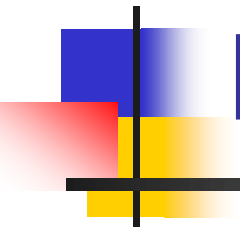


Fouille de données (FD) et Recherche d'Information (RI)

Cours 1 : Généralités

Plan de présentation

- Vocabulaire
- Notion de fouille de données
- Histoire
- Motivations
- Applications
- Types de données
- Tâches de fouille de données
- Recherche d'information
- Exemples

Vocabulaire

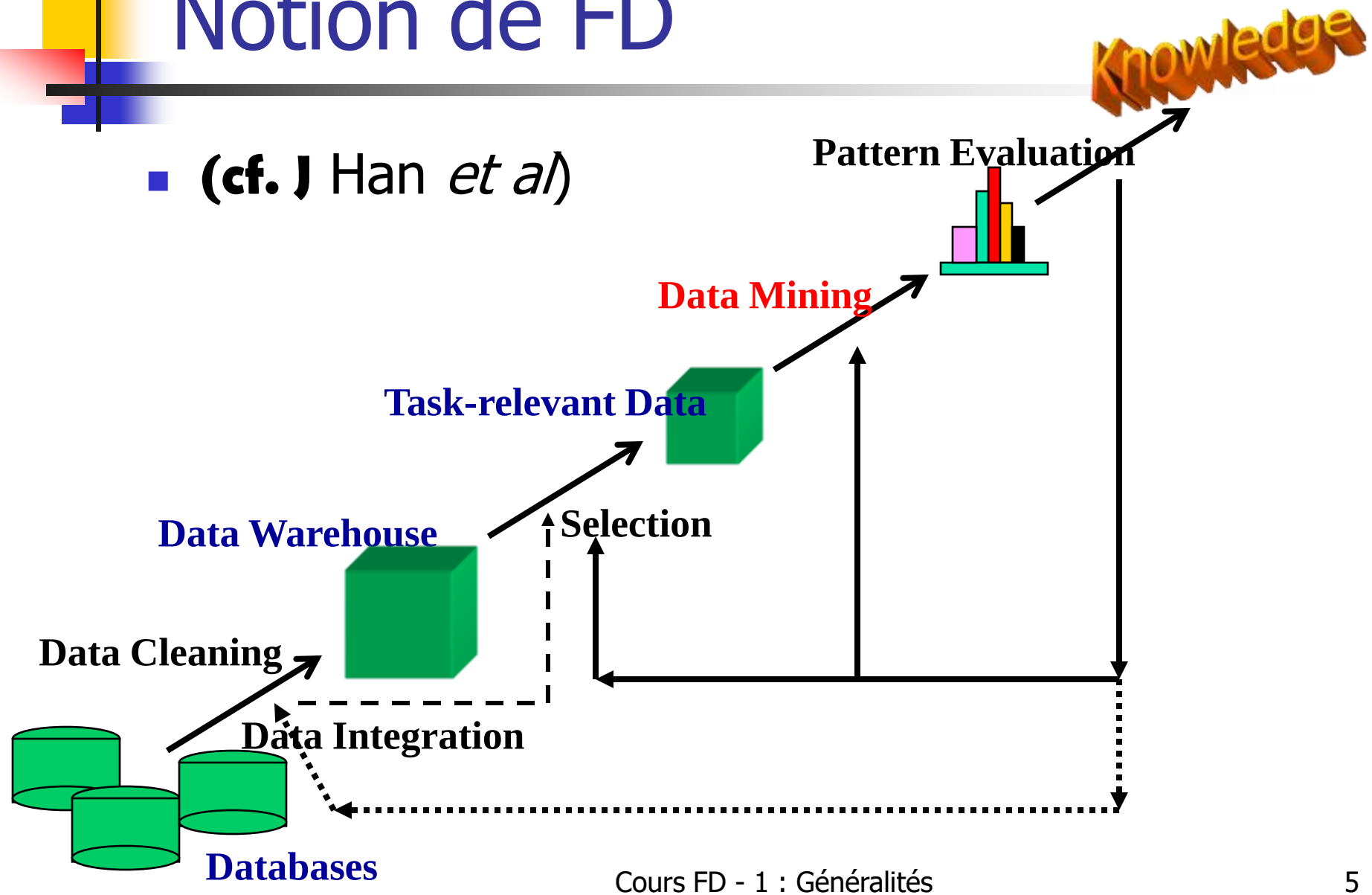
- FD - Fouille de Données / *DM - Data Mining*
- ECD - Extraction des Connaissances à partir des Données / *KDD – Knowledge Discovery in Databases*
- RI – Recherche d'Information / *IR – Information Retrieval*
- EI – Extraction d'Information / *IE – Information Extraction*

Notion de FD

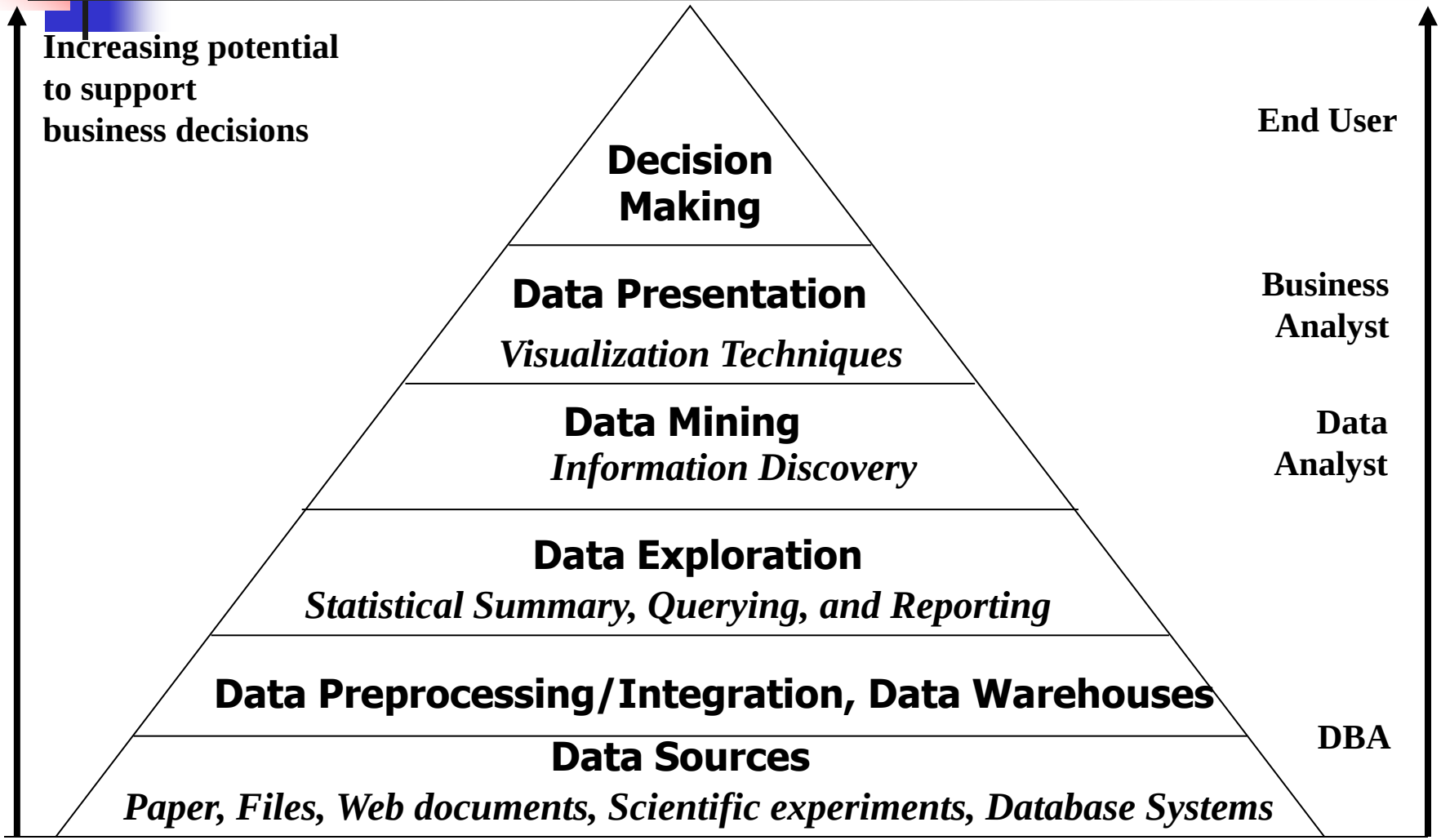
- La FD est le processus de recherche de nouvelles connaissances potentiellement utiles à partir des données (Fayyad *et al.*, *AI Magazine* 1996)
- La FD, ou encore l'ECD, est l'extraction à partir des données, d'informations implicites ignorées jusque là et potentiellement utiles. Cela inclut différentes approches techniques comme le *clustering*, la réduction des données, l'apprentissage de règles de classification, la découverte de réseaux de dépendance, les analyses de variations et la détection d'anomalies. (William J. Frawley, Gregory Piatetsky-Shapiro et Christopher J. Matheus)

Notion de FD

- **(cf. J** Han *et al*)

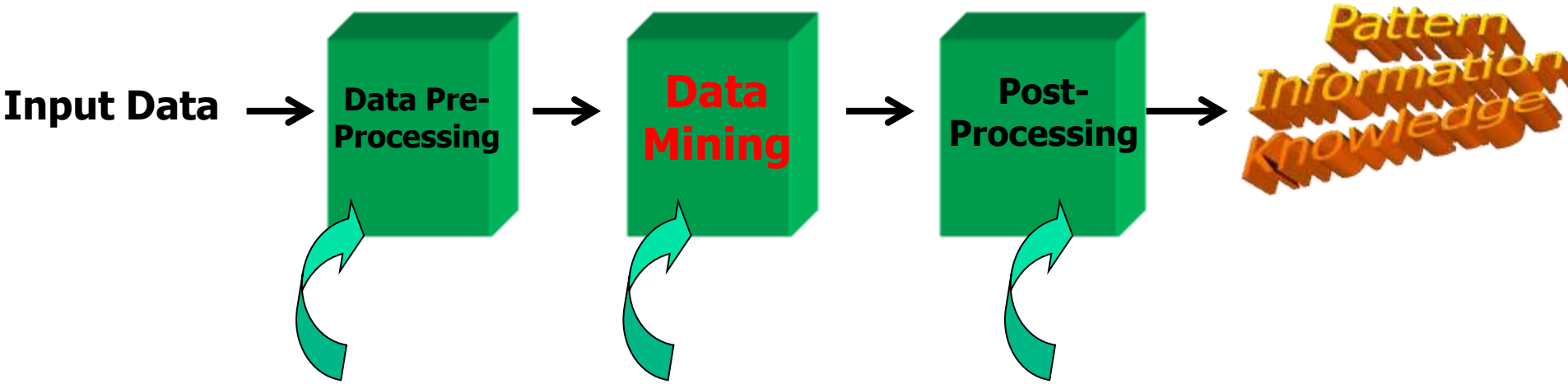


Notion de FD


Increasing potential
to support
business decisions
Decision
Making
End User
Data Presentation
Visualization Techniques
Business
Analyst
Data Mining
Information Discovery
Data
Analyst
Data Exploration
Statistical Summary, Querying, and Reporting
Data Preprocessing/Integration, Data Warehouses
DBA
Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems

Notion de FD

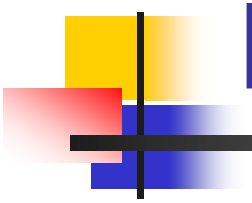
Vue de l'apprentissage automatique et de la statistique



Data integration
Normalization
Feature selection
Dimension reduction

Pattern discovery
Association & correlation
Classification
Clustering
Outlier analysis
… … … …

Pattern evaluation
Pattern selection
Pattern interpretation
Pattern visualization



Histoire

- Première conférence sur la Fouille de données en 1995
- Première revue « Data Mining and Knowledge Discovery Journal » en 1997

Pourquoi la fouille de données

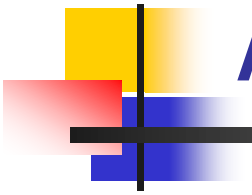
- Explosion de données : de teraoctets à petaoctets, maintenant on parle souvent de zettaoctets
  - Disponibilité de grands ensembles de données
    - seule une partie est analysée
  - Sources abondantes de données dans de grandes dimensions
    - Business: Web, e-commerce, transactions, bourse, …
    - Science: Télédétection, bioinformatique, simulation scientifique, …
    - Société/individu : journaux, caméras numériques, YouTube
- On est noyé dans les données, mais manque de connaissances !
- FD – Analyse automatique de grandes masses de données

Améliorations techniques

- Stockage des données
- Bases de données relationnelles
- Bases de données distribuées
- Entrepôts de données
- Bases de données multimédia, Web, etc.
- Taux de transfert dans les réseaux
- Algorithmes
- Calcul haute performance
- etc.

FD – multidisciplinaires

- Domaines concernés
  - Statistique : Méthodes, théorie de l'apprentissage
  - Informatique : Bases de données, Algorithmique, théorie de l'apprentissage, interfaces de visualisation
  - Linguistique
  - ...

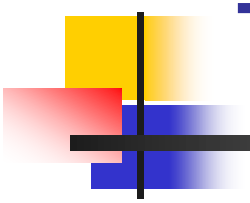


Applications

- Analyse de données et aide à la décision
  - Gestion et analyse de marché
    - Segmentation de marché, gestion de la relation client (CRM - *customer relationship management*), Analyse de ventes croisées, etc.
  - Gestion et analyse du risque
  - Détection de fraude et de "hors-norme" (*outliers*)

Applications

- D'autres applications
  - Fouille de textes (nouvelles d'agences de presse, email, documents) et de Web (*Text/Web Mining*)
  - Fouille d'images : RI, détection de copies illégales, etc.
  - Fouille de réseaux d'information
  - Fouille de flux de données (*Stream data mining*)
  - Bio-informatique
  - Analyse de données scientifiques



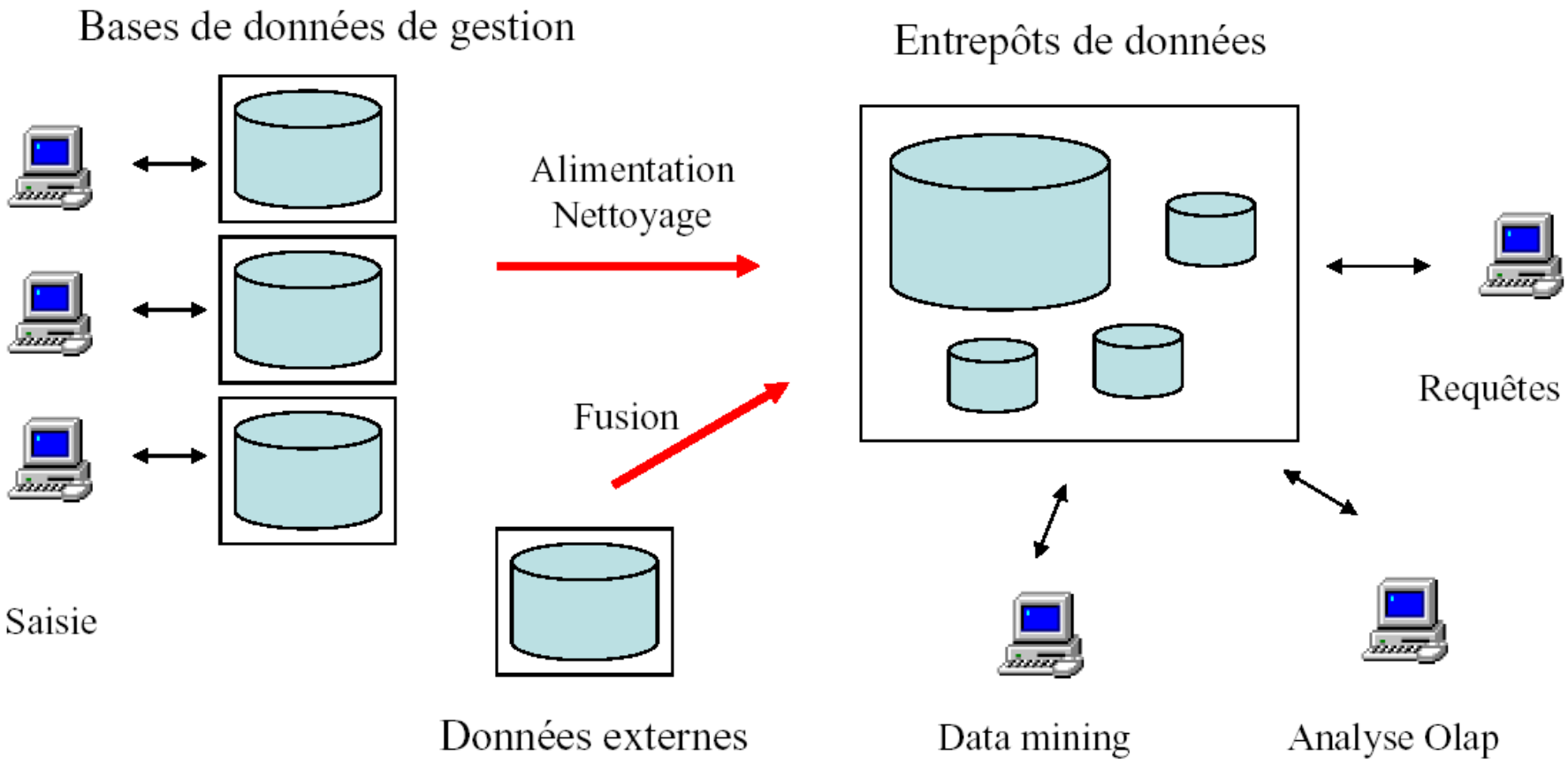
Types de données

- Numérique
- Qualitative
- Textuelle
- Séries chronologiques (médecine, sciences économiques, etc.)
- Graphes
- Images (fixes et vidéo)
- Données hétérogènes (web, dossiers médicaux, etc.)

FD : une étape dans le processus d'ECD

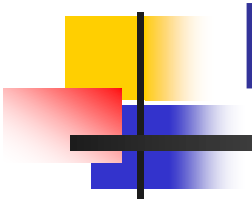
- Processus ECD
  - Préparation : sélection, nettoyage, intégration, codage
  - Fouille : exploration, découverte de structure, modélisation, validation
  - Récupération des connaissances et de l'information

Système d'information



Etapes de la FD

- Exploration : visualisation, analyse de données exploratoire et analyse de données multidimensionnelles
- Découverte de « formes » (clustering, règles d'association)
- Classification et prédiction (modélisation)
- Validation et choix de modèle



Exploration

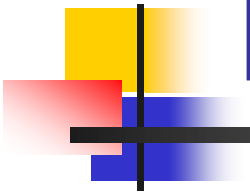
- Statistiques élémentaires, tri-à-plat
- Histogrammes, diagrammes de dispersion
- OLAP (*online analytical processing*) : une catégorie d'applications et de technologies permettantde collecter, stocker, traiter et restituer des données multidimensionnelles, à des fins d'analyse : hypercubes pour des données agrégées
- Visualisation
- Analyse de données multidimensionnelles : analyse en composantes principales, analyse des correspondances, projections révélatrices

Exploration : Visualisation

- La visualisation de l'information et les représentations graphiques
  - permettent de manipuler des données de nature diverse, non-homogènes, bruitées
  - n'exigent pas de prérequis sophistiqués en maths ou en algorithmes statistiques
  - fournissent une vue d'ensemble des données

Découverte de « formes »

- Chercher des preuves de la présence de structures locales appelées « formes » (recherche des règles d'association, analyse de séquences génomiques, détection de fraude etc.)
  - Détection d'items fréquents
  - Méthodes de classification non supervisée (*clustering*)



Modélisation

- Trouver des structures globales (modèles) dans les données (apprentissage statistique ou non)
  - Analyse discriminante
  - Méthodes de régression
  - Graphes inductifs
  - Réseaux neuronaux
  - SVM (*Support Vector Machine*) : Séparateurs à vaste marge
  - etc.
- Prédiction pour des nouvelles données

Validation et choix des modèles

- Il existe des indicateurs pour sélectionner les modèles
- On peut comparer les modèles
  - Taux d'erreur en généralisation (précision, rappel, F-mesure)
  - Courbes de lift ou courbes ROC (*Receiver Operating Characteristic Curves* - courbes de caractéristiques d'efficacité)

KDnuggets Polls

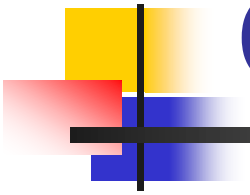
https://www.kdnuggets.com/polls/index.html

Quelques termes

- *Clustering* : classification
- *Classification* : classement, analyse discriminante
- *Decision trees* : arbres de décision, segmentation

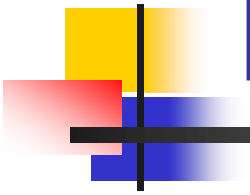
Recherche d'information

- Plus liée à l'analyse de données textuelles mais aussi à l'analyse de base d'images ou de vidéos
- Classiquement, tout système de RI se décompose en 3 parties :
  - Processus d'indexation
  - Processus d'interrogation
  - Modélisation de la connaissance



Quelques exemples

- cf. M. Tenenhaus



Résumé

- Notions de FD, ECD, RI
- Tâches de la FD
- Quelques exemples