- Déroulement
  - Supports cours/TP sur le site 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

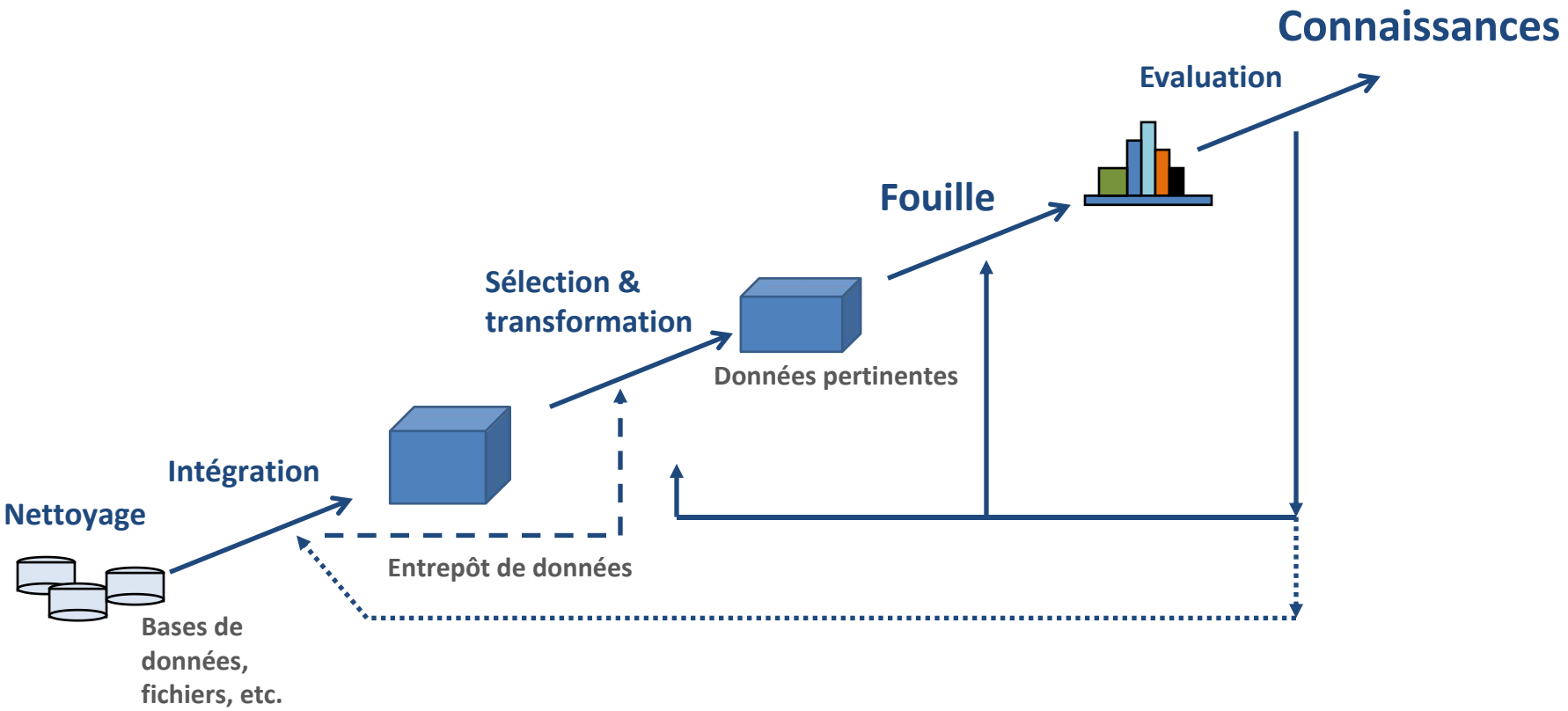
Data Mining = Knowledge Discovery in Databases (KDD) = Fouille de données

- Définition :

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

- 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, …
  - “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

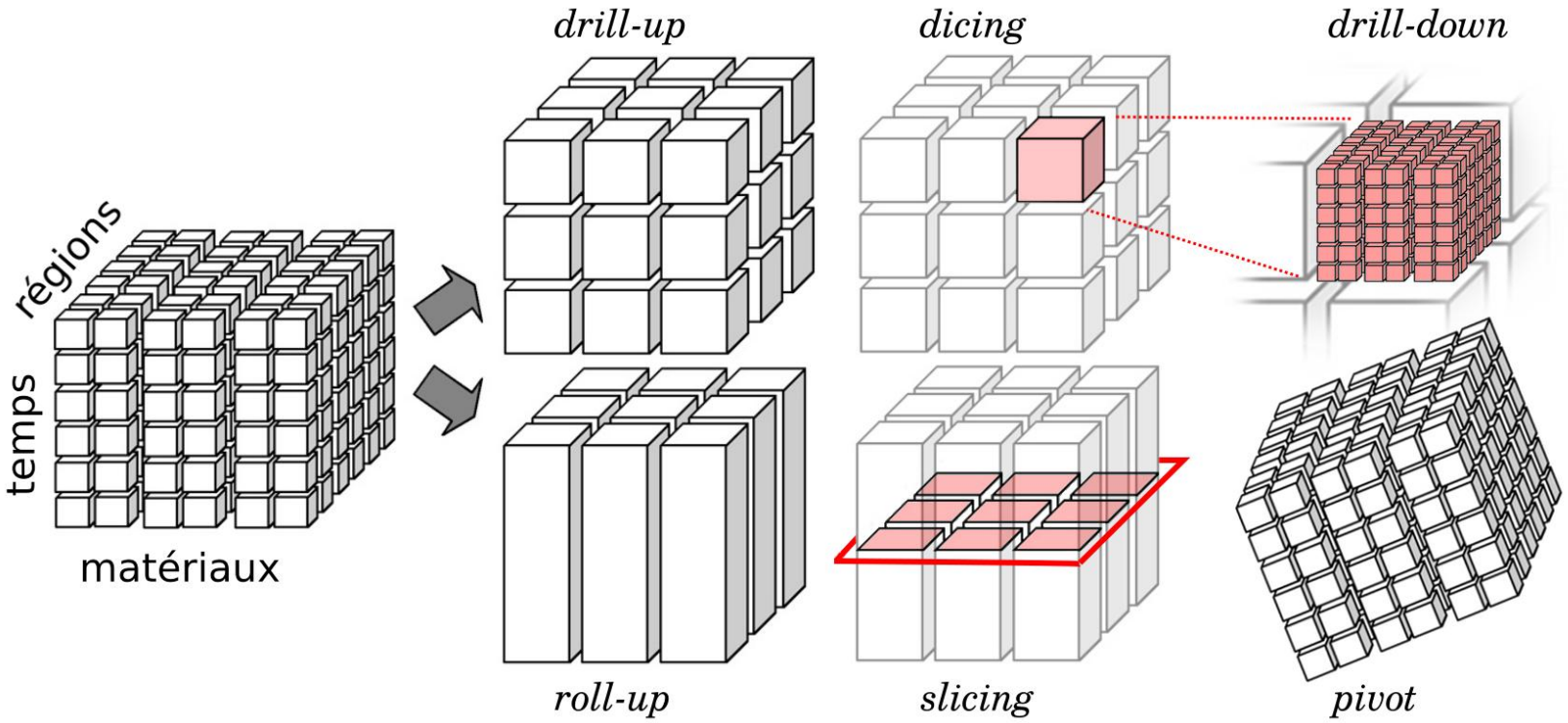
Tâches et méthodes

- 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

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, NGS (génomes, 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

Data warehousing et on-line analytical processing (OLAP)



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, données massive peu/pas structurées (NoSQL)

https://www.kdnuggets.com/

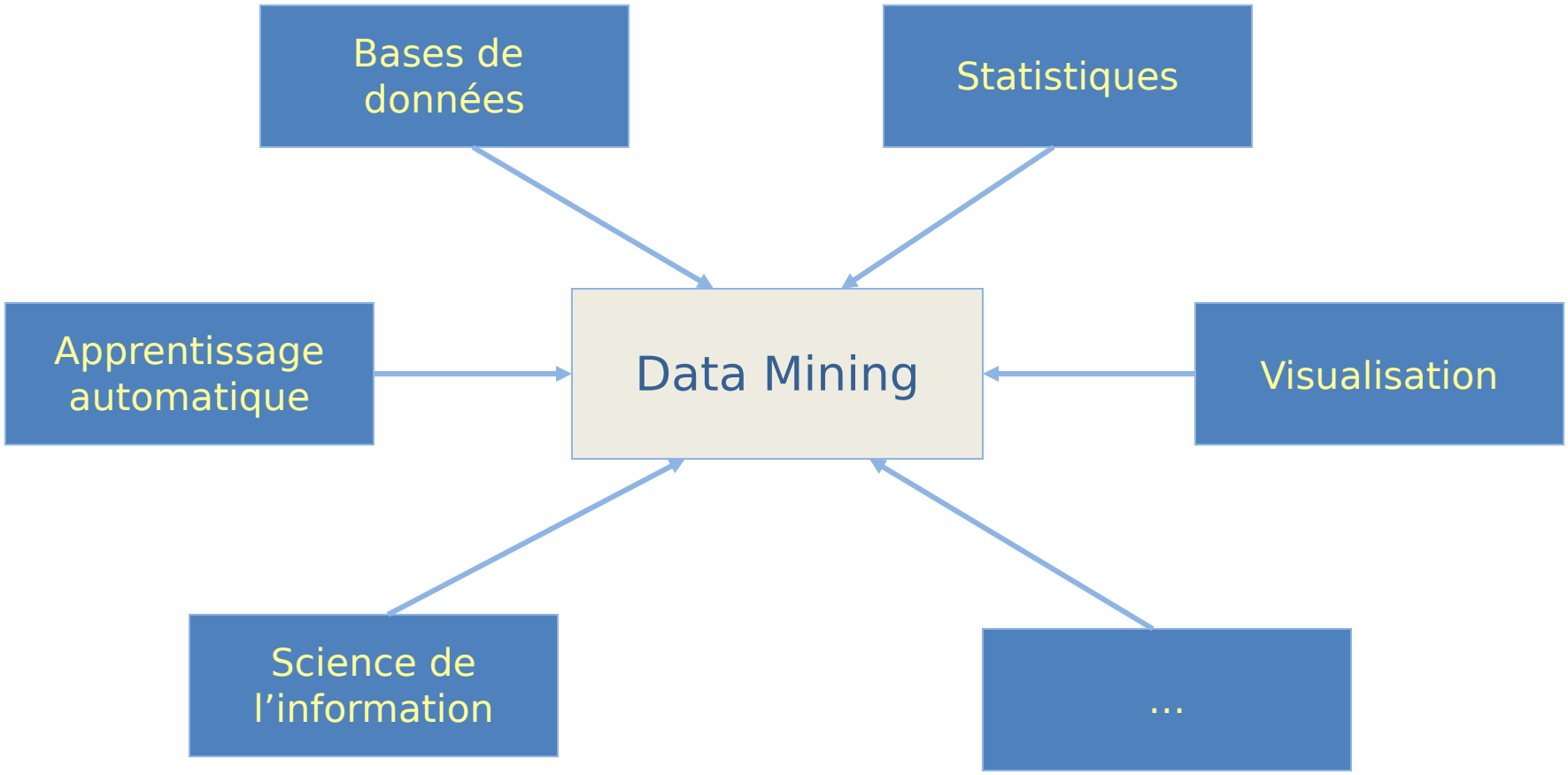
Qu’est ce que le data mining ?

- 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

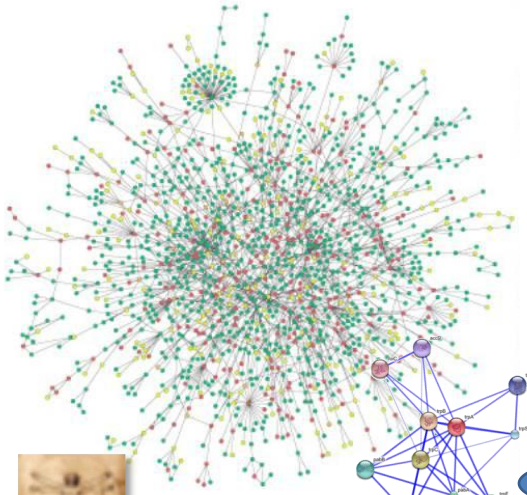
- Autres appellations :
  - Data mining : est-ce judicieux ?
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, information harvesting, business intelligence, apprentissage automatique (machine learning), *etc.*

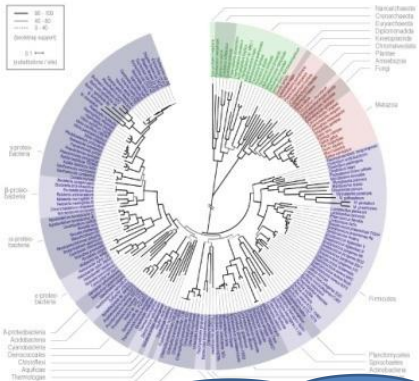
- Ce qui n’est pas du data mining
  - (Deductive) query processing
  - Systèmes experts

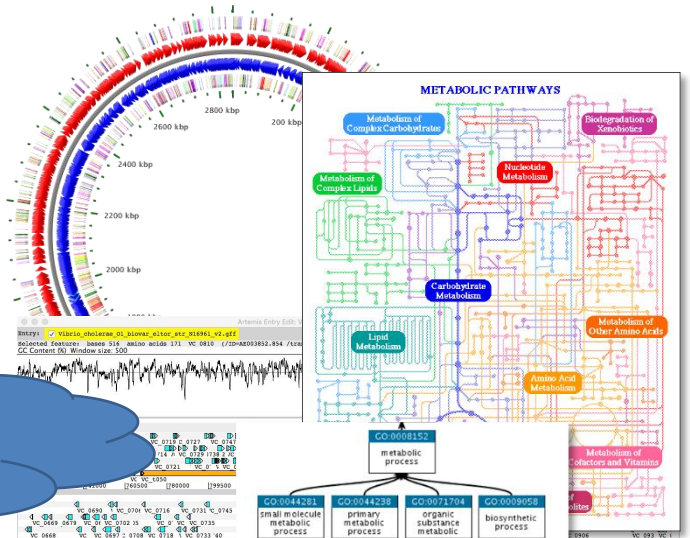
Data Mining : union de disciplines variées



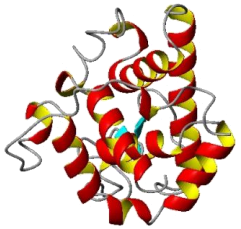
Un déluge de données

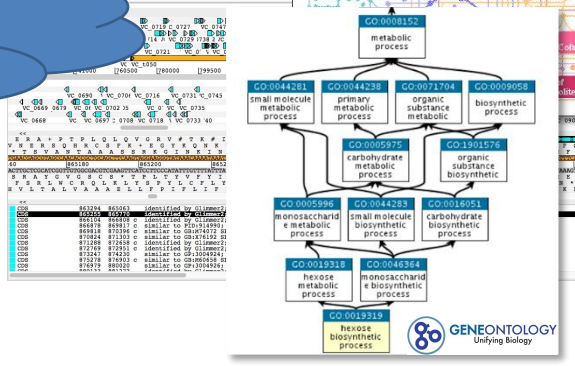


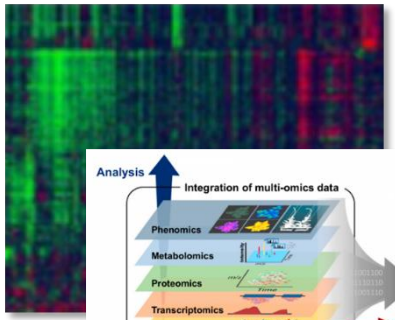


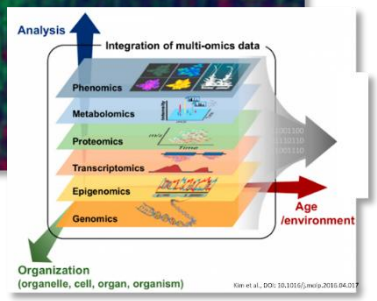


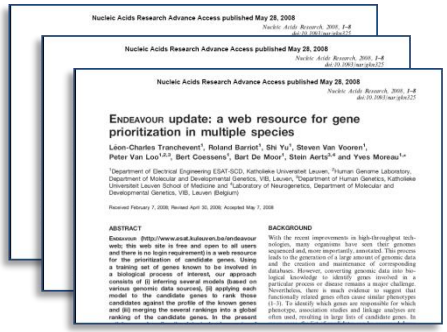
Comment exploiter au mieux ces données ?

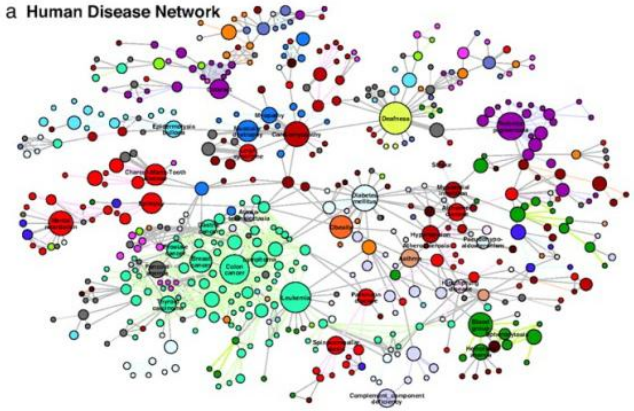




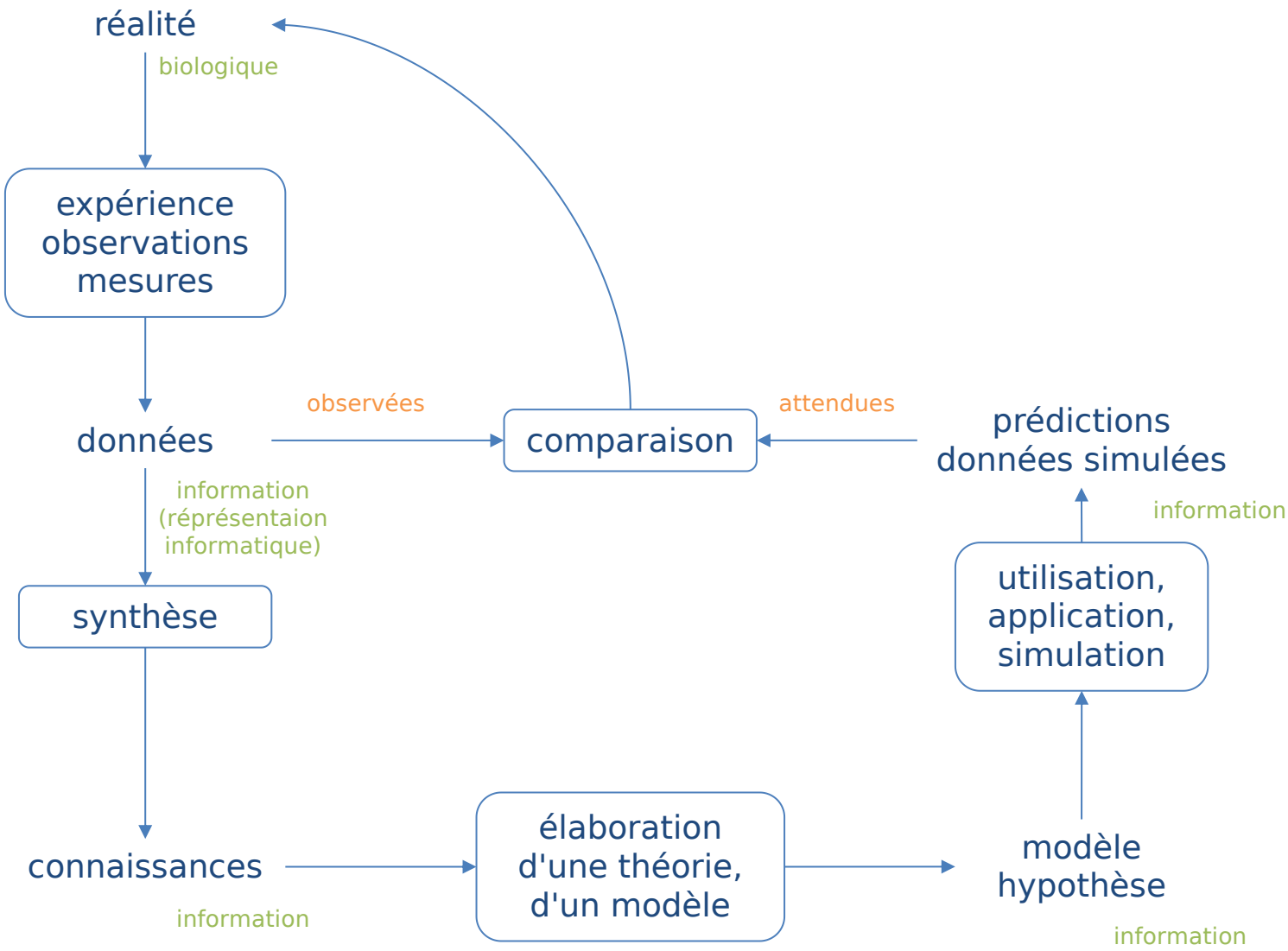




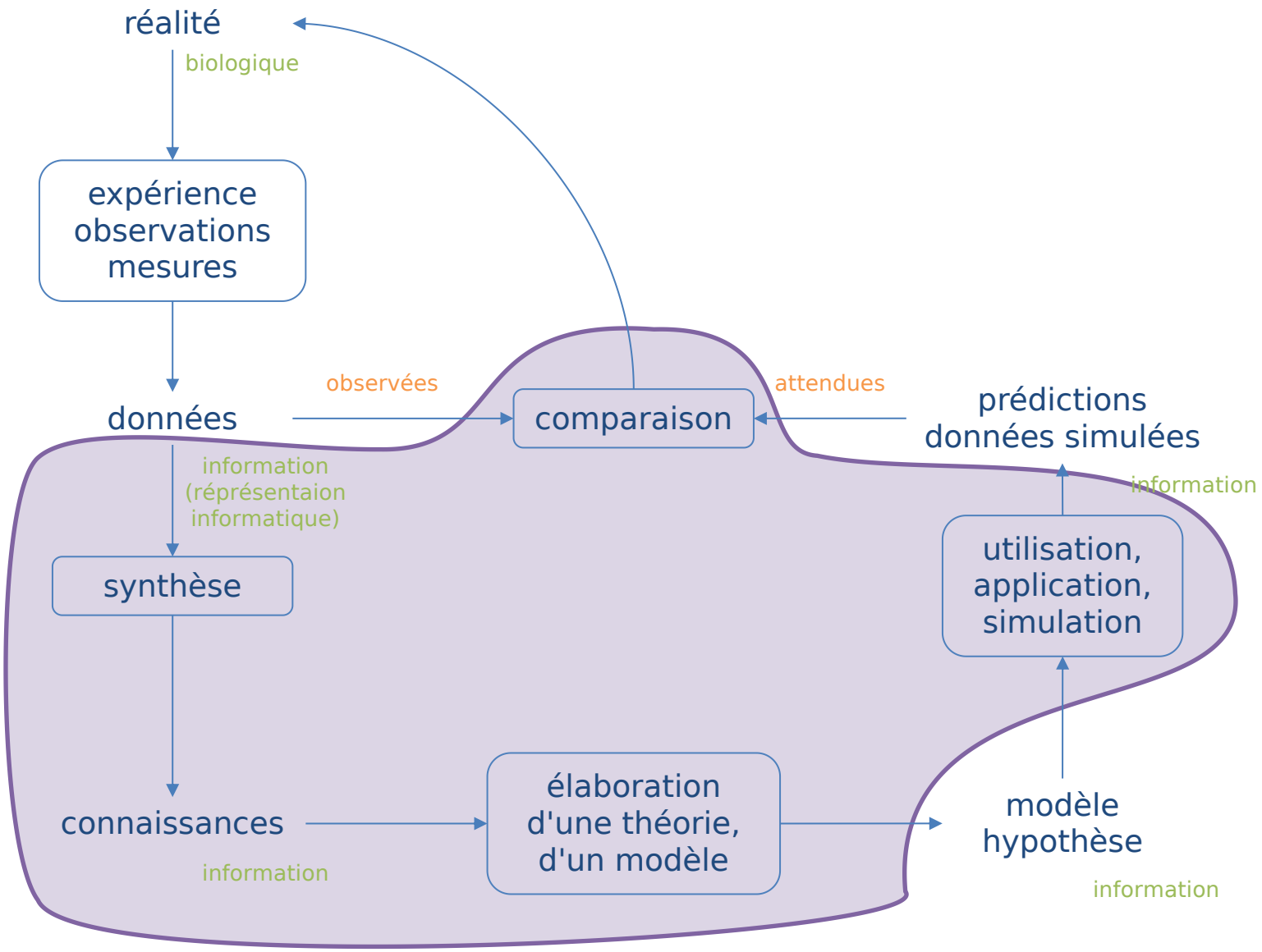




Méthode scientifique



Méthode scientifique



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

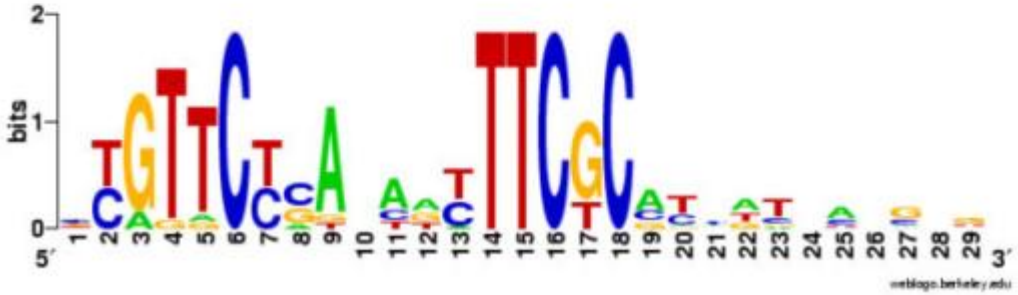
Recherche de motifs intéressants

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 ctgttccaaaatttttcgctcactgtgcg  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 ctgttccgagcgctttctcccataatggt  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



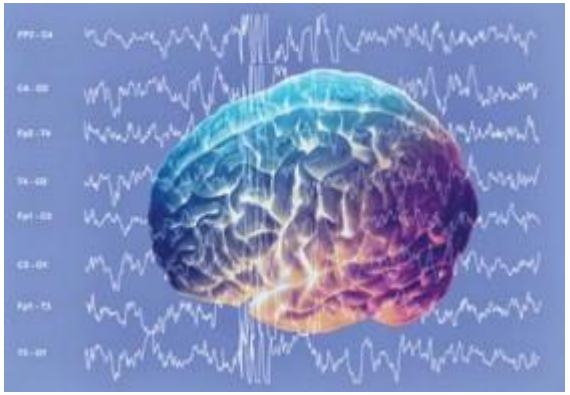
Utilisation du modèle sur des séquences nouvelles pour effectuer des prédictions (ex: sites de fixation de facteur de transcription)

Deep Brain Stimulations

- Maladie de Parkinson
  - Crise de tremblements

PHOTO: Dr. Helen Mayberg

Stimulation électrique pour empêcher la crise

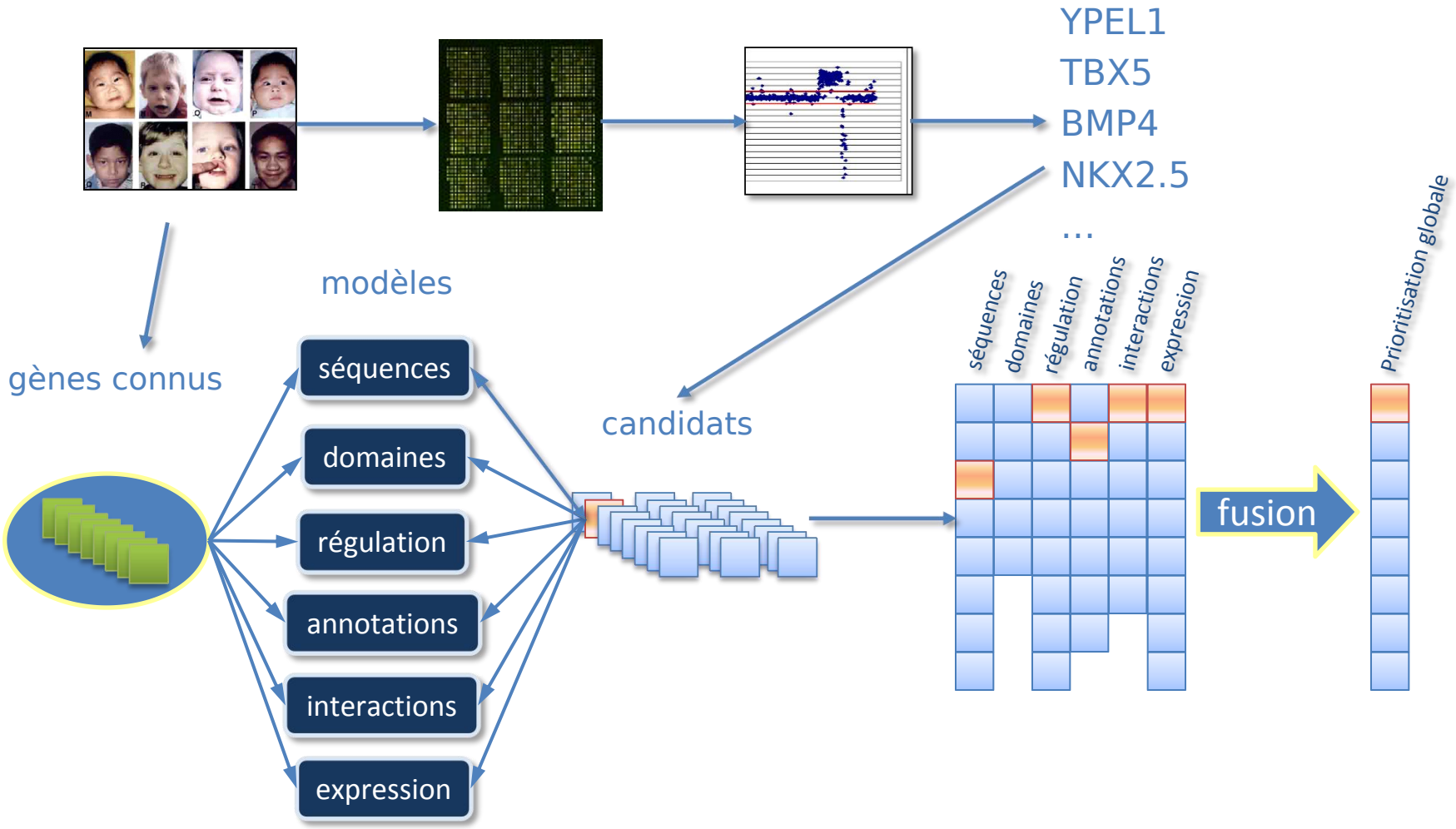


Activité cérébrale précédant une crise

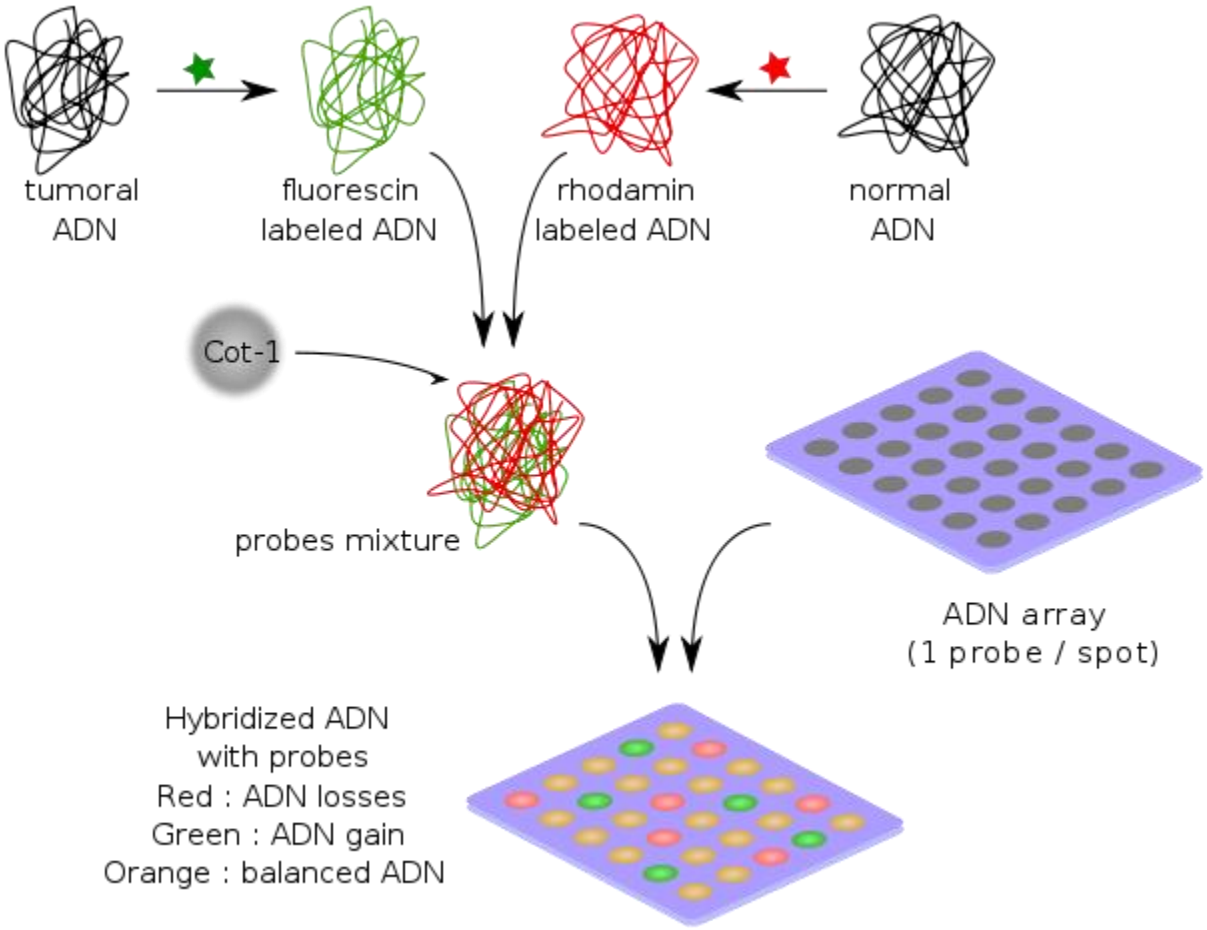
Modèle pour la détection et la prédiction en temps réel

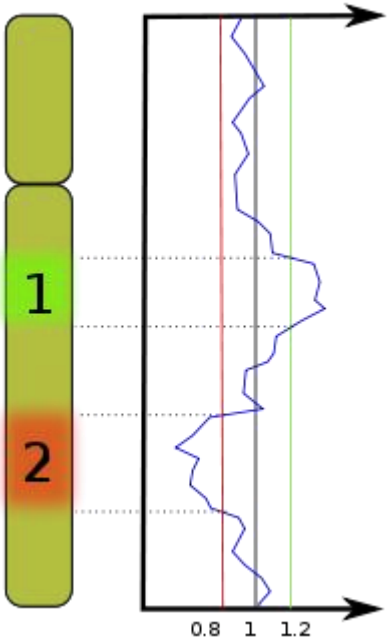
Implantation sur une puce électronique reliée à des électrodes de stimulation

Exploitation des données disponibles



- Comparative Genomic Hybridization

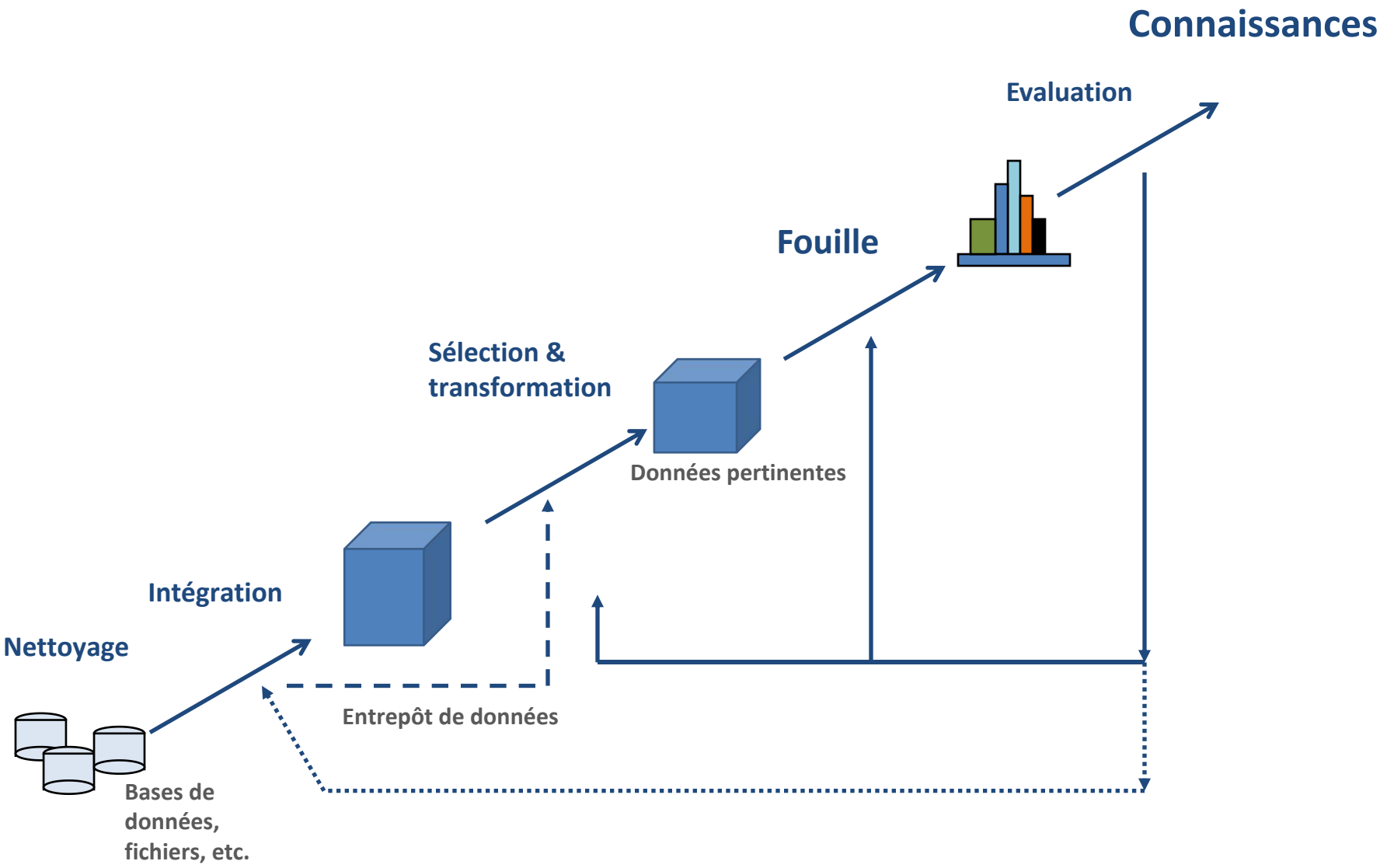




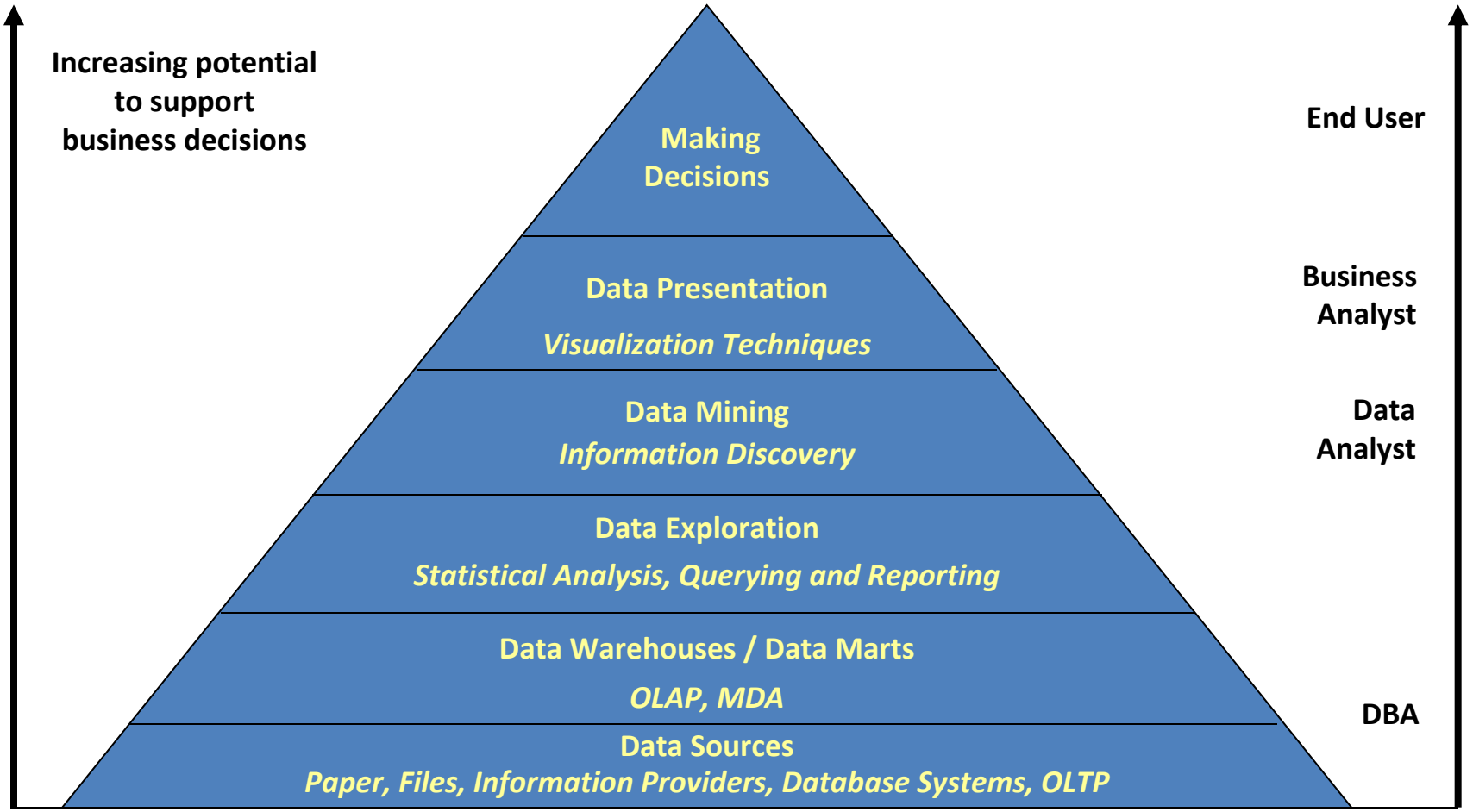
Étapes impliquées dans le processus de découverte de connaissances

- Apprentissage du domaine d'application :
  - Connaissances nécessaires et **objectifs à atteindre**
- 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



Data Mining and Business Intelligence



- 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%]

- 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

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
…	…	…	…	…	…	…	…
Removed	Retained	Sci,Eng, Bus	Country	Age range	City	Removed	Excl, VG,..

Relation initiale

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

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

Tableau croisé

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 : 3ème cycle (Σ =120)

gender	major	birth_country	age_range	gpa	count
M	Science	Foreign	<20	Very_good	188
F	Business	Canada	<20	Fair	200
M	Business	Canada	<20	Fair	222
F	Science	Canada	20-25	Fair	244
M	Engineering	Foreign	20-25	Very_good	22
F	Engineering	Canada	<20	Excellent	24

Relation candidate pour la classe contrastante : 1er et 2nd cycle (Σ =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 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 ?

- **Problème : Un système de data mining peut générer des milliers de patterns/motifs/règles**
  - 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 de tests avec un certain degré de certitude
  - potentiellement utile
  - nouveau
  - ou encore s'il sert à valider une hypothèse que l'utilisateur cherche à confirmer
- **Mesure 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 ou 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

Principaux défis en data mining (2)

- 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)

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*.