

Université Mohamed Boudiaf – M'sila
Département Informatique
Master Intelligence Artificielle
2020-2021

Chapitre 3
Data Mining: Concepts et
Méthodologie

Dr. Mehenni Tahar

Un Problème ...

- Vous êtes Responsable du Marketing dans une société de courtage
 - Problème: **Renoncement** très élevé des clients
 - > Taux de renoncement 40%
(La période de six mois de courtage est finie)
 - Les clients reçoivent des **encouragements** lorsque l'affaire est ouverte
 - Donner de nouveaux encouragements à chaque client est très coûteux (parfois inutile)
 - Ramener un client après son départ est à la fois difficile et coûteux

Une Solution ...

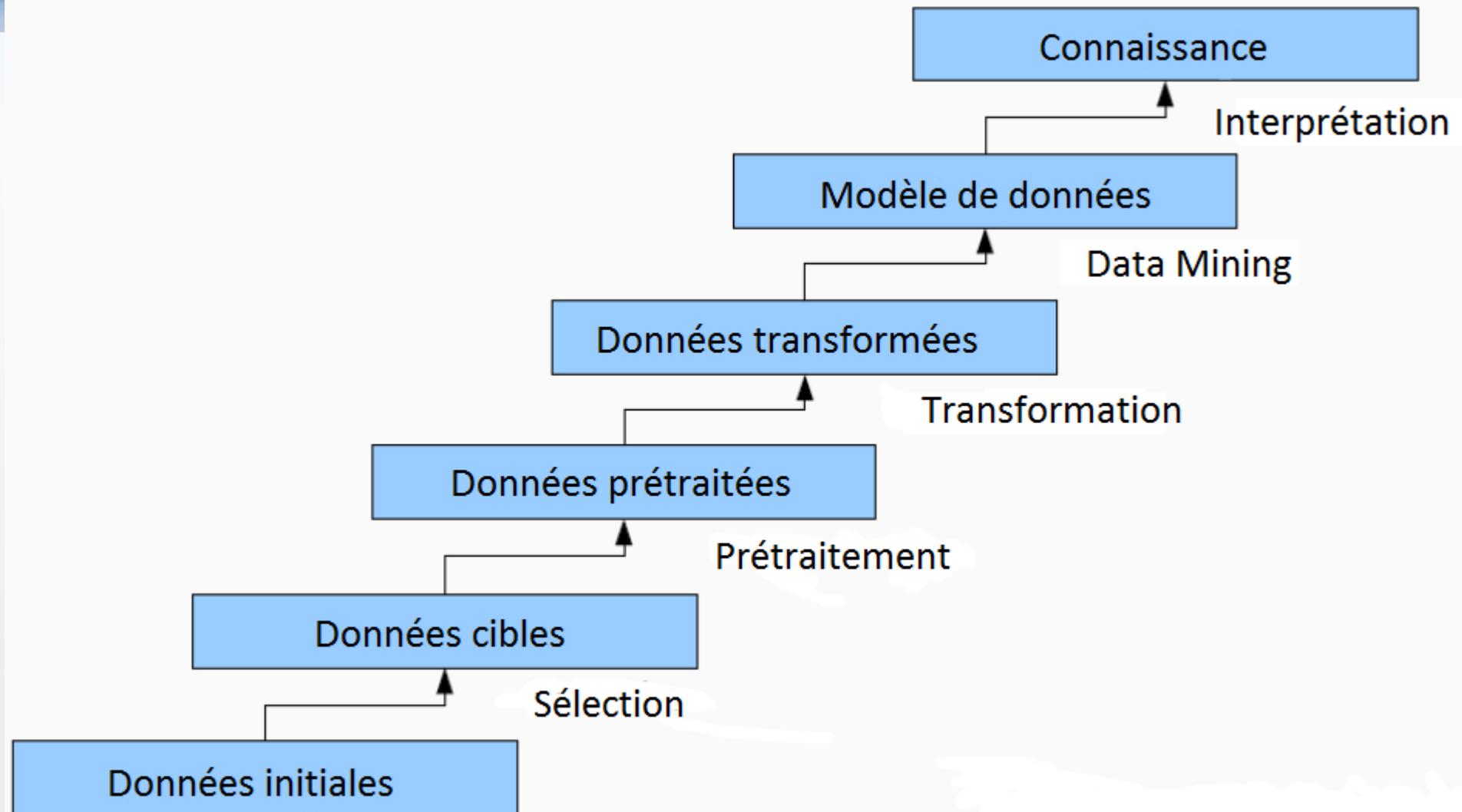
- Un mois avant l'expiration de la période de courtage, **prévoir** (prédire) les clients qui vont quitter (renoncer)
- Si on veut garder un client prévu au renoncement, on lui **offre** quelque chose en se basant sur sa **valeur prédite**
 - > Les autres clients non prévus au renoncement n'auront besoin de rien
- Si on ne veut pas garder un client prévu au renoncement, on ne le laisse partir
- Comment prévoir un futur comportement?
 - > Les Cartes?
 - > Les boules magiques?
 - > Autres choses?



Processus du KDD

- **Découverte des Connaissances dans les Bases de Données ou Knowledge Discovery in Databases (KDD)** est un processus multi-étapes pour la découverte d'informations utiles et de patterns à partir des données
- **Fouille de Données ou Data Mining** est l'utilisation d'algorithmes pour l'extraction d'informations et de patterns, conduite par le processus du KDD.
- Plusieurs documents considèrent le KDD et le Data Mining comme étant un même processus, mais il est également possible de traiter le Data Mining en tant qu'étape de découverte du processus de KDD.

Étapes du processus KDD



Etapes du processus KDD

1. Sélection-

Extraction de données –Obtenir les données à partir des sources de données hétérogènes –Bases de données, Data warehouses ou Entrepôts de données, World wide web ou autres récipients de données.

2. Prétraitement-

Nettoyage des données- Incomplètes , incohérentes inconsistantes données à nettoyer – données absentes peuvent être ignorées ou prédites, données erronées à supprimer ou à corriger.

3. Transformation-

Intégration des données- Combiner les données issues des différentes sources dans une forme cohérente –Codage des Données sous des formats communs, normalisation, réduction.

Etapes du processus KDD

4. Data mining –

Application d'algorithmes sur les données transformées pour l'extraction d'informations et de patterns

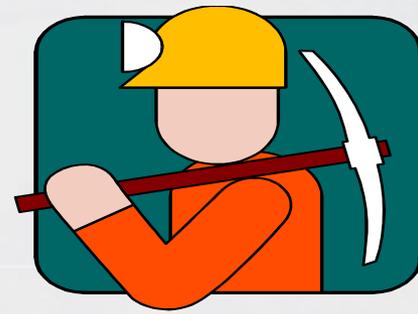
5. Interprétation/évaluation des patterns

Evaluation des patterns- Evaluer l'intérêt des patterns trouvés ou appliquer des mesures d'interressement pour les filtrer.

Présentation des informations/connaissances découvertes-

Les techniques de visualisation peuvent être utilisées.

Qu'est-ce que le Data Mining?



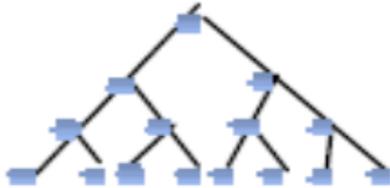
Quelques Définitions

- “L'**extraction non triviale**, à partir des **données**, d'**informations** implicites, inconnues a priori et potentiellement **utiles**” (Piatetsky-Shapiro)
- “...L'**extraction automatique** ou adéquate de **patterns** représentant des **connaissances** implicitement stockées ou capturées dans des **bases de données** volumineuses, des entrepôts de données, du Web, ... ou des flôts de données.” (Han, pg xxi)
- “...Le processus de **découverte** de **patterns** à partir des **données**. Le processus doit être **automatique** ou (généralement) semi-automatique. Les patterns découverts doivent être **significatifs**...” (Witten, pg 5)
- “...**trouver** l'**information** cachée dans les **bases de données**.” (Dunham, pg 3)

Pourquoi le Data Mining?

- On dit souvent que le Data Mining est ...compliqué. Pourquoi donc doit-on l'apprendre?
- Pourquoi ne pas utiliser seulement les bases de données relationnelles? Que veut-on obtenir de plus en suivant ces étapes [extra-compliquées]?
- Ce processus n'est-t-il pas coûteux? Il semble qu'il a besoin de beaucoup de talent, de programmation, de temps de calcul et d'espace de stockage.
- Où est-il le profit?
- Data Mining n'est pas seulement un exercice académique malin, il a des applications réelles très profitables. Pratiquement toutes les grandes compagnies et plusieurs gouvernements développent le data mining comme une tâche de leur planification et analyse.

Data Mining c'est ...



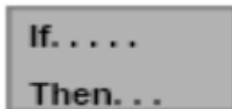
- Arbres de décision



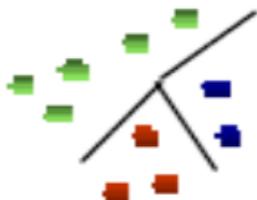
- Classification par Voisinage



- Réseaux de neurones



- Règles d'association



- Clustering

Qu'est-ce qui n'est pas du Data Mining?

Ce n'est pas du Data Mining

- Chercher un numéro de téléphone dans un annuaire
- Requête d'un moteur de recherche internet à propos d'informations sur "Amazon"

C'est du Data Mining

- Certains noms sont populaires dans certaines régions de Mumbai (Kulkarni, Shah, Iyer...)
- Grouper les documents similaires retournés par un moteur de recherche selon leur contexte (e.g. Amazon rainforest, Amazon.com,)

Traitement BDD VS Data Mining

- **Requête**

- Bien définie
- SQL

Données

- **Données
Opérationnelles**

Sortie

- **Précise**
- **Un sous-ensemble de la base de données**

- **Requête**

- Mal définie
- Aucun langage de requête précis

Données

- **Données non
opérationnelles**

Sortie

- **Probable**
- **Ce n'est pas un sous-ensemble de la base de données**

Fonctions de Data mining

**Data
Mining**

Prédictif

Classification: affecter les données à leurs classes prédéfinies

Régression: affecter les données à une fonction

analyse des séries chronologiques: analyse de données au cours du temps, ...

Apprentissage supervisé

Descriptif

Clustering: Trouver des groupes ou items similaires

Règles d'association: Trouver des relations (liaisons) entre les items

....

Apprentissage non supervisé

Types des problèmes de Data Mining

- **Database**
- Trouver tous les demandeurs de crédits ayant le nom "Mehenni"
- Identifier les clients ayant acheté plus de 10,000 DA le mois précédent
- Trouver tous les clients ayant acheté du lait
- **Data Mining**
- Trouver tous les demandeurs de crédit qui ont un risque de crédit mineur. (**classification**)
- Identifier les clients qui ont des habitudes d'achat similaires. (**Clustering**)
- Trouver tous les items qui sont fréquemment achetés avec le lait (**règles d'association**)

Types de problèmes de Data Mining

- **Chapitre 4 : Clustering**
- **Chapitre 5 : Règles d'association**
- **Chapitre 6 : Arbres de décision**

