Samir Loudni (samir.loudni@imt-atlantique.fr)
TASC – DAPI, IMT Atlantique

Chap I -
Introduction à la Fouille de Données

Plan

- Data Science – Définition
- Fouille de données – Définition
- Fouille de données – Exemples d’applications
- Processus KDD – Définition
- Fouille de données – Panorama des méthodes
- Logiciels de fouille de données

Data science – Une nouvelle discipline ?

(sources : Wikipedia)

- La science des données est un domaine interdisciplinaire qui utilise des méthodes, des processus, des algorithmes et des systèmes scientifiques pour extraire des connaissances et des idées de nombreuses données structurelles et non structurées.

- Elle utilise des techniques et des théories tirées de nombreux domaines dans le contexte des mathématiques, des statistiques, de l'informatique, de la théorie et des technologies de l'information.

- Parmi ces techniques et modèles, on retrouve les modèles probabilistes, l'apprentissage automatique, l'apprentissage statistique, la programmation informatique, l'ingénierie de données, la visualisation de données, la modélisation d'incertitude, le stockage de données, la compression de données et le calcul à haute performance.

Data science – A la croisée de trois disciplines

Connaître et maîtriser les techniques de modélisation d'analyse de données ..., savoir exploiter les régularités « cachées » dans les données

STATISTIQUES

INFORMATIQUE

CONNAISSANCES METIER

Maîtriser les outils pour manipuler les données, développer de nouveaux modèles pour gérer la profusion de l'information

Toute analyse s'inscrit dans un domaine qu'il faut connaître pour savoir interpréter et exploiter les résultats.

Data science – Pourquoi une telle effervescence?

1) Nous sommes à l'heure des « data » ... qui arrivent de partout et que l'on sait collecter et conserver
2) Prise de conscience collective... surtout des entreprises... de la valeur ajoutée que l'on peut en tirer
3) Indéniablement, il y a un effet de mode. Les éditeurs de solutions informatiques n'y sont pas étrangers.

Statistique / Analyse de données → Data Mining → Data Science / Big Data Analytics

La progression s'accompagne d'une évolution des techniques/technologies et des sources d'information.

Des statistiques ...

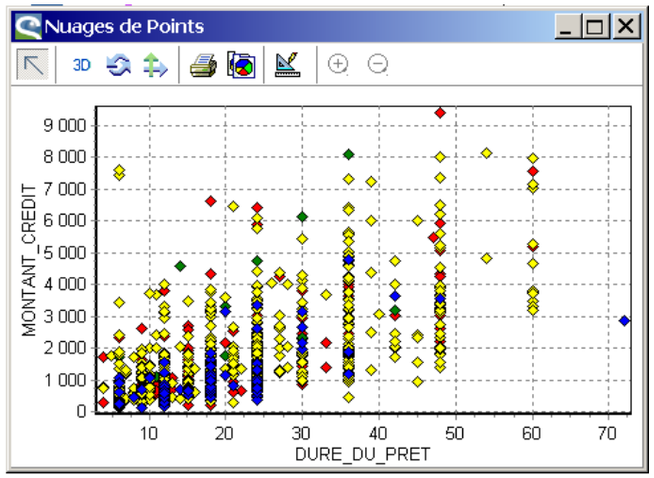
- **L’approche classique de la statistique :**
  - basée sur le paradigme des tests d’hypothèses
  - fortes hypothèses sur les lois statistiques suivies
  - les modèles sont issus de la théorie et confrontés aux données
  - quelques centaines d’individus et quelques variables recueillies (échantillonnage, etc)
  - processus trop coûteux : une hypothèse est proposée, une expérience est faite pour collecter des données, et ensuite les données sont analysées par rapport à cette hypothèse.

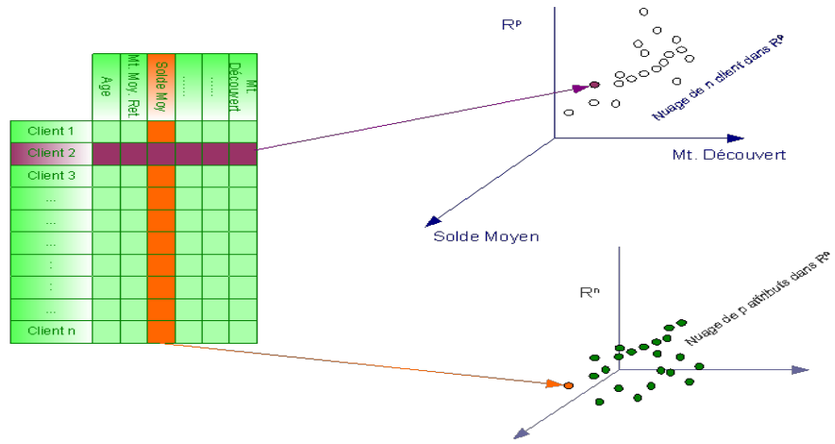
- **Analyse de données :**
  - quelques dizaines de variables
  - construction des tableaux « Individus x Variables »
  - importance du calcul et de la représentation visuelle

L’analyse de données – Traitement statistique des données

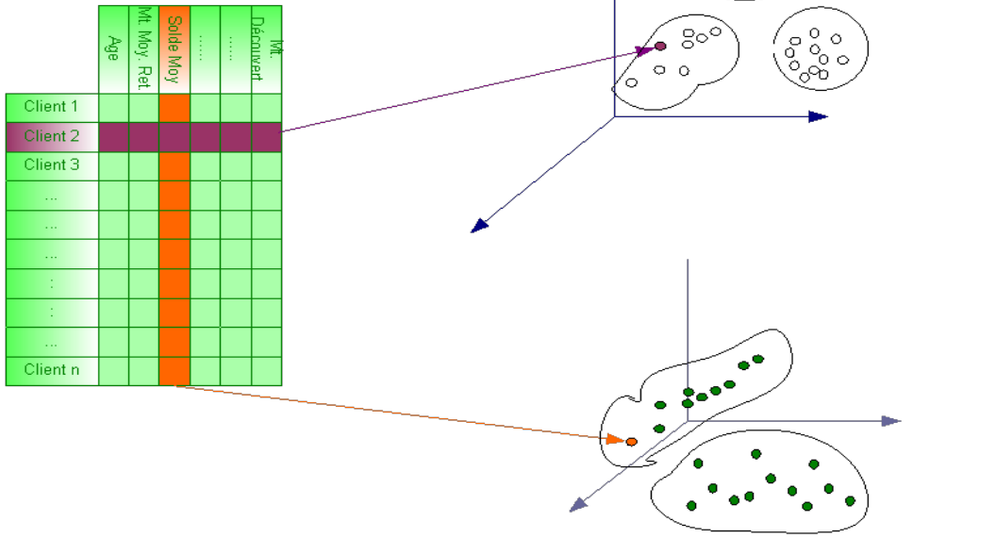
Application des techniques de modélisation et de statistique

Un ensemble limité de caractéristiques mesurées sur un ensemble restreint d’individus.





Données généralement recueillies à des fin d’étude (enquêtes, etc)



Au data Mining

- Data Mining :
  - plusieurs millions d’individus et plusieurs centaines de variables
  - nombreuses variables non numériques, parfois textuelles
  - données recueillies avant l’étude, et souvent à d’autres fins
  - données imparfaites, avec des erreurs de saisie, de codification, des valeurs manquantes, aberrantes
  - population constamment évolutive (difficulté d’échantillonner)
  - nécessité de calculs rapides, parfois en temps réel
  - on ne recherche pas toujours l’optimum mathématique, mais le modèle le plus facile à appréhender par des utilisateurs non-statisticiens
  - les modèles sont issus des données et on en tire des éléments théoriques
  - méthodes statistiques, d’intelligence artificielle et de théorie de l’apprentissage (« machine learning »)

Fouille de données – DATA MINING

- Le data mining fait passer de **l'analyse confirmatoire** à **l'analyse exploratoire**

- La métaphore «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** (classification, visualisation avec des méthodes comme l'ACP, etc.).

- L'émergence du data mining dues à la conjonction des plusieurs facteurs :
  - l'accroissement exponentiel, dans les entreprises, de données liées à leur activité (données sur la clientèle, les stocks, la fabrication, la comptabilité, la gestion, les ressources humaines, etc.)
  - l'évolution des SGBD vers l'informatique décisionnelle avec les entrepôts de données (Data Warehouse).
  - les progrès fulgurant des matériels et logiciels informatiques en capacité de stockage et d'analyse pour un coût de plus en plus faible.

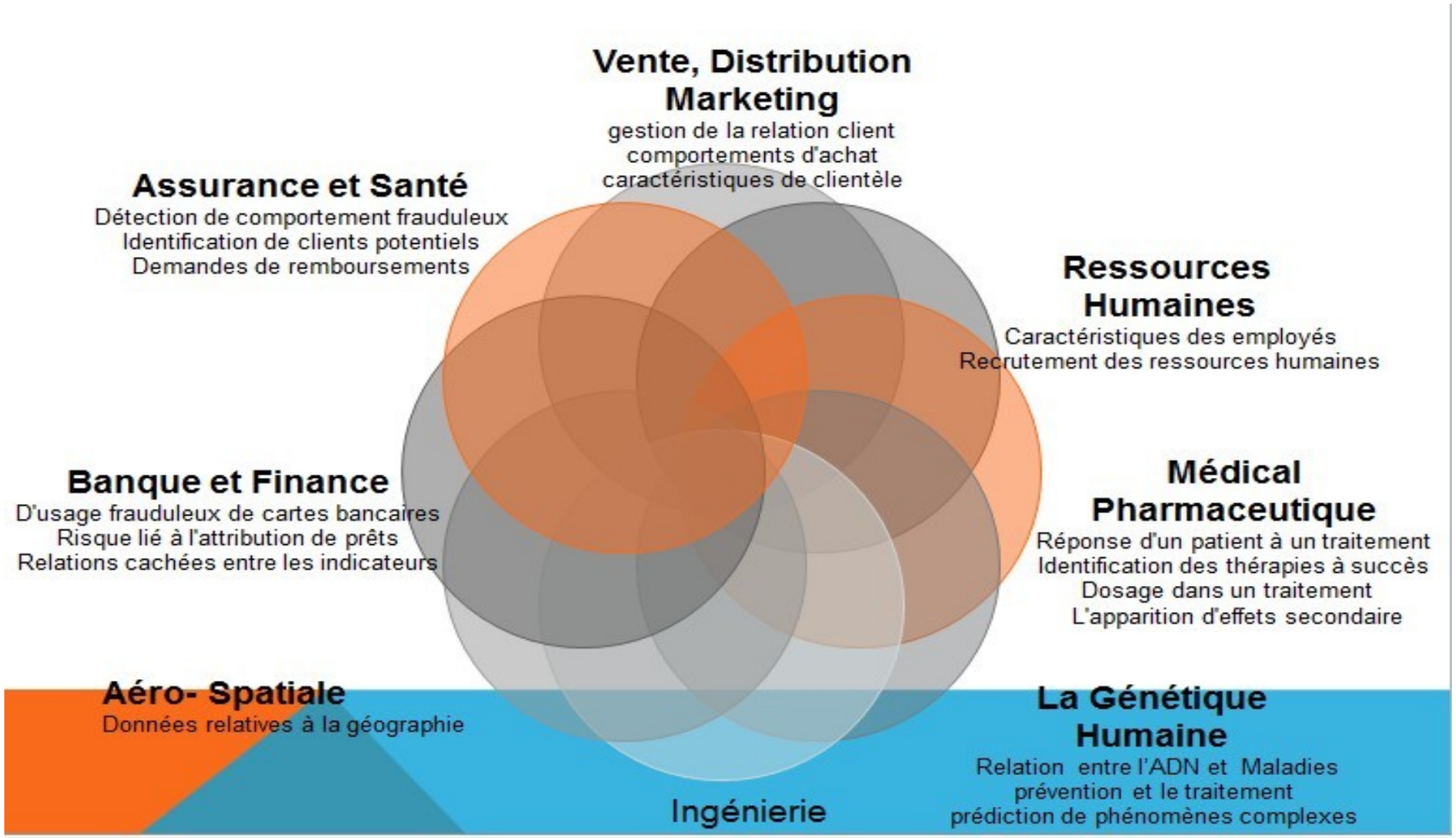
La fouille de données : qu'est-ce que c'est ?

- Processus inductif, itératif et interactif de découverte dans les BD larges de modèles de données valides , nouveaux , utiles et compréhensibles.
  - Itératif : nécessite plusieurs passes
  - Interactif : l'utilisateur est dans la boucle du processus
  - Valides : valables dans le futur
  - Nouveaux : non prévisibles
  - Utiles : permettent à l'utilisateur de prendre des décisions
  - Compréhensibles : présentation simple

- **Induction** : généralisation d'une observation ou d'un raisonnement établis à partir de cas singuliers.
  - Utilisée en Data mining pour tirer une conclusion à partir d'une série de faits, pas sûre à 100%
  - La clio a 4 roues, La Peugeot 106 a 4 roues, La BMW M3 a 4 roues, La Mercedes 190 a 4 roues => Toutes les voitures ont 4 roues

Domaines d’application

(sources : Wikipedia)



Exemple 1 – Vente, distribution / Marketing

- On ne veut plus seulement savoir :

  « Combien de clients ont acheté tel produit pendant telle période ? »

- Mais :
  - « Quel est leur profil ? »
  - « Quels autres produits les intéresseront ? »
  - « Quand seront-ils intéressés ? »

Exemple 1 – Vente, distribution / Marketing

- La gestion de la relation client consiste en l’ensemble des activités visant à cibler, attirer et conserver les "bons" clients.

- Profils des consommateurs, modèle d’achat, effet des périodes de solde.

- Analyse du ticket de caisse dans les grandes surfaces pour déterminer les produits souvent achetés simultanément, et agencer les rayons et organiser les promotions en conséquence

- Découverte de caractéristiques de clientèle.

- Prédiction de probabilité de réponse aux campagnes de mailing

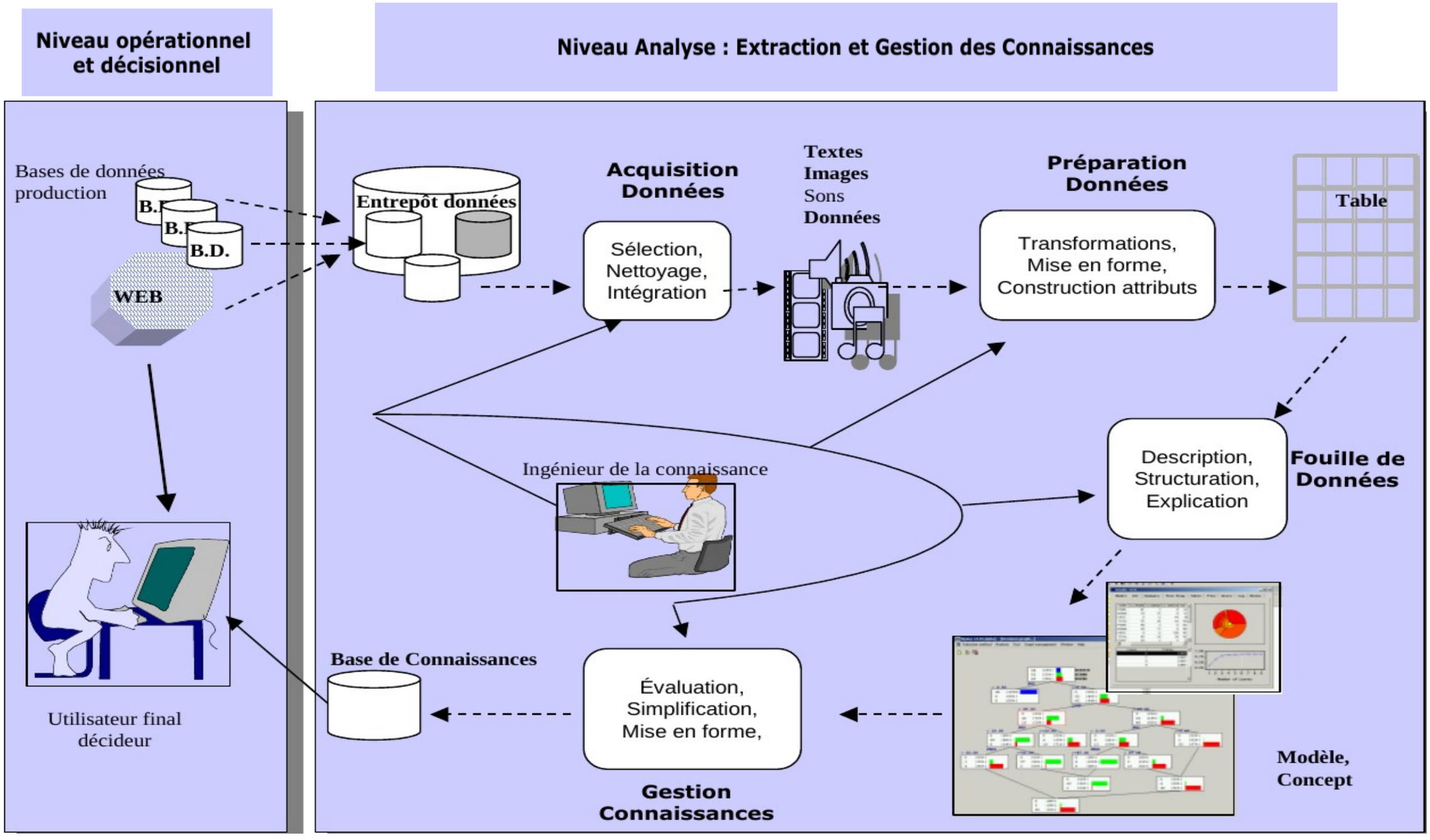
Exemple 2 – Médical / Pharmaceutique

- Pronostic des infarctus et des cancers (décès, survie)
- Décryptage de génome
- Étude des corrélations entre le dosage dans un traitement et l'apparition d'effets secondaire.
- Diagnostic : découvrir d’après les symptomes du patient sa maladie.
- Gestion de parcours de soins d’un patient.
- Élaboration de nouveaux médicaments, etc.

Exemple 3 – Banques / Finances

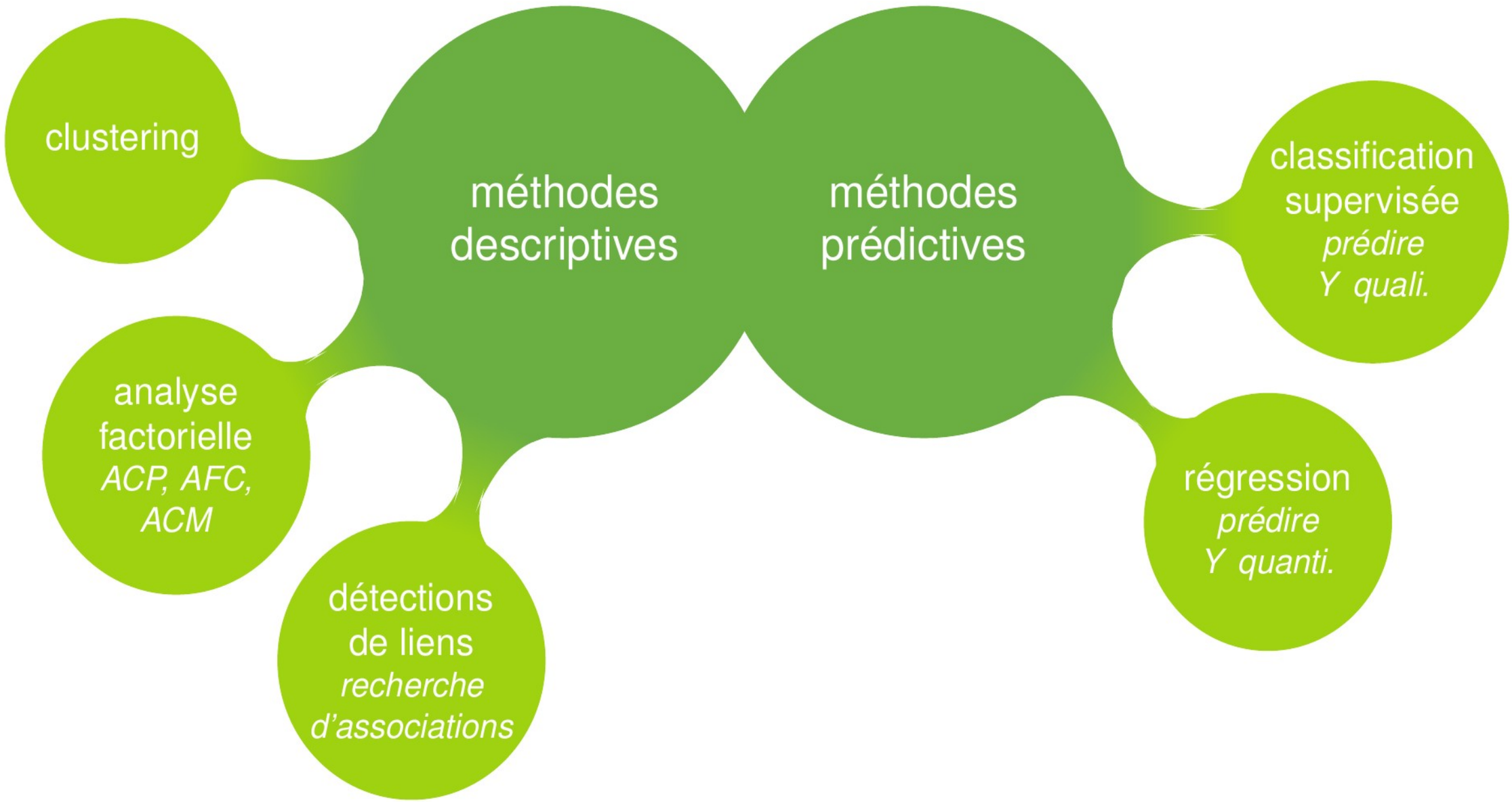
- Détection d’usage frauduleux de cartes bancaires.
- Gestion du risque lié à l'attribution de prêts bancaires par l‘utilisation du score de risque.
- Découverte de relations cachées entre les indicateurs financiers.
- Détection de règles de comportement boursier par l'analyse des données du marché.
- Détection de comportement frauduleux.

Processus KDD – Knowledge discovery in Databases



Sources : D. ZIGHED & R. RAKOTOMALALA, Techniques de l'Ingénieur, 2002.

Fouille de données – panorama des méthodes



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

- **Les méthodes descriptives** (ou exploratoires) visent à mettre en évidence des informations présentes, mais cachées par le volume des données (c'est le cas des classifications automatiques d'individus et des recherches d'associations de produits ou de médicaments) ;

- **Les méthodes prédictives (ou explicatives)** visent à extrapoler de nouvelles informations à partir des informations présentes, ces nouvelles informations pouvant être qualitatives (classement, discrimination, etc).

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 prédiction.
  - 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).
  - Étant 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 des 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 : exemples d’applications

Marketing direct

- But : réduire le coût du mailing en ciblant un ensemble de consommateurs qui achèteront vraisemblablement un nouveau téléphone portable

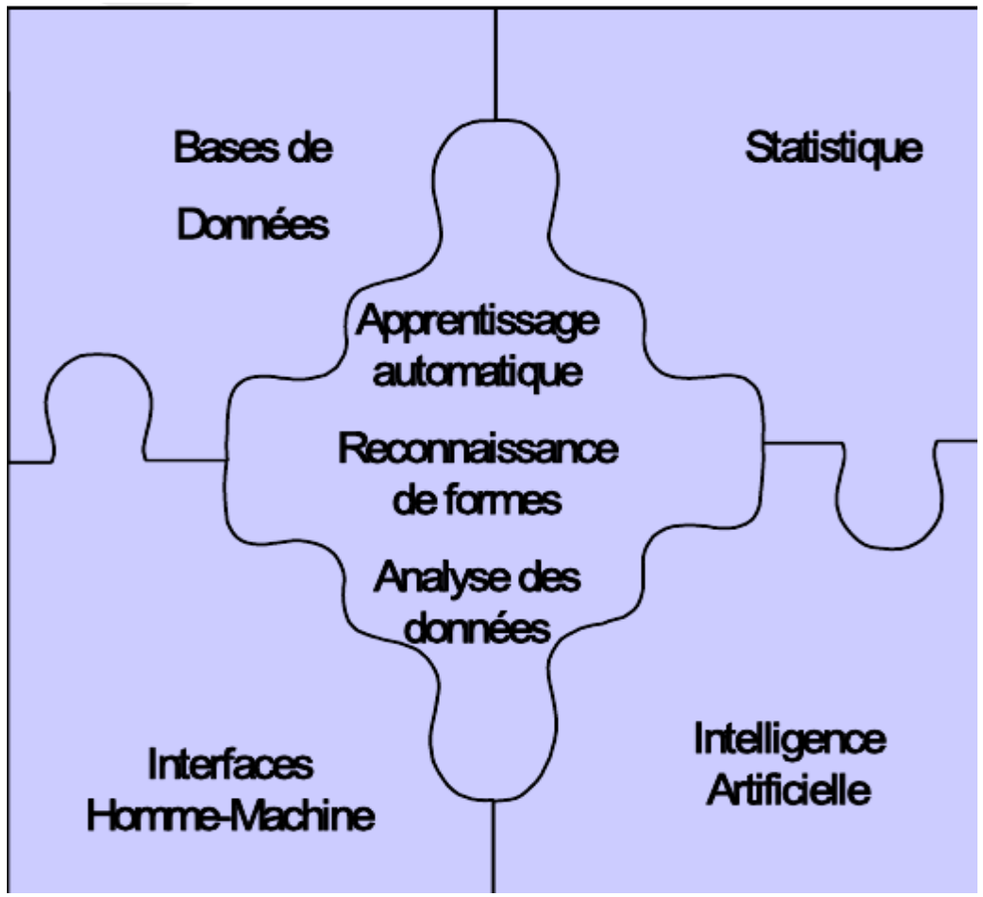
- Approche :
  - Utiliser des données pour un produit similaire.
  - On sait quels consommateurs ont acheté. La décision (Achat - Pas achat) est l’attribut classe.
  - Collecter diverses informations sur ce type de consommateurs.

- Cette information représente les entrées du classifier.

Segmentation

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

FD : La rencontre de plusieurs disciplines



Sources : D. ZIGHED & R. RAKOTOMALALA, Techniques de l'Ingénieur, 2002.

- Le data mining trouve ses principaux composants au sein de deux disciplines :

  (1) les **mathématiques appliquées,** avec l'analyse statistique des données,

  (2) **l'informatique** et **l'intelligence artificielle**, avec l'apprentissage numérique et symbolique et les bases 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.
- Attribut – valeur : la valeur d'un attribut est un nombre ou un symbole
  - **Quantitative** (numérique, exprime une quantité)
    - Discrète (ex : nombre d'étudiants dans un cours) ou continue (ex :longueur)
    - Échelle proportionnelle (chiffre d'affaires, taille), ou échelle d'intervalle (QI)
  - **Qualitative**
    - Variable ordinale (classement à un concours, échelle de satisfaction client)
    - Variable nominale (couleur de yeux, diplôme obtenu, sexe)

Logiciels de fouille de données

- Logiciels gratuits
  - Rstudio
  - 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.
  - WEKA : http://www.cs.waikato.ac.nz/ml/weka/
  - Bibliothèque Scikit-Learn de Pyhton
  - Knime : https://www.knime.org/

Références

- Bernard ESPINASSE (AMU) – Introduction à la Fouille de Données
- http://cedric.cnam.fr/~saporta
- http://eric.univ-lyon2.fr/~ricco/data-mining/
- http://data.mining.free.fr