Data Mining

Mohammed Fethi KHALFI

Fethi.Khalfi@yahoo.fr

Introduction générale

Université Djilali Liabés- Sidi Bel Abbes

Data Mining: Concepts et Techniques

Introduction

- Motivation: Pourquoi le Data mining?
- Ce qu'est le Data mining?
- Data Mining: Sur quels types de données?
- Fonctionnalités du Data mining
- Intérêt des motifs (patterns)
- Classification des systèmes de Data mining
- Problèmes rencontrés

Motivation: Le besoin crée l'invention

- Problème de l'explosion de données
  - Les outils automatiques de collecte de données font que les Bases de Données (BD's) contiennent énormément de données (Ex: La base de données des transactions d'un super marché)
- Beaucoup de données mais peu de connaissances !
- Solution: Data warehousing et data mining
  - Data warehousing et OLAP (On Line Analytical Processing)
  - Extraction de connaissances intéressantes (règles, régularités, patterns, contraintes) à partir de données

Le data mining : des data...

Les données X_{ij} peuvent être de différent types

- quantitatif (mesurables)
- catégoriel (nominales, ordinales)
- mais également
  - textes, images, réseaux...



  - tout en même temps


Big Data : augmentation sans cesse de données générées

Twitter : 50M de tweets /jour (=7 téraoctets)
Facebook : 10 téraoctets /jour
Youtube : 50h de vidéos uploadées /minute
2.9 million de mail /seconde

Evolution des Bases de Données

- 1960s:
  - Collecte des données, création des BD's, le modèle réseau
- 1970s:
  - Modèle et SGBD's relationnels, SQL,
- 1980s:
  - Modèles de données et SGBD's avancés (relationnel étendu, OO, déductifs, etc.) et SGBD's dédiés (spatial, génomique, engineering, etc.)
- 1990s—2000s:
  - Data mining et data warehousing, BD's multimédia, BD's sur le WEB

Ce qu'est le Data Mining

- Data mining :
  - Extraction d'informations intéressantes préalablement inconnues et potentiellement utiles) à partir de grandes bases de données.
- Autres appellations:
  - ECD (Extraction de Connaissances à partir de Données)
  - KDD (Knowledge Discovery from Databases)
  - Analyse de données/patterns, business intelligence, fouille de données, etc …

Des statistiques …

- Statistique
  - Quelques centaines d'individus
  - Quelques variables recueillies
  - Fortes hypothèses sur les lois statistiques suivies
- Analyse de données
  - Quelques dizaines de milliers d'individus
  - Quelques dizaines de variables
  - Construction de tableaux: Individus * Variables
  - Importance du calcul et de la représentation visuelle

… au datamining

- Datamining
  - Quelques millions d'individus
  - Quelques centaines de variables
  - Nombreuses variables non numériques
  - Population *constamment* évolutive (difficulté de l'échantillonage)
  - Nécessité de calcul rapide
  - On ne cherche pas nécessairement l'optimum mathématique mais plutôt un modèle qu'un non statisticien pourrait appréhender

Pourquoi faire ?
Applications potentielles

- Analyse de données et aide à la décision
  - Analyse de marché
    - Marketing ciblé, gestion des relations client, analyse des achats des clients,
  - Détection de fraudes
- Autres Applications
  - Text mining : news groups, emails, documents Web.

Analyse de marché et management (1)

- Les sources de données à analyser ?
  - Transactions avec carte de crédit, carte de fidélité, sondages

- Marketing ciblé
  - Trouver un « modèle » pour regrouper les clients partageant les mêmes caractéristiques. Pour chaque groupe, adopter une démarche marketing particulière

Applications

- L'analyse d'une BD de transactions d'un supermarché permet d'étudier le comportement des clients :
  - réorganiser les rayons
  - Ajuster les promotions

- L'analyse de données médicales :
  - Support pour la recherche

Applications

- Exemples
  - Assurances auto: détecter les personnes qui collectionnent les accidents et les remboursements
  - Blanchiment d'argent: détecter les transactions suspectes (US Treasury's Financial Crimes Enforcement Network)

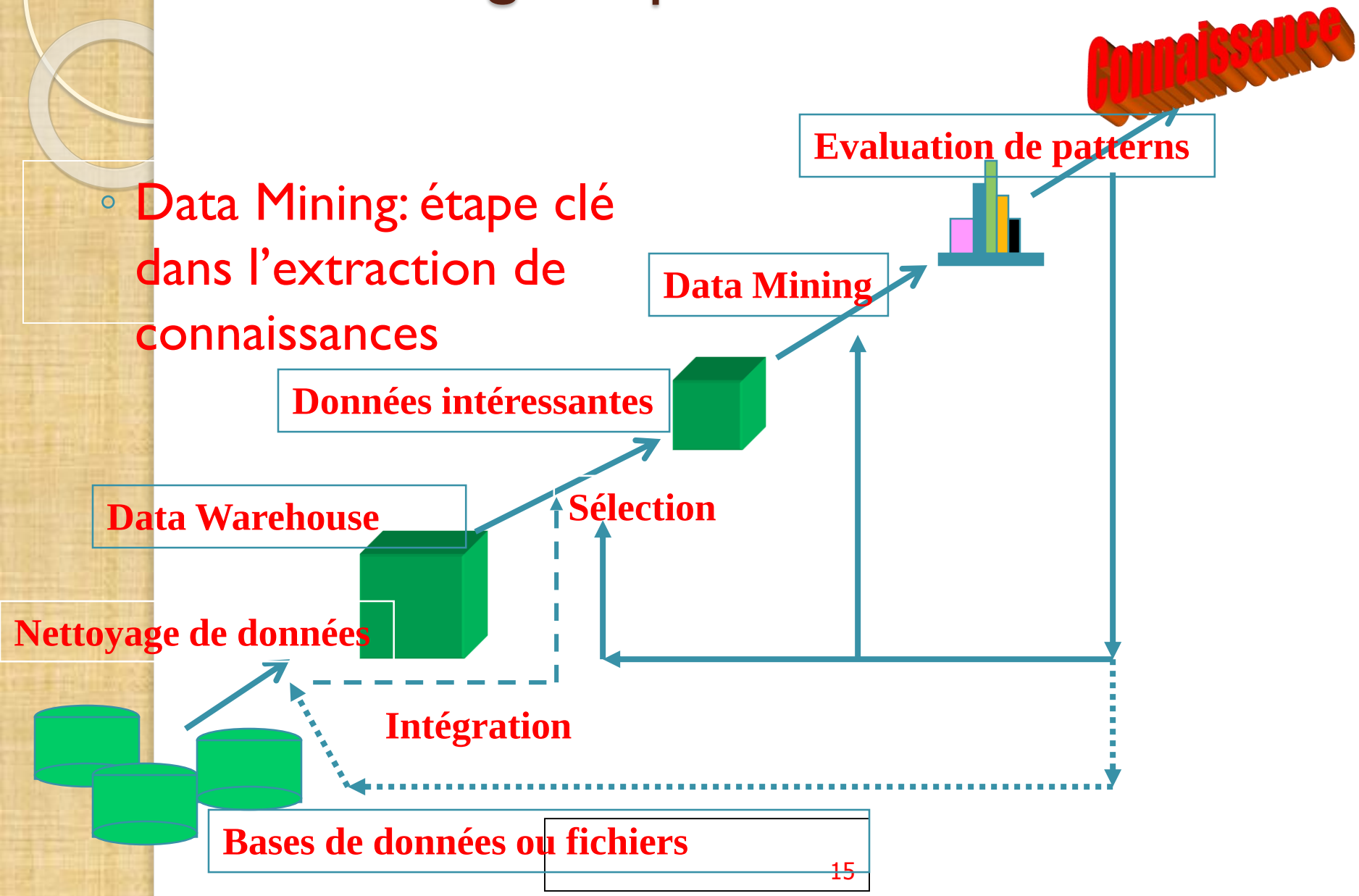
Applications

- Vente, marketing
  - gestion de la relation client (scoring, score d'appétence)
  - segmentation de la clientèle
- Banque, finance, assurance
  - détection de fraude (comportements atypiques)
  - score de risque (attribution ou non d'un crédit)
- Technologie
  - reconnaissance faciale dans une image
  - reconnaissance de la parole
- Médecine, industrie pharmaceutique
  - réponse d'un patient vis-à-vis d'un traitement
  - identification des facteurs de risques
- Energie, transport...
  - prévision de consommation d'électricité
  - prévision de traffic routier

Le Data Mining peut s'appliquer à tout phénomène dont on peut mesurer des observations et dont on souhaite appréhender les caractéristiques et / ou prévoir le comportement

Datamining: Un processus dans l'ECD

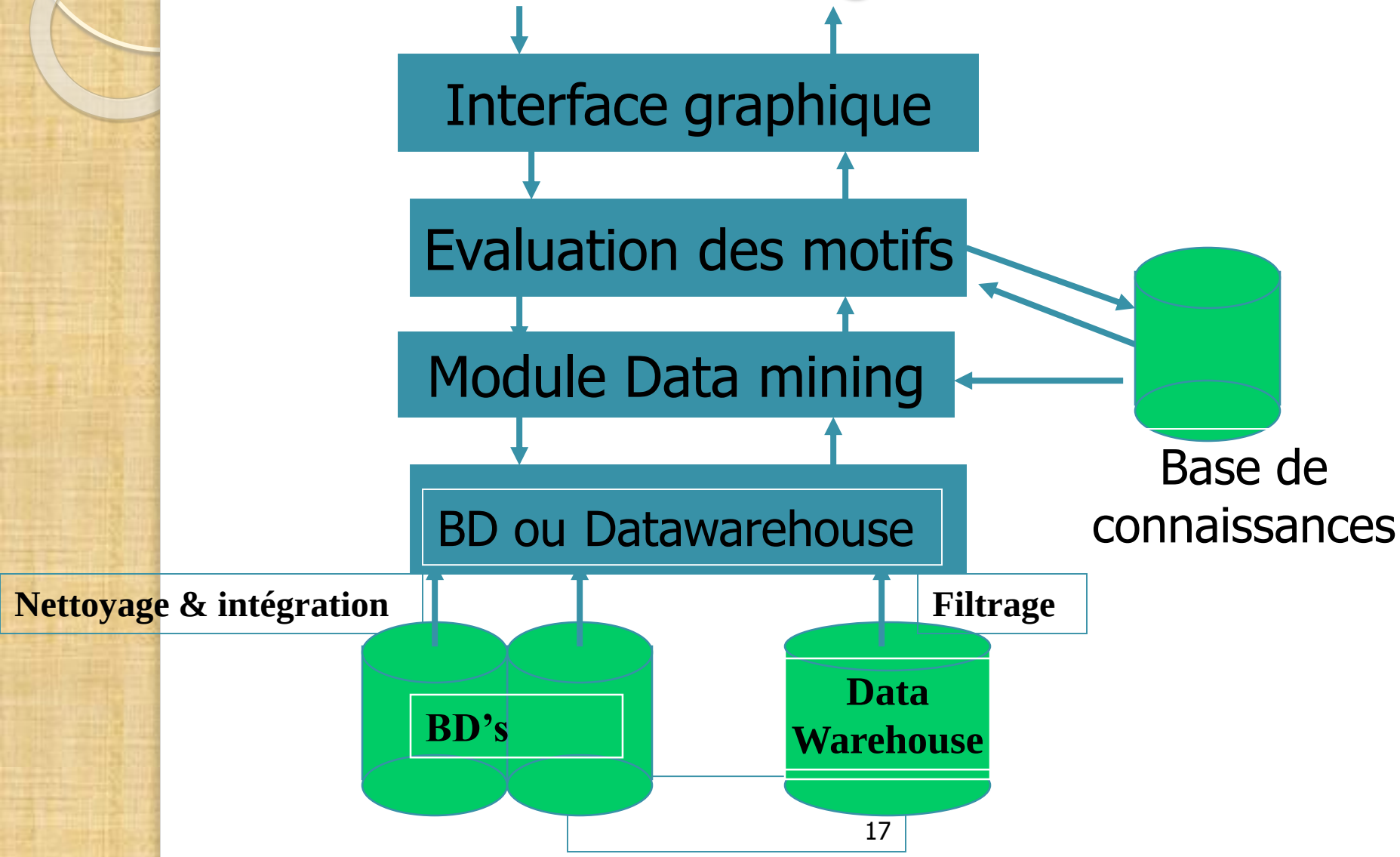
- Data Mining: étape clé dans l'extraction de connaissances



Etapes du processus d'ECD

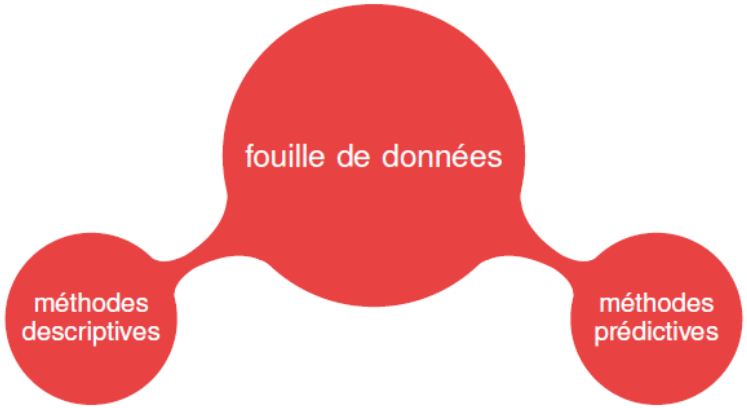
- Comprendre le domaine d'application
- Création d'un ensemble de données (sélection)
- Nettoyage et pré-traitement des données (peut prendre 60% de l'effort)
- Choix des fonctionnalités du data mining
  - classification, consolidation, régression, association, clustering.
- Choix de(s) l'algorithme(s) d'extraction
- Datamining: Recherche des motifs (patterns) intéressants
- Evaluation des Patterns et présentation
  - visualisation, transformation, suppression des patterns redondants, etc.
- Utilisation de la connaissance extraite

Architecture typique d'un système de Data mining



Fonctionnalités du Data Mining

- On distingue deux grandes familles de tâches réalisées en datamining
  - Description : consiste à trouver les caractéristiques générales relatives aux données fouillées
  - Prédiction : consiste à faire de l'inférence à partir des données actuelles pour prédire des évolutions futures



Description

- Il s'agit de **mettre en évidence des informations présentes mais cachées par le volume des données**
- Réduit, résume et synthétise les données
- Il n'y a pas de variable cible à prédire

Techniques descriptives

- Regroupement (ou segmentation, ou clustering)
- Recherche d'associations, de corrélations
- Recherche de séquences similaires

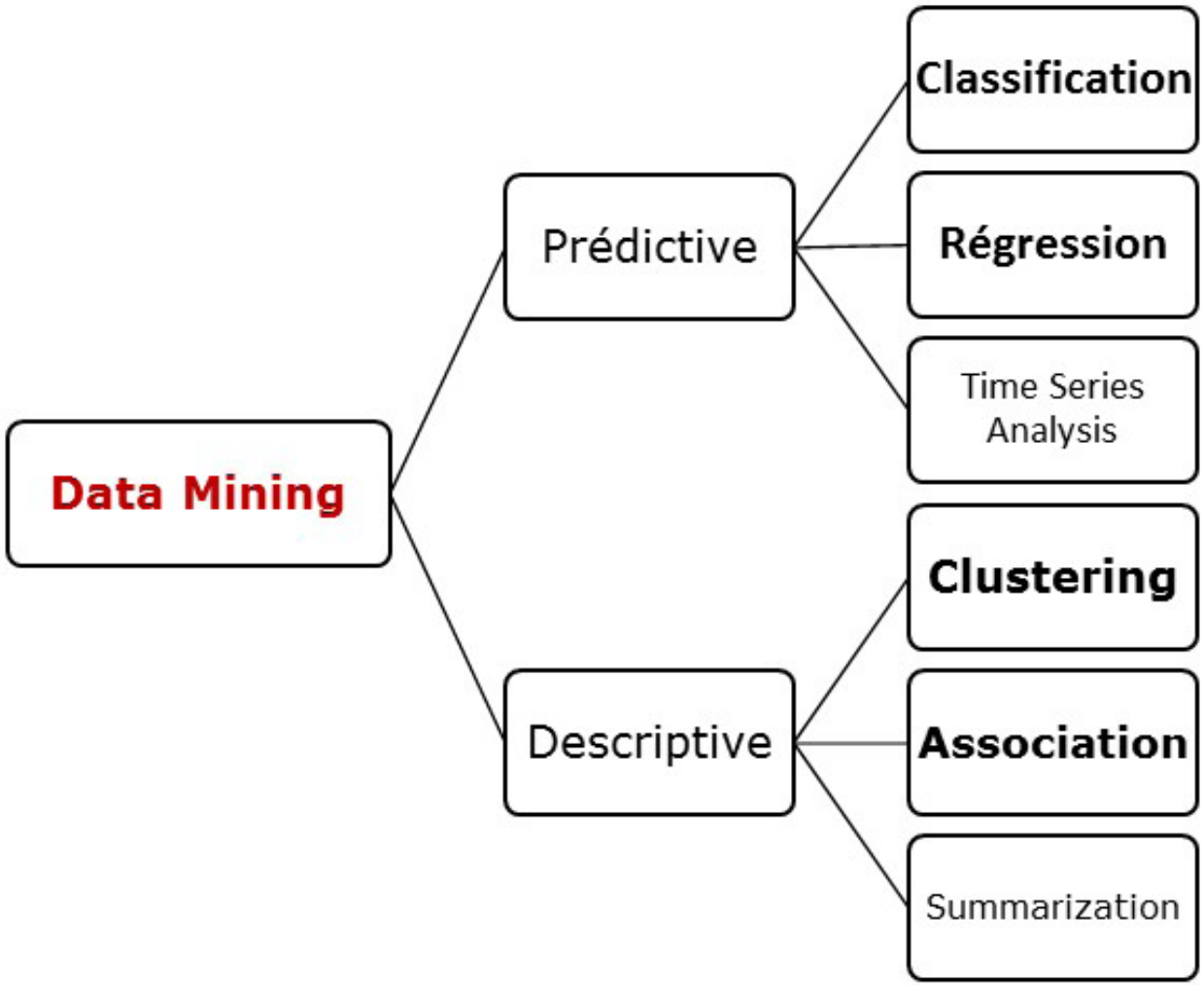
Prédiction

- Vise à **extrapoler de nouvelles informations** à partir d'informations déjà présentes
- Explique les données
- Il y a une variable cible à prédire

Techniques prédictives

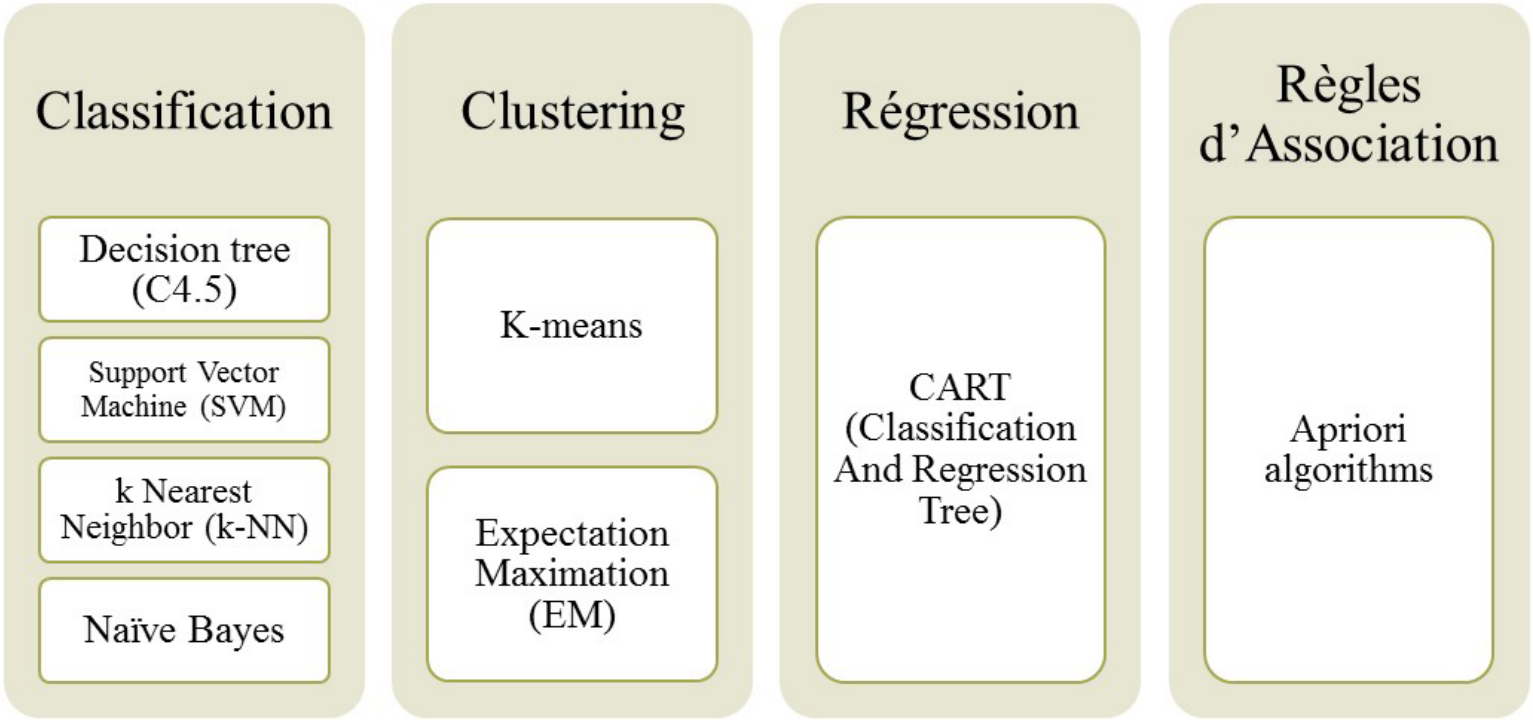
- Classification
  - Arbres de décision
  - Classification bayésienne
  - Réseaux neuronaux
  - Méthodes SVM (support vector machine)
  - Régression
  - …

Fonctionnalités du Data Mining



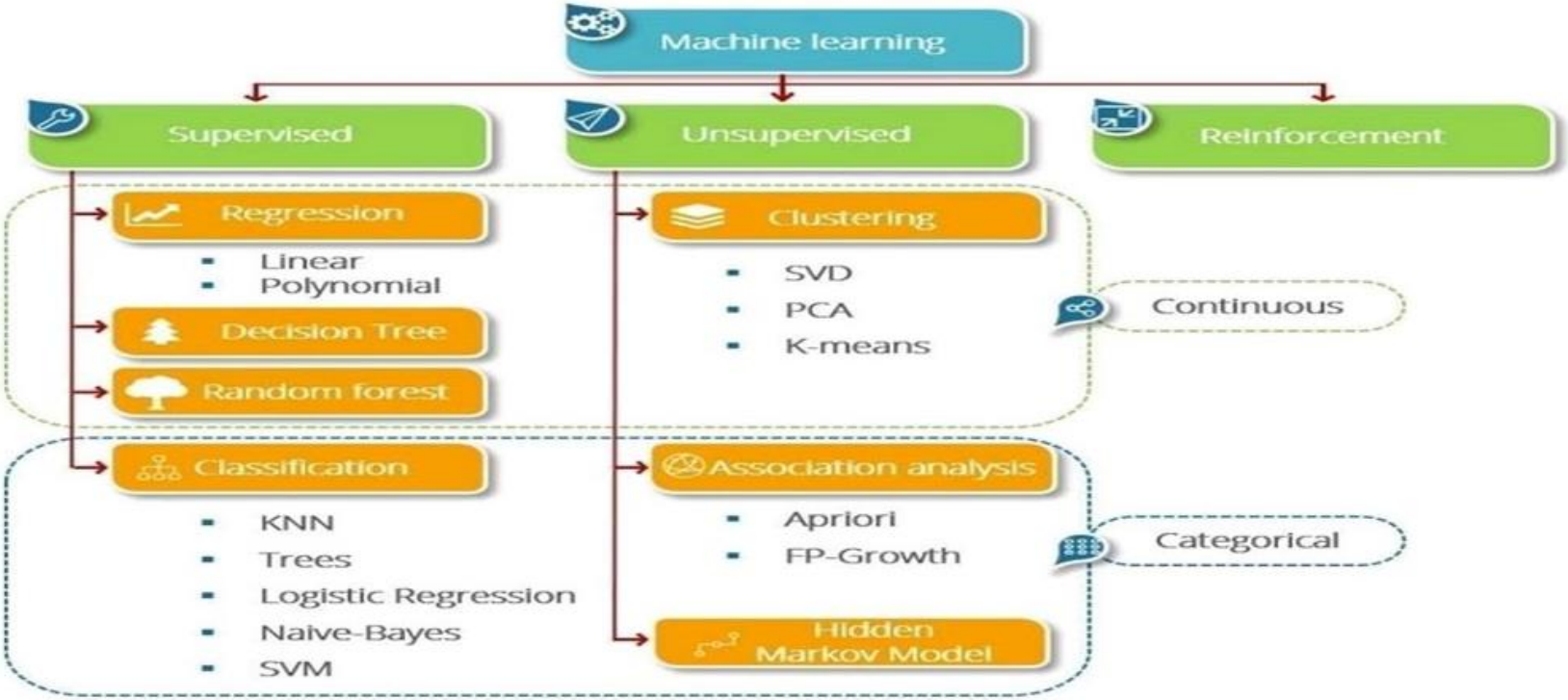
Fonctionnalités du Data Mining

Les algorithmes basiques du Data Mining



Fonctionnalités du Data Mining

Classification des algo. De machine learning



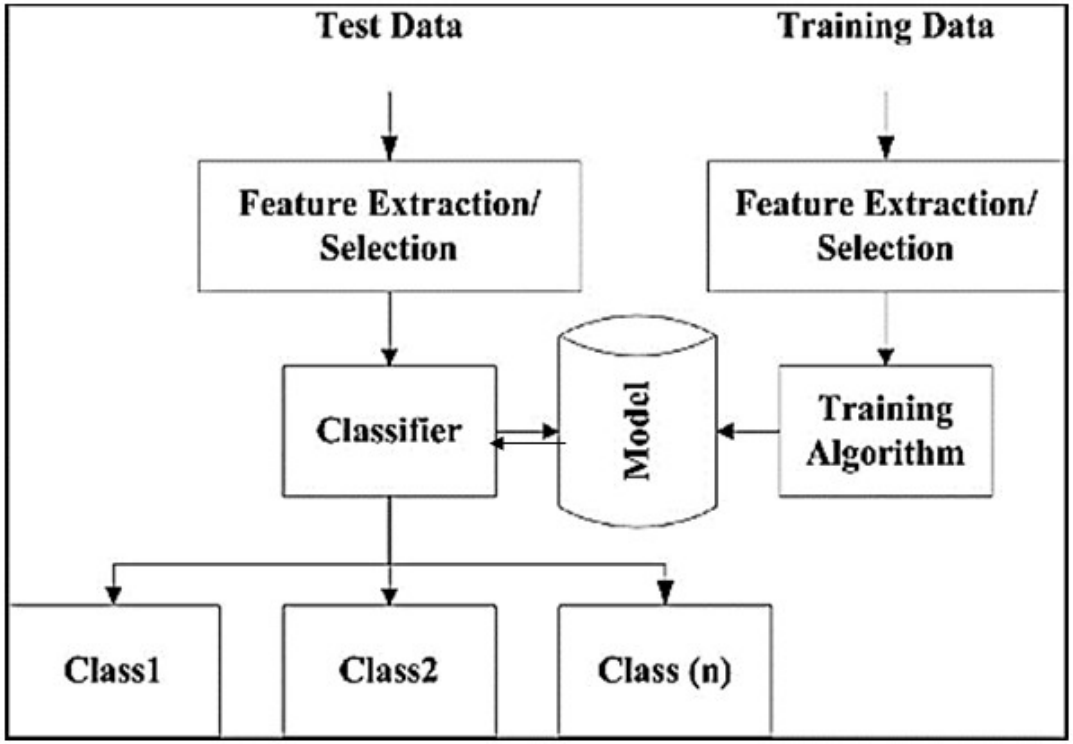
Fonctionnalités du Data Mining

- La classification supervisé et non supervisé (clustering) font parties des tâches de l'apprentissage machine (Machine Learning).

- Classification supervisé consiste à examiner les caractéristiques d'un objet nouvellement présenté (**Test dataset**) afin de l'affecter à une classe (class or labels) d'un ensemble prédéfini (**Training dataset**).
  - Le modèle généré permet de prédire ou estimer la valeur manquante en utilisant l'algorithme de classification supervisé comme référence.
  - Les algorithmes les plus utilisés : SVM, k-NN, Naive bayes, C4.5 …

Fonctionnalités du Data Mining

- Supervised classification workflow



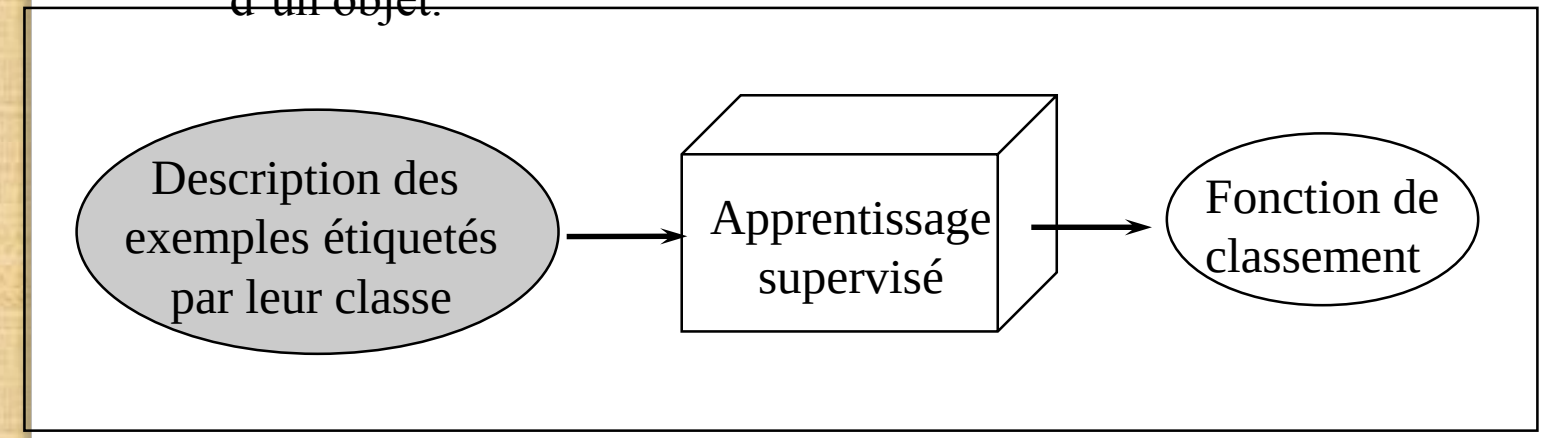
Fonctionnalités du Data Mining

L'apprentissage supervisé

• On dispose d'un ensemble de données étiquetées par un expert
⇨la base d'apprentissage

• Objectif de l'apprentissage supervisé :
construire à partir de la base d'apprentissage des fonctions de classement

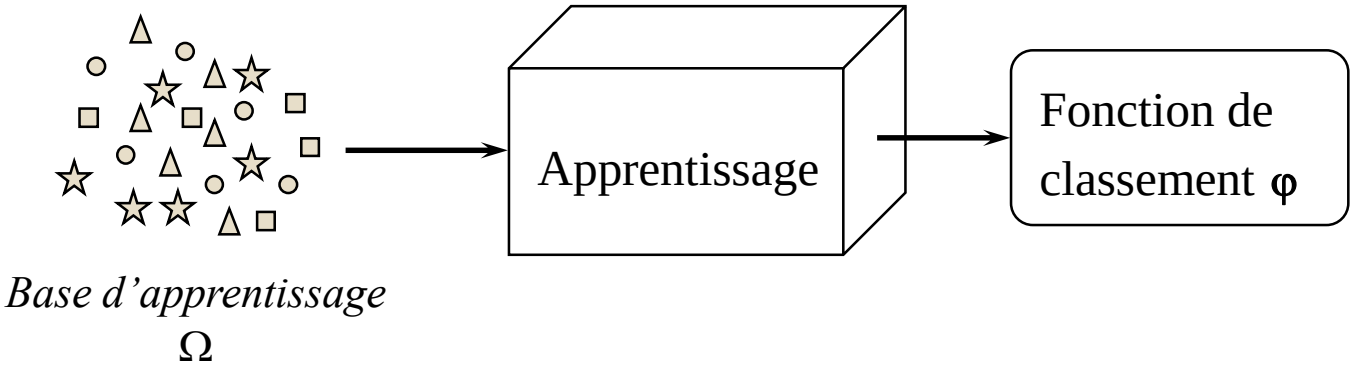
•Fonction de classement :
reconnaît un attribut particulier (la classe) à partir de la description d'un objet.



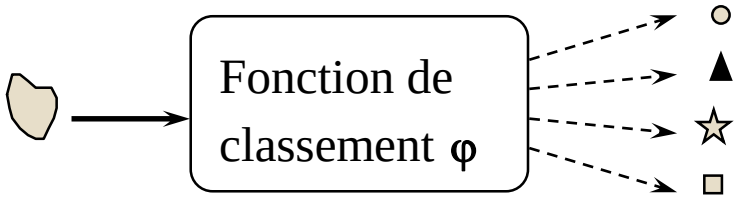
Fonctionnalités du Data Mining

L'apprentissage supervisé

è La phase d'apprentissage



è La phase de reconnaissance



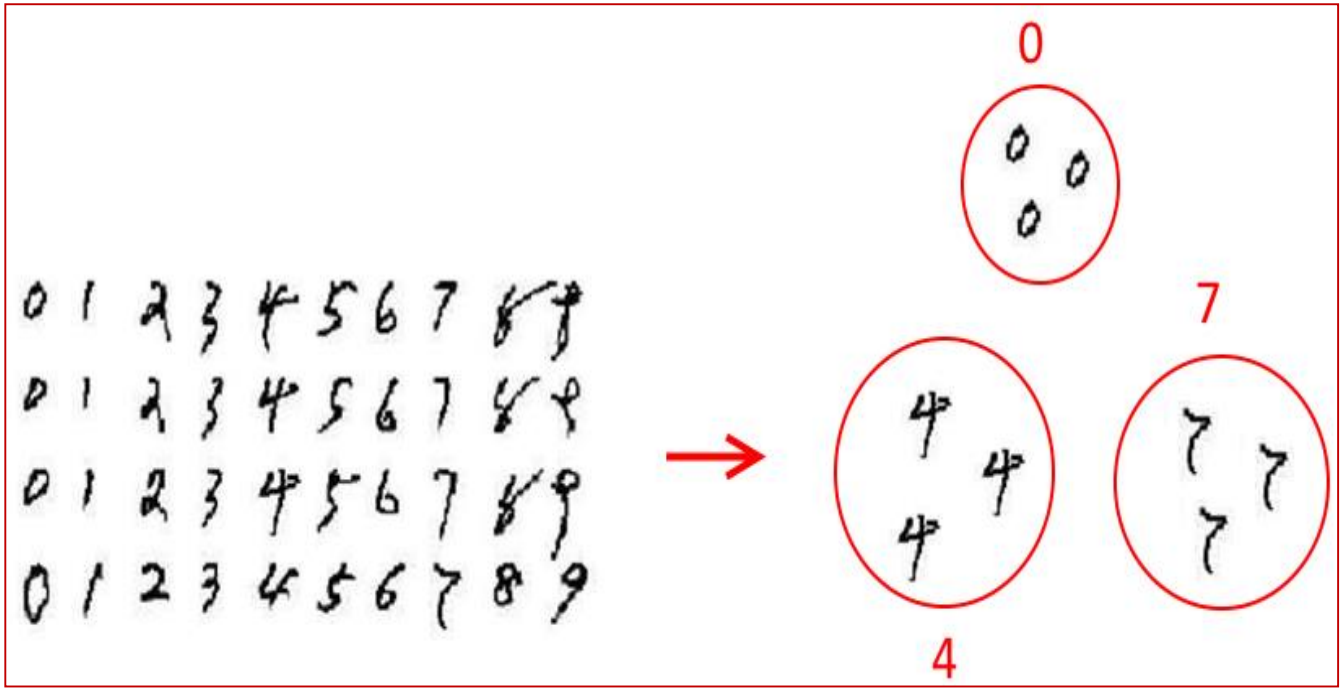
Fonctionnalités du Data Mining

Clustering ou classification non supervisé

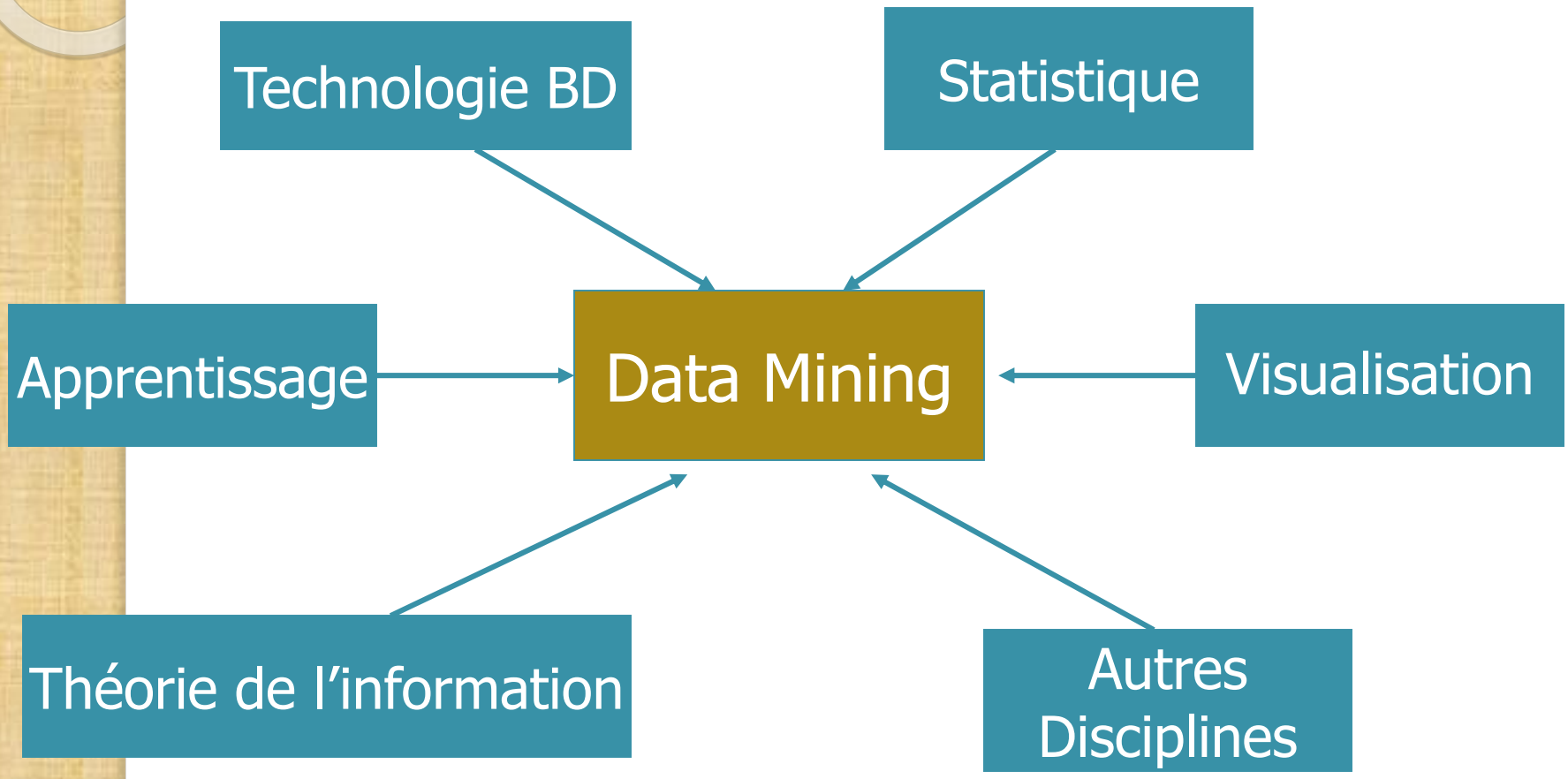
- Clustering (regroupement) lorsque un ensemble de données avec des étiquettes de classe ne sont pas disponibles, par exemple lors de l'introduction d'un nouveau produit.
- Le clustering est un processus de partitionnement d'un ensemble de données (ou d'objets) en un ensemble de sous-classes significatives, appelées **clusters**.
  - Peut aider les utilisateurs à comprendre le regroupement naturel ou la structure dans un ensemble de données.
- Clustering: **classification non supervisée**: pas de classes prédéfinies.
- Utilisé soit comme **outil autonome** pour obtenir un aperçu de la distribution des données, soit comme étape de prétraitement pour d'autres algorithmes.

Fonctionnalités du Data Mining

Clustering ou classification non supervisé



Data Mining: Confluence de plusieurs Disciplines



Quelques logiciels de fouille de données

- **Logiciels commerciallsés :**
  - S-PLUSTM de Insight,
  - AliceTM de Isoft,
  - Predict TM de Neuralware,
  - R (version gratuite de S-PLUS)
- **Logiciels gratuits :**
  - **Weka**
  - **Tanagra**
  - **Orange**
- **Intérêts :**
  - faciles à installer, utiliser, prix abordable
  - adaptés aux PME car ils peuvent gérer plusieurs dizaines de milliers voire plusieurs centaines de milliers d'individus
- **Limites :**
  - ils ne permettent pas de traiter de très grandes bases de données
  - ils ne mettent souvent en œuvre qu'une ou deux techniques (excepté les produits S-PLUS, R, Tanagra et Weka)

Quelques logiciels de fouille de données

- **Logiciels commerciallsés :**
  - S-PLUSTM de Insight,
  - AliceTM de Isoft,
  - Predict TM de Neuralware,
  - R (version gratuite de S-PLUS)
- **Logiciels gratuits :**
  - **Weka**
  - **Tanagra**
  - **Orange**
- **Intérêts :**
  - faciles à installer, utiliser, prix abordable
  - adaptés aux PME car ils peuvent gérer plusieurs dizaines de milliers voire plusieurs centaines de milliers d'individus
- **Limites :**
  - ils ne permettent pas de traiter de très grandes bases de données
  - ils ne mettent souvent en œuvre qu'une ou deux techniques (excepté les produits S-PLUS, R, Tanagra et Weka)

Quelques logiciels de fouille de données

- **Weka :**
  - Weka (Waikato Environment for Knowledge Analysis) est un ensemble de classes et d'algorithmes en Java développé à l'Université de Waikato en Nouvelle Zélande
  - Weka implémente les principaux algorithmes de la fouille, notamment :
    - **les arbres de décision**
    - **les réseaux de neurones**
  - il est téléchargeable (versions Unix et Windows) à l'adresse : http://www.cs.waikato.ac.nz/ml/weka
  - développé en complément du livre : Data Mining par I. Witten et E. Frank (éditions Morgan Kaufmann).
  - peut être utilisé de plusieurs façons :
    - par l'intermédiaire d'une interface utilisateur (comme utilisée en TP)
    - sur la ligne de commande.
    - par l'utilisation des classes fournies à l'intérieur de programmes Java (classes documentées)

Quelques logiciels de fouille de données

- **Tanagra :**
  - TANAGRA est un logiciel gratuit développé à l'Université de Lumiere Lyon 2, laboratoire ERIC, par Ricco Rakotomalala
  - Il est destiné à l'enseignement et à la recherche, et téléchargeable à l'adresse : http://chirouble.univ-lyon2.fr/~ricco/cours/index.html
  - Il implémente diverses méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'apprentissage automatique et des bases de données, …
- **Orange :**
  - est développé par Blaz Zupan, à la Faculty of Computer and Information Science, de l'Université de Ljubljana en Slovenie
  - Il est destiné à l'enseignement et à la recherche, et téléchargeable à l'adresse : http://www.ailab.si/orange
  - Il implémente aussi diverses méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'apprentissage automatique et des bases de données, …

Quelques systèmes

- Intelligent miner d'IBM (couplé avec le SGBD DB2)
  - Classification, association, régression, analyse de séquences, regroupement
- Entreprise miner de SAS
  - Multiples outils d'analyse statistique, classification, …
- Mine set de Silicon graphics.
  - Classification, association et divers outils statistiques. Très puissant en terme de visualisation
- Clémentine de SPSS
  - En plus des fonctionnalités classiques, l'utilisateur peut y rajouter ses propres algorithmes
- DBMiner de DBMiner technologie.
  - Il se distingue par le fait qu'il incorpore les fonctionnalités d'OLAP

Conclusion

- Utiliser un système de datamining est intéressant quand on sait
  - Quelles actions nous voulons entreprendre
  - Quelles types d'information nous devons rechercher
- Pour chaque type d'information, il existe plusieurs techniques qui ne sont dans la plupart des cas, pas équivalentes mais complémentaires