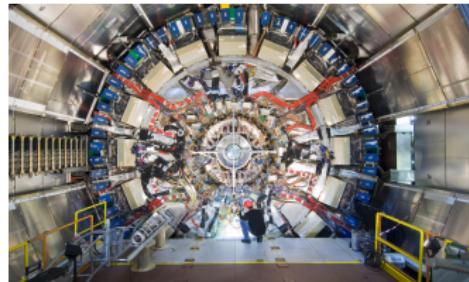
Pour la science

## fMRI

- ▶ functional Magnetic Resonance Imaging
- ▶ variation de pression sanguine en réponse à des stimuli
- ▶ brain computer interface



## Big Science



- ▶ détecteur ATLAS du CERN
- ▶ 40M événements par secondes, 25Mo par événement
- ▶ 1Po de données générées par secondes à analyser
- ▶ même situation en biologie, astronomie,  
...

# Pourquoi l'AFD ?

Pour la société



- ▶ tous les textes et discussion du parlement européen sont disponibles...
- ▶ ...avec leur traduction/interprétation
- ▶ **corpus parallèle** : les phrases sont **alignés**
- ▶ utilisable pour apprendre :
  - ▶ des dictionnaires
  - ▶ des systèmes de traduction automatique
  - ▶ des mémoires de traduction
- ▶ ⊕ analyse « politique » des données

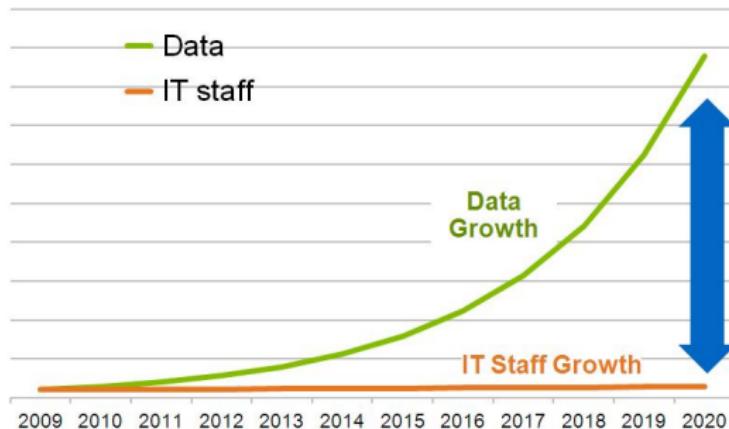
## Et encore

- ▶ historique des achats
- ▶ historique des passages de frontières
- ▶ liste des appels
- ▶ analyse de traffic
- ▶ test A/B pour choisir les prix
- ▶ pollution
- ▶ données médicales
- ▶ ...

# Pourquoi l'AFD ?

## Le fossé des données (**data gap**)

### The Data Management Gap



Une grosse quantité de données qui n'est jamais analysée  
⇒ mettre en place des mécanismes d'analyse automatique.

**Big Data**

# AFD : composants de base

Grande quantité de données + algorithmes efficaces

Un domaine qui s'appuie sur :

- ▶ **La disponibilité de grandes quantités de données**
  - ▶ Si ensemble trop petit, les structures peuvent ne résulter que du hasard.
  - ▶ On peut espérer qu'un gros volume de données représente bien l'univers (échantillon).
- ▶ **Des algorithmes sûrs et efficaces**
  - ▶ Algorithmes sûrs : fondés théoriquement, corrects.
  - ▶ Efficaces en temps et en espace.
  - ▶ Résultats interprétables.
  - ▶ Paramètres ajustables facilement et rapidement.

# AFD : Un exemple

Issu du livre de Adriaans and Zantige (d'après B. Espinasse)

- ▶ Un éditeur vend 5 sortes de magazines : sport, voiture, maison, musique, cinéma.
- ▶ Il veut étudier ses clients pour découvrir de nouveaux marchés ou vendre plus à ses clients habituels.

## Quelques questions

1. Combien de personnes ont pris un abonnement à un magazine de cinéma cette année ?
2. A-t-on vendu plus d'abonnement de magazines de sport cette année que l'année dernière ?
3. Est-ce que les acheteurs de magazines de musique sont aussi amateurs de cinéma ?
4. Quelles sont les caractéristiques principales des lecteurs de magazine de cinéma ?
5. Peut-on prévoir les pertes de client et prévoir des mesures pour les diminuer ?

## AFD : Un exemple

1 : Combien de personnes ont pris un abonnement à un magazine de cinéma cette année ?

Requête SQL à partir des données opérationnelles suffit si les tables concernées ont été suffisamment indexées.

2 : A-t-on vendu plus d'abonnement de magazines de sport cette année que l'année dernière ?

- ▶ Nécessite de garder toutes les dates de souscription, même pour les abonnements résiliés.
- ▶ Requêtes multidimensionnelles de type OLAP.

# AFD : Un exemple

3 : Est-ce que les acheteurs de magazine de musique sont aussi amateurs de cinéma ?

- ▶ Exemple simplifié de problème où l'on demande si les données vérifient une règle.
- ▶ Réponse formulée par une valeur estimant la probabilité que la règle soit vraie.
- ▶ Utilisation d'outils statistiques.

# AFD : Un exemple

## 4. Quelles sont les caractéristiques principales des lecteurs de magazine de cinéma ?

Question plus ouverte, il s'agit de trouver une règle et non plus de la vérifier ou de l'utiliser.

## 5 : Peut-on prévoir les pertes de client et prévoir des mesures pour les diminuer ?

Question ouverte : il faut disposer d'indicateurs comme durée d'abonnement, délai de paiement, ...

C'est pour ce type de questions que sont mis en oeuvre les outils d'analyse et de fouille de données

# Les données ?

Les données peuvent être vues comme une collection d'objets (enregistrements) et leurs attributs.

- ▶ Un attribut est une propriété et ou une caractéristique de l'objet.
- ▶ Un ensemble d'attributs décrit un objet.

**Attributes**

The diagram illustrates data as objects and attributes. A large curly brace on the left side groups the rows of the table, labeled 'Objects'. Above the table, another curly brace groups the columns, labeled 'Attributes'. The table itself has columns: Tid, Refund, Marital Status, Taxable Income, and Cheat. The data rows are numbered 1 through 10.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

# Attribut - valeur

- ▶ La valeur d'un attribut est un nombre ou un symbole.
- ▶ Ne pas confondre attribut et valeur

## Types

- ▶ Quantitative (numérique, exprime une quantité)
  - ▶ Discrète (ex : nombre d'étudiants dans un cours) ou continue (ex : longueur)
  - ▶ Echelle proportionnelle (chiffre d'affaires, taille), ou échelle d'intervalle (température, QI)
- ▶ Qualitative
  - ▶ Variable ordinaire (classement à un concours, échelle de satisfaction client)
  - ▶ Variable nominale (couleur de yeux, diplôme obtenu, CSP, sexe)
- ▶ Les **modalités** d'une variable sont l'ensemble des valeurs qu'elle prend dans les données  
ex : les modalités de notes sont  $\{0, 1, 2, \dots, 20\}$  les modalités de couleur sont {bleu, vert, noir, ...}

# Exemple de données disponibles

- ▶ Transactions.
- ▶ Bases de données des entreprises.
- ▶ Téléphone portable.
- ▶ Satellites : espace et la terre.
- ▶ Données temporelles : cours de la bourse, météo.
- ▶ Génomique.
- ▶ Données du web.
- ▶ Données textuelles.
- ▶ ...

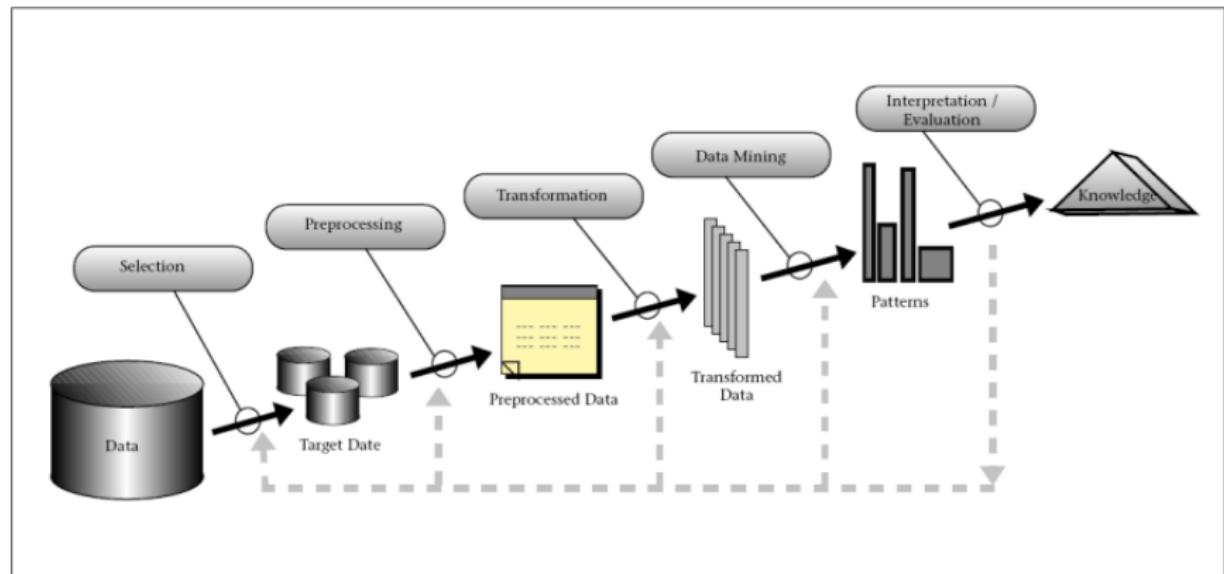
# Types de connaissances extraites

Connaissances sous la forme de modèles de description permettant de

- ▶ **décrire le comportement actuel** des données et/ou
- ▶ **prédir le comportement futur** des données.
- ▶ **Analyses**
  - ▶ e.g. distribution du trafic routier en fonction de l'heure
- ▶ **Règles**
  - ▶ e.g. si un client a acheté un produit alors il sera intéressé par un autre.
- ▶ **Attribution de scores de qualité**
  - ▶ e.g. score de fidélité au client
- ▶ **Classification d'entités**
  - ▶ e.g. mauvais payeurs.

# Processus ECD (extraction de connaissances à partir de données)

Un processus découpé en 5 étapes, une dernière étape étant l'utilisation du modèle.



# Processus ECD (extraction de connaissances à partir de données)

## Un déroulement non linéaire

- ▶ On constate souvent à l'étape de validation que :
  - ▶ les performances obtenues sont insuffisantes.
  - ▶ les utilisateurs du domaine jugent l'information inexploitable.
  - ▶ ...
- ▶ Il faut donc :
  - ▶ Choisir une autre méthode de fouille.
  - ▶ Remettre en cause l'étape de transformation.
  - ▶ Enrichir les données

Dans un projet d'ECD, le temps passé à l'étape de fouille de données ne représente souvent que 20% du temps.

# Etape de sélection des données

Elle consiste à

- ▶ Obtenir des données en accord avec les objectifs de l'ECD.
- ▶ Ces données proviennent le plus souvent (mais pas toujours) de bases de production ou d'entrepts.
  - ▶ Par l'utilisation d'outils de requêtage (SQL, OLAP, ...).
  - ▶ Copie sur une machine adéquate (pour pouvoir les modifier et pour des questions de performance)
- ▶ Structuration des données en champs typés.

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bémol	Rue du moulin, Paris	7/10/2006	Voiture
23134	Bémol	Rue du moulin, Paris	12/5/2006	Musique
23134	Bémol	Rue du moulin, Paris	25/7/2005	BD
31435	Bodinoz	Rue Verte, Nancy	11/11/1111	BD
43342	Airinair	Rue de la source, Brest	30/05/2005	Sport
25312	Talonion	Rue du Marché, Paris	25/02/2007	NULL
43241	Manvussa	NULL	14/4/2006	Sport
23130	Bémolle	Rue du moulin, Paris	11/11/1111	Maison

# Etape de prétraitement

Elle consiste à

- ▶ Nettoyer les données
  - ▶ Corrections des doublons, des erreurs de saisie.
  - ▶ Contrôle sur l'intégrité des domaines de valeurs : détection des valeurs aberrantes.
  - ▶ Détection des informations manquantes
- ▶ Enrichissement des données

# Etape de prétraitement

## Corrections des doublons, des erreurs de saisie

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bémol	Rue du moulin, Paris	7/10/2006	Voiture
23134	Bémol	Rue du moulin, Paris	12/5/2006	Musique
23134	Bémol	Rue du moulin, Paris	25/7/2005	BD
31435	Bodinoz	Rue Verte, Nancy	11/11/1111	BD
43342	Airinair	Rue de la source, Brest	30/05/2005	Sport
25312	Talonion	Rue du Marché, Paris	25/02/2007	NULL
43241	Manvussa	NULL	14/4/2006	Sport
23130	Bémolle	Rue du moulin, Paris	11/11/1111	Maison

# Etape de prétraitement

## Intégrité de domaine

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bémol	Rue du moulin, Paris	7/10/2006	Voiture
23134	Bémol	Rue du moulin, Paris	12/5/2006	Musique
23134	Bémol	Rue du moulin, Paris	25/7/2005	BD
31435	Bodinoz	Rue Verte, Nancy	11/11/1111	BD
43342	Airinair	Rue de la source, Brest	30/05/2005	Sport
25312	Talonion	Rue du Marché, Paris	25/02/2007	NULL
43241	Manvussa	NULL	14/4/2006	Sport
23130	Bémol	Rue du moulin, Paris	11/11/1111	Maison

# Etape de prétraitement

## Information manquante

- ▶ Cas où les champs ne contiennent aucune donnée.
- ▶ Parfois intéressant de conserver ces enregistrements car l'absence d'information peut être informative (e.g. fraude).

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bémol	Rue du moulin, Paris	7/10/2006	Voiture
23134	Bémol	Rue du moulin, Paris	12/5/2006	Musique
23134	Bémol	Rue du moulin, Paris	25/7/2005	BD
31435	Bodinoz	Rue Verte, Nancy	NULL	BD
43342	Airinair	Rue de la source, Brest	30/05/2005	Sport
25312	Talonion	Rue du Marché, Paris	25/02/2007	NULL
43241	Manvussa	NULL	14/4/2006	Sport
23134	Bémol	Rue du moulin, Paris	NULL	Maison

# Etape de prétraitement

## Enrichissement

- ▶ Recours à d'autres bases de données souvent pour ajouter de nouveaux champs en conservant le même nombre d'enregistrements.
- ▶ Plusieurs difficultés :
  - ▶ Relier les données, parfois hétérogènes, entre elles.
  - ▶ Introduction de nouvelles valeurs manquantes et/ou aberrantes.

Client	Date naissance	Revenus	Propriétaire	Voiture
Bémol	13/1/50	20 000	Oui	Oui
Bodinoz	21/5/70	12 000	Non	Oui
Airinair	15/06/63	9 000	Non	Non
Manvussa	27/03/47	15 000	Non	Oui

# Etape de transformation (codage et normalisation)

Une étape très dépendante du choix de l'algorithme de fouilles de données utilisés.

- ▶ Regroupements.
  - ▶ Cas où les attributs prennent un très grand nombre de valeurs discrètes (e.g. adresses que l'on peut regrouper en 2 régions (Paris - Province))
- ▶ Attributs discrets.
  - ▶ Les attributs discrets prennent leurs valeurs (souvent textuelles) dans un ensemble fini donné (e.g. colonne magazine de l'exemple).
  - ▶ Deux représentations possibles : représentation verticale ou représentation horizontale ou éclatée (plus adaptée à la fouille de données)
- ▶ Changements de types pour permettre certaines manipulations comme par exemple des calculs de distance, de moyenne (e.g. date de naissance)
- ▶ Uniformisation d'échelle.
  - ▶ Certains algorithmes sont basés sur des calculs de distance entre enregistrements :
    - ▶ Variations d'échelle selon les attributs peuvent perturber ces algorithmes.

# Etape de transformation (codage et normalisation)

Représentation horizontale ou éclatée

Client	Magazine
23134	Voiture
23134	Musique
23134	BD
31435	BD
43342	Sport
43241	Sport
23134	Maison

Client	Sport	BD	Voiture	Maison	Musique
23134	0	1	1	1	1
31435	0	1	0	0	0
43342	1	0	0	1	0
43241	1	0	0	1	0

# Etape de transformation (codage et normalisation)

## Etape de transformation sur l'exemple

Client	Sport	BD	Voiture	Maison	Musique	DN	Rev	Prop	Voit	PP	DA
23134	0	1	1	1	1	50	20	oui	oui	1	4
31435	0	1	0	0	0	30	12	non	oui	0	null
43342	1	0	0	1	0	37	9	non	non	1	5
43241	1	0	0	1	0	53	15	non	oui	null	4

- Avec :
  - DN : date de naissance
  - Rev : revenus
  - Prop : Propriétaire
  - Voit : possède une voiture
  - PP : Paris ou province
  - DA : date d'abonnement

# Etape de fouille de données

Etape :

- ▶ Au coeur même du processus ECD.
- ▶ Difficile à mettre en oeuvre.
- ▶ Coûteuse.
- ▶ Aux résultats devant être interprétés et relativisés.

## Approche traditionnelle

1. Regarder, explorer.
2. Etablir une hypothèse, un modèle.
3. Essayer de contredire ou de vérifier de modèle.

# Etape d'évaluation et de validation

## Deux modes de validation

- ▶ Par statistique et / ou.
- ▶ Par expertise.

# Typologie des méthodes de fouilles de données

## Typologie selon l'objectif

- ▶ **Classification** : examiner les caractéristiques d'un objet et lui attribuer une classe.  
e.g. diagnostic ou décision d'attribution de prêt à un client.
- ▶ **Prédiction** : prédire la valeur future d'un attribut en fonction d'autres attributs.  
e.g. prédire la qualité d'un client .
- ▶ **Association** : déterminer les attributs qui sont corrélés.  
e.g. analyse du panier de la ménagère
- ▶ **Segmentation** : former des groupes homogènes à l'intérieur d'une population.

# Typologie des méthodes de fouilles de données

## Typologie selon le type de modèle obtenu

- ▶ Modèles prédictifs.
  - ▶ Utilisent les données existantes et des résultats connus sur ces données pour développer des modèles capables de prédire les valeurs d'autres données.  
e.g. *Prédire les clients qui ne rembourseront pas leur crédit.*
  - ▶ Utilisés principalement en classification et prédiction.
- ▶ Modèles descriptifs.
  - ▶ Proposent des descriptions de données pour aider à la prise de décision.
  - ▶ Souvent en amont de la construction de modèles prédictifs.
  - ▶ Utilisés principalement en segmentation et association.

# Typologie des méthodes de fouilles de données

## Typologie selon le type d'apprentissage utilisé

- ▶ Apprentissage supervisé : fouille supervisée
  - ▶ Processus qui prend en entrée des exemples d'apprentissage contenant à la fois des données d'entrée et de sortie.
  - ▶ Les exemples d'apprentissage sont fournis avec leur classe.
  - ▶ But : classer correctement un nouvel exemple.
  - ▶ Utilisés principalement en classification et prédition.
- ▶ Apprentissage non supervisé : fouille non supervisée
  - ▶ Processus qui prend en entrée des exemples d'apprentissage contenant que des données d'entrée
  - ▶ Pas de notion de classe
  - ▶ But : regrouper les exemples en paquets (clusters) d'exemples similaires.
  - ▶ Utilisés principalement en segmentation et association.

# Classification

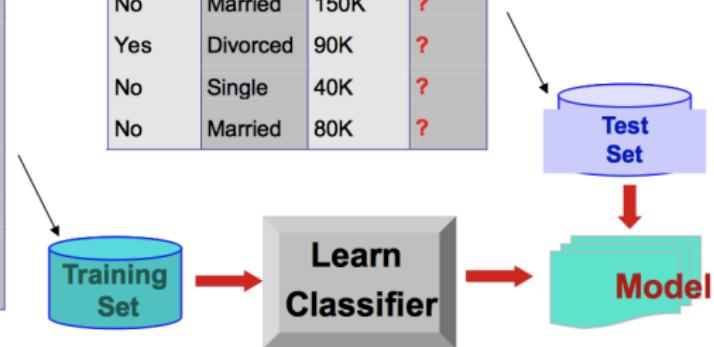
Examiner les caractéristiques d'un objet et lui attribuer une classe (un champ particulier à valeurs discrètes).

- ▶ Etant donnée une collection d'enregistrements (**ensemble d'apprentissage**).
  - ▶ Chaque enregistrement contient un ensemble d'attributs et un de ces attributs est sa classe.
- ▶ Trouver un modèle pour l'attribut classe comme une fonction de la valeurs des autres attributs
- ▶ But : permettre d'assigner une classe à des enregistrements inconnus de manière aussi précise que possible.
  - ▶ **Un ensemble de test** est utilisé pour déterminer la précision du modèle.

# Classification : exemple

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



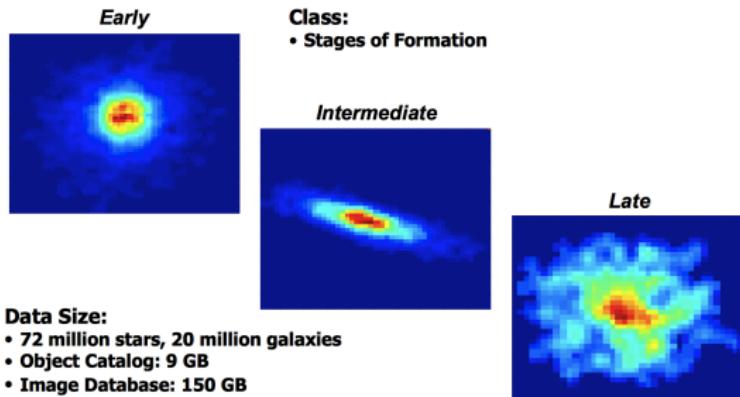
# Classification : exemples d'applications

## Marketing direct

- ▶ But : réduire le coût du mailing en ciblant un ensemble de consommateurs qui achèteront vraisemblablement un nouveau téléphone portable.
- ▶ Approche :
  - ▶ Utiliser des données pour un produit similaire.
  - ▶ On sait quels consommateurs ont acheté. La décision (Achat - Pas achat) est l'attribut classe.
  - ▶ Collecter diverses informations sur ce type de consommateurs.
  - ▶ Cette information représente les entrées du classifier.

# Classification : exemples d'applications

- ▶ Détection de fraudes à la carte bancaire à l'aide des transactions et d'informations sur le porteur du compte.
- ▶ Détection de désabonnement à l'aide des données sur d'autres consommateurs présents ou passés.
- ▶ Catalogage du ciel : classification des objets du ciel à l'aide d'images.



# Segmentation

Former des groupes homogènes à l'intérieur d'une population

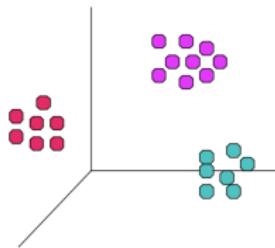
- ▶ Etant donné un ensemble de points, chacun ayant un ensemble d'attributs, et une mesure de similarité définie sur eux, trouver des groupes tels que :
  - ▶ Les points à l'intérieur d'un même groupe sont très similaires entre eux.
  - ▶ Les points appartenant à des groupes différents sont très dissimilaires.
- ▶ Le choix de la mesure de similarité est important.

# Segmentation : illustration

☒ Euclidean Distance Based Clustering in 3-D space.

Intracluster distances  
are minimized

Intercluster distances  
are maximized



# Segmentation : exemples d'applications

- ▶ Segmentation de marchés .
- ▶ Segmentation de documents.
- ▶ ...

# Association

## Entrée : Un ensemble de tickets de caisse

- ▶ Une observation = un caddie, un ticket de caisse.
- ▶ Non prise en compte de la fréquence des produits.
- ▶ Un grand nombre de produits, un grand nombre de caddies (petit sous ensemble de l'ensemble de produits).

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

## Sortie : Des règles

### Rules Discovered:

{Milk}  $\rightarrow$  {Coke}

{Diaper, Milk}  $\rightarrow$  {Beer}

# Association : exemples d'application

- ▶ Marketing et promotions sur des produits.
- ▶ Gestion du supermarchés : rayonnage.
- ▶ Inventaire.
- ▶ ...

# Résumé



masse de données (corpus)



- ▶ connaissances
- ▶ informations
- ▶ prédictions

# Logiciels d'AFD

## Logiciels commerciaux

- ▶ Suites logicielles SAS  
(<http://www.sas.com/offices/europe/france/>)
- ▶ SPSS d'IBM (<http://www-01.ibm.com/software/fr/analytics/spss/>)
- ▶ Solution Analytics de SAP (<http://www.sap.com/pc/analytics/strategy.html>),  
KXEN
- ▶ ...

# Logiciels de data mining

## Logiciels gratuits

- ▶  **Weka** : <http://www.cs.waikato.ac.nz/ml/weka/>
  - ▶ Ensemble de classes et d'algorithmes JAVA développés par l'Université de Waikato en Nouvelle Zelande.
  - ▶ Principaux algorithmes de data mining.
  - ▶ Utilisable en ligne de commande, à l'aide d'une interface utilisateur, par l'API.
- ▶ **TANAGRA** : <http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html>
  - ▶ Destiné à l'enseignement et à la recherche. Développé à l'Université de Lyon 2
  - ▶ Principaux algorithmes de data mining.
- ▶ **ORANGE** : <http://orange.biolab.si/>
- ▶ ...