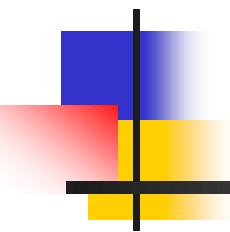


Fouilles de Données

Introduction à la fouille de données (partie 1)



Par Dr Azizi Nabiha

Professeur, Université Badji Mokhtar –
Annaba-



Données et information

- Les entreprises produisent des quantités énormes de données
 - Sources: affaires, science, géographie, environnement...
- Les données sont des ressources de valeur
- Besoin de techniques pour analyser les données et extraire des informations / connaissances automatiquement
 - Données = faits
 - Information = modèle / motif (pattern) dans les données

Data science – A la croisée de 3 profils

Connaître et comprendre les techniques de modélisation, d'analyse de données, d'inférence... savoir exploiter les régularités « cachées » dans les données, pourvoyeuses de connaissances

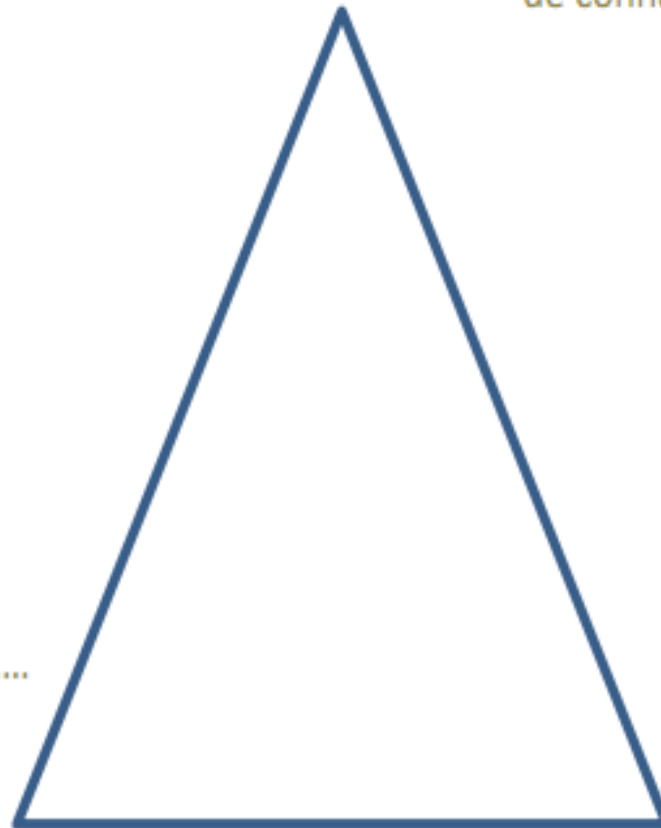
STATISTICIEN

Maîtriser les outils pour manipuler les données, développer des stratégies nouvelles pour gérer la profusion de l'information,...

INFORMATICIEN

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

CONNAISSANCES METIER



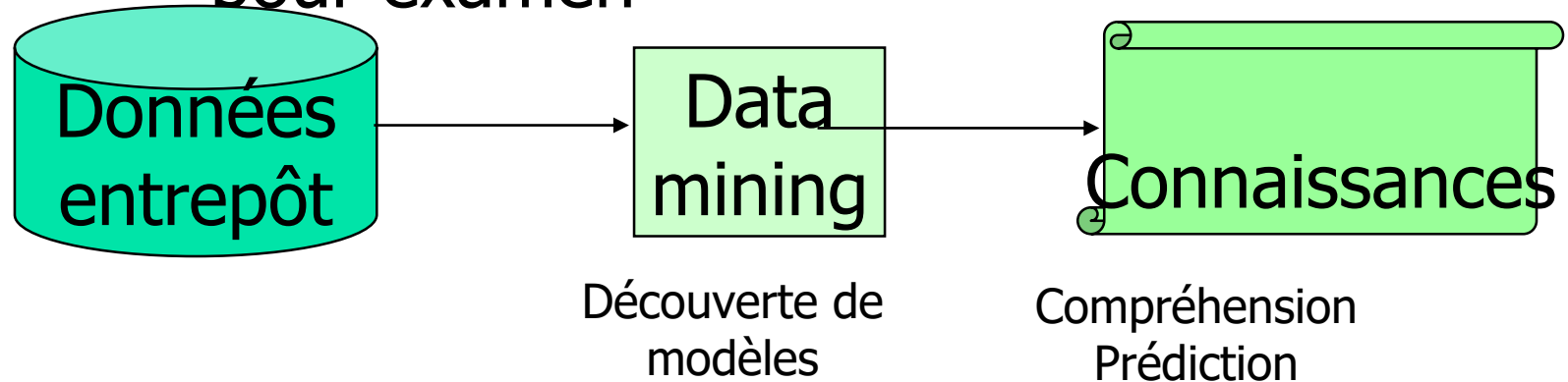


Fouille de données (FDD)

- Fouille / découverte de connaissances dans les données ou création d'un modèle des données
- Processus qui emploie des techniques d'apprentissage automatiques et intelligentes pour analyser et extraire des connaissances, de grandes quantités de données
- 98% de ce que les humains apprennent proviennent de reconnaissance des formes [Kurzweil]

Définition

- Data mining
 - ensemble de techniques d'exploration de données afin d'en tirer des connaissances (la signification profonde) sous forme de modèles présentés à l'utilisateur averti pour examen

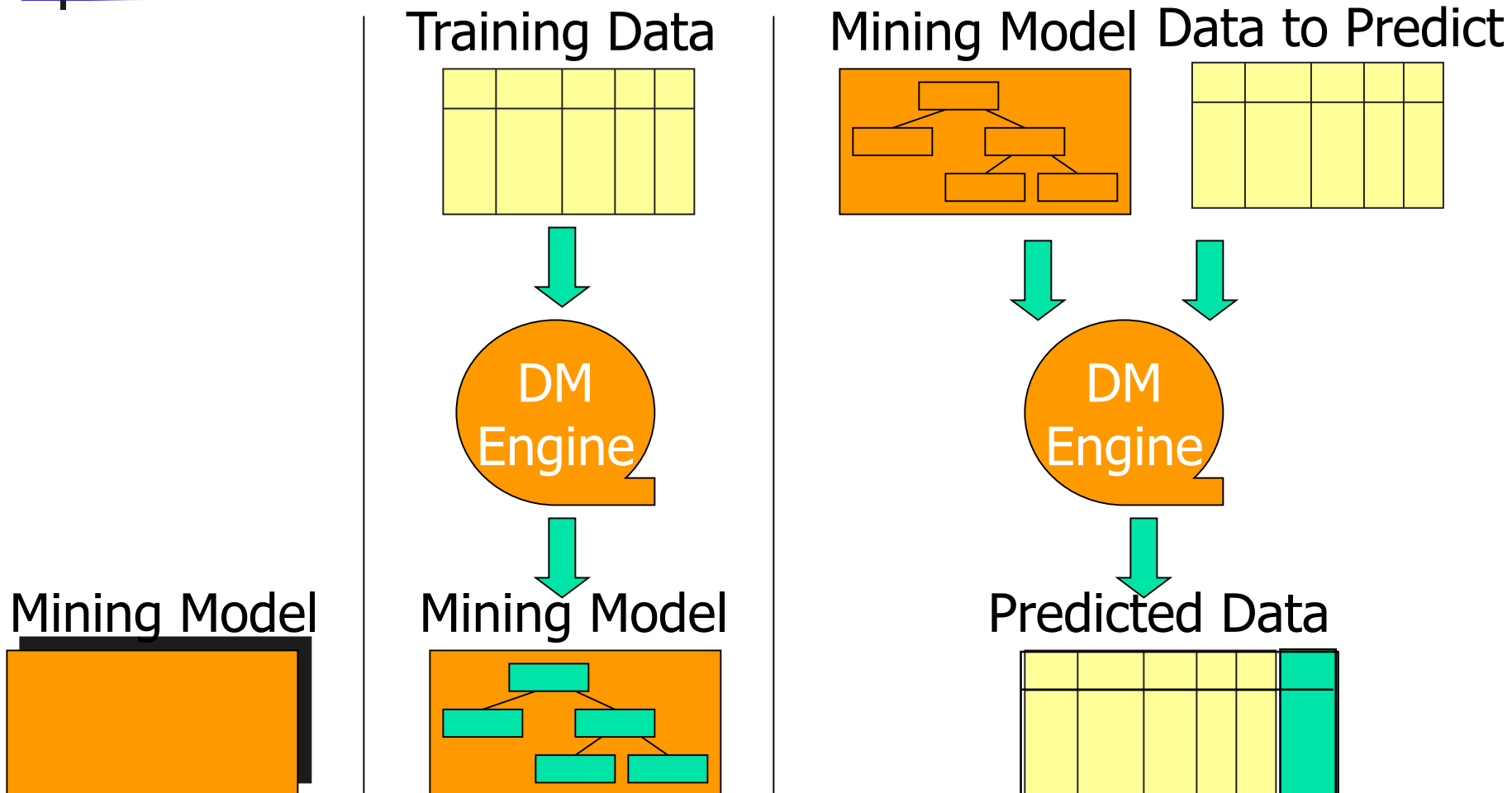




Extraction de Connaissances dans les Bases de Données

- extraction non triviale de connaissances implicites, inconnues au préalable, intéressantes, utiles à partir d'informations stockées dans des bases de données
- Quels types de connaissances ?
 - **Champ de recherche multidisciplinaire**
 - Bases de Données (BD)
 - Statistiques
 - Apprentissage Statistique / Symbolique
 - Visualisation

Découverte et Exploitation





Pourquoi la FDD maintenant?

- Des machines plus puissantes
- Existence d'algorithmes de fouille de données
- Collections et sauvegardes des données améliorées
- Différents domaines: base de données, statistiques, intelligence artificielle, visualisation, parallélisme...
- Domaine pluridisciplinaire

Pourquoi cette émergence ?

- un univers de données en plein expansion, mais peu de connaissances
- de nombreuses bases de données, de plus en plus volumineuses
- connexion des BDs par les réseaux Internet
- besoin vital d'outils d'aide à la décision pour interpréter les données
- développement des techniques d'apprentissage automatique

Les méthodes selon les finalités

Description :

trouver un résumé des données qui soit plus intelligible

- statistique descriptive
- analyse factorielle

Ex : moyennes conditionnelles, etc.

Structuration :

Faire ressurgir des groupes « naturels » qui représentent des entités particulières

- **classification** (clustering, apprentissage non-supervisé)

Ex : découvrir une typologie de comportement des clients d'un magasin

Méthodes de Data Mining

```
graph TD; A[Méthodes de Data Mining] --> B[Description]; A --> C[Structuration]; A --> D[Explication]; A --> E[Association];
```

Explication :

Prédire les valeurs d'un attribut (endogène) à partir d'autres attributs (exogènes)

- régression
- **apprentissage supervisé**

Ex : prédire la qualité d'un client (rembourse ou non son crédit) en fonction de ses caractéristiques (revenus, statut marital, nombre d'enfants, etc.)

Association :

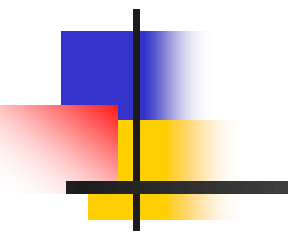
Trouver les ensembles de descripteurs qui sont le plus corrélés

- **règles d'association**

Ex : rayonnage de magasins, les personnes qui achètent du poivre achètent également du sel

Les méthodes sont le plus souvent complémentaires !

La démarche DATA MINING

- 
- formalisation des objectifs
 - acquisition des données
 - préparation des données
 - apprentissage – application des méthodes
 - interprétation – explication
 - évaluation et validation
 - déploiement