# Fouille de Données et Apprentissage

## C1- Introduction générale

Lina Soualmia

Université de Rouen
LITIS - Équipe TIBS-CISMeF
lina.soualmia@chu-rouen.fr

18 janvier 2016

# Organisation des séances

- Module : Fouille de données (5 séances ; 9 h cours ; 9 h TP) et Apprentissage (9 h cours ; 9 h TP)
- Cours (lundi) :
  - Fouille : L. Soualmia du 18/01 au 29/02 puis Apprentissage : J. Grosjean
- TPs (mardi) :
  - Fouille : L. Soualmia du 19/01 au 01/03 puis Apprentissage : J. Grosjean
  - Mise en œuvre des concepts vus en cours
  - Logiciel : Weka (java)
- Évaluations :
  - Pratique 40% (TPs + un mini projet)
  - Écrit (examen) 60%

# Contenu pédagogique

- Fouille de données : (L. Soualmia)
  - extraction de connaissances à partir de bases de données
  - découverte et extraction de motifs
  - extraction de règles d'associations (Algorithmes)
  - classification (introduction)
- Apprentissage automatique : (J.Grosjean)
  - classification
  - clustering
  - SVM (Support Vector Machines)
  - Naïve Bayes
  - knn
- ...

# Crédits

- P.Poncelet (Montpellier)
- S.Ullman (Stanford)
- R.Rakotomalala (Lyon)
- N.Pasquier (Nice)
- parmi d'autres ...

# Plan

- Pourquoi fouiller les données ?
- Le processus d'extraction
- Quelques domaines d'application
- Un aperçu de quelques techniques

# Pourquoi fouiller les données ?

- De nombreuses données sont collectées et entreposées
  - Données du Web, e-commerce
  - Achats dans les supermarchés
  - Transactions cartes bancaires
- Les puces à ADN génèrent des expressions de gènes
- Les techniques traditionnelles ne sont pas adaptées
- Volume de données trop grands (trop de tuples, trop d'attributs)
  - Comment explorer des millions d'enregistrements avec des milliers d'attributs ?
- Requêtes traditionnelles (SQL) impossibles
  - Rechercher tous les enregistrements indiquant une fraude
- Explosion du volume des données : nécessité d'en tirer des connaissances utiles

# Définition

- La fouille de données (data-mining) est l'étape centrale du processus d'extraction de connaissances des bases de données (ECBD ou KDD pour Knowledge Discovery in Databases).
- Processus non trivial d'extraction de connaissances d'une base de données pour obtenir de nouvelles données, valides, potentiellement utiles, compréhensibles.
- Exploration et analyse, par des moyens automatiques ou semi-automatiques de grandes quantités de données en vue d'extraire des motifs intéressants.

# Objectifs

- Un processus non trivial d'extraction de modèles valides, nouveaux, potentiellement utiles et compréhensibles à partir de grands volumes de données
- Objectifs :
  - Compréhension des données et des phénomènes sous-jacents (liens, récurrences, etc.)
  - Extrapolation d'informations pour la prédiction d'événements
  - Construction de modèles (calculs) pour la prédiction de valeurs (données)

## Marketing :

- CRM (Customer Relationship Management), ventes croisées, Segmentation des marchés
  - Quels types de clients achètent quels types de produits ?
  - Y-a-t-il des liens de causalité entre l'achat d'un produit A et d'un autre produit B ?
  - Quel est le comportement des clients au cours du temps ?
- Utiliser des données recueillies pour un produit similaire
  - Chercher des associations/corrélations entre produits
  - Chercher des segments dans les données décrivant les clients

## Objectifs :

- profils de consommateurs et modèles d'achats
- constitution des rayonnages
- marketing ciblé

# Domaines d'application

**Assurances/Domaine bancaire**

- Analyse et gestion des risques : accord de crédit
- Détection de fraudes : cartes de crédit
  - Peut-on caractériser les assurés qui font des déclarations d'accident frauduleuses ?
  - Peut-on détecter un groupe de patients et un réseau de médecins qui ont des comportements anormaux ?
  - Quels sont les clients "à risque" pour l'accord de crédit ?

# Domaines d'application

- Santé, Médecine
  - Aide au diagnostic
  - Étude de l'influence de certaines médications sur l'évolution d'une maladie
  - Recherche des médicaments les plus efficaces
- Biologie
  - Identifier des similarités dans des séquences d'ADN
  - Identifier les fonctions des gènes
- Astronomie
  - Identifier le type d'un objet à partir d'images stellaires données en pixels

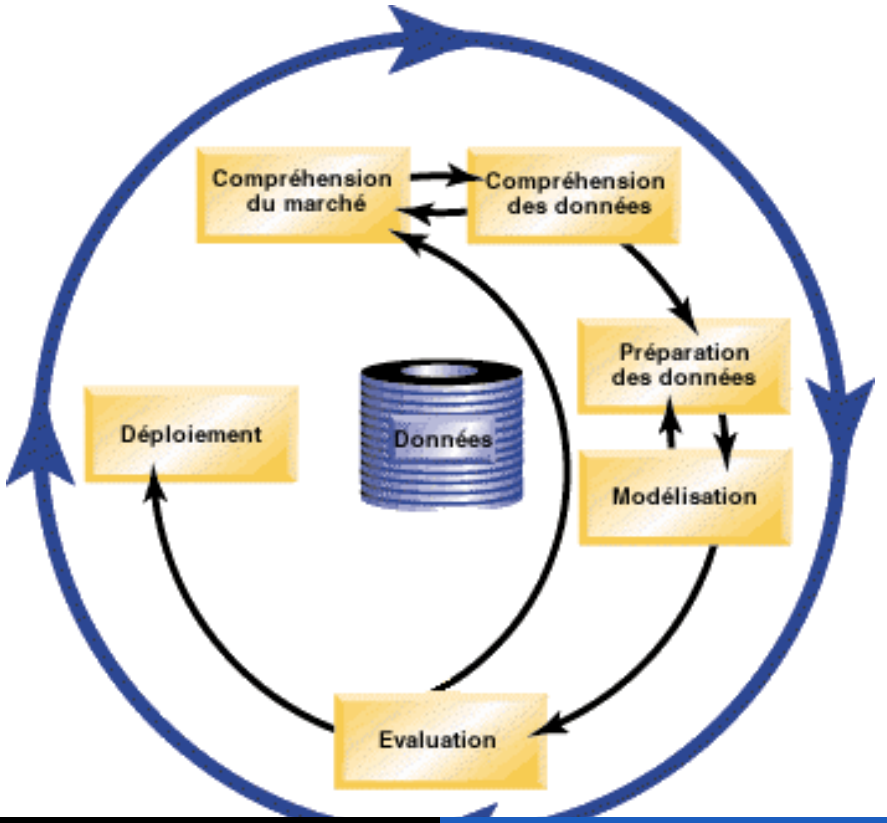
# Domaines d'application

- Web Mining
  - Etudier le contenu, la structure ou l'usage des pages web
- Text mining (news group, email, documents)
  - Routage de courrier, rapports d'activité, etc.
  - repose sur des données textuelles non structurées (vs. data mining qui repose sur des bases relationnelles ; et traite des données structurées)

# Méthode CRISP-DM

## Cross Industry Standard Process for Data Mining

- Modèle de cycle de vie à six phases
- Arcs : dépendances les plus importantes et fréquentes
- Séquence non stricte : dans la plupart des projets, on passe d'une phase à l'autre en fonction des besoins



# Processus

Étapes du processus :

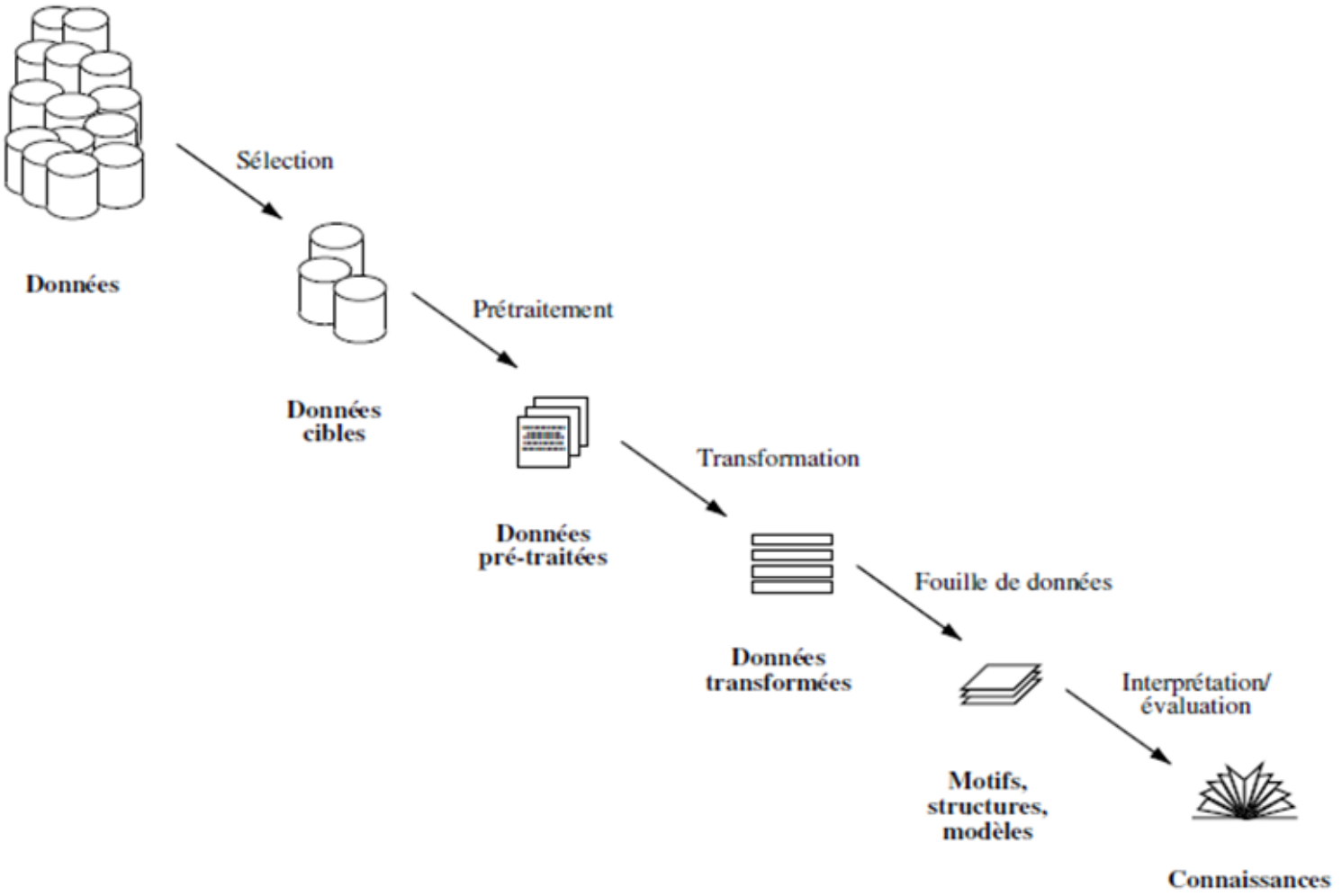
- Comprendre le problème
  - Connaissance du domaine, buts poursuivis, données disponibles, déploiement des résultats
- Explorer : visualiser, questionner
- Créer la table de données (jeu de données)
  - Nettoyage et Intégration (60% du travail)
  - Réduction et Transformation
- Choisir la ou les fonctionnalités
  - Description, classification, régression, clustering, extraction d'associations, séries chronologiques

# Processus

Étapes du processus

- Choisir la (les) méthodes (algorithmes)
- Effectuer l'extraction
  - Recherche des modèles intéressants
- Évaluation du modèle
- Présentation des résultats

## Processus interactif et itératif



# Base de données et base de connaissances

- Données :
  - Interprétées pour ce qu'elles sont (~faits)
- Connaissances :
  - Interprétées par des raisonnements
- Raisonnements déductifs :
  - S'appuient sur des connaissances
  - Souvent sous la forme de règles logiques
    - Socrate est un homme, tout homme est mortel donc Socrate est mortel
    - Homme $\Longrightarrow$ mortel

# Les connaissances

- Décrites dans une syntaxe particulière :
  - Un langage
- Ont une sémantique :
  - Leur donne un sens

**Exemples :**

- $[0; 3[$
  - Syntaxe : chaîne de caractères « [ 0 ; 3 [»
  - Sémantique : ensemble des réels $x$ tels que $0 \leq x < 3$
- $Chat(x)$ *et* $Dort(x) \Longrightarrow Ronronne(x)$
  - Syntaxe : la formule elle-même (le texte de la formule)
  - Sémantique : si $x$ est un chat qui dort, alors $x$ ronronne
  - $C \cap D \subseteq R$ ; avec $C$ l'ensemble des chats, $D$ les dormeurs, $R$ les ronronneurs

# Base de données/connaissances

**Différence BD et BC :**

Quand on interroge :

- Une base de données :
  - on n'obtient que des informations qui y sont déjà de façon explicite
- Une base de connaissances :
  - on peut obtenir des informations qui n'y sont pas explicitement

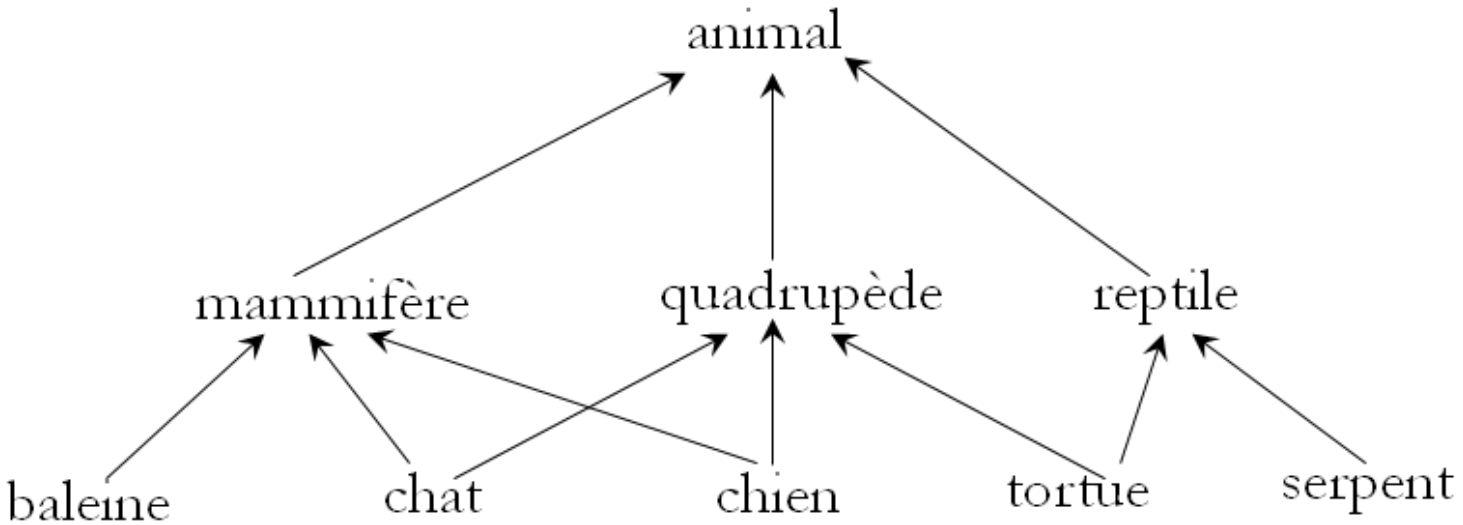
## Exemple de base de connaissances :

- Règles
  - père(x,y) $\Longrightarrow$ parent(x,y) ; mère(x,y) $\Longrightarrow$ parent(x,y)
  - père(x,y) $\wedge$ parent(y,z) $\Longrightarrow$ grand-père(x,z)
  - parent(x,y) $\Longrightarrow$ascendant(x,y)
  - parent(x,y) $\wedge$ ascendant(y,z) $\Longrightarrow$ascendant(x,z)
- Faits
  - père(Marcel,Maurice) ; mère(Henriette,Maurice) ; père(Maurice,Léon) ; mère(Cunégonde,Léon)
- Exemple de requêtes sur cette base de connaissances
  - grand-père(Marcel,Léon); grand-père(?,Juliette); ascendant(?,Léon); ascendant(Henriette,?); ascendant(?,?)

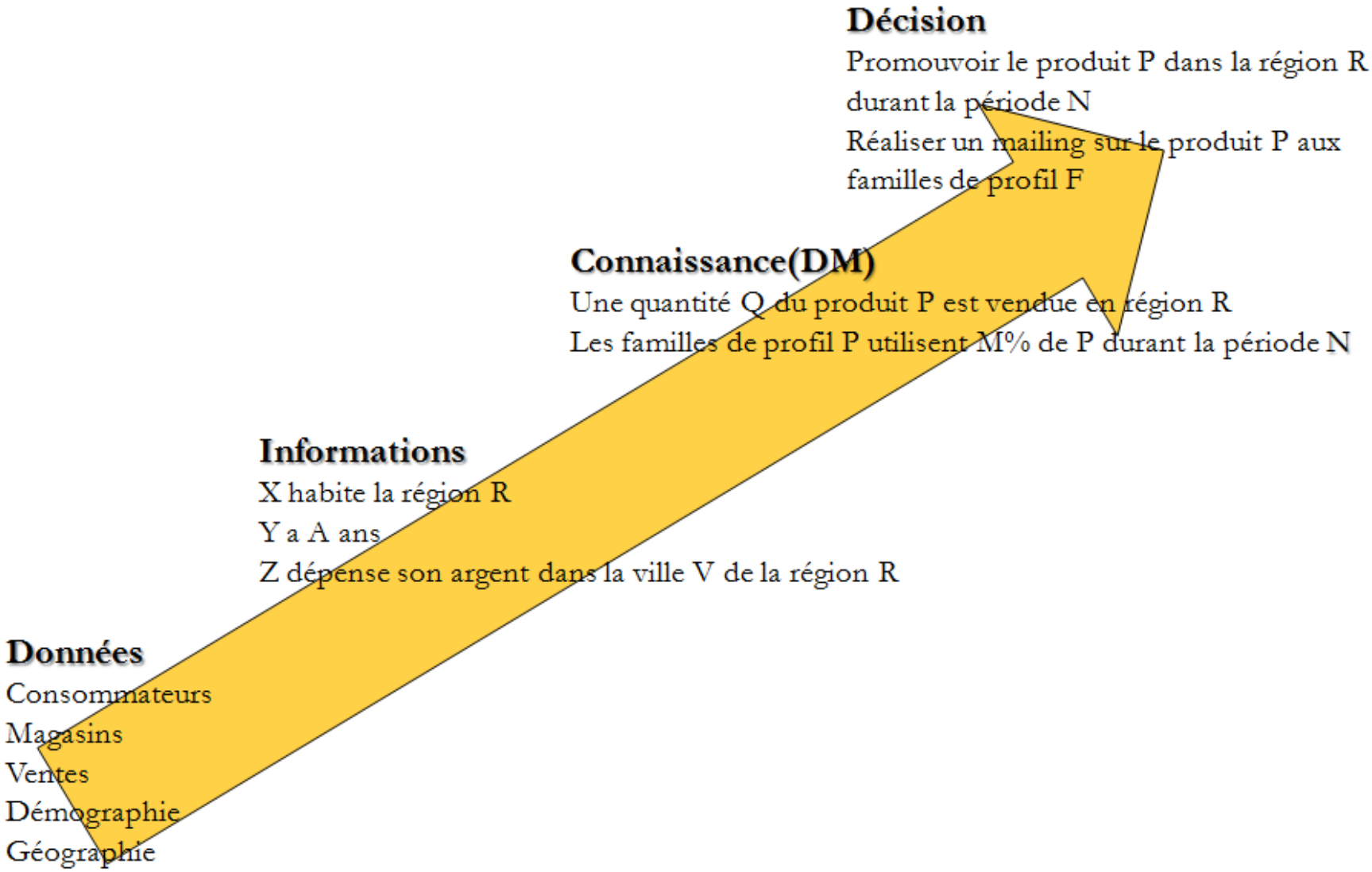
## Autre exemple de base de connaissances : les ontologies

- Une ontologie décrit un domaine par un ensemble de concepts de ce domaine et par les liens entre ces concepts.
- Parmi ces liens, le plus fréquemment utilisé est le lien de généralité entre concepts qui indique qu'un concept A est plus général qu'un concept B ce qui signifie que tous les individus dénotés par A le sont par B.
- Les concepts sont organisés en hiérarchie.

## Exemple simple : zoologie



# Processus


Décision
Promouvoir le produit P dans la région R durant la période N
Réaliser un mailing sur le produit P aux familles de profil F
Connaissance(DM)
Une quantité Q du produit P est vendue en région R
Les familles de profil P utilisent M% de P durant la période N
Informations
X habite la région R
Y a A ans
Z dépense son argent dans la ville V de la région R
Données
Consommateurs
Magasins
Ventes
Démographie
Géographie

## Data Mining ou non ?

- Rechercher le salaire d'un employé
  - NON
- Les supporters achètent de la bière le samedi et de l'aspirine le dimanche
  - OUI
- Interroger un moteur de recherche Web pour avoir des informations sur le Data Mining
  - NON
- Regrouper ensemble des documents retournés par un moteur de recherche en fonction de leur contenu ?

## Extraction de connaissances à partir de BD

- Des années 60 à fin 80 : Développement de l'informatique et des bases de données relationnelles, méthodes de modélisation, optimisation des requêtes
- Fin des années 80 : situation *data rich but information poor* : disposition de grandes sources de données inexploitées, besoin d'outils d'analyse puissants pour aider les décideurs : naissance de l'informatique décisionnelle
- Depuis : les besoins en extraction de connaissances continuent d'augmenter avec le Web, le séquençage du génome humain ...etc.

## Data Mining vs. Statistiques

- Data Mining :
  - intégration de techniques issues de plusieurs domaines : BD, Stats, Apprentissage automatique, analyse de données, visualisation ...
- Data Mining vs. Statistiques :
  - Découvrir plutôt que vérifier
- Data Mining vs. Machine Learning
  - Manipuler des BDs volumineuses plutôt que de petits ensembles d'apprentissage

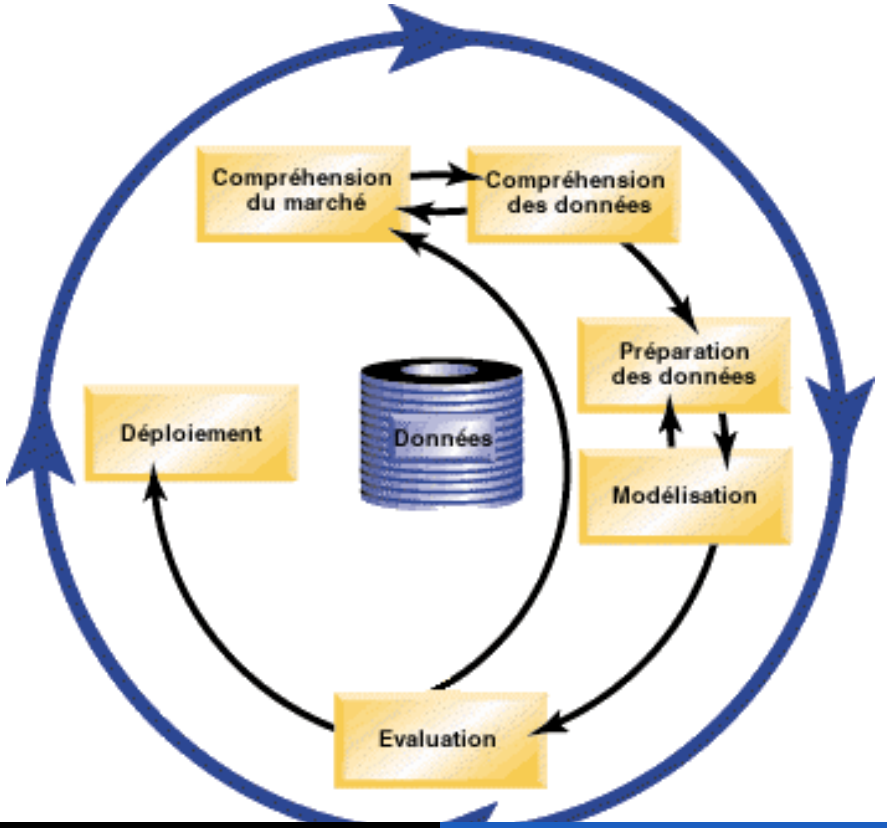
## Data Mining vs. Statistiques

La fouille de données se démarque des analyses statistiques par la puissance de calcul offerte :

- Fouille de données :
  - traitement de millions d'individus avec des centaines de variables (numériques et textuelles) ;
- Statistiques :
  - quelques centaines d'individus avec un nombre réduit de variables (numériques).
- Recueil des données :
  - Les données recueillies pour la statistiques l'ont été dans un cadre précis.
  - La fouille de données s'affranchit de cette contrainte (les systèmes effectuent un apprentissage des données qui leur sont fournies en entrée).

Cross Industry Standard Process for Data Mining

- Modèle de cycle de vie à six phases
- Arcs : dépendances les plus importantes et fréquentes
- Séquence non stricte : dans la plupart des projets, on passe d'une phase à l'autre en fonction des besoins



# Phase 1

- Adaptable : modèle aisément personnalisable
- Ces six phases couvrent la totalité du processus de Data Mining, y compris la façon de l'incorporer aux activités

***Compréhension du problème* : étape primordiale qui vise à**

- déterminer des objectifs (commerciaux ou autres),
- analyser la situation,
- déterminer les objectifs en termes de Data Mining,
- établir une planification du projet.

***Compréhension du problème* :**

- Identifier les différents objectifs à atteindre
- Déterminer les facteurs importants qui peuvent influencer l'aboutissement du projet
- Inventorier :
  - les ressources,
  - les contraintes,
  - les hypothèses.
- Déterminer les objectifs de la fouille de données en termes techniques
- Description exhaustive de toutes les étapes à venir

***Compréhension des données* : étape de sensibilisation à l'importance de bien maîtriser les ressources de données et leurs caractéristiques**

- Elle aborde :
  - la collecte de données initiale,
  - la description et l'exploration des données,
  - la vérification de la qualité de ces données.

# Phase 2

***Compréhension des données* :**

- Charger les données
  - Rapport sur la nature, la localisation, les méthodes de récupération, les problèmes éventuels
- Examen rapide et superficiel des données
  - Les données satisfont-elles les conditions requises ?
- Comprendre les données :
  - Utilisation de requêtes, outils de visualisation et de reporting.
  - Déterminer les attributs importants et leurs relations (redondantes)
  - Premiers résultats statistiques (graphiques, répartitions, etc.)
- Qualité des données
  - Données manquantes, erronées, incertaines ?

# Phase 3

***Préparation des données* : transformation des données à explorer afin d'assurer leur adéquation à la problématique et la pertinence des connaissances extraites**

- Phase de préparation comprend
  - la sélection,
  - le nettoyage,
  - la construction,
  - l'intégration,
  - le formatage des données

# Phase 3

***Préparation des données*** **:**

- Sélection des variables et instances
  - Selon l'objectif, la qualité des données, les contraintes techniques (catégorielles, numériques, etc.)
- Traitements des données bruitées
  - Correction des erreurs, estimation de valeurs
- Construction
  - Créer des valeurs dérivées, transformer des valeurs (normalisation, discrétisation)
- Combiner les données de diverses sources, supprimer les données redondantes
- Représentations adaptées, contraintes techniques

# Phase 4

***Modélisation* : élaboration des méthodes d'analyse qui seront utilisées pour extraire des connaissances à partir des données (cœur du processus)**

- Cette étape consiste à
  - sélectionner des techniques de modélisation,
  - générer des conceptions de test,
  - construire et évaluer des modèles.

# Phase 4

***Modélisation*** **:**

- Sélectionner les techniques (extraction d'associations, classification, clustering)
- Définir les mécanismes pour évaluer la qualité et la validité du modèle
- Définition de jeux d'apprentissage et de test (échantillonnage)
- Exécution des algorithmes
  - Documenter les résultats (données, paramètres, modèles)
- Classer les modèles par interêt
- Faire appel aux experts du domaine

# Phase 5

***Évaluation* : évaluer l'aide apportée par l'utilisation des modèles définis pour la concrétisation des objectifs poursuivis**

- Cette étape aborde
  - l'évaluation des résultats,
  - la vérification du processus de Data mining,
  - la prise de décision des étapes à suivre.

# Phase 5

***Évaluation*** **:**

- Évaluer l'adéquation des modèles aux objectifs métier
  - Rapport d'analyse des résultats selon les critères de succés
  - Approbation des modèles
- Vérification du processus
  - Vérifier qu'aucun facteur important n'a été oublié
  - Vérifier le respect des critères de qualité
  - Évaluer l'utilité potentielle des données non utilisées
  - Les données utilisées seront-elles toujours disponibles ?
- Décider des étapes suivantes
  - Le projet est-il prêt à être déployé ?
  - Nouvelles itérations du processus nécessaires ?

***Déploiement* : étape de rentabilisation des efforts**

- Objectif : intégrer les nouvelles connaissances aux processus quotidiens pour résoudre le problème initial / améliorer l'activité
- Elle comprend
  - le déploiement du plan,
  - la surveillance et la maintenance,
  - la production d'un rapport final,
  - la révision du projet.

# Phase 6

- Analyser les évaluations
  - Comment déployer les modèles dans l'organisation ?
  - Comment analyser les bénéfices du modèle ?
- Rapport final
  - Résumer le projet et l'expérience acquise
  - Présenter de façon compréhensible les résultats du data mining
- Bilan du processus
  - Analyser le processus et observer ce qui c'est bien ou mal déroulé
  - Les utilisateurs sont-ils pleinement satisfaits ? Ont-ils besoin de support ?

## Principaux logiciels

- Logiciels « stand-alone »
  - SPSS Clementine
  - Salfort Systems CART
  - Weka, RapidMiner, Orange, Tanagra (logiciels libres)
- Outils intégrés
  - IBM Intelligent Miner (IBM DB2)
  - Oracle Data Mining (Oracle 10g)
  - DB Miner (IBM BD2)
  - SQL Server (Microsoft)

## Interfaces logicielles

- Logiciels utilisant la méthode CRISP-DM
  - Représentations graphique des traitements
  - Graphes des flux de données et traitements enchaînés
  - Définition de « run-times » pour automatiser les traitements
  - Présentation des résultats sous forme de rapports
- Logiciels ou modules intégrés aux SGBD
  - Requêtes de Data Mining
  - Stockage relationnel des données
  - Langages SQL étendu, DMQL, opérateurs spécifiques
  - Modules de génération de rapports du SGBD
  - Interface avec les modules de data warehousing

## Types de données

- Origine des données
  - BD relationnelles
  - Data Warehouses : relationnel, cubes multi-dimensionnels
  - Données de transactions
  - BD orientées objets, spatiales, multimédia, textuelles
  - Données temporelles et séries temporelles
  - Données du Web
- Mais le plus souvent, pré-traitées et intégrées dans une table unique sur laquelle la recherche d'un modèle est réalisée.

# Extraction de connaissances à partir de BD

## Définition de l'EC-BD

C'est un processus d'extraction de connaissances :

- Nouvelles :
  - c'est- à dire pas déjà connues
  - On extraira des connaissances connues mais ce n'est pas l'objectif ...
- Potentiellement utiles : réutilisables dans un processus de raisonnement
- et ayant un degré de plausibilité : on cherche à contrôler la plausibilité des connaissances extraites
- dans de grands volumes de données :
  - Nécessitent des processus automatiques
  - Permettent une certaine validité statistique des connaissances extraites

Exemple

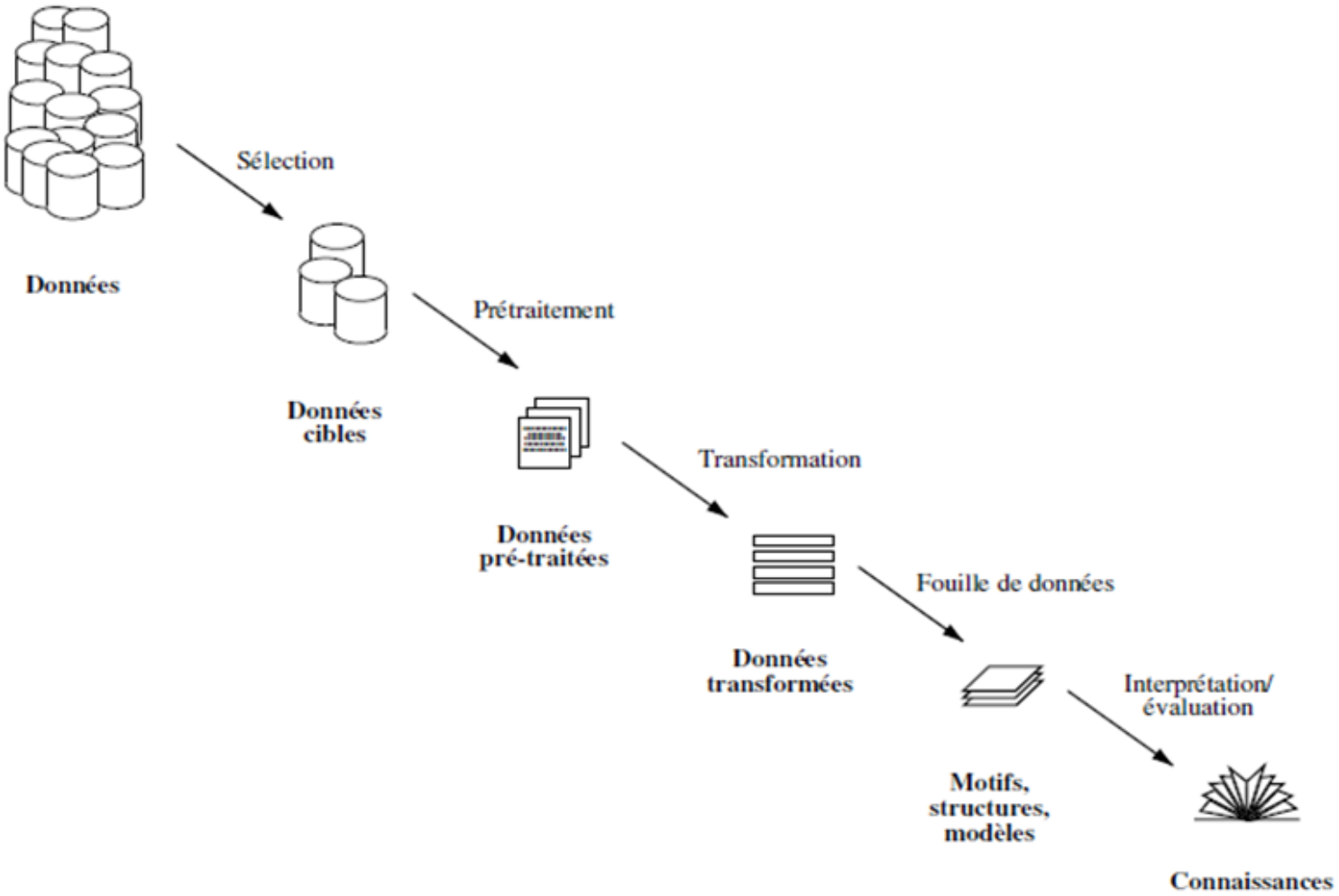
- Supposons qu'on extrait la règle suivante d'une base de données :
  - Blanc $\Longrightarrow$ né aux Etats Unis avec une probabilité de 99,07%
- Cette règle est totalement dépendante des données étudiées
- Elle donne des connaissances sur la base de données considérée
- C'est le même problème que celui de l'échantillonnage en statistique

## Restons simples

- Données binaires
- Recherche de motifs fréquents : « lait, oeufs et farine sont souvent achetés ensemble »
- Recherche de règles d'association : « dans 75 % des cas, l'achat de couches va de pair avec l'achat de bière »
- Pas de notion du temps
- Ni disjonctions, ni négations

## Principe

- Entrée : des données
- Sortie : des connaissances
- Les données ont des structures hétérogènes
  - On les trouve dans des bases de données relationnelles
  - Bases de données objet
  - Sous forme de liste de fichiers
  - Sur des pages Web
  - Dans des textes


Sélection
Données
Prétraitement
Données cibles
Transformation
Données pré-traitées
Fouille de données
Données transformées
Interprétation/ évaluation
Motifs, structures, modèles
Connaissances

# Étapes

- Intégration des sources de données et sélection des données intéressantes
- Transformation et formatage :
  - Données brutes $\rightarrow$ données préparées et formatées
  - Transformation : enlever les données bruitées, traiter les données manquantes
  - Formatage : sous forme traitable par les algorithmes de fouille
- Fouille de données :
  - données préparées et formatées $\rightarrow$ éléments d'information extraits
  - C'est l'étape algorithmique du processus d'ECBD
  - mise en œuvre de méthodes de fouilles
- Interprétation
  - Éléments d'information extraits $\rightarrow$ unités de connaissances

# Processus

- Itératif : se fait en plusieurs passes
- Interactif : l'analyste est dans la boucle et il oriente le processus selon ses connaissances du domaine des données et selon son intuition
- L'analyste est un expert du domaine des données, s'il connait les principes de l'informatique et de l'ECBD c'est un plus
- C'est l'analyste qui effectue l'interprétation
- Les éléments d'information extraits des processus de fouille doivent être sous une forme compréhensible par l'analyste (intelligibilité)
- Il est utile d'avoir des connaissances du domaine représentées et manipulables informatiquement pour assister l'analyste. Ces connaissances du domaine sont souvent sous la forme d'une ou plusieurs ontologies.
- Si une connaissance extraite n'est pas nouvelle et qu'elle est représentée en machine, inutile de demander une confirmation de l'analyste

# Objectifs : décrire ou prédire

- Caractérisation :
  - Identification de critères d'appartenance à une classe
- Discrimination :
  - Identification de critères discriminants entre l'appartenance à deux classes
- Extraction de règles d'association
  - Recherche de relations « attribut-valeur » qui occurrent ensemble fréquemment
- Classification
  - Construction d'un modèle ou d'une fonction à partir du jeu de données, pour permettre la prédiction de l'appartenance du nouvel objet à une classe
- Clustering
  - Regroupement d'objets de sorte à minimiser la distance entre objets similaires et à maximiser la distance entre objets différents

**Caractérisation–Discrimination**

- Requêtes SQL
- Requêtes OLAP (Online Analytical Processing)
- Description analytique
- Mesures statistiques

**Analyse d'association (corrélation et causalité)**

- Découvrir des règles d'association de la forme X $\rightarrow$Y où X et Y sont des conjonctions de termes attributs-valeurs ou prédicats
- Les mesures de support et confiance indiquent la portée et la précision de la règle
- Exemples :
  - Achat=pain et Achat=café $\rightarrow$ Achat=beurre (support = 5%, confiance = 70%)
  - Age>20 et Age<29 et Revenu>1000 $\rightarrow$ Achète_PC="oui" (support = 2%, confiance = 60%)

## Clustering

- Trouver des groupes ou classes d'objets tels que la similarité intra-classe est élevée et la similarité inter-classes est faible
- Pas de variable identifiant la classe
- Classification (apprentissage) non-supervisé : classes inconnues à l'avance
- Exemples :
  - segmentation des clients
  - cluster d'étoiles caractérisées par leur luminosité et leur température

## Classement et prédiction (classification supervisée)

- Apprendre un modèle qui associe un objet à une classe prédéfinie
- Apprendre une fonction permettant de prédire la valeur d'une variable numérique
- Exemples :
  - Classer des patients selon leur risque de développer une malade en fonction de leurs symptômes
  - Évaluer le risque d'un incident de remboursement en fonction des caractéristiques des demandeurs de crédit
- Forme du résultat :
  - Arbres de décision, règles de classification, réseaux de neurones, classifieurs Bayésiens ...etc.

# Fonctionnalités

## Analyse de déviations

- Déviation : objet qui n'est pas conforme au comportement général
- Donnée bruitée ou exception : information utile dans le cas de détection de fraudes ou d'évènements rares

## Recherche de corrélations

- Analyse statistique
- Recherche de motifs séquentiels
- Analyse de régression

## Méthodes de fouille de données

- Il existe plusieurs moyens de différencier les méthodes de fouille de données.
- Un moyen classique est de les distinguer selon leur objectif
- Les méthodes descriptives
  - Caractérisation, discrimination
  - Extraction de règles d'association
  - Analyse de cas extrêmes
- Les méthodes prédictives
  - Classification
  - Clustering

## Méthodes de fouille de données

- On peut les distinguer en fonction du type de données qu'elles traitent en entrée
- Méthodes symboliques
  - Extraction de motifs fréquents
  - Extraction de règles d'associations
  - Classification par treillis
  - Classification par arbres de décision
- Méthodes numériques issues de la reconnaissance des formes
  - Méthodes statistiques
  - Analyse de données (classification automatique, classification par composantes principales)
  - Modèles de Markov caché d'ordre 1 et 2 (très bons résultats en fouille de données en génomique)
  - Les méthodes neuronales, algorithmes génétiques
- Fouille de textes :
  - Fouille de base de données bibliographiques d'un domaine pour le décrire

# Limites

Deux points critiques au niveau des données :

- A cause du volume de données étudiées, il faut faire un passage à l'échelle pour passer :
  - Des algorithmes d'apprentissage standards qui ne sont applicables en pratique que sur des volumes de données relativement faibles
  - Aux algorithmes qui prennent en compte de très grands volumes de données
- Qualité des données :
  - Problèmes dus à l'absence de données (il en manque) et aux bruits (données erronées)
  - Nécessite d'avoir des méthodes pour améliorer cette qualité notamment par filtrage

# Préparation des données

- Données réelles imparfaites/endommagées
  - Incomplètes
  - Bruitées
  - Incohérentes
- Nécessité de préparer les données
  - Nettoyage
  - Intégration et transformation
  - Réduction
  - Discrétisation

# Types de données

- Numériques linéaires :
  - Ex : poids, taille, longitude, vitesse, etc.
- Binaires : une valeur parmi deux possibles
  - 0 : la variable est absente, 1 : la variable est présente
- Nominales : valeur prise dans une liste finie
  - Ex : couleur peut être « vert, bleu, rouge, jaune, noir »
- Ordinales : l'ordre des valeurs est plus important
  - Ex : résultat d'un concours
- Par ratios : variables numériques sur une échelle exponentielle/logarithmique
  - Valeurs non-linéaires

# Préparation des données

- Nettoyage
  - Compléter les valeurs manquantes, lisser les données bruitées, supprimer les déviations et corriger les incohérences
- Intégration
  - Intégrer des sources de données multiples
- Transformation
  - Normaliser
- Réduction
  - Réduire le volume des données (agréger, supprimer une dimension, etc.)
- Discrétisation
  - Pour les attributs numériques, permet de réduire le volume

# Nettoyage-Valeurs manquantes

- Les valeurs manquantes peuvent être codées par diverses valeurs :
  - <Vide>, « 0 », « . », « NA », « ? » , « NULL »
  - Il est nécessaire d'uniformiser le code
- Valeurs manquantes sont interdites
  - Ignorer le tuple
  - Compléter la valeur à la main
  - Utiliser une constate globale
  - Utiliser la valeur moyenne
  - Utiliser la valeur moyenne pour les exemples d'une même classe
  - Utiliser la valeur la plus probable

# Nettoyage - Données bruitées

On peut :

- Trier et partitionner (discrétiser)
- Classifier (exceptions)
- Appliquer un modèle de prédiction (ex : une fonction de régression)

# Partitionnement et lissage

- Les valeurs triées sont réparties en largeur (distance)
  - La suite triée est partitionnée en N intervalles de même amplitude
  - Amplitude de chaque intervalle $W = \frac{(max - min)}{N}$
  - Solution la plus simple mais les exceptions peuvent dominer
- Les valeurs triées sont réparties en profondeur (fréquence)
  - La suite triée est partitionnée en N intervalles contenant le même nombre de valeurs

# Partitionnement

Exemple :

- Données triées : 4 8 9 15 21 21 24 25 26 28 29 34
- Partition en profondeur :
  - Part 1 : 4 8 9 15
  - Part 2 : 21 21 24 25
  - Part 3 : 26 28 29 34

# Lissage

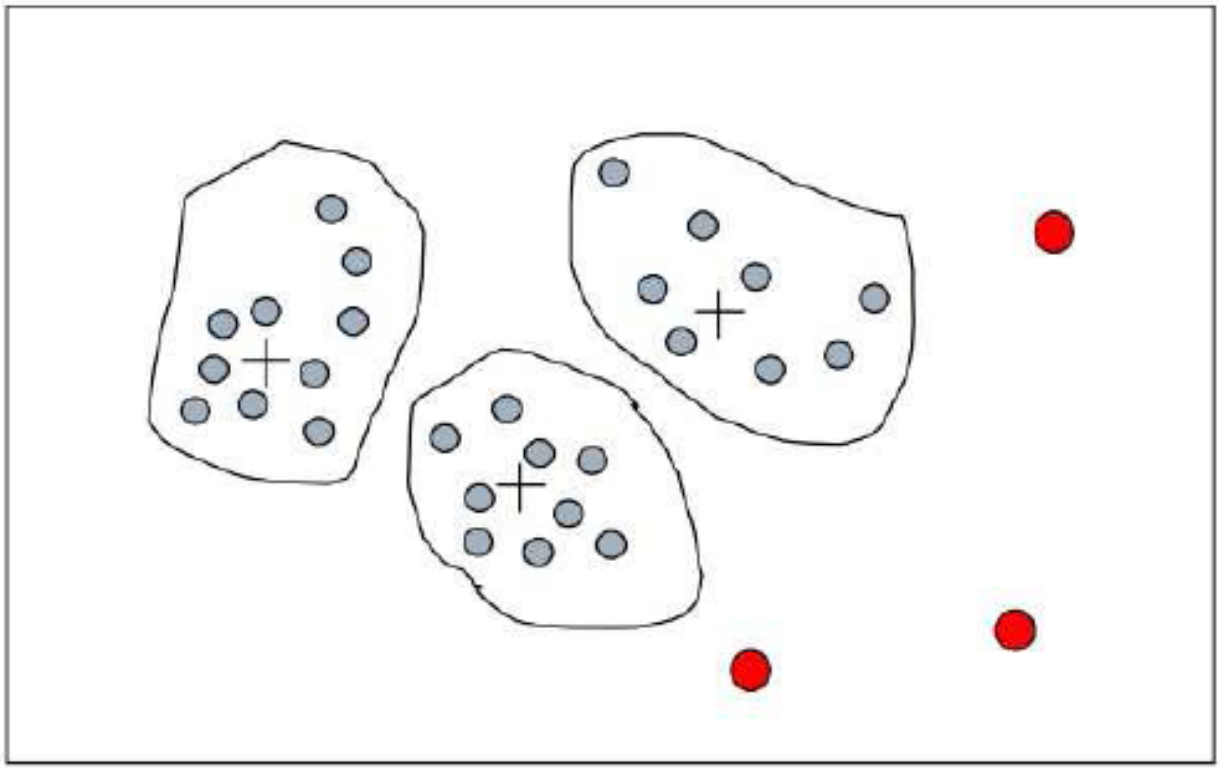
**Exemple :**

- Lissage par les moyennes : chaque valeur de la partition est remplacée par la moyenne
  - Part 1 : 9 9 9 9
  - Part 2 : 23 23 23 23
  - Part 3 : 29 29 29 29
- Lissage par les extrêmes : chaque valeur de la partition est remplacé par la valeur extrême la plus proche
  - Part 1 : 4 4 4 15
  - Part 2 : 21 21 25 25
  - Part 3 : 26 26 26 34

# Nettoyage - Supprimer les déviations

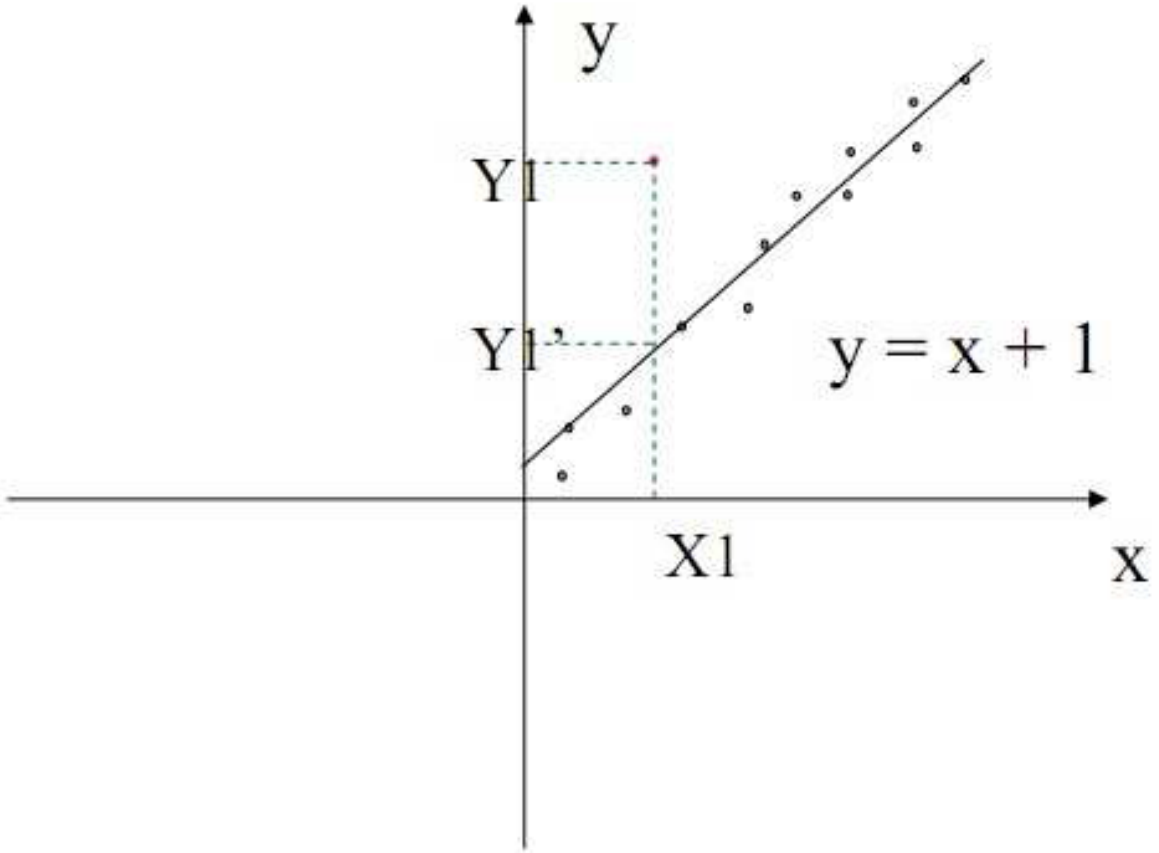
## Classifier

- Les valeurs similaires sont organisées en classes
- Les valeurs hors-classes sont considérées comme des déviations



# Lissage par régression

- Les données sont lissées de manière à approcher une fonction
  - Régression linéaire
  - Régression linéaire multiple



## Intégration

- Combinaison de données issues de différentes sources
- Intégration de schémas
  - Identifier les entités similaires
  - Ex : ID, Code, Matricule
- Détection et résolution de conflits
  - Résoudre les problèmes d'attributs symbolisant les mêmes entités avec des représentations différentes, des unités différentes
  - Ex : âge et date de naissance

**Redondances**

- Détection de données redondantes par analyse de corrélation
- Par ex. redondance entre attributs : Prix HT et TTC
- Mesure de la corrélation entre les attributs A et B

**Transformation**

Les transformations appliquées :

- Le lissage qui supprime les données bruitées
- L'agrégation qui calcule des sommes, moyennes . . . etc.
- La généralisation qui remonte dans une hiérarchie de concepts (taxonomie, ontologie)
- La normalisation qui ramène les valeurs dans un intervalle donné
- La construction d'attributs

## Réduction

- Permet d'obtenir une représentation réduite d'ensembles volumineux de données
- Stratégies appliquées
  - Agrégation
  - Réduction de dimensions
  - Compression
  - Discrétisation

## Réduction de dimensions

- Suppression d'attributs : la présence d'attributs non pertinents détériore les performances des algorithmes
  - Par exemple, les algorithmes d'induction d'arbres
- Pour assurer de bonnes performances aux algorithmes d'extraction :
  - Supprimer les données non pertinentes
    - Identifiants, attributs à valeur unique
  - Supprimer les données redondantes
    - Données déductibles d'autres données
    - Âge et année de naissance

**Réduction - Discrétisation**

- Permet de réduire le nombre de valeurs d'un attribut continu en divisant le domaine de valeurs en intervalles.
- Utile pour la classification, extraction d'associations.
- Des techniques de discrétisation peuvent être appliquées récursivement pour fournir un partitionnement hiérarchique de l'attribut.

# Exemple – jeu de Tennis

Attribut de classe : Play (yes/no)

- Yes : jeu possible ;
- No : jeu impossible.

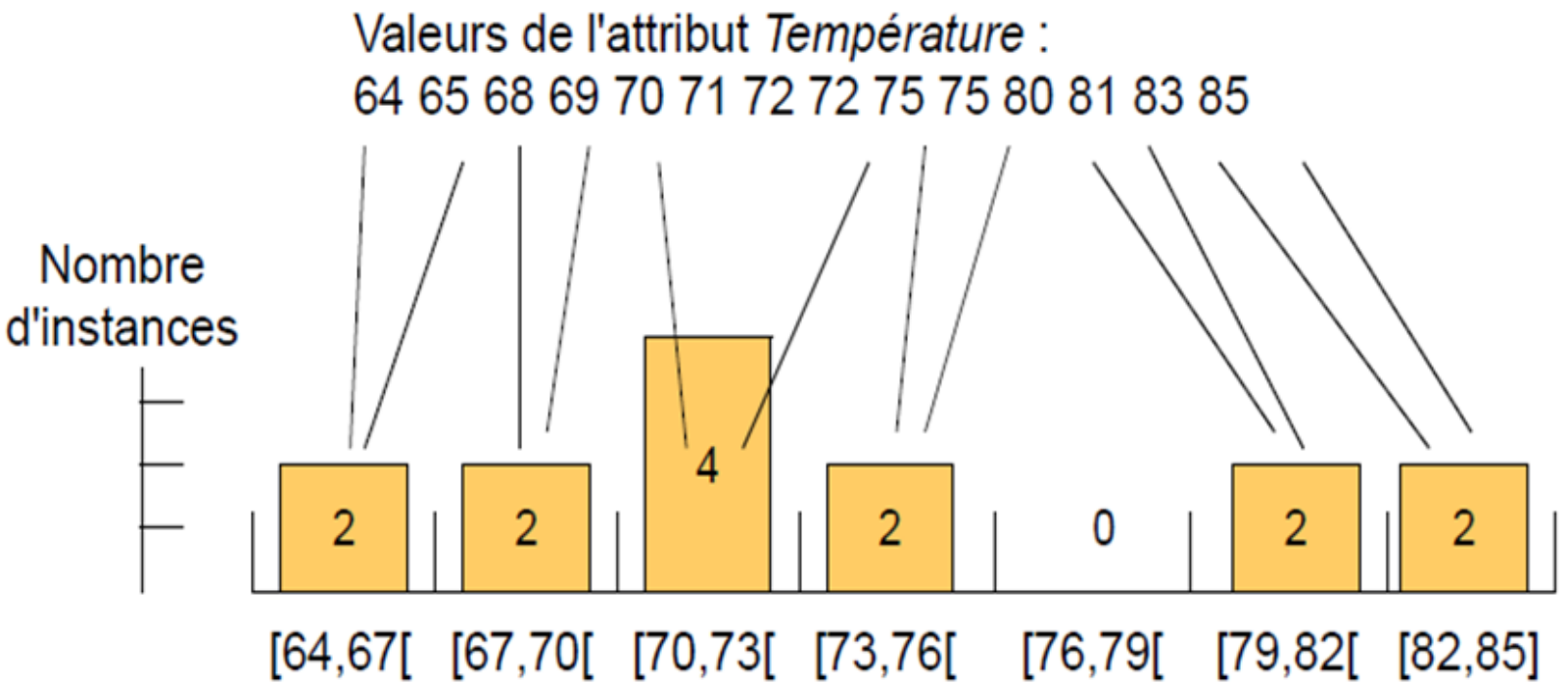
Autres attributs : météo

- Outlook : ciel (sunny, overcast, rainy)
- Température : degrés K
- Humidity : tx d'humidité
- Windy : présence de vent (Vrai/Faux)

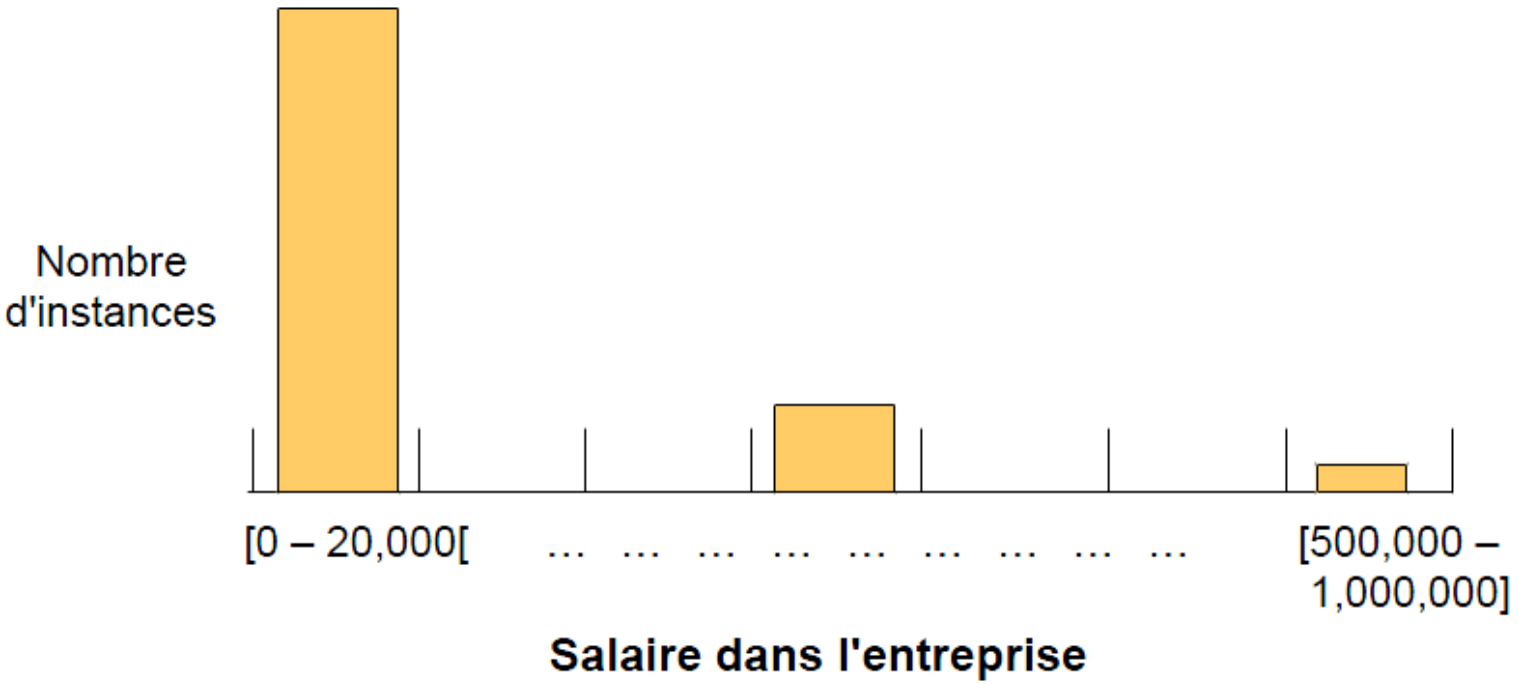
| outlook | temperature | humidity | windy | play |
|---|---|---|---|---|
| sunny | 85 | 85 | false | no |
| sunny | 80 | 90 | true | no |
| overcast | 83 | 86 | false | yes |
| rainy | 70 | 96 | false | yes |
| rainy | 68 | 80 | false | yes |
| rainy | 65 | 70 | true | no |
| overcast | 64 | 65 | true | yes |
| sunny | 72 | 95 | false | no |
| sunny | 69 | 70 | false | yes |
| rainy | 75 | 80 | false | yes |
| sunny | 75 | 70 | true | yes |
| overcast | 72 | 90 | true | yes |
| overcast | 81 | 75 | false | yes |
| rainy | 71 | 91 | true | no |

# Discrétisation en largeur

- Discrétisation par intervalles égaux des valeurs

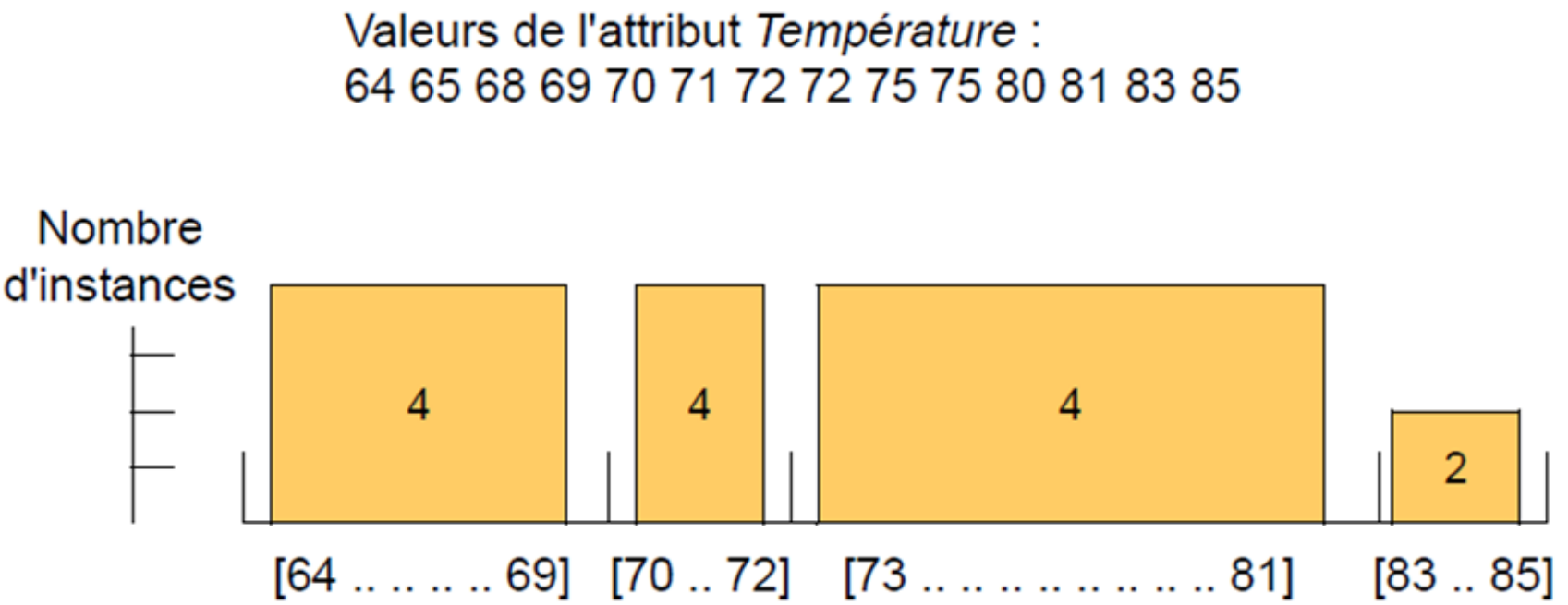


- Peut entraîner des inégalités importantes dans les effectifs



# Discrétisation en profondeur

- Discrétisation par effectifs égaux
- Tailles égales excepté pour le dernier intervalle



# Discrétisation : autres méthodes

- Présence de seuils significatifs
  - Ex : Age $>$ 18 ans
- Discrétisation supervisée
  - Prend en compte la classification

| temperature | 64 | 65 | 68 | 69 | 70 | 71 | 72 | 73 | 74 | 81 | 83 | 85 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| class | yes | no | yes<br>no | yes | yes<br>no | no | no<br>yes | yes<br>yes | no | yes | yes | no |

- Utilise l'entropie pour mesurer l'information et obtenir un critère de « pureté »

| temperature | 64 | 65 | 68 | 69 | 70 | 71 | 72 | 73 | 74 | 81 | 83 | 85 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| class | yes | no | yes<br>no | yes | yes<br>no | no | no<br>yes | yes<br>yes | no | yes | yes | no |

- Comptage des occurrences de yes et no pour class
- Les intervalles maximisent les co-occurrences des valeurs

# Jeu de données discrétisé

- Valeurs catégorielles uniquement
- Simplifie l'interprétation du résultat
- Améliore les performances (temps de calcul)

| outlook | temperature | humidity | windy | play |
|---|---|---|---|---|
| sunny | hot | high | false | no |
| sunny | hot | high | true | no |
| overcast | hot | high | false | yes |
| rainy | mild | high | false | yes |
| rainy | cool | normal | false | yes |
| rainy | cool | normal | true | no |
| overcast | cool | normal | true | yes |
| sunny | mild | high | false | no |
| sunny | cool | normal | false | yes |
| rainy | mild | normal | false | yes |
| sunny | mild | normal | true | yes |
| overcast | mild | high | true | yes |
| overcast | hot | normal | false | yes |
| rainy | mild | high | true | no |

## Sites et outils :

- LA référence : KDNuggets
  - http ://www.kdnuggets.com/
- TheData Mine
  - http ://www.the-data-mine.com/
- Weka : en TPs
  - http ://www.cs.waikato.ac.nz/~ml/
- SPSS (SPSS Clementine)
  - http ://www.spss.com/SPSSBI/Clementine/
- IBM (Intelligent Data Miner)
  - http ://www.ibm.com/Stories/1997/04/data1.html