

Matière: *Fouille de Données (FD)*



Master Académique Informatique
Spécialité : **Intelligence Artificielle (IA)**
M2 Semestre3

Plan (Programme)

- ❑ Introduction à la fouille de données (définitions, processus, applications, modèles, ...)
- ❑ Rappels des statistiques (types de variables, représentations graphiques, ...)
- ❑ Classification automatique
- ❑ Recherche de règles d'associations
- ❑ Arbres de décision
- ❑ Réseaux Bayésiens
- ❑ Validation de la qualité des connaissances

Chapitre 1

Introduction à la fouille de données

Généralités sur les systèmes décisionnels

L'informatique décisionnelle :

On qualifie d'informatique décisionnelle (en anglais « Business intelligence », parfois appelé tout simplement « le décisionnel ») l'exploitation des données de l'entreprise dans le but de faciliter la prise de décision par les décideurs, c'est-à-dire la compréhension du fonctionnement actuel et l'anticipation des actions pour un pilotage éclairé de l'entreprise.

Le Contexte du décisionnel :

- **Besoin**: prise de décisions stratégiques et tactiques
- **Pourquoi**: besoin de réactivité
- **Qui**: les décideurs (non informaticiens)
- **Comment**: répondre aux demandes d'analyse des données, dégager des informations qualitatives nouvelles ...etc.

Problématique et Solution:

Le **problème** posé est comment répondre aux demandes des décideurs en se basant sur les données à la base décrites précédemment en donnant un accès rapide et simple à l'information stratégique et en donnant du sens aux données?

La **solution** consiste à mettre en place un système d'information dédié aux applications décisionnelles: généralement c'est les **entrepôts de données (ED) ou data warehouses (DW)** sur lesquels sont appliquées des techniques de fouille et d'extraction de données : c'est le domaine du **Data Mining**.

Systèmes transactionnels

« Le système transactionnel (opérationnel) est généralement une base de données, développée par application, stockant les **données courantes** d'une organisation, c'est-à-dire qu'il n'y a pas de données d'archives dans les systèmes transactionnels » (*Bédard et al. 1997*)

Remarque : ce type de système est appelé aussi :

ODS : **O**perational **D**ata **S**to**r**e

Systèmes transactionnels

Le **transactionnel** réfère à un mode d'exploitation de données tourné vers la saisie, le stockage, la mise à jour, la sécurité et l'intégrité des données.

Par exemple:

- ❑ les systèmes de gestion des transactions boursières ou bancaires;
- ❑ les systèmes d'inventaire dans les magasins;
- ❑ les réservations d'hôtel;
- ❑ ... etc.

Systèmes transactionnels

La plupart des systèmes transactionnels sont implantés selon une **structure relationnelle normalisée** :

- Redondance minimum
- Intégrité des données
- Facilité de mise à jour

Systèmes transactionnels et analyse

Les systèmes transactionnels ont des obstacles d'analyse du fait de leur forte normalisation, en effet nous avons:

- Un Nombre élevé de tables donc un nombre élevé de jointures nécessaires entre les tables (performance pauvre)
- Un Temps de traitement long
- Une Élaboration complexe des requêtes

Systèmes transactionnels et analyse

De plus, les types d'analyses servant aux processus de décision des organisations nécessitent :

- Données sommaires (agrégées ou résumées) sur l'ensemble de l'organisation (provenant des différentes BD dispersées de l'organisation et intégrées)
- Données historiques
- Réponses rapides (requêtes surtout de type agrégatif), interfaces à l'utilisateur faciles à utiliser

Conclusion :

On a Besoin de systèmes dédiés à l'analyse, c'est **les entrepôts de données**.

Définition d'un entrepôt de données :

La notion d'entrepôt de données a été introduite par Bill Inmon(1996) qui lui a donné la définition suivante:

Un entrepôt de données est une collection de données **orientées sujet**, **intégrées**, **non volatiles** et **historisées**, organisées pour le support d'un processus d'aide à la décision.

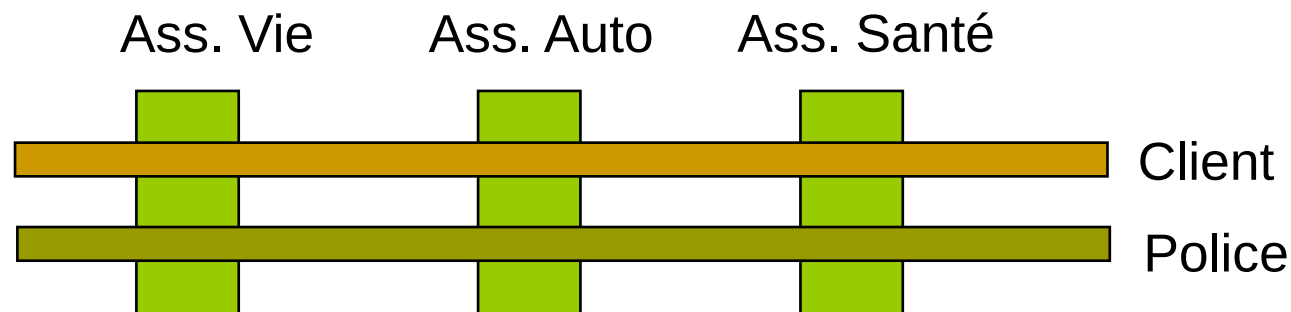
A data warehouse is a “**subject-oriented, integrated, time-variant, and nonvolatile** collection of data in support of management's decision-making process.”—W. H. Inmon

Inmon → Bill William H. Inmon

Définition d'un DW : Les données orientées sujet

- Organisées autour de sujets majeurs comme consommateur, produit, ventes
- Données pour l'analyse et la modélisation en vue de l'aide à la décision, et non pas pour les opérations et transactions journalières
- Vue synthétique des données selon les sujets intéressant les décideurs

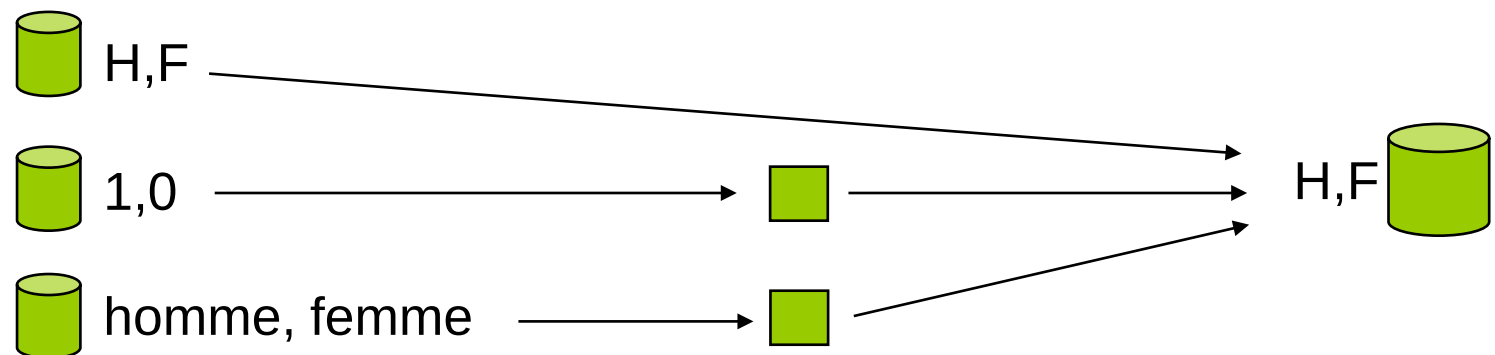
Exemple :



Définition d'un DW : Les données intégrées

- Construites en intégrant des sources de données multiples et hétérogènes
- Nettoyées et intégrées
 - ✓ Consistence dans les noms des champs, le codage des données issues de plusieurs sources.
 - ✓ La conversion se fait quand les données sont transférées dans le DW.

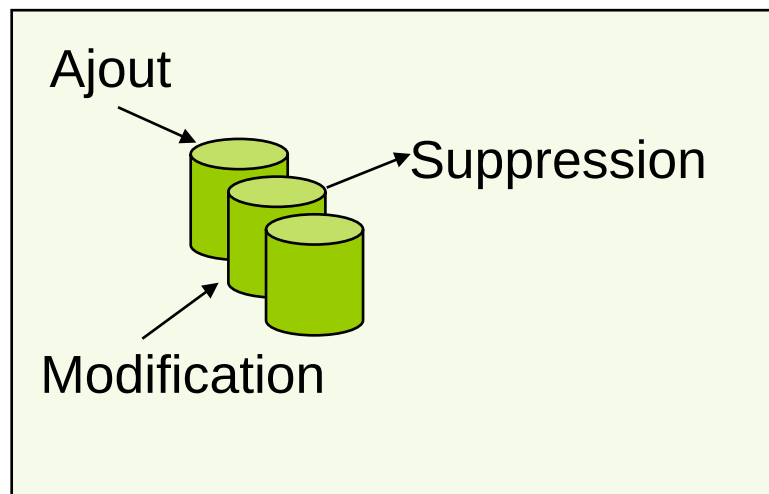
Exemple :



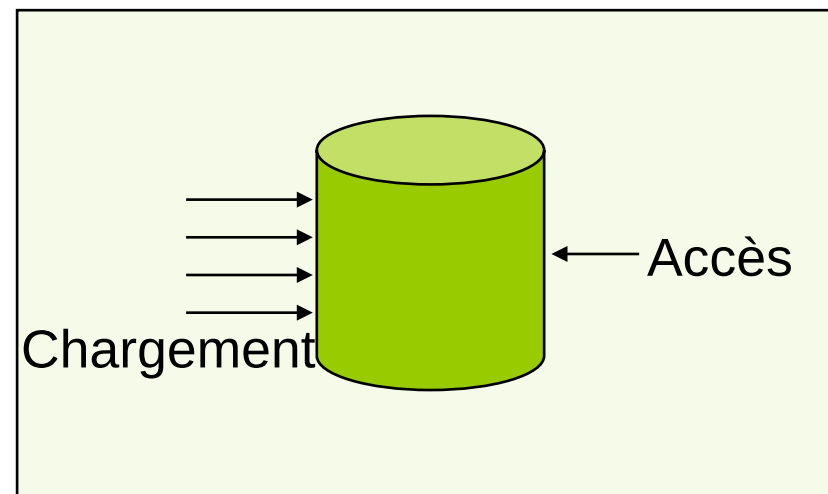
Définition d'un DW : Les données non volatiles

- Stockage indépendant des BD opérationnelles
- Pas de mises à jour des données dans le DW, 2 actions sont possibles sur le DW:
 - Alimentation du DW à partir des données des BD opérationnelles
 - Accès (lecture) de ces données

Bases de production



Entrepôts de données



Définition d'un DW : Les données historisées

- L'échelle de temps dans le DW est beaucoup plus longue que dans les BD
 - BD opérationnelle: valeur courante des données
 - DW: information dans une perspective historique (ex: les 5 dernières années)
- Chaque structure dans le DW contient un élément décrivant le temps

Définition d'un DW : Les données historisées

Exemple :

Base de
production

Image de la base en Mai 2008

Répertoire

Nom	Ville
Med	Alger
Yahia	Jijel

Image de la base en Juin 2009

Répertoire

Nom	Ville
Med	Oran
Yahia	Jijel

Entrepôt
de
données

Calendrier

Code	Année	Mois
1	2008	Mai
2	2009	Juin

Répertoire

Code	Nom	Ville
1	Med	Alger
1	Yahia	Jijel
2	Med	Oran

Définition d'un DW :

Remarques :

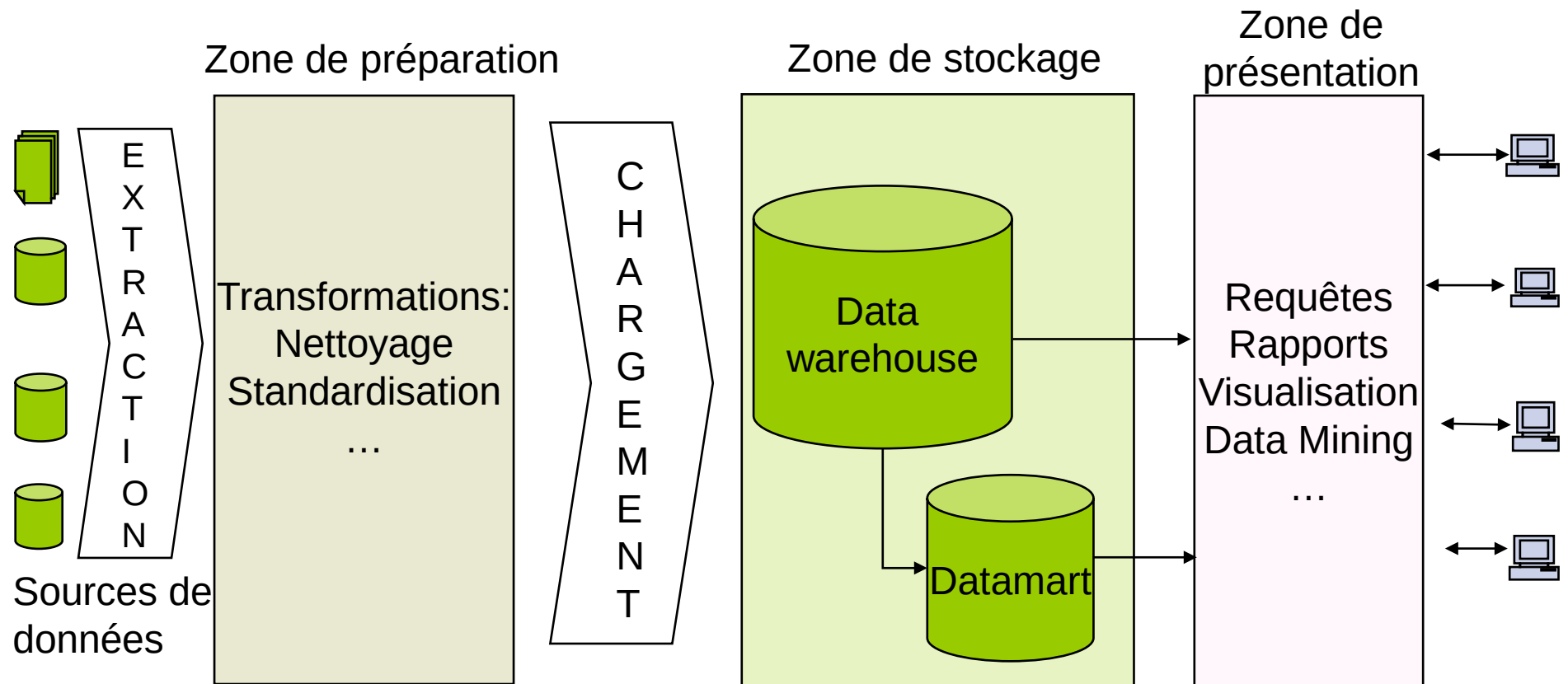
- Le Data warehouse : Ce n'est pas un produit c'est une solution.
- Le data warehouse ne s'achète pas il se construit.

Définition : Data Warehousing

Entreposage de données (Data warehousing) :

- C'est le processus de construction et d'utilisation d'un entrepôt de données.

Architecture générale du DW



Les différentes zones de l'architecture

1. Zone de préparation (Staging area):

- ❑ Zone temporaire de stockage des données extraites
- ❑ Réalisation des transformations avant l'insertion dans le DW:
 - Nettoyage : Gestion des erreurs de saisie, de l'intégrité des domaines (les dates par exemple), des Données manquantes
 - Format **Exemple**: Type des données
 - Consolidation **Exemple**: Choix des unités et des représentations
 - Uniformisation d'échelle
Exemple: Homogénéisation des échelles
- ❑ Données souvent détruites après chargement dans le DW

Les différentes zones de l'architecture

2. Zone de stockage (DW, DM)

- ☐ On y transfère les données nettoyées
- ☐ Stockage permanent des données

Les différentes zones de l'architecture

3. Zone de présentation

- ❑ Donne accès aux données contenues dans le DW
- ❑ Peut contenir des outils d'analyse programmés:
 - Outils de création de Rapports
 - Outils OLAP (Analyse et les Requêtes)
 - Outils de recherche d'information (**Data mining**)
 - ... etc.

Notion de Méta données(Metadata):

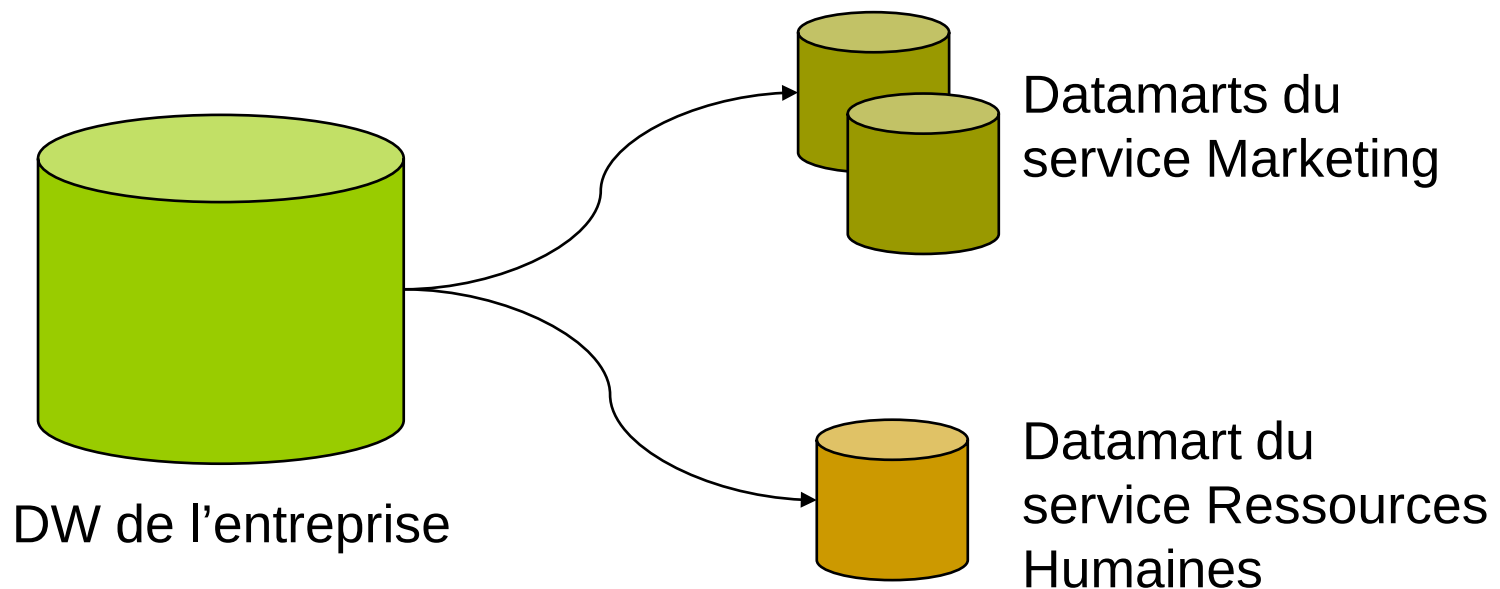
C'est une description des données:

- Elles permettent de connaître le contenu du DW ;
- Ainsi de faciliter la recherche de données;
- Elles contiennent les informations sur l'ensemble du DW.

Notion de Magasin de données (Datamart):

Un **Magasin de données** est un petit entrepôt de données ciblé sur un sujet :

- ▣ Sous-ensemble d'un entrepôt de données
- ▣ Destiné à répondre aux besoins d'un secteur ou d'une fonction particulière de l'entreprise
- ▣ Point de vue spécifique selon des critères métiers



Entrepôt de données Vs Base de données :

La base de données classique supporte des requêtes transactionnelles de type **OLTP** (**O**n**L**ine **T**ransaction **P**rocessing). A la différence de ça un entrepôt de données est conçu pour supporter des requêtes d'analyse et de prise de décision de type **OLAP** (**O**n**L**ine **A**nalytical **P**rocessing).

Le tableau suivant montre la différence entre un système doté d'une base de données classique et un entrepôt de données:

Entrepôt de données Vs Base de données :

	Systèmes opérationnels	Entrepôt de données
Type de requêtes	OLTP	OLAP
But	Exécution d'un processus métier	Evaluation d'un processus métier
Interaction / l'utilisateur	Insertion, modification, interrogation, supp.	Interrogation
Données	Courantes	Courantes et historiques
Usage	Support de l'opération de l'entreprise	Support de l'analyse de l'entreprise
Requêtes	Simple, prédéterminées	Complexe, Ad-hoc
Type utilisateur	Employé	Décideur
Conception	Troisième forme normale	Multidimensionnelle

Représentation d'un ED :

La représentation d'un entrepôt de données est vue selon trois niveaux qui sont:

- **Représentation conceptuelle**

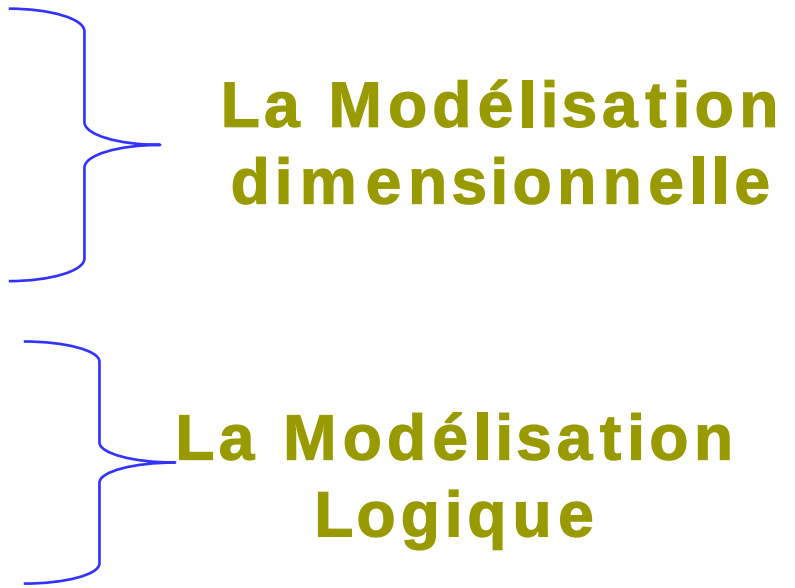
- schémas en étoile
- schémas en flocons
- schémas en constellation

- **Représentation logique**

- tables
- Cubes

- **Représentation physique**

- vues matérialisées



La Modélisation dimensionnelle

La Modélisation Logique

Modélisation des DW :

- ❑ Nouvelle méthode de conception autour des concepts métiers
 - Ne pas normaliser au maximum
- ❑ Introduction de nouveaux types de table:
 - Table de faits
 - Table de dimensions
- ❑ Introduction de nouveaux modèles:
 - Modèle en étoile
 - Modèle en flocons
 - Modèle en constellation

Table de faits :

- ❑ Table principale du modèle dimensionnel
- ❑ Contient les données observables (les faits) sur le sujet étudié selon divers axes d'analyse (les dimensions)

Exemple:

Clés étrangères
vers les
dimensions

Faits

Table de faits des ventes	
{	Clé date (CE)
	Clé produit (CE)
	Clé magasin (CE)
{	Quantité vendue
	Coût
	Montant des ventes

Table de dimension :

Elle représente l'axe d'analyse selon lequel vont être étudiées les données observables (faits). Elle Contient donc le détail sur les faits.

Exemple :

Clé de substitution

Attributs de la
dimension

{

{

{

{

{

Dimension produit
Clé produit (CP)
Code produit
Description du produit
Famille du produits
Marque
Emballage
Poids

La table de dimension contient souvent un grand nombre de colonnes mais elle contient en général beaucoup moins d'enregistrements qu'une table de faits

Table de dimension :

Remarque:

La dimension temps est commune à l'ensemble du DW.
Elle est reliée à toute table de faits.

Dimension Temps
Clé temps (CP)
Jour
Mois
Trimestre
Semestre
Année
Num_jour_dans_année
Num_semaine_ds_année

Le modèle (Schéma) en étoile (Star) :

C'est un modèle d'ED basé sur une table de fait centrale et des dimensions. Les dimensions n'ont pas de liaison entre elles.

Avantages:

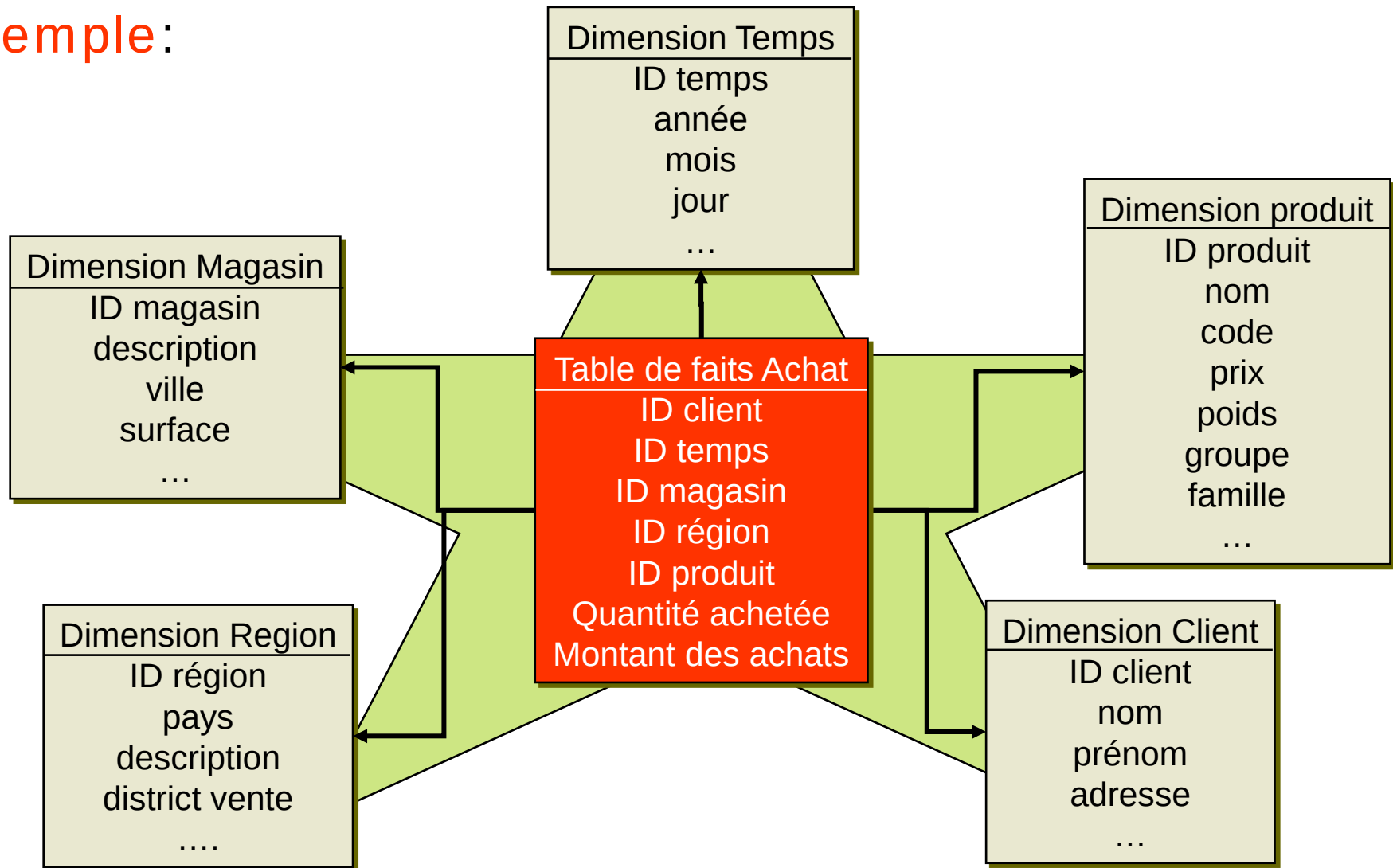
- Facilité de navigation
- Performances, nombre de jointures limité
- Fiabilité des résultats

Inconvénients:

- Redondance dans les dimensions
- Toutes les dimensions ne concernent pas les mesures
- Alimentation complexe

Le modèle en étoile :

Exemple :



Le modèle (schéma) en flocons (Snow-Flake) :

C'est un modèle basé sur une table de fait et des dimensions décomposées en sous hiérarchies. On a un seul niveau hiérarchique dans une table de dimension. La table de dimension de niveau hiérarchique le plus bas est reliée à la table de fait. On dit qu'elle a la granularité la plus fine

Avantages:

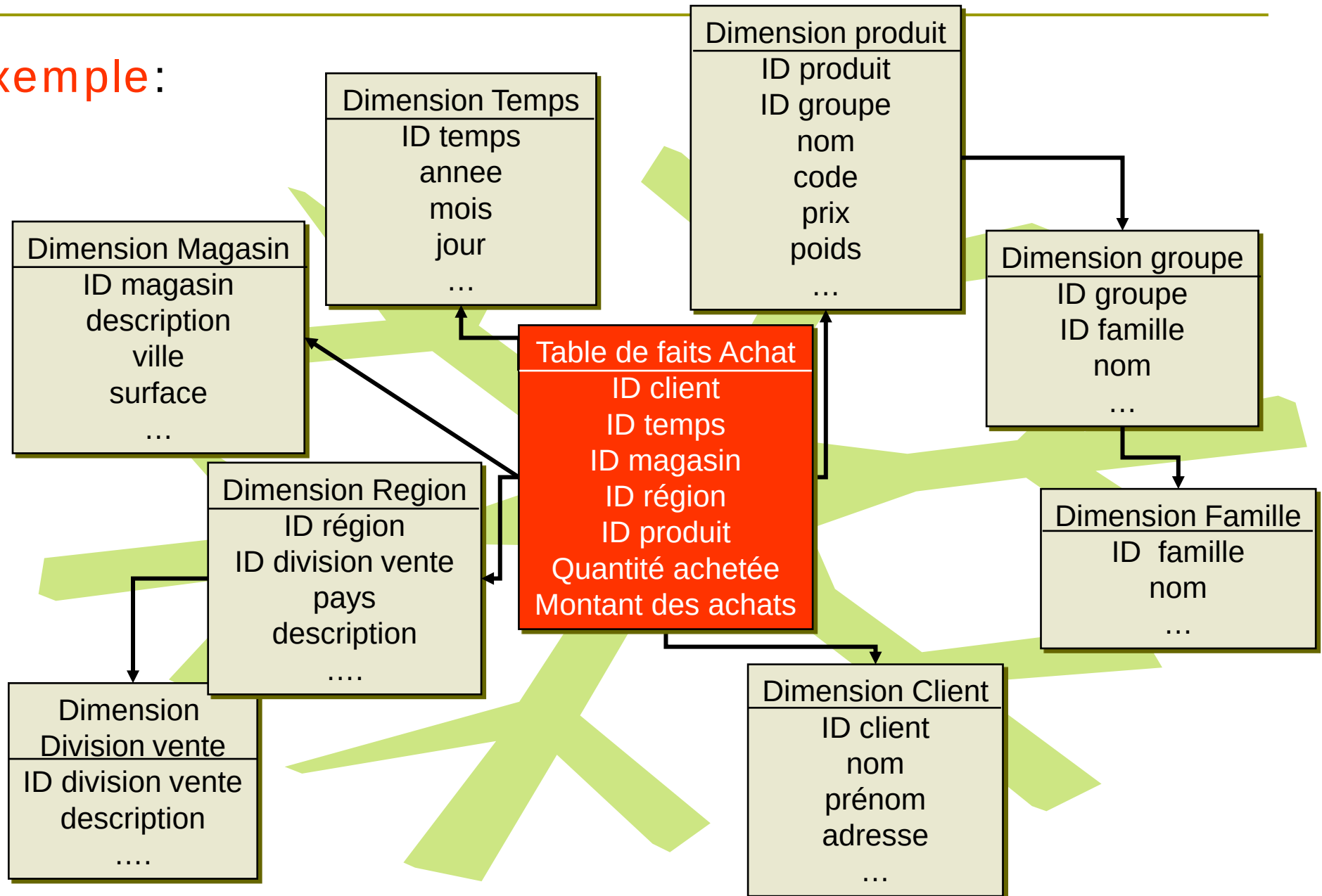
- ❑ Normalisation des dimensions
- ❑ Économie d'espace disque

Inconvénients:

- ❑ Modèle plus complexe (jointure)
- ❑ Requêtes moins performantes

Le modèle en flocons :

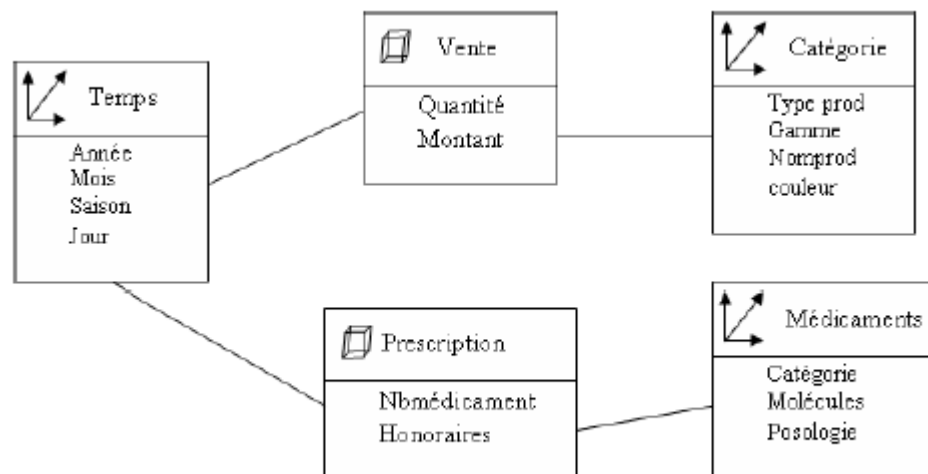
Exemple :



Le modèle (schéma) en constellation (Fact-flaked):

Ce modèle est basé sur plusieurs tables de faits pour décrire plusieurs séries de données sur le métier étudié et partageant les tables dimensionnelles.

Exemple:



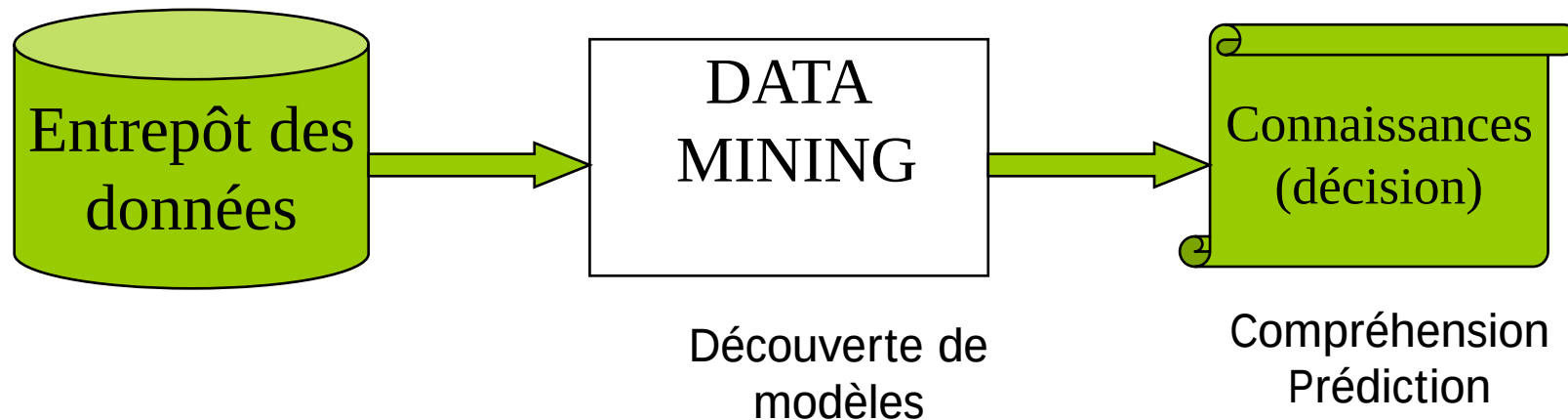
Processus global de fouille de données

Introduction:

- 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

Définition : data mining

Le **Data Mining**, ou **fouille de données**, est l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de données informatiques, afin d'en tirer des connaissances, des règles, des associations, des structures particulières restituant l'essentiel de l'information.



Remarque :

Le terme «*data mining*», ou «fouille de données», a été habituellement utilisé par les statisticiens, les analystes de données, et la communauté de systèmes de gestion de l'information, son équivalent est le **KDD** ou **ECD** qui a été principalement utilisé par les chercheurs en intelligence artificielle et apprentissage automatique.

Définition : ECD ou KDD

- ECD (Extraction de Connaissances à partir de Données)

C'est le processus d'extraction d'informations intéressantes (**non triviales, implicites, préalablement inconnues et potentiellement utiles**) à partir de grandes bases de données.

- Autres appellations:

- KDD (Knowledge Discovery from Databases)
- Analyse de données/patterns (motifs), business intelligence, etc...

Remarque :

- Le terme **KDD** (ou **ECD**), est utilisé depuis 1989 pour signifier le processus d'extraire de la connaissance à partir des données, et pour souligner le «*high-level*» dans l'application de méthodes de fouille de données (*data mining*)

KDD & Data mining

Extraction de connaissance à partir de données

(Knowledge Discovery in Databases – KDD) :

- ❑ cycle de découverte d'information regroupant la conception de grandes bases de données ou entrepôts de données (Data Warehouse)
- ❑ tous les traitements à effectuer pour extraire de l'information des données
- ❑ L'un de ces traitement est la Fouille de données (Data Mining)

Fouille de données (Data Mining) :

Ensemble de techniques d'exploration de données permettant d'extraire d'une base de données des connaissances sous la forme de modèles de description afin de :

décrire le comportement actuel des données et/ou

prédire le comportement futur des données

KDD & Data mining

- **KDD** pour signifier l'ensemble du processus
 - En effet *KDD* désigne *l'ensemble* du processus d'extraction de connaissances utiles à partir des données
- Tandis que **data mining** se réfère
 - à l'application d'algorithmes pour l'extraction de formes à partir des données,
 - sans considérer les étapes où il est question d'incorporer la connaissance du domaine et
 - de l'interprétation des résultats

Type de connaissances extraites :

- Quel type de connaissances à extraire par l'ECD?
 - Règles d'associations
 - Règles de classification
 - Clusters
 - Formules
 - Descriptions

Architecture typique d'un système KDD/DM

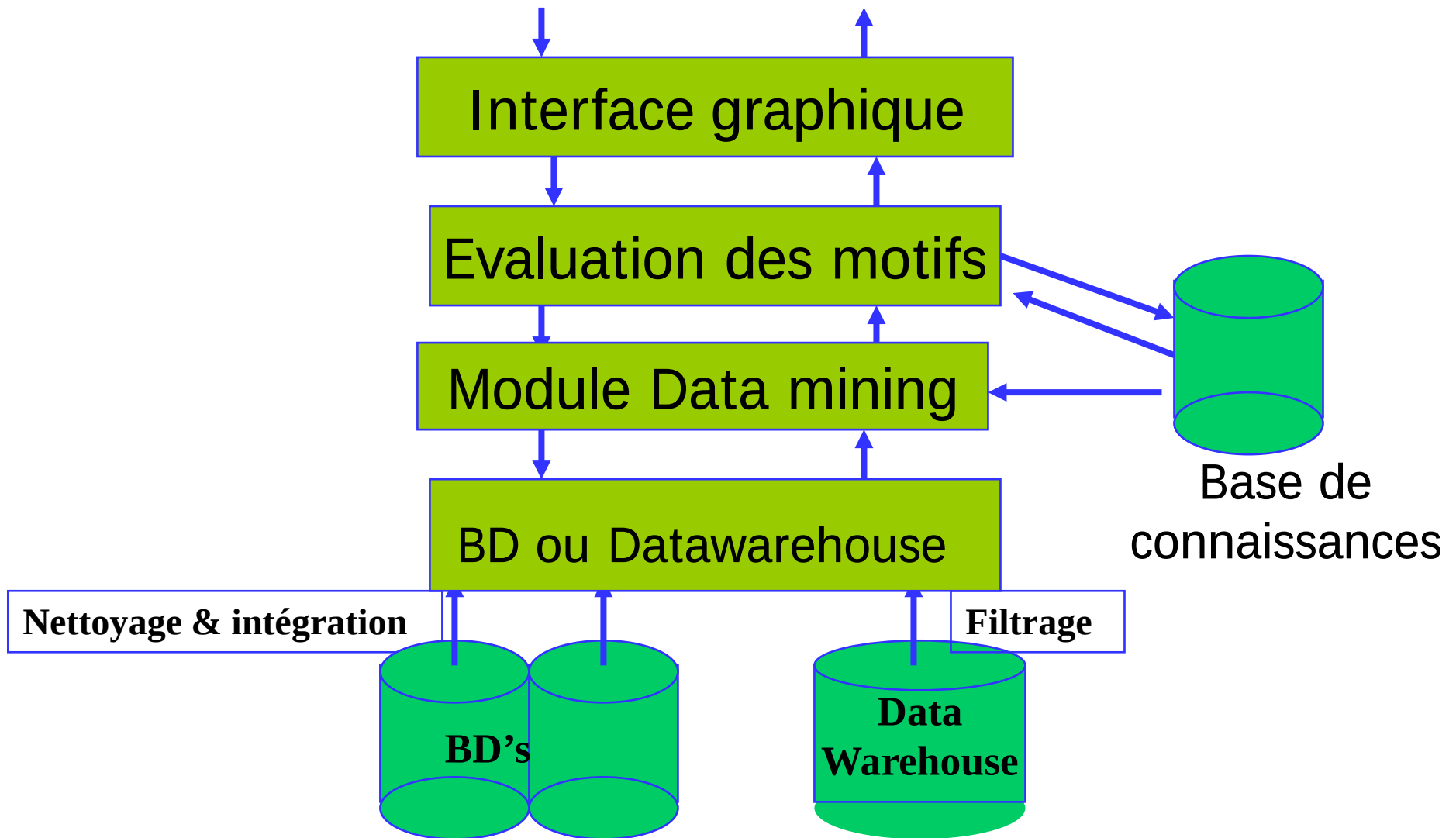
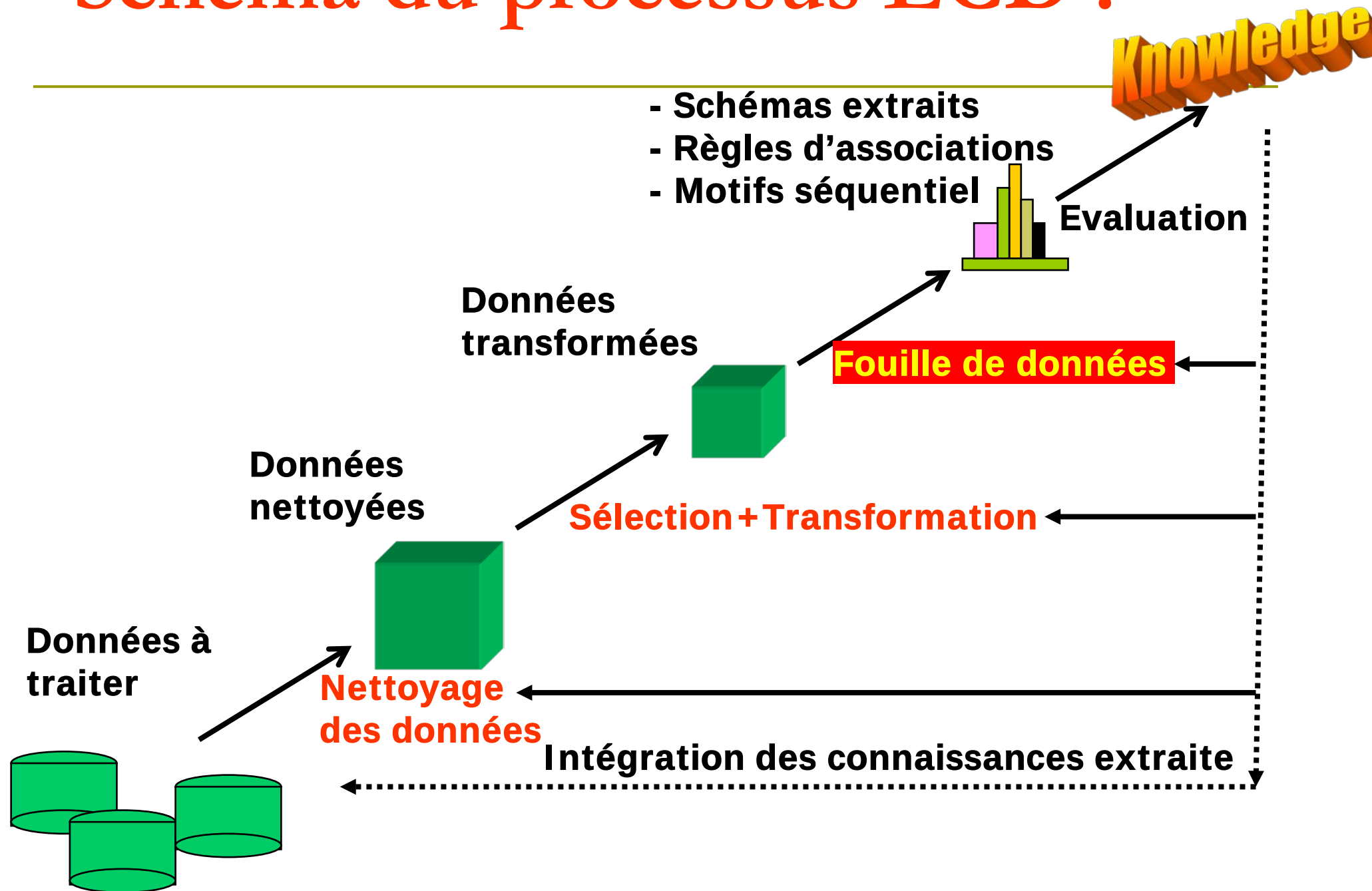


Schéma du processus ECD :



Etapes de l'ECD :

1. Compréhension du domaine d'application
2. Création du fichier cible (target data set)
3. Traitement des données brutes (data cleaning and preprocessing)
4. Réduction des données (data reduction and projection)
5. Définition des tâches de fouille de données
6. Choix des algorithmes appropriés de fouille de données
7. Fouille de données (data mining)
8. Interprétation des formes extraites (mined patterns)
9. Validation des connaissances extraites

Etapes de l'ECD :

Remarque :

Les étapes précédentes d'un processus ECD peuvent se résumer en trois grandes phases:

- ☐ Phase 1 : Prétraitements
- ☐ Phase 2 : Fouille de données
- ☐ Phase 3 : Post-traitements

Etapes de l'ECD :

Phase 1 : Prétraitements

▶ Recueil

- ✓ Connaître le domaine d'application

▶ Sélection

- ✓ faire des échantillons représentatifs de la population d'origine
- ✓ Épurier les données et éliminer le bruit et les redondances

▶ Transformation

- ✓ assure la bonne correspondance des formats entre la base de données d'origine et le traitement de cette base par le Data Mining.
- ✓ Fusion
- ✓ Agrégation

Etapes de l'ECD :

Phase 2 : Fouille de données

- ✓ arbres et lois de décision
- ✓ modèles linéaires
- ✓ modèles non linéaires (réseaux de neurones)
- ✓ méthodes de description probabilistes graphiques (réseaux bayesiens)
- ✓ méthodes relationnelles,
- ✓ etc.

Etapes de l'ECD :

Phase 3 : Post-traitements

► Visualisation

- ✓ graphiques 2-D et 3-D,
- ✓ représentations hiérarchiques
- ✓ règles en langage naturel,
- ✓ etc.

► Interprétation & Intégration

- ✓ évaluer les échantillons des données obtenus
- ✓ transformer les informations utiles obtenues en termes compréhensibles par l'utilisateur.
- ✓ utiliser la connaissance découverte,
- ✓ implanter cette connaissance à l'intérieur de la base de connaissance du domaine analysé
- ✓ vérifier qu'il n'y a pas de conflit

Applications pour l'ECD/DM

- ❑ L'analyse d'une BD de transactions d'un supermarché permet d'étudier le comportement des clients :
 - réorganiser les rayons
 - Ajuster les promotions
- ❑ En VPC (Vente Par Correspondance), regrouper les clients selon certains critères :
 - Cibler les « mailings »
- ❑ L'analyse de données médicales :
 - Support pour la recherche
- ❑ L'analyse de données financières :
 - Prédire l'évolution des actions
 - Dresser des profils de clients

Applications pour l'ECD/DM

□ Détection de fraudes

- en santé, services de cartes de crédit, télécommunications, etc.

Approche

- Utiliser les données historiques pour construire des modèles de comportements frauduleux puis utiliser les techniques de datamining pour retrouver des instances similaires

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 pour l'ECD/DM

□ Astronomie

- Le laboratoire JPL a découvert 22 quasars (galaxies, astres) en utilisant les techniques de datamining

□ Web

- IBM a appliqué des algorithmes de data mining pour réorganiser leurs sites WEB afin de faciliter la navigation.
- Améliorer le WEB marketing

□ Social

- Données démographiques, votes, résultats des élections,

Applications pour l'ECD/DM

□ Informatique

- IHM,
- Réseau,
- Data-Warehouse, Data Mart,
- Internet (moteurs intelligent, profiling, text mining, ...)

Sous domaines DM

Web mining

- optimisation des sites
- meilleure connaissance des internautes
- croisement avec les bases de données de l'entreprise

Text mining

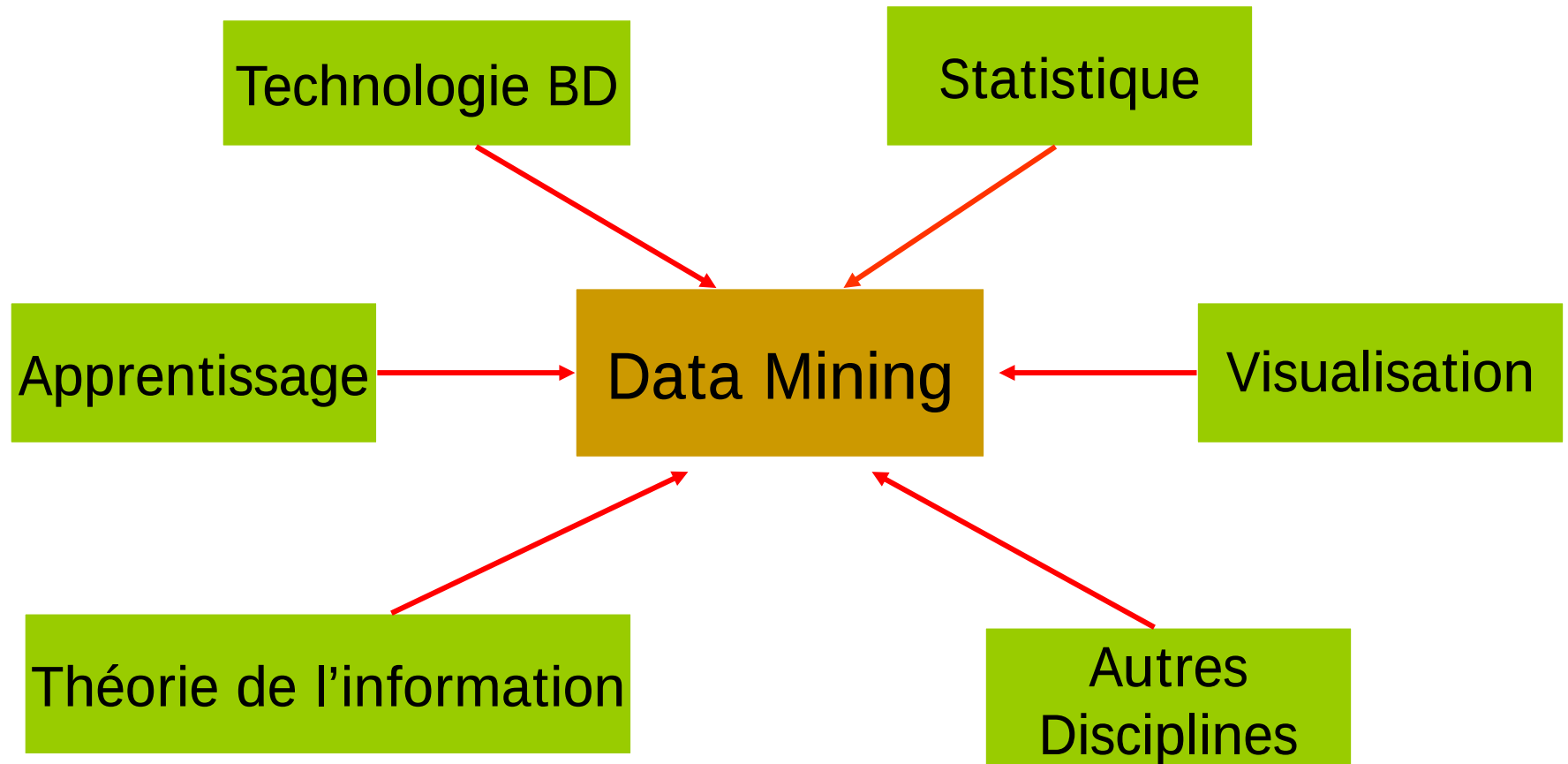
- statistique lexicale pour l'analyse des courriers, courriels, dépêches, comptes-rendus, brevets (langue naturelle)

Image mining

- reconnaissance automatique d'une forme ou d'un visage
- détection d'une échographie anormale,

Etc...

KDD/DM: Confluence de plusieurs Disciplines



Logiciels pour DM :

Logiciels commercialisés :

- S-PLUSTM de Insight,
- AliceTM de Isoft,
- Predict TM de Neuralware,

Logiciels gratuits :

- Weka
- Tanagra
- Orange
- R (version gratuite de S-PLUS)

Logiciels pour DM :

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 oeuvre qu'une ou deux techniques (excepté les produits S-PLUS, R, Tanagra et Weka)

Principaux Modèles

Typologies des méthodes de fouille de données :

- ❑ Typologie selon objectifs
- ❑ Typologie selon le type de modèle obtenu
- ❑ Typologie selon le type d'apprentissage
- ❑ Quelques méthodes de fouille de données

Typologies selon objectifs :

Types d'objectifs :

▣ **Classification :**

Consiste à examiner les caractéristiques d'un objet et lui attribuer une classe :

- Attribuer ou non un prêt à un client
- établir un diagnostic

▣ **Prédiction :**

Consiste à prédire la valeur future d'un attribut en fonction d'autres attributs :

- Prédire la "qualité" d'un client en fonction de son revenu, de son nombre d'enfant,...

▣ **Association :**

Consiste à déterminer les attributs qui sont corrélés :

- Analyse du panier de la ménagère: ex. Lait et Café

▣ **Segmentation :**

Consiste à former des groupes homogènes à l'intérieur d'une population

- Tâche souvent faite avant les précédentes pour trouver les groupes sur lesquels appliquer la classification

Typologies selon le type de modèle obtenu :

Types de modèles obtenus :

▣ **Modèles prédictifs :**

- utilisent les données avec des résultats connus pour développer des modèles permettant de prédire les valeurs d'autres données
- Ex: modèle permettant de prédire les clients qui ne rembourseront pas leur crédit
- utilisé principalement en classification et prédiction

▣ **Modèles descriptifs :**

- proposent des descriptions des données pour aider à la prise de décision
- aident à la construction de modèles prédictifs
- utilisés principalement en segmentation et association

Typologies selon le type de l'apprentissage :

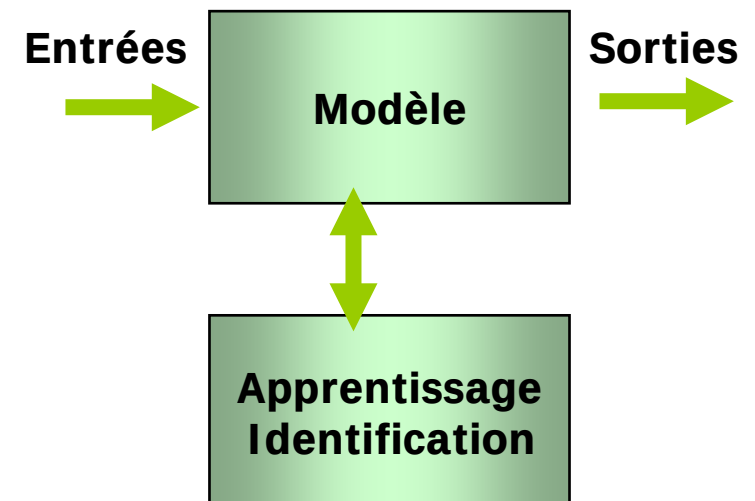
Types d'apprentissage utilisés dans les méthodes de fouille :

- ▣ **Apprentissage supervisé - Fouille supervisée**
- ▣ **Apprentissage non supervisé - Fouille non supervisée**

Typologies selon le type de l'apprentissage :

□ Apprentissage supervisé - Fouille supervisée :

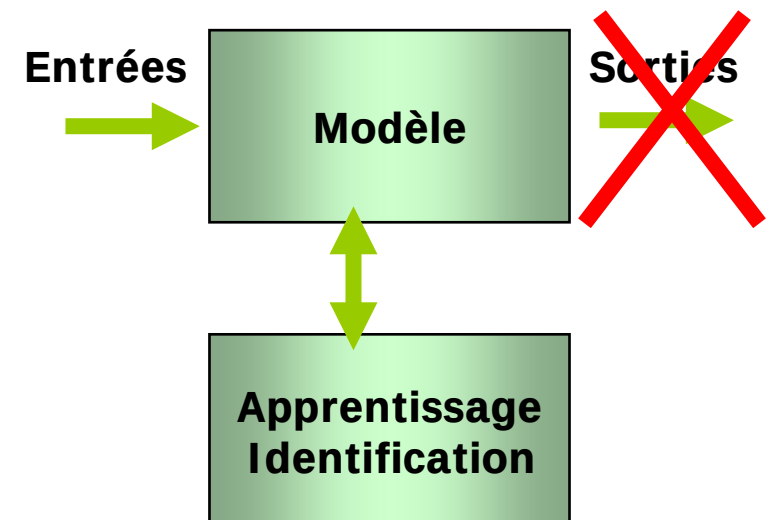
- processus dans lequel l'apprenant reçoit des exemples d'apprentissage comprenant à la fois des **données d'entrée et de sortie**
- les exemples d'apprentissage sont fournis avec leur classe (valeur de sortie prédite)
- But : classer correctement un nouvel exemple (généralisation)
- utilisées principalement en classification et prédiction
- **Algorithmes**
 - Arbres de décision,
 - classifications,
 - algorithmes génétiques,
 - régression linéaire et non linéaire



Typologies selon le type de l'apprentissage :

□ Apprentissage non supervisé - Fouille non supervisée :

- processus dans lequel l'apprenant reçoit des exemples d'apprentissage
- ne comprenant que des **données d'entrée**
- pas de notion de classe
- But : regrouper les exemples en « paquets » (clusters) d'exemples similaires (on peut ensuite donner un nom à chaque paquet)
- utilisé principalement en association et segmentation
- **Algorithmes**
 - Segmentation,
 - regroupement,
 - découverte d'associations et de règles



Techniques de l'ECD :

Auteurs anglo-saxon	Auteurs traduits et certains auteurs francophones	Analyse des données à la française
Clustering	Segmentation	Classification
Classification	Classification	Classement, analyse discriminante
Decision trees	Arbres de décision	Segmentation

La technique de Classification :

- Définition:

- Opération qui permet de placer chaque individu de la population étudiée dans une classe, parmi plusieurs classes prédéfinies.

- Exemple:

- Client acheteur, non acheteur
 - Client bon payeur, mauvais payeur
 - Client fidèle, infidèle à l'entreprise

- Méthodes

- Arbres de décision
 - Réseaux de neurones
 - Analyse factorielle et analyse en composantes principales

La technique de Segmentation :

■ Définition:

- Regrouper les individus d'une population en un nombre limité de groupes, les segments (ou cluster, ou partitions)
- Les clusters sont découverts automatiquement
- Les clusters regroupent des individus ayant des caractéristiques similaires et séparent les individus ayant des caractéristiques différentes.
- La segmentation est utile:
 - Pour avoir une compréhension d'ensemble de la population
 - Pour constituer des panels représentatifs

■ Méthodes

- Segmentation neuronale
- Segmentation hiérarchique ascendante
- Segmentation relationnelle