Data Mining, fouille de données: Concepts et techniques

Marius Fieschi

Faculté de Médecine de Marseille

Data Mining, fouille de données: Concepts et techniques

Ce cours est très proche du cours diffusé sur le net par

Jiawei Han et Micheline Kamber

Intelligent Database Systems Research Lab School of Computing Science

Simon Fraser University, Canada

http://www.cs.sfu.ca

Qu'ils en soient remerciés

Introduction

- Motivation: Pourquoi le *data mining* (fouille de données)?
- Qu'est-ce que le *data mining*?
- Fouille de données: Sur quel type de données?
- Fonctionnalités de la fouille de données
- Classification des systèmes de *data mining*

Pourquoi la fouille de données?

- L'explosion des données

  Les outils de collecte automatique des données et les bases de données conduisent à d'énormes masses de données stockées dans des entrepôts

- Submergés par les données, manque de connaissance!

- Solution: Entrepôts de données et fouille de données

  ✓ Entrepôts de données et analyse on-line

  ✓ Extraction de la connaissance intéressante (règles, régularités, patterns, contraintes) à partir de grandes bases de données

Evolution de la technologie des bases de données

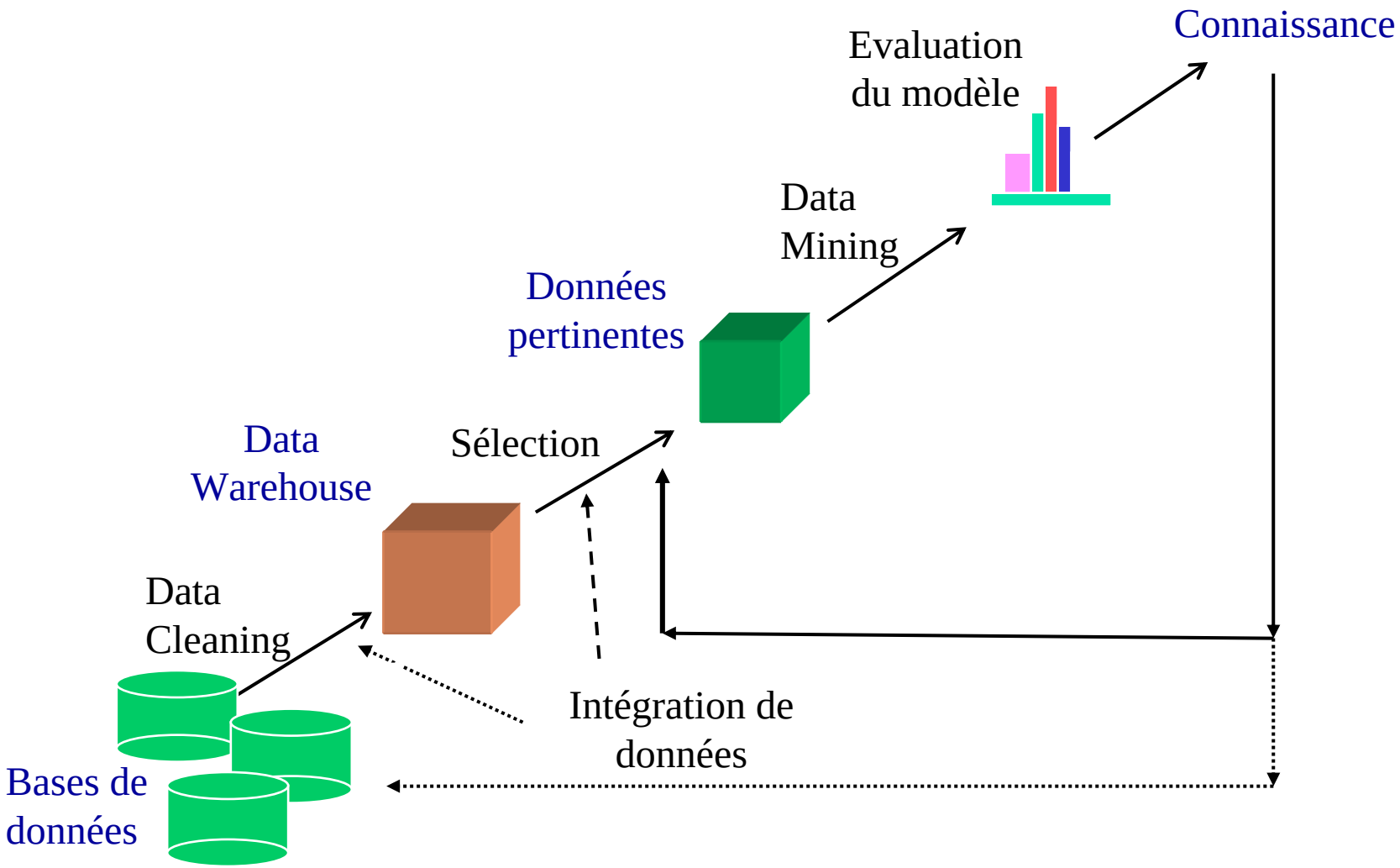
- 1970…: Bases de données relationnelles (RDBMS)
- 1980…: RDBMS, modèles de données avancés (extension du relationnel, OO, ...) et DBMS orientés application (spatial, scientifique, …)
- 1990 - 2000: Fouilles de données et entrepôts de données, BDD multimédia, bases de données Web

Qu'est-ce que la fouille de données?

Fouille de données (découverte de connaissance dans de grandes bases de données):

- ✓ Extraction d'information intéressante (non triviale, implicite, non connue précédemment et potentiellement utile) ou de patterns
- ✓ Découverte de connaissance (mining) dans des Bdd, extraction de connaissance, analyse de données/pattern.
- ✓ Propose des résumés d'information (rapports multidimensionnels, résumés statistiques)

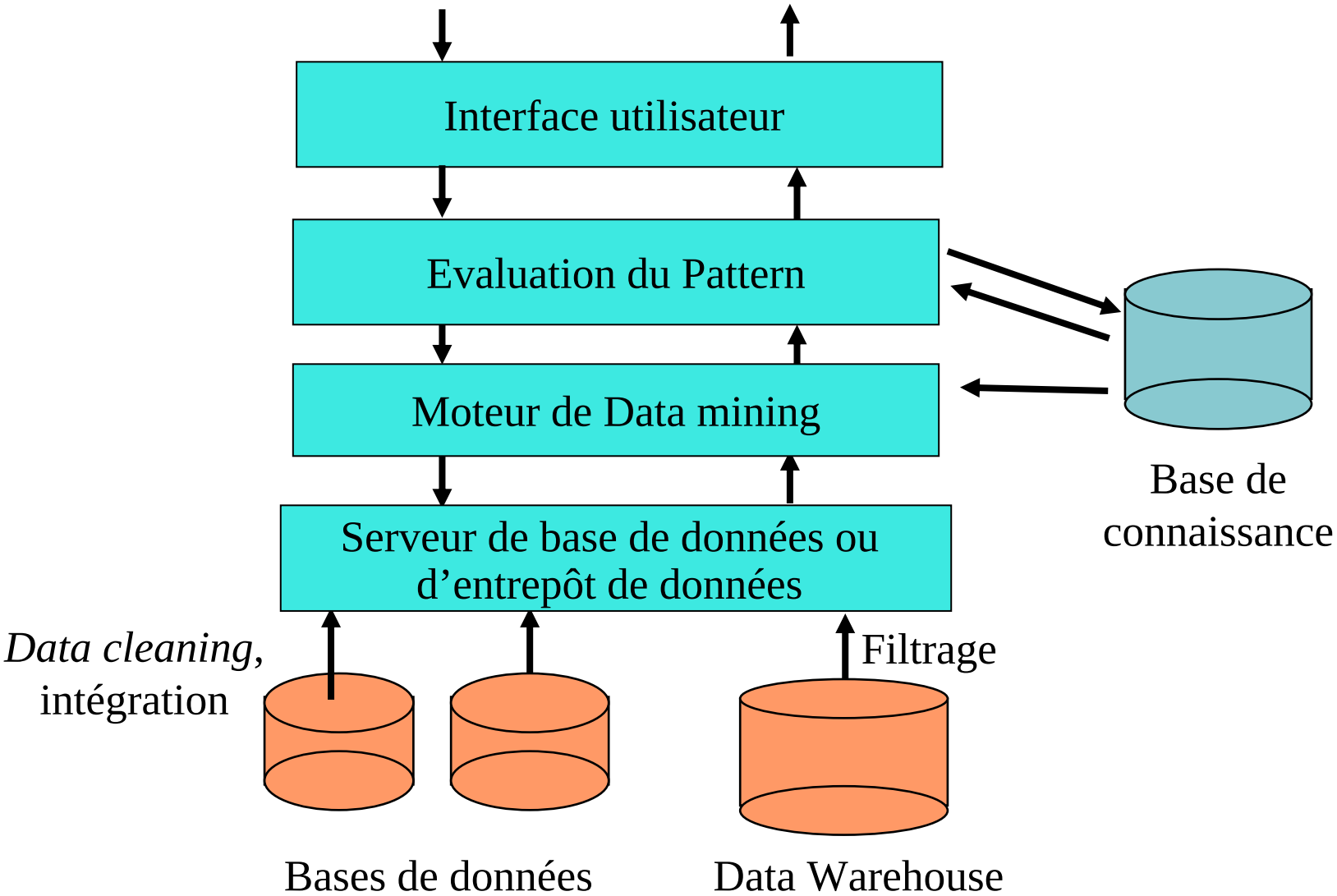
Data Mining: Un processus de découverte de connaissance



Etapes d'un processus de découverte de connaissance

- Connaître le domaine d'application
  Connaissance pertinente déjà établie et buts de l'application
- Sélection des données cibles
- *Data cleaning*, pré traitement
- Réduction de données et transformation:
- Choix des fonctions du *data mining*
  Synthèse, résumé, classification, régression, association, clustering.
- Choix des algorithmes de fouille
- *Data mining*:
  Recherche des modèles intéressants
- Evaluation des pattern et présentation de la connaissance
  Visualisation, transformation, etc.
- Utilisation de la connaissance

Architecture d'un système type de data mining



Data mining: Sur quel type de données?

- Bases de données relationnelles
- Data warehouses / entrepôts de données
- Réservoir de données
  - ✓ Orientées Objet
  - ✓ Bases de données spatiales
  - ✓ Données chronologiques et données temporelles
  - ✓ Bases textuelles et multimédia
  - ✓ WWW

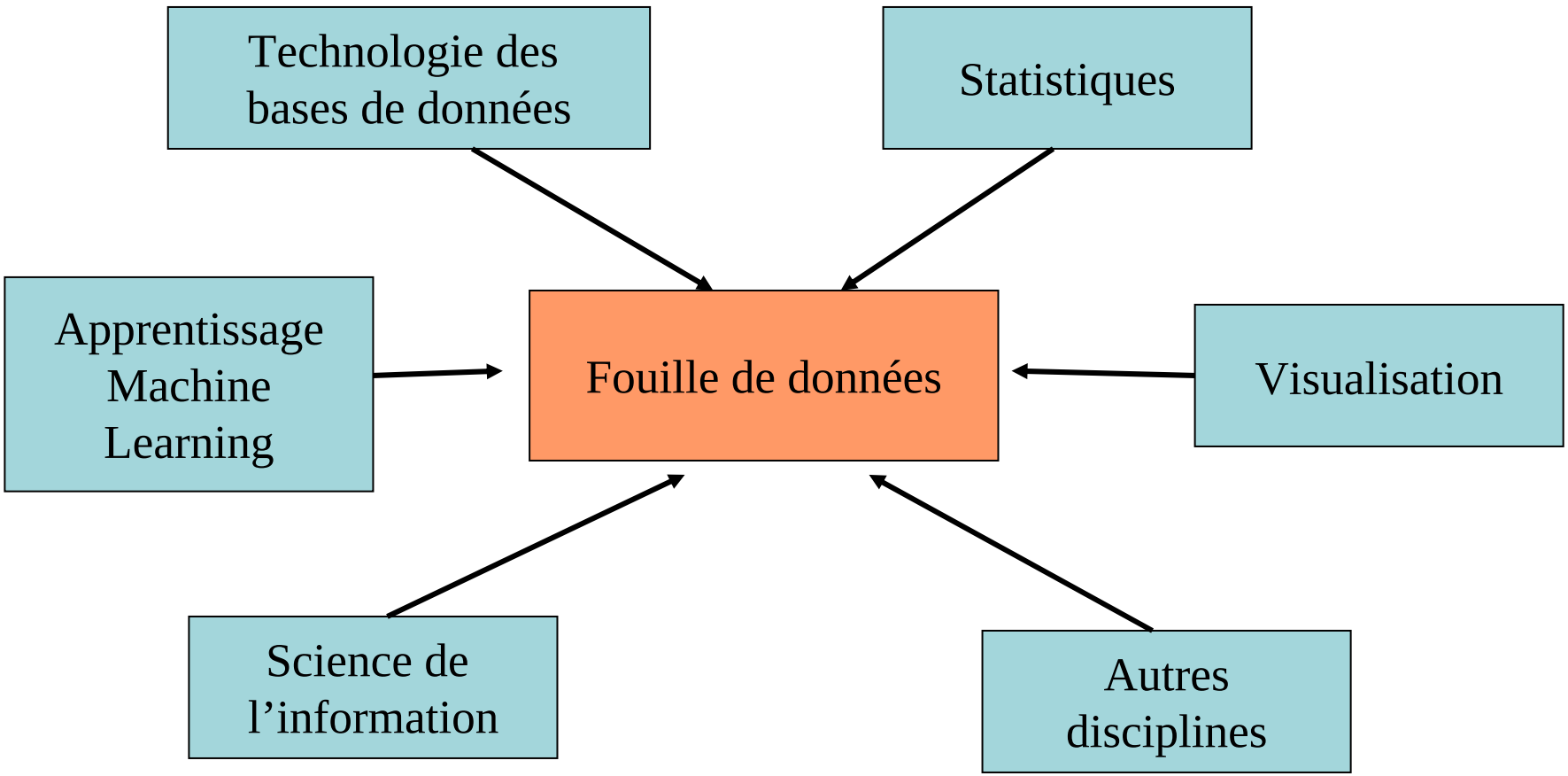
Intérêt des modèles (patterns) découverts

- Un système de data mining génère des milliers de patterns, tous ne sont pas intéressants.
- Intérêt

  Un pattern est intéressant si il est

  - ✓ facilement compris par les humains,
  - ✓ valide sur données nouvelles ou testées avec un certain degré de certitude,
  - ✓ potentiellement utile,
  - ✓ nouveau, ou validant certaines hypothèses que l'on cherche à confirmer
- Objectif vs. subjectif
  - ✓ Objectif: basé sur des statistiques et des structures de patterns
  - ✓ Subjectif: basé sur des croyances des utilisateurs

La fouille de données



La fouille de données

- **Bases de données à fouiller**

  Relationnelles, transactionnelles, orientées objet, spatiales, séries chronologiques, textuelles, multi-media, WWW, etc.

- **Connaissance à fouiller**

  Caractérisation, discrimination, association, classification, déviation et analyse des *outliers*

- **Techniques utilisées**

  Bases de données, data warehouse (OLAP), *machine learning*, statistiques, visualisation, réseaux de neurones.

Entrepôts de données *(data warehousing)* et technologies pour la fouille de données (*data mining*)

Data Warehouse: les applications

Trois types d'applications pour les data warehouse

- Traitement de l'information

  Pour requêtes, analyse statistique de base, rapports, tableaux croisés, diagrammes, graphiques

- Traitement analytique ++

  Analyse multidimensionnelle des données

- Data mining

  Découverte de connaissances et de modèles

  Pour réaliser des classifications, des analyses de prédiction.

Data Warehousing et technologies pour la fouille de données

- Qu'est-ce qu'un *data warehouse*?
- Un modèle de données multi dimensionnelles
- Architecture du *data warehouse*
- Implémentation d'un *data warehouse*
- Du *data warehousing* à la fouille de données

Qu'est-ce que le Data Warehouse?

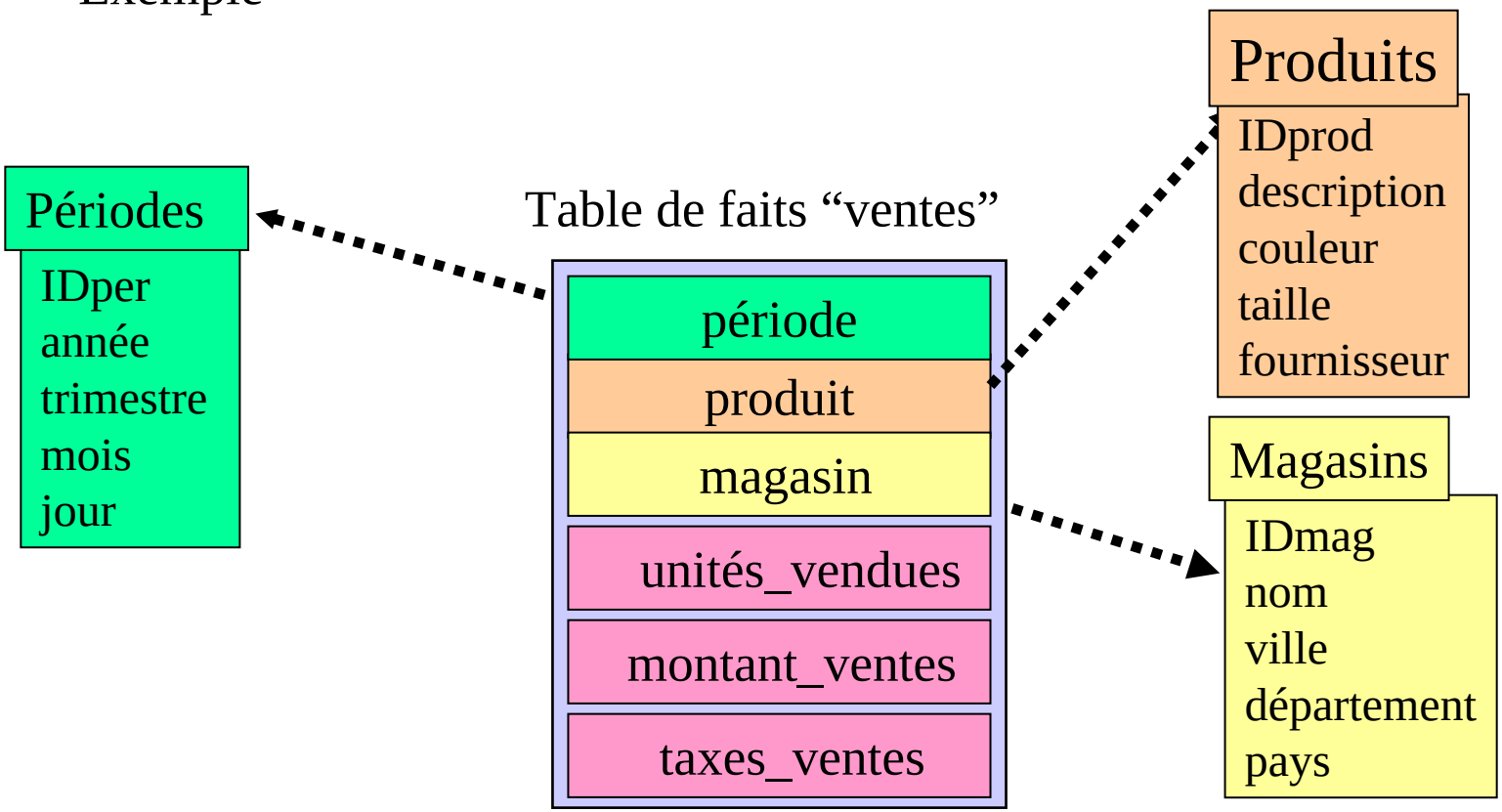
- Une base de données d'aide à la décision qui est entretenue de manière séparée de la base de données opérationnelle de l'organisation
- Aide au traitement de l'information en fournissant une plateforme de données historiques consolidées pour l'analyse.
- *Data warehousing*: Le processus de construction et d'utilisation du *data warehouse*

L'approche "*Subject-Oriented*" du data warehouse

- Organisé autour des sujets majeurs, tels que personne, client,…
- Sujet= Faits + dimensions
- Centré sur la modélisation et l'analyse de données pour les décideurs, non pour des opérations quotidiennes
- Fournit une vue simple, concise sur des sujets particuliers en excluant des données inutiles dans le processus d'aide à la décision
- Construit par intégration de sources de données multiples et hétérogènes

Le schéma en étoile

- Une table de faits encadrées par N tables de dimensions
- Exemple



Data Warehouse

- La consolidation des données

  Application de techniques de « *data cleaning* » et de «*data intégration* »

- La cohérence des données doit être assurée

  Nommer les règles de codage, les mesures, les attributs,… pour les différentes sources de données

- La conversion des données intégrées au *data warehouse*
- L'importance du « temps » pour le *data warehouse*

  Toutes les structures « clés » dans le *data warehouse* contiennent un élément de temps, explicitement ou implicitement

Data Warehouse

- La mise à jour de données opérationnelles n'est pas réalisée dans le *data warehouse*

  Ne demande pas de transactionnel et mécanismes de contrôle d'accès concurrentiels

  Demande uniquement deux opérations en accès aux données:

  Chargement initial de données et accès aux données.

- Intégration traditionnelle de bases de données (bdd) hétérogènes:

  Construction de wrappers/médiateurs au dessus des bdd hétérogènes

Deux types de traitements: OLAP / OLTP

- OLTP (*on-line transaction processing*)

  Tâche majeure des Bdd relationnelles traditionnelles

  Opérations quotidiennes enregistrées.

- OLAP (*on-line analytical processing*)

  Tâche majeure des systèmes de data warehouse

  Analyse de données et décision

Le data warehouse: OLAP

OLTP vs. OLAP

	OLTP	OLAP
Utilisateurs	employé, professionnel	Analyste connaissance
Fonction	Opérations au jour le jour	Aide à la décision
Conception de la Bdd	Orientée application	Orientée sujet
Donnée	courante, détaillée, simple relationnel	historique, résumée, multidimensionnelle, intégrée, consolidée
Usage	répétitif	ad-hoc
Accès	read/write index/hash sur clé primaire	multiples
Unité de travail	court, transaction simple	Requête complexe
Enregistrements accès.	dizaines	millions
Nb utilisateurs	milliers	centaines
Taille de la Bdd	100MB-GB	100GB-TB
Métrique	transaction	requête

Pourquoi séparer le *Data Warehouse*?

Haute performance pour les deux systèmes

- DBMS: performance pour OLTP

  méthodes d'accès, index, accès concurrentiels, restauration

- Warehouse: performance pour OLAP

  requêtes complexes, vue multidimensionnelle, consolidation

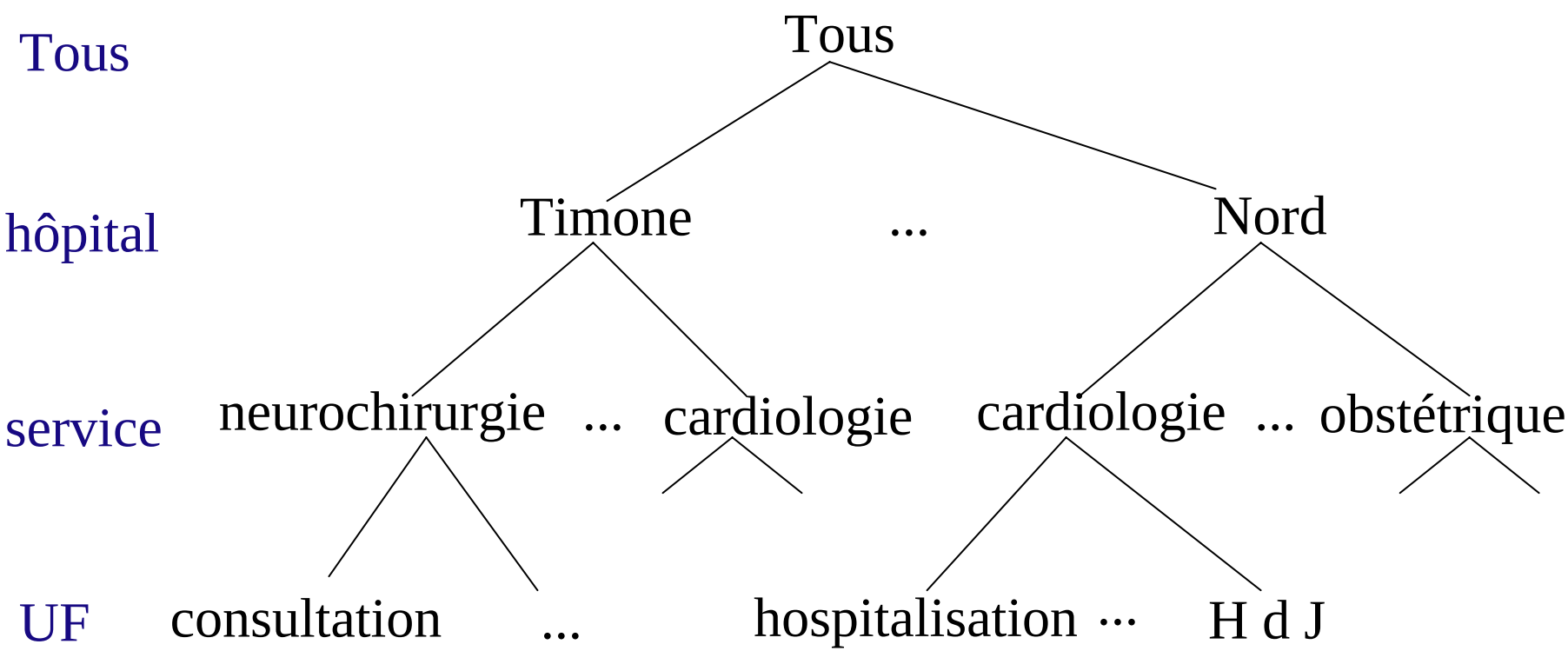
Différentes fonctions et différentes données

- Données manquantes:
  L'aide à la décision (AAD) demande des données historiques. Les Bdd opérationnelles ne les gèrent pas toujours
- Consolidation de données:
  L'AAD demande la consolidation (agrégation, résumé) de données issues de sources hétérogènes
- Qualité des données:
  Habituellement différentes sources utilisent des représentations de données non cohérentes, des codes et des formats à «réconcilier»

Les cubes de données

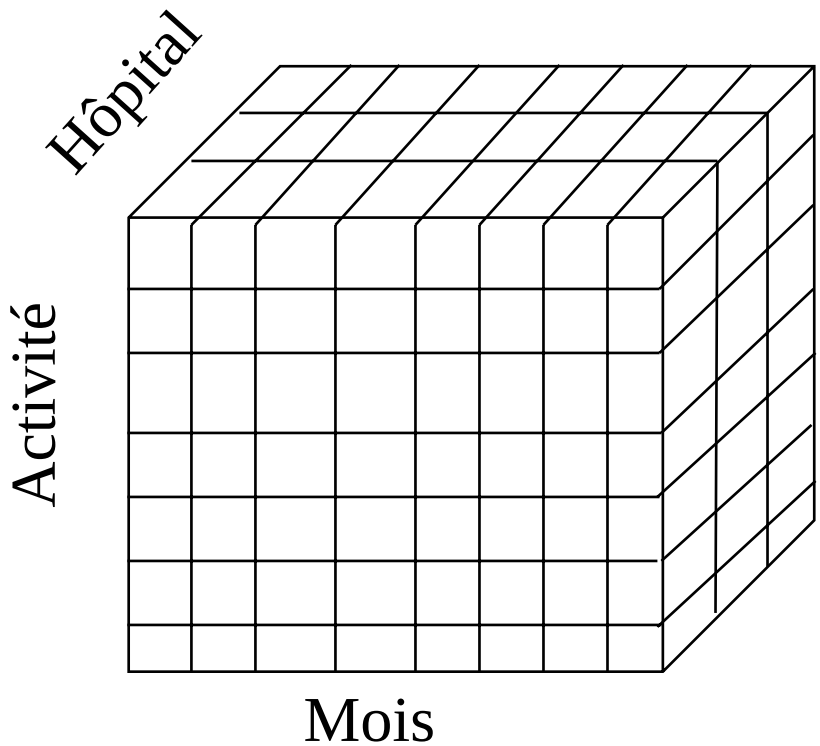
- Un data warehouse est basé sur un modèle de données multidimensionnel qui « voit » les données sous forme de «data cube »
- Un « data cube », comme par exemple les ventes, permet de modéliser et de voir les données relatives aux ventes en de multiples dimensions

Une hiérarchie de concepts: Dimension (lieu)

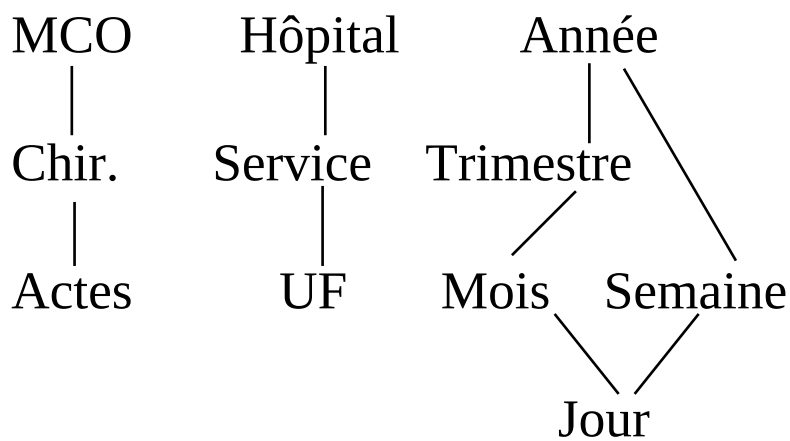


Données multidimensionnelles

Volume des factures, une fonction de l'activité, mois, et de l'hôpital



Dimensions: Activité, Lieu, Temps
Synthèses hiérarchiques



Un exemple de cube de données