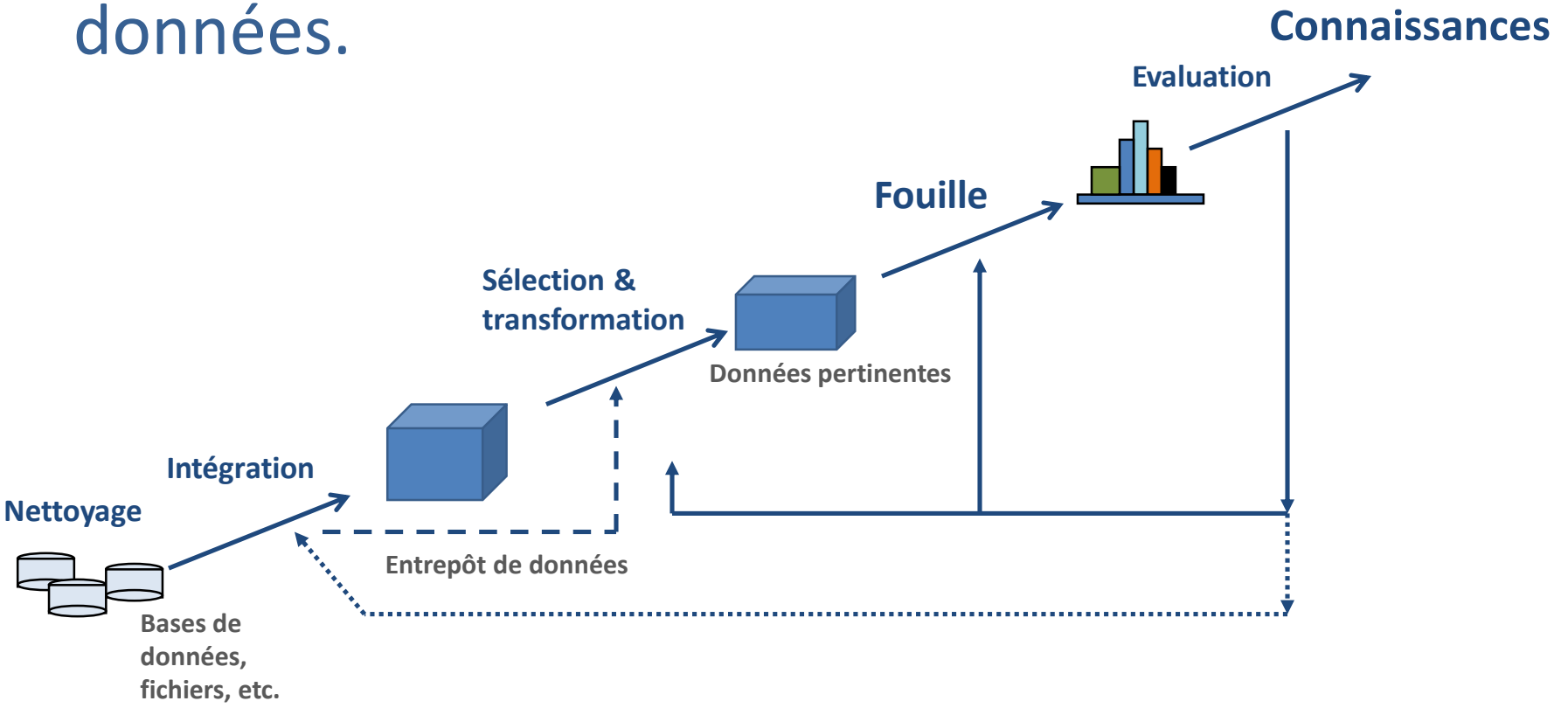
- Déroulement
  - Supports et emploi du temps sur le site silico.biotoul.fr
- Objectifs
  - Aperçu de la discipline, regard critique
  - Utilisation de méthodes existantes
  - Implémentation et évaluation de certaines méthodes
- Evaluation
  - Projet (CC) + exam

- Définition :

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



Omni-présence du data mining

- Données numériques disponibles
  - Monétaires : Comptes bancaires, CB, cartes de fidélité
  - Réseaux sociaux
  - Localisation et déplacement : cartes de transports, géolocalisation GPS
  - Santé : Carte vitale, Sécu, génomes
  - Scientifique : astrophysique, biologie, …
- Fouille : comment extraire du sens ?
  - Profils clients, sécurité du territoire, risque économique, santé, …
  - Météo
  - Modèles et hypothèses scientifiques

Tâches et méthodes

- Statistiques (descriptives, inférentielles, …)
- Prétraitement des données
- Caractérisation
- Classification et prédiction
- Clustering
- Règles d'association
- Evaluation des performances
- Optimisation
- Fouille de texte
- Cubes de données et OLAP (On-Line Analytical Processing)
- Big data

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, spectres de masse, métabolomique
- Données *vs.* connaissances
- Solution : entrepôts de données et data mining
  - Data warehousing et on-line analytical processing (OLAP)
  - Extraction de connaissances (règles, régularités, motifs, contraintes) à partir de grosses bases de données

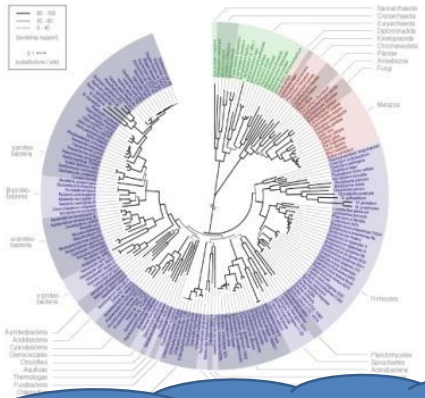
Evolution

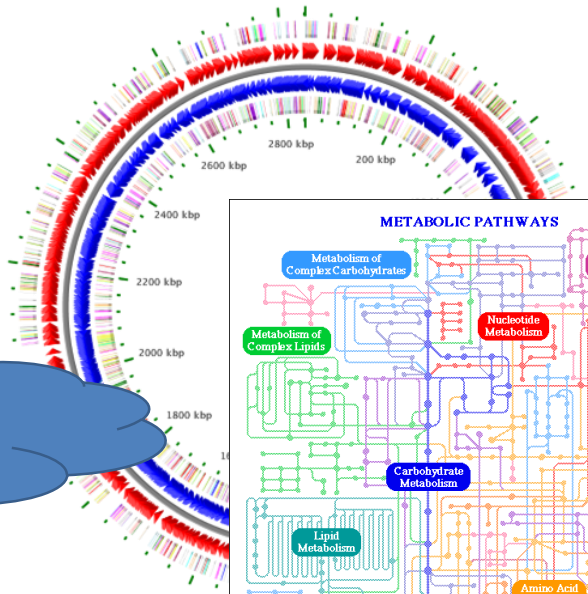
- 1960 :
  - Systèmes de gestion de fichiers, collection de données, bases de données (modèle réseau)
- 1970 :
  - Émergence du modèle relationnel et de son implémentation
- 1980 :
  - SGBD relationnels, modèles avancés (relationnel étendu, OO, déductif, etc.) et orientés application (spatial, scientifique)
- 1990 :
  - Data mining et entrepôts de données, multimédia, et Web
- 2000 :
  - Données biologiques puis réseaux sociaux
  - Workshop BioKDD (2001), Journal BioData mining (2008)
- 2010 :
  - Big Data, deep learning, cloud computing

Qu'est ce que le data mining ?

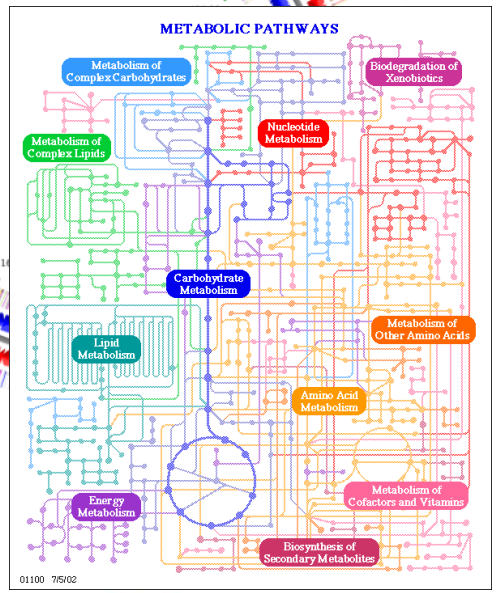
- Data mining (découverte de connaissances dans les bases de données) :
  - Extraction d'informations ou de motifs intéressants (non triviaux, implicites, inconnus auparavant et potentiellement utiles) à partir de grandes bases de données
- Autres appellations :
  - Data mining : est-ce judicieux ?
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

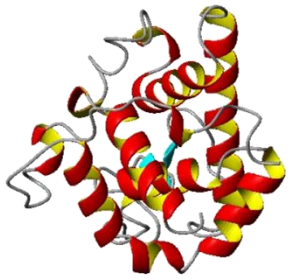
Un déluge de données

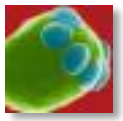


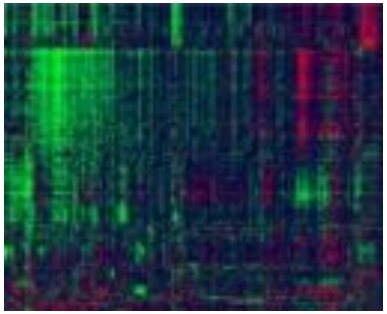


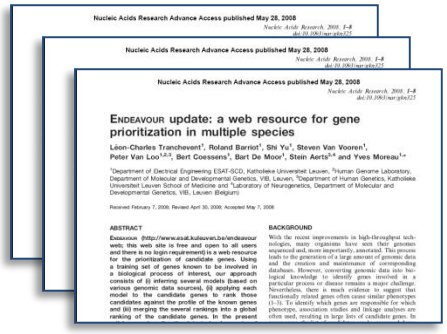
Comment exploiter au mieux ces données ?











Nucleic Acids Research Advance Access published May 28, 2008

Nucleic Acids Research Advance Access published May 28, 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[1], 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,*]

[1]Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, [2]Human Genome Laboratory, Department of Molecular and Developmental Genetics, VIB, Leuven, [3]Department of Human Genetics, Katholieke Universiteit Leuven School of Medicine and [4]Laboratory of Neurogenetics, Department of Molecular and Developmental Genetics, VIB, Leuven (Belgium)

Received February 7, 2008; Revised April 30, 2008; Accepted May 7, 2008

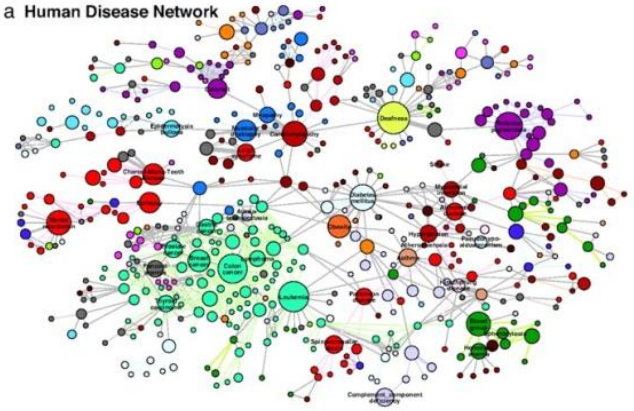
ABSTRACT

ENDEAVOUR (http://www.esat.kuleuven.be/endeavour web), this web site is free and open to all users and there is no login requirement) is a web resource for the prioritization of candidate genes. Using a training set of genes known to be involved in a biological process of interest, our approach consists of (i) inferring several models (based on various genomic data sources), (ii) applying each model to the candidate genes to rank those candidates against the profile of the known genes and (iii) merging the several rankings into a global ranking of the candidate genes. In the present

BACKGROUND

With the recent improvements in high-throughput technologies, many organisms have seen their genomes sequenced and, more importantly, annotated. This process leads to the generation of a large amount of genomic data and the creation and maintenance of corresponding databases. However, converting genomic data into biological knowledge to identify genes involved in a particular process or disease remains a major challenge. Nevertheless, there is much evidence to suggest that functionally related genes often cause similar phenotypes [1–3]. To identify which genes are responsible for when phenotype, association studies and linkage analyses are often used, resulting in large lists of candidate genes. In

a Human Disease Network



Classification non supervisée

- Pfam
  - Alignement des séquences
  - Clustering des séquences et domaines
  - Définition des familles
  - Associations familles/fonctions

Classification supervisée, prédiction

  - Nouvelle séquence, détection des domaines présents

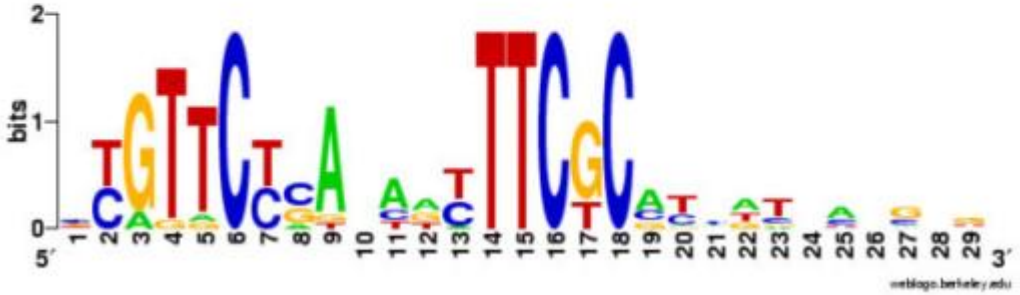
Recherche de motifs intéressants

A

```
>HHT2_HHF2_IR_S_cerevisiae
314 ccgttccgagcacttcgcattaagcgcgt  286 -  (25.64520) (5.586871e-10)
381 acgttctgggagcttcgcgtctcaagctt  409 +  (9.891990) (1.132719e-02)
349 ctagaccgagagttttcgcatttgtatggc  377 +  (8.797430) (2.536818e-02)
>HTA2_HTB2_IR_S_cerevisiae
347 ctgtgccccaccgcttcgcctaataaagcg  375 +  (27.34320) (3.455151e-11)
334 gtgttcccattattttctcaaagtgatgcg  306 -  (13.35890) (7.270649e-04)
376 gtgttctcaaaattttctcccgttttcag  404 +  (10.77820) (5.300732e-03)
291 gtgttctctgaaattcgcatcactttgag  319 +  (10.14830) (8.379708e-03)
>HTB1_HTA1_IR_S_cerevisiae
443 ccattccaatagcttcgcacagtgaggcg  415 -  (25.60310) (7.318539e-10)
400 ctgttccaaaatttttcgcctcactgtgcg  428 +  (12.09860) (2.459747e-03)
298 ctgttctcacttttttcgcgcgttgcaccc  270 -  (11.52740) (3.554757e-03)
244 tcgttctcatttttttcgcggaagaaaggg  272 +  (10.85850) (5.873168e-03)
>HHF1_HHT1_IR_S_cerevisiae
278 ctgttccgagcgcttctcccataatggt  250 -  (27.53840) (2.261761e-11)
391 tcgttctcacaatttctcacatttccttg  419 +  (12.23800) (1.720110e-03)
357 tcgttctcacatttttcgcattgtcccata  385 +  (11.22860) (3.671231e-03)
313 gcgttctgaaaacttcgcatcttcacata  285 -  (8.294790) (2.770252e-02)
```

B

Inférence d'un modèle

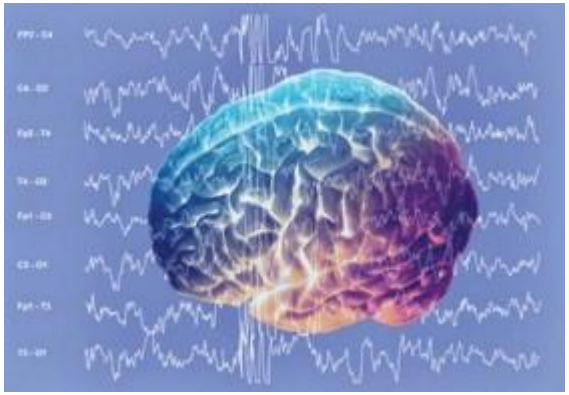


Utilisation du modèle sur des séquences nouvelles pour effectuer des prédictions (ex: sites de fixation de facteur de transcription)

- Maladie de Parkinson
  - Crise de tremblements

PHOTO: Dr. Helen Mayberg

Stimulation électrique pour empêcher la crise

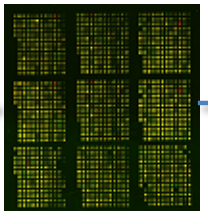


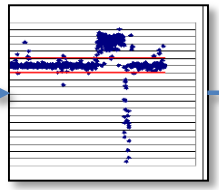
Activité cérébrale précédent une crise

Modèle pour la détection et la prédiction en temps réel

Implantation sur une puce électronique reliée à des électrodes de stimulation

Exploitation des données disponibles





YPEL1
TBX5
BMP4
NKX2.5
…

gènes connus

modèles

séquences
domaines
régulation
annotations
interactions
expression

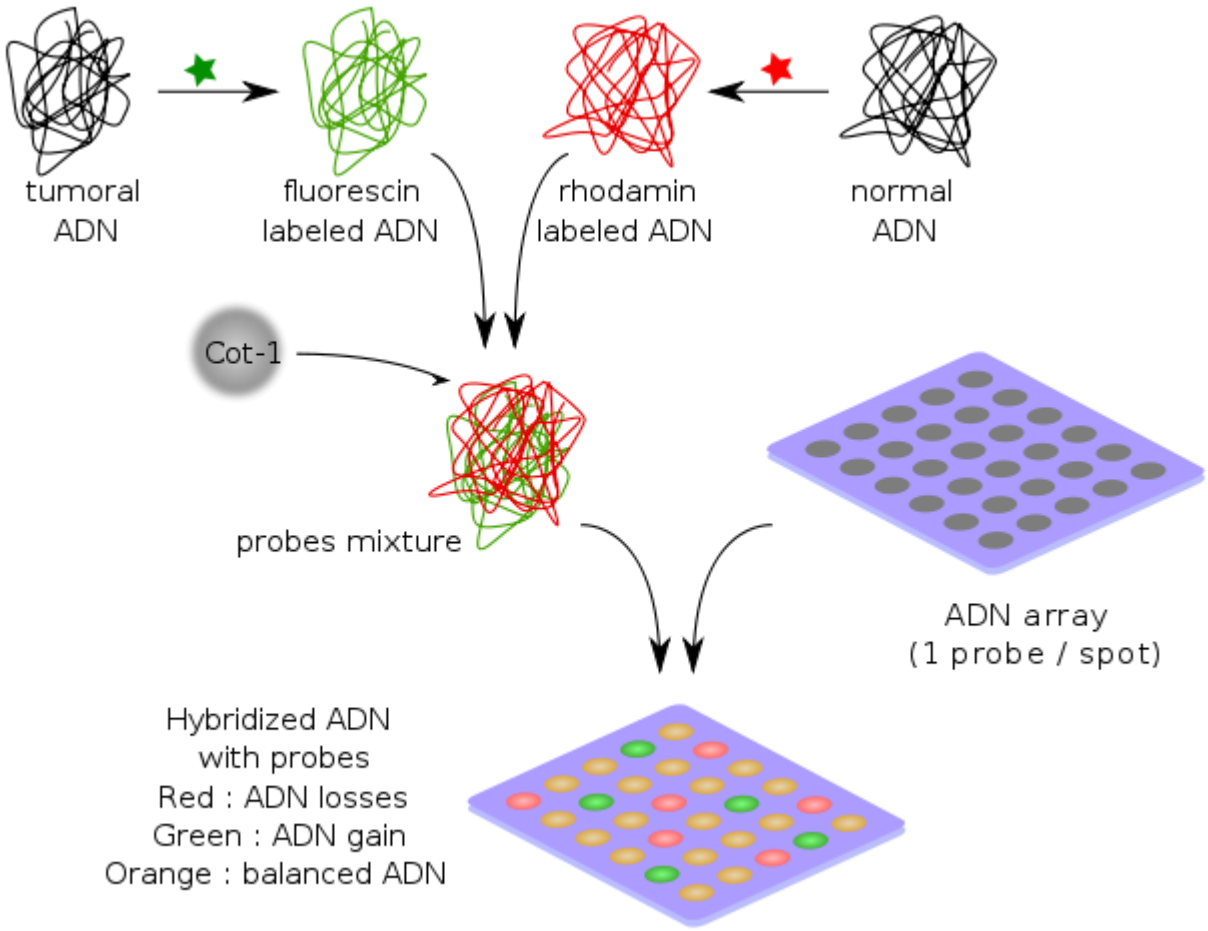
candidats

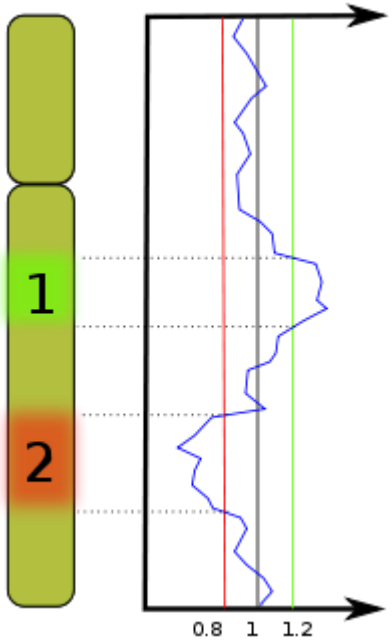
séquences
domaines
régulation
annotations
interactions
expression

fusion

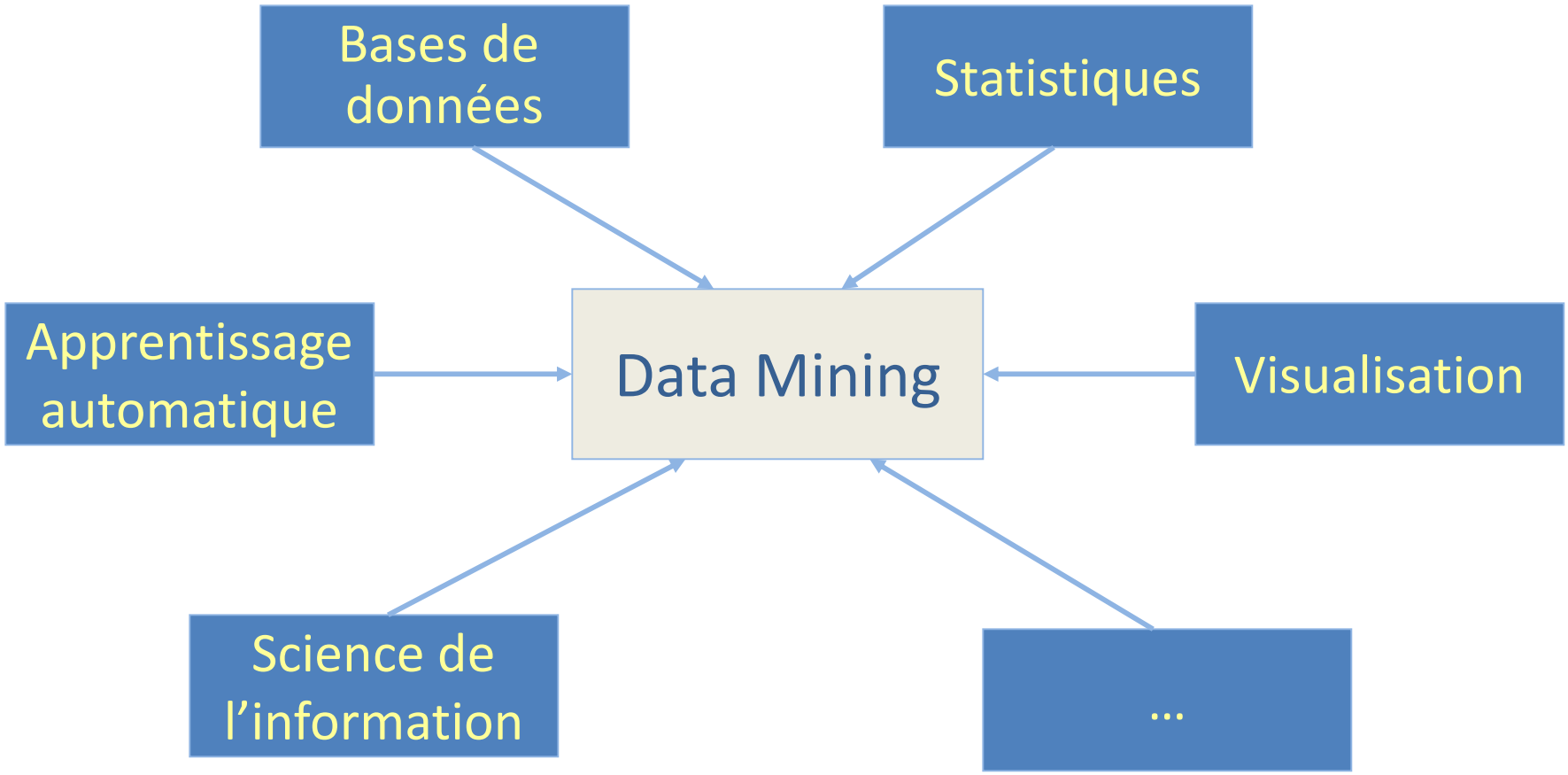
Prioritisation globale

- # Comparative Genomic Hybridization





Data Mining : union de disciplines variées



Pour faire quoi ?

- 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
  - Réseaux sociaux

- Quelles sources de données ?
  - Transactions bancaires (CB), coupons de réduction, cartes de fidélité, service clients (plaintes), et aussi les études publiques de style de vie, historique de navigation
- Cible marketing
  - Trouver des groupes « modèles » de clients qui partagent les mêmes caractéristiques : intérêts, revenus, habitudes de consommation, *etc*.
    - ex : famille de gènes associées à des processus biologiques
- Déterminer les profils d'achat des clients au cours du temps
  - Ex : compte joint après le mariage, assurances
- 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) et autres omics
  - Interaction protéiques
  - réseau métabolique
  - régulation génétique
  - PubMed
- Applications
  - Prédiction de structure 3D
  - Prédiction des séquences codantes
  - Prédiction de fonction, d'interaction, de localisation, …
  - Découverte de motifs sur/sous représentés, répétitions
  - Analyse de données d'expression
  - Aide au diagnostic, médecine personnalisée
  - Méthodes de classification, clustering, *etc*.

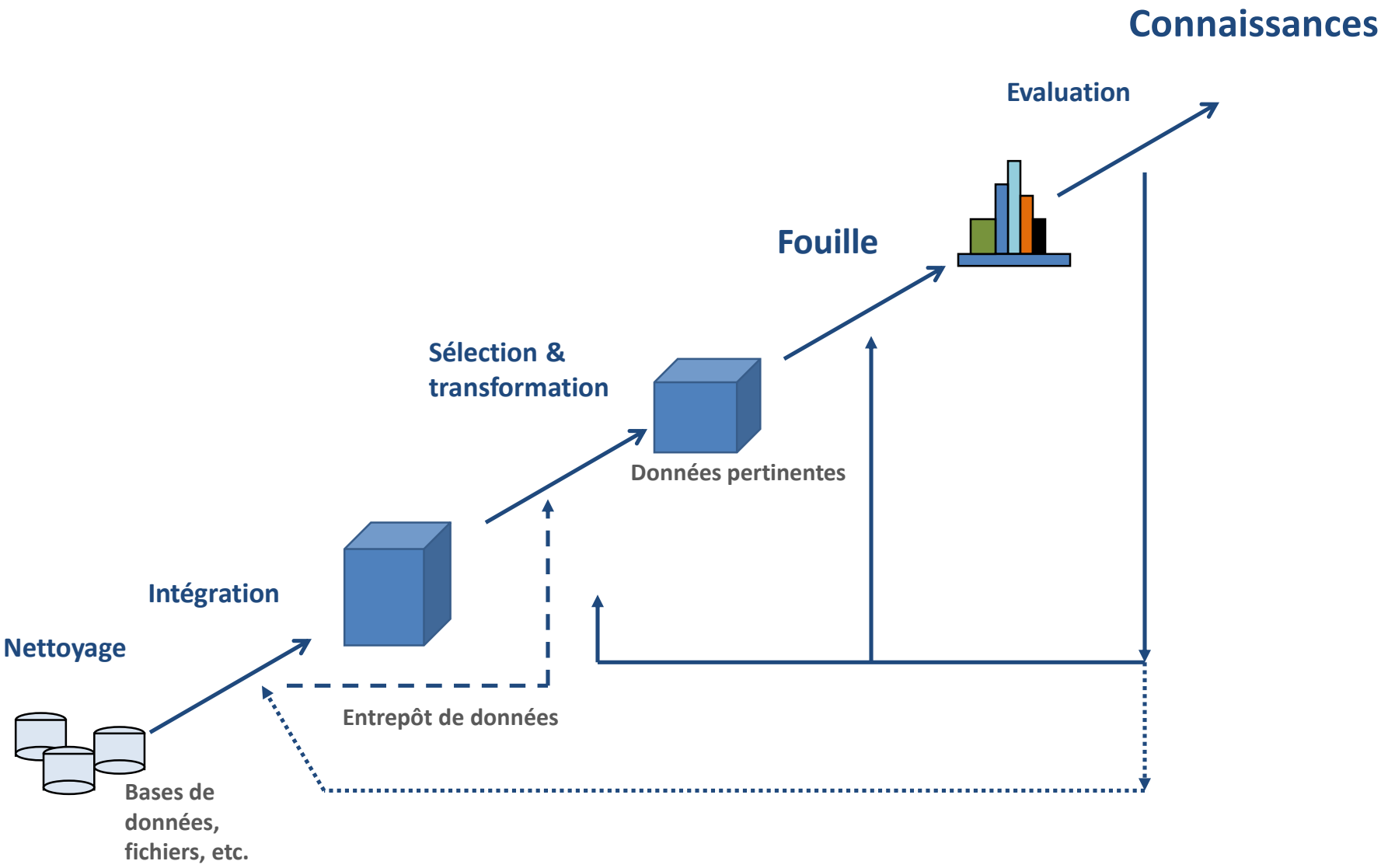
Autres applications

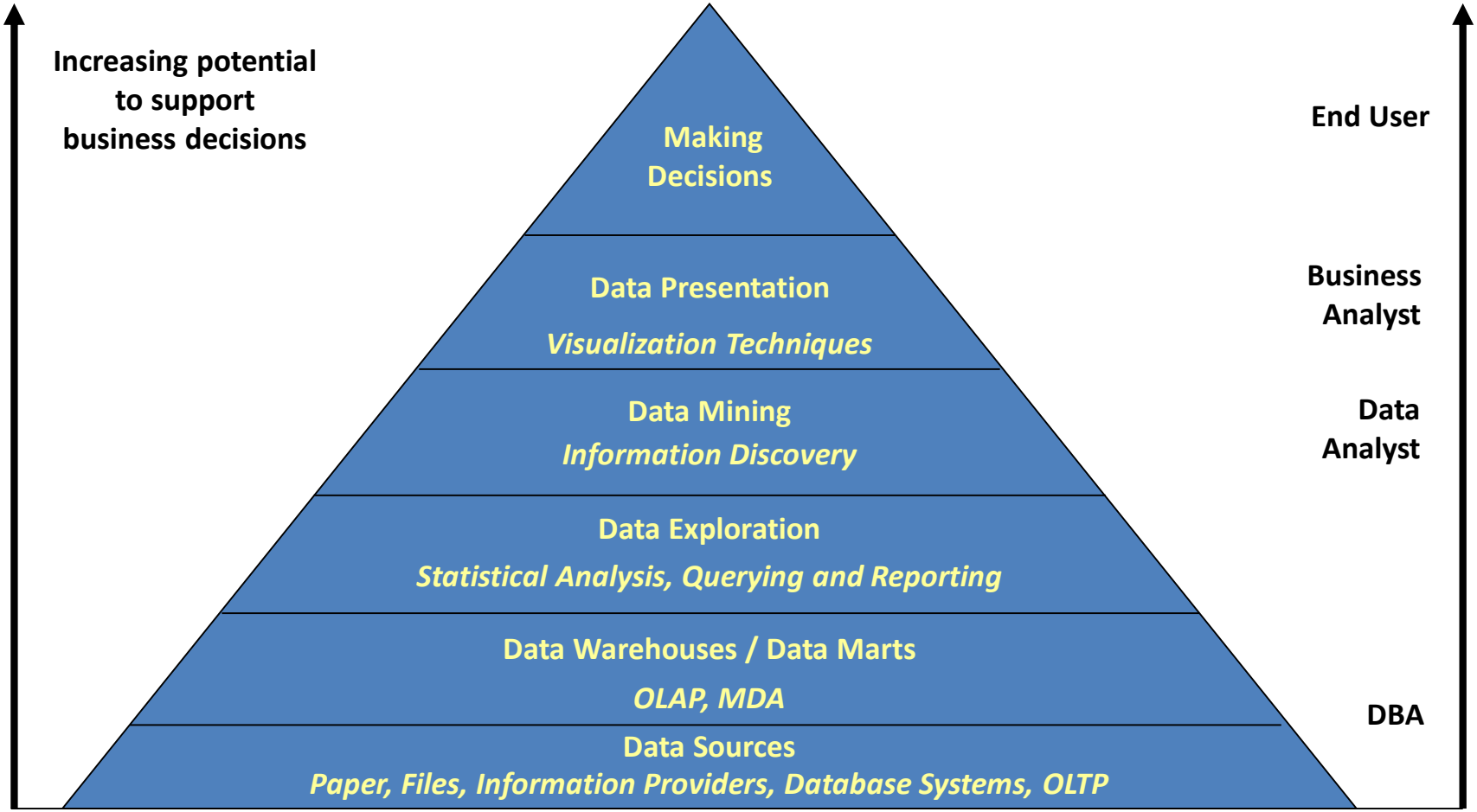
- Astrophysique
  - Découverte de 22 quasars par le Jet Propulsion Laboratory et le Palomar Observatory
- Organisation de sites Web
  - Algorithmes de data mining appliqués aux journaux d'accès aux pages commerciales afin d'identifier les préférences et les comportements des clients et d'analyser les performances du marketing Web et l'organisation du site. Ex: IBM Surf-Aid, GoogleAnalytics
- Réseaux sociaux
  - Qu'est-ce qui fait le buzz
  - Prédiction de l'orientation sexuelle à partir des contacts Facebook (Source http://www.kdnuggets.com/news/2009/n18/28i.html)

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

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
  - Twits

- Description de concepts : Caractérisation et discrimination
  - Généraliser, résumer, et contraster les données caractéristiques

    ex : régions sèches vs. Humides, gènes différentiellement exprimés
- Association (corrélation et causalité)
  - Association multidimensionnelle *vs.* monodimensionnelle
  - âge(X, “20..29”) ^ revenu(X, “20..29K”) → achète(X, “trottinette”)

    [support = 2%, confiance = 60%]
  - contient(T, “trottinette”) → contient(x, “lunette”)

    [support = 1%, confiance = 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, domaines, 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

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

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
  - Temps réel : micro-trading, guidage, twits

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 et génomiques)

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'entrepôts) de données
- **Fonctionnalités :** caractérisation, discrimination, association, classification, clustering, analyse des tendances et des outliers, *etc*.