Ministère de l'Enseignement supérieur et de la recherche scientifique
Université Abderrahmane Mira de Bejaïa
Faculté des Sciences Exactes
Département de Mathématiques

Module: Data Mining **Année Universitaire: 2022-2023**
Nature de document: Cours **N.BERMAD**
Niveau : L3STID
Chapitre:1
Titre: Introduction au Data Mining et à la Découverte de Connaissances

1. Introduction

- Récemment, nos possibilité de produire et de recueillir des données avaient augmenté rapidement.Des millions de bases de données ont été employés dans la d'entreprise, l'administration de gouvernement, la gestion des données scientifiques, et beaucoup d'autres applications, le nombre de telles bases de données ne cessent d'augmenter jours après jours.
- Cette croissance explosive des données et des bases de données a produit des besoins urgents pour les nouvelles techniques et les outils qui peuvent intelligemment et automatiquement transformer les données traitées en informations utiles et connaissances.

1

- Le Data Mining (Fouille de données) est un domaine qui consiste à comprendre les données, généralement par le moyen de méthodes statistiques .En d'autre termes, le data Mining cherche à identifier des tendances parmi les données .Comme ce processus peut être très difficile, il est souvent comparé au minage de l'or dans les rivières: le gravier des alluvions représente l'énorme quantité de données et les pépites d'or représentent la connaissance cachées que l'on veut trouver.

2

2. Définition de datamining

- La fouille de données consiste à rechercher et extraire de l'information (utile et inconnue, non trivial,implicites) de gros volumes de données stockées dans des bases ou des entrepôts de données.
- C'est l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de grandes bases de données informatiques, de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en déduisant la quantité de données permettant d'étayer des prises de décisions.

3

4. Motivation

Explosion de données

- Masse important de données (millions de milliards d'instances)
  - ✓ **BD très large(VLDB)**
- Données multidimensionnelles (milliers d'attributs)
  - ✓ **BD denses**
- Inexploitable par les méthodes d'analyse classiques.
- Collecte de masses importantes de données(Gbytes/heurs)
  - ✓ **Données satellitaire, génomique, simulation scientifiques, etc...**
- Besoin de traitement en temps réel de ces données

4

Améliorer la productivité

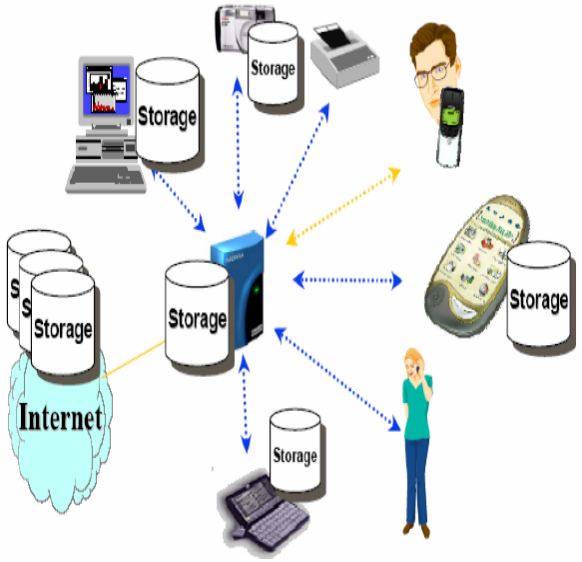
- Forte pression due à la concurrence du marché
- Brièveté du cycle de vie des produits
- Besoin de prendre des décisions stratégiques efficaces
  - ✓ **Exploiter le vécu (données historiques) pour prédire le futur et anticiper le marché.**
  - ✓ **Individualisationdes consommateurs.**

Croissances en puissances /coût des machines capables.

- ✓ **De supporter de gros volumes de données**
- ✓ **D'exécuter le processus intensif d'exploration**
- ✓ **Hétérogénéité des supports de stockages**.

5

Masse importantes de données/ support hétérogènes



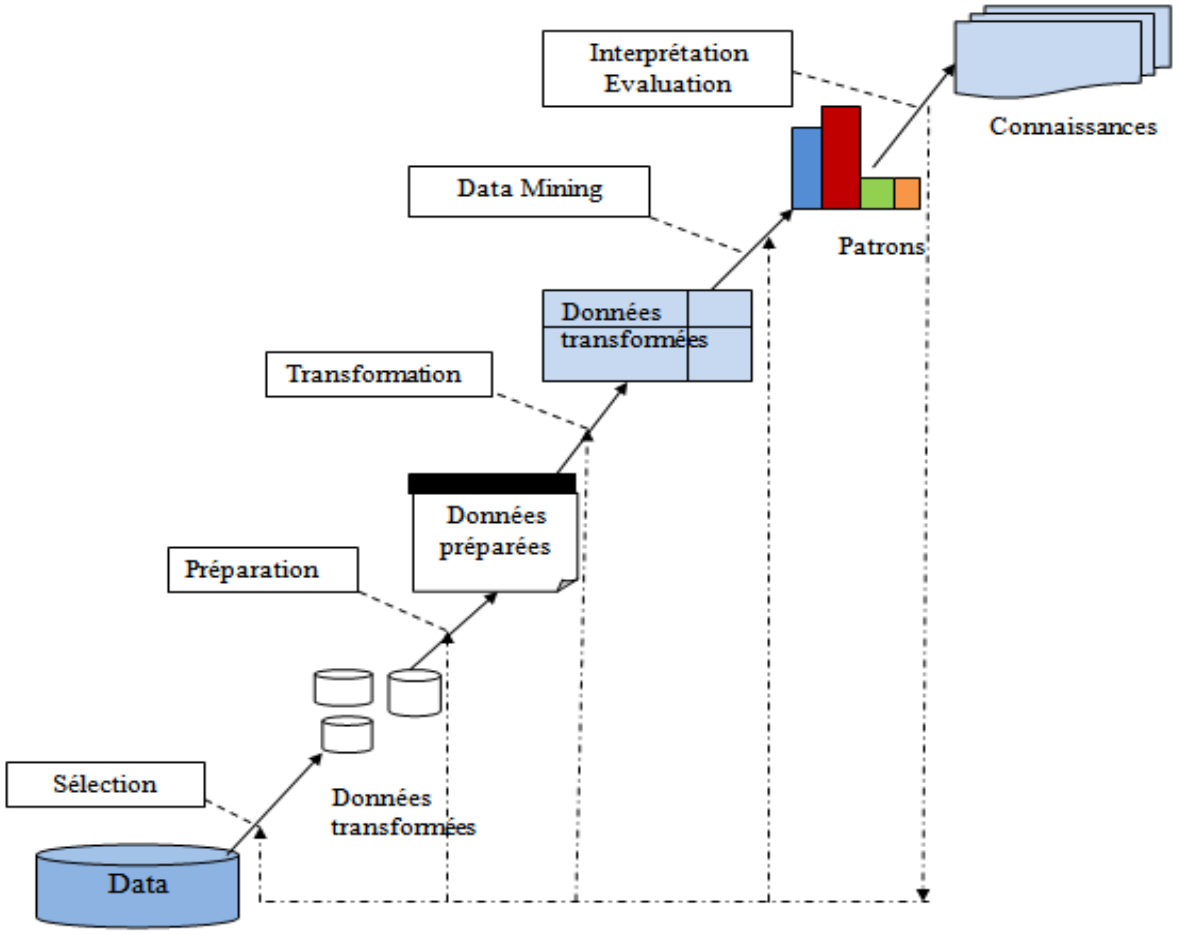
6

5. Processus d'extraction de connaissances à partir de données (ECD)

Une confusion subsiste encore entre le "fouille de données", et "l'extraction des connaissances à partir des données. Le data mining est l'un des maillons de la chaine de traitement pour la découverte des connaissances à partir des données .Sous forme imagée, nous pourrions dire que l'ECD est un véhicule dont le data mining est le moteur.

Le ECD est un processus semi -automatique et itératif, constitué de plusieurs étapes allant de la sélection et préparation des données jusqu'à l'interprétation des résultats, en passant par la phase de recherche des connaissances (le data mining), les différents étapes de ce processus sont présentées dans la figure ci-dessous:

7



- **Sélection de données:** Le but de cette phase est l'extraction à partir d'un plus grand stock de données seulement celles qui sont appropriées à l'analyse d'exploitation de données. Cette extraction de données aide à rationaliser et accélérer le processus
- **Préparation de données:** Cette phase d'ECD est concernée par les données nettoyant et les tâches de préparation qui sont nécessaire pour assurer les résultats corrects.
- **Transformation de données:**Les données sélectionnées dans l'étape précédente vont subir une transformation dont le but est de les rendre dans une forme appropriée pour les méthodes et les techniques de datamining.
- **Data mining:** Le but de la phase d'exploitation de données est d'analyser les données par un ensemble approprié d'algorithmes afin de découvrir les modèles et les règles significatifs et produire les modèles prédictifs. C'est l'élément de noyau du cycle d'EDC.

- **Interprétation et évaluation:** Tandis que les algorithmes de data mining ont le potentiel de produire un nombre illimité de modèles cachés dans les données, beaucoup de ces derniers peuvent ne pas être significatifs ou utiles. Cette phase finale est visée en choisissant ces modèles qui sont valides et utiles pour prendre de futures décisions économiques.

6. La démarche d'une ETUDE DATA MINING:

1. **Nettoyage des données** : Erreurs, Données manquantes, Valeurs atypiquesattributs sansvaleur, ayant une valeur.

2. **Transformation des données**: Normalisation, Linéarisation, Compression.

3. **Explicitation de l'objectif et de la stratégie d'analyse :** Exploration, Classification, Discrimination, Segmentation, Recherche de singularités, Modélisation, Prévision, etc...

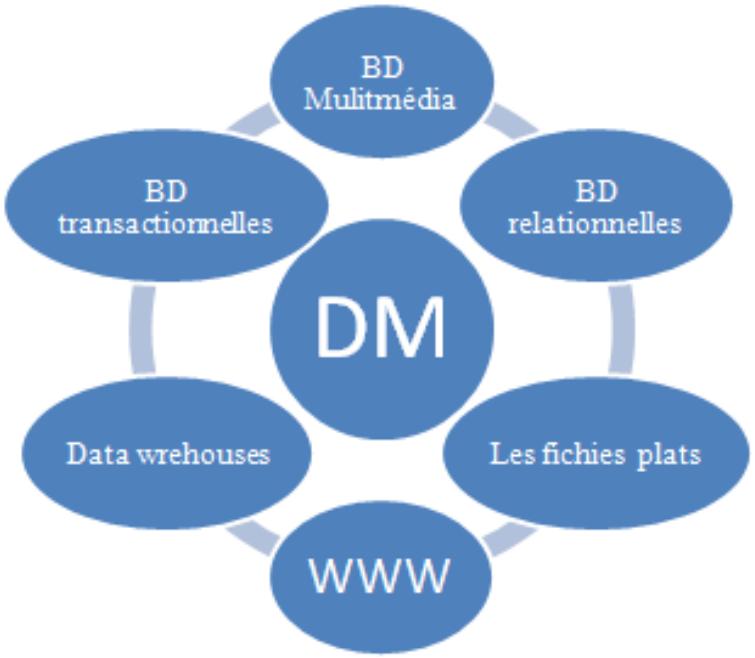
4. Choix des méthodes et algorithmes et mise en œuvre informatique.

5. Test et Validation (Choix des critères)

6. Conclusion (Enseignements, Prévision)

7. Data Mining sur quels types de données?

Le Data Mining est applicable à n'importe quel type de données, il n'est pas spécifique à un type particulier de données.



- **Les bases de données multimédias :**Les bases de données multimédia incluent la vidéo, les images et l'audio. Elles peuvent être stockées sur des BD orienté objets ou relationnelles.

- **Le Word Wide Web:** C'est le dépôt le plus hétérogènes et le plus dynamique disponible. Les données dans le Word Wide Web sont organisées dans des documents interconnectés.

- **Les fichiers plats:** Les fichiers plat sont des fichiers de données simples dans un format binaire ou textuelle avec une structure connue par l'algorithme du data ming qui va-t-être y appliqué.

- **Les bases de données relationnelles:** Une base de données se compose d'un

ensemble de tables qui ont des colonnes et des lignes, où les colonnes représentent les attributs et les lignes représentent les tuples.

- **Le data warehouse (entrepôt de données):**C'est un support de données dans laquelle est centralisé un volume important de données consolidées à partir des différents sources de données hétérogènes.

- **Les bases de données transactionnelles:**

  En général, une base de données transactionnelle est un fichier ou chaque enregistrement représente une transaction (par exemple : les achats d'un client lors d'une visite)

8. Domaines d'application

Le data Miningest devenu plus en plus applicable dans différents domaines et secteurs d'activités parmi lesquels:

- **Gestion et analyse de marchés:**Analyse des comportements des consommateurs, recherche de ses similarités en fonction de critères, prédiction, des ventes croisées, optimisation des réapprovisionnements.
- **Banques:** Recherche de formes d'utilisation de cartes caractéristiques d'une fraude, prédictive des clients partants, détermination de pré-autorisations de crédits.
- **Assurance:** Modèles de sélection et de tarification, analyse des sinistres, recherche des critères explicatifs du risque ou de fraude,

prévision d'appels sur les plates-formes d'assurance directe.

- **Médecine et pharmacie:** Modélisation comportementale et prédiction de médication ou de visites, identification des thérapies pour différentes maladies.
- Aéronautique, automobile et industries: Contrôle de qualité des défauts prévisions des ventes, dépouillement d'enquêtes de satisfaction.
- **Télécommunications, eau et énergie:** Simulation des tarifs, détection des formes de consommation frauduleuse, classification des clients selon la forme de l'utilisation des services de prévisions de ventes.
- **Education**: Analyse des facteurs d'échec...

Ministère de l'Enseignement supérieur et de la recherche scientifique
Université Abderrahmane Mira de Bejaïa
Faculté des Sciences Exactes
Département de Mathématiques

Module: Data Mining **Année Universitaire: 2022-2023**
Nature de document: Cours **N.BERMAD**
Niveau : L3STID
Chapitre:1
Titre: Introduction au Data Mining et à la Découverte de Connaissances

Exemple1: Marketing

Vous êtes gestionnaire marketing d'un opérateur de télécommunications mobiles:

- Les clients reçoivent une téléphonie gratuite (valeur 150 £) avec un contrat d'un an; vous payer une commission de vente de 250£ par contrat
- **Problème:**Taux de renouvellement (à la fin du contrat) est de 25%
- Donner un nouveau téléphone à toute personne ayant expiré son contrat coûte cher.
- Faire revenir un client après avoir quitté est difficile et coûteux

Exemple2: Assurance

Vous êtes un agent d'assurance et vous devez définir un paiement mensuel adapté à un jeune de 18 ans qui a acheté une Ferrari

- Qu'est ce qu'il faut faire?
- Analyser les données de tous les clients de la compagnie.
- La probabilité d'avoir un accident est basée sur...?
  - ✓ Sexe du client (M/F) et l'âge.
  - ✓ Modèle de la voiture, âge, adresse,....
  - ✓ etc.
- Si la probabilité d'avoir un accident est supérieure à la moyenne, initialiser la mensualité suivant les risques.

Exemple3:Web

- Les logs des accès web sont analysés pour...
  - ✓ Découvrir les préférences des utilisateurs
  - ✓ Améliorer l'organisation du site Web
- De manière similaire...
  - ✓ L'analyse de tous les types d'informations sur les logs
  - ✓ Adaptation de l'interface utilisateur/service.

8. Outils logiciel de datamining

- NeuroOnline
- ARMiner
- Tminer
- Bayda
- Cluto
- YALE
- Rainbow
- MDR
- C5
- Clementine
- RapideMiner
- Etc…