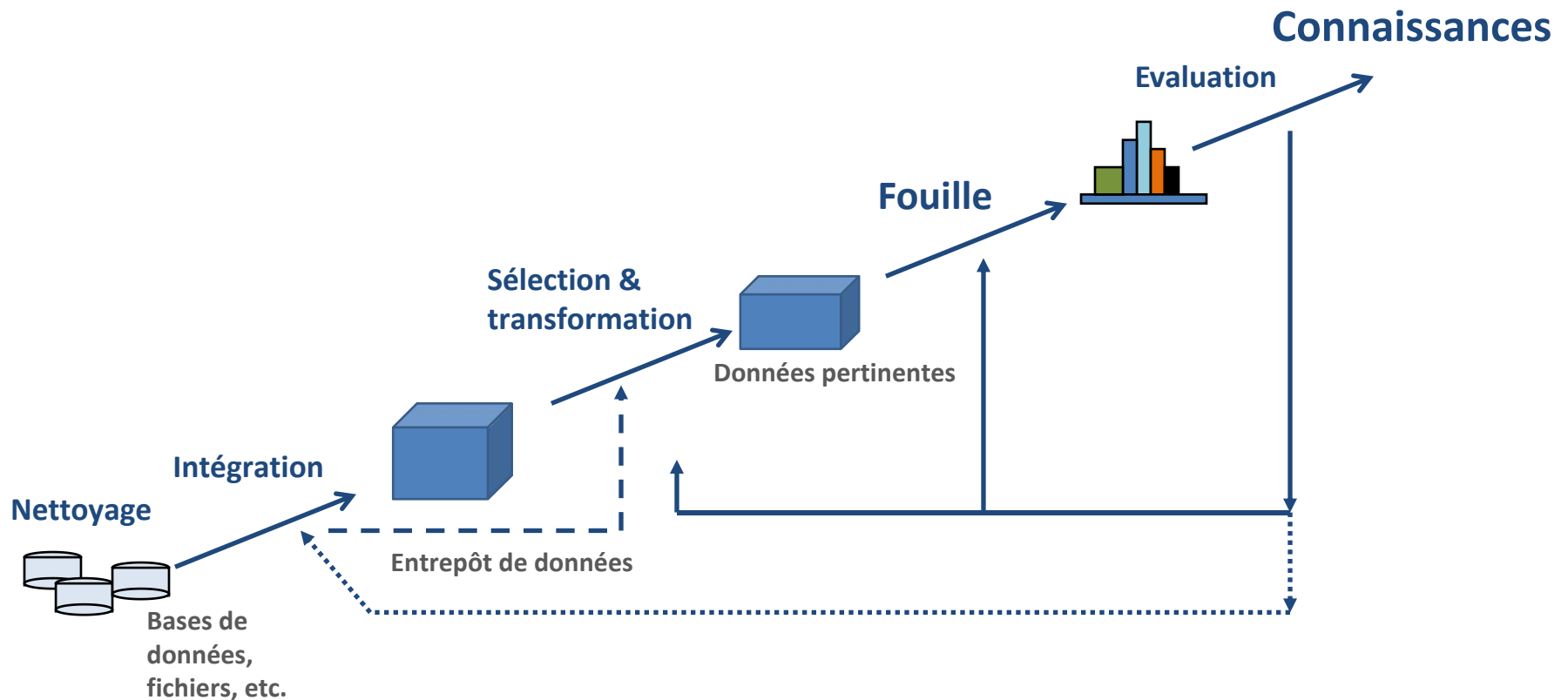


- Dérroulement
  - Supports cours/TP sur le site [silico.biotoul.fr](http://silico.biotoul.fr)
  - 14h cours + 16h TP
- Objectifs
  - Aperçu de la discipline, regard critique
  - Utilisation de méthodes existantes
  - Implémentation et évaluation de certaines méthodes
- Evaluation
  - CC 40% : projets, 9 groupes de 2 personnes
  - CT 60% : exam 2h

- Définition :

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



- 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, ...
  - “Optimisation” (navigation site Web, ergonomie des applications, publicités ciblées)
- Fouille : comment extraire du sens ?
  - Profils clients, publicité, prospection, sécurité du territoire, bourse, risque économique, santé, élections, ...
  - Météo, climat
  - Modèles et hypothèses scientifiques

## Tâches et méthodes

---

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

- **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, NGS (génomés, RNAseq, CHIPseq, ...)
- **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

---

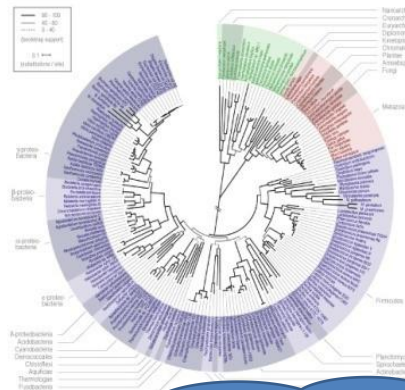
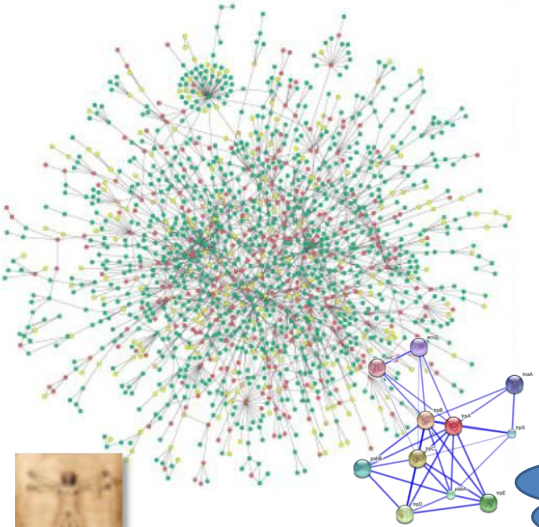
- 1960 :  
Systèmes de gestion de fichiers, collection de données, bases de données (modèle réseau)
- 1970 :  
Émergence du modèle relationnel et de son implémentation
- 1980 :  
SGBD relationnels, modèles avancés (relationnel étendu, OO, déductif, etc.) et orientés application (spatial, scientifique)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

## Qu'est ce que le data mining ?

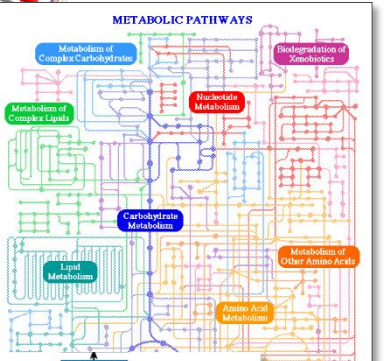
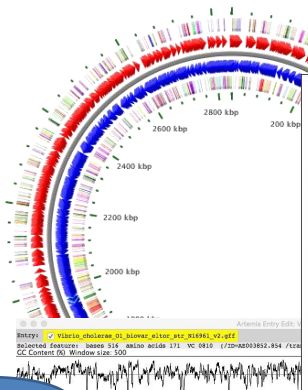
---

- 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

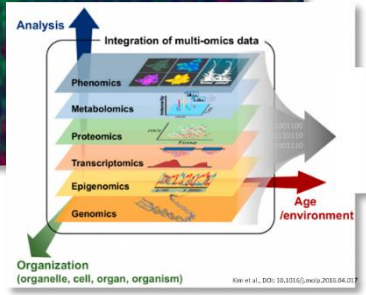
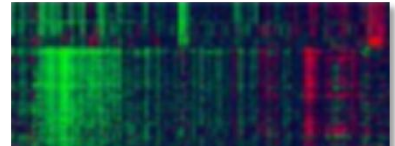
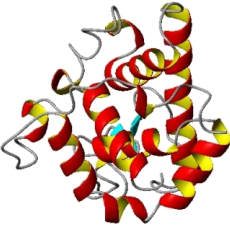
# Un déluge de données



Agrobacterium tumefaciens strain C58 circular chromosome, co...



Comment exploiter au mieux ces données ?



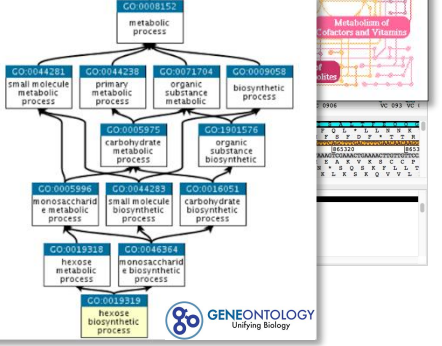
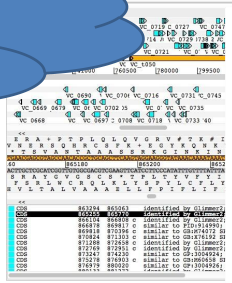
Nucleic Acids Research Advance Access published May 28, 2008  
 Nucleic Acids Research Advance Access published May 28, 2008  
 Nucleic Acids Research Advance Access published May 28, 2008

**ENDEAVOUR update: a web resource for gene prioritization in multiple species**  
 Léon-Charles Tranchesi<sup>1</sup>, Roland Barret<sup>1</sup>, Xia Wu<sup>2</sup>, Steven Van Ypersele<sup>1</sup>, Peter Van Looy<sup>3,4</sup>, Bert Coessens<sup>4</sup>, Bart De Moor<sup>4</sup>, Steen Aerts<sup>4</sup> and Yves Moreau<sup>1\*</sup>

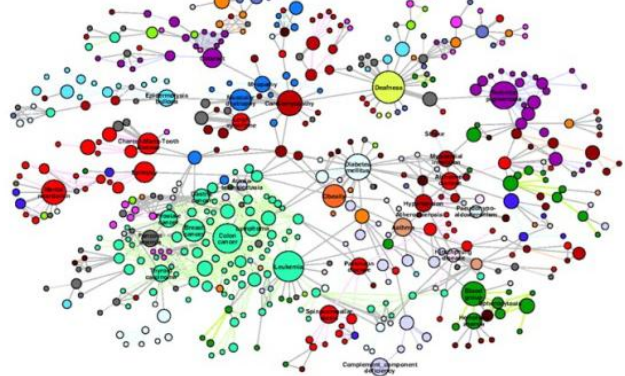
<sup>1</sup>Department of Electrical Engineering (ESAT-ICD), Katholieke Universiteit Leuven, <sup>2</sup>Human Genome Laboratory, Department of Medical and Developmental Genetics, VIB Leuven, <sup>3</sup>Department of Human Genetics, Radboud University Nijmegen School of Medicine and <sup>4</sup>Laboratory of Neurogenetics, Department of Molecular and Developmental Genetics, VIB Leuven (Belgium)

Received February 3, 2008; Revised April 30, 2008; Accepted May 7, 2008

**ABSTRACT** **BACKGROUND**  
 With the recent improvements in high-throughput technologies, large amounts of data have been generated and, more importantly, analyzed. This process leads to the generation of a large amount of genetic data and the creation and maintenance of corresponding databases. However, existing genetic data also facilitate knowledge in identifying genes involved in a particular process or disease through a major challenge: how to search for genes that are most relevant to a particular process or disease. We have developed a web resource that allows users to search for genes that are most relevant to a particular process or disease. It includes a list of genes and their associated phenotypes, a list of genes and their associated phenotypes, and a list of genes and their associated phenotypes.



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

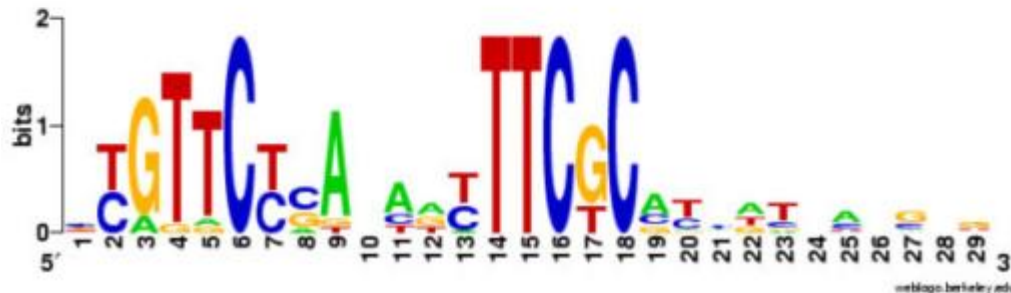
```

>HHT2_HHF2_IR_S_cerevisiae
314 ccgttccgagcacttcgcattaagcgcgt 286 - (25.64520) (5.586871e-10)
381 acgttctgggagcttcgcgtctcaagcct 409 + (9.891990) (1.132719e-02)
349 ctagaccgagagttcgcatattgtatggc 377 + (8.797430) (2.536818e-02)
>HTA2_HTB2_IR_S_cerevisiae
347 ctgtgcccaaccgttcgcctaataaagcg 375 + (27.34320) (3.455151e-11)
334 gtgttcccattattttctcaaagtgatgcg 306 - (13.35890) (7.270649e-04)
376 gtgttctcaaaattttctccccgttttcag 404 + (10.77820) (5.300732e-03)
291 gtgttctcctgaaatttcgcatcactttgag 319 + (10.14830) (8.379708e-03)
>HTB1_HTA1_IR_S_cerevisiae
443 ccattccaatagcttcgcacagtgagggcg 415 - (25.60310) (7.318539e-10)
400 ctgttccaaaattttcgcctcactgtgcg 428 + (12.09860) (2.459747e-03)
298 ctgttctcactttttcgcgcggttgcaccc 270 - (11.52740) (3.554757e-03)
244 tcgttctcattttttcgcggaagaagggg 272 + (10.85850) (5.873168e-03)
>HHF1_HHT1_IR_S_cerevisiae
278 ctgttccgagcgcttctccccataatggt 250 - (27.53840) (2.261761e-11)
391 tcgttctcacaattttctcacatttcottg 419 + (12.23800) (1.720110e-03)
357 tcgttctcacattttcgcattgtcccata 385 + (11.22860) (3.671231e-03)
313 gcgttctcgaaaacttcgcatcttcacata 285 - (8.294790) (2.770252e-02)

```

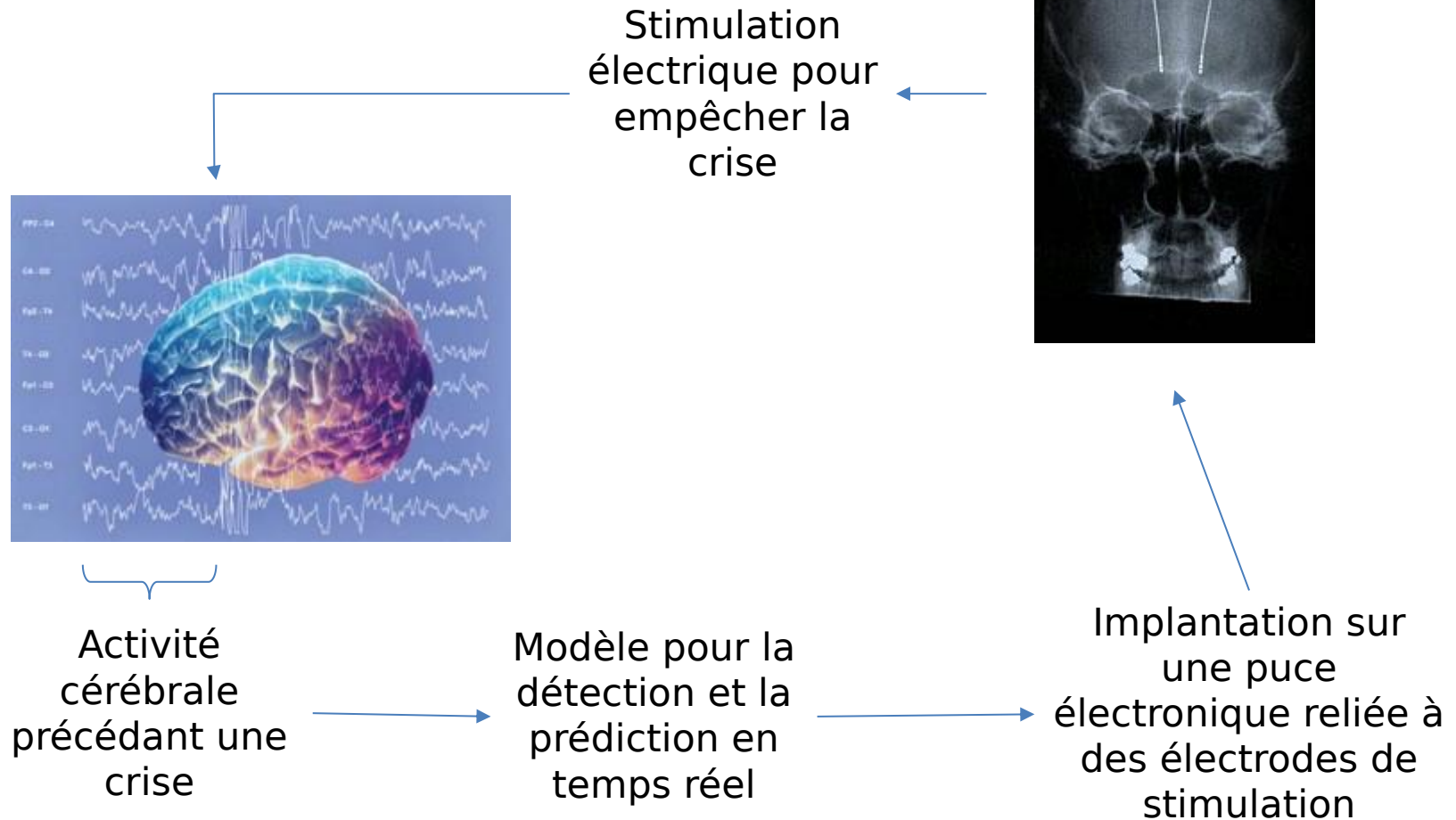
B

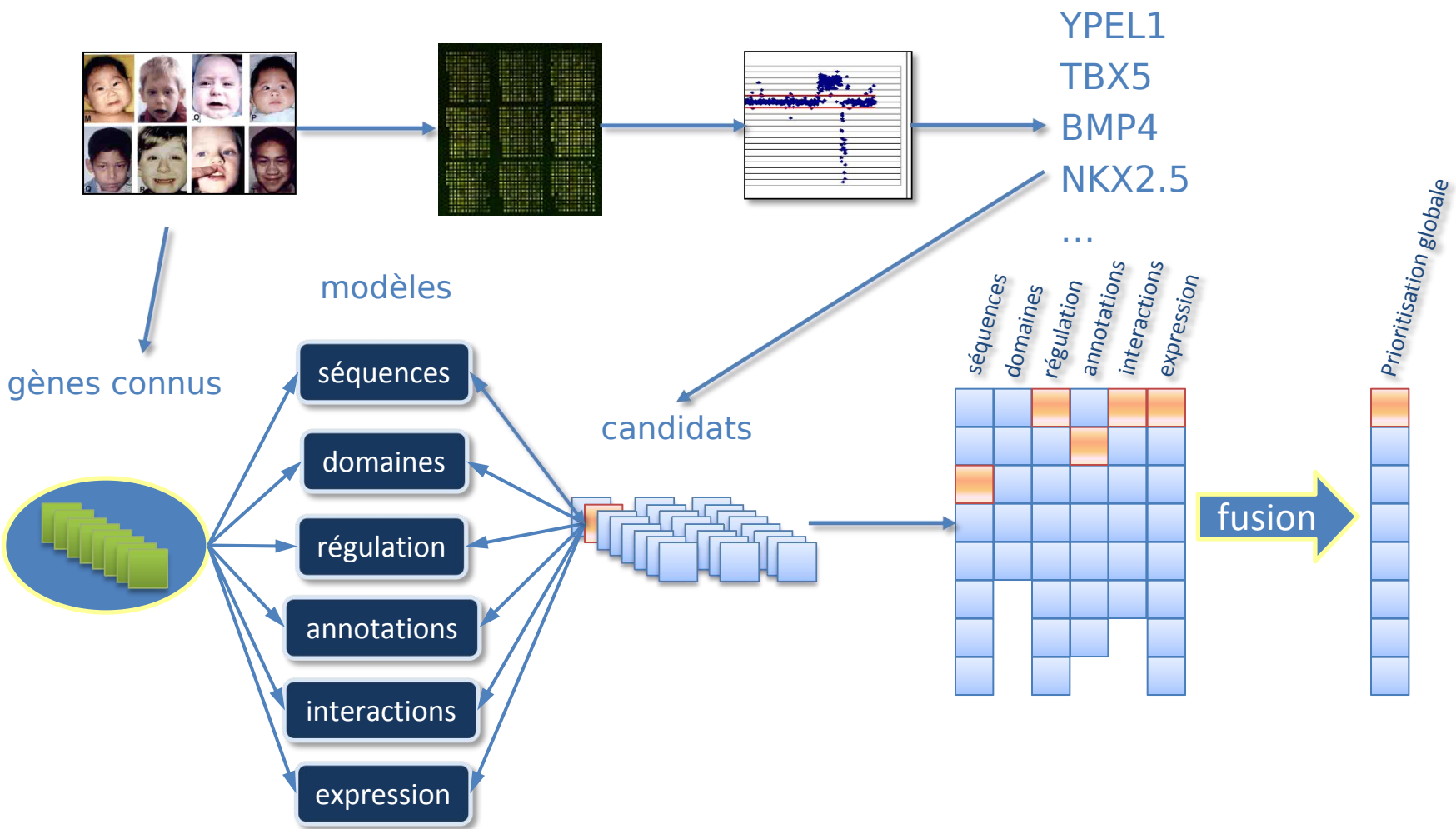
Inférence d'un modèle



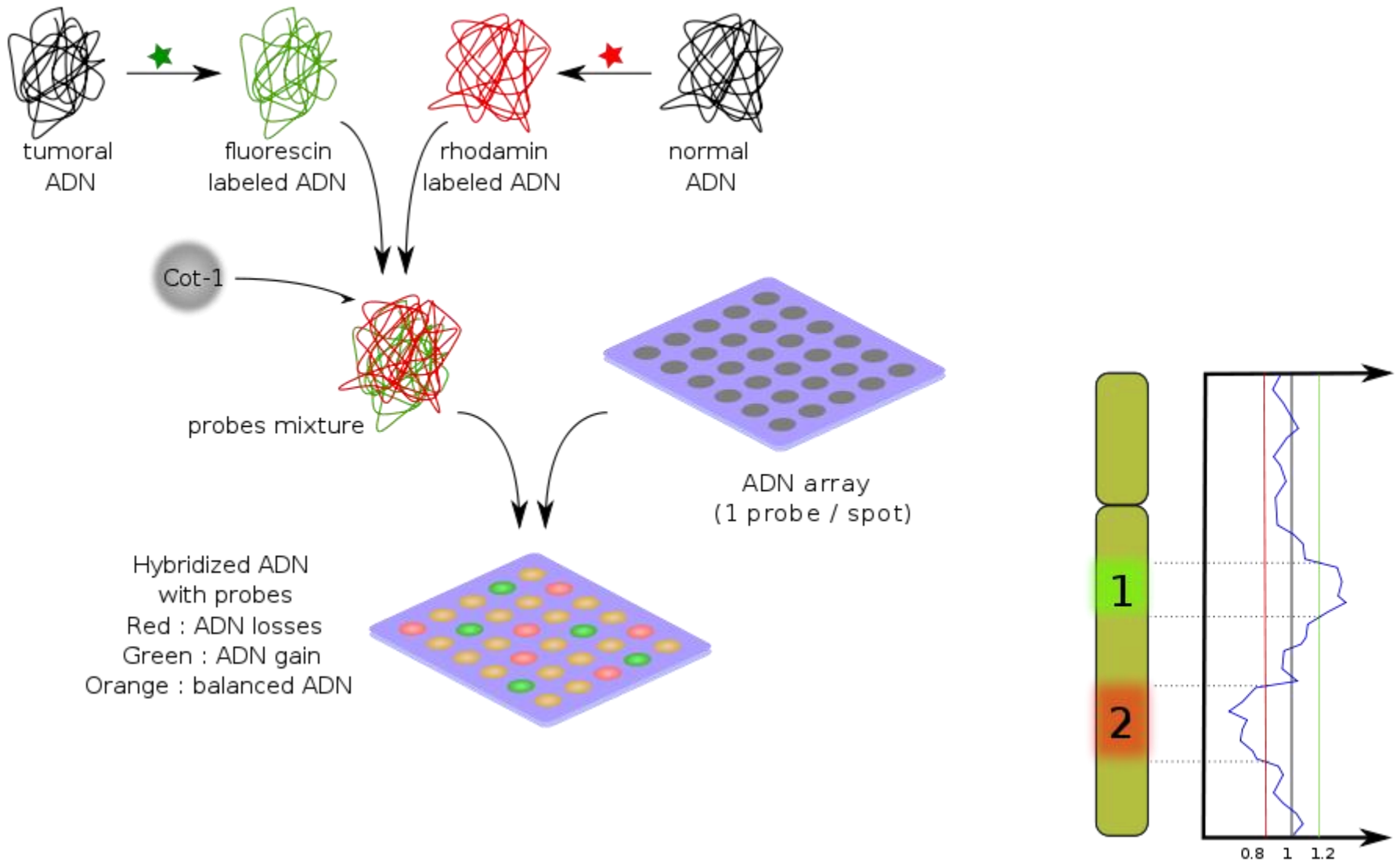
Utilisation du modèle sur des séquences nouvelles pour effectuer des prédictions (ex: sites de fixation de facteur de transcription)

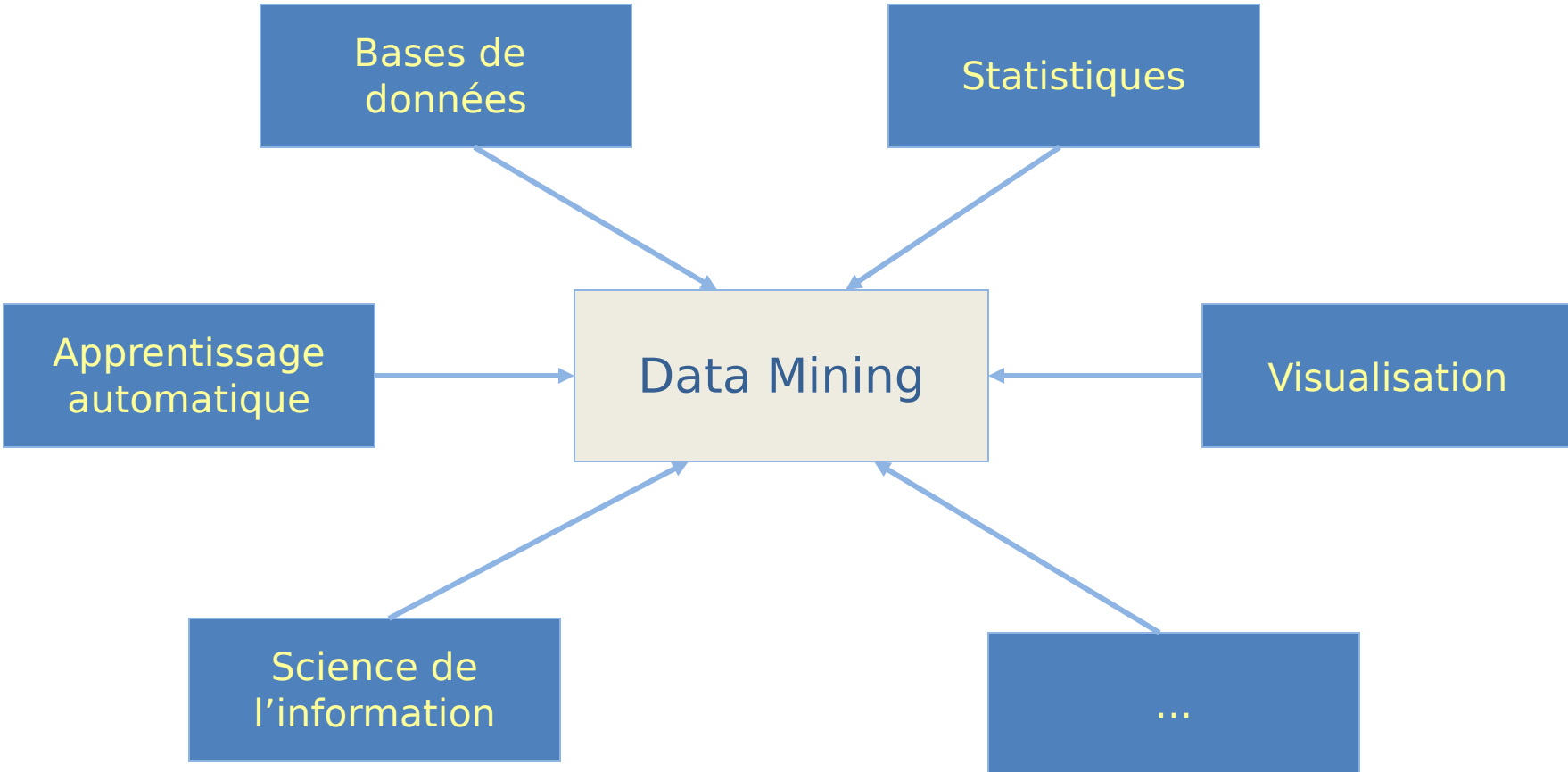
- Maladie de Parkinson
  - Crise de tremblements





# Comparative Genomic Hybridization



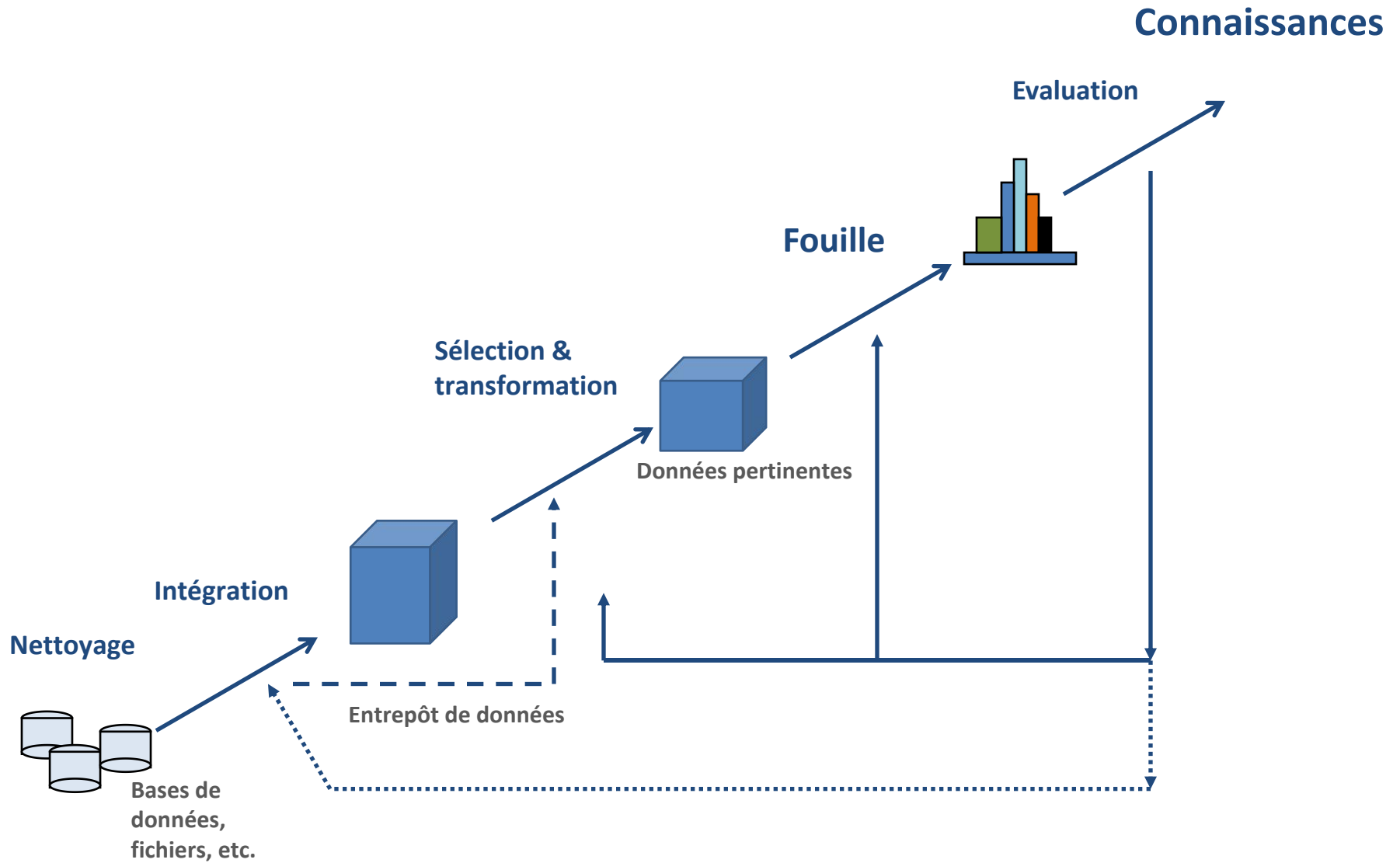


# Étapes impliquées dans le processus de découverte de connaissances

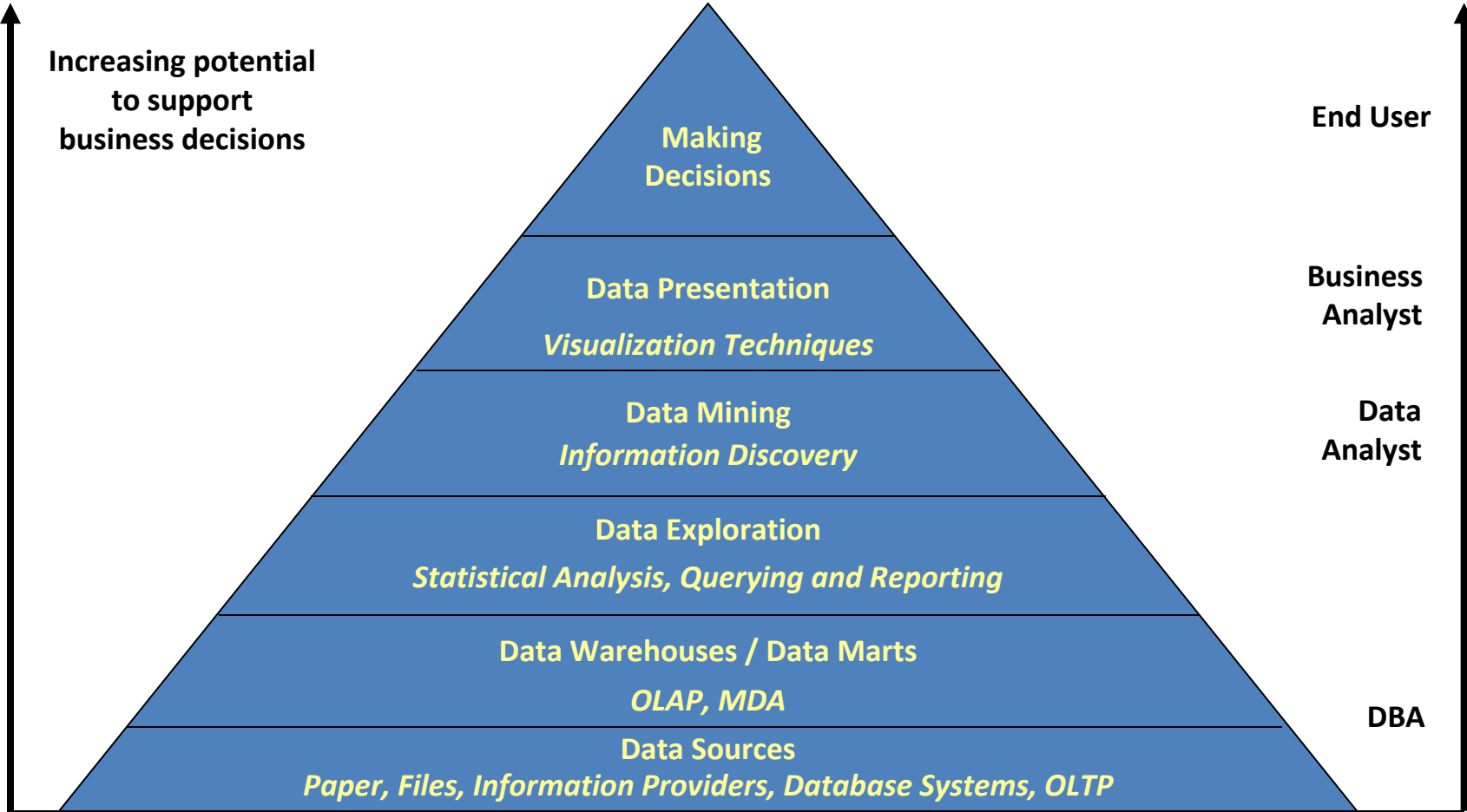
---

- 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







- 
- Description de concepts : Caractérisation et discrimination
    - Généraliser, résumer, et contraster les données caractéristiques  
ex : régions sèches vs. humides, gènes différentiellement exprimés, processus biologiques caractérisant un ensemble de gènes
  - Association (corrélation et causalité)
    - Association mono- vs. multi-dimensionnelle  
contient(T, "raclette") → contient(T, "patate")  
[support = 1%, confiance = 75%]  
âge(X, "20..29") ^ revenu(X, "10..19K€") → loisir(X, "musée")  
[support = 2%, confiance = 60%]

# Exemple pour la caractérisation/généralisation

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC, Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
...	...	...	...	...	...	...	...
<b>Removed</b>	<b>Retained</b>	<b>Sci,Eng, Bus</b>	<b>Country</b>	<b>Age range</b>	<b>City</b>	<b>Removed</b>	<b>Excl, VG,..</b>

RelationInitiale

Gender	Major	Birth_region	Age_range	Residence	GPA	Count
M	Science	Canada	20-25	Richmond	Very-good	16
F	Science	Foreign	25-30	Burnaby	Excellent	22
...	...	...	...	...	...	...

Généralisée Primaire

Tableau croisé

Birth_Region \ Gender	Canada	Foreign	Total
M	16	14	30
F	10	22	32
Total	26	36	62

# Exemple : caractérisation analytique / discrimination

gender	major	birth_country	age_range	gpa	count
M	Science	Canada	20-25	Very_good	16
F	Science	Foreign	25-30	Excellent	22
M	Engineering	Foreign	25-30	Excellent	18
F	Science	Foreign	25-30	Excellent	25
M	Science	Canada	20-25	Excellent	21
F	Engineering	Canada	20-25	Excellent	18

**Relation candidate pour la classe cible : 3<sup>ème</sup> cycle ( $\Sigma = 120$ )**

gender	major	birth_country	age_range	gpa	count
M	Science	Foreign	<20	Very_good	18
F	Business	Canada	<20	Fair	20
M	Business	Canada	<20	Fair	22
F	Science	Canada	20-25	Fair	24
M	Engineering	Foreign	20-25	Very_good	22
F	Engineering	Canada	<20	Excellent	24

**Relation candidate pour la classe contrastante : 1<sup>er</sup> et 2<sup>nd</sup> cycle ( $\Sigma = 130$ )**

Analyse de pertinence  
→ pouvoir discriminant des attributs



major	age_range	gpa	count
Science	20-25	Very_good	16
Science	25-30	Excellent	47
Science	20-25	Excellent	21
Engineering	20-25	Excellent	18
Engineering	25-30	Excellent	18

puis caractérisation / généralisation

---

- Classification et Prédiction

- Construire 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

- **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 événements 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, tweets

- **Liées à la diversité des types de données**
  - Données relationnelles et types complexes
  - Bases de données hétérogènes et système global d'informations (WWW)
  - Données peu/non structurées
- **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)

- 
- **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.*