Introduction à l'Extraction de Connaissances

(From Data to Knowledge)

Alexandre Saidi
ECL-3A
Master Informatique
ECL - LIRIS - CNRS

Introduction

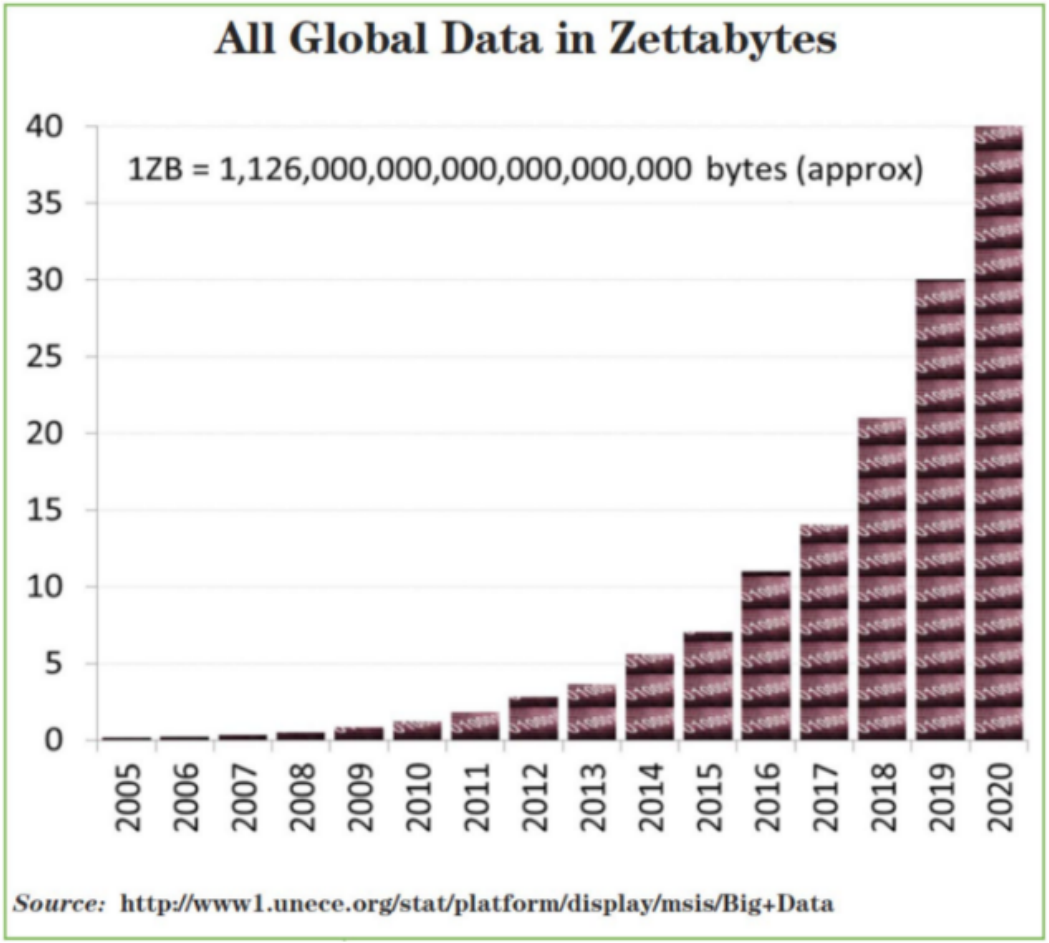
- L'*Extraction de Connaissances* (KD, ML) :
  - Croissance d'utilisation de (grosses) BDs par les ordinateurs
  - Besoin accru d'analyse de données économiques, sociales, ...
- Évolution des techniques et des besoins :

Etape d'Evolution	**Questions**	**Technologies**
Collection de Data (1960s)	"Montant de mes revenus ... dans les 5 dernières années ?"	Ordinateurs, bandes, disques
Data Access (1980s)	"Quels sont les résultats des agences ... dans le Rhône", "Et en Mars dernier ?"	Ordinateurs + rapides, moins chers avec + de stockage, DBs relationelles
Data Warehousing & Decision Support (Fin 80s) : Algorithmes d'Analyse de D.	"Quels étaient les résultats des agences dans le Rhône en Mars dernier ?" → (Résultats des requêtes/analyses stockés) "idem pour l'Alsace." → (analyser, comparer Rhône /Alsace)	Ordinateurs + rapides, moins chers avec + de stockage, DBs multidim., Data warehouses (en ligne) **OLAP** (*On-Line Analytical Processing*) Datacubes, Hiérarchies sur les attribut
Data Mining (Fouille de données)	"Quelles **prévisions** pour l'agence ... d'Alsace ? " "Le mois prochain ? Et **pourquoi** ?"	Ordinateurs + rapides, moins chers avec + de stockage, Algorithme avancés

- Les techniques statistiques ont été utilisées depuis longtemps,
  Les techniques algorithmiques d'analyse sont plus récentes.

Introduction (suite)

- Rythme de progression du volume des données (1ZB = 10^{21} Octets)



Data Warehouses et M.L.

• Dans le passée, les données servaient principalement à établir des rapports et aux visualisations (relativement) simples.

• Les analyses (statistiques) tentaient de vérifier des hypothèses sur des données (souvent de faible volume) !

• Les Data Warehouses (depuis env. 25 ans) contiennent des données

- simples (montant d'un ticket de caisse),
- structurées (données d'un étudiant, d'un client),
- semi-structurées (Web, avis sur un produit, log des clics), etc

Après "préparations", elles font l'objet du processus *Machine Learning*

• L'essor récent inclue les techniques algorithmiques : recherche "blanche".

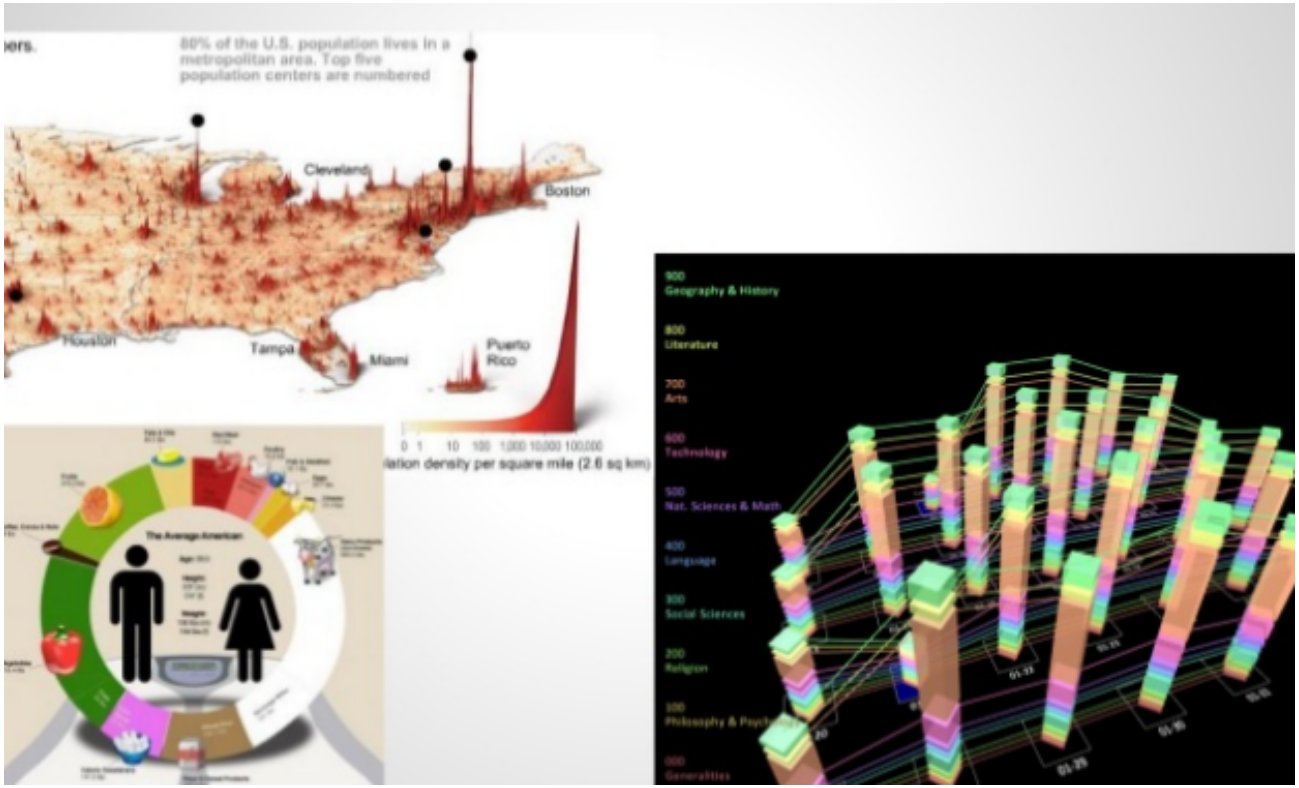
• **Facilité d'accès et Open-data :**
Récemment, Amazone *Redshift*, Google *BigQuery* ou Microsoft *Azure* ont transféré les data warehouses vers le cloud (accès plus ouvert, plus simple).

Data Warehouses et M.L. (suite)

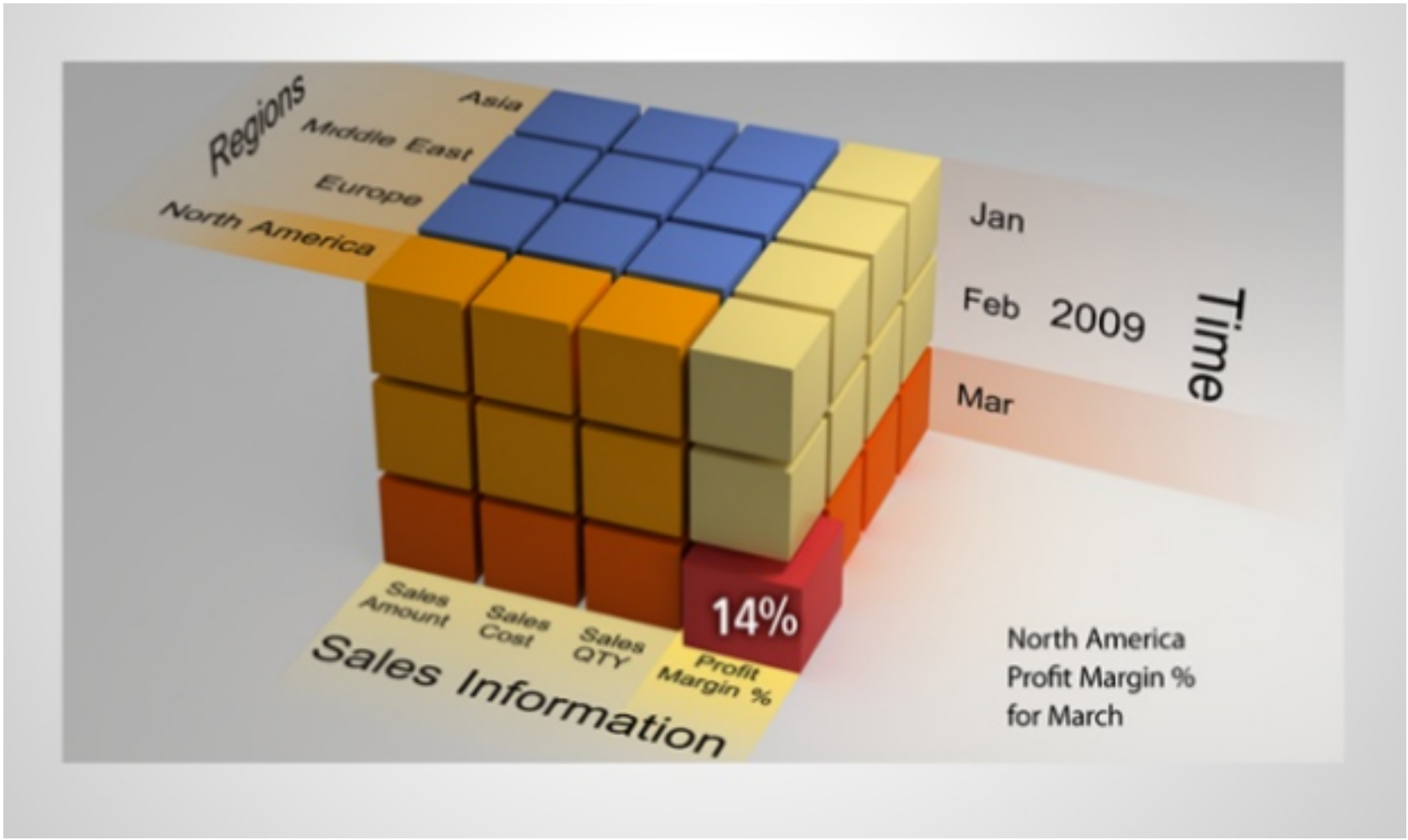
- Les questions auxquelles M.L. peut répondre (vs. une BDR ?) :
  - Evaluation des risques dans les investissements
  - Prédiction des résultats des campagnes de pub.
  - Détection de fraude (tél, banque, impôts, aides sociales, sécu, ...)
  - Prédiction des changement de profils / comportement des clients
  - Prédiction des préférences des clients (attrait pour tel ou tel produit)
  - Prédiction des ventes / revenus / de charge / ...
  - Prédiction des résultats d'étudiants (évolution des promotions d'élèves)
  - Prédiction de la criminalité, probation, ...
  - Prédiction des maladies qu'un individu pourrait développer (génétique)
  - Traitement / Résumé de données massives, ...

Visualisation

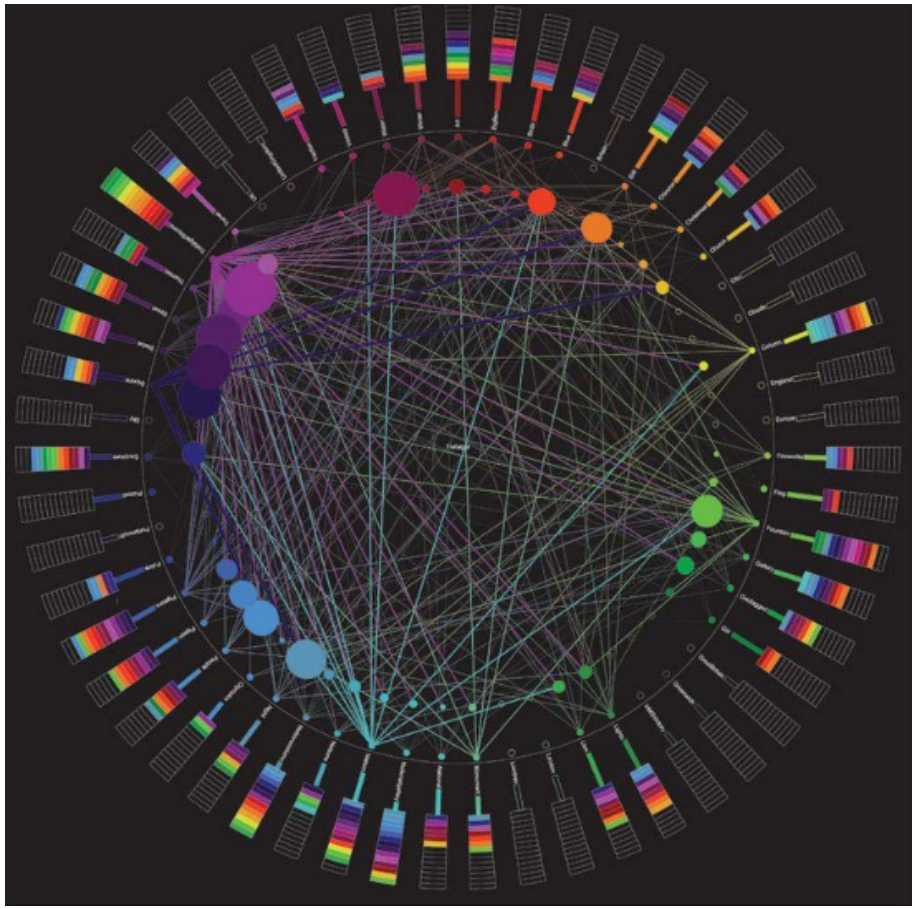
☞ Ne pas négliger la **visualisation**



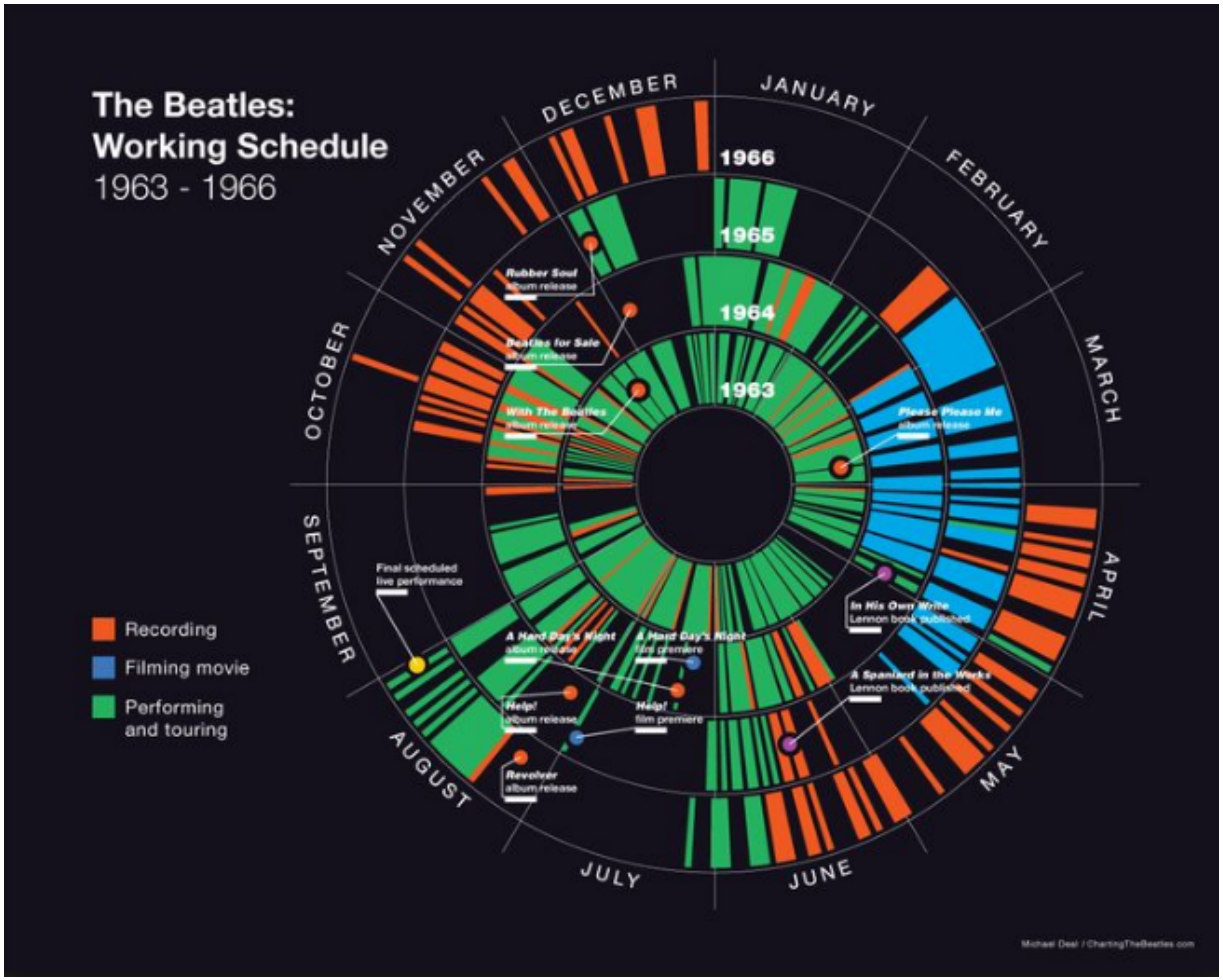
Visualisation (suite)



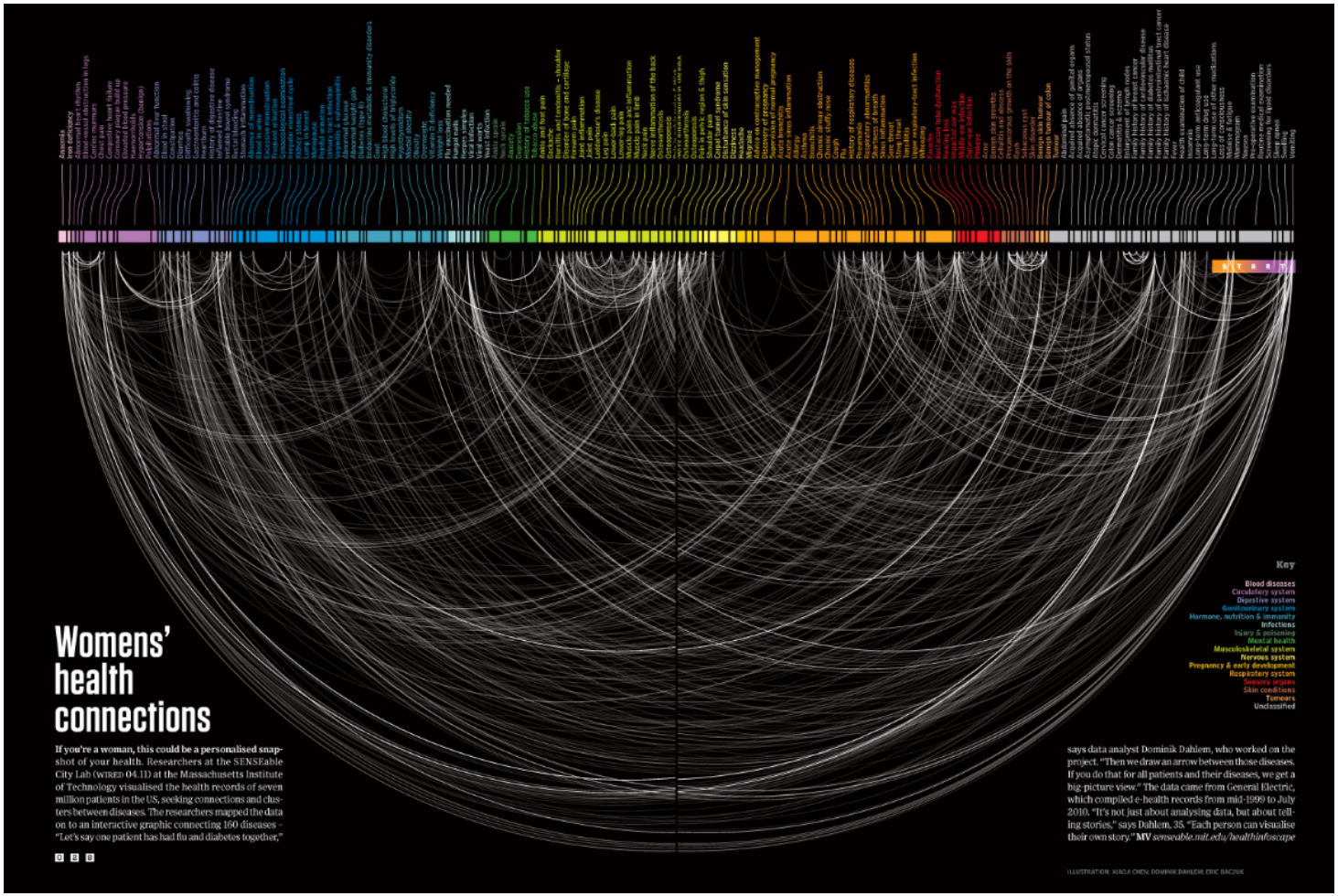
Visualisation (suite)



Visualisation (suite)



Visualisation (suite)

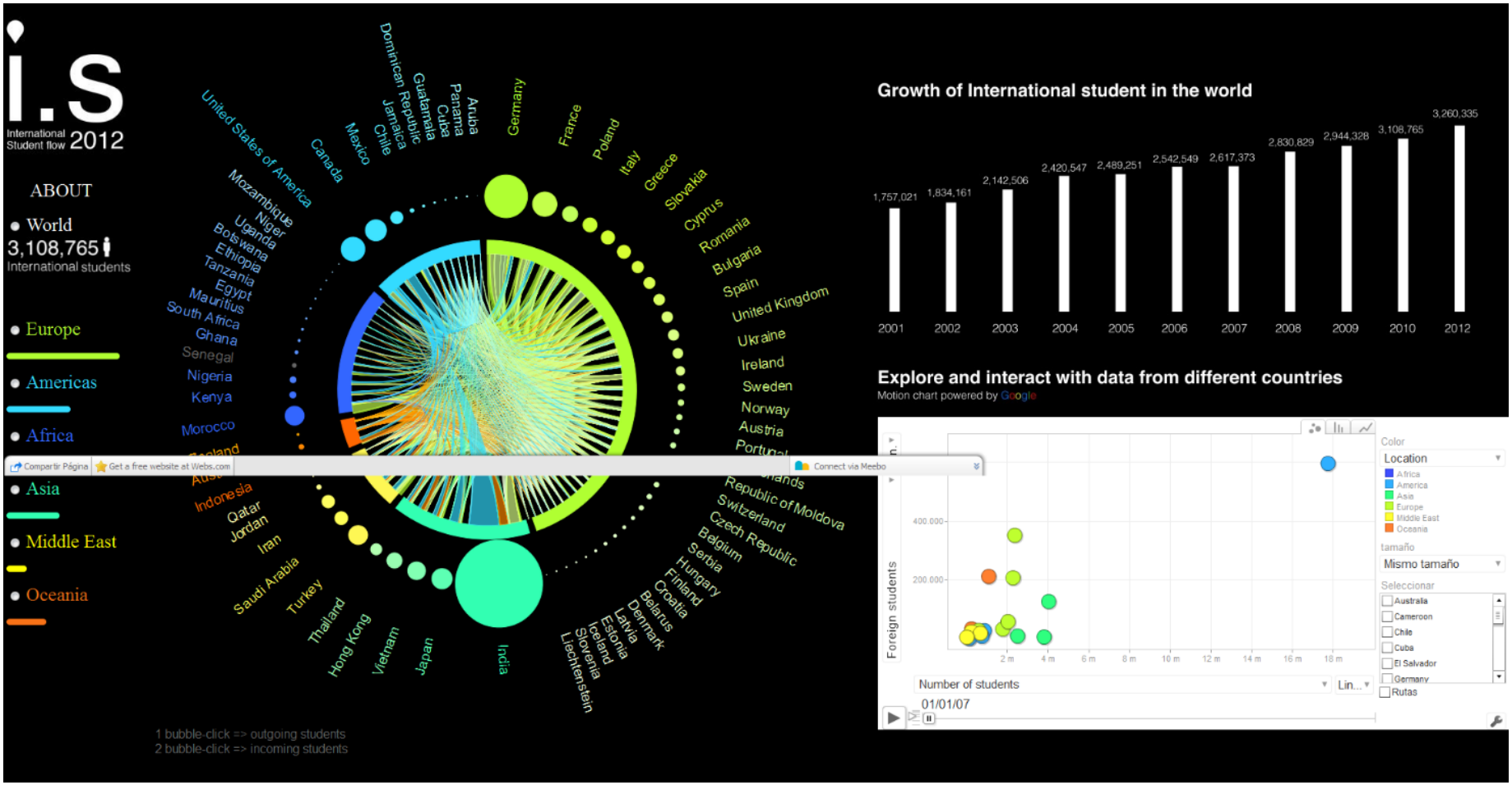


Visualisation (suite)

- Texte :

Visualisation (suite)

- Mouvements mondiaux des étudiants (2012)



EC = Apprentissage automatique

Constat :

- Quantité et disponibilité d'informations (Data) dans tous domaines
  - En général, les données brutes (+ bruits !) sont accessibles à tous
  - Des données de qualité sont plus rares
  - L'exploitation des données (cf. motifs pertinents) est un atout

    ➡ Données → Connaissances → Décisions

- Domaines d'application concernés : Tous !
  économie, éducation, finances, commerces, Industrie, santé, jeux, ...
- **L'apprentissage** = méthode pratique de découverte de **concepts**.
- L'humain (dès l'enfance) utilise des instances de concepts pour se représenter des **animaux, plantes, homme, femme, jouet, ...**

EC = Apprentissage automatique (suite)

➥ On apprend des instances particulières → on choisit des attributs
➥ On forme des *modèles de classification*
➥ On utilise ces modèles pour identifier des objets similaires.

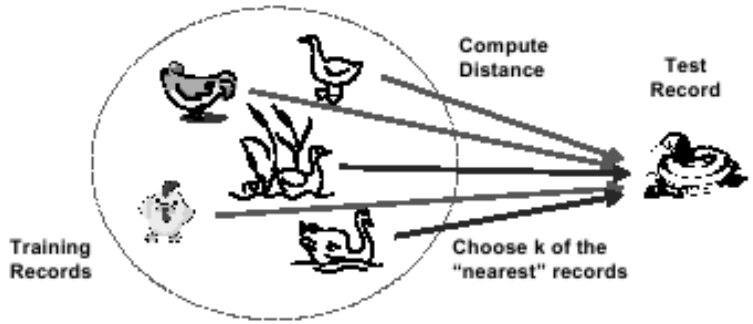
- Le cerveau apprend en permanence.

  Exemple : deviner le nombre suivant :

  1,2,3,4,5... 1,2,3,5,8,13,...

  1,2,3,5,7,11,13,... 5 ?5 ?5=6

➥ Les données (data) contiennent potentiellement des connaissances



Fouille de données : quelques exemples indicatifs

Quel rapport ?

Pourquoi beaucoup de tournois de Golf télévisés sont sponsorisés par des courtiers (brokers) en ligne ?

➥ Découvert dans les fichiers :

⇒ *plus de 70% des investisseurs en ligne sont des hommes d'âge environ 40 ans et qui jouent au Golf.*

⇒ *60% de tous les investisseurs en stock option sont des Golfeurs.*

Fouille de données : quelques exemples indicatifs (suite)

Let's shake it !

Est il utile, pour une compagnie de musique, de faire de la pub pour la musique ***Rap*** *dans des magazines pour les* ***seniors*** *?*

➔ Oui : les seniors offrent souvent de la musique Rap à leur petits enfants (adoles).

The big brother !

Comment les banques (service CB) peuvent-elles suspecter une carte volée, même si le propriétaire n'est pas conscient du vol ?

➔ Beaucoup de ces compagnies (de crédits) stockent un modèle général de nos *habitudes* d'achat avec la carte.

➔ Ce modèle alerte la compagnie d'un vol possible lors d'une transaction qui ne rentre pas dans le profil général de nos *habitudes*.

Fouille de données : quelques exemples indicatifs (suite)

Satisfait ou remboursé !

Comment et de quelle durée étendre la période "satisfait ou remboursé" ?

➔ Données complètes concernant les pannes et leurs couts + enquêtes.

La fertilisation *in vitro* de vaches :

Comment choisir les "meilleurs" oeufs avec un maximum de chance de survie.

- Collecter des oeufs + fertilisation + transfert vers l'utérus de la "porteuse"
- Env. 60 caractéristiques (**attributs ou variables**) ➔ complexité
  → la morphologie, oocyte, follicule (taille oeuf), qualité sperme, etc.
  → difficile de trouver une corrélation pour *prédire* la survie de l'embryon.
- Utilisation des techniques **Apprentissage automatique** pour aider à la sélection (en Angleterre).

Fouille de données : quelques exemples indicatifs (suite)

Etc... → Vers le Big data

Caractéristiques du Big Data : théorème de HACE

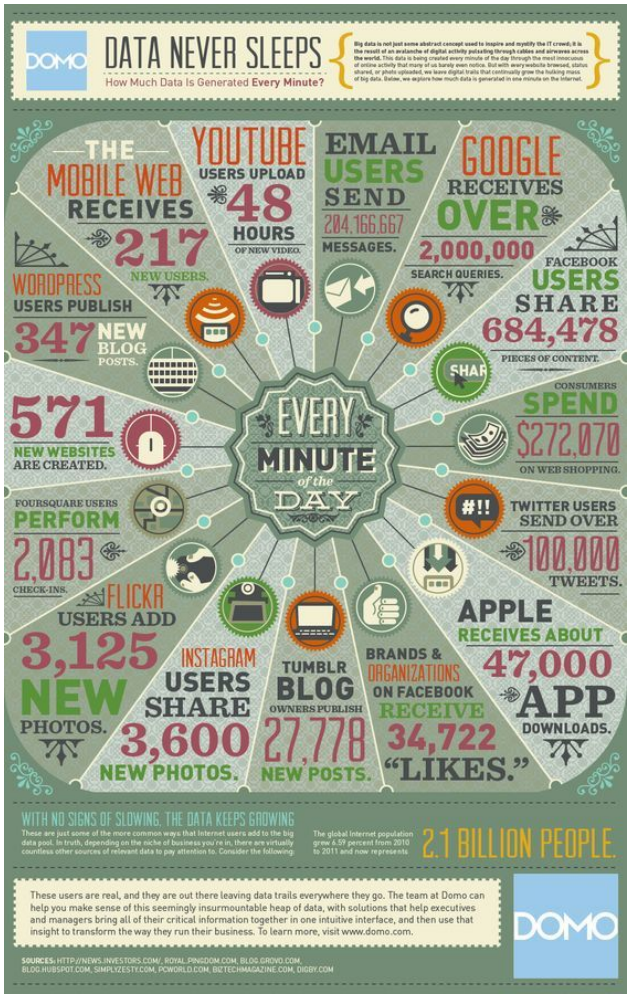
Big Data est caractérisé par un volume large et **h**étérogène, de sources **a**utonomes avec un contrôle distribué et décentralisé cherchant à explorer des relations **c**omplexes et **e**volutives entre ces données.

Vers le Big Data

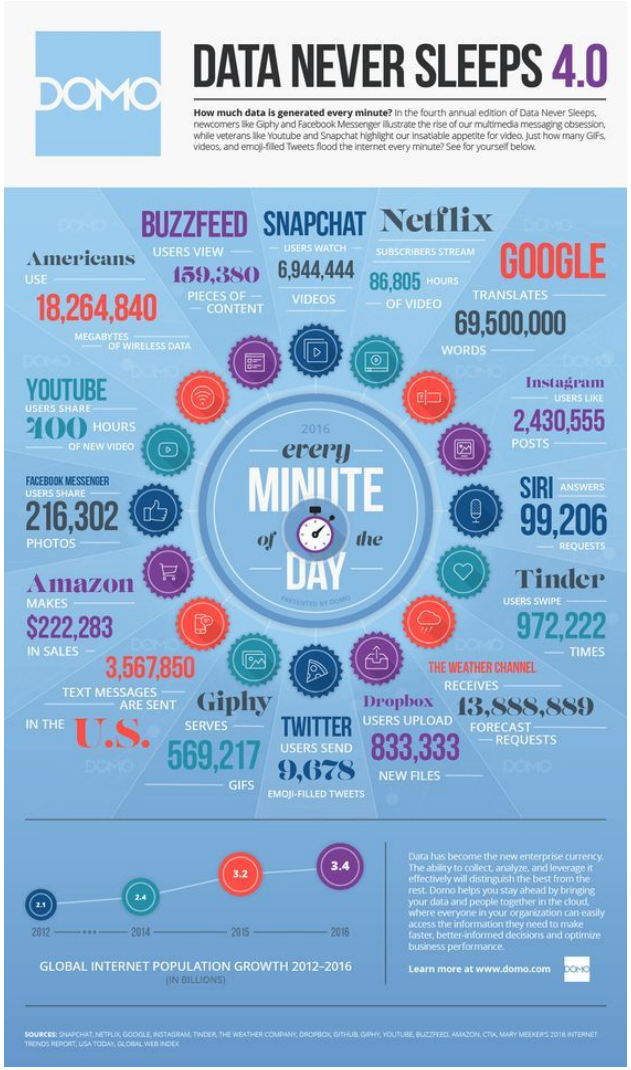
Le processus Big Data :

- Les domaines récents et significatifs (en plus des domaines "classiques") :
  - ➛ **Biologie**,
  - ➛**Nature** et espace,
  - ➛ **Réseaux sociaux** (par la taille des données), IoT, ...
  - ➛ **Automatisation** (industrie, conduite automatique en ville)

Vers le Big Data (suite)



Vers le Big Data (suite)



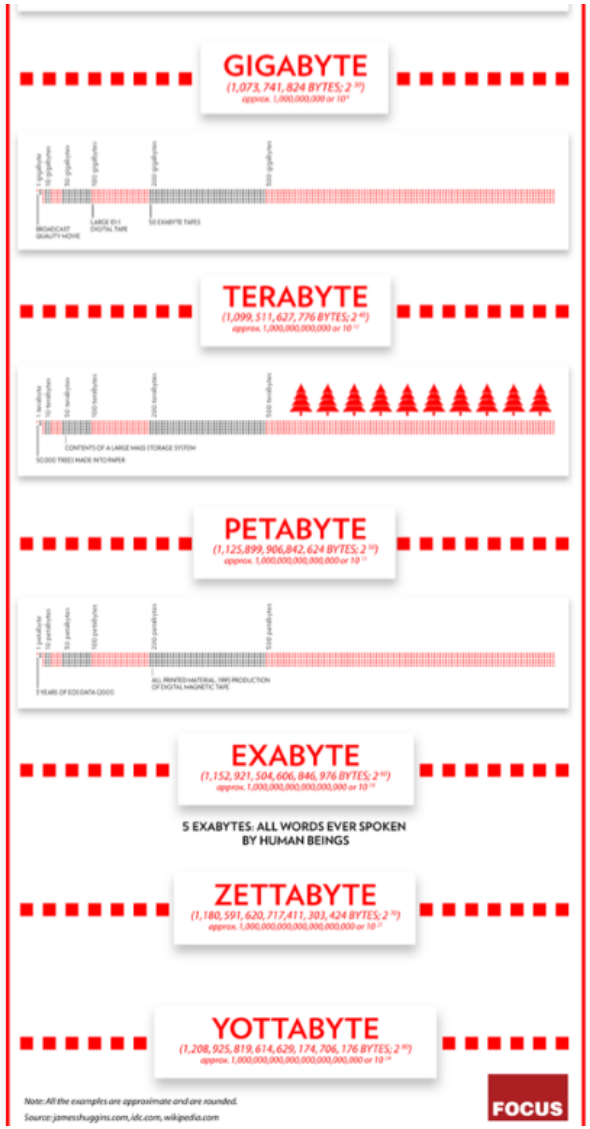
Vers le Big Data (suite)

Une idée de la taille des données

- *GigaByte* : 10^9 octets = 1000 *Mo*
- *TeraByte* : 10^{12} octets = 1000 *GO*
- *PetaByte* : 10^{15} octets = 1000 *terabytes*
- *ExaBytes* : 10^{18} octets
- *ZettaBytes* : 10^{21} ou 2^{70} octets
- *YottaByte* ou *Yobibyte* : 10^{24} octets

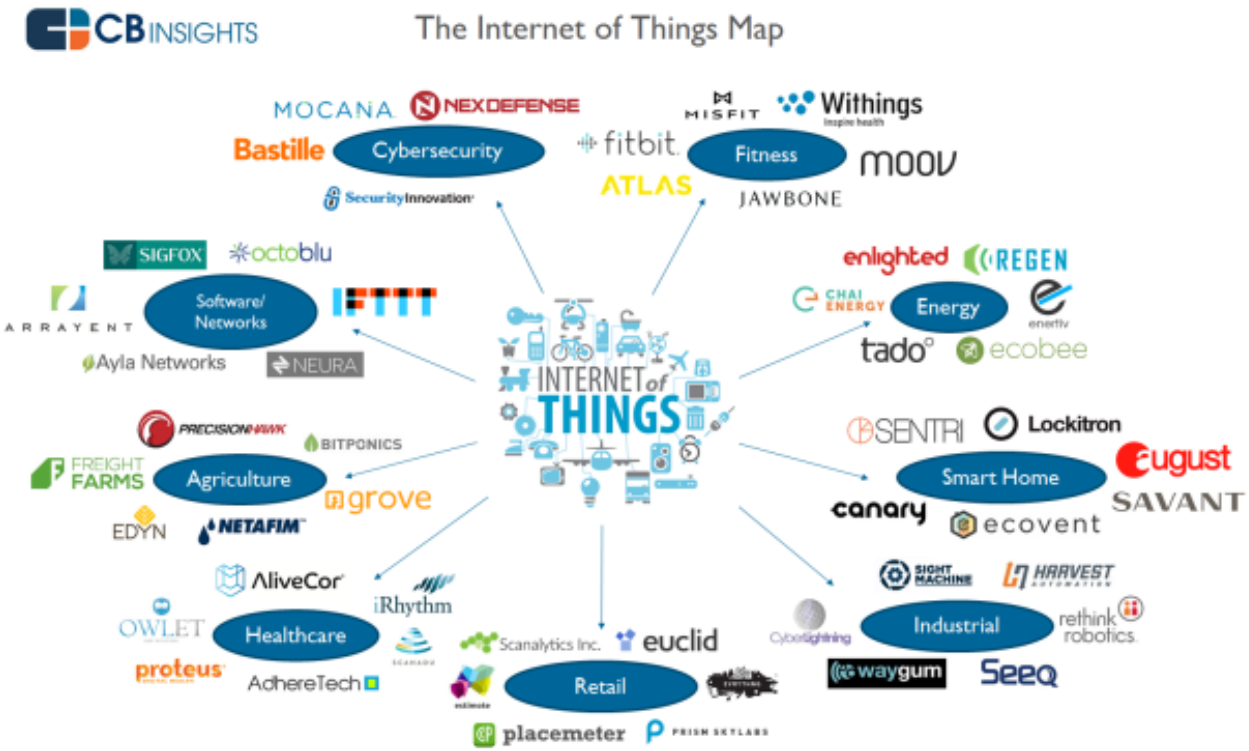
• L'humanité n'a prononcé que ➙
5 *ExaBytes* de mots (depuis l'origine !)

• Le volume de données mondiales a été évalué à 2 *ZettaBytes* en 2016 ...



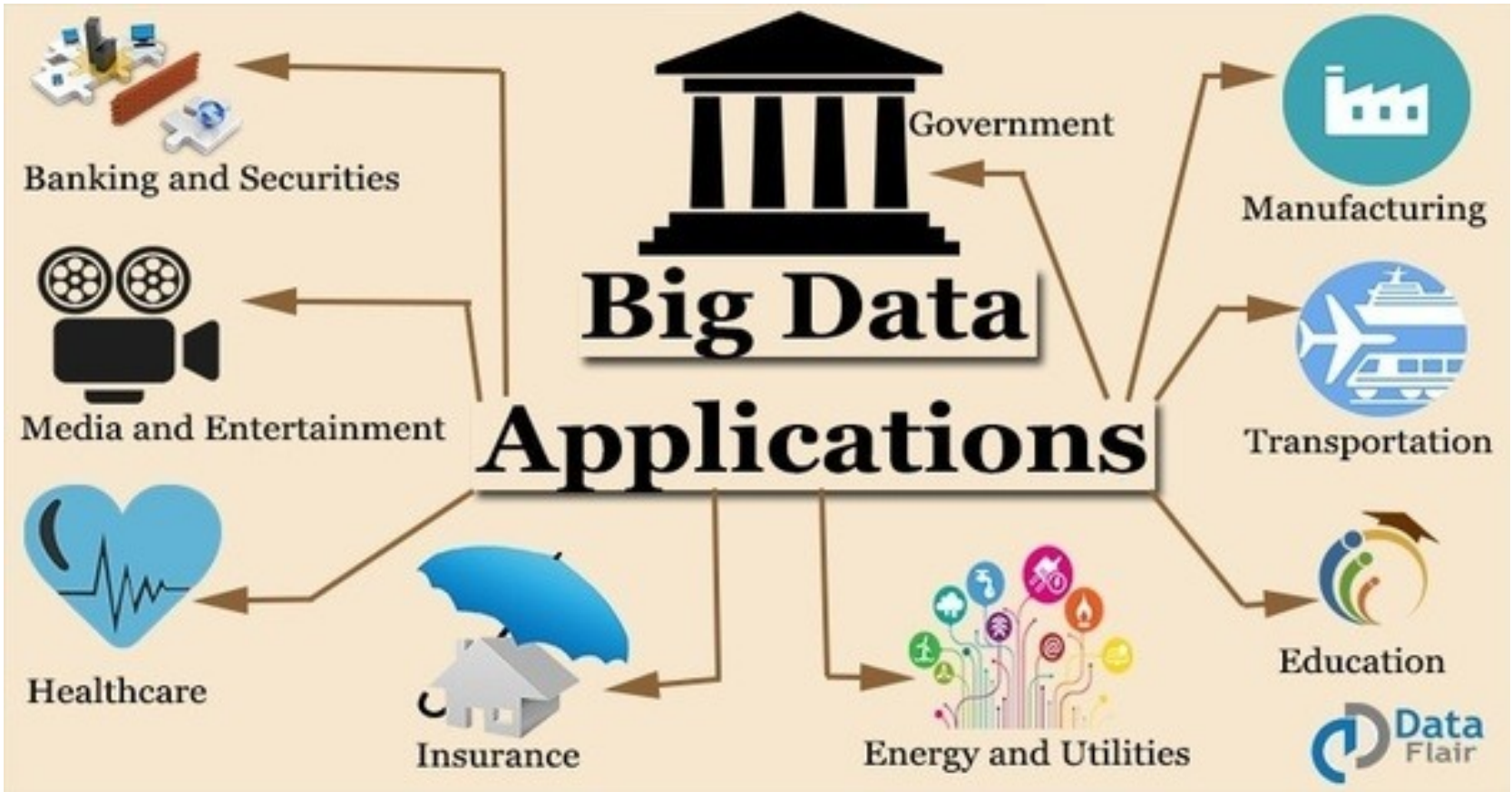
Vers le Big Data (suite)

IoT (*Thanks to https ://www.cbinsights.com*)



Vers le Big Data (suite)

(*Thanks to https ://www.quora.com*)



Vers le Big Data (suite)

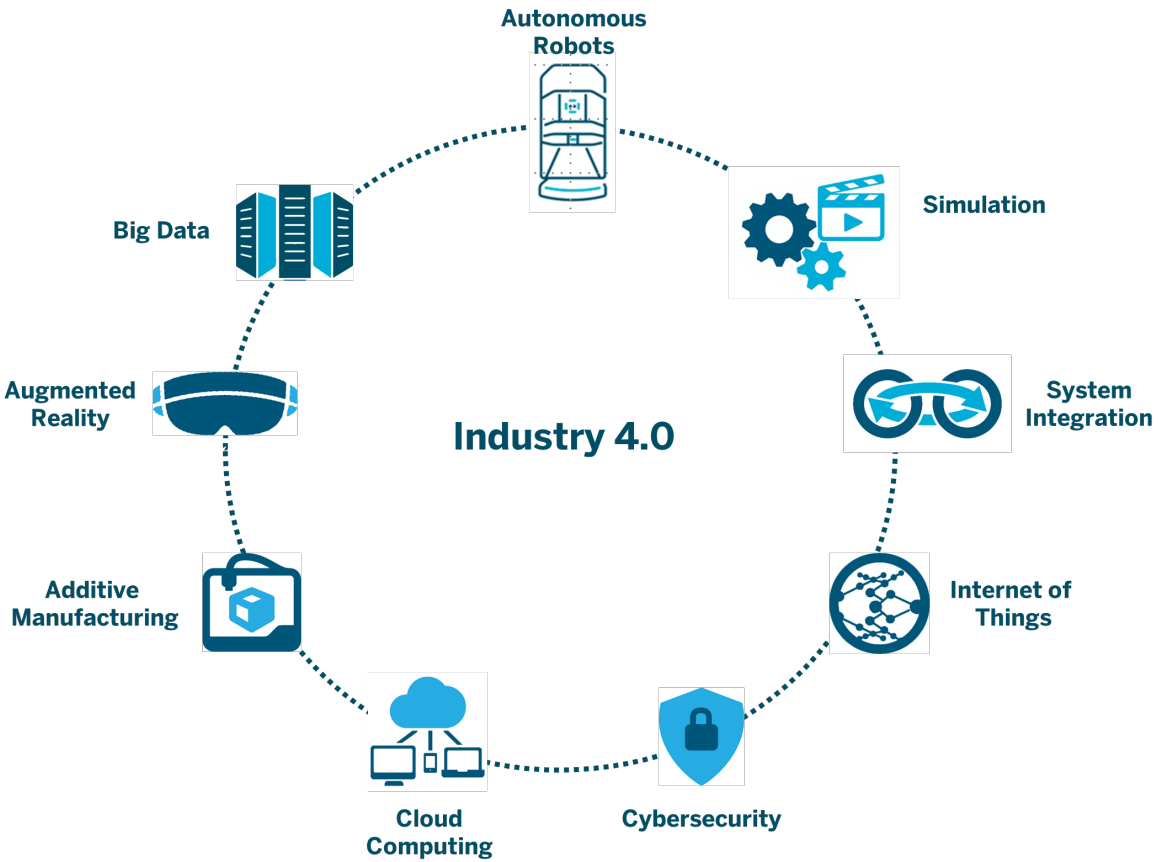
La nature (et l'espace) sont des sources de données du big-big-data

☞ Savez vous que chacun de nous est traversé par des milliards de *neutrinos* à la seconde !
"neutrino" (ou petit "neutron") émis par les galaxies et l'infinité d'étoiles (soleil compris)
➛ On ne sait pas quelles actions ces neutrinos ont sur les êtres vivants mais on sait qu'en nous traversant (à la vitesse de la lumière), une partie d'entre eux disparaît !

➛ **Vous voulez faire du Big-Data ? !**

Vers le Big Data (suite)

L'industrie nouvelle 4.0 (il manque les "finances")



Google page Ranking

- Google détient 67% du marché de recherche d'information sur internet (18% pour microsoft *Bing*, *Yahoo* 11%, le reste pour *Ask*, *AOL*, etc.).

- L'algorithme de Google travaille avec des *spiders* (ou *crawler* : araignée / robot d'indexation) et maintient un très grand indexe de mots clés + url.

- Critère principal : le nombre et la qualité des liens vers les pages
  → l'importance du site
- Le raisonnement : plus un site est référencé, plus il est susceptible d'en recevoir !

- Autres critères : le choix de la page par les internautes dans les résultats, la fréquence et la position des mots clés dans 1 page, la durée de vie de la page, etc.

- La pondération de chacun de ces critères connus (et inconnus ?) fait la force de Google (biaisé néanmoins par des considérations commerciales).

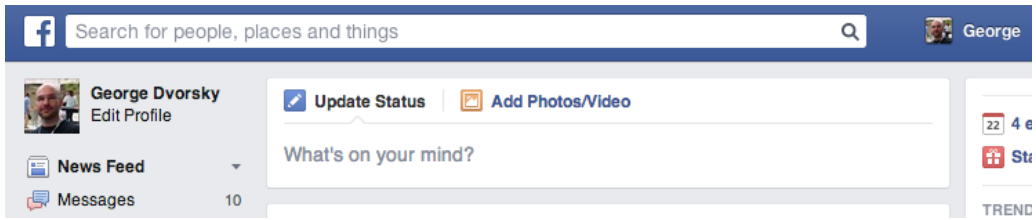
Google page Ranking (suite)

Les sources inspiratrices de Google (*Sergueï Brin, Larry Page*) :

- Au départ, une idée sur les indices de citations (**sociologie** : G. Pinsky et F. Narin - 1976) a inspiré Google.
  ➙ Ont appliqué cette idée aux pages web (les liens In / Out des pages)
- Cette idée est formalisée en **Mathématiques / Algèbre** par les (très grandes) matrices + calculs de valeurs propres exploitée par le théorème de *Perron-Frobenius* (1900, application en chaînes de Markov et en théorie de Graphes).
  ➙ Calcul de *pagerank* (Google)
- Un système **Informatique** distribué efficace et adapté implante et prend cet ensemble en charge.

Facebook's News Feed

• Pré-sélectionne (pour nous) les articles choisis par (l'algorithme de) Facebook (sauf si vos préférences demandent de tout afficher dans l'ordre chronologique).



• Pour sélectionner ces articles, l'algorithme de Facebook s'appuie (entre autres)
  - sur le nombre de commentaires,
  - qui a posté l'article suivant une classification des personnes plus ou moins populaires (ou ceux avec qui vous avez eu des contacts ou ceux dont vous avez déjà lu des choses),
  - le type de l'article (photo, vidéo, état, MAJ, etc.), ...

• Pour placer des pubs contextuels, Google / Facebook / ... utilisent des algorithmes pour épier nos habitudes, usages, achats, questions posées et les mots choisis.

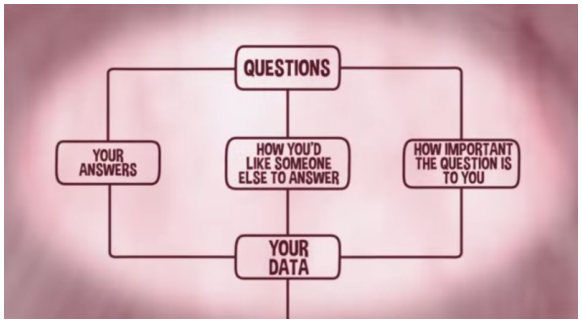
OKCupid Date Matching

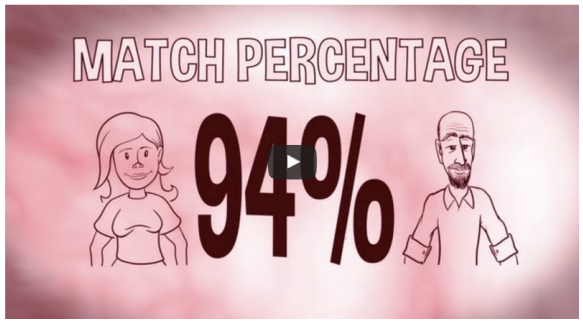
• Un marché de 2 milliards avec une progression constante de 3% depuis 2008 !

• Les gens n'ont plus le temps de trouver l'âme soeur comme avant et 1/3 des mariages serait du à ces sites.

• On prétend que ces couples sont plus solides que les 'classiques'.

• La méthode utilisée par *OKCupid* est basé sur le *matching* des préférences, centres d'intérêts, tendances, gouts, ...





• Mais au lieu de faire du matching brute des intérêts communs, OKCupid analyse les réponses de chacun, attache à ces réponses une pondération suivant son importance pour la personne ainsi que pour un/une âme soeur.

NSA Data Collection, Interpretation, and Encryption

• Vous croyez être surveillé par des caméras ? Non, vous êtes épiés par des algos.

• *Edward Snowden* a dévoilé que la NSA surveille les données de milliards de personnes.

• Ses révélations ont montré qu'il existe des algorithmes d'analyse des données de surveillance récupérées par 5 pays : USA, CA, GB, Aust., Nlle-Z.

• Données provenant des écoutes tél-ques, mails, webcams et vidéo, Géo loc, ...

• La NSA est drôle : elle dit ne pas collecter d'info puisque par (leur) définition :

> ☞ *une information n'est collectée que si elle est effectivement analysée et transformée sous une forme intelligible par un membre de la NSA pendant ces missions officielles.*
>
> ➙ *C-à-d. quelqu'un qui possède chez lui une bibliothèque pleine de livres ne collecte pas de livres. Les seuls livres qu'il aurait collectés sont ceux qu'il a déjà lus !*

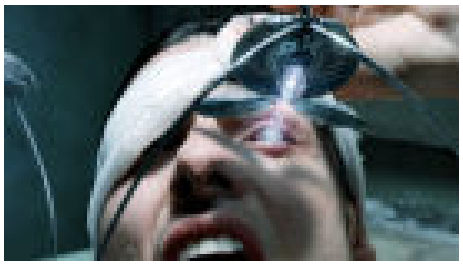
➙ Le problème : collecter des infos nous concernant et à notre insu, quand bien même toutes ces données ne sont pas encore exploitées.

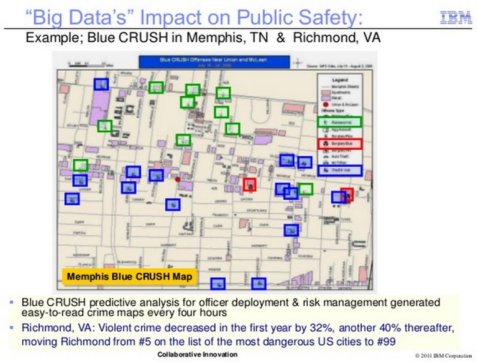
IBM'CRUSH

• Pas encore trop rependu mais de plus en plus de départements de police aux USA utilisent un outil d'*analyse prédictive* (cf. le filme *Minority Report*).

• Le CRUSH d'IBM (*Criminal Reduction Utilizing Statistical History*) aurait permis à la police de Memphis de réduire les crimes de 30% depuis 2006 (chiffres publiés).

➛(D'autres états US et plusieurs autres pays sont fortement intéressés.)





• CRUSH exploite les techniques statistiques et des algorithmes de déduction et de prédiction

(voir *http ://www.ibm.com/smarterplanet/us/en/leadership/memphispd/*). ../..

IBM'CRUSH (suite)

• En matière de la lutte prédictive contre le crime (Crime Predictive Fighting), CRUSH permet d'analyser les *incidents* pour prédire les *points chauds* où un crime est susceptible d'avoir lieu

➙ Pour déployer les ressources et les hommes adéquats à ces endroits.

• Ces d'algorithmes seront (bientôt) déployés pour surveiller les surfs sur Internet, les données GPS, celles des PDAs et smartphones (écoute téléphonique), la signature biologique (e.g. de la rétine) pour une analyse en temps réel.

IBM'CRUSH (suite)

Reprendre le cours 1 de la formation et prendre les images et pages dans ce coin.

Recherche de motifs dans les données

- **Pas une nouvelle science/discipline** :

  ➙ l'homme a toujours recherché des motifs (patterns) :

  - les chasseurs : motifs (habitudes répétées) dans la migration des oiseaux ;
  - les fermiers : motifs dans l'élevage de bétail ;
  - les politiciens : pattern dans l'opinion des votants ;
  - les scientifiques : découvert de motifs qui expliquent les phénomènes ϕ

    $\rightarrow$ proposition de théories pour **prédire** une nouvelle situation.
  - les entrepreneurs : identification et exploitation des opportunités / comportement du marché (bourse)
  - même les amoureux : patterns dans les réponses des partenaires.

- Les économistes, prévisionnistes(e.g. météo), statisticiens ont déjà travaillé sur cette idée (e.g. *régression(s)*)

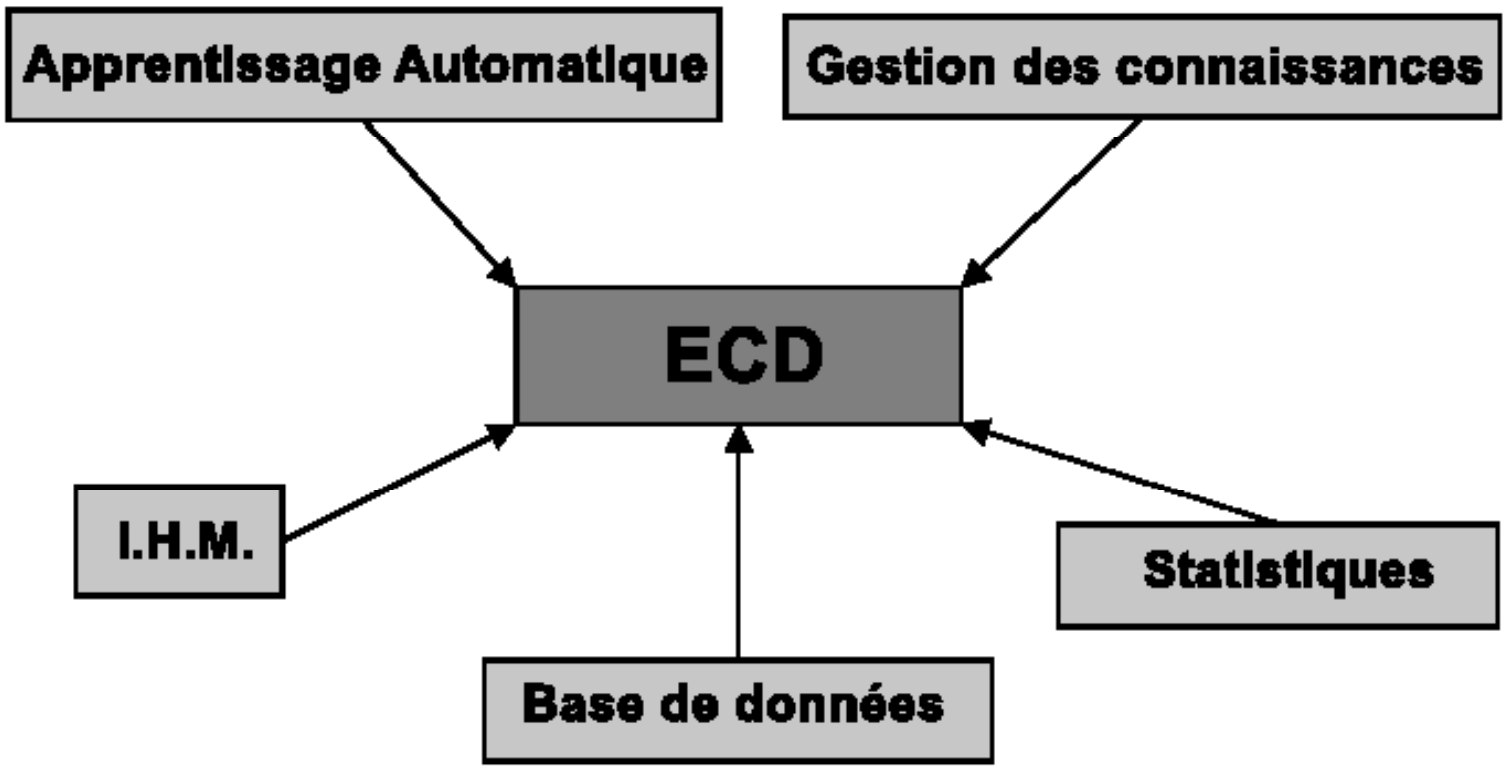
Recherche de motifs dans les données (suite)

➙ Ce qui est nouveau : **Quantité, Techniques et moyens**

☞ Le volume de données mondiales **double tous les 20 mois** ;

- Gros supports peu chers, la manie de tout sauvegarder plusieurs fois !
  ➥ Nos choix / décisions (achats dans les supermarchés), nos habitudes financières, nos aller-et-venus, nos navigations (WEB), Caméras, ...
- Opportunités Économiques, Médicales, Industrielles, Bio-xx ....
  ➥ *Modélisation* et *Généralisation* vs. vérification d'hypothèses
  ➥ Traitement par des méthodes algorithmiques
- Recherche de **Motifs (Patterns)** est un des objectifs de l'EC
  ➥ *algorithmes Génétiques*, *Réseaux de Neurones*, les méthodes algorithmiques pures (*ID3, C45, ...* → arbres / règles) , *inférence et réseaux Bayesiens*
- Nouveauté : **combinaison des méthodes** statistiques, algo et IA.

EC : Domaine multi disciplinaire



Apprentissage = Recherche de motifs

Recherche des motifs (pattern recognition)

• Processus de formation de définitions de concepts généraux en observant des exemples spécifiques des concepts à apprendre.

☞ 98% de l'apprentissage humain = Reconnaissance de Motifs.

(Selon *Ray Kurzwell*, le père de *Pattern/Voice Recognition*)

• La connaissance extraite des données dans une session d'apprentissage
= un **modèle** ou une **généralisation** des données.

• Quasi toutes les méthodes d'Extraction de Connaissances sont **inductives**
➙ **Induction-based Learning**
➙ On crée des modèles à partir d'**instances positifs** et **négatifs**.

• Par contre, l'application de ce qui est appris est **déductive**.

Démarche Déductive / Indictive

Exemple naïf :

- les pigeons : sont des oiseaux, ont des ailes et savent voler
- les aigles : sont des oiseaux ont des ailes et savent voler
- les cormorans : sont des oiseaux ont des ailes et savent voler
- .....

➙ **Induction** du concept "*oiseau*" : ont des ailes et savent voler

- Les merles sont des *oiseaux*.

➙ **Déduction** : les merles ont des ailes et savent voler

☞ Quid des cas négatifs comme les *autruches* (kiwi, dodo, weka, ...) !

Démarche Déductive / Indictive (suite)

• Le processus d'apprentissage (Algorithmique, Statistique) suit en général une approche **inductive** tout en utilisant des étapes préliminaires **déductives** :

1. Définir le problème.
2. Assembler des données sur le phénomène, les *explorer*
3. Émettre une ou plusieurs hypothèses (plutôt Stat.)
4. Assembler des (nouvelles) données expérimentales
5. **Induction** : Analyser les données
6. Tester les hypothèses sur ces (nouvelles) données et Interpréter les résultats
7. Synthétiser les conclusions tirées àpd toutes ces données + intuitions
8. Former de nouvelles hypothèses (pour de futurs tests)
9. Recommencer

➙ Les étapes inductives (3,7-8).

DM et Statistiques

Tentative de définition du EC. (ML.) :

Extraction non-trivial d'information implicite, inconnue d'avance et potentiellement utile à partir de (larges) données (Frawley et al., 1991).

• **Modélisation Statistique** : utilisation des algos paramétriques statistiques pour grouper ou prédire un résultat / un evt. àpd données spécifiquement assemblées pour ce propos (e.g. en réponse à certaines questions).

• **Data mining** : utilisation des algos de ML pour trouver des motifs ou relations (e.g. corrélation) entre les éléments sur de grandes quantités de données éventuellement erronées, bruitées ou incomplètes menant à des décisions/actions qui améliorent le diagnostic, la prédiction, la détection, etc.

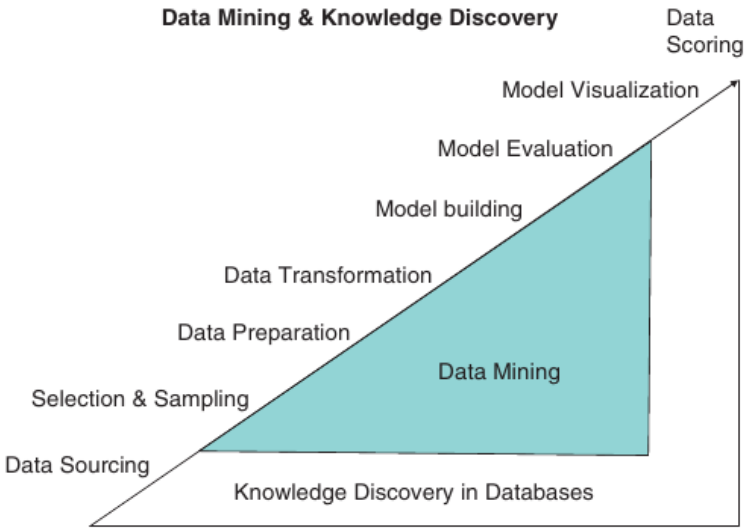
DM et Statistiques (suite)

- Pour la prédiction, le DM utilise les techniques de **Reconnaissance de Formes** développées en ML, avec beaucoup de données éventuellement bruitées, données parfois avec plus d'attributs que de données, ...
  - Parfois, de manière *non-supervisée* : on ne sait pas ce qu'on découvrira.
  - Le DM ne s'intéresse pas toujours au processus qui a généré les données
- Les Stats (sous domaine des Maths) utilisent la probabilité et l'optimisation.
  - Elles s'intéressent à comprendre le processus qui a généré les données.
  - En Stats, on commence en général avec un modèle préalable (hypothèses) dont seront dérivés des procédures pour améliorer (et optimiser) ce modèle.
    ➙ cf. la technique de base de de régression (utilisée également en DM).
  - L'analyse statistique traditionnelle suit une méthode **déductive** dans la recherche de relations dans une BD.

DM, Stats et Extraction de Connaissances (KD)

• **Knowledge Discovery** (KD englobant DM) : le processus complet d'accès, exploration, préparation, modélisation de données et déploiement et pilotage (monitoring) de modèles ➙ KDD inclue les techniques statistiques.

• KD sur les BDs est un processus non-trivial d'identification de nouveaux motifs valides, utiles et compréhensibles.



Analyses et modélisations en DM

- **Analyse exploratoire des données** :
  ➙ techniques visuelles et interactives qui permettent de ”voir” les données sous forme de résumés statistiques pour ”se faire une idée” des motifs, tendances (et données *ex calibur*).
- **Modélisation prédictive : classification et régression**
  ➙ Le but est de prédire la valeur d'une variable en fonction d'autres.
- **Découverte de motifs et de règles** :
  ➙ Création de règle d'association / de classification.
- **Extraction par contenu :**
  ➙ Recherche de ”motifs” similaires à la requête.
  ➙ Souvent textuel/image.

Analyses et modélisations en DM (suite)

- **Modélisation Descriptive :**

  ➙ Une vue de niveau plus élevé des données qui peut contenir :

  - La détermination d'une *distribution* de probabilité générale
    → Par ex. une estimation de densité(s)
  - Les Modèles décrivant les *relations* entre les variables
    → modélisation des dépendances, corrélations
    → sous forme de règles, arbres, équations, vraisemblance, RN, BN, &c.
  - Le Partitionnement des données en groupes par l'analyse de *clusters* ou par *segmentation*.

- Dans le cas de *Clustering*, il n'y a pas de *a priori* dans la recherche des groupes "naturels" sauf peut être le nombre de cluster précisé d'avance.

- Dans le cas de la *Segmentation*, on veut trouver des groupes homogènes en relation à une variable (e.g., des segments de clients les plus dépensiers).

Quelques autres exemples d'applications récentes

- Sécurité en aviation (méthodes DM basées sur les rapports d'accidents)
- Prédiction de ventes en Box-office (films)
- Détection de clients insatisfaits (banques)
- Evaluation de risques de crédit (Cartes crédits, en Allemagne)
- Analyse de clients "perdus" sur une période donnée,
- Étude et prédiction de marchés d'automobiles (à base de texte)
- Prédiction de processus industriels / pannes / maintenance / **Fiabilité**
- Administration et gestion dans l'industrie pharmaceutique :
  ➙ Prédiction du nbr de jours d'hospitalisation pour les maladies mentales
- Psychologie clinique : meilleure thérapie pour un patient donné
- Découverte de relations structurelles entre le niveau d'éducation (études) et leadership industriel (exemples universités).
- Dentaire : étude de douleurs faciales (selon 84 variables de prédiction)
- Prédiction de maladie sur la base d'auto-évaluation (via Internet)
- Déchiffrement génétique (découverte de gènes resp. de maladies),
- ....

Deux exemples représentatifs

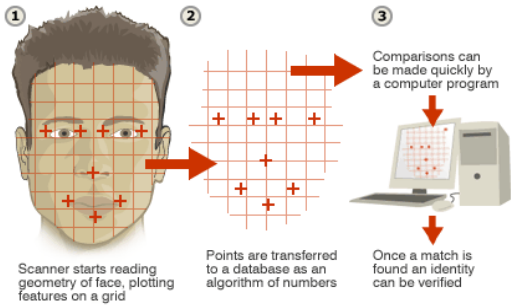
- **Un exemple** (domaine séquencement génétique, difficile) :

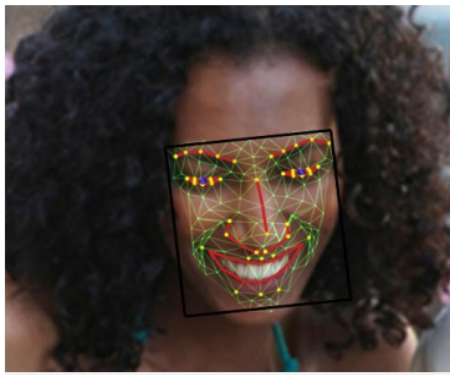
```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKA
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKA
                                     ****: .***:  * *:** * :****.:* ******

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHN
                 **** *:************:***:**.: .***************     :  *.: :
```

➙ Fait appel aux techniques avancées du "Text Mining".

- **Un autre exemple** (reconnaissance faciale, difficile) :





Objectifs de l'Extraction de Connaissances

• La tâche de découverte de connaissances à partir de données (KDD) est divisée en 2 catégories principales :

➤ **Prédictive :** prédiction des valeurs des attributs (dépendantes) en fonction des variables d'explication (indépendantes)

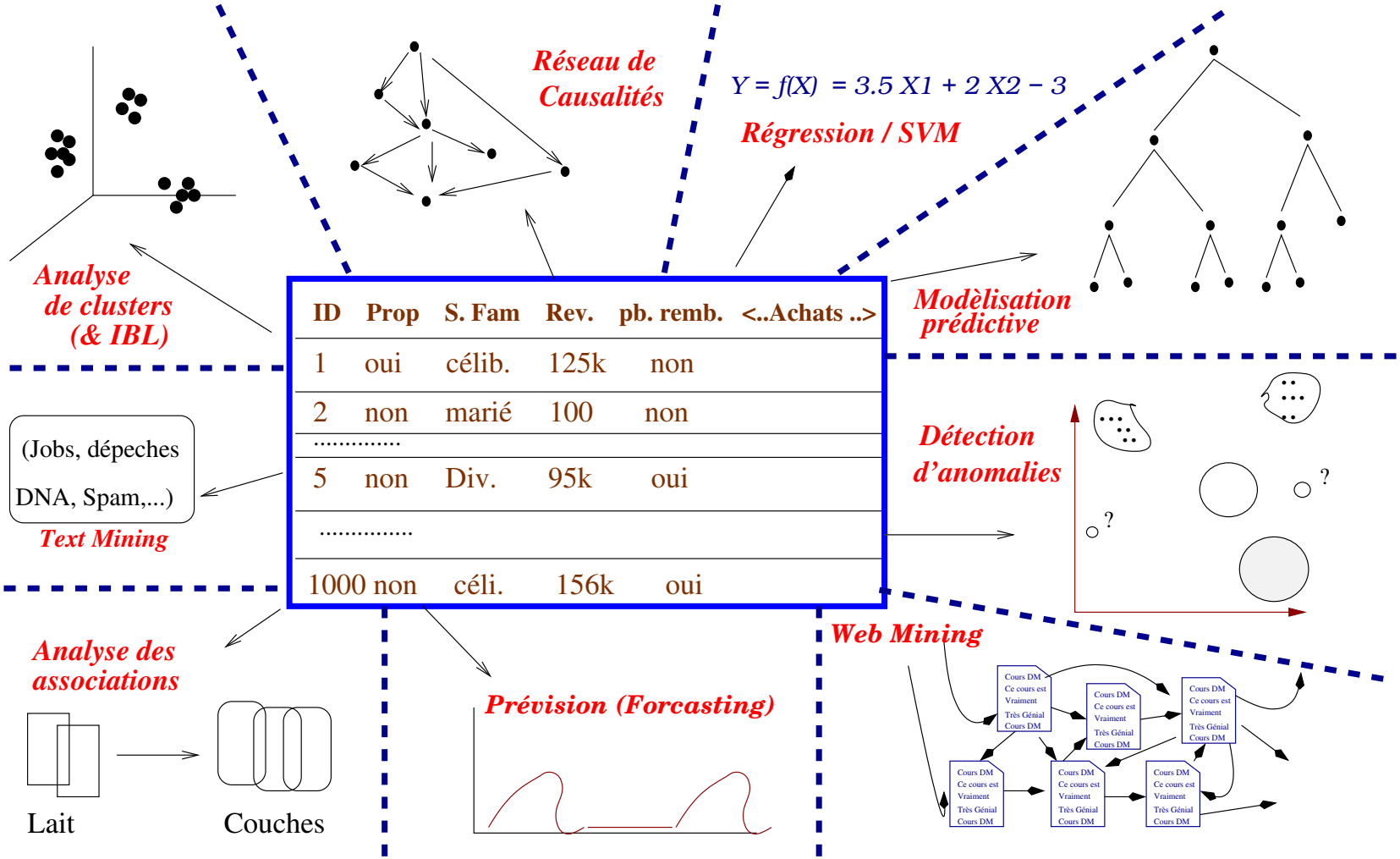
➤ **Descriptive :** Extraction de motifs (corrélations, clusters, tendances ou anomalies, ...) qui résument les relations entre les données.

• Cette apprentissage peut se faire de deux manières principales :

➤ **Supervisé** = création de modèles en formant des *définitions de concepts* à partir des données contenant des classes prédéfinies.

➤ **Non supervisé** = création de modèles à partir de données sans l'aide de classes prédéfinies

Principales sorties de l'EC



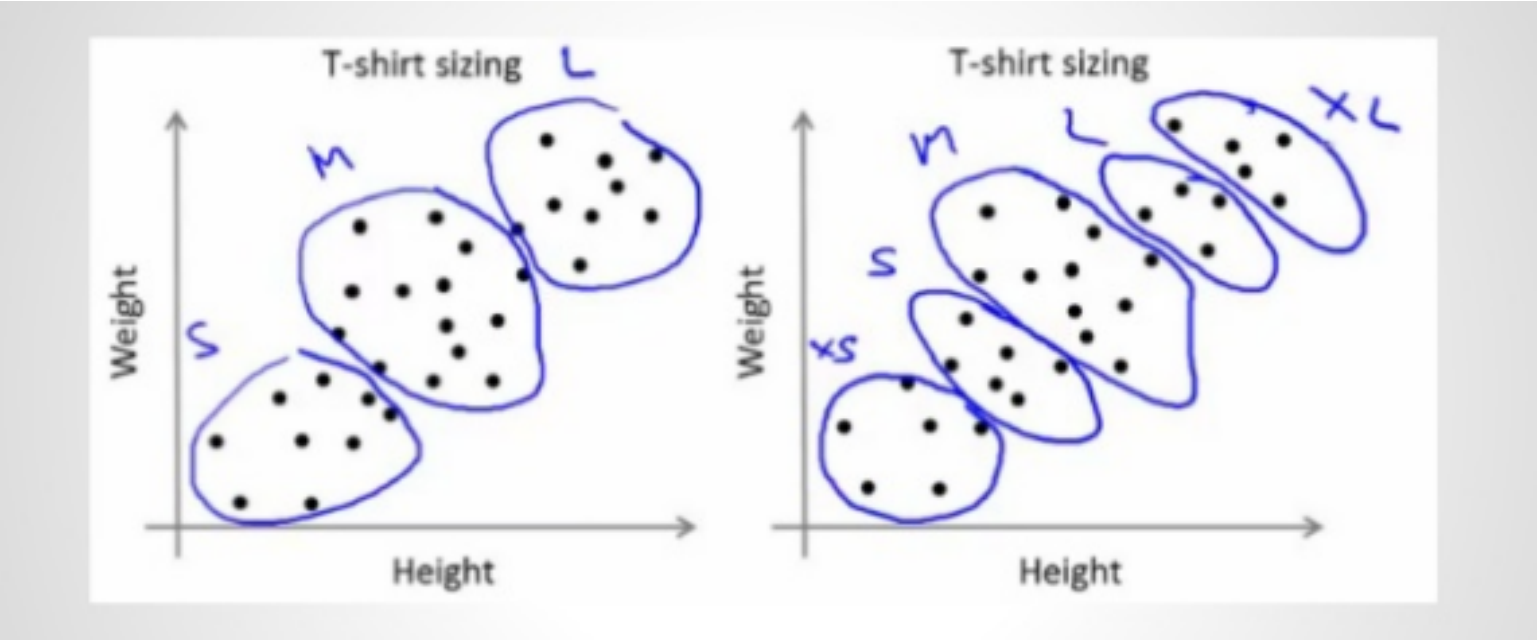
Base de données

Les données peuvent être :

- transactionnelles (supermarchés, banques, données saisonnières, etc.)
- images (pixels),
- son (signaux),
- texte (mail, texte littéraire, séquences génétiques, des nombres, etc.),
- les pages WEB comme données multimédia (au sens "plus d'un médium"),
- courbes, histogrammes, fonctions, relations, graphes ou arbres, etc.

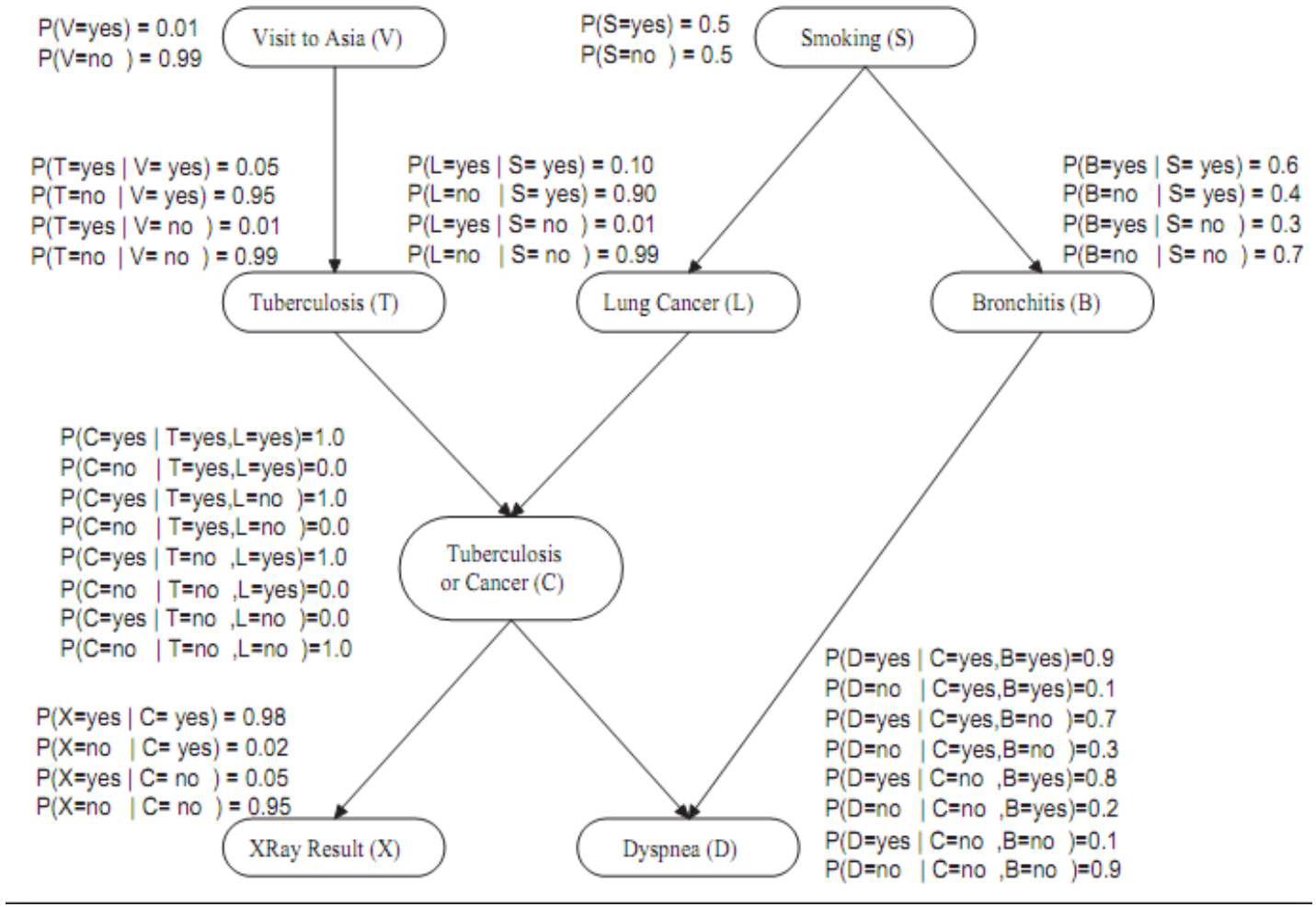
➙ Quelques exemples de modèles ../..

Clustering & IBL

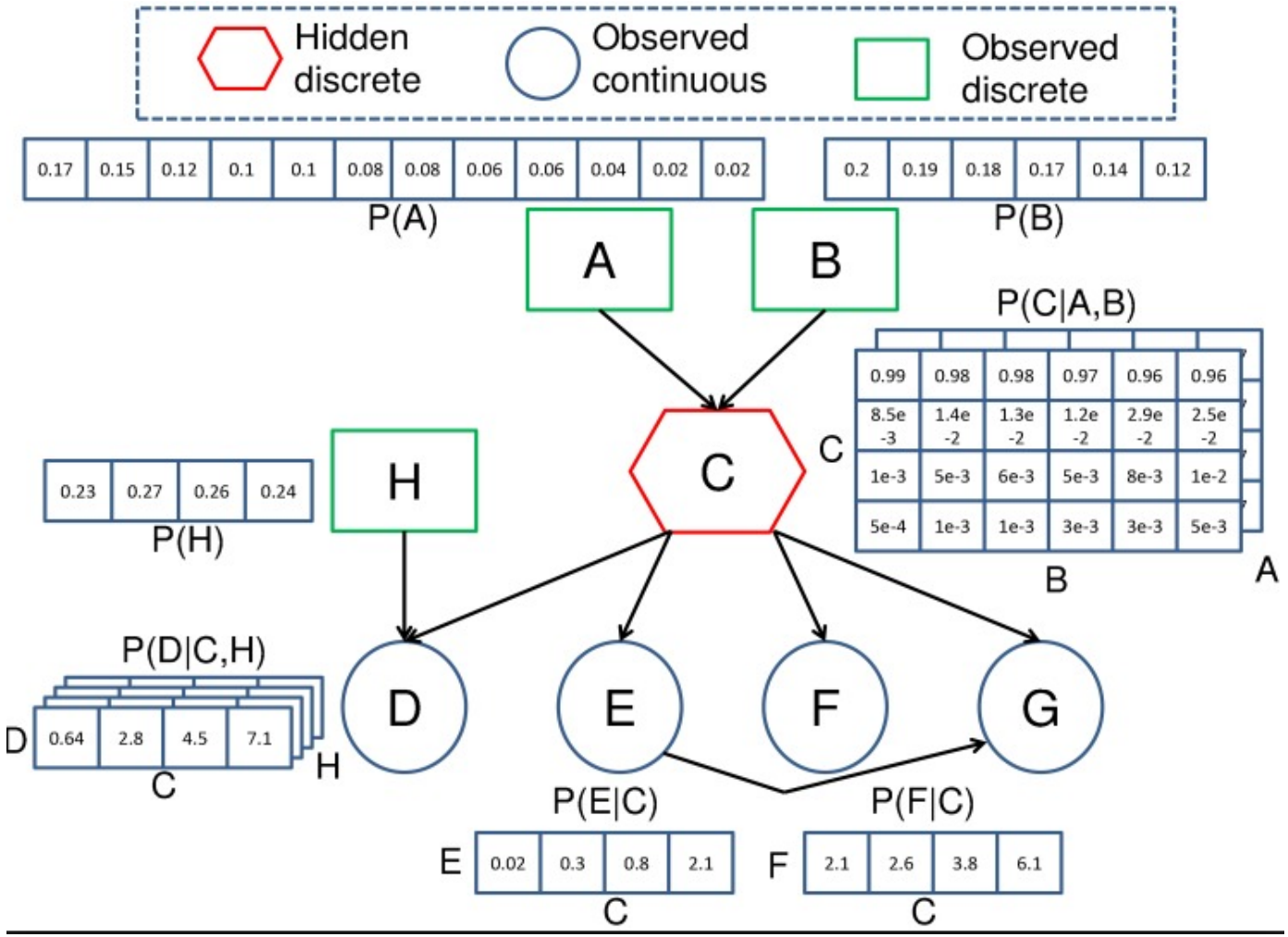


- IBL (généralisé en CBL) sur la base de KNN.

Construction de réseaux de "causalités" (BN)

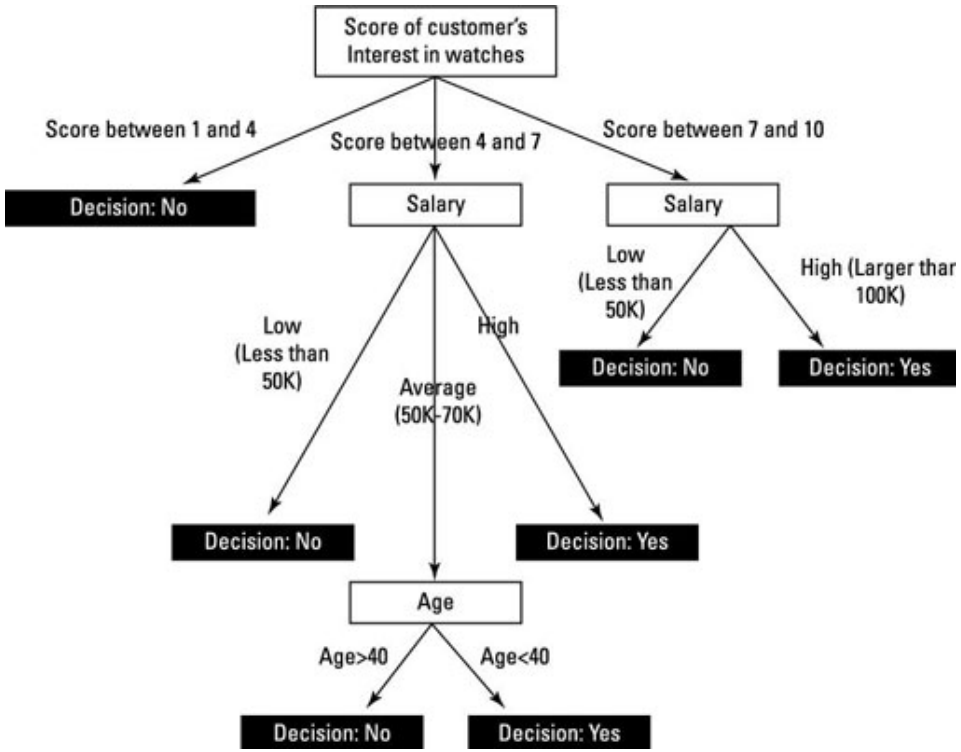


Construction de réseaux de "causalités" (BN) (suite)



Arbre de décision

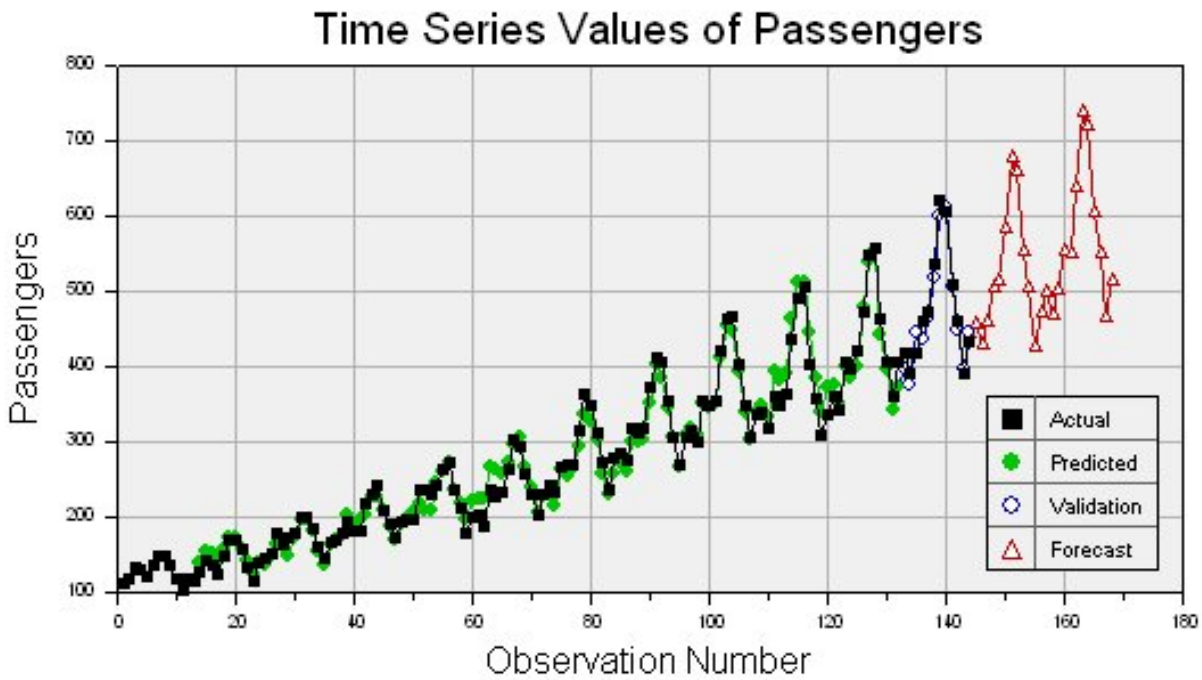
- Construction d'arbre de décision (peut être transformé en règles)



Détection d'anomalies (et d'outsiders)

Données saisonnières

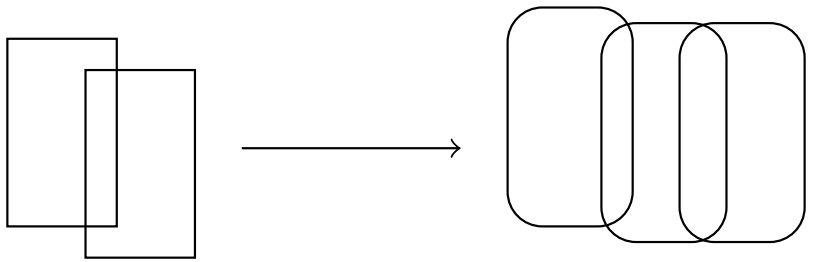
• Prévision sur des données saisonnières (par ex. en météo, en Finances, prévision de charges, etc.)



Recherche d'associations

- Recherche d'associations / corrélations dans les données

Analyse des associations



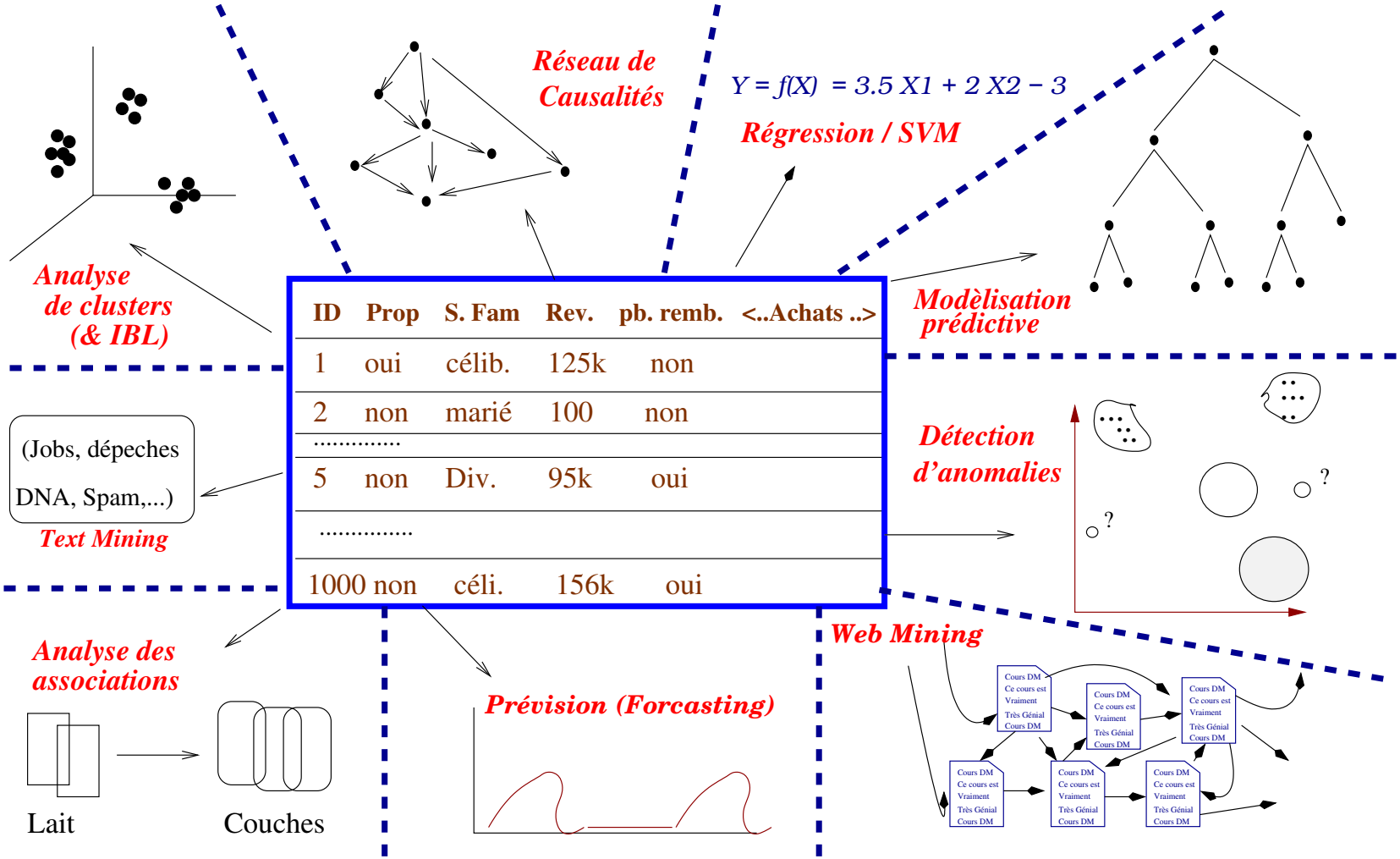
Text Mining

Recherche de motifs dans le texte (ou une donnée séquentielle) :

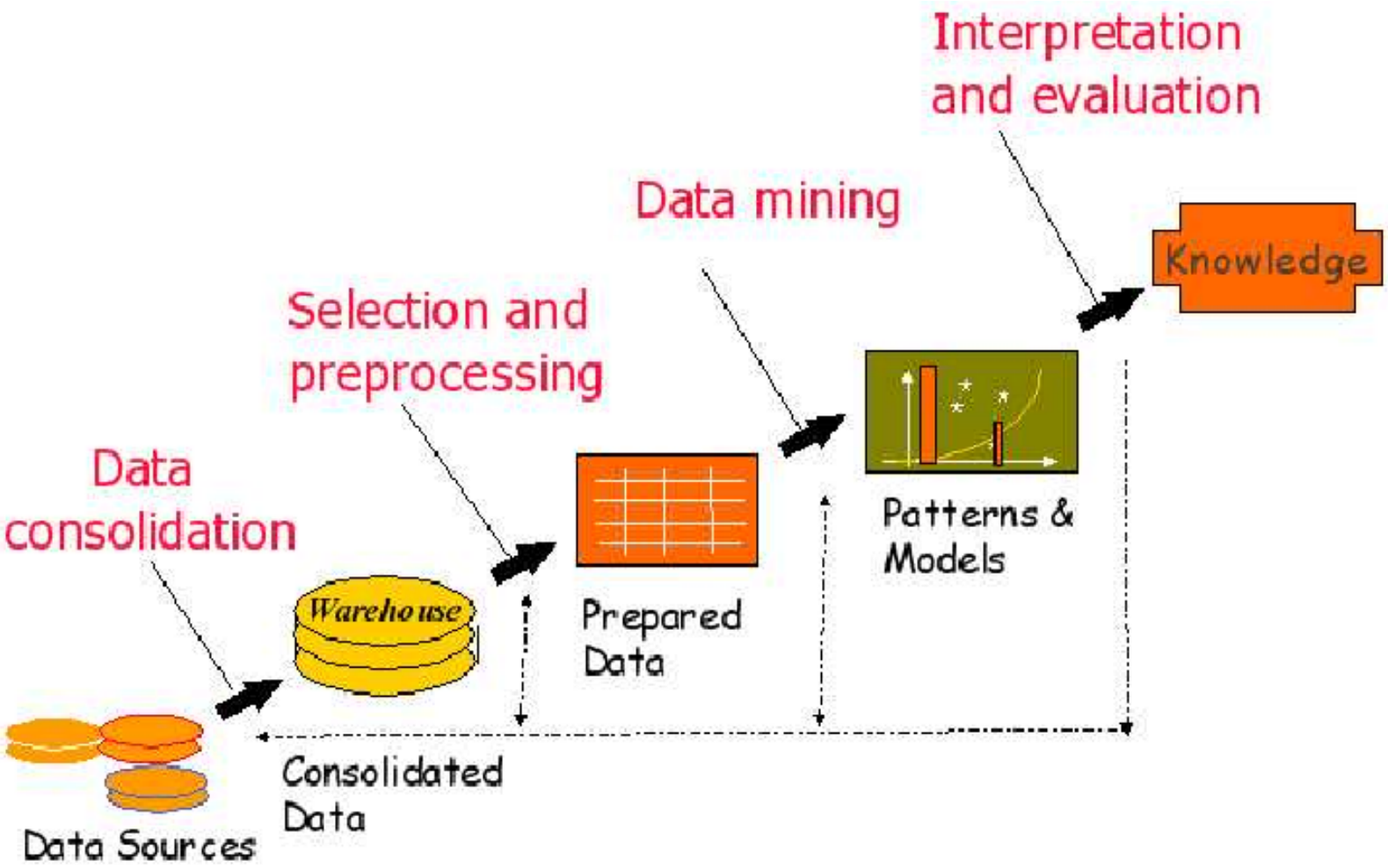
➛ Jobs, SPAM, DNA, dépêches, ...

Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications (2012)

Vue d'ensemble

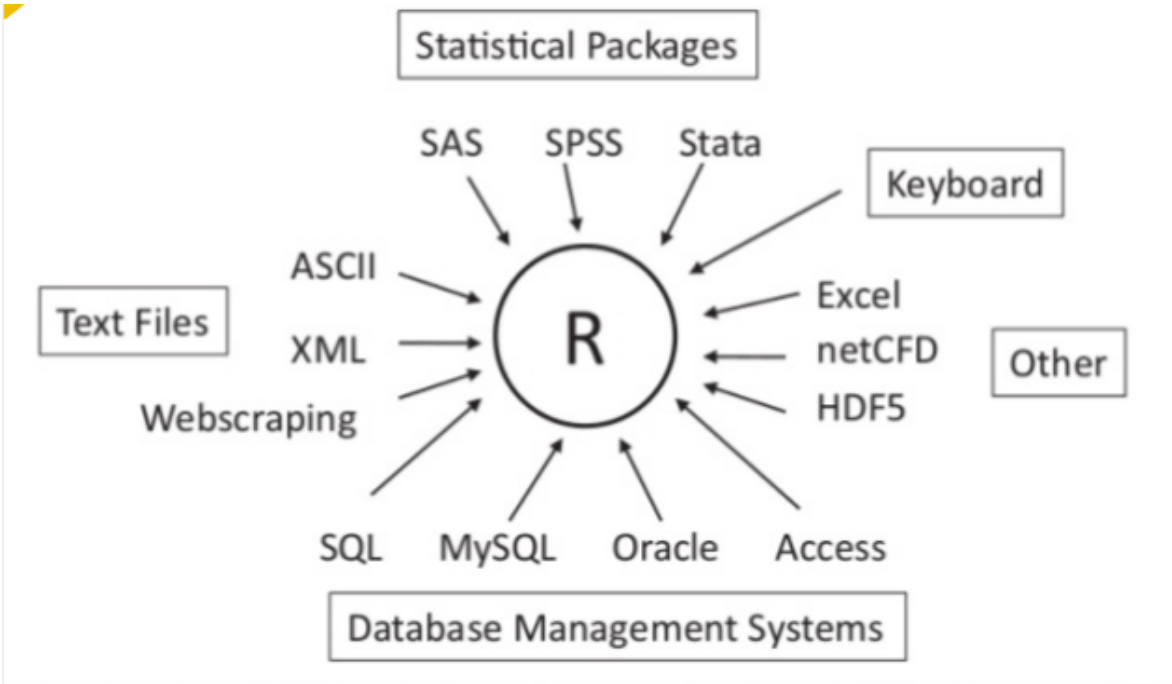


Processus KDD



Quelques Outils

- Outils utilisés : Weka, R, Python, (TANAGRA, Octave, ...)
- R :



- DeepNet : outils Deep Learning pour R

Big Data : eldorado ?

- Big Data ? quantité vs. qualité ?
- CNN , TextMining.
- Quelques outils de *Machine learning* Open-source :
  - *TensorFlow* de Google (graphe de noeuds – opérateurs – et des arêtes – tables de données=tensors – + calculs massivement parallèles)
  - *Sci-Kit Learn*, *Theano, Keras* de Python
  - *Torch* (à base de Lua), *PyTorch* (Torch en Python)...
  - Voir *http ://www.kdnuggets.com/2015/12/deep-learning-tools.html*
- Émergence de **"ML-as-a-service"**
  → *Cloud Prediction API* de Google
  → *Azure Machine Learning platform* de Microsoft

Qu'est-ce que les ordinateurs peuvent apprendre

- Le processus d'apprentissage est complexe.

➙ On manipule et apprend :

Fait : une simple expression de la "vérité" (fait avéré)
$\rightarrow 2 \times 2 = 4$

Concepts : un ensemble d'objets, symboles ou événements groupés ensemble qui ont quelques caractéristiques en commun
$\rightarrow$ *un oiseau*

Procédure : une séquence pas à pas d'actions pour résoudre un problème
$\rightarrow$ *scenario de construction de ... X*

Principes : niveau supérieur d'apprentissage des vérités générales / des lois basiques utiles à d'autres faits/vérités.
$\rightarrow$ *la règle de 3*

Qu'est-ce que les ordinateurs peuvent apprendre (suite)

Les ordinateurs apprennent bien les concepts :

- **Concept** = l'issu d'un processus de Extraction de Connaissances .

→ Exemples : *maison, homme, automobile, bon client, comestible, panne ...*

- La forme des concepts appris est imposé par l'outil DM choisi.
- Les *concepts appris* sont représentés par :

  arbres, règles, réseaux, équations/fonction, ...

- Les concepts appris peuvent être **expliqués ou non** :
  - ➙ Les arbres (de décision) et les règles (compréhensibles pour l'humain)
  - ➙ Idem pour les réseaux Bayésiens (graphes)
  - ➙ Les réseaux (de neurones) et les équations, ...(moins évidents !)
- Le "comment / pourquoi" :
  - ➙ Modèles *Black-Box* vs. *Explicatifs*.

Recherche de Concepts par Généralisation

Enumération de l'espace de concepts : Espace de Versions

Une définition théorique qui permet de comprendre.

- L'espace de description de concepts consistants :
  - ➥ **L** (least) et **G** (greatest) general descriptions

    L : les descriptions *les plus spécifiques* qui couvrent tous les exemples positifs et aucun exemple négatif (spécifique).

    G : les descriptions *les plus générales* qui ne couvrent aucun exemple négatif mais tous les exemples positifs (générale).

- On a juste besoin de maintenir à jour **L** et **G**

- **Utilité :** définition théorique *générative*
  - ➥ C'est encore couteux en temps de calcul et.
  - ➥ Ne résout pas de problème pratique !
  - ➥ Est rendue praticable par des algos + heuristiques

Recherche de Concepts par Généralisation (suite)

Exemple :

Soit le vocabulaire : couleurs ∈ {rouge, vert}, animaux ∈ {vache, poule}

Ex. positifs	Ex. négatifs	L	G
<>	<>	{ }	{< * , * >} (a priori)
<vache, verte>		{<vache, verte>}	{< * , * >}
	<poule, rouge>	{<vache, verte>}	{<*, verte>, <vache, *>}
<poule, verte>		{<*, verte>}	{<*, verte>, <vache, *>}

• Si l'on ajoute <vache, rouge> ? :

→ Si ex. positif : ajouter <vache, *> à L

→ Si ex. négatif : retirer <vache, *> de G

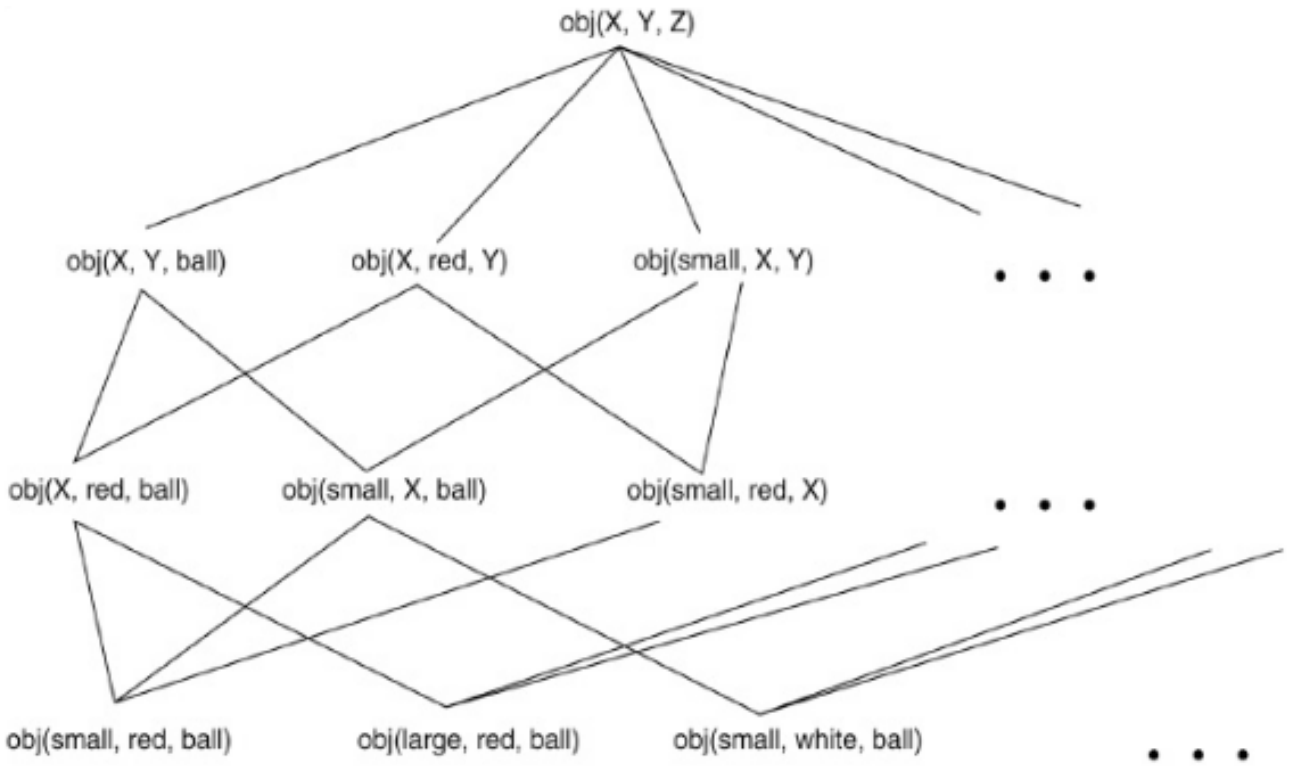
• Le modèle appris sera *inconsistance* si les exemples positifs et négatifs se contredisent.

• L'ensemble G dénote les phrases qu'on pourrait construire sur un alphabet.

• Définition claire et théorique mais peu opérationnelle .

Espace de version : un exemple synthétique

Taille : {large, small}
Couleurs : {red, white, blue}
Obj : {ball, brick, cube}



Expression opérationnelle des concepts

Le modèle appris peut être présenté par différentes vues :

■ 1- Une **vue classique** : définition claire et déterministe du concept

➙ Tous les exemples d'un concept particulier sont, de manière égale, représentatifs de ce concept.

➙ On peut voir (tester) si un individu est un exemple d'un concept particulier.

• Exemple : définir un risque acceptable pour un prêt (un bon risque = un bon client) :

Si les revenus annuel $\geq$ *30,000*
& *L'ancienneté dans l'entreprise* $\geq$ *5 ans*
& *Possède sa maison = vrai*
Alors Risque de prêt acceptable = vrai

Expression opérationnelle des concepts (suite)

■ 2- Une **vue probabiliste** : les concepts représentés par des propriétés probables après la généralisation des **instances**.

• Exemple : une vue probabiliste d'un risque acceptable (pour un prêt) :

> • *La moyenne annuelle des revenus des individus qui paient à temps leurs mensualités est de 30000.*
> • *La plupart des individus qui sont des bons risques de crédit ont travaillé au moins 5 ans pour la même entreprise.*
> • *La majorité des bons clients (risques) de crédit possèdent leur propre maison.*

➥ Des grandes lignes des caractéristiques d'un bon risque
→ aide à la décision.

• Une vue probabiliste associe une probabilité à une classification :

> ⇒ *Un client possédant sa maison avec un revenu annuel de 27000, embauché dans la même entreprise depuis 4 ans peut être classé à 0.85 comme un bon risque.*

Expression opérationnelle des concepts (suite)

■ 3- Une **vue à base d'instance** :

⇒ une nouvelle **instance** est une instance (exemple) d'un concept si elle est **assez proche** d'un ensemble d'un ou plusieurs exemples connus du concept

- L'humain utilise souvent cette vue.
- Exemple : une liste possible de bons risques :

Instance 1 *: revenus annuel = 32000 & dans la même boîte=6 & Possède sa maison*
Instance 2 *: revenus annuel = 52000 & dans la même boîte=16 & Locataire*
Instance 3 *: revenus annuel = 28000 & dans la même boîte=12 & Possède sa maison*

- On peut associer une probabilité (d'appartenance au concept) à chaque classification.

Aperçu des méthodes étudiées dans ce cours

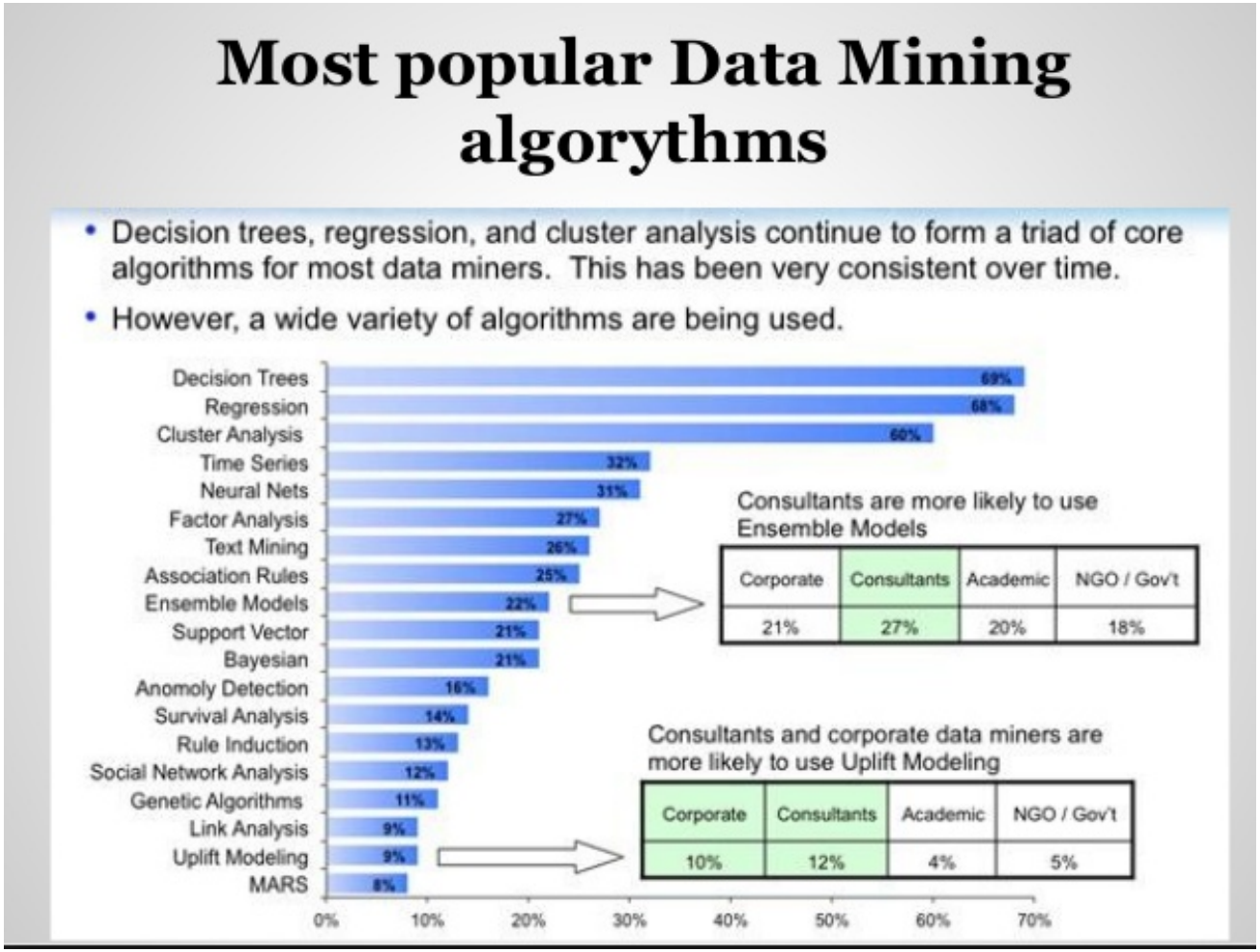
- Différents types de Motifs (Patterns) ?
- La forme des entrées ?
- Comment les décrire ?

... Voyons un exemple .... mais notons :

- L'expertise est indispensable en apprentissage (supervisé ou non supervisé)
- Un résultat est toujours accompagné d'indicateurs (erreur, intervalle de confiance, mesures diverses, ...)

Aperçu des méthodes étudiées dans ce cours (suite)

Les algorithmes :



Exemple d'un jeu

- Conditions pour jouer à un jeu (tennis, golf, météo).

Num	Temps(S)	Temperature(T)	Humidité(H)	Vent(V)	Jouer
1	ensoleillé	Elevée	Elevée	non	Non
2	ensoleillé	Elevée	Elevée	oui	Non
3	nuageux	Elevée	Elevée	non	Oui
4	pluvieux	Moyenne	Elevée	non	Oui
5	pluvieux	Faible	Normale	non	Oui
6	pluvieux	Faible	Normale	oui	Non
7	nuageux	Faible	Normale	oui	Oui
8	ensoleillé	Moyenne	Elevée	non	Non
9	ensoleillé	Faible	Normale	non	Oui
10	pluvieux	Moyenne	Normale	non	Oui
11	ensoleillé	Moyenne	Normale	oui	Oui
12	nuageux	Moyenne	Elevée	oui	Oui
13	nuageux	Elevée	Normale	non	Oui
14	pluvieux	Moyenne	Elevée	oui	Non

TABLE 1: Données météo (jeu)

Exemple d'un jeu (suite)

- Les exemples (*instances*) sont décrits dans une BD par des **attributs**.
- Données : 5 attributs : {*temps, température, humidité, vent*}
  - ➥ Et la décision = la **sortie = la classe** : peut-on jouer (oui/non) ?
- 36 combinaisons d'attributs possibles (3 x 3 x 2 x 2 = 24) dont 14 dans la BD

Résultats d'une 1er tentative d'apprentissage :

- *Si aspect=ensoleillé & humidité=forte alors jouer=non*
- *Si aspect = pluvieux & vent = vrai alors jouer=non*
- *Si aspect = nuageux alors jouer=oui*

...

Entrées : attributs mixtes

- Les attributs *température* et *humidité* sont numériques

Num	Temps(S)	Temperature(T)	Humidité(H)	Vent(V)	Classe
1	ensoleillé	81	78	non	N
2	ensoleillé	80	90	oui	N
3	nuageux	83	80	non	P
4	pluvieux	75	96	non	P
5	pluvieux	69	75	non	P
6	pluvieux	64	70	oui	N
7	nuageux	65	65	oui	P
8	ensoleillé	72	83	non	N
9	ensoleillé	68	72	non	P
10	pluvieux	71	74	non	P
11	ensoleillé	75	69	oui	P
12	nuageux	70	77	oui	P
13	nuageux	85	70	non	P
14	pluvieux	73	82	oui	N

- Un exemple de règle :

  Si aspect = ensoleillé & humidité > 83 Alors jouer=non

- **Cas d'attributs numériques** et **mixtes** ..

Prédiction : Règles d'association

- Les exemples de règles ci-dessus sont des **règles de classification**
  - ➥ elles (prédisent) la classe d'un exemple → "jouer, ne pas jouer".
- Au lieu de classifier, on peut trouver des règles qui associent les attributs
- Exemple de *règles d'association* dérivables de la même table :

> - *Si température=normale Alors humidité = normale.*
> - *Si humidité = normale & vent = faux Alors jouer=oui*
> - *Si aspect = ensoleillé & jouer=non Alors humidité=forte.*
> - *Si vent = faux & jouer=non Alors aspect=ensoleillé & humidité =forte.*

- Toutes ces règles sont justes (sur les exemples donnés) :
  - ➥ pas de fausse prédiction.
  - ➥ Les 2 premières règles couvrent 4 cas ; la suivante à 3 et la 4e à 2 cas.

Prédiction : Règles d'association (suite)

- Les règles d'association peuvent **prédire** tout attribut
  - ➥ Elles peuvent même prédire plus d'un attribut
  - ➥ Par exemple, la règle :

Si vent = faux & jouer=non Alors aspect=ensoleillé & humidité = forte.

prédit que l'aspect du temps sera ensoleillé et l'humidité sera forte.

- Il y aura beaucoup plus de règles que pour la classification
  - ➥ On peut en trouver env. 60 règles applicables à au moins 2 cas.
  - ➥ Et beaucoup plus si l'on accepte celles pas tout à fait correctes (à 100%)
    → notion de degré de **confiance**

Résumé des données

- **Objectif** : prescription de lentilles de contact par un opticien

Age (ophtalmo.)	Prescription (diagnostique)	Anomalie de réfraction	Effet Lacrymal	Type lentilles
Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

TABLE 2: de contact

Résumé des données (suite)

- Lentilles de contact $\in$ {*souple , rigide, rien*} : (soft, hard, none).
- But : correction d'une anomalie visuelle :

  {*myopie, hypermétropie, astigmatisme et presbytie*}.
- **Attributs** : {*âge, problème, astigmatisme, effet-lacrymal , prescription*}.
- **Sorties** : une règle = une description structurelle (attributs)

Si effet-lacrymal = réduit Alors recommandation=non
Sinon si âge=jeune et astigmate=non Alors recommandation=soft.

- On a 5 attributs dont 4 *explicatives* : 3 x 2 x 2 x 2 = 24 combinaisons
- La table 2 : 24 lignes, combinaisons→ tous les cas présents ! (sans la classe)
  - ➥ Table simple, les règles résument les données. (cas particulier)
  - ➥ Dans ces cas, on tente de **résumer** les données en les **généralisant**.

Arbres de décisions

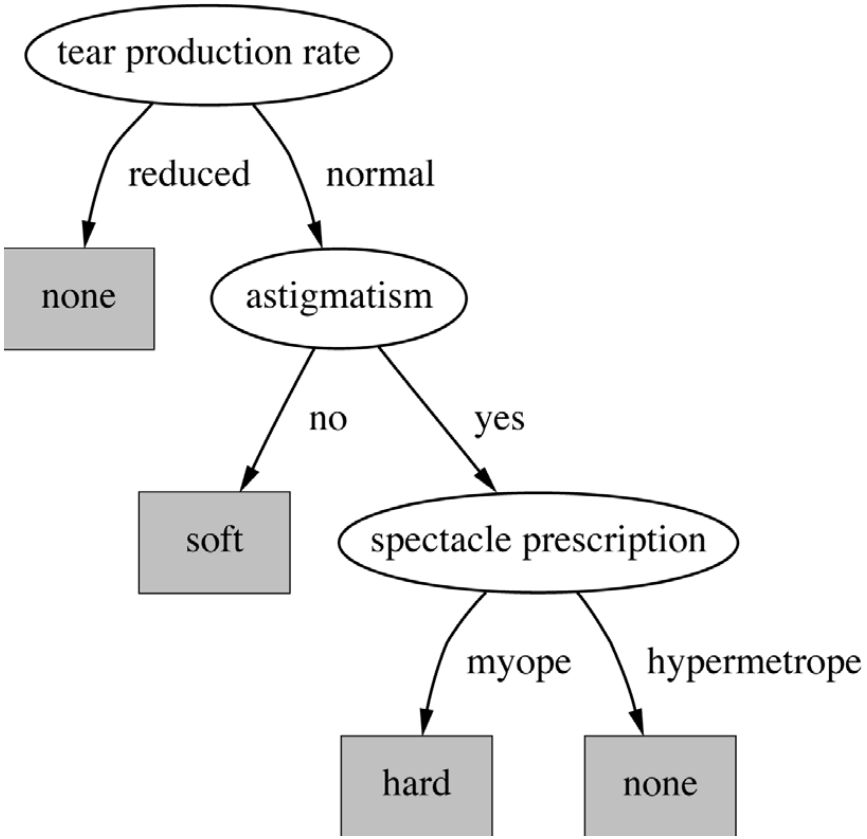


FIGURE 8.1: Arbre de Décision pour le problème des lentilles

Arbres de décisions (suite)

• **Un arbre de Décision** est plus compacte et plus lisible que les règles de classification.

• Les branches sont calculées à l'aide de l'*entropie*.

• Chaque décision peut être accompagnée d'un taux d'erreur (incertitude).

• On se posera souvent une question importante :

quelle représentation préférer ?

⇒ Classification (règle ou arbre) ou association ? (hormis le clustering)

⇒ Données explicatives (*support/confiance, etc.*) ou black-box ?

Prédiction numérique

- Performances relatives des PC selon certains attributs.

	Main Memory (Kb)				Channels		
	Cycle Time (ns)	*Min*	*Max*	Cache (KB)	*Min*	*Max*	Performance
	MYCT	*MMIN*	*MMAX*	*CACH*	*CHMIN*	*CHMAX*	*PRP*
1	125	256	6000	256	16	128	198
2	29	8000	32,000	32	8	32	269
3	29	8000	32,000	32	8	32	220
4	29	8000	32,000	32	8	32	172
5	29	8000	16,000	32	8	16	132
...							
207	125	2000	8000	0	2	14	52
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

TABLE 3: Données Performances CPU

$$PRP = -55.9 + 0.0489MYCT + 0.0153MMIN + 0.0056MMAX + 0.6410CACH - 0.2700CHMIN + 1.480CHMAX$$

- 209 configurations différentes(une ligne = une configuration)

  $\Rightarrow$ Les attributs et la sortie sont tous **numériques**.

Modélisation Bayésienne

Rappel de la BD :

Outlook	**Temp.**	**Humidity**	**Windy**	**Play**
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Modélisation Bayésienne (suite)

- Deux hypothèses : les attributs sont
  - ➙ d'importance égale,
  - ➙ distribution normale
  - ➙ statistiquement indépendantes (vis à vis de la classe)

  - ➥ Indépendance : les Connaissances sur la valeur d'un attribut particulier ne disent rien sur la valeur d'un autre attribut (pour une classe connue)

La probabilité conditionnelle de Bayes

- L'hypothèse H et l'évidence E basée sur H :

$$\boxed{\Pr[\mathrm{H} \mid \mathrm{E}] = \frac{Pr[E|H] \times Pr[H]}{Pr[E]}}$$

➥ **Hypothèse** (naïve) de Bayes (indépendance) :

l'évidence E se décompose ici en ses composantes indépendantes p/r à la classe (les attributs de l'instance) :

$$\boxed{\Pr[\mathbf{H} \mid \mathrm{E}] = \frac{Pr[E|H] \times Pr[H]}{Pr[E]} = \frac{Pr[E_1|H] \times Pr[E_2|H] \times \cdots \times Pr[E_n|H] \times Pr[\mathbf{H}]}{Pr[E]}}$$

- Pour classer (e.g. play=yes) un nouvel exemple dépendant des 4 attributs :

$$\boxed{\Pr[\mathbf{yes} \mid \mathrm{E}] = \frac{Pr[Outlook|yes] \times Pr[Temp|yes] \times Pr[Hum|yes] \times Pr[Windy|yes] \times Pr[\mathbf{yes}]}{Pr[E]}}$$

On ne peut multiplier les probabilités que sous l'hypothèse de l'indépendance.

La probabilité conditionnelle de Bayes (suite)

- Par exemple (ex. d'évidence E) :

Outlook	Temperature	Humidity	Windy	Play
Sunny	Cool	High	True	? ?

- Quelle est la probabilité de "yes" pour la nouvelle instance (E ci-dessous) ?

$$\textbf{Pr[yes | E]} =$$

$$\textbf{Pr[Outlook = Sunny | yes]} \times \textbf{Pr[Temperature = Cool | yes]} \times \textbf{Pr[Humidity = High | yes]} \times \textbf{Pr[Windy = True | yes]} \times \frac{Pr[Yes]}{Pr[E]}$$

$$= \frac{2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14}{Pr[E]}$$

$\Rightarrow$ On fait le même calcul pour $Pr[Play = no|E] = \dfrac{3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14}{Pr[E]}$

N.B. : Ici, $Pr[E]$ est la proba marginale de E et sert à la normalisation.

- Bayes et données numériques : hypothèse gaussienne et PDF.

Le réseau Bayésien de l'exemple météo

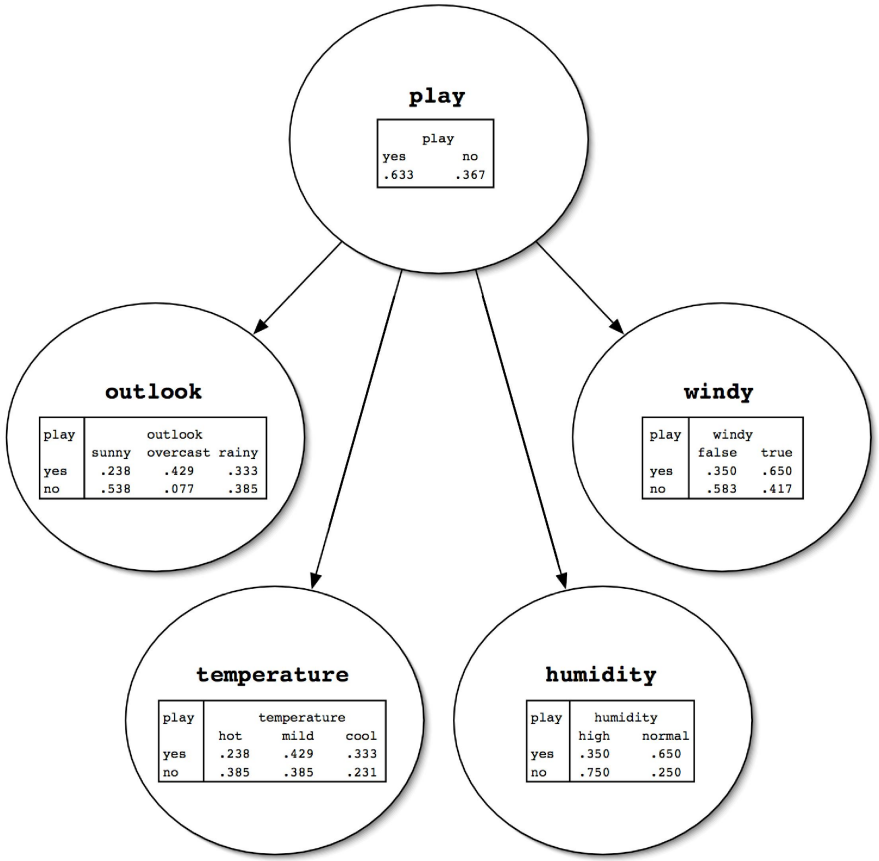


FIGURE 10.1: BN pour les données météo avec un seul parent

Le réseau Bayésien de l'exemple météo (suite)

Et un BN plus complexe :

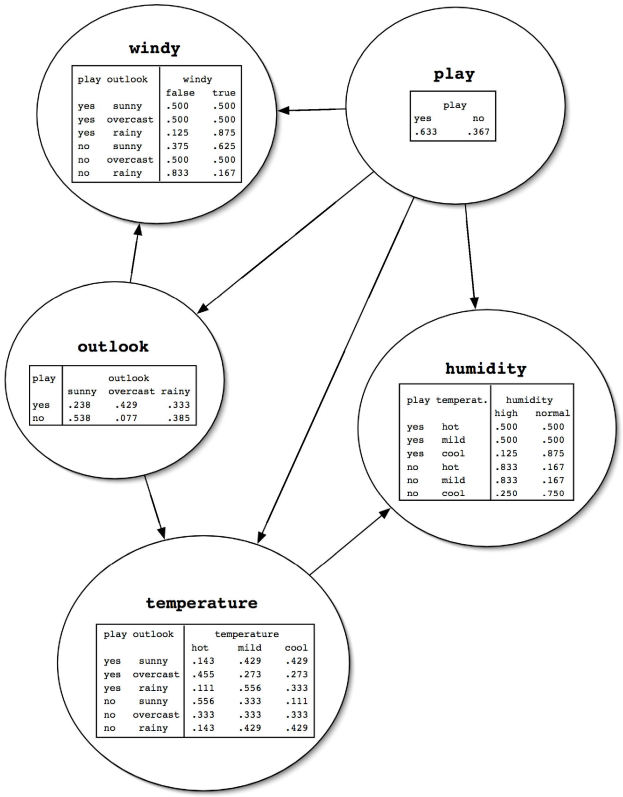


FIGURE 10.2: BN pour les données météo avec plus d'un parent

Non Supervisé : Clustering

- Construction d'un modèle sans classes prédéfinies.
- Groupage des données sur la base des **similarités**
- A l'aide de techniques d'évaluation, on définit/interprète le sens des classes formées.

ID Client	Type Cpte	Cpte Bourse	Méth. Transact.	Op/ mois	Sexe	Age Age	Sport favori	Rev./ an
1005	joint	non	en ligne	12.5	F	30-39	Tennis	40-59k
1013	ép-enf	non	courtier	0.5	F	50-59	Ski	80-99k
1245	joint	non	en ligne	3.6	M	20-29	Golf	20-39k
2110	indiv	oui	courtier	22.3	M	30-39	Pêche	40-59k
1001	indiv	oui	en ligne	5.0	M	40-49	Golf	60-79k

TABLE 4: Données de l'Investisseur de Stock ACME

- Exemple : table 4 de 5 clients :
- Pour un compte de type "Bourse", la banque prête de l'argent liquide.
- On veut trouver des "motifs" dans ces données.

Non Supervisé : Clustering (suite)

- **On se pose les questions suivantes** ($\rightarrow$ pub / clients actuels) :

1. Profil général d'un investisseur (client) en ligne ?
   ➙ Quelles caractéristiques ? Comment les distinguer d'un courtier ?
2. Peut-on savoir si un client sans "compte bourse" est susceptible d'en ouvrir un ?
3. Modèle de prédiction du nombre moyen de transactions (stocks achetés) / mois pour un nouveau client ?
4. Quelles différences entre les clients hommes et femmes ?, etc...

Mais aussi :

5. Quelles similitudes d'attributs peut grouper les clients de cette entreprise ?
6. Quelles différences d'attributs peuvent segmenter ces données ?

Non Supervisé : Clustering (suite)

Exemple de clusters : obtenus par une méthode non supervisée :

** Si Compte-bourse=oui & Age=20-29 & Revenus-annuels=40-59k*
Alors Cluster=1 {justesse=0.80, couverture=0.50} *(1)*

** Si Type-compte=épargne-enfant & Sport-favori=Ski & Revenus-annuels=80-99k*
Alors Cluster=2 {justesse=0.95, couverture=0.35} *(2)*

** Si Type-compte=joint & Nb-transactions > 5 & Méthode-Transaction=en-ligne*
Alors Cluster=3 {justesse=0.82, couverture=0.65} *(3)*

- Le cluster 1 : logique (bon sens) :
  ➙ clients plus jeunes, revenus raisonnables, approche moins conservative.
- Cluster 3 : ce n'est pas une découverte ! Mais
- **Cluster 2 : peut être intéressant.**

La compagnie ACME (American Company Making Everything) peut consacrer une partie de son budget pub dans les magazines de Ski avec promo sur les comptes type épargne-enfant (comptes sous tutelle).

Clustering Probabiliste

• D'un point de vu statistique, la tâche de Clustering est de trouver un ensemble de clusters les plus **vraisemblables**, étant donné une BD.

• En l'absence d'assez d'évidences (sur la décision de placer une instance définitivement dans tel ou tel cluster), on peut plutôt considérer la probabilité pour une instance d'être dans tel ou tel cluster .

Modèle de Mélange (Mixture) Finie

- La base de la méthode de Clustering Statistique est le modèle "mixture finie".
- Une **mixture** est un ensemble de K distributions de probabilités qui représenteront les K clusters.

  Chaque distribution est celle des attributs d'un cluster.

  ➥ Plus exactement, chaque distribution Θ_i donne la probabilité pour une instance particulière d'avoir un certain ensemble de valeurs d'attributs *"S'il devait appartenir réellement au cluster C_i"*.

- Les clusters ne sont pas semblables.
  ➥ Chaque cluster a une distribution qui représente sa population.
  ➥ Chaque instance appartient à un seul cluster mais l'on ne sait pas lequel.

Un exemple simple de Mélange

• Un seul attribut numérique, 2 Clusters, Distrib. Normales (ici 2 pour A et B), $P_A + P_B = 1$,

A	51	B	62	B	64	A	48	A	39	A	51
A	43	A	47	A	51	B	64	B	62	A	48
B	62	A	52	A	52	A	51	B	64	B	64
B	64	B	64	B	62	B	63	A	52	A	42
A	45	A	51	A	49	A	43	B	63	A	48
A	42	B	65	A	48	B	65	B	64	A	41
A	46	A	48	B	62	B	66	A	48		
A	45	A	49	A	43	B	65	B	64		
A	45	A	46	A	40	A	46	A	48		

FIGURE 13.1: Données de l'Exemple de Mixture finie des clusters A et B (connus)

Un exemple simple de Mélange (suite)

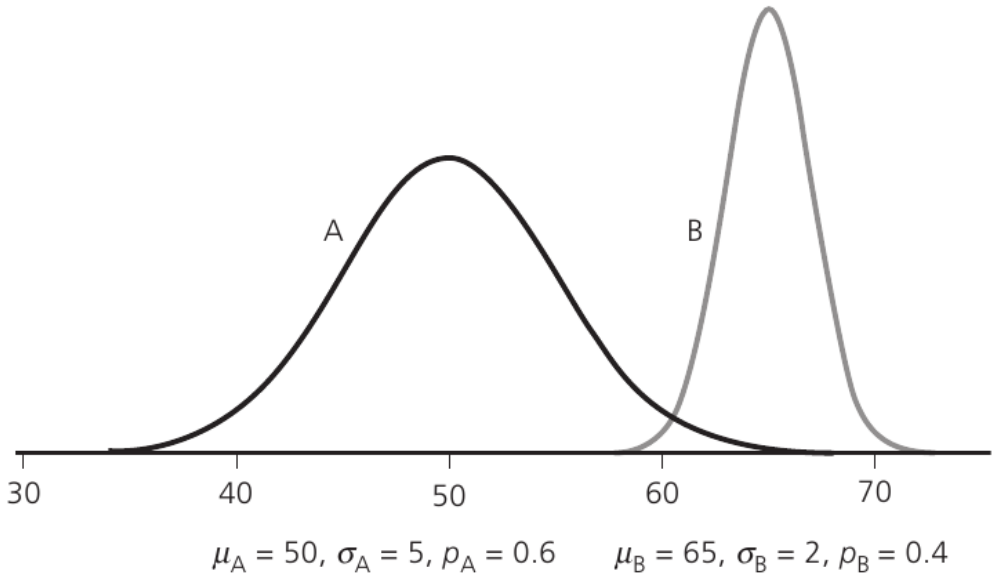


FIGURE 13.2: Distributions Normales de l'Exemple de Mixture finie

• L'objectif du clustering probabiliste est de calculer les clusters de sorte que chaque instance ait la meilleure probabilité d'appartenir à son cluster.

➙ Le cluster dont la distribution aura généré cette instance.

Clustering Hiérarchique

- Produit un ensemble de clusters imbriqués organisé comme un arbre Hiérarchique
- Peut être visualisé comme un **dendrogram**
- Deux méthodes : Agglomérative (ascendant) et Divisive (descendant)

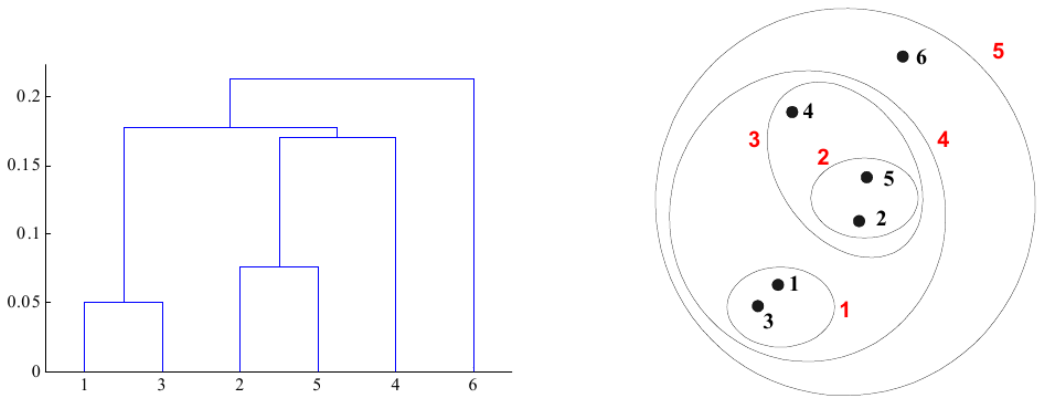


FIGURE 13.3: Exemple de classifications hiérarchiques

IBL

ID Patient	Inflam.	Fièvre	Glndes. enflées	Congest.	Mal de tête	Diag-nostic
1	oui	oui	oui	oui	oui	pharyngite
2	non	non	non	oui	oui	allergie
3	oui	oui	non	oui	non	rhume
4	oui	non	oui	non	non	pharyngite
5	non	oui	non	oui	non	rhume
6	non	non	non	oui	non	allergie
7	non	non	oui	non	non	pharyngite
8	yes	non	non	oui	oui	allergie
9	non	oui	non	oui	oui	rhume
10	oui	oui	non	oui	oui	rhume

TABLE 5: Données hypothétiques pour le diagnostic d'une maladie

- Les entrées (attributs) = symptômes : *inflammation de gorge, fièvre, glandes enflées (inflammation des glandes), congestion, mal de tête*

- La sortie : les maladies possibles : *pharyngite*, *rhume* ou une *allergie*.

IBL et la recherche simple

- **IBL** est une variante (simpliste) du clustering :

  ➥ Chaque instance est un cluster !

  ➥ Pour chaque nouvelle instance X à classifier (définir l'attribut classe) :

  - Calculer une **distance** de X par rapport à chaque instance de la table
  - Donner à la nouvelle instance la même classe que celle de l'instance **la plus proche**
  - **Ajouter** la nouvelle instance à la table

  ➔ Cette approche = **le plus proche voisin** à une instance

  ➔ Cette méthode stock les instances (plutôt que leur généralisation)

Mieux :

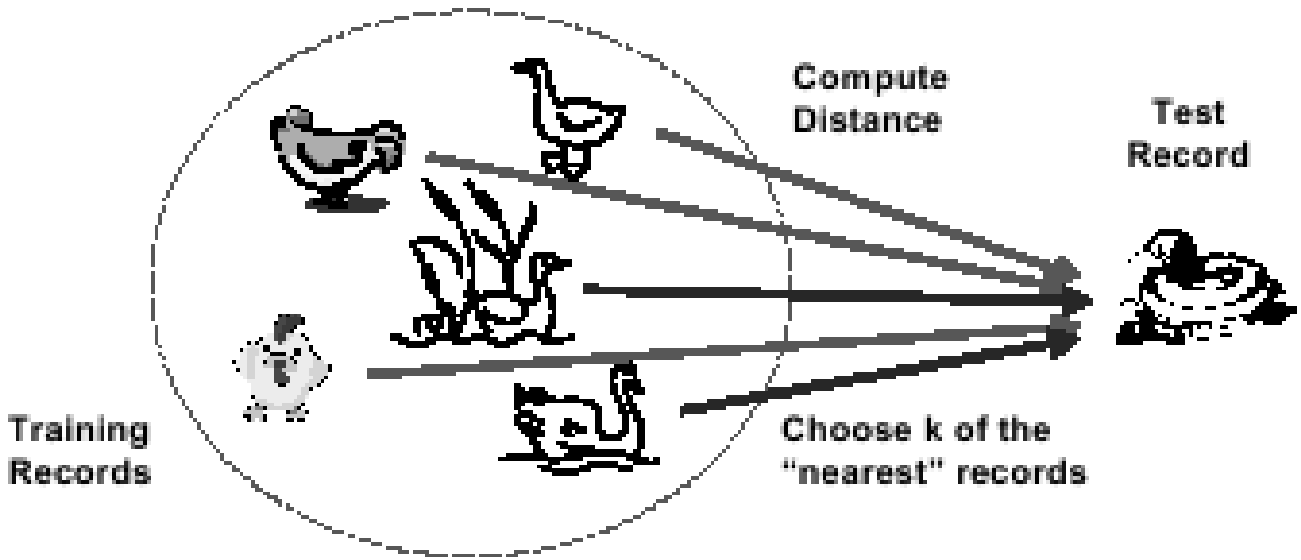
⇒ **Plus Proche Voisin** (PPV) → notion de distance

IBL et la recherche simple (suite)

KPPV ➙ K-means

⇒ On prend la classe la plus commune (majoritaire) des K voisins.

➥ Empêche le mauvais classement dû à une instance atypique de la table.



Régression

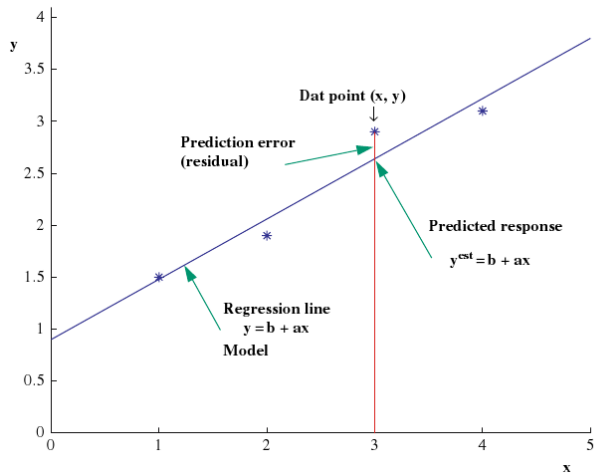
La **Rég. lin.** apprend les (k+1) pondérations w_j en **minimisant** la somme des carrés des différences entre les classes connue ($y^{(i)}$) et prédite ($\sum w_j\, a_j^{(i)}$).

➔ Avec n instances d'apprentissage, la somme des carrés des différences :

$$\sum_{i=1}^{n} \left(y^{(i)} - \sum_{j=0}^{k} w_j\, a_j^{(i)} \right)^2$$

$y^{(i)}$ = la classe connue,

$\sum w_j\, a_j^{(i)}$ = prédiction



➔ La valeur entre parenthèses = l'erreur pour la classe de la *ième instance*.

☞ La régression linéaire est ensuite adapté au problème de classification.

Régression (suite)

Régression Logistique (pour les classes binaires)

→ Dans la variante **régression logistique (linéaire)**, on utilise :

$$\boxed{log(\frac{P}{1-P}) = w_0\, a_0 + w_1\, a_1 + w_2\, a_2 + ... + w_k\, a_k}$$

avec $P = Pr[1|a_1, a_2, ..., a_k]$ = la probabilité d'être dans la classe visée.

• **La Régression Logistique fait une estimation directe de la probabilité de la classe.**

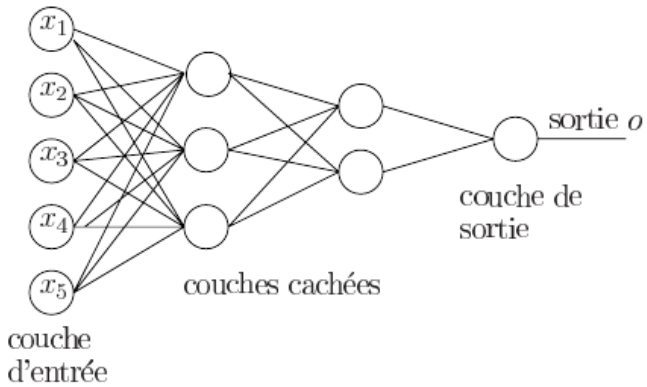
RNs & Perceptrons

• Les neurones élémentaires ont une structure simple avec une fonction d'activation (discrète, sigmoïde, etc.) .

• Dans le cas simple, il n'y a pas de retour vers une couche précédente.

• Chaque couche peut contenir de 1 à plusieurs neurones, y compris la couche de sortie.

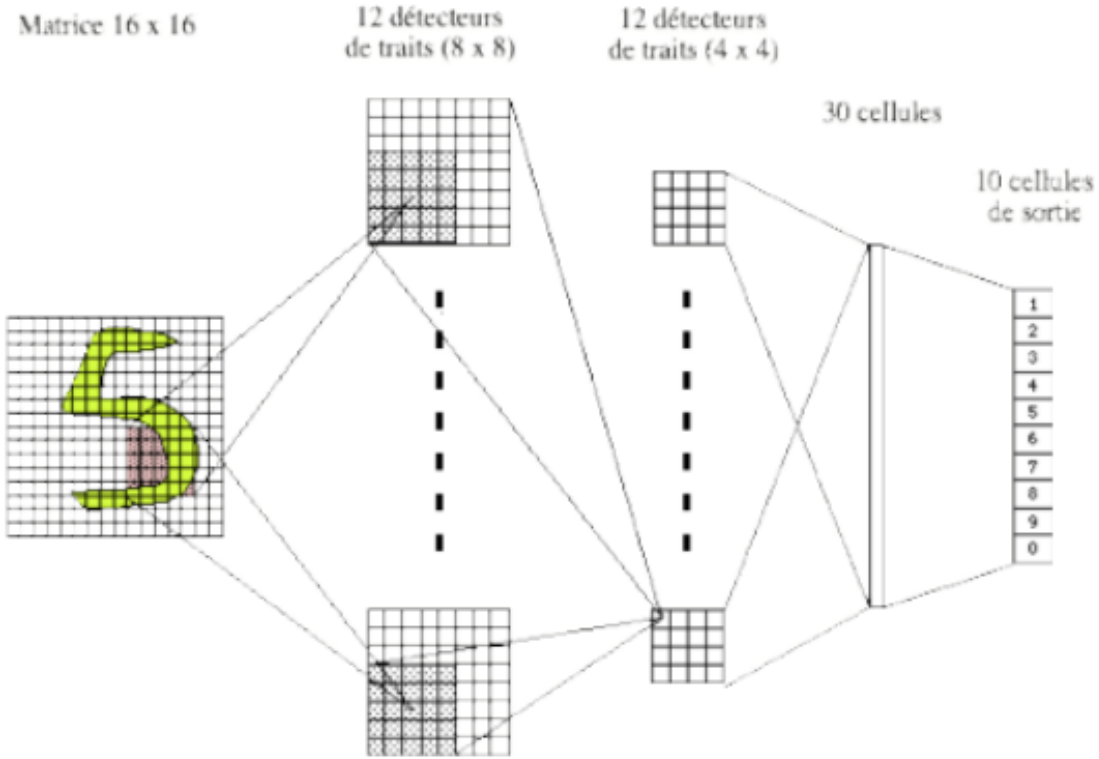
➙ C'est le problème à traiter qui définira le nombre de neurones en entrée et en sortie.



N.B. : les RNs existent depuis longtemps mais c'est dans les années 80s que l'algorithme de *rétro-propagation du gradient* [Rumelhart] a permis leur développement.

RNs & Perceptrons (suite)

Un exemple (CNN, Kernel)



Méthodes à base de noyaux

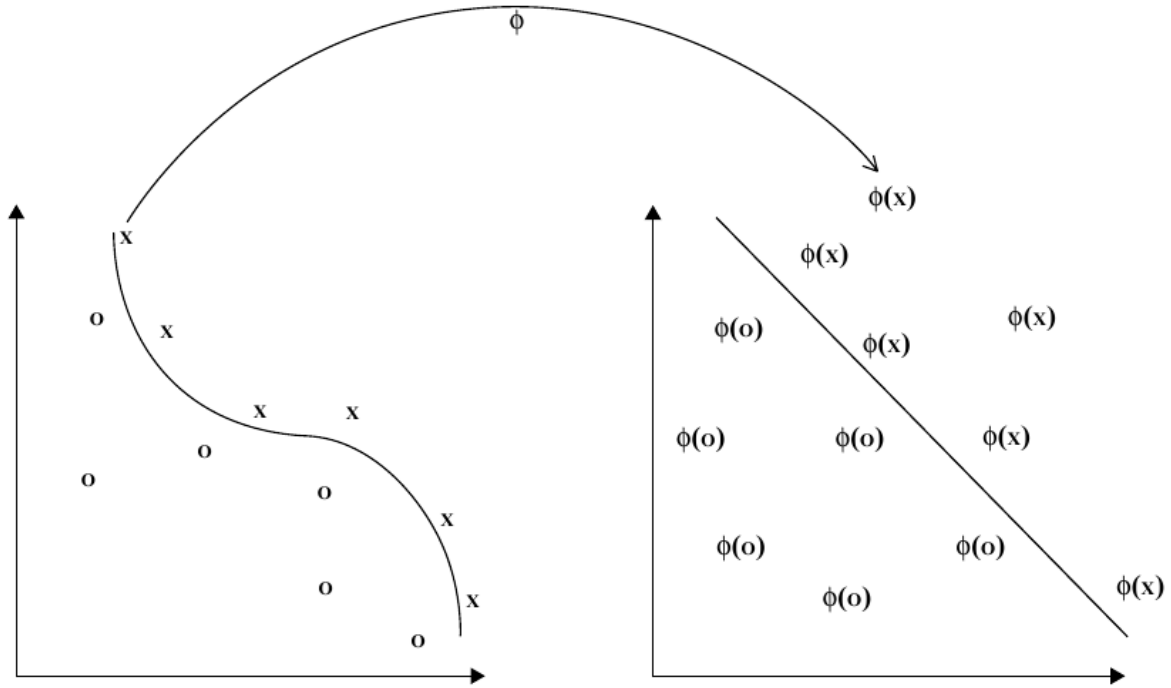


FIGURE 14.1: La fonction ϕ envoie les données dans un EdC où les motifs non-linéaires (d'origine) paraissent linéaires. Le noyau calcule ensuite des produits matriciels dans EdC directement àpd des instances d'origine.

De l'importance et qualité des données

Fouillez, vous saurez ! :

➙ la qualité des données, bruits et informations (connaissances) extraites

Exemple :

• *Crimes et chatiments en Floride-USA* (1973-78) : Paradoxe de *Yule-Simpson*

Meurtrier	**Jugement**	
	peine de mort	**autre**
Black	59	2547
Blanc	72	2185

➥ 3.2% des meurtriers blancs ont été condamnés à la peine de mort

➥ 2.3% des meurtriers blacks ont été condamnés à la peine de mort

➙ On ne peut franchement pas accuser ! ../..

De l'importance et qualité des données (suite)

- **Données de meilleure qualité (détails)** :

Victime	Meurtrier	Jugement	
		peine de mort	**autre**
Black	Black	11	2309
	Blanc	**0**	111
Blanc	Black	48	238
	Blanc	72	2074

➥ Pour les victimes Blacks, 4,7% des meurtriers Blacks sont condamnés à la peine de mort vs. **0%** pour les meurtriers Blancs

➥ Pour les victimes Blancs, **16,8%** des meurtriers Black sont condamnés à la peine de mort vs. 3,4% pour les meurtriers Blancs

☞ La qualité (détails) de ces données → des analyses et prédictions différentes.

Exemple d'application : Décision bancaire

- La décision de la banque en 2 temps :
  - 1) méthodes statistiques → acceptation/refus franc.
  - 2) puis le jugement (humain) pour le reste.
- Méthodes statistiques → une note à la demande de prêt (90% des cas) :

  ⇒ Si la note $>$ *bsup* ($<$ *binf*) → accord (refus) → Décision directe ;

  ⇒ Sinon (le reste) : on prend une décision.
- Données historiques importantes :

  $\frac{1}{2}$ des clients marginaux approuvés ont fait défaut.

  ⇒ On peut ne pas leur prêter.

  ⇒ Mais (constatation) : ce sont des clients intéressants

  ➥ des clients actifs de l'institution de crédits avec une finance volatile

  → ils empruntent souvent, finissent par payer, même avec retard.
- important : bien déterminer la tendance de la situation d'un client marginal

Exemple d'application : Décision bancaire (suite)

- **Echelle typique d'Apprentissage automatique dans ce cas réel** :
  ⇒ 1000 exemples d'apprentissage de cas marginaux + s'ils ont "finalement" payé

  ⇒ 20 attributs étudiés : age, nbr. années dans la même boîte, nbr. années à la même adresse, années avec la banque, autres cartes de crédits, ...

- Un algorithme simple de classification → 2/3 des cas MARGINAUX correctement prédits (testé avec un ensemble de cas aléatoires).

  ➥ Aide à la décision des banques + l'explication (le **pourquoi**) de la décision.

Application : screening

Satellites → images → recherche de tâches de pétrole → ..
→ catastrophe écologique.

• Les satellites radars → surveillance des côtes jour et nuit (par tout temps).
⇒ Ces tâches changent d'aspect et de taille avec le temps / les conditions de la mer.

• Mais il y a des tâches (semblables) dûes au climat local/vent très fort.
➥ Il faut des très spécialistes entrâinés pour diagnostiquer

• Système d'apprentissage automatique **supervisé** entraîné sur des exemples positifs / négatifs (fausses alertes).

• Pas de classification mais un schéma d'apprentissage directement déployé.

• Utilisation par gouvernements/individus/géographes/...

Application : Prévision de charges

- Prévision de Charge électrique (demande à venir) aussi loin que possible.
- Prévoir le max/min de la charge par heure/jour/mois/saison/année
  - ➥ Électricité non stockable → Économies sur la réserve opérationnelle,
  - ➥ planification de la maintenance, gestion de fuel...
- Comment a-t-on fait auparavant ? → un modèle sophistiqué (à la main).
- Création d'un outil de prévision de charge (*load forecasting*) automatique pour les prévisions de charges par heure avec 2 jours en avance.

  ⇒ 3 composantes : charge de base de l'année, périodicité sur l'année, effet des vacances.

Application : détection de pannes

• Prévention de pannes et maintenance dans l'industrie sur les machines électromécaniques : → Importance des coûts et le fonctionnement.

• Problèmes classiques : pièces mal alignés, perte mécaniques, comportement erroné, pompes mal balancées, ...

⇒ Habituellement : techniciens expérimentés inspectent les machines et (pré)voient les pannes

⇒ Mesure des vibrations + analyse par la transformée de Fourier.

⇒ Tâche et conditions difficiles (noisy informations)

→ il faut un expert pour diagnostiquer

→ il faut répéter tout pour chaque machine.

Application : détection de pannes (suite)

- Dans ce domaine :

  - *Systèmes Experts* : les règles de production fonctionnent bien
  - *Apprentissage automatique* quand la création de règles de production est difficile.

- Un système ML proposé :

  ✓ 600 points de défaut
  ✓ chacun avec un ensemble de mesures
  ✓ + diagnostic de l'expert (20 ans d'expérience dans le domaine).

☞ Le but du système produit : déterminer le type d'un défaut avéré ($\rightarrow$ diagnostic).

$\Rightarrow$ 1/2 des cas non satisfaisants par les méthodes classiques.

$\Rightarrow$ 1/2 restante $\rightarrow$ données pour l'Apprentissage automatique.

Application : détection de pannes (suite)

• L'expert a ajouté des connaissances (fonctions sur les attributs)

• Un algorithme d'induction → un ensemble de règles (complexes) de production.

• Les règles acceptées par l'expert + ses connaissances en mécanique.
  ⇒ 1/3 des règles produites = ce qu'il utilisait
    ➥ Il les a utilisé pour en trouver d'autres et clarifier son expertise.
• Les règles générées ont été "meilleures" que celles trouvées manuellement
    ➥ Elles ont été insérées dans un Système Expert.

• Utilisation dans le domaine de la chimie industrielle (processus très complexe).

Application : Marketing

- Un domaine très actif de Extraction de Connaissances, beaucoup de succès.
- Volume très important de données.
- L'objectif recherché : **la prédiction**,
  ➥ Important : la découverte de la structure pour prendre des décisions.
- Les applications bancaires :

  - Fidélisation des clients (par traitement/promotion particulier).
  - Les banques et les demandes de crédits.
  - comportement/insatisfaction de clients → changement de banque.
  - changement dans la vie/ville, ... → peut entrainer un changement de banque.

Application : Marketing (suite)

• Exemples d'éléments intéressants pour les banques :

- insatisfaction des clients qui font leur opérations par téléphone avec un service téléphonique lent.
- des clients qui retirent peu de cash sur leur carte sauf vers noël ou pour les frais de fin d'année.
- des clients de retour des vacances → découvert !
- des clients que l'on peut charger !

• Analogues aux compagnies de téléphonie cellulaire qui détectent de changement de téléphone ou envie de nouveaux services.

• Pour les banques : une opération générale de promo coûte très cher

➙ Détecter des motifs dans les comportements des clients

➙ Cibler les opérations pour maximiser les résultats (bénéfices) à l'aide des outils DM

Application : Market Basket analyse

- Techniques pour trouver des motifs/groupes de produits qui vont ensemble.
- Ces données sont souvent la seule source d'info sur les clients.
- Une analyse classique ne fait pas sortir la relation *bière* $\rightarrow$ *chips*.
- Par ex., on a trouvé que des clients achetaient (le jeudi) des déguisements, figures autocollantes, couches bébé + bière.
  - ➥ Semble étrange, mais des jeunes couples commencent à préparer le WE!!.
- Informations utiles :
  - ➥ rangement des rayons, limiter des promos sur un seul des articles du groupe, offrir des coupons pour l'autre quand l'un est acheté seul, ...
- Beaucoup d'intérêt à connaître les habitudes des clients.
  - $\Rightarrow$ Prolifération de discount/coupon/carte fidélité
  - $\Rightarrow$ Identification individuelle des clients (pour les cibler).
  - $\Rightarrow$ Informations beaucoup plus importantes que le prix du discount.

Application : Market Basket analyse (suite)

- **Direct Marketing** : un autre domaine

⇒ Les promos sont chères et ne sont intéressantes que si elles "marchent".

⇒ On utilise des courriers/méls directs ou l'on demande à appeler un No gratuit.

⇒ Il y a peu de réponse de la part des clients.

➥ Important : réduire au maximum le nombre des clients visés pour obtenir les bonnes réponses (éviter ceux qui ne seront pas intéressés).

➥ On peut travailler sur les données démographiques (basées sur le code zip) pour trouver des corrélations avec des clients futurs.

➥ Les habitants d'un même quartier risquent d'avoir les mêmes comportements !

D'autres exemples d'applications

Détection de fraude

- AT&T pour détecter les appels internationaux frauduleux
- Falcon fraud assessment system (développé par HNC Inc.) pour la détection sur les cartes de crédit.
- Détection de blanchissement d'argent à travers les transactions importantes par FAIS (Financial Crimes Enforcement Network AI Systems)
- Classification non supervisée par Aspect (Advanced Security for Personnal Communications) en Europe pour la détection de fraude dans les réseaux de tél mobile. Stockage de profil pour les utilisateurs, comparaison entre l'utilisation et le profil.

D'autres exemples d'applications (suite)

Santé

- Plusieurs systèmes de détection urgent de Césarienne (biblio *Mitchel*) ;
- Recherche de médicaments génériques (par Merck)
- Décodage du genome / DNA et bio informatique

Finance et commerce

- Risk Management pour les crédits (pub. Wall Street Journal). Une corrélation trouvée : ceux qui ont une 2e voiture de sport (la 1e non sport) ne représente pas plus de risque !
- Bank of America pour le marketing et promo pour les clients
- US West Communications avec 25 millions de clients, DM pour caractériser les clients (familles) et proposition de services.

D'autres exemples d'applications (suite)

- 20th Century Fox pour l'analyse des recettes pour déterminer quels acteurs/films peuvent être mieux reçus dans telle région, ...
- Définition de "customer intrinsic/actual values" pour l'étude poussée des clients dans divers type de commerces (banque, bourse, ...)

D'autres exemples d'applications (suite)

Applications Scientifiques

- Etude des rayons Gamma (rayons hors système solaire), classification : une 3e classe de ces rayons a été trouvée.
- Analyse d'image du ciel, location des volcans sur Venus, détection de tremblement de Terre (Fayyad 1996).

Sport et Jeux

- Equipe de Basket de Toronto, outil DM pour NBA pour trouver une organisation la mieux adaptée des compétitions
- Profil des joueurs de Casinos, organisation et placement des jeux, ...

D'autres exemples d'applications (suite)

Exemples industriels

- Aide à la décision en temps réel (NASA : suivi des moteurs de positionnement orbital)
- Ricoh : dépannage (assistance aux opérateurs)
- General Electrique : analyse des performances des moteurs d'avion
- Hugin : système de contrôle du véhicule sous-marin UUP (Lock Heed)

Opportunité d'application de l'EC

- Difficile de décider si l'EC est adaptée à tel ou tel problème ;
- On peut se demander :
  - ✓ Peut-on clairement définir le problème ?
  - ✓ Est-ce que de données potentiellement significatives existent ?
  - ✓ Ces données contiennent-elles des Connaissances implicites (cachées) ou est-ce simplement des données factuelles = un rapport ?
  - ✓ Est-ce que le coût du traitement est inférieur p/r aux gains (étant données des projets DM semblables) ?

Opportunité d'application de l'EC (suite)

Défis et Challenges de l'Extraction de Connaissances :

- Echelle Larges (Scalability)
- Réduction de Dimensions
- Données Complexes et hétérogènes
- Qualité de données
- La propriété et la distribution des données
- Préservation de confidentialité (anonimisation)
- Streaming de données, etc.

bibliographie

• Ce cours s'est inspiré de multiples documents sur l'Extraction de Connaissances ainsi que les outils employés (outils Mathématiques, Probabilité et Statistiques, etc.).

• Aussi, un nombre important d'articles et des rapports de recherche peuvent être consultés (*http ://citeseer.ist.psu.edu/cs*).

• A titre d'indication, les ouvrages suivants peuvent être approfondis :

Data Mining : Introductory and Advanced Topics : Margaret H. Dunham. Prentice Hall 2002.

Data Mining : Concepts, Models, Methods and Algorithms. M. Kantardzic. IEEE 2001

From Data Mining to Knowledge Discovery in Databases U. Fayyad, G. Piatetsky-Shapiro, P. Smyth 1996 & 2008 (document PDF)

Principles of Data Mining : D. Hand, H. Manila, P. Smyth. MIT Press, Cambridge, 2001

ET bien d'autres !

tabmat

1. Introduction
   - L'apprentissage
   - Fouille de données : quelques exemples indicatifs
2. Vers le Big Data
3. Quelques autres exemples
   - De nos jours : Google page Ranking
   - Facebook's News Feed
   - OKCupid Date Matching
   - NSA Data Collection
   - IBM'CRUSH
4. Recherche de motifs dans les données
   - Domaine multi disciplinaires
   - Apprentissage et Recherche de motifs
   - Induction / Déduction
   - KDD
   - Exemples d'application
5. Extraction de Connaissances
   - Missions de l'Extraction de Connaissances
   - Principales formes de concepts sorties de l'EC
6. KDD : le Processus d'Extraction
   - Qu'est-ce que les ordinateurs peuvent apprendre
7. Recherche de Concepts par Généralisation (Induction)
   - Un exemple simple
   - Expression opérationnelle des concepts
8. Aperçu d'extraction de Motifs structurels
   - Exemple d'un jeu

tabmat (suite)

- Entrées : attributs mixtes
- Prédiction : Règles d'association
- Résumé des données : lentilles de contact
- AD : Une forme importante de sortie
- Prédiction numérique : exemple des Performances

9 Modélisation Probabiliste
- Application à l'exemple jeu
- La probabilité conditionnelle de Bayes

10 Le réseau Bayésien de l'exemple météo

11 Supervisé ou Non Supervisé

12 Clustering Probabiliste

13 Modèle de Mélange (Mixture) Finie
- Pour comprendre : un exemple simple de Mélange

14 Apprentissage à base d'instances (IBL)
- IBL et la recherche simple
- Régression
- SVM & kernel based

15 Fouillez, vous saurez !
- La qualité des données

16 Quelques exemples et domaines d'applications
- Décision nécessitant un jugement : prêt bancaire
- Exemple : détection/nettoyage (screening) d'images
- Autre domaine : Prévision (de charge)
- Autre domaine : Diagnostic et prévision de pannes
- Domaine de Marketing & Ventes

tabmat (suite)

17 Quelques autres exemples concrètes d'applications

18 Opportunité de l'Extraction de Connaissances

19 Quelques références Bibliographiques