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

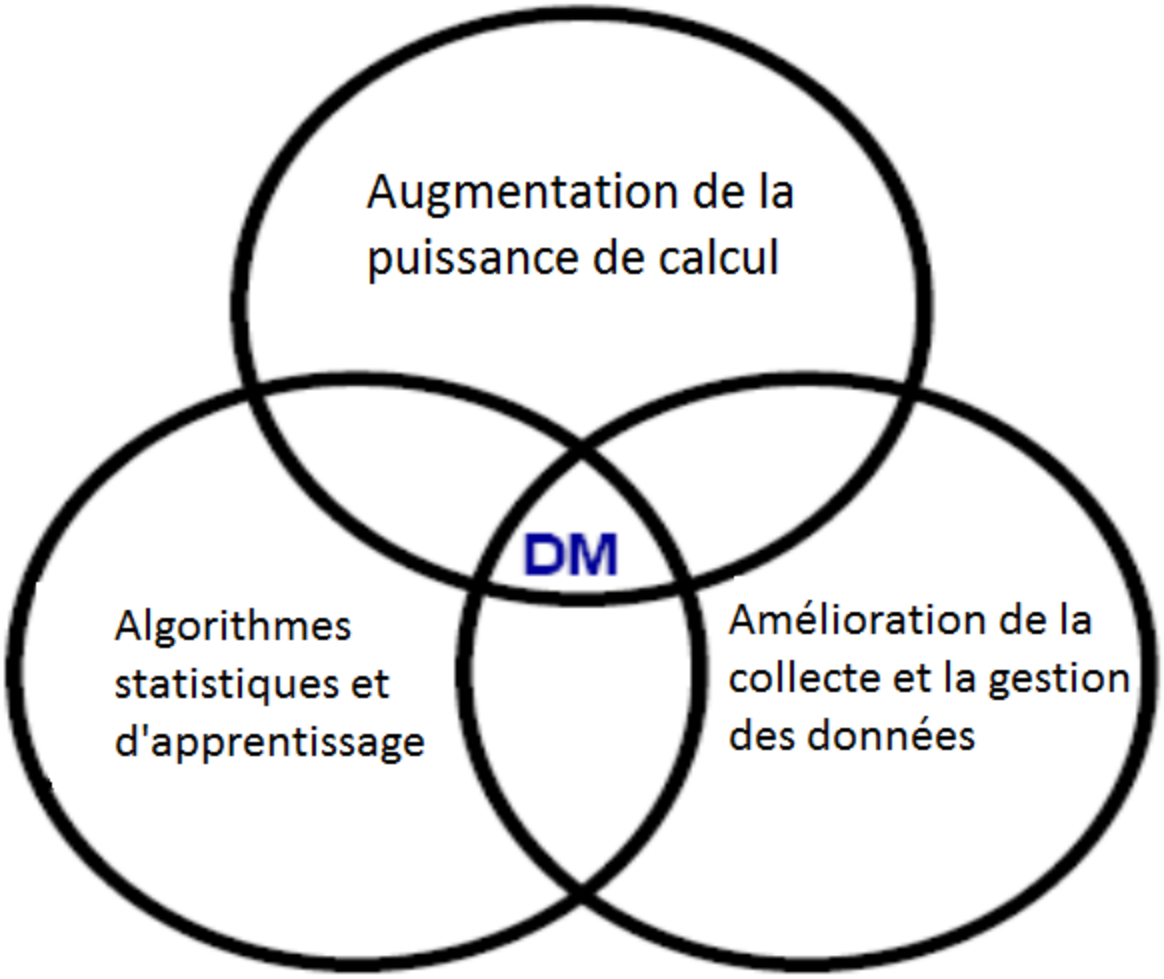
Données

- Données non opérationnelles

Sortie

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

Convergence de 3 Technologies clés



1. Augmentation de la puissance de calcul

— Loi de Moore: la puissance de calcul double chaque 18 mois

— Stations de travail puissantes: de plus en plus fréquentes

— Serveurs rentables fournissent un traitement parallèle au marché grand public

1. Explosion des données

- Le taux de création des données est en accélération chaque année. En 2003, UC Berkeley a estimé que l'année précédente (2002) a généré 5 exaoctets de données, dont 92% sont stockée sur des média électroniques.
- Mega < Giga < Tera < Peta < Exa ... Toutes les données des livres de la bibliothèque du Congrès américain sont. ~136 Teraoctets
- VLBI Telescopes produit 16 Gigaoctets de données chaque seconde.
- Google effectue sa recherche dans 18 billions+ de pages web.

1. Implications de l'explosion des données

- **Lorsque le volume des données augmente, la proportion d'information décroit.**
- **Comme les données sont de plus en plus générées automatiquement, on a besoin de solutions automatiques pour transformer ces données brutes en information.**
- **Les compagnies ont besoin de tirer profit des leurs données stockées ... Sinon, pourquoi doivent-elles les stocker?**

2. Amélioration de la collecte et la gestion des données

Jan	Feb	Mar	Apr	May	Jun
12.0	18.0	19.0	7.0	17.0	6.0
3.0	5.0	2.0	7.0	5.0	25.0
6.0	1.5	2.4	2.1	6.9	12.3

— **Collecte des données? Accès ? Navigation ? Mining**
— **Plus on a les données plus c'est bon! (généralement)**

3. Algorithmes de statistiques et d'apprentissage

— Les techniques ont souvent attendu les technologies de calcul pour les saisir

— Les Statisticiens réalisaient manuellement les tâches de data mining

— Un bon algorithme d'apprentissage automatique (machine learning) est une aplication intelligente des processus statistiques

— La recherche en data mining s'interesse à l'ajustement des techniques existantes pour réaliser un pourcentage modeste de gains

3.Donnée/Information/Connaissance/Sagesse

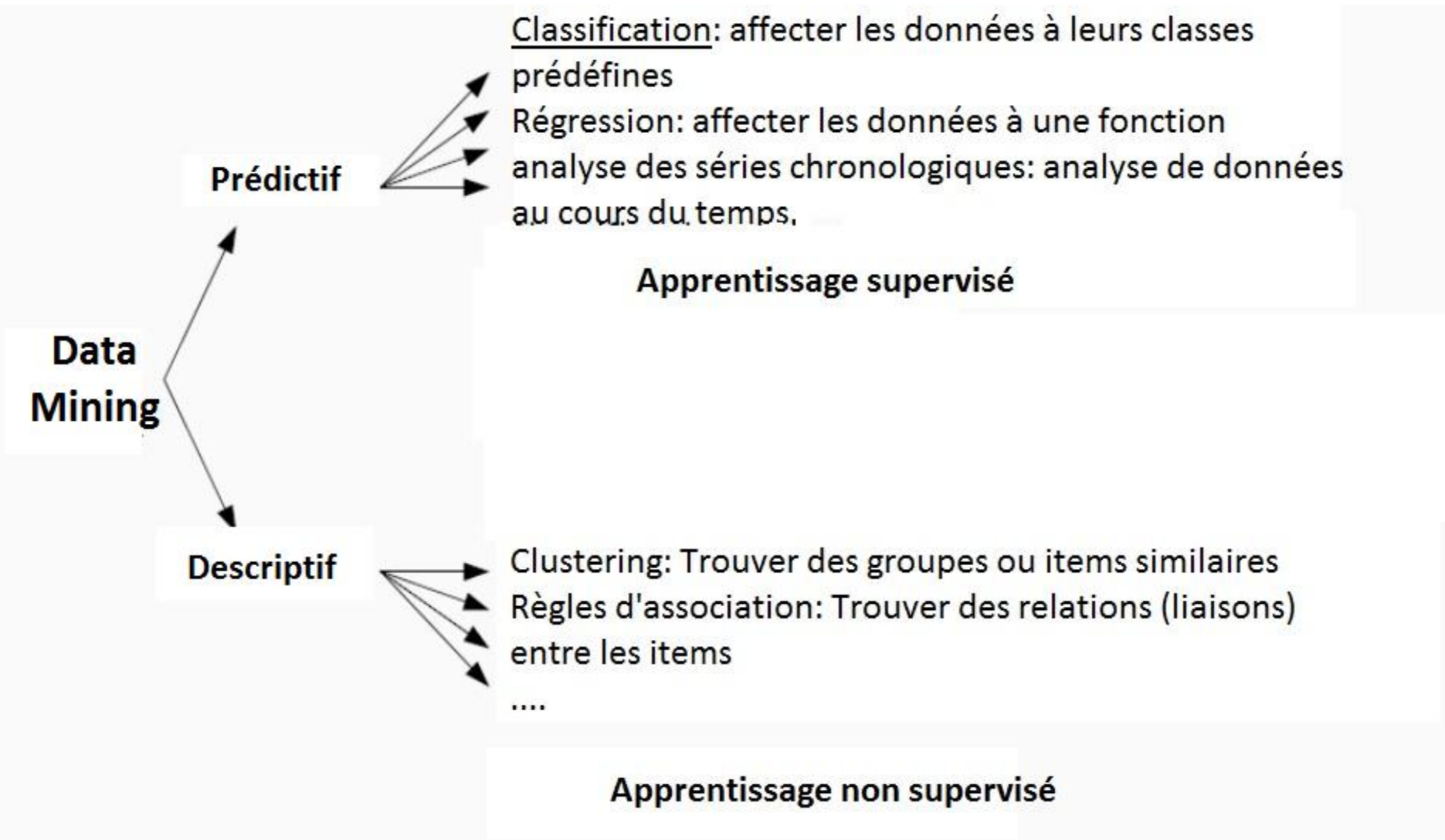
- **Par exemple, l'application du data mining peut nous dire qu'il y a une corrélation entre l'achat des magasines de musique et du café, mais ne peut pas nous dire coment exploiter cette connaissance. Devons-nous les arranger l'un à côté de l'autre pour renforcer la tendance, ou les arranger n'importe où, les gens vont les acheter ensemble peu importe leur endroit de stockage.**

- **Le Data mining peut aider les gestionnnaires à l'élaboration de strategies pour leur entreprises mais ne peut pas leur donner les strategies.**

Fonctions de Data mining

- Toutes les fonctions de Data Mining peuvent être considérées comme une tentative de trouver un modèle pour ajuster les données.
- Chaque fonction a besoin de critères pour créer un modèle plutôt qu'un autre.
- Chaque fonction a besoin d'une technique pour comparer les données.
- **Deux types de modèles:**
  - **Les modèles prédictives: Prédire une valeur inconnue sur la base de données connues**
  - **Les modèles descriptives: Identifier les patterns des données**

Fonctions de Data mining



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

- Classification
- Clustering
- Règles d’association