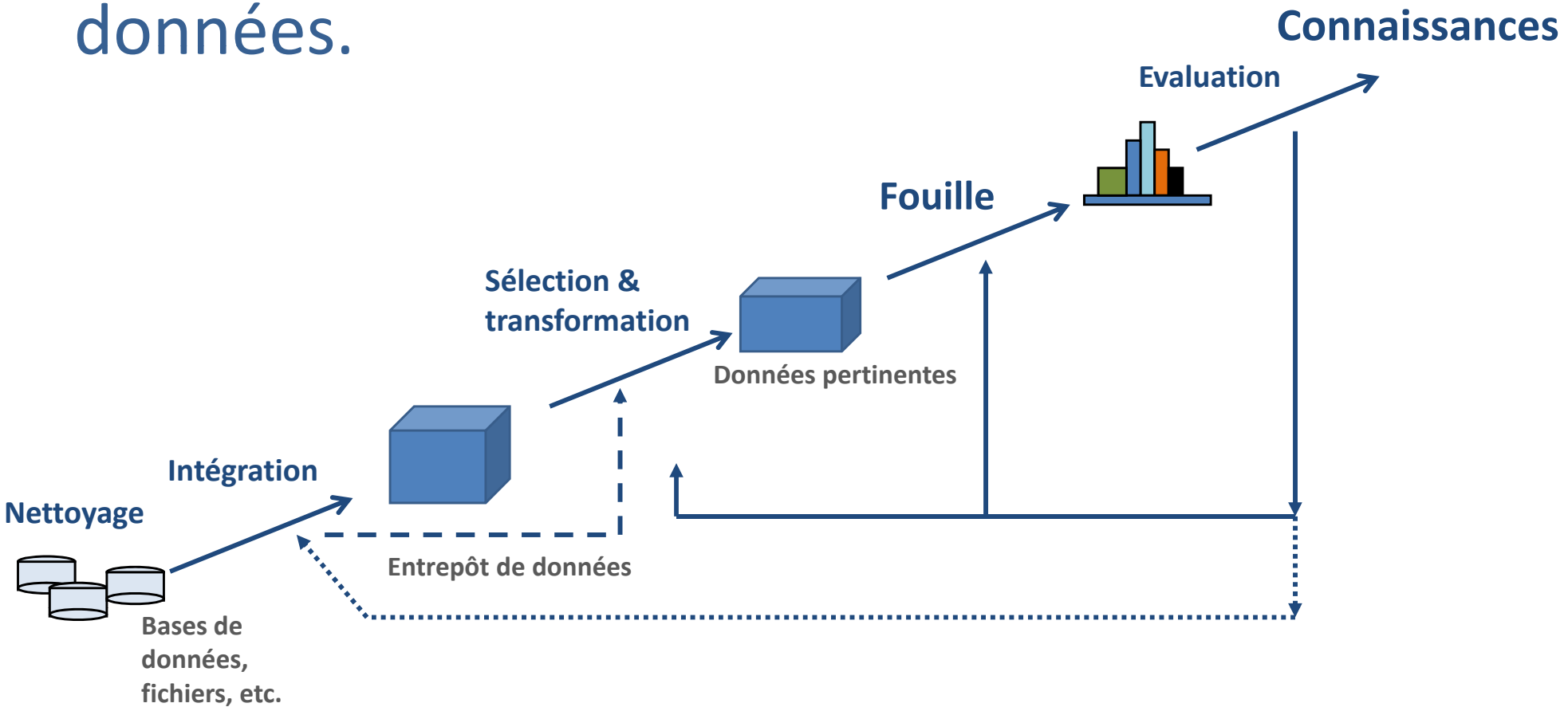
- Définition :

  Processus ou méthode qui extrait des connaissances « intéressantes » ou des motifs (patterns) à partir d'une grande quantité de données.



Au programme

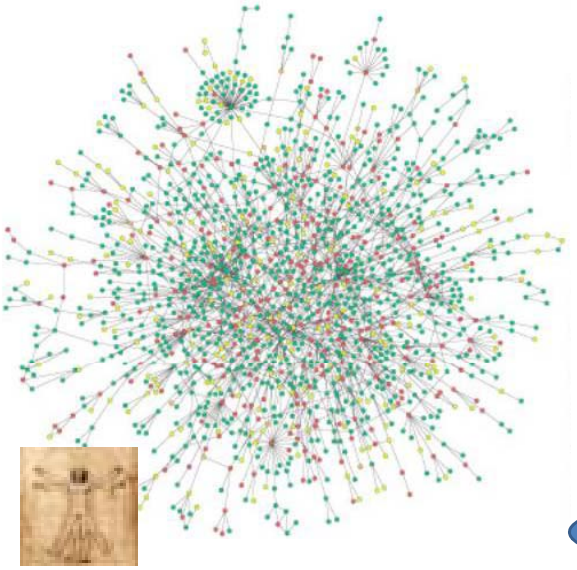
- Statistiques descriptives
- Prétraitement des données
- Caractérisation
- Classification et prédiction
- Clustering
- Statistiques

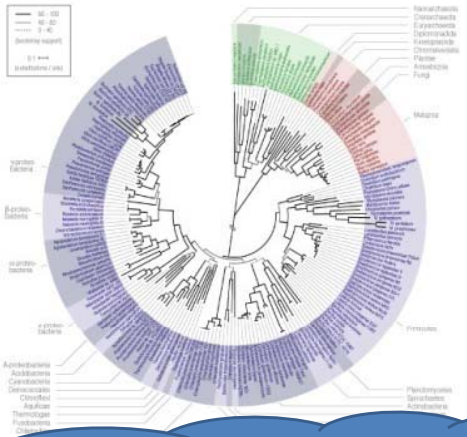
- Règles d'association
- Cubes de données et OLAP (On-Line Analytical Processing)

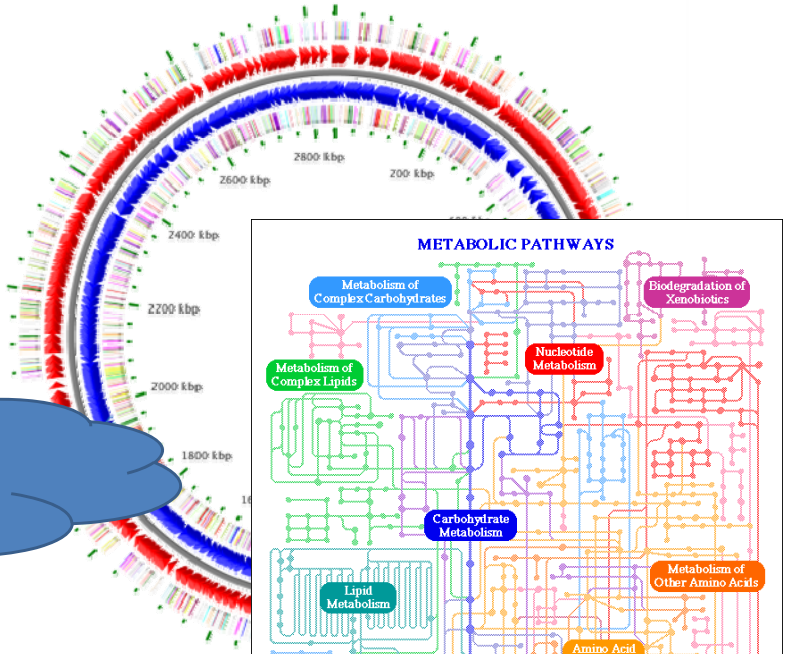
Motivation

- Masses de données
  - Outils automatisés de collecte de données
  - Maturité des SGBD
  - Entrepôts de données (data warehouses et information repositories)
  - ex : Génomes (complets), PubMed, données d'expression
- Données *vs.* connaissances
- Solution : entrepôts de données et data mining
  - Data warehousing and on-line analytical processing (OLAP)
  - Extraction de connaissances (règles, régularités, motifs, contraintes) à partir de grosses bases de données

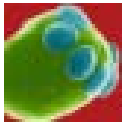
Un déluge de données mais peu de connaissances

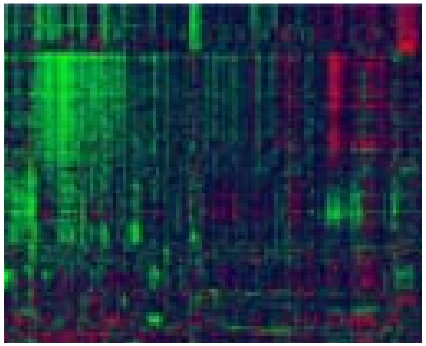


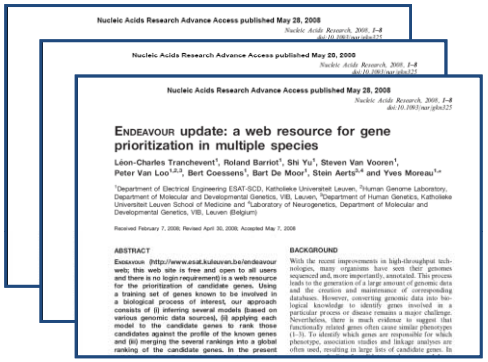




Comment analyser ces données ?






Nucleic Acids Research Advance Access published May 28, 2008

Nucleic Acids Research Advance Access published May 20, 2008

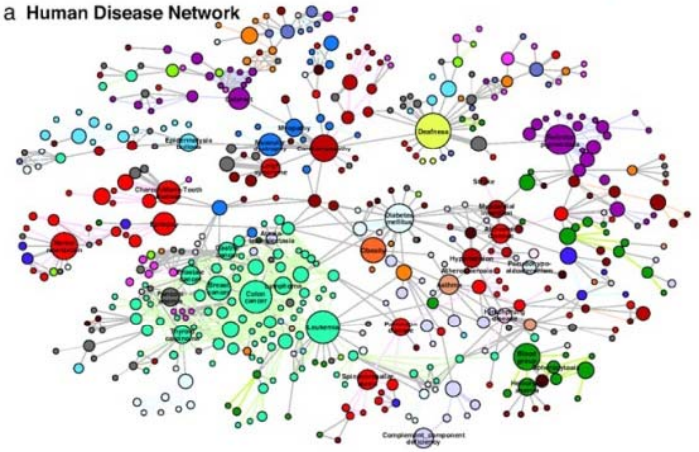
Nucleic Acids Research Advance Access published May 28, 2008

ENDEAVOUR update: a web resource for gene prioritization in multiple species

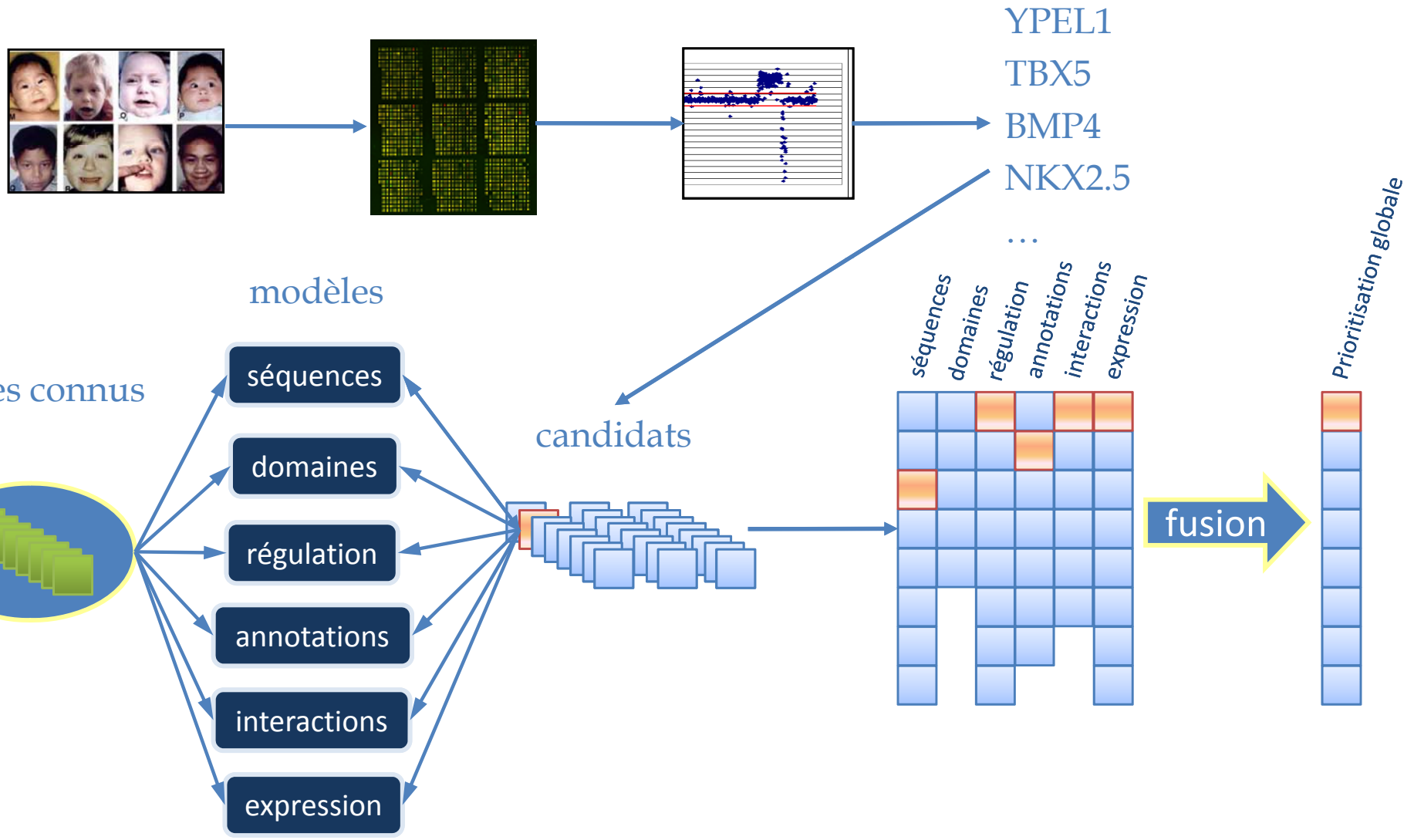
Léon-Charles Tranchevent[1], Roland Barriot[2], Shi Yu[1], Steven Van Vooren[1], Peter Van Loo[1,2,3], Bert Coessens[1], Bart De Moor[1], Stein Aerts[3,4] and Yves Moreau[1,*]

ABSTRACT

BACKGROUND



Recherche de motifs intéressants



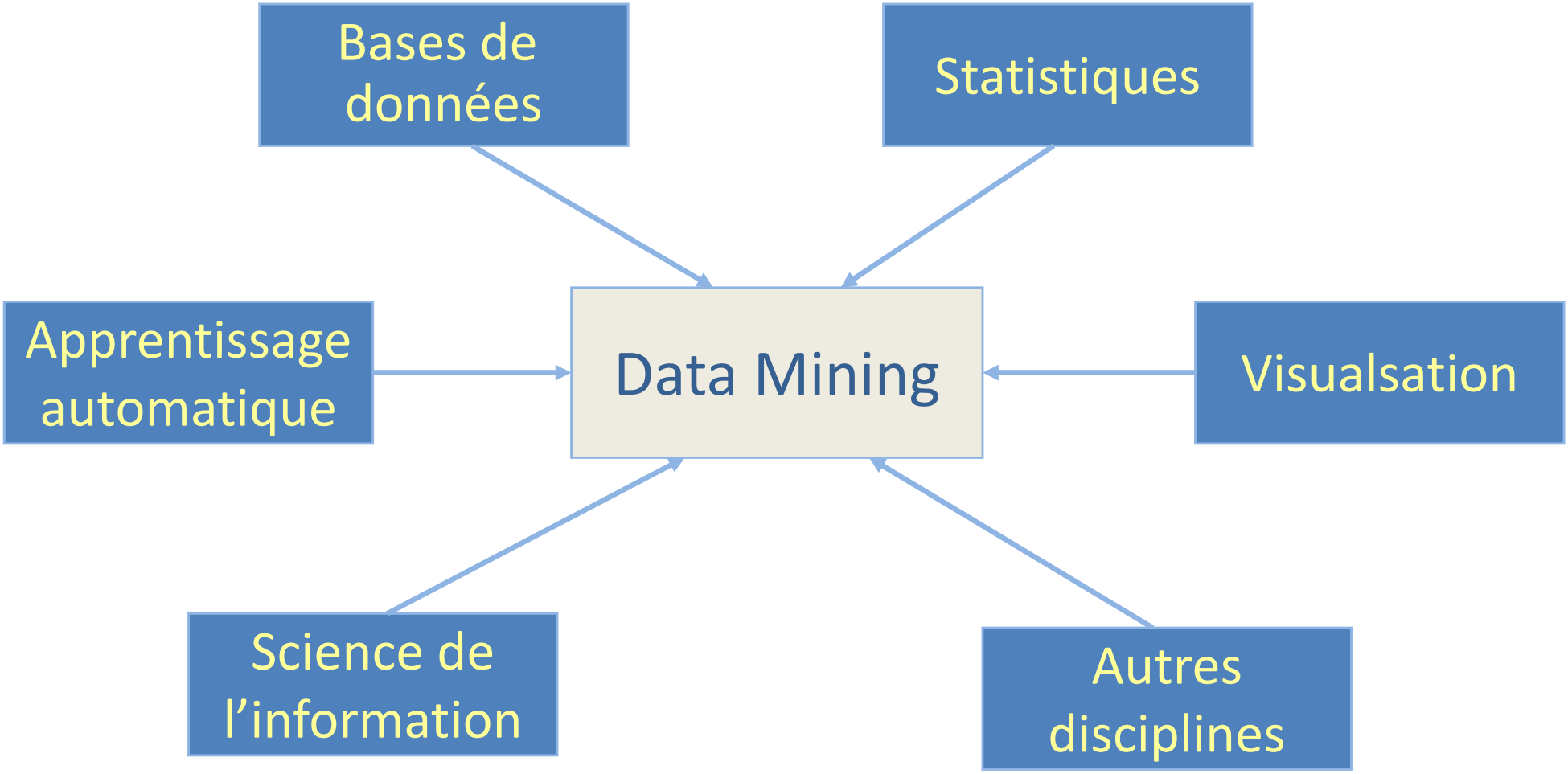
Evolution

- 1960 :
  - Systèmes de gestion de fichiers, collection de données, bases de données (modèle réseau)
- 1970 :
  - Émergence du modèle relationnel et de son implémentation
- 1980 :
  - SGBD relationnels, modèles avancés (relationnel étendu, OO, déductif, etc.) et orientés application (spatial, scientifique)
- 1990 :
  - Data mining et entrepôts de données, multimédia, et Web

Qu'est ce que le data mining ?

- Data mining (découverte de connaissances dans les bases de données) :
  - Extraction d'informations ou de motifs intéressants (non triviaux, implicites, inconnus auparavant et potentiellement utiles) à partir de grandes bases de données
- Autres appellations :
  - Data mining : est-ce judicieux ?
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, information harvesting, business intelligence, etc.
- Ce qui n'est pas du data mining
  - (Deductive) query processing
  - Systèmes experts

Data Mining : union de disciplines variées



Pour quoi faire ?

- Analyse des bases de données et aide à la décision
  - Analyse du marché et management
    - cible marketing, gestion de la relation client, analyse du panier de la ménagère, segmentation du marché
  - Analyse de risques et management
    - Prévisions, fidélisation du client, mises en avant améliorées, contrôle qualité, analyses de compétitivité
  - Détection des fraudes et management
- Bio-informatique
- Autres applications
  - Text mining (news group, email, documents, PubMed) et Web
  - Intelligent query answering

- Quelles sources de données ?
  - Transactions bancaires (CB), coupons de réduction, service clients (plaintes), et aussi les études publiques de style de vie
- Cible marketing
  - Trouver des groupes « modèles » de clients qui partagent les mêmes caractéristiques : intérêts, revenus, habitudes de consommation, etc.
- Déterminer les profils d'achat des clients au cours du temps
  - Ex : compte joint après le mariage
- Cross-market analysis
  - Associations/corrélations des ventes entre produits
  - Prédictions basées sur les associations d'information

Analyse du marché (suite)

- Profils client
  - Quels types de clients achètent quels produits (clustering ou classification)
- Identifier les besoins des clients
  - Identifier les meilleurs produits pour des clients différents
  - Utiliser la prédiction pour trouver quels facteurs vont attirer des nouveaux clients
- Fournir une synthèse de l'information
  - Rapports multidimensionnels variés
  - Rapports statistiques (tendance générale des données et variation)

Détection de fraude

- Applications
  - carte bancaire
- Approche
  - Utiliser les données d'historique pour construire des modèles de comportements frauduleux puis rechercher par data mining des instances similaires
- Exemples
  - Assurances : détecter les groupes de personnes qui déclarent des accidents/vols pour les indemnités
  - Blanchiment d'argent : détecter les transactions suspectes (US Treasury's Financial Crimes Enforcement Network)
  - Assurance maladie : détecter les patients professionnels et les docteurs associés

Bio-informatique

- Sources de données
  - Séquences (ADN, acides aminés)
  - Structures tri-dimensionnelle (PDB)
  - Puces à ADN (expression, aCGH)
  - Interaction protéiques
  - réseau métabolique
  - régulation génétique
  - PubMed
- Applications
  - Prédiction de structure 3D
  - Prédiction des séquence codantes
  - Prédiction de fonction, d'interaction, de localisation, …
  - Découverte de motifs sur/sous représentés, répétitions
  - Analyse de données d'expression
  - Aide au diagnostic
  - Méthodes de classification, clustering, etc.

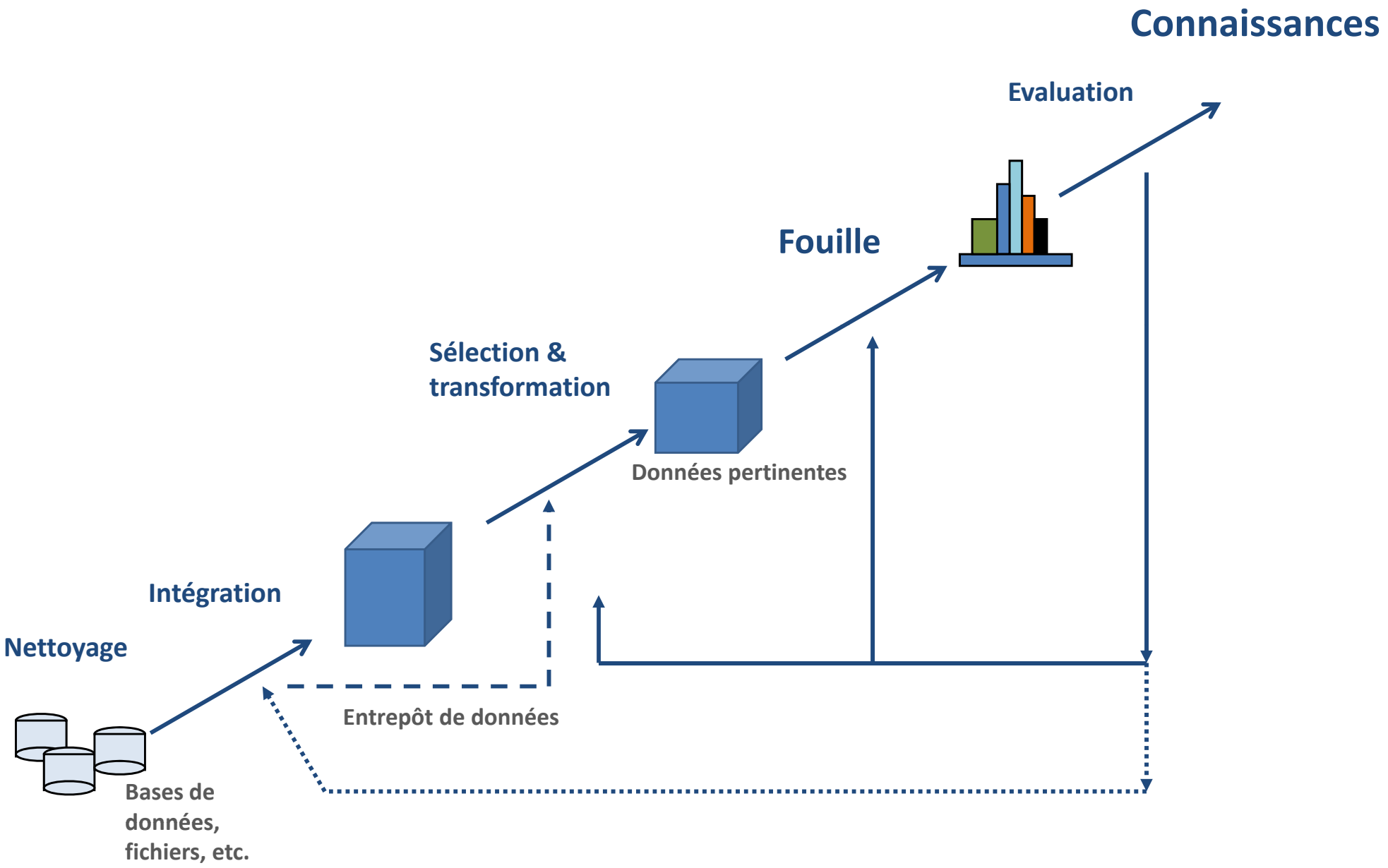
Autres applications

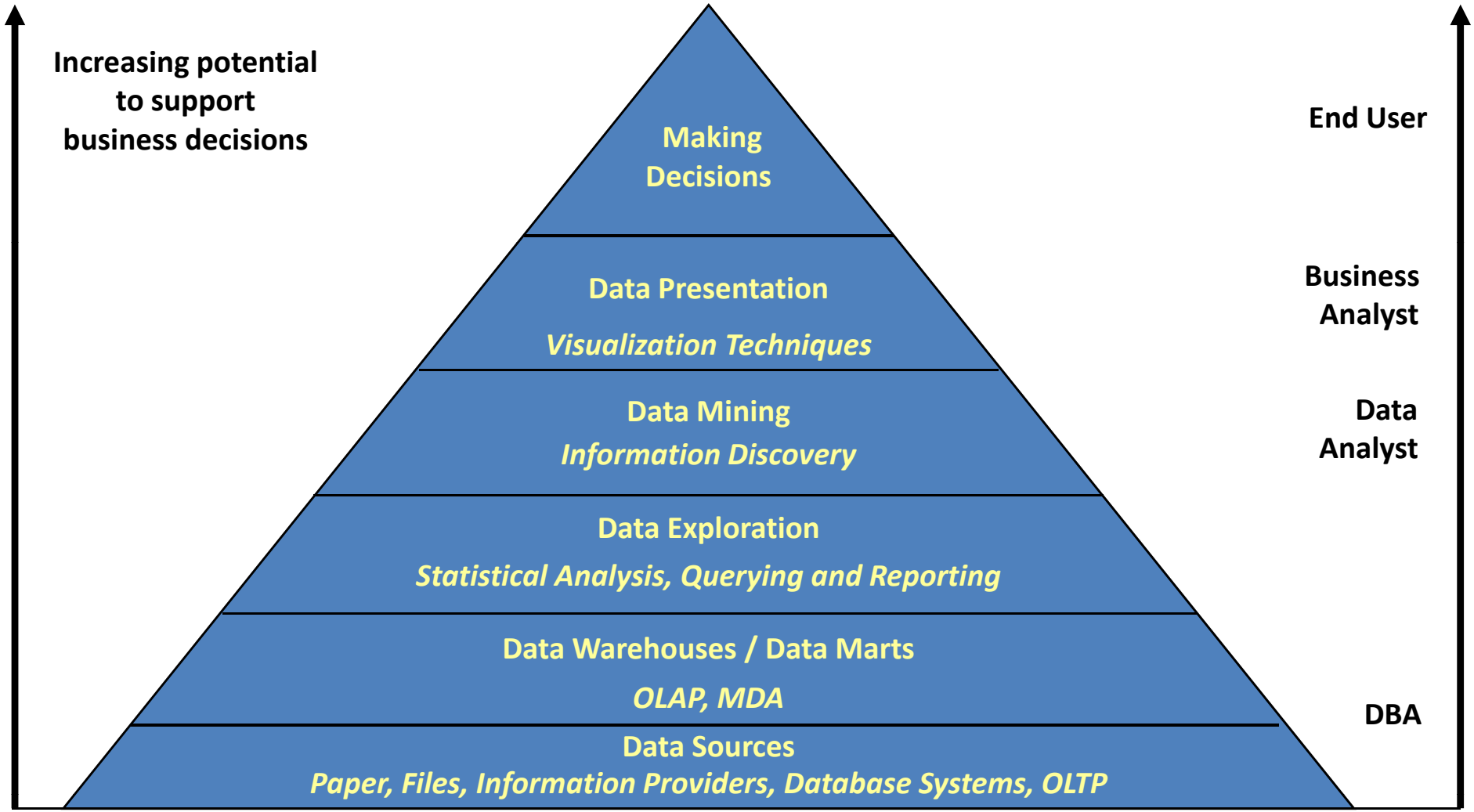
- Astrophysique
  - JPL and the Palomar Observatory discovered 22 quasars with the help of data mining
- Organisation de sites Web
  - Algorithmes de data mining appliqués aux journaux d'accès aux pages commerciales afin d'identifier les préférences et les comportements des clients et d'analyser les performances du marketing Web et l'organisation du site. Ex: IBM Surf-Aid, GoogleAnalytics

Étapes impliquées dans le processus de découverte de connaissances

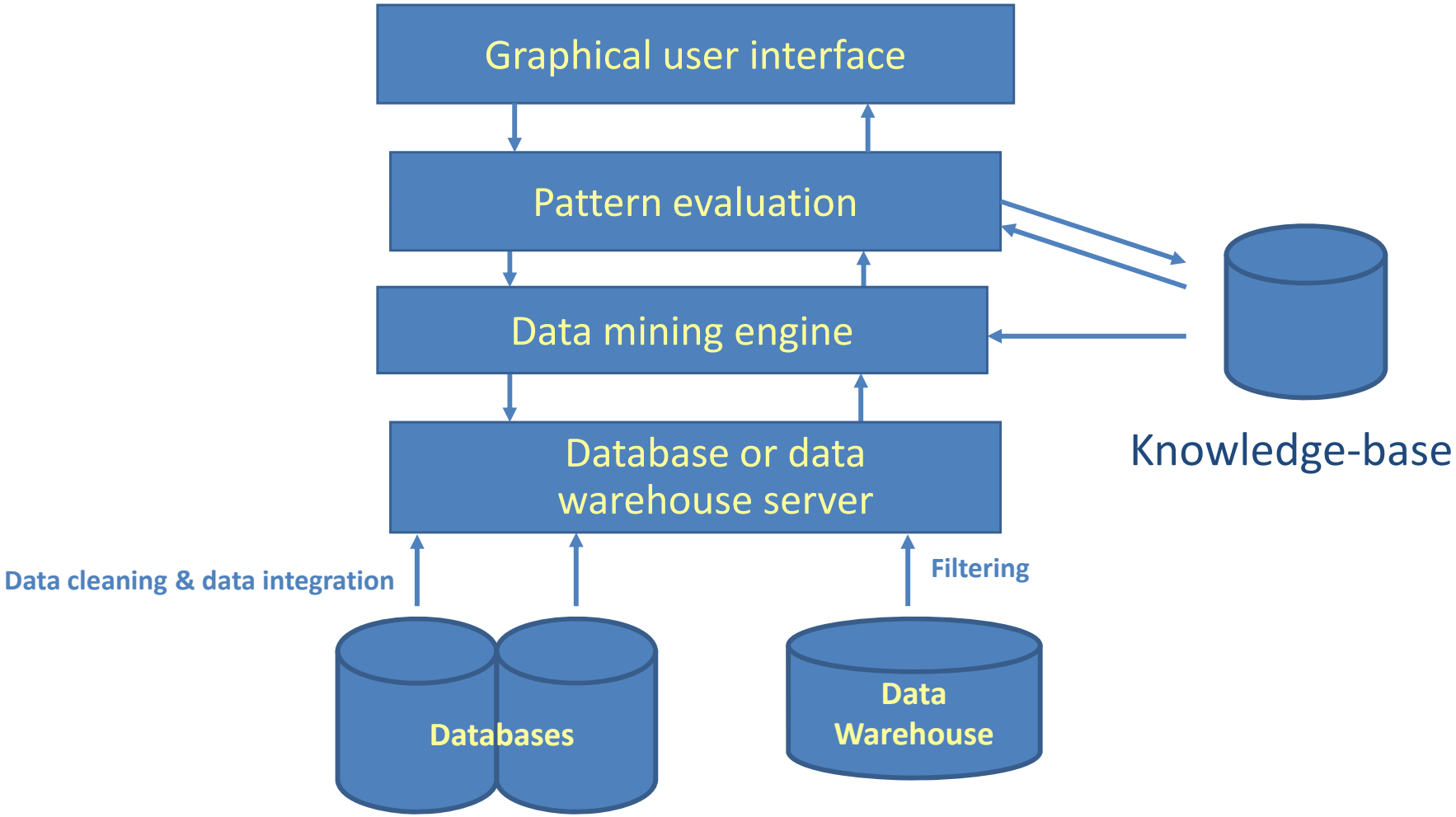
- Apprentissage du domaine d'application :
  - Connaissances nécessaires et buts de l'application
- Création du jeu de données cible : sélection des données
- Nettoyage et prétraitement des données (jusqu'à 60% du travail !)
- Réduction et transformation des données
  - Trouver les caractéristiques utiles, dimensionnalité/réduction des variables
- Choix des fonctionnalités data mining
  - synthèse, classification, régression, association, clustering
- Choix des algorithmes
- Data mining : recherche de motifs (patterns) intéressants
- Évaluation des motifs et représentation des connaissances
  - visualisation, transformation, élimination des motifs redondants, etc.
- Utilisation des connaissances découvertes

Data mining: a KDD process




Increasing potential to support business decisions
Making Decisions
Data Presentation
Visualization Techniques
Data Mining
Information Discovery
Data Exploration
Statistical Analysis, Querying and Reporting
Data Warehouses / Data Marts
OLAP, MDA
Data Sources
Paper, Files, Information Providers, Database Systems, OLTP
End User
Business Analyst
Data Analyst
DBA

Architecture of a Typical Data Mining System



Quelles sortes de données

- Fichiers plats
- Bases de données relationnelles
- Entrepôts de données (data warehouses)
- Bases de données transactionnelles
- Bases de données avancées et entrepôts de données (data repositery)
  - Bases de données orientées objets, et relationnelles objets
  - Spatiales
  - Données temporelles
  - Textes et multimédia
  - WWW

- Description de concepts : Caractérisation et discrimination
  - Généraliser, résumer, et contraster les données caractéristiques, ex : régions sèches vs. humides
- Association (corrélation et causalité)
  - Association multidimensionnelle vs. monodimensionnelle
  - âge(X, “20..29”) ^ revenu(X, “20..29K”) → achète(X, “PC”) [support = 2%, confiance = 60%]
  - contient(T, “PC”) → contient(x, “logiciel”) [1%, 75%]

Fonctionnalités (suite)

- Classification et Prédiction
  - Trouver des modèles (fonctions) qui décrivent et distinguent des classes ou concepts pour la prédiction future
  - ex : séquences codantes, aide au diagnostic
  - Présentation: arbre de décision, règles de classification, réseaux de neurones
  - Prédiction: Prédire des valeurs inconnues ou manquantes
- Clustering
  - Pas de classes prédéfinies : grouper les données pour former des classes nouvelles, ex : familles de protéines basées sur la similarité des séquences
  - Principe : maximiser la similarité intra-classe et minimiser la similarité inter-classes

Fonctionnalités (suite)

- Outlier analysis
  - Outlier: un objet qui se distingue du comportement général des données
  - Peut être considéré comme du bruit ou une exception
  - Utile à la détection de fraudes et évènements rares
- Analyse des tendances et de l'évolution
  - Tendance et déviation : analyse de régression
  - Découverte de motifs séquentiels, analyse de périodicité
  - Analyses basées sur la similarité
- Autres analyses basées sur des motifs ou sur des statistiques

Est-ce que tous les patterns découverts sont intéressants ?

- **Pb: Un système de data mining peut générer des milliers de patterns**
  - Approches suggérées : centré sur l'utilisateur, basé sur des requêtes
- **Mesures du niveau d'intérêt** : un motif est intéressant si il est :
  - facile à comprendre par un humain
  - valide sur des nouvelles données ou données test avec un certain degré de certitude
  - potentiellement utile
  - nouveau
  - ou encore s'il sert à valider une hypothèse que l'utilisateur cherche à confirmer
- **Mesures objective vs. subjective :**
  - Objective : basée sur des statistiques et sur les structures des motifs, ex : support, confiance
  - Subjective : basée sur les sentiments de l'utilisateur, ex : inattendu, nouveau

Peut-on trouver tous les motifs intéressants et seulement ceux là ?

- Complétude : trouver tous les patterns intéressants
  - Est-ce qu'un système peut trouver tous les patterns intéressants ?
  - Association *vs.* classification *vs.* clustering
- Optimisation : trouver seulement les patterns intéressants
  - Est ce qu'un système peut trouver seulement les patterns intéressants ?
  - Approches
    - Générer tous les patterns et filtrer ceux intéressants
    - Générer seulement des patterns intéressants

Schéma de classification des systèmes de data mining

- Fonctionnalité générale
  - Descriptif
  - Prédictif
- Vues différentes, classifications différentes
  - Types de bases de données à fouiller (relationnelles ou OO. Texte ou multimédia)
  - Types de connaissances à découvrir (association, clustering)
  - Types de techniques utilisées (automatique, guidée par l'utilisateur, exploratoire)
  - Types d'applications spécifiques

Classification des systèmes de data mining (suite)

- **Types de bases de données**
  - Relationnelles, transactionnelles, orientée objet, relationnelle objet, spatial, temporelles, texte, multimédia, WWW, etc.
- **Types de connaissances à découvrir**
  - Caractérisation, discrimination, association, classification, clustering, tendance, déviation et outlier analysis, etc.
  - Fonctions Multiples/intégrées et data mining sur plusieurs niveaux
- **Techniques utilisées**
  - Orienté bases de données, entrepôt de données (OLAP), apprentissage automatique(machine learning), statistiques, visualisation, réseaux de neurones, etc.
- **Applications spécifiques**
  - télécommunication, banque, analyse de fraude, Bio-informatique, bourse, Web mining, Weblog, etc.

Principaux défis en data mining (1)

- Méthodologie et interactions utilisateur
  - Fouille de différents types de connaissances dans les bases de données
  - Fouille interactive à des niveaux multiples d'abstraction
  - Incorporation de connaissances *a priori* (background knowledge)
  - Langages de requêtes pour le data mining
  - Expression et visualisation des résultats
  - Prise en compte du bruit ou de données manquantes/incomplètes
  - Évaluation des patterns : le problème du niveau d'intérêt
- Performance et mise à l'échelle
  - Efficacité et mise à l'échelle des algorithmes de data mining
  - Parallélisation, distributivité et possibilités incrémentales des méthodes de fouille

Principaux défis en data mining (2)

- Liées à la diversité des types de données
  - Données relationnelles et types complexes
  - Bases de données hétérogènes et systèmes global d'information (WWW)
- Liées aux applications et aux nouvelles connaissances
  - Applications
    - Création d'outils domaine-spécifique
    - Intelligent query answering
    - Contrôle de processus et aide à la décision
  - Intégration des connaissances découvertes avec celles existantes : problème de fusion des connaissances
  - Protection des données : sécurité, intégrité, et données privées (informatique et libertés, données cliniques)

Bilan

- Data mining: découverte de motifs intéressants à partir de données massives
- Évolution naturelle des technologies des bases de données, large demande, beaucoup d'applications
- **Le processus de découverte implique le nettoyage, l'intégration, la sélection, la transformation et la fouille des données, suivies de l'évaluation des motifs extraits et de leur représentation**
- La fouille peut s'effectuer sur une grande variété d'entrepôt de données
- Fonctionnalités : caractérisation, discrimination, association, classification, clustering, analyse des tendances et des outliers, etc.