

# Outils Modernes de Conception

## Ingénierie de la Fiabilité

(Extraction de Connaissances & Fiabilité)

Ecole Centrale de Lyon - Master GI

2023-24

Cours 1

Alexander Saidi

UMR LIRIS - CNRS

Alexandre.Saidi@liris.cnrs.fr

# 0.1 Introduction

## 0.1.1 *Propos*

- Cadre : "cycle de Vie" de produits industriel
- Prévvision de la "Fiabilité" et de la "Défaillance" par des **méthodes informatiques modernes**
  - Phase *design optimisé* par l'analyse de la fiabilité des systèmes
  - Phase *exploitation* par la prédiction de la fiabilité des systèmes
- Le propos sera principalement : *observer, apprendre et prédire* la Fiabilité.
  - Relève de la *Rétro-Conception (Reverse Engineering)*

## La fiabilité de nos jours :



- On a **besoin de la fiabilité** pour répondre aux questions :
  - de sûreté, des risques, des normes et statuts, des responsabilités,
  - les garanties, les exigences des clients, la pression du marché, la concurrence,
  - management et gestion, la prévision (et prédiction), remporter des compétitions, ...

## 0.1.2 *Plan*

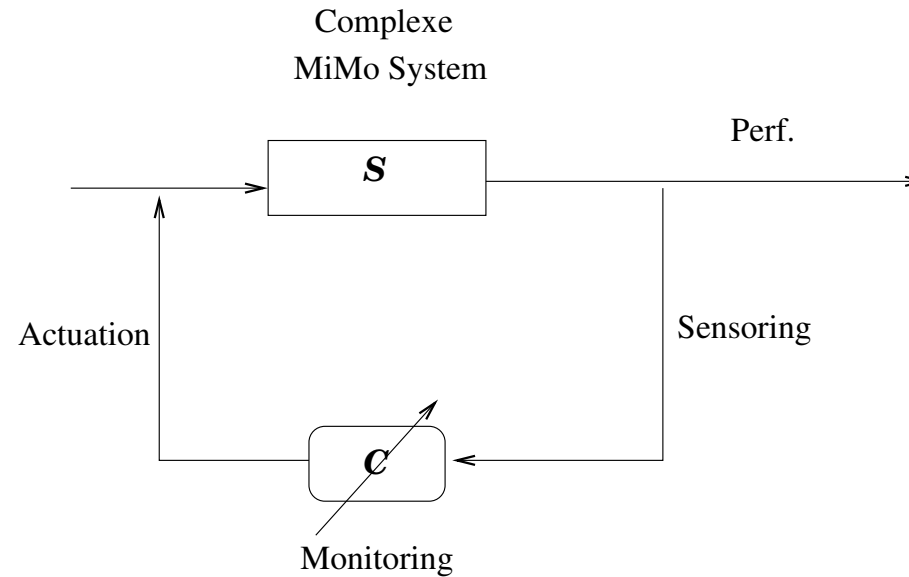
**Cadre général** : apprendre avec les données.

Deux séances de cours :

- **Séance 1** : Introduction & Aperçu des méthodes d'apprentissage Automatique (*ML*)
  - *Extraction de Connaissances* à travers des exemples simples
- **Séance 2** : Lois de défaillance et de fiabilité des dispositifs / processus complexes
  - Fiabilité et Cycle de vie de produits
  - Métriques et Distributions
  - Utilisation du cadre **Bayésien**, Réseaux de Neurones, etc.
- L'objectif (parmi d'autres) : **données + hypothèse (*a priori*) → prédiction (*a posteriori*)**



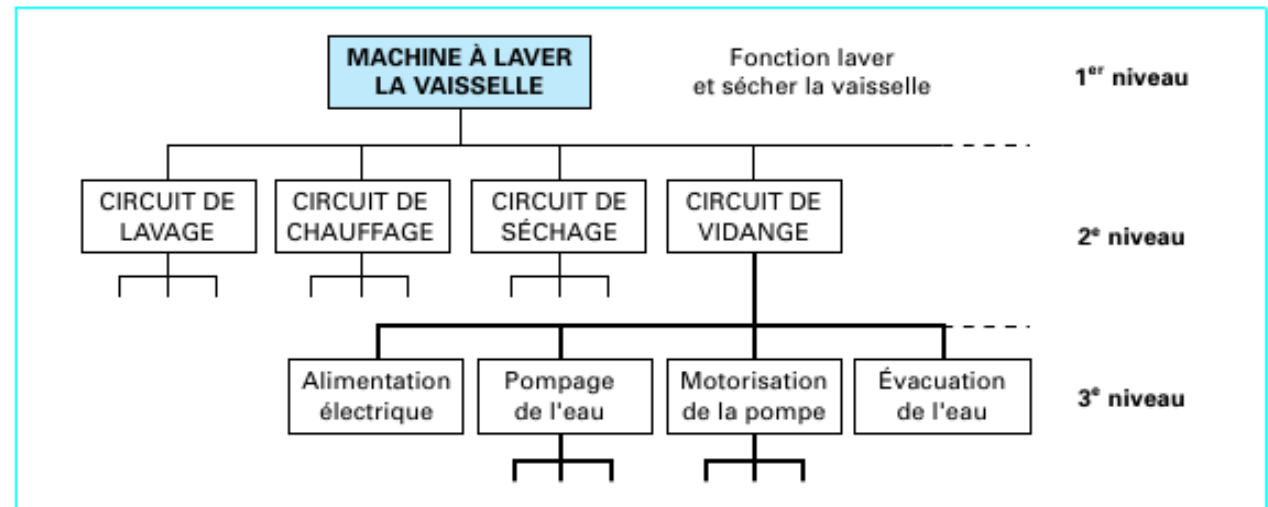
### 0.1.3 Une vue synthétique de la problématique "fiabilité"



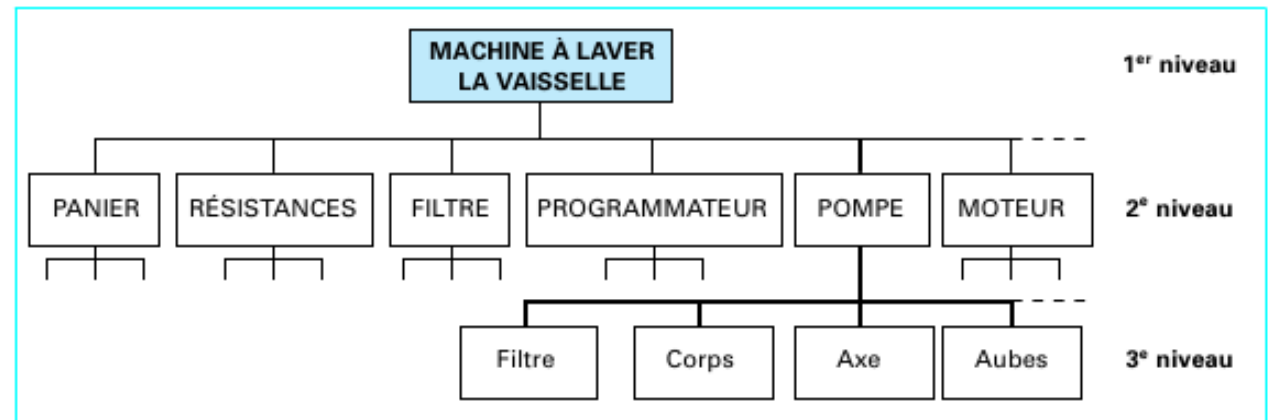
- Analyse Fiabiliste : principalement le système **S** (voir séance 2)
- Synthèse Fiabiliste : Fiabilité de la fonction *Perf.* (*plus complexe*)
  - Un système complexe comporte des structures en série / parallèle
- Exemples : ABS, Suspension Hydro-active, etc.

## Un exemple de système :

- Le calcul de la fiabilité des organes est abordable,
  - ➔ Mais celle de leur **combinaison** et des **fonctions** est plus difficile.
  - ➔ Dépend des modalités d'utilisation / de défaillances / etc.



- Parfois une modélisation propre n'est pas possible,
  - ➔ Elle peut dépendre de paramètres inconnus (Contexte, Maintenance, Modèle non immuable, ...)



## 0.2 Aperçu de la terminologie fiabiliste

- Notions de : Survie, Défaillance, Fiabilité, ...
- Diverses formulations de la durée de vie (fiabilité vs. défaillance) d'un système dans le temps :
  1. Directe et par la loi de **Défaillance**  $f(t)$  ( $\propto \simeq$  un taux de défaillance dans les cas simples)
  2. Par la fonction de **Fiabilité** (ou de **Survie**)  $R(t)$
  3. Par la fonction de distribution cumulative (CDF) de défaillance
    - fonction de **défaillance**  $F(t)$  avec  $F(t) = 1 - R(t)$
  4. Par la fonction de *Hasard* ou **taux de défaillance instantanée**  $h(t) = \frac{f(t)}{R(t)}$
- ☞ Lorsque les lois et leurs paramètres sont (supposés) connus, les calculs ont lieu directement.
  - Phase de design & conception
- ☞ Notre propos : estimation du modèle et ses paramètres à partir des observations (ReX)

## Exemples de lois

### 1. Calcul directe par une loi de défaillance $f(t)$

→ Exemple de loi **exponentielle**

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & t > 0, \lambda > 0 \\ 0 & t \leq 0 \end{cases}$$

- On dira que la variable aléatoire  $T$  a une distribution exponentielle :  $T \sim \exp(\lambda)$

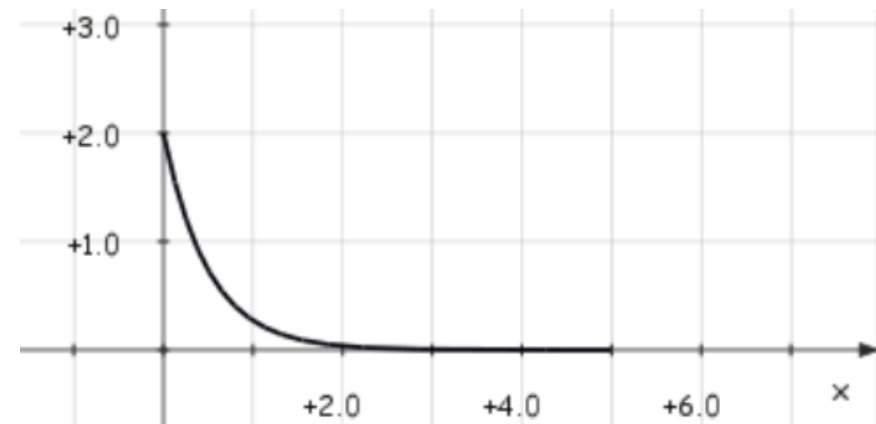
☞  $f(t)$  représente la probabilité de défaillance d'un élément à l'instant  $t$ .

→ Exemple (figure) avec la constante  $\lambda = 2$

- $f(t)$  satisfait les conditions d'une fonction de densité de probabilité (PDF) :

$$f(t) \geq 0, \forall t \quad \text{et}$$

$$\int_{-\infty}^{+\infty} f(t) dt = 1, \quad \forall t > 0$$



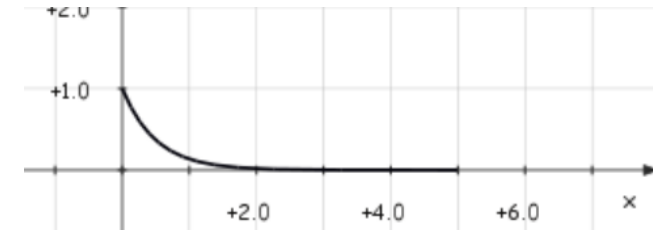
☞ Rappel :  $f(t)$  représente la probabilité de la défaillance d'un élément à l'instant  $t$ .

## 2. Calcul de la fonction de **Fiabilité** (de survie) :

$$R(t) = P(T > t) = \int_t^{+\infty} f(s) ds, \quad R(t) \in [0..1]$$

◦ Exemple pour la loi exponentielle (courbe  $\lambda = 2$ )

$$R(t) = \int_t^{+\infty} \lambda e^{-\lambda s} ds = e^{-\lambda t}$$

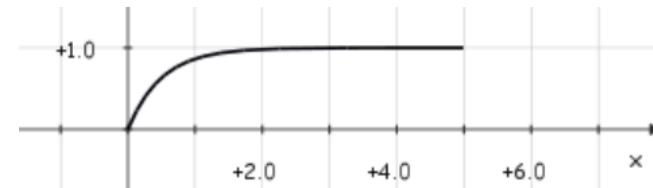


## 3. La fonction de distribution cumulative (CDF) de la défaillance $f(.)$ = fonction de **défaillance**

$$F(t) = P(T \leq t) = \int_{-\infty}^t f(s) ds$$

◦ Exemple pour la loi exponentielle (courbe  $\lambda = 2$ )

$$F(t) = P(T \leq t) = \int_0^t \lambda e^{-\lambda s} ds = 1 - e^{-\lambda t}$$



☞ On remarque  $F(t) = 1 - R(t)$

## 4. ... Et la fonction de Hasard ou **taux de défaillance instantanée** $\left( h(t) = \frac{dF(t)/dt}{R(t)} = \frac{f(t)}{R(t)} \right)$

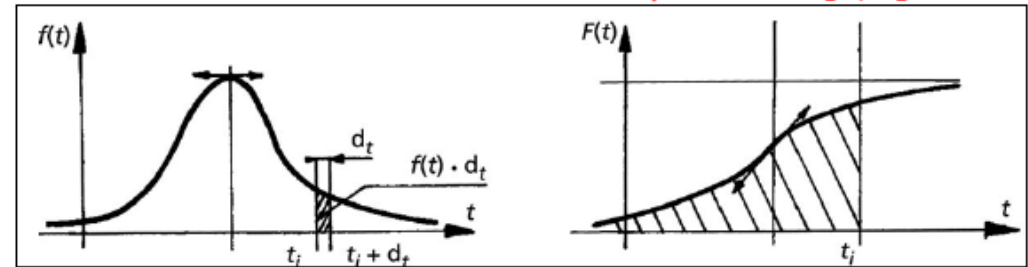
- Passage des unes aux autres (voir détails en séance 2).

**A propos de :**  $F(t) = P(T \leq t) = \int_0^t f(t) dt$

→  $F(t)$  = la probabilité de subir une défaillance à l'instant  $T$  compris entre  $[0, t]$

→ Sachant que  $f(t) = \frac{dF(t)}{dt} = -\frac{dR(t)}{dt}$

$f(t).dt = dF(t)$  = la variation de  $F(t)$  en  $dt$



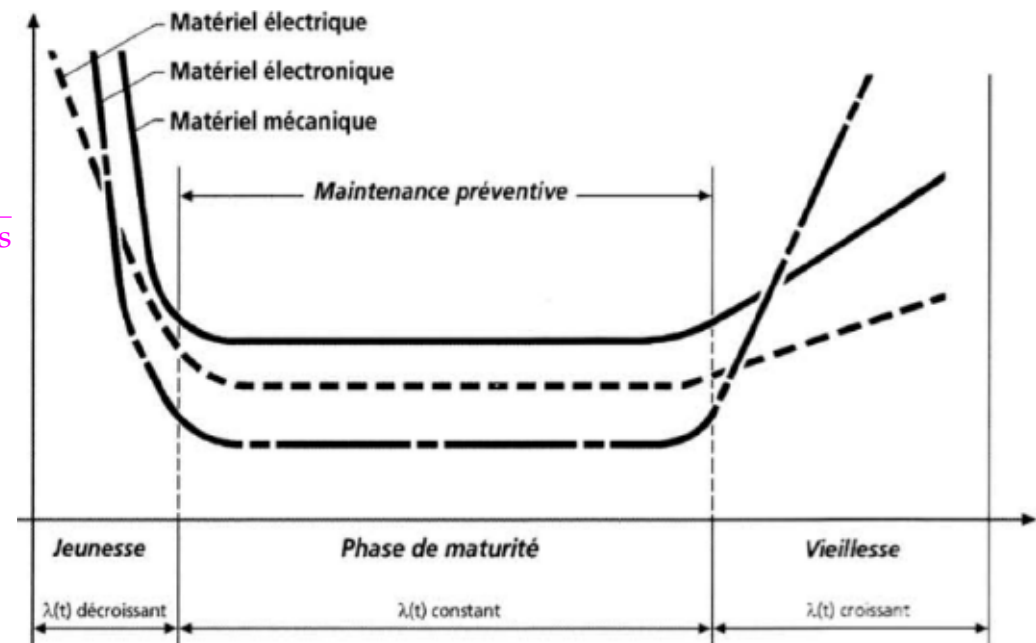
**A propos de**  $\lambda(t)$

- La variation ( $\lambda$  constante dans certains cas)

$$\lambda(t) = \frac{\text{nbr. défaillance sur un intervalle de temps}}{\text{nbr. de survivants au début de période} \times \text{intervalle de temps}}$$

représente l'évolution du cycle de vie des équipements.

- Figure : quelques courbes de défaillance (et l'évolution de  $\lambda(t)$ ) selon le type du matériel.



## 0.3 Un exemple classique de calcul de fiabilité

**Vérification de la Garanti sur une machine** (les données ReX ont permis la calcul du  $\lambda$ )

- La durée de **vie** (défaillance / fiabilité) en heures de certains composants d'une machine suit une fonction continue de densité de probabilité (*exponentielle*)  $f(x) = \lambda e^{-x\lambda}$  avec  $x \geq 0$
- **La machine contient 5 composants identiques** (tous ont la même distribution de défaillance).

- **Le fabricant veut offrir la garanti** suivante :

*Pas plus de deux des composants ne seront à remplacer pendant les 1000 premières heures d'utilisation* (les composants vieillissent indépendamment)

- **Il se demande** quelle serait la probabilité de violer cette garanti?

☞ Question de maintenance : le remplacement d'un des composants sera d'une bonne qualité permettant une (nouvelle) durée de vie de 1000h.

☞ **Notons que  $\lambda$  a été fixé à partir de données ReX.**

**Calculs :**

pour chaque composant, la probabilité d'usure dans les 1000 premières heures d'utilisation suit une loi exponentielle :  $f(x) = \lambda \exp(-x\lambda) \quad x \geq 0$ , ici  $\lambda = 1/1000$ .

$$\text{Et } F(x) = P(x \leq 1000) = \int_0^{1000} \frac{1}{1000} e^{-x/1000} dx = \left[ -e^{-x/1000} \right]_0^{1000} = 1 - e^{-1} = 0.6321$$

→ Il y a "2/3 chance" qu'un composant "tombe en panne" avant 1000h.

Indicatif : si  $\lambda = 1/10000$ , alors cette probabilité sera de l'ordre de  $\frac{1}{10}$ .

• **La garanti proposée** : savoir si plus de 2 parmi 5 des composants défailiront : 3, 4 ou 5.

◦ Il faudra calculer  $P(x = 3) + P(x = 4) + P(x = 5)$

◦ Loi binomiale (où la probabilité de défaillance de chaque composant = 0.6321) :

$$\frac{5!}{3!2!}(0.6321)^3(0.3679)^2 + \frac{5!}{4!1!}(0.6321)^4(0.3679)^1 + \frac{5!}{5!0!}(0.6321)^5(0.3679)^0 = 0.736$$

→ La garanti proposée a **une probabilité de 0.736 d'être violée** avant 1000 heures.

• **Que peut faire le fabricant ?**



## 0.4 Fiabilité et Fouille de données

- Les principes sous-jacents dans un phénomène physique ne sont pas toujours bien connus,
- Parfois le développement d'un modèle mathématique immuable est trop complexe.
- **Une solution** : à partir de Rex, chercher des liens significatifs entre les variables du système,
  - ➔ Trouver des dépendances importantes entre les entrées et les sorties du système.
- Les données deviennent alors le nerf de la guerre : les données riches d'informations sont recherchées dans tous les domaines de science et d'ingénierie (rareté, secrets, ...).
- **Autre élément important** : savoir extraire des connaissances utiles à partir de ces données.
  - ➔ *Data Mining* : le processus (souvent itératif) de recherche d'information nouvelle, inattendue, intéressante et non triviale à partir des (grands volumes de) données.
  - ➔ Les modèles obtenus permettent la **Prédiction** et/ou la **Description**

- **Prédiction ou explication** : variables *Explicatives* et *Cibles*.

→ On peut aussi découvrir des "variables cachées" (*concepts latents*).

- **Le Data Mining fait parler les données** :

*Accès aux connaissances (knowledge) à partir des données (Data).*

- Permet d'accéder aux données et aux connaissances par **analyse + présentation**.
  - Les résultats peuvent être stockés dans un *entrepôts de données (Data Warehouse)*
- **Domaine multi disciplinaires** : techniques statistiques + apprentissage automatique + analyse de données + techniques d'IA + ...
  - Application dans le domaine du **contrôle** en ingénierie de systèmes et en processus industriels.

- **Processus du Data Mining** : observer les couples entrées-sortie pour déterminer un modèle mathématique (appelé **identification de systèmes**) pour prédire le comportement du système et expliquer les interactions et les relations entre les variables du système.

- **Identification de la structure** : découverte du modèle (expertises, traitements)

*les sorties  $y = f(u = \text{vecteur des entrees}, \theta = \text{vecteur des parametres})$*

- Lors de l'identification de la structure, on procède à l'identification des paramètres en appliquant des techniques d'optimisation pour découvrir  $y^* = f(u, \theta^*)$  et décrire le système.

- Le Data Mining se nourrit de l'analyse de données, des techniques statistiques pour procéder à l'**inférence de connaissances** par des méthodes (éventuellement itératives) statistiques / algorithmiques + IA.

## 0.5 Introduction à l'EC

- Data-Science à la mode

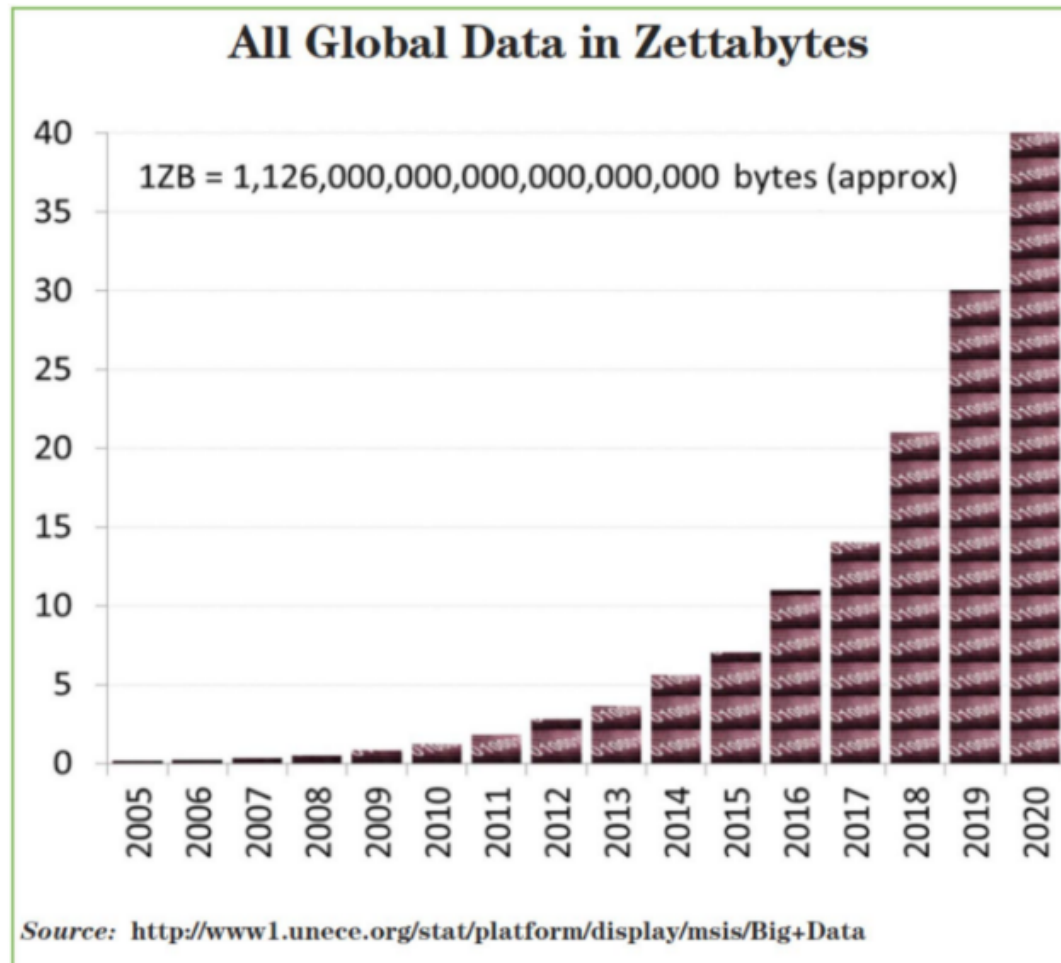
→ Qu'est-ce ?

*Extraction de Connaissances, KD, ML, ...*

- Raisons de ce boum ? :

- Croissance de la puissance de calcul des ordinateurs
- Besoin accru d'analyse de données techniques, économiques, sociales, ...
- Croissance d'utilisation de (grosses) BDs par les ordinateurs
- Besoin d'appuyer le *prévisionnel* par le *prédictif* :
  - P. Ex. Fiabilité prévisionnelle vs. prédictive via le Rex (retour d'expérience)
- Besoin de connaissances **a priori** vs. **a posteriori**
- ....

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



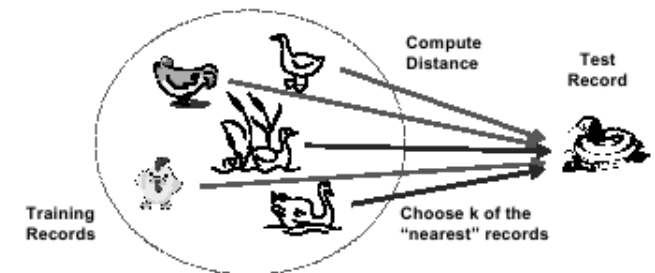
**Constat :**

- Quantité et disponibilité de "Data" dans tous domaines
  - En général, les données brutes (+ bruits !)
  - Les données de qualité sont plus **rares**
  - Les données de qualité contiennent potentiellement des connaissances
- L'exploitation des données (cf. motifs pertinents) est un atout
  - ➔ Données → Connaissances → Décisions
- Une mine d'argent pour qui peut **les avoir** et **les exploiter** (réellement!)
- ☞ Méfiez-vous des outils "presse-boutons"!
- Domaines d'application concernés : Tous!  
économie, éducation, finances, commerces, Industrie, santé, jeux, vie courante, ...

- Les questions auxquelles M.L. peut répondre (vs. une BDR ?) :
  1. Évaluation des risques dans les investissements
  2. Prédiction des résultats des campagnes de pub.
  3. Détection de fraude (tél, banque, impôts, aides sociales, sécu, ...)
  4. Prédiction des changement de profils / comportement des clients
  5. Prédiction des préférences des clients (attrait pour tel ou tel produit)
  6. Prédiction des ventes, revenus, charge, etc.
  7. Prédiction des résultats d'étudiants (évolution des promotions d'élèves)
  8. Prédiction de la criminalité, probation, ...
  9. Prédiction des maladies qu'un individu pourrait développer (génétique)
  10. Traitement / Résumé de données massives,
  11. Compréhension de texte, analyse d'opinions / sentiments, Chatbots
  12. ...

## 0.6 Apprendre des concepts

- **L'apprentissage** = méthode pratique de définition de **concepts**.
- L'humain (enfant) utilise des instances de concepts pour se représenter le monde :
  - ➔ animaux, plantes, homme, femme, jouet, ...
- On apprend des instances particulières → On choisit des attributs
- On forme des modèles de classification
- On utilise ces modèles pour identifier des objets similaires (par **analogie**).
- Deviner la suite : 1,2,3,4,5... 1,2,3,5,8,13,...  
 1,2,3,5,7,11,13,... ou (un cas de déduction) 5? 5? 5=6  
 ➔ Les données (chiffres) contiennent des connaissances



Les machines ne peuvent pas (encore) tout apprendre.



## 0.7 Les données : Big Data

## Chaque minute, on génère :

- 208K personnes en conf Zoom
- 52K personnes en conf MS. Teams
- 147K photos chargées sur Facebook
- 156K messages partagés sur Facebook
- 41,6M messages Whatsapp
- 1,4M appels vidéos
- 479K contenus publiés sur Reddit
- 347K 'histoires' sur Instagram
- 70K candidatures (jobs) sur LinkedIn
- 500 heures vidéo sur Youtube
- 139K pubs cliquées sur Instagram
- 1M\$ dépensés on-line (dont 2.8K\$ via un mobile)

• • •



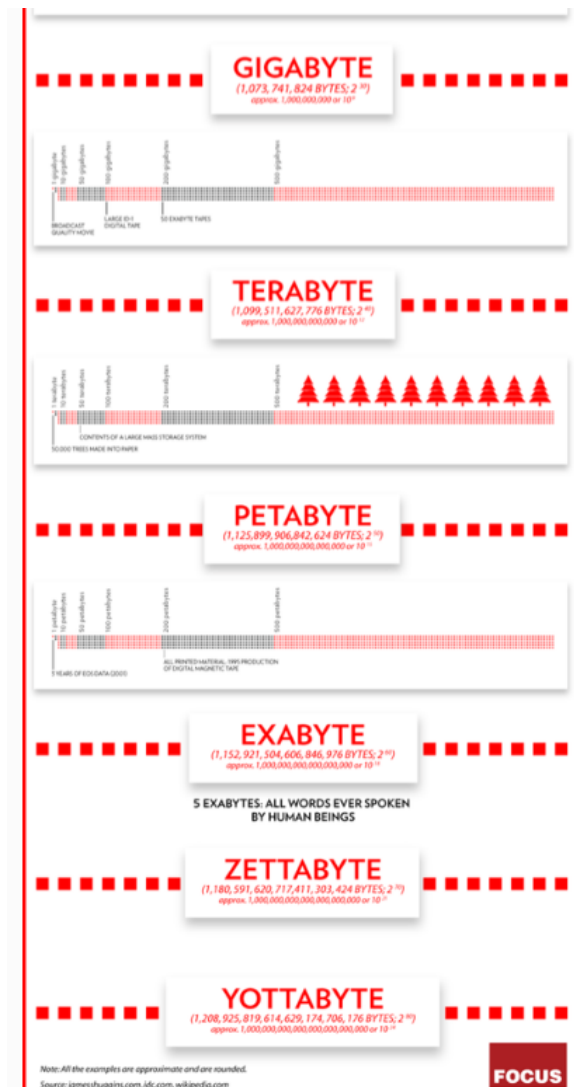
<https://www.mobilemarketingreads.com/>

## Une idée de la taille des données

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

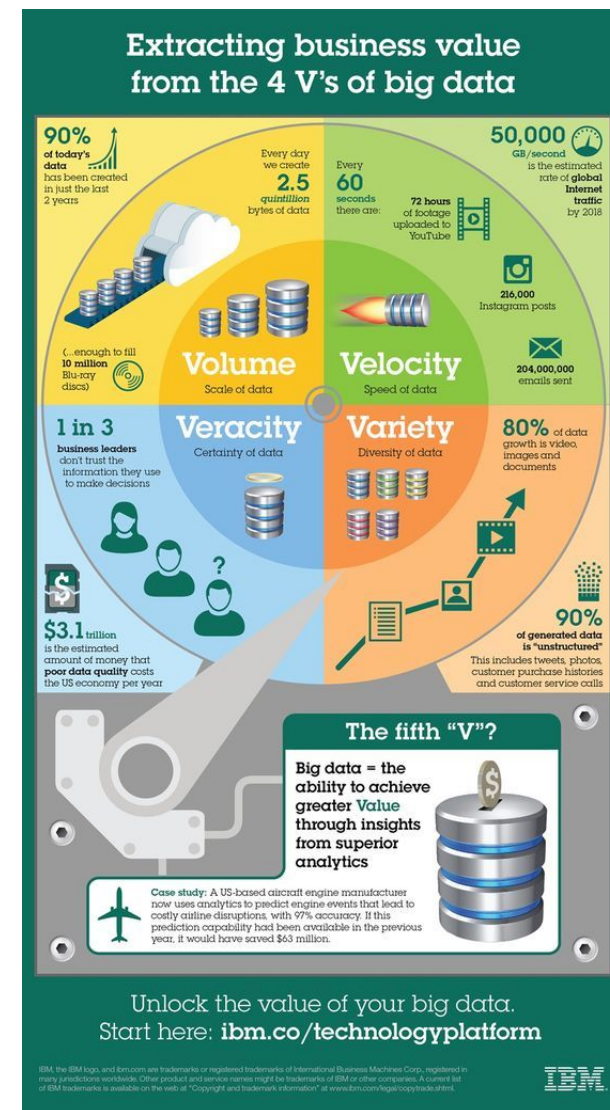
→ Il se double tous les 2 ans !

- GigaByte* :  $10^9$  octets = 1000MO
  - TeraByte* :  $10^{12}$  octets = 1000GO
  - PetaByte* :  $10^{15}$  octets = 1000tera
  - ExaBytes* :  $10^{18}$  octets
  - ZettaBytes* :  $10^{21}$  ou  $2^{70}$  octets
  - YottaByte* ou *Yobibyte* :  $10^{24}$  oct.
- L'humanité n'a prononcé que 5 *ExaBytes* de mots (depuis l'origine!)



## A propos du Big Data : les (3+1+1) "V"s du Big data

- Volume,
  - Variété,
  - Vélocité
- ☞ Un 4e V (Véracité) est souvent proposé.
- ☞ Le 5e V (\$\$\$\$) est sous-jacent.
- On estime que seul 10% des données mondiales est structuré
    - ➔ (présenté dans une BD)
  - Le 90% restant : données non structurées (Web en particulier).



### 0.7.1 D'où viennent les données

L'homme, la nature (et l'espace) sont des sources de données du big-big-data





- Les domaines récents et significatifs qui viennent s'ajouter aux domaines "classiques" :
  - **Biologie & médicale,**
  - **Nature** et espace,
  - **Réseaux sociaux** (par la taille des données),
  - **IoT, IoE,**
  - **Automatisation** (industrie, conduite automatique en ville),
  - **Reconnaissance** de toutes sortes (cf. la Chine), ...

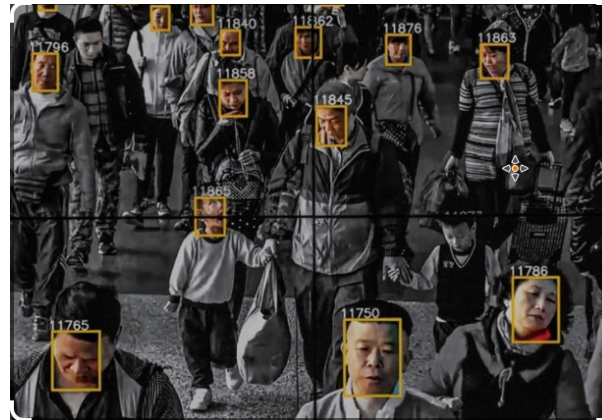
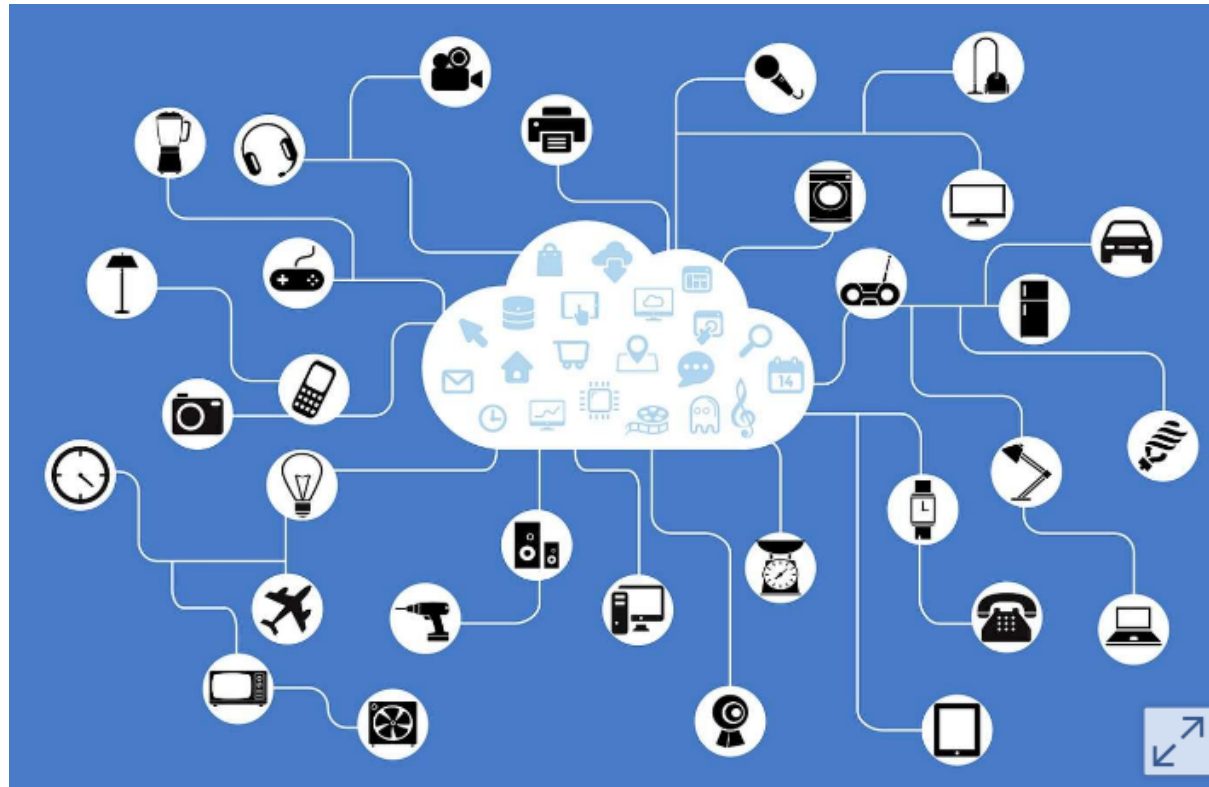


FIGURE 1 – (<https://www.telepro.be/>)

Dernier arrivée : **IoT** ( Thanks to <https://www.futura-sciences.com/>)

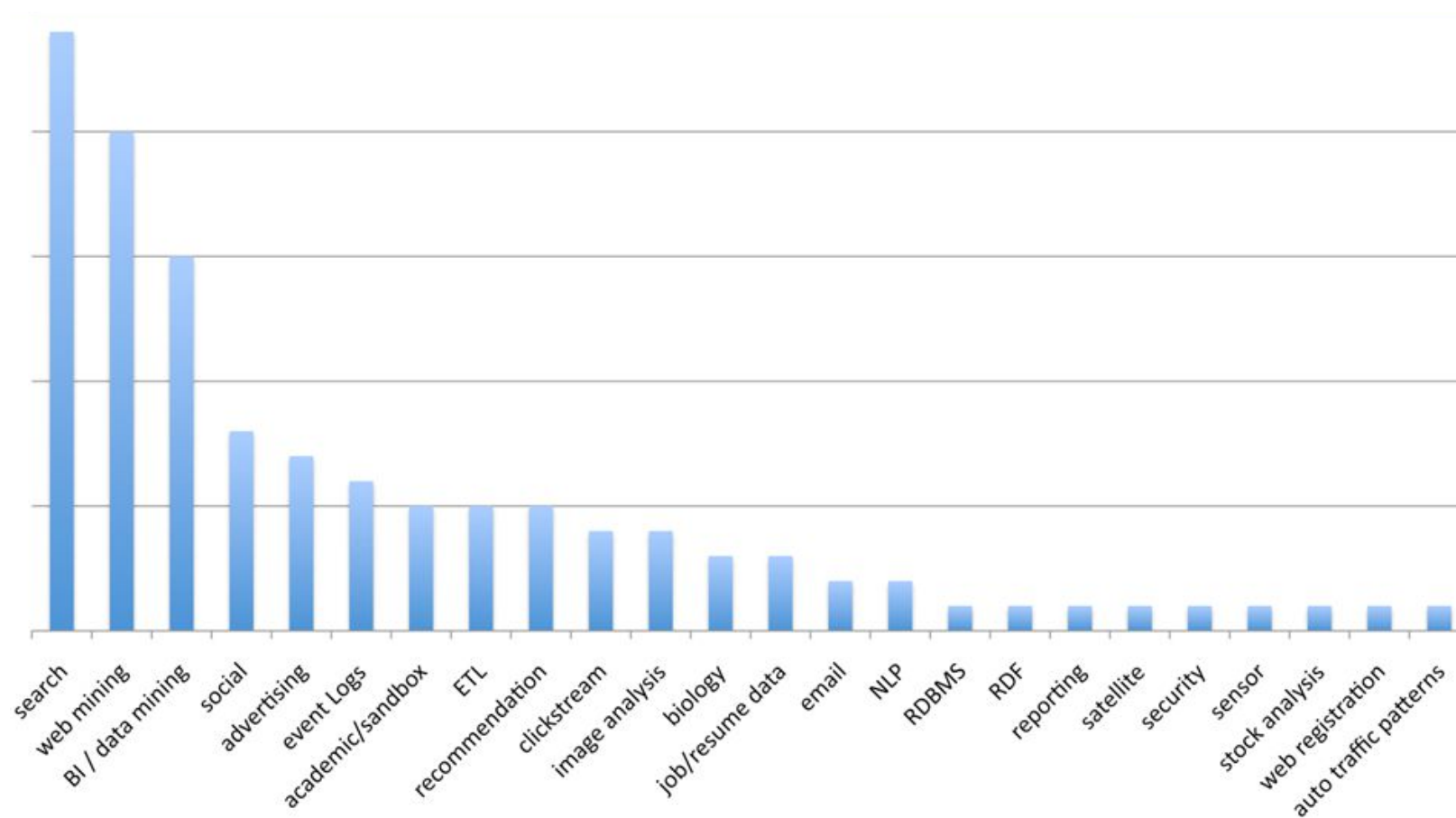


## 0.8 Quelques métiers

### DM & Big-Data : les métiers *Data Scientist* les plus en vogue

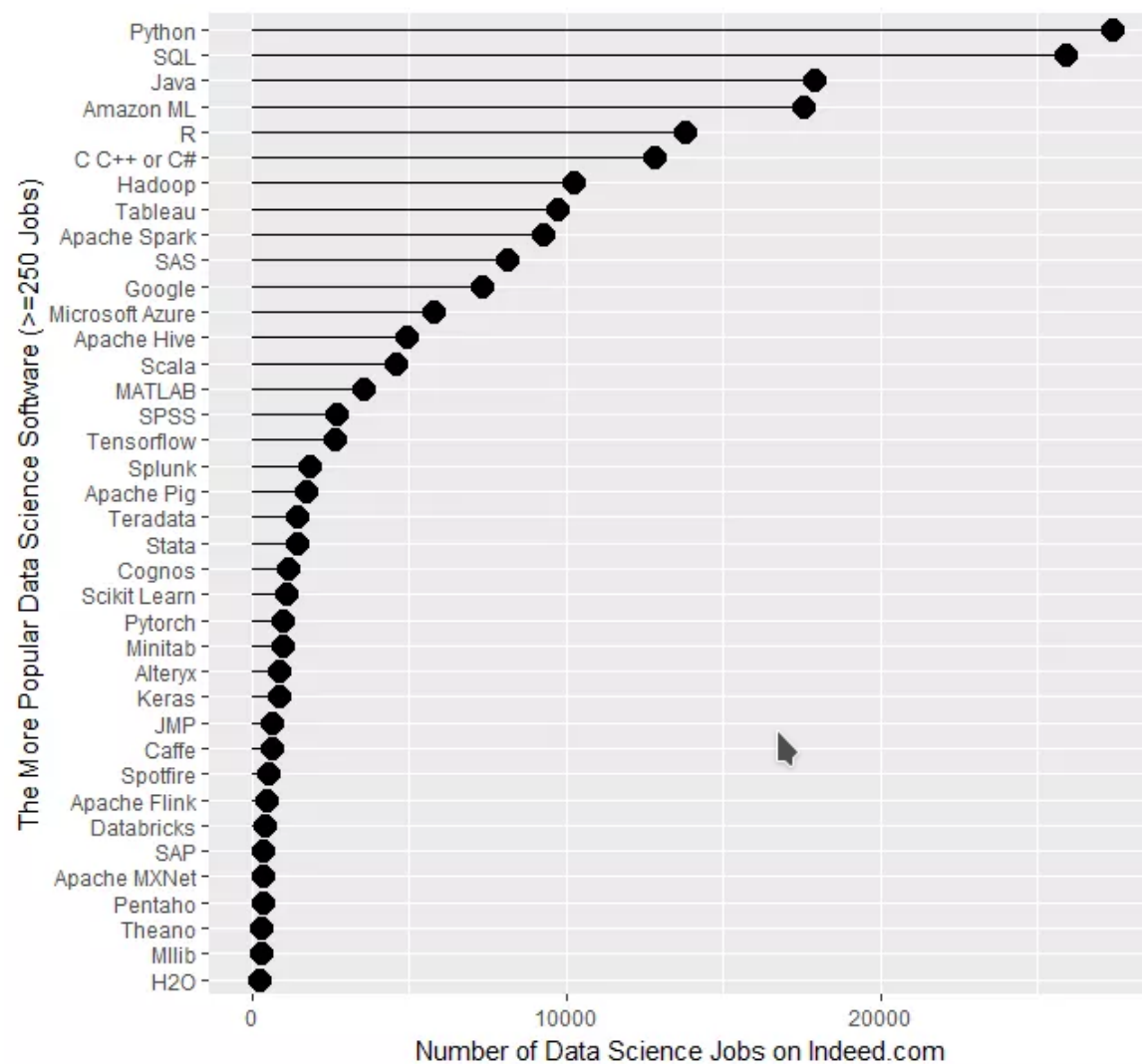
1. **Architecte Big Data** : s'occupe des données (phase pré-ML)
  2. **Ingénieur Big Data** : gestionnaire de données à large échelle
  3. **Ingénieur Data Scientist** : Data Analyste
  4. **Global data analytics** : utilise la *Data Visualization*, ...
- Selon un sondage, 61% des entreprises sont persuadées qu'elles doivent se focaliser davantage sur l'analyse de données pour ne pas être dépassées par la concurrence.
  - Le **salaire** annuel d'un *Data Scientist* en début de carrière est estimé entre 65000 et 130000 euros (le site *Datajobs*).

## Répartition par secteurs d'application (Corvelle Consulting)



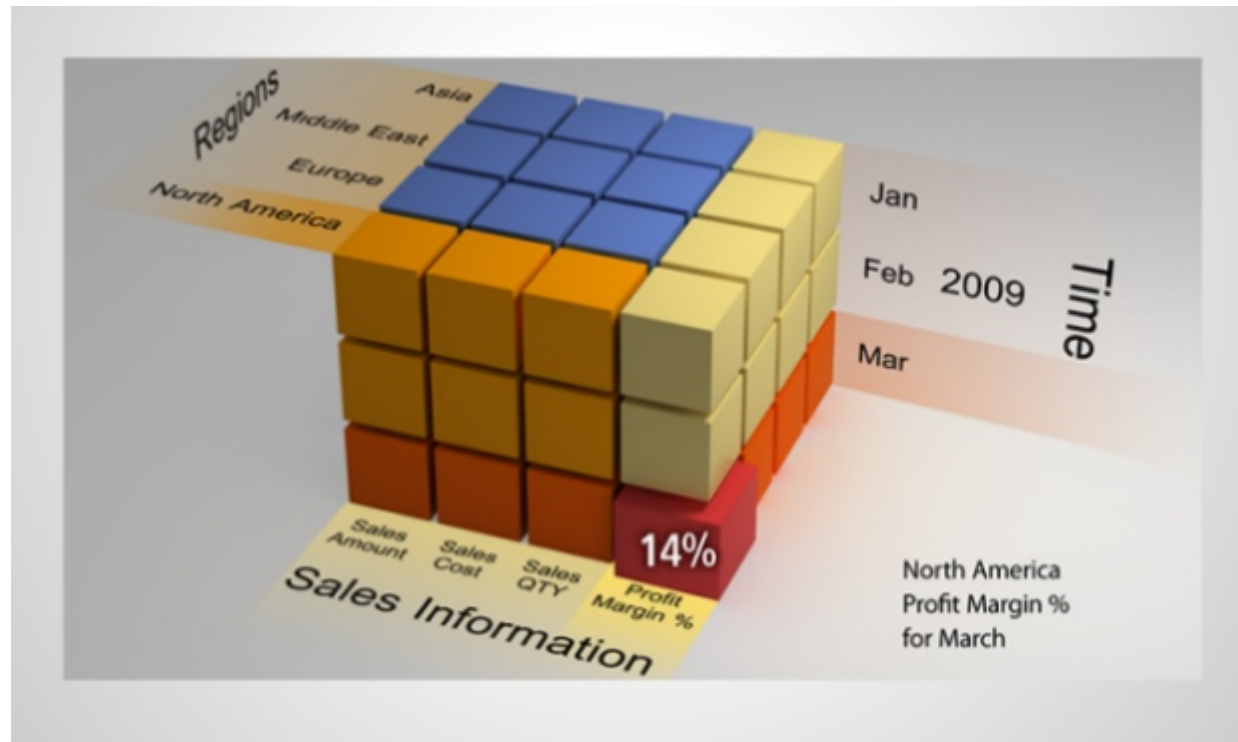


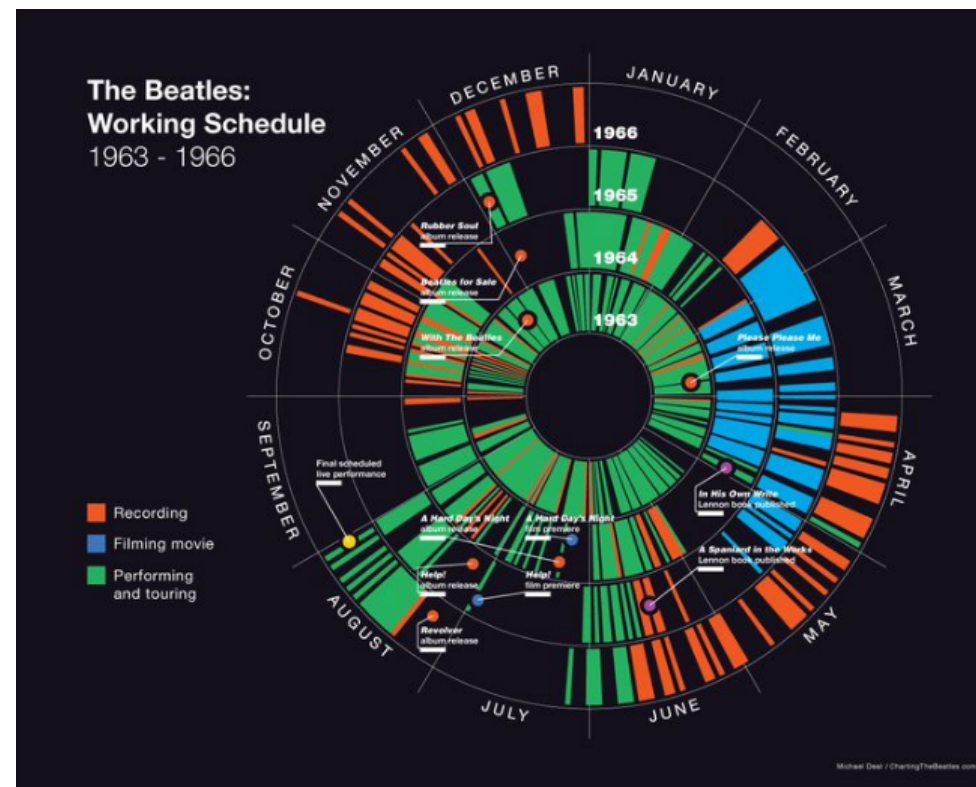
## Jobs pour Data Scientist / langage (sur *Indeed.com*)



## 0.9 Visualisation

☞ Plus récent (boosté par le Big-Data)





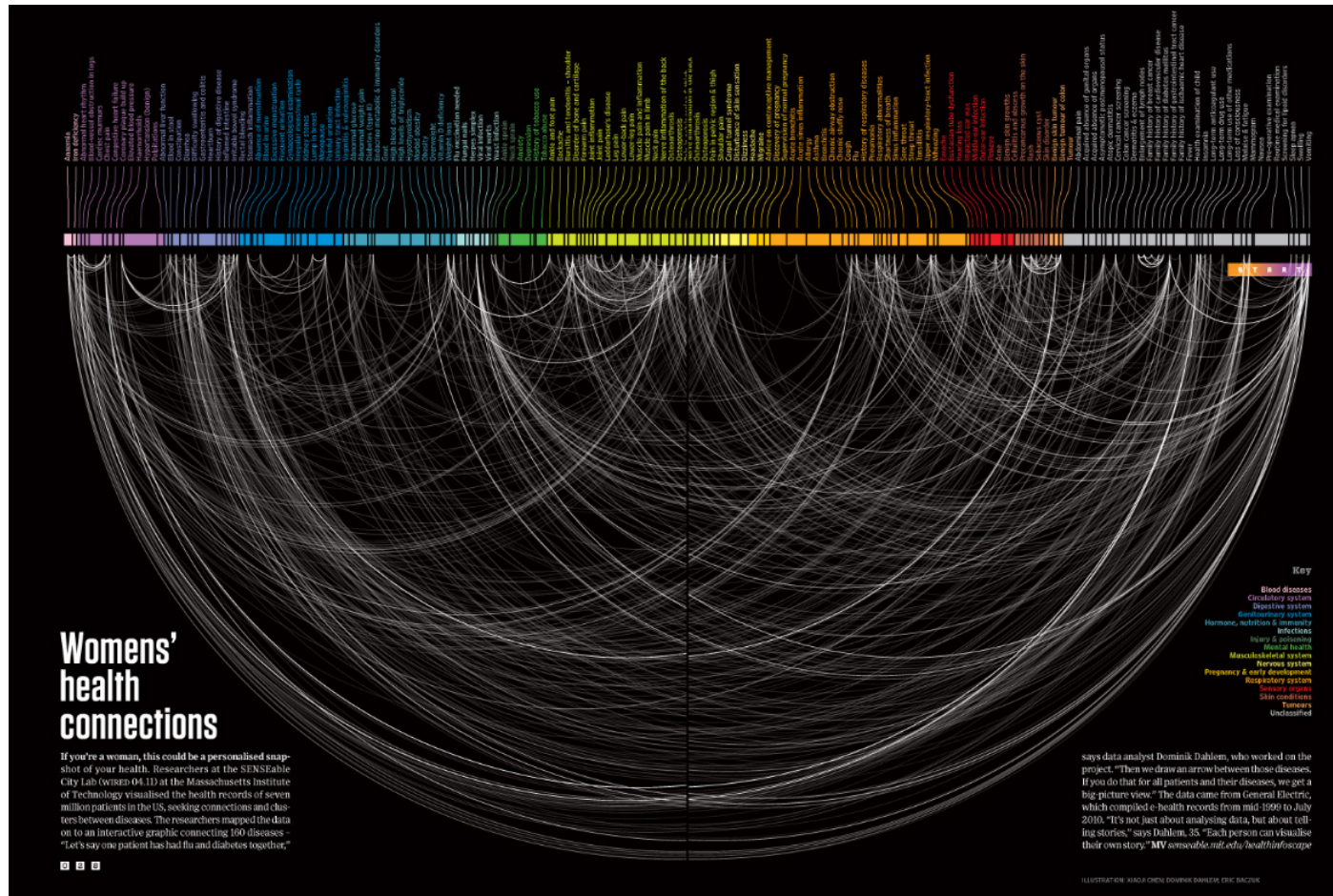


ILLUSTRATION: MARIO CHEN; COURTESY DAHLEM, ERIC BACCUX

- Cas de corpus textuel :



## 0.10 Fouille de données : quelques exemples intuitifs

**Exemple :** *Quel rapport ?*

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

**Exemple :** *Let's shake it !*

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

**Exemple :** *The big brother !*

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

## 0.11 Des données brutes à la Connaissance (information)

- **Pourquoi** : traiter et analyser de grandes quantités de données → connaissances
  - ↳ Données (BDs) économiques, médicales, industrielles, scientifiques, vie, ...
- En matière de **fiabilité**, les données REX très importantes (parfois rares).
- **Techniques** : *Extraction des connaissances* (KD)
  - ↳ Statistiques (*observation* → *loi*) & Algorithmiques (modèles)
    - + Intelligence Artificielle (et la logique) pour manipuler les connaissances → **Décisions**
  - ↳ Convergence des techniques Stat / IA / Algo.
- **Exemples d'application** : contrôle de véhicules autonomes, modélisation des risques opérationnels, analyse et localisation des gènes, Basket Analysis, Prédiction, météo, finances, jeux ...

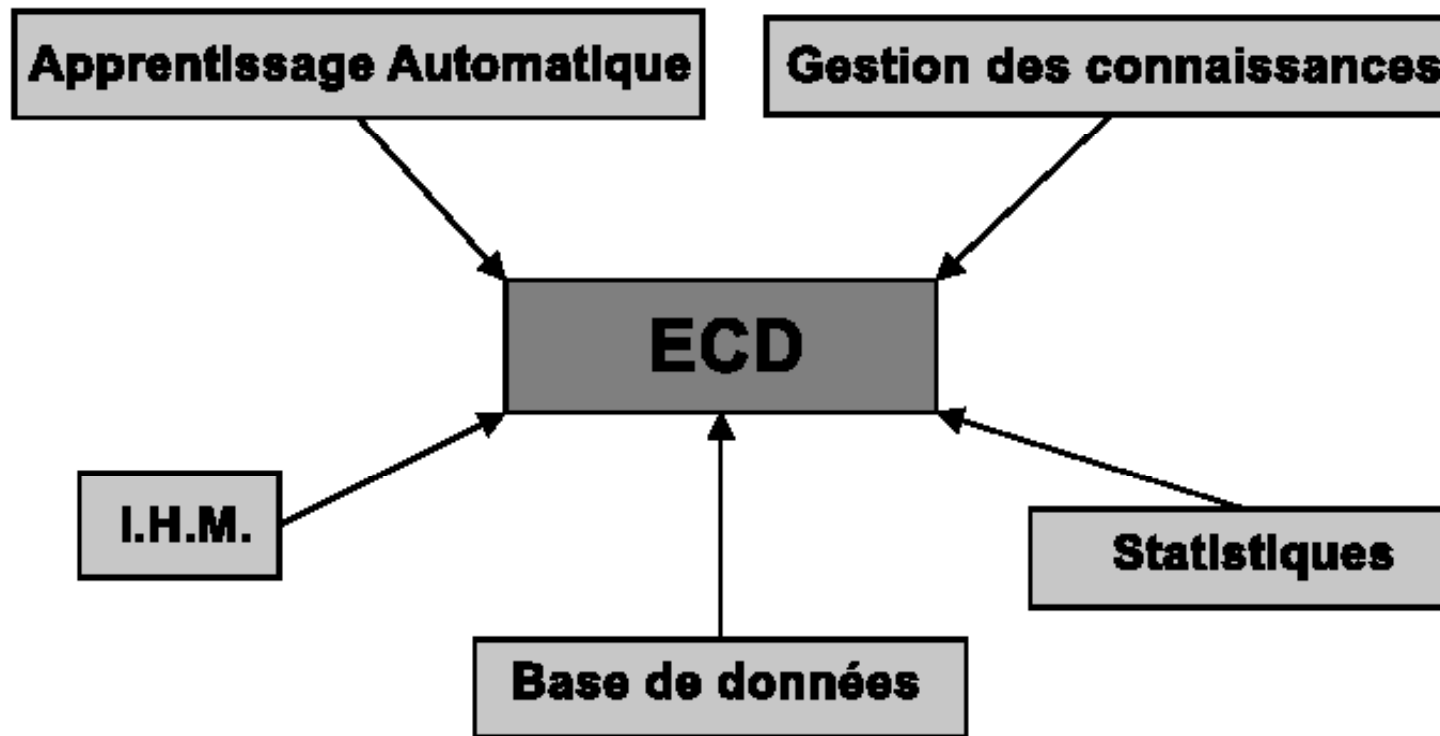


## 0.12 Quelques exemples d'application (Real World)

- Détection/nettoyage (screening) d'images
  - ➔ Détection automatique de catastrophes marines (tâches sur images)
- Prévion de charge (données saisonnières, rare, ...)
- Diagnostic et prévision de pannes
- Régulation d'embarquement dans les aéroports.
- Domaine de Marketing & Ventes
- Finances, banques
- Météo ...
- Résultats (dit aussi la *sortie* ou le *modèle*) :
  - un Arbre de décision, une équation de régression, hyperplan de séparation, etc.
- ➔ Voir plus loin pour le développement de ces exemples.



## 0.13 Extraction de Connaissances : Domaine multi disciplinaire



- Dans l'idéal, un apprentissage se fait à base d'exemples **positifs** et **négatifs**.

*La construction du modèle est **inductive**, son application est **déductive**.*

## 0.14 Recherche de motifs dans les données

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

→ l'homme a toujours recherché des motifs (patterns).

→ Ce qui est nouveau : **Qté données, Techniques et moyens**

1. Techniques : **combinaison des méthodes** statistiques, algorithmique et IA.

- Modélisation et Généralisation (vs. vérification d'hypothèses)

- S'ajoutent aux méthodes statistiques, on a :

Réseaux de Neurones, inférence et réseaux Bayesiens , algorithmes Génétiques, méthodes algorithmiques pures (ID3, C45, ...), extraction de règles, ...

## 0.15 Objectifs de l'Extraction de Connaissances

- La tâche de l'Extraction de Connaissances est divisée en 2 catégories :
  - **Predictive** : 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.
- **ECD** : Apprentissage de deux manières principales :
  - Apprentissage **supervisé** = création de modèles en formant des *définitions de concepts* à partir de données contenant des classes prédéfinies.
  - Apprentissage **non supervisé** = création de modèles  $\tilde{A}$  partir de de données sans l'aide de classes prédéfinies
  - Apprentissage **semi-supervisé**, par **Réenforcement**, ...

### 0.15.1 *Induction / Déduction*

- La 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**.
- ☞ Importance des données d'apprentissage **négatives** (mais rare).

## 0.15.2 Ex. de démarche Déductive / Inductive

### Exemple naïf de l'importance d'ex. négatifs :

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

→ **Induction** du concept "*oiseau*" : les oiseaux ont des ailes et savent voler

### Utilisation par Déduction :

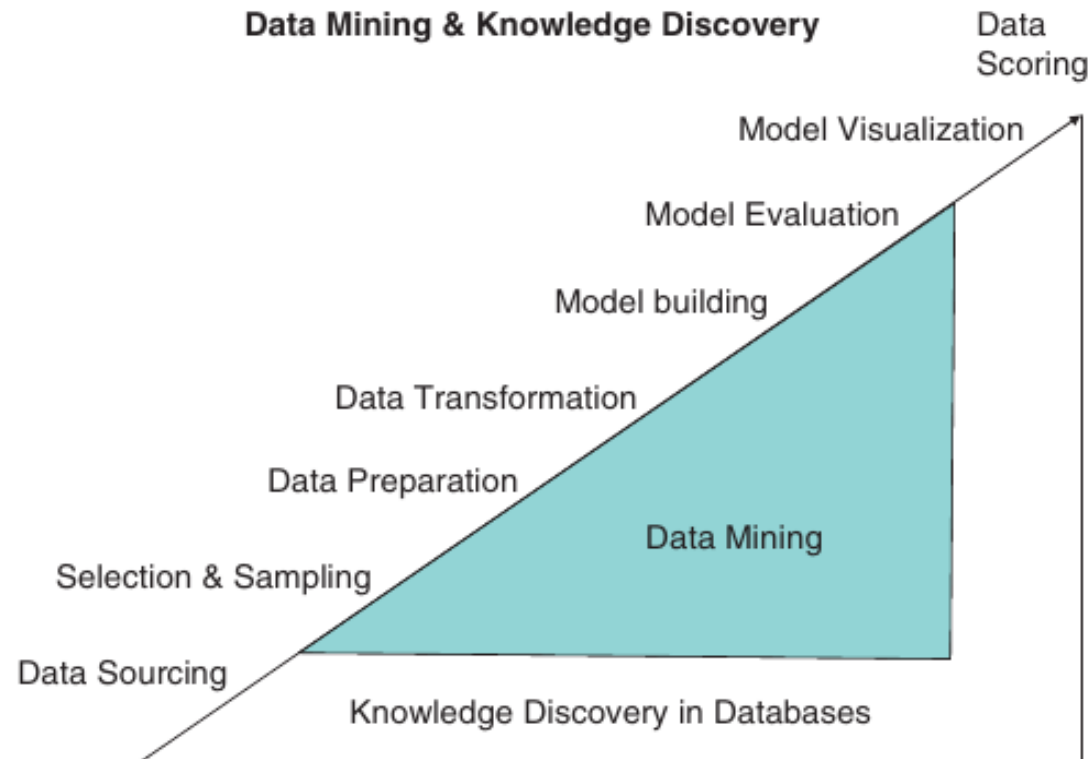
- Les merles sont des *oiseaux*.
  - **Par déduction** : les merles ont des ailes et savent voler
- Les Faucons Pèlerins ont des ailes et savent voler
  - **Par induction** ce sont des oiseaux

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

→ Erreur? *Drift* (changement du domaine?)

## 0.16 Analyses et modélisations en DM

Le processus KD / KDD :



- **Analyse exploratoire des données** : visuel / 0-R / 1-R / ..
  - **Modélisation Descriptive** :
    - Détermination d'une *distribution* de probabilité générale
    - Description des *relations* entre les variables (par les modèles)
    - Partitionnement des données en *clusters* (groupes)  
(groupes "naturels" inconnus d'avance) ou par *segmentation*.
  - **Modélisation prédictive : classification et régression**
    - prédire la valeur d'une variable en fonction d'autres.
- La découverte règles** est une variante :
- La création de règles d'association / de classification.
- **Extraction par contenu (& IBL)** :
    - Recherche de "motifs" / "instance" similaires à la requête

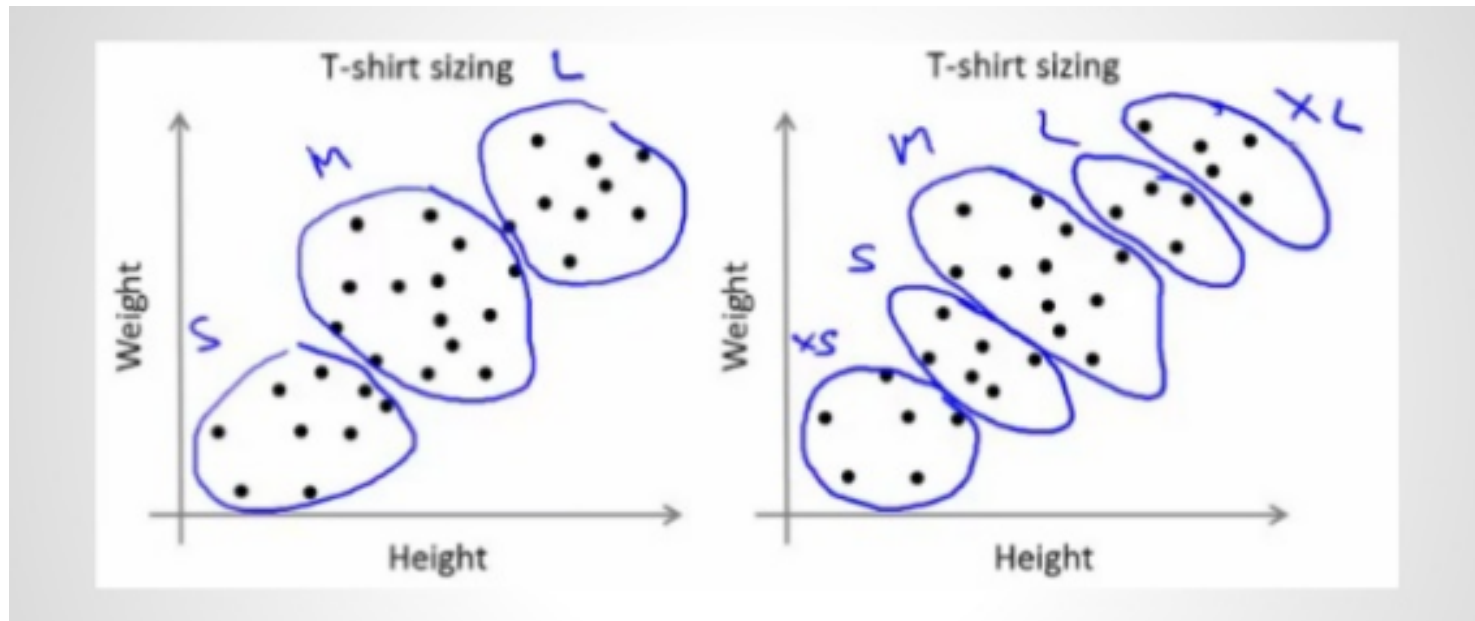
## 0.17 Différentes formes de Modèles

**Tout dépend des données** (qui 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...

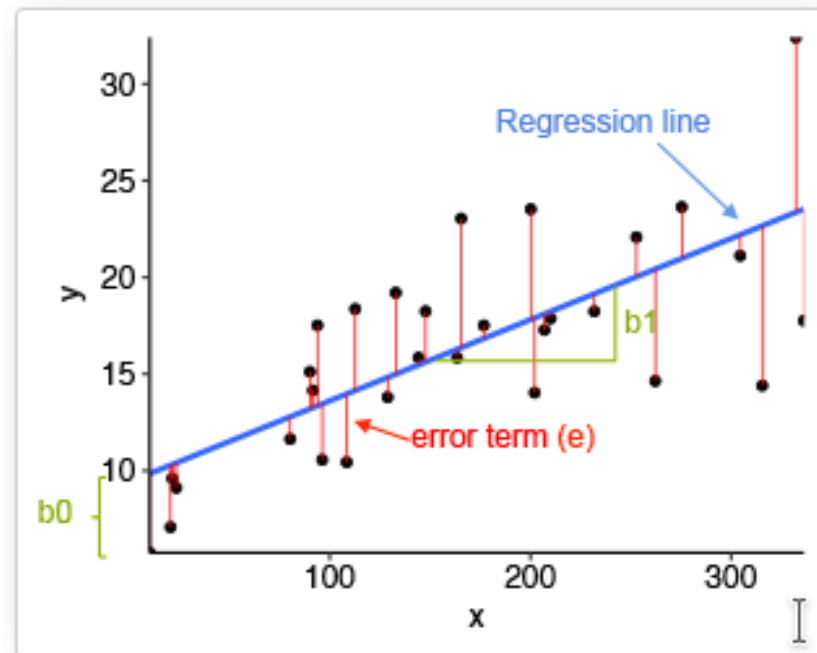


### 0.17.1 Exemple : Clustering



## 0.17.2 Régression Linéaire

- $y \simeq b_0 + b_1 * x$  : la meilleure droite de régression
- $b_0$  : l'intercepte (l'ordonnée à l'origine),  $b_1$  : la pente
- $e$  : l'erreur



### 0.17.3 Exemple : réseaux de "causalités" (BN)

Ce BN permet de simplifier  $P(I, D, G, S, L) = P(I)P(D)P(G|I, D)P(S|I)P(L|G)$ .

On peut calculer la chance d'obtenir une "bonne lettre de recommandation"!

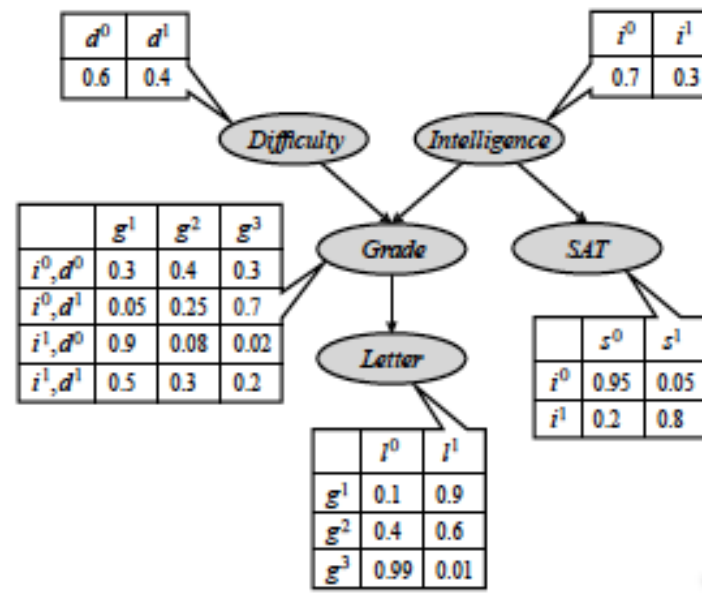
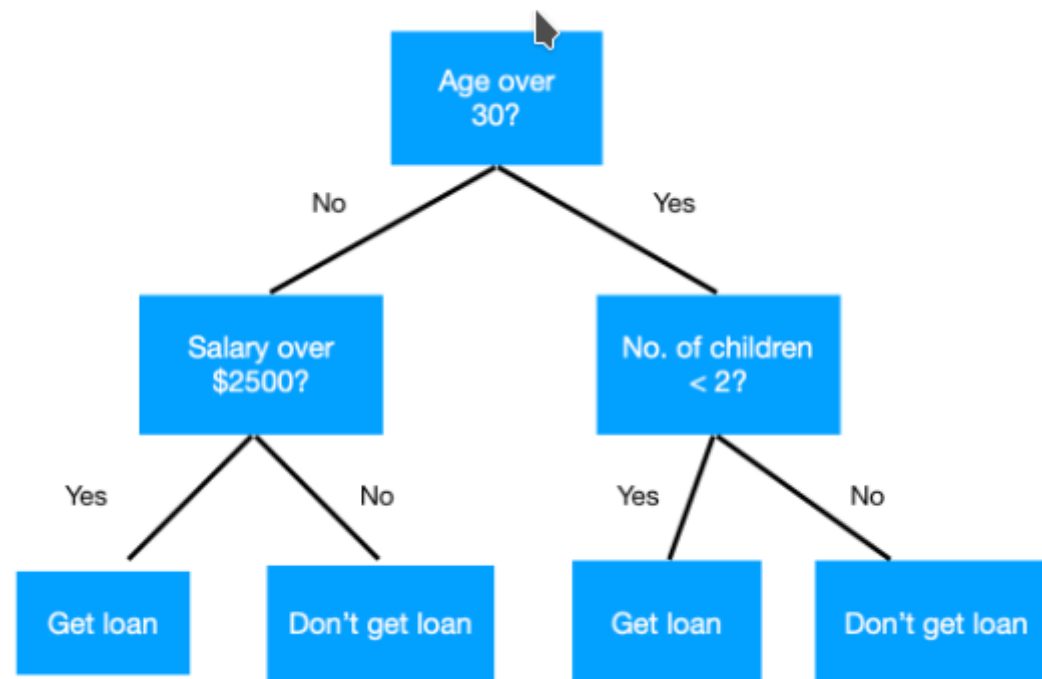


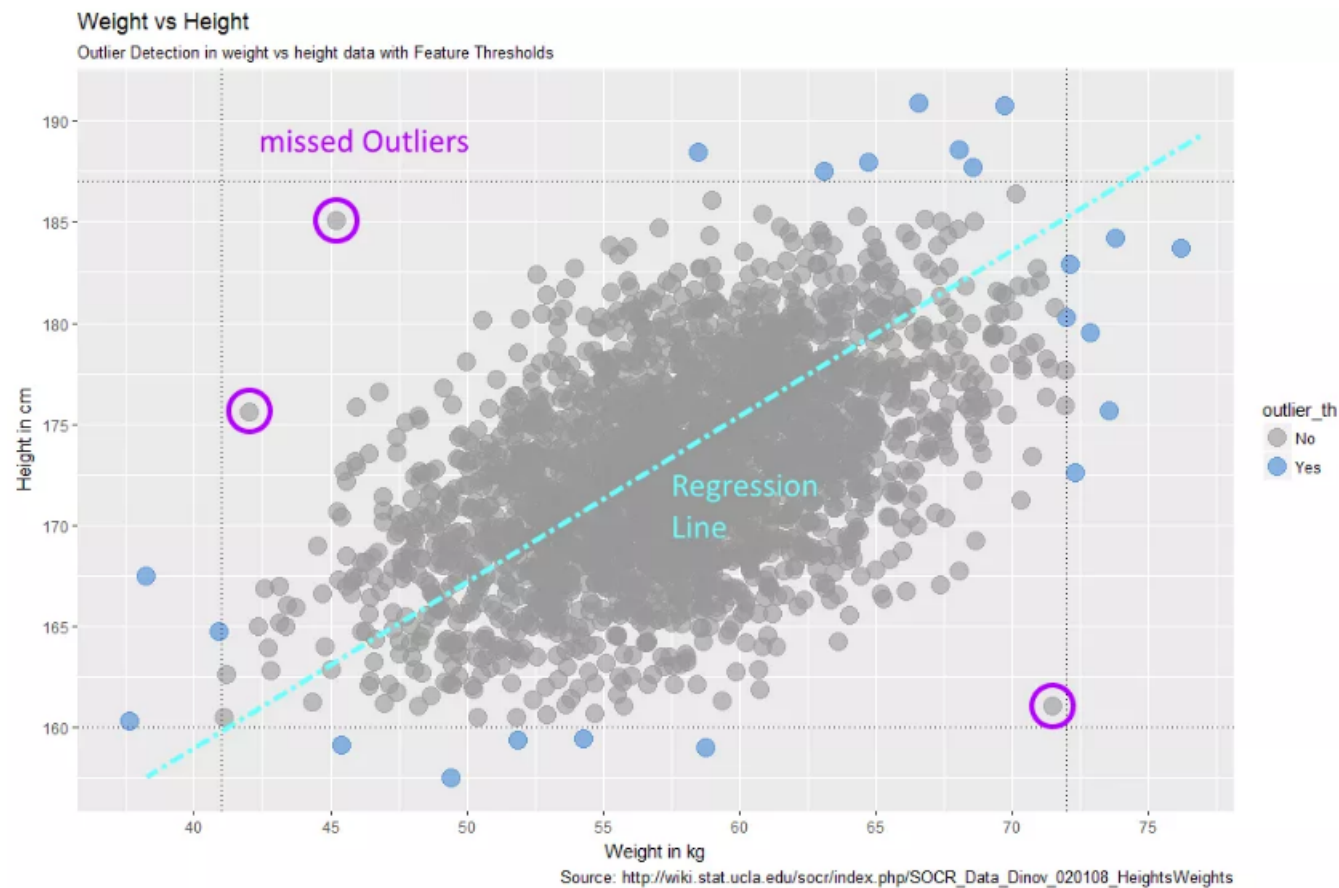
FIGURE 2 – Réseau Bayésien d'étudiants :  $X^{0,1} : X \in \{F, T\}$

### 0.17.4 *Arbre de décision*

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

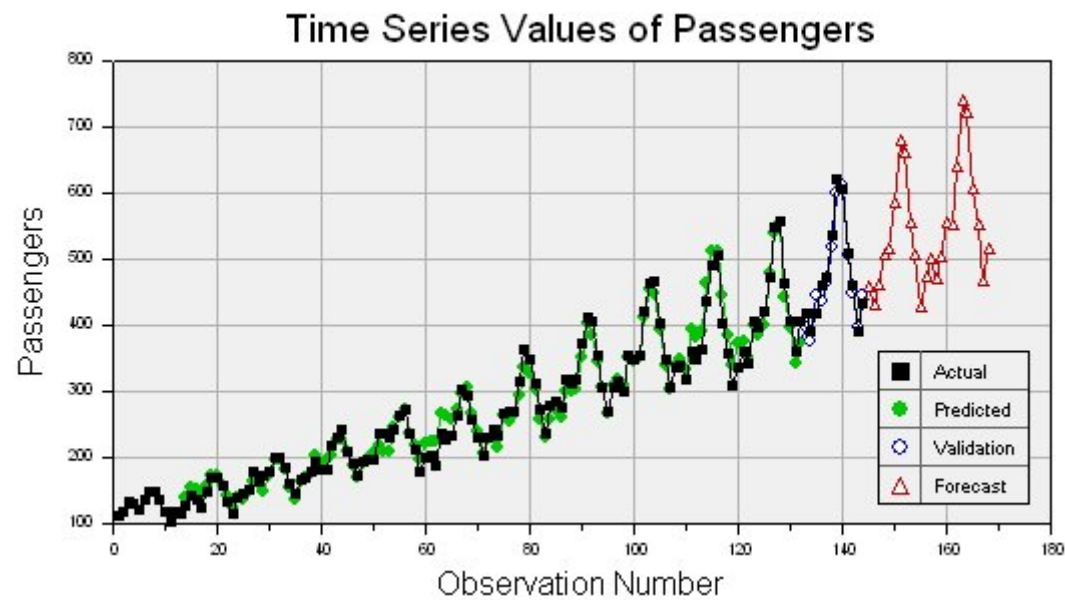


### 0.17.5 Détection d'anomalies (et d'outsiders)



### 0.17.6 Données saisonnières

- Prédiction sur des données saisonnières (par ex. en météo, en Finances, prédiction de charges, etc.)



### 0.17.7 Recherche d'associations

- Recherche d'associations / corrélations dans les données

## Analyse des associations

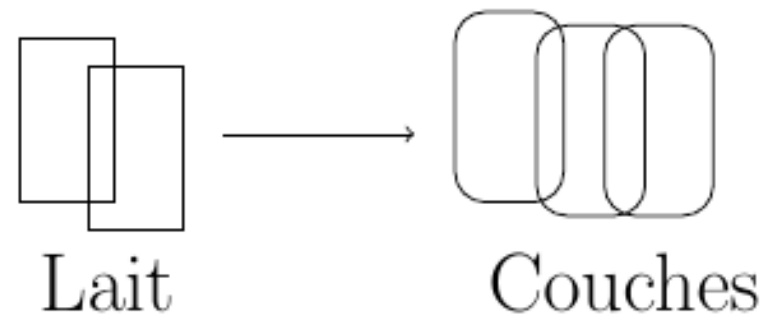


FIGURE 3 – règles d'association

## 0.17.8 Cas de données séquentielles et Images

- Séquencement génétique :

```

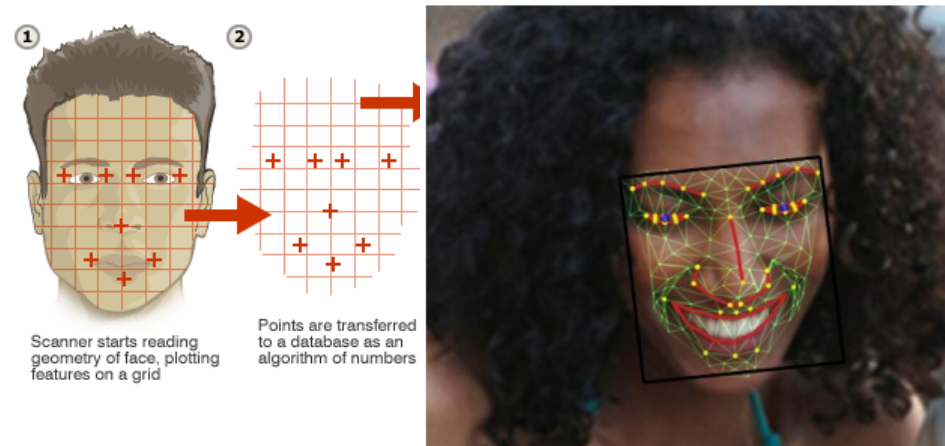
AAB24882      TYHMCQFHCRCYVNNHSGEKLIECNERSKAFSCPSHLQCHKRRQIGETHEHNQCGKA
AAB24881      -----YECNQCGKAFQAQSSSLKCHYRTHIGEKPYECNQCGKA
               **** : ,***: * *:* * * :***** :* *****

AAB24882      PSHLQYHERHTHTGKPYECHQCGQAFKKCSLLQPHKRTHTGKPYE-CNQCGKAFQAQ
AAB24881      HSHLQCHKRTHTGKPYECNQCGKAFSQHGLLQPHKRTHTGKPYMNVINMVKPLHN
               ***** :***** :***** :***** :***** :***** :

```

→ Fait appel aux techniques avancées du "Sequence Mining".

- Image : p. ex. la reconnaissance faciale :





### 0.17.9 Réseaux de Neurones

Exemple d'un réseau de neurones de reconnaissance de chiffres manuscrits :

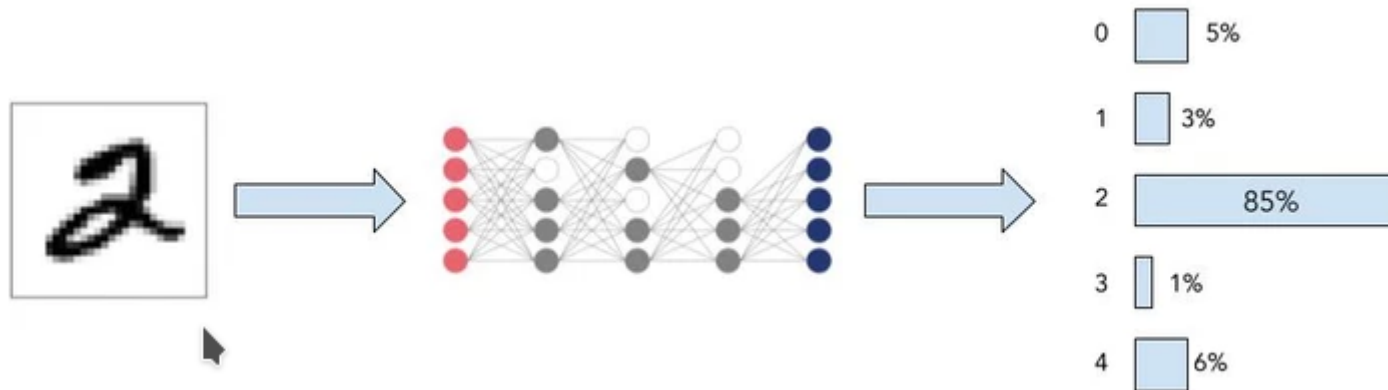


FIGURE 4 – (Thanks to <https://www.datagenius.fr>)

### 0.17.10 Apprentissage de règles

Exemple : un banquier a appris à définir un "risque" acceptable pour un prêt

- Définition du concept par des **règles** telles que :

*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*

- Une vue à base d'instance :

**Instance 1** : *revenus annuel = 32000 & dans la même boîte=6 & Possède sa maison*

**Instance 2** : *revenus annuel = 52000 & dans la même boîte=16 & Locataire*

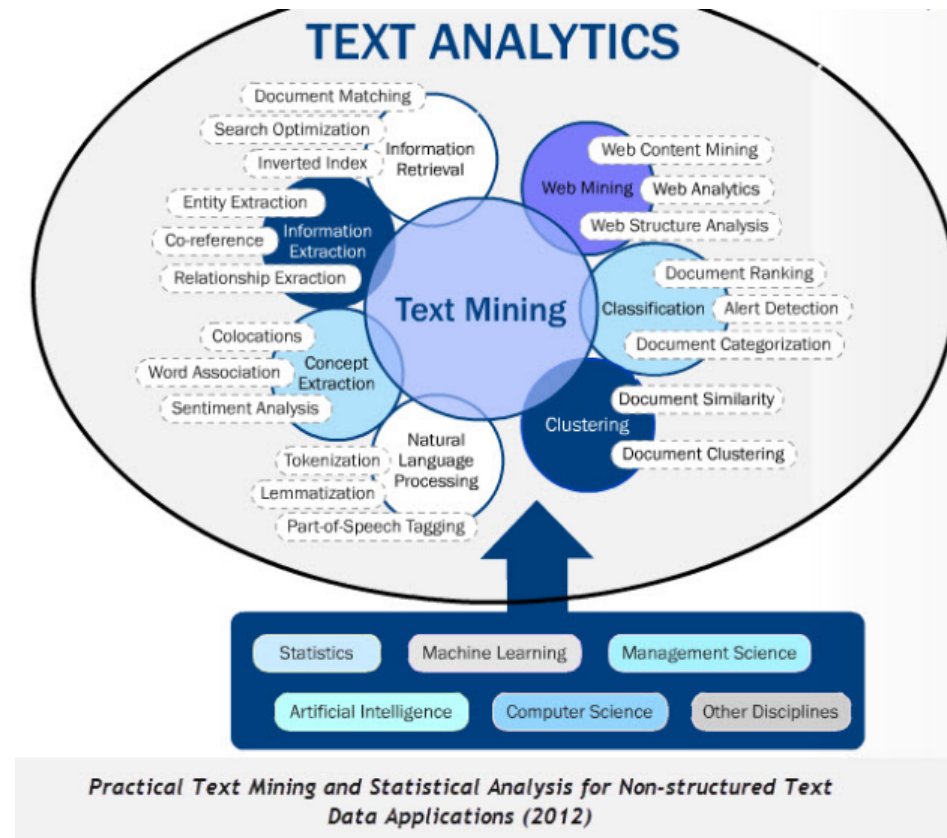
**Instance 3** : *revenus annuel = 28000 & dans la même boîte=12 & Possède sa maison*

- Il existe d'autres formes ...

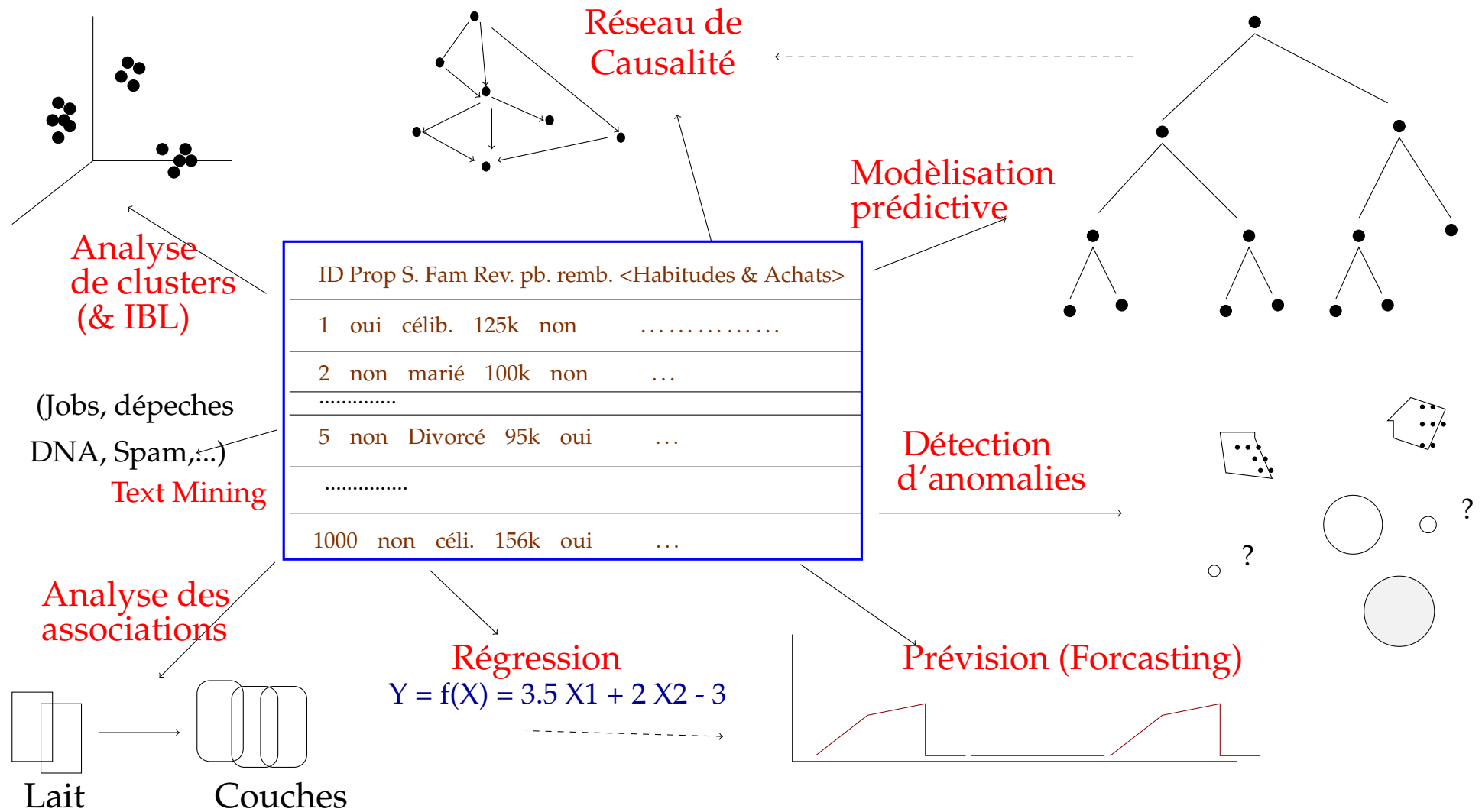
### 0.17.11 Text Mining

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

→ Jobs, SPAM, DNA, dépêches, ...



## 0.18 Résumé des principales tâches de l'EC

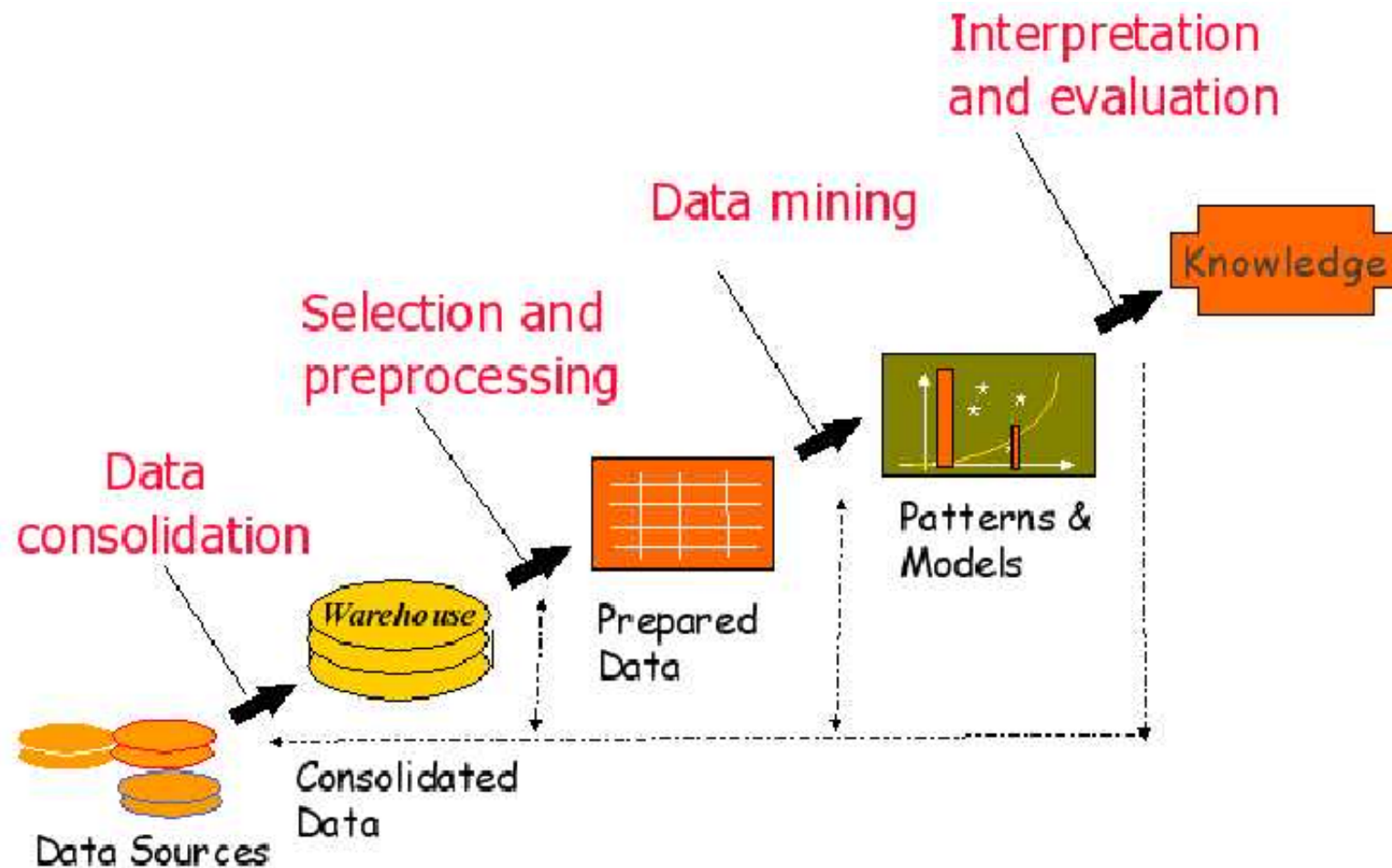


- La base de données est le point de départ

ID Prop S. Fam Rev. pb. remb. <Habitudes & Achats>					
1	oui	célib.	125k	non	.....
2	non	marié	100k	non	...
.....					
5	non	Divorcé	95k	oui	...
.....					
1000	non	céli.	156k	oui	...

- Une base de données peut contenir des données
  - transactionnelles (supermarchés, banques, données saisonnières, etc.)
  - images (pixels), son (signaux),
  - texte (mail, texte littéraire, séquences génétiques, des nombres, etc.),
  - données multimédia (au sens "plus d'un médium") comme les pages WEBs,
  - courbes, histogrammes, fonctions, relations, graphes ou arbres, etc.
- ➔ Voir plus loin quelques exemples de modèles ../..

## 0.19 Processus d'Extraction de Connaissances dans les BDs

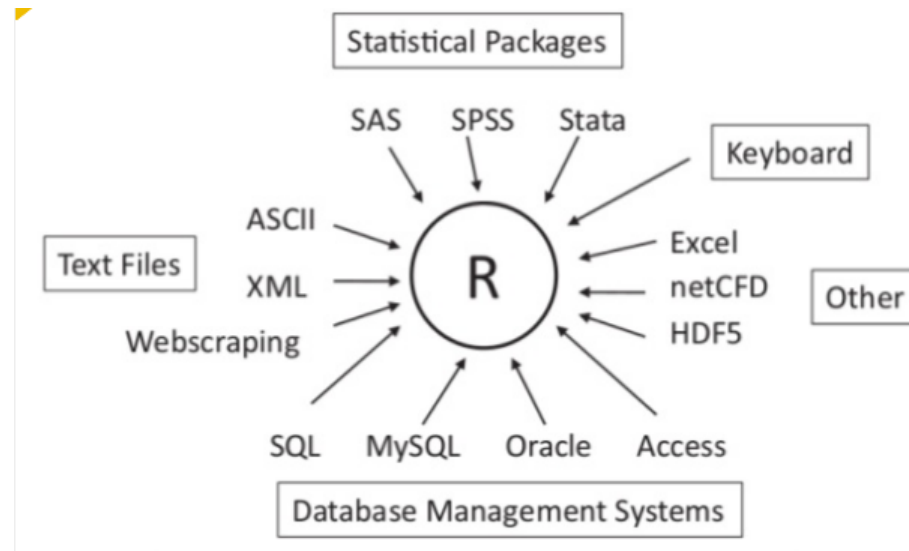


- L'ensemble du processus : **KDD** (Knowledge Discovery in Data)

## 0.20 Quelques Outils

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

- R :



- DeepNet : outils Deep Learning pour R

- Quelques outils de *Machine learning* Open-source :
  - *Sci-Kit Learn*, *Theano*, *Keras* de Python
  - *TensorFlow* de Google
    - graphe de noeuds = opérateurs et des arêtes = tables de données=tensors + calculs massivement parallèles
  - *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



## 0.21 Qu'est-ce que les ordinateurs peuvent apprendre

- Le processus d'apprentissage est complexe.

**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 des problèmes  
 $\rightarrow$  scénario 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

☞ Les ordinateurs apprennent bien les concepts.

- **Concept** : la sortie 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.
- Ils sont représentés par : *arbres, règles, réseaux, équations/fonction, ....*
  - ⇒ Les arbres (de décision) et les règles (compréhensibles pour l'humain)
  - ⇒ Idem pour les réseaux Bayesiens (graphes)
  - ⇒ Les réseaux (de neurones) et les équations, ... (moins évidents!)

→ le "comment / pourquoi" : black-Box vs. explication

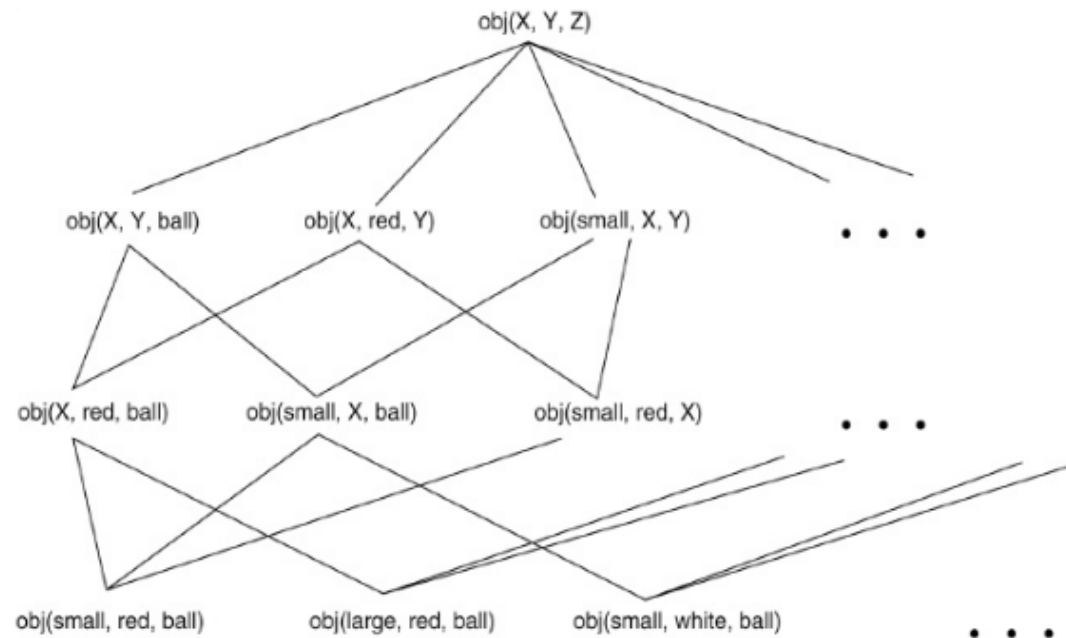
## 0.22 Recherche de Concepts par Généralisation (Induction)

### Vue générative : exemple d'Espace de versions

Taille : {large, small}

Couleurs : {red, white, blue}

Obj : {ball, brick, cube}



## Enumération de l'espace de concepts : Espace de Versions (théorique)

- L'espace de description de concepts consistents

↳ **L** (least) et **G** (greatest) general descriptions

**L** : les descriptions *les plus spécifiques* qui couvrent tous les Exemples positifs et aucun Exemple négatif (de manière spécifique).

**G** : les descriptions *les plus générales* qui ne couvrent aucun Exemple négatif mais tous les Exemples positifs (de manière générale).

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

☞ L'objectif de l'apprentissage automatique et de modéliser **G** par induction (extrapolation) sur **L** en éliminant les exemples négatifs.

- **Inconvénients** : coûteux en temps de calcul, ne résout pas de problème pratique!

→ Mais donne une définition théorique et claire ; permet l'apprentissage on-line, ...

## Exemple

- Soit le vocabulaire : couleurs  $\in \{\text{rouge, vert}\}$ , animaux  $\in \{\text{vache, poule}\}$
- Et quelques observations (au départ, toute combinaison est *possible sauf* contre indication) :

Obs. positifs	Obs. négatifs	L	G
$\langle \rangle$		$\{\}$	$\{\langle *, * \rangle\}$
$\langle \text{vache, verte} \rangle$		$\{\langle \text{vache, verte} \rangle\}$	$\{\langle *, * \rangle\}$
	$\langle \text{poule, rouge} \rangle$	$\{\langle \text{vache, verte} \rangle\}$	$\{\langle *, \text{verte} \rangle, \langle \text{vache}, * \rangle\}$
$\langle \text{poule, verte} \rangle$		$\{\langle *, \text{verte} \rangle\}$	$\{\langle *, \text{verte} \rangle, \langle \text{vache}, * \rangle\}$

- Si l'on ajoute  $\langle \text{vache, rouge} \rangle$  ? :
  - Si ex. positif : ajouter  $\langle \text{vache}, * \rangle$  à L
  - Si ex. négatif : retirer  $\langle \text{vache}, * \rangle$  de G



Inconsistance si les exemples **positifs** et **négatifs** se contredisent.

**Rappelez-vous** (cas d'une langue parlée (difficile)) :

- On a des lettres a..z
  - A priori, on peut former n'importe quel mot ,
    - P. Ex. "Honkre !"
    - Des règles **lexicales** nous limitent (vérification dans un dico? apprentissage?)  
"Honkre!" n'existe pas!
- On a des mots ...
  - A priori, on peut former n'importe quelle séquence (phrase),
    - P. Ex. "Le soupe mange La chien !"
    - Des règles **syntaxiques** (grammaticales) nous limitent (apprentissage?)  
"La chien", "Le Soupe"
    - Des règles **sémantiques** nous limitent (apprentissage?)  
"Une soupe ne mange pas un chien!"
- On a des phrases ...
  - Des règles **sémantiques discursives** nous limitent (apprentissage?)
  - P. Ex. les phrases de "Lorem Ipsum" ou le discours d'un bébé de 3 ans

## 0.23 Les questions qui se posent

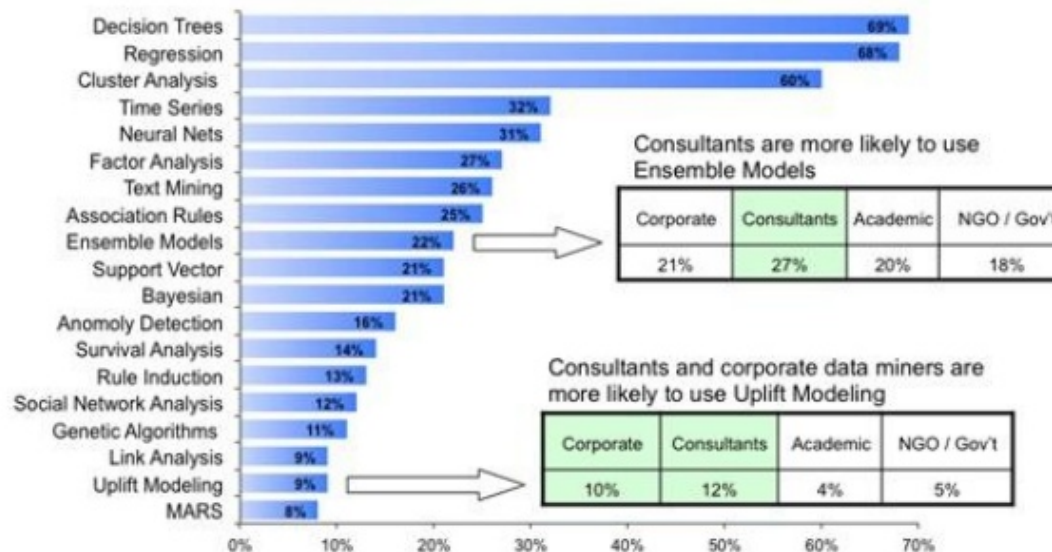
Avant toute chose, il faut se poser quelques questions :

- La forme des entrées ?
  - Hypothèse de représentativité
- La forme des sorties ?
  - Quels types de Motifs (Patterns) ?
  - Comment les décrire ?
- **Evaluation** : un résultat est toujours accompagné d'indicateurs (erreur, intervalle de confiance, mesures diverses, ...)
- L'expertise est indispensable en apprentissage

## Les algorithmes :

### Most popular Data Mining algorithms

- Decision trees, regression, and cluster analysis continue to form a triad of core algorithms for most data miners. This has been very consistent over time.
- However, a wide variety of algorithms are being used.





Parmi ces algorithmes :

- Régression (différentes variantes)
- Classification : Arbres de décision (C 4.5)
- Classification : Bayes
- Clustering : Kmeans
- Classification : Support Vector Machine
- Règles d'association : Apriori
- Classification : PageRank (importance d'un objet p/r aux autres)
- Classification : Boosting (AdaBoost)
- Classification : PCA
- Recommandation : Collaborative filtering
- Classification : Bootstrap Aggregating (RF)

## 0.24 Extraction de Motifs (Patterns) sur une BD.

### 0.24.1 Exemple d'un jeu in-door

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

## 0.24.2 *Extraction de règles*

### **Résultats d'une première tentative d'apprentissage : (règles de classification)**

Un 1er ensemble de règles apprises (indicatif) : une liste de décision

- |   |     |
|---|-----|
| - Si <i>aspect=ensoleillé &amp; humidité=forte</i> alors <i>jouer=non</i> | (1) |
| - Si <i>aspect = pluvieux &amp; vent = vrai</i> alors <i>jouer=non</i>    | (2) |
| - Si <i>aspect = nuageux</i> alors <i>jouer=oui</i>                       | (3) |
| - Si <i>humidité = normale</i> alors <i>jouer=oui</i>                     | (4) |
| - Si <b><i>aucun ci-dessus</i></b> alors <i>jouer=oui</i>                 | (5) |

La même base d'instance avec deux attributs **numériques** :

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

- Attributs mixtes dans la BD → **discrétisation**

<i>Si aspect = ensoleillé &amp; humidité &gt; 83 Alors jouer=non</i>
--

### 0.24.3 *Arbre de décision*

- Stratégie *Diviser et Régner* (Divide & Conquer)

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

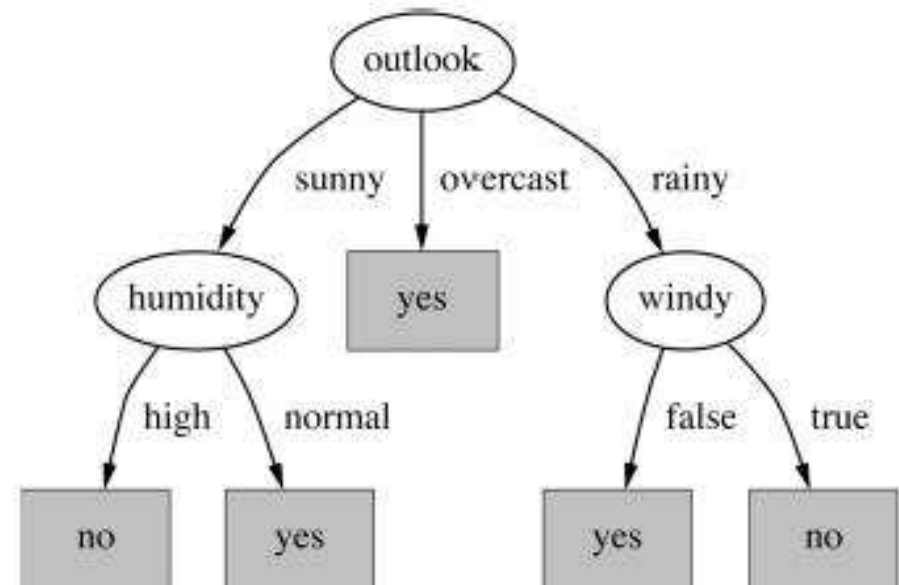
TABLE 1 – Données météo (jeu)

- Problème : quel attribut choisir ? → notion d'information / pureté

../..

L'Arbre de décision final :

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



### 0.24.4 Le calcul de l'information

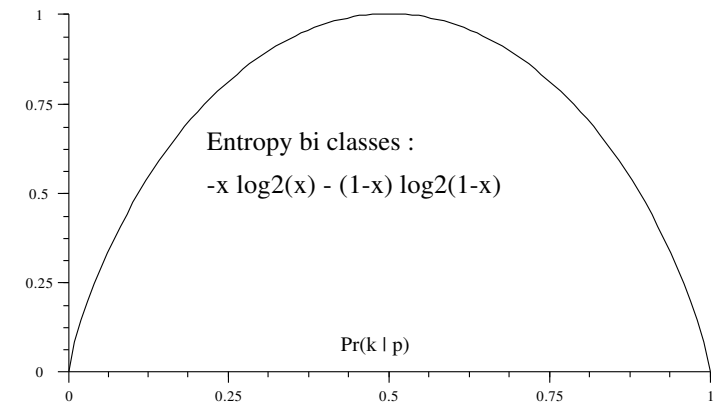
$$\text{entropie}(p_1, p_2, \dots, p_n) = -p_1 \times \log p_1 - p_2 \times \log p_2 \dots - p_n \times \log p_n$$

➡ Les arguments  $p_i$  sont des fractions et la somme des  $p_i = 1$ .

Étant donné une distribution de probabilités (ici fréquences), **l'information nécessaire pour prédire un évènement** est *l'entropie de la distribution*

- Étant données une position  $p$  dans l'arbre et  $c$  classes que l'on cherche à prédire, l'entropie associée à  $p$  est donnée par

$$H(p) = - \sum_{k=1}^c \text{Pr}(k|p) \log_2(\text{Pr}(k|p))$$

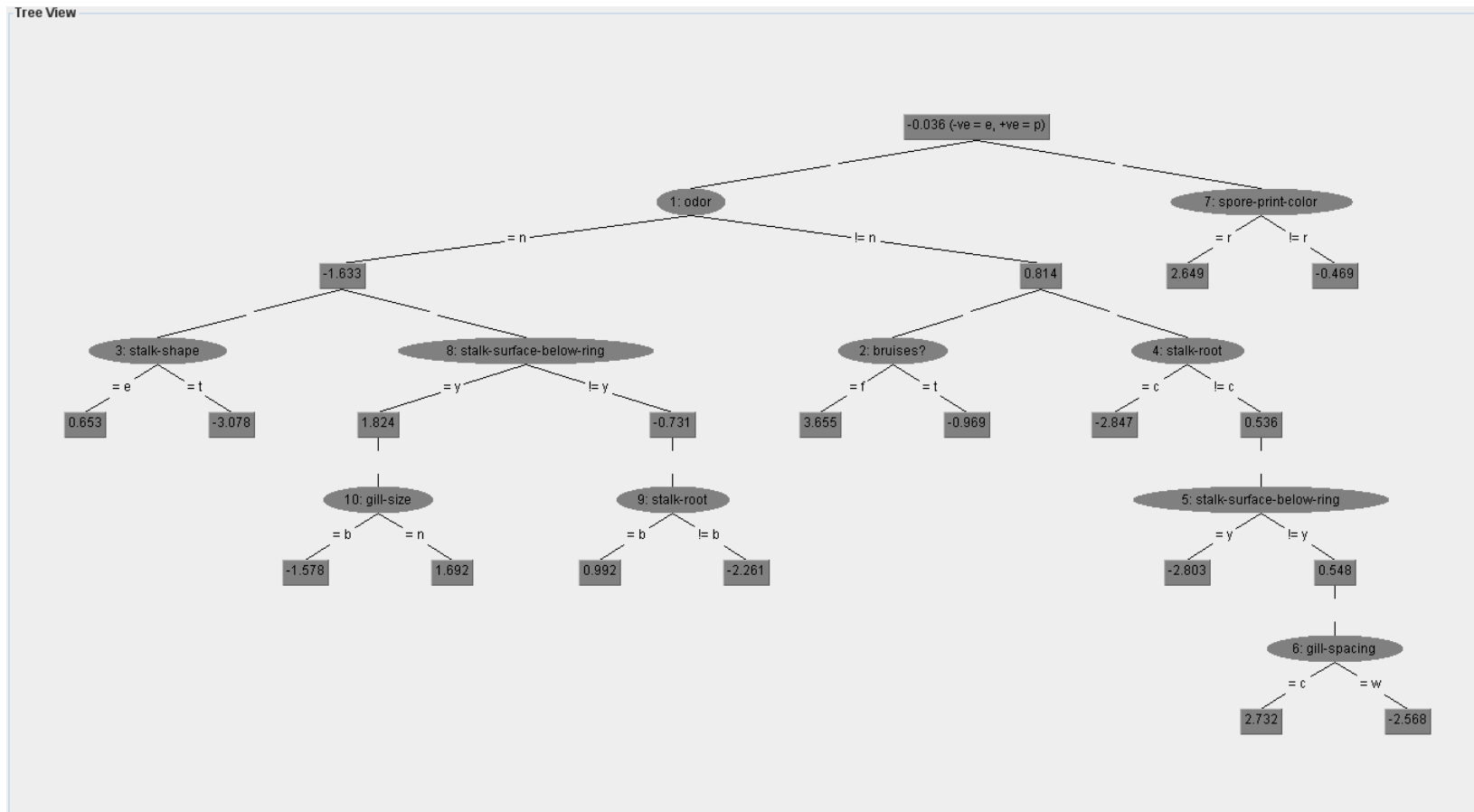


dim sep 30 15:20:37 2007

L'**entropie** permet de mesurer l'incertitude relative à l'appartenance des objets aux différentes classes (si tous les objets appartiennent à une seule classe, l'incertitude est nulle).

### 0.24.5 Un autre exemple d'arbre de décision

- BD de **8124** champignons avec **107** attributs dont les types *comestible* et *non-comestible*.
- Les règles de classification (99,5% de succès, 14 règles)





## 0.24.6 Régression

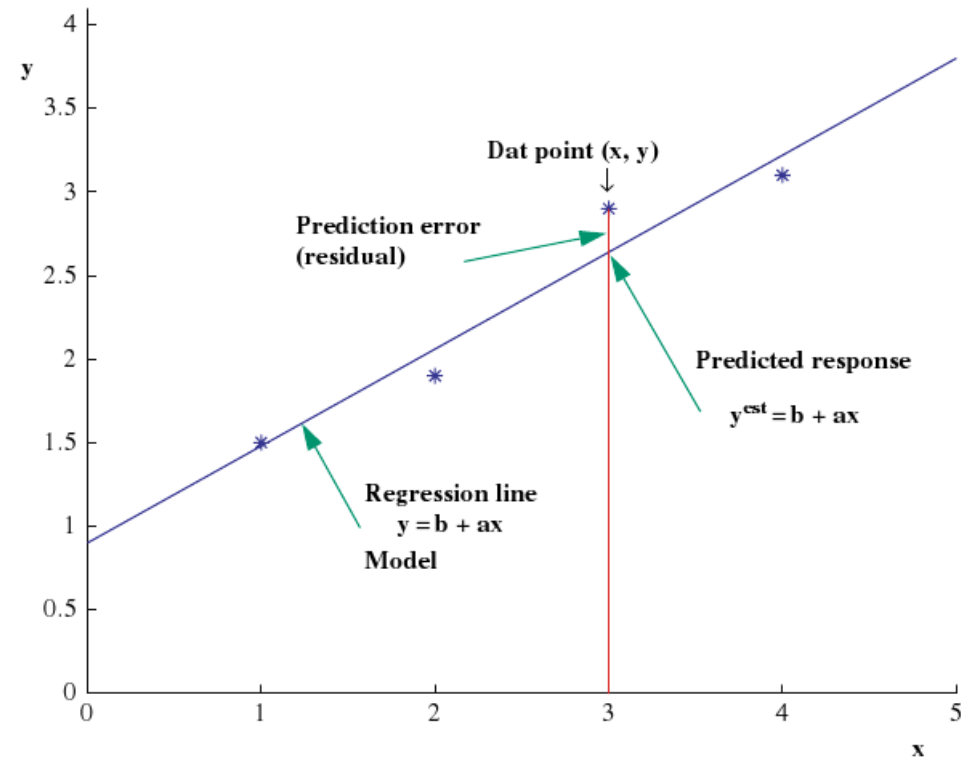
La **Rég. lin.** apprend les  $(k+1)$  pondérations  $w_j$  en **minimisant** la somme des carrés des différences entre les classes connue ( $y^{(i)}$ ) et prédite ( $\sum w_j a_j^{(i)}$ ).

→ 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 Logistique

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

$$\log\left(\frac{P}{1-P}\right) = w_0a_0 + w_1a_1 + w_2a_2 + \dots + w_ka_k$$

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

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

## Difficultés et limites des modèles de régression :

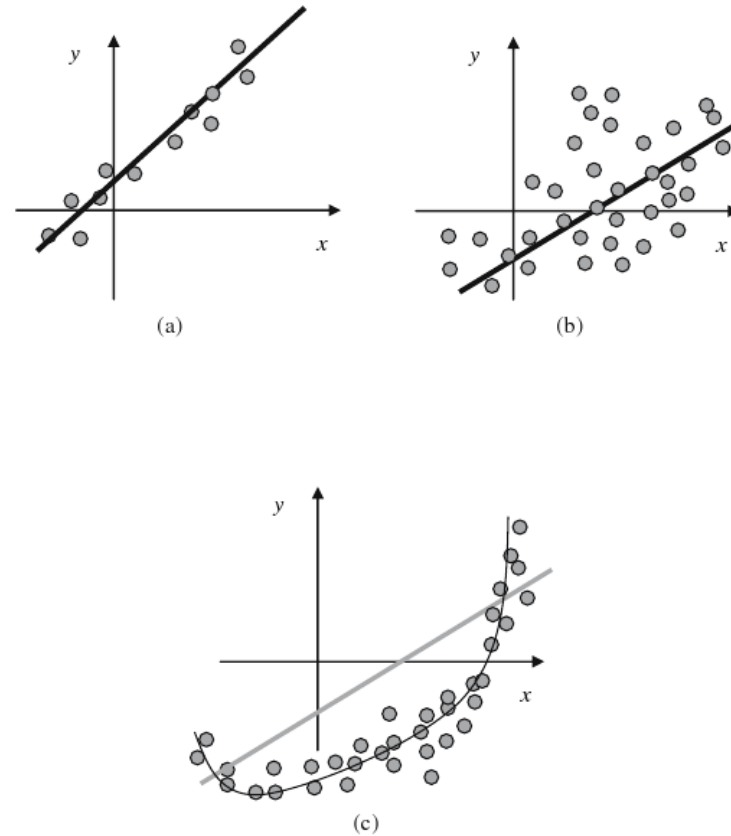


FIGURE 5 – (a) : classifieur linéaire avec une bonne dispersion, (b) : linéaire avec une dispersion importante, (c) : non linéaire avec une dispersion importante

## Une autre exemple de régression : prédiction des Performances

- Performances relatives des PC selon certains attributs.

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMA	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
...							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

- 209 configurations différentes (une ligne = une configuration)  
 ⇒ Les attributs et la sortie sont tous **numériques** (vs. cas général : mixtes).

$$PRP = -55.9 + 0.0489 MYCY + 0.0153 MMIN + 0.0056 MMA + 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX.$$

### 0.24.7 Extraction de règles d'association

- Soit  $support \geq 2$

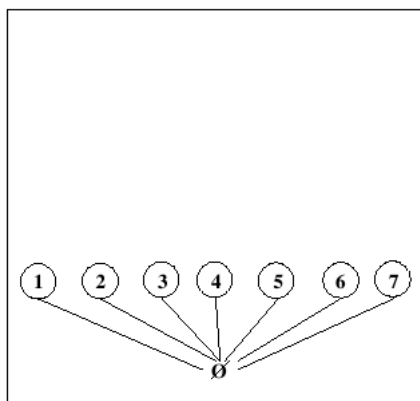
**BD trans.**

Id Trans.	Les items
50	4, 6, 7
100	4, 5
120	2, 3, 5
150	3, 4, 5
200	1, 2, 3, 5

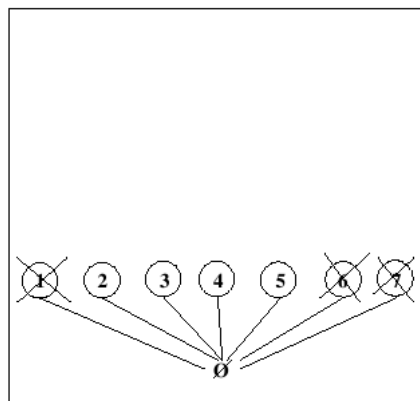
→ **Ens.  $C_1$**

Itemset	Supp.
{1}	1
{2}	2
{3}	3
{4}	3
{5}	4
{6}	1
{7}	1

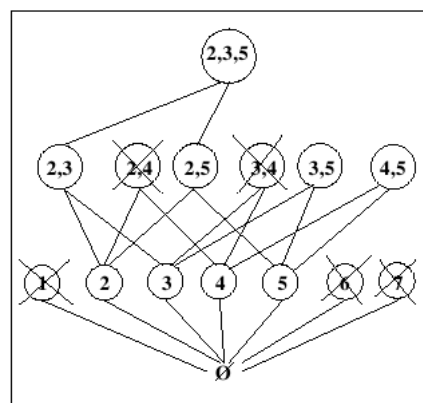
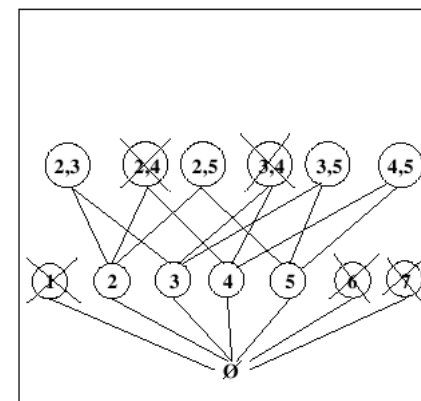
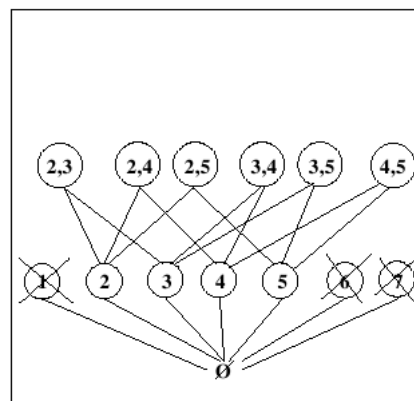
## Itemsets : vision par Treillis



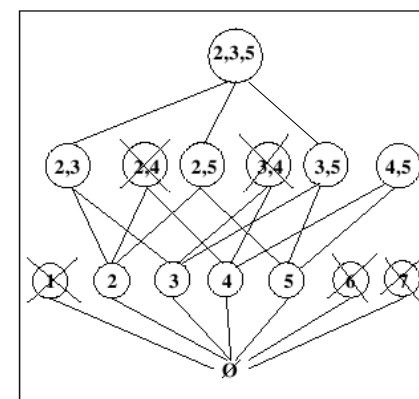
$$C_1 = \{1, 2, 3, 4, 5, 6, 7\}$$



$$L_1 = \{2, 3, 4, 5\}$$



→



$$L = L_1 \cup L_2 \cup L_3 = \{\{2\}, \{3\}, \{4\}, \{5\}, \{2, 3\}, \{2, 5\}, \{3, 5\}, \{4, 5\}, \{2, 3, 5\}\}$$

## Exemple pour météo

- On transforme chaque item-set en (un ensemble éventuellement vide de) règles avec la **confiance** minimum spécifiée.

⇒ Par Exemple, **Humidity=normal, windy=false, play=yes (4)**

donne  $7(2^N - 1)$  règles potentielles avec une valeur de confiance.

if humidité=normale and windy=false then play=yes	(4/4)
if humidité=normale and play=yes then windy=false	(4/6)
if windy=false and play=yes then humidité=normale	(4/6)
if humidité=normale then windy=false and play=yes	(4/7)
if windy=false then humidité=normale and play=yes	(4/8)
if play=yes then humidité=normale and windy=false	(4/9)
if True then humidité=normale and windy=false and play=yes	(4/14)

### 0.24.8 Apprentissage à base d'instances (IBL)

- Stockage des instances
- Fonction de **distance** pour déterminer la classe d'une nouvelle instance

⇒ La fonction de *distance* définit ce qui est appris

⇒ La fonction de distance simple si les attributs numériques.

⇒ La plupart des schémas *IBL* utilisent une **distance Euclidienne**.

⇒ La distance entre les instances avec des valeurs de  $k$  attributs :

1<sup>e</sup> instance  $a^{(1)} : a_1^{(1)}, a_2^{(1)}, \dots, a_k^{(1)}$       2<sup>e</sup> instance  $a^{(2)} : a_1^{(2)}, a_2^{(2)}, \dots, a_k^{(2)}$

$$\text{La distance Euclidienne} = \sqrt{\left(a_1^{(1)} - a_1^{(2)}\right)^2 + \left(a_2^{(1)} - a_2^{(2)}\right)^2 + \dots + \left(a_k^{(1)} - a_k^{(2)}\right)^2}$$

➡ Quand on compare les distances, la racine carré est inutile

- Il existe une pléthore de distances.

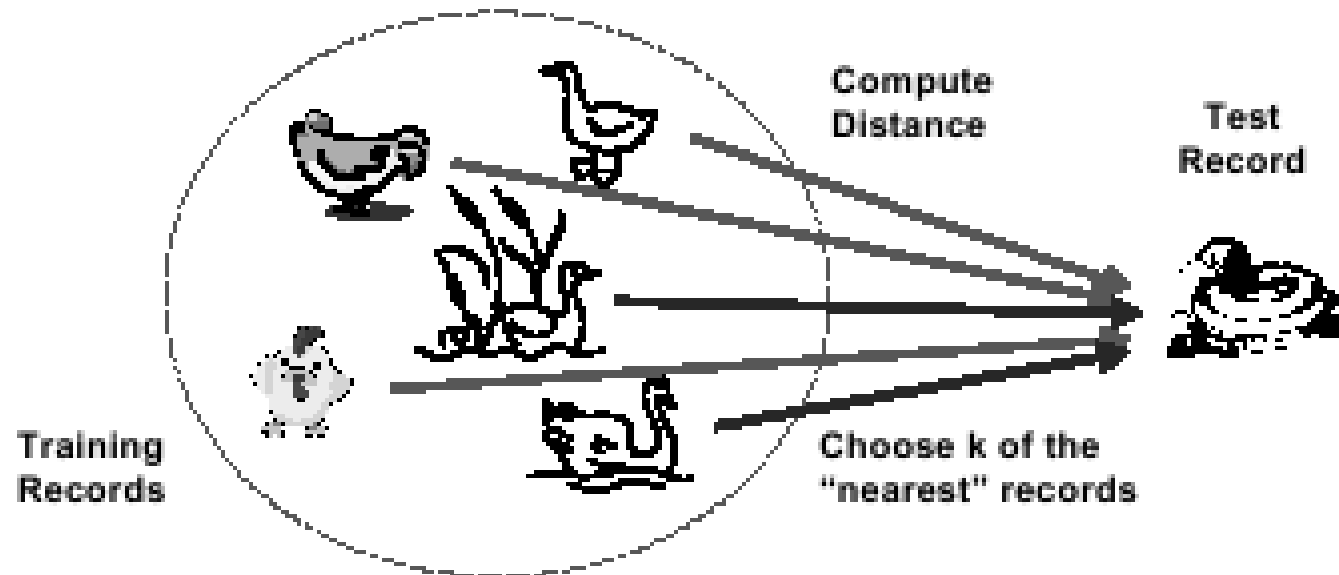


- L'approche **1-plus proche voisin (1-NN)** insuffisante.

☞ **Une solution : K-plus proches voisins**

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

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



### 0.24.8.1 Un exemple : la cueillette de champignons

- Un groupe d'amis se propose d'aller ramasser des champignons le week-end prochain.

Mais les champignons sont capricieux et seules de bonnes conditions météorologiques dans la journée qui précède favorisent leur croissance.

Aussi, ils préfèrent analyser leur cueillette des années précédentes pour savoir si le ramassage permettra ou non une fricassée au dîner.

Ils disposent de l'agenda suivant :

		Temps	Humidité	Vent	Bonne cueillette
E1	WE 1, année-1	nuageux	haute	oui	oui
E2	WE 2, année-1	nuageux	basse	non	non
E3	WE 3, année-1	nuageux	basse	oui	non
E4	WE 1, année-2	soleil	haute	oui	oui
E5	WE 2, année-2	pluvieux	basse	oui	non
E6	WE 3, année-2	nuageux	haute	non	oui
E7	WE 4, année-2	soleil	basse	non	non

TABLE 2 – Anals champignons

Ce week-end : **ensoleillé, humide et sans vent**. → Ira-t-on cueillir des champignons ?

## 0.24.9 Modélisation Statistique

### Règles de produit et règle de Bayes

- **Règle de produit :**  $P(A, B|C) = P(A|B, C) P(B|C) = P(B|A, C) P(A|C)$

➡ Justification : avec  $P(X, Y) = P(X|Y).P(Y)$  → poser  $Y = (B, C)$  puis ré-appliquer

$$P(A, B|C) = \frac{P(A, B, C)}{P(C)} = \frac{P(A|B, C).P(B, C)}{P(C)} = P(A|B, C) P(B|C)$$

- **Règle de Bayes :**  $P(A|B, C) = \frac{P(B|A, C) P(A|C)}{P(B|C)}$

➡ Justification : appliquer la 1e égalité de la règle produit puis la 2e.

- Plus général : la formule de Bayes dans un context (background)  $c$  :

$$\boxed{Pr[H|E, c] = \frac{Pr[E|H, c] \times Pr[H|c]}{Pr[E|c]}} \rightarrow \boxed{P(\text{Modèle}|Data) = \frac{P(\text{Modèle}).P(Data|\text{Modèle})}{P(Data)}}$$

### 0.24.9.1 Prédiction Bayésienne sur l'Exemple Météo

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

TABLE 3 – BD Exemple "météo"

- Deux hypothèses : les attributs sont d'importance égale,
  - ➔ elles 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)

### 0.24.9.2 Application de de Bayes

- L'hypothèse  $H$  et l'évidence  $E$  basée sur  $H$  : 
$$\Pr[H \mid E] = \frac{\Pr[E|H] \times \Pr[H]}{\Pr[E]}$$

➡ **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) :

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

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

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



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

- Pour une nouvelle instance à classer, l'évidence **E** :

Outlook	Temperature	Humidity	Windy	Play
Sunny	Cool	High	True	??

⇒ Probabilité de "yes" pour la nouvelle instance :

$$\begin{aligned}
 \Pr[\text{yes} \mid E] &= \Pr[\text{Outlook} = \text{Sunny} \mid \text{yes}] \times \\
 &\quad \Pr[\text{Temperature} = \text{Cool} \mid \text{yes}] \times \\
 &\quad \Pr[\text{Humidity} = \text{High} \mid \text{yes}] \times \\
 &\quad \Pr[\text{Windy} = \text{True} \mid \text{yes}] \times \frac{\Pr[\text{Yes}]}{\Pr[E]} \\
 &= \frac{2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14}{\Pr[E]} = 0.205
 \end{aligned}$$

⇒ N.B. Ici,  $\Pr[E]$  ( $= \Pr[\text{yes}|E] + \Pr[\text{no}|E]$ ) disparaîtra avec la normalisation.

### 0.24.9.3 Valeurs Numériques dans Bayes

- Hypothèse : elles ont une distribution **normale** ou **Gaussienne** de probabilités
- La *fonction de densité de probabilité* pour une distribution normale avec :

$$\mu \text{ la moyenne : } \mu = \frac{1}{n} \sum_{i=1}^n x_i \quad \sigma \text{ l'écart type : } \sigma^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n - 1}$$

→ la fonction de densité (loi normale/gaussienne) :

$$f(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x - \mu)^2}{2\sigma^2}}$$

⇒ Le "-1" sur "n" concerne le **degré de liberté** dans les instances

⇒ Les valeurs manquantes (dans une instance) n'interviennent pas dans  $\mu$  et  $\sigma$

- Pour calculer la probabilité (e.g. pour "yes") pour une nouvelle instance , on utilisera la fonction de densité pour les numériques et les fréquences pour les nominaux

### 0.24.10 *Un exemple d'analyse Bayésienne (à partir du REX)*

#### **Un cas d'analyse des données de fiabilité du type succès/échec.**

- **Contexte de l'exemple** : données ReX disponibles sur des générateurs d'urgence (moteur diesel) utilisés en dépannage dans une centrale nucléaire (pour le refroidissement des bassins).
    - On a observé le comportement (démarrage ou non) de ces générateurs lors d'une sollicitation (utilisation non continue) dans la situation de remplacement d'urgence.
  - **Modélisation Bayésienne** : identifier une *vraisemblance* et une probabilité *a priori* (*belief*)
    - Principe de Bayes : **Vraisemblance (Rex) x Hypothèse (*a priori*) → Prédiction (*a posteriori*)**
    - Les conditions d'utilisation de ce cas : par sollicitation et bien informé (voir cours 2).
    - Les données succès/échec sont traitées par une distribution **Binomiale** (voir cours 2).
- ☞ Rappel : la *Binomiale* est appropriée dans le cas de test d'un nombre fixe  $n$  de composants testés, où les tests sont supposés **indépendants**, étant donné une **même** probabilité de succès  $\pi$ .



## Remarques :

- La fonction de densité de proba (du *likelihood*) est dans ce cas :  $f(x|n, \pi) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$

Où  $0 \leq \pi \leq 1$  = la proba de succès, pour  $x$  (parmi  $n$ ) succès observés.

☞ la loi binomiale s'applique puisque tous les évènements ont la même probabilité de succès et les tests sont indépendants;

### ☞ Remarques

- La distribution de **Bernoulli** est un cas spécial de la Binomial pour  $n = 1$ .
- Variant dans ces calculs : on peut utiliser le temps d'échec (défaillance) comme une donnée de type succès/échec par rapport à un temps spécifié  $t$ .
  - ➔ C'est à dire :  $x$  sera le nbr de défaillances parmi  $n$  items testés avant le délai  $t$ .
  - ➔ Voir séance 2.

## Précisions :

- Objectif : estimer la probabilité  $\pi$  qui est le paramètre inconnu (par le modèle Binomial).
- Dans le modèle succès/échec, la **vraisemblance**  $f(x|n, \pi) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$  est comme une fonction de  $\pi$  pour  $x$  succès observés,
- Pour faire une **prédiction** (proba. *a posteriori*), il faut spécifier une distribution **a priori** pour  $\pi$ .
- La distribution adaptée pour  $\pi$  est celle qui est conjuguée à la vraisemblance (v. crs 2).
- Cette *distribution a priori conjuguée* pour une vraisemblance **binomiale** est une distribution **Beta** :
 
$$p(\pi|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1} (1 - \pi)^{\beta-1} \quad 0 \leq \pi \leq 1, \alpha > 0, \beta > 0$$
  - Avec  $\alpha$  le nbr *a priori* de succès des composants testés et  $\beta$  le nbr *a priori* des échecs ;
  - Donc  $\alpha + \beta$  est comme la taille *a priori* des données (la BD de référence).
- Bayes : on a notre *vraisemblance* + *a priori*  $\rightarrow$  on peut faire une prédiction (*a posteriori*) .. / ..

**La prédiction** (ajustement de l'*a priori*) :

- Après calculs, la distribution *a posteriori* de  $\pi$  (conditionnée par  $x$  succès observés parmi  $n$  tests)

est de la forme :  $p(\pi|x) \propto \pi^{\alpha+x-1} (1 - \pi)^{\beta+n-x-1}$

→ Cela veut dire que la distribution *a posteriori* de  $\pi$  sachant  $x$  est :

$$\pi|x \sim \text{Beta}(\alpha + x, \beta + n - x)$$

**Appliquons cela à l'exemple des générateurs de secours** et les données *ReX* :

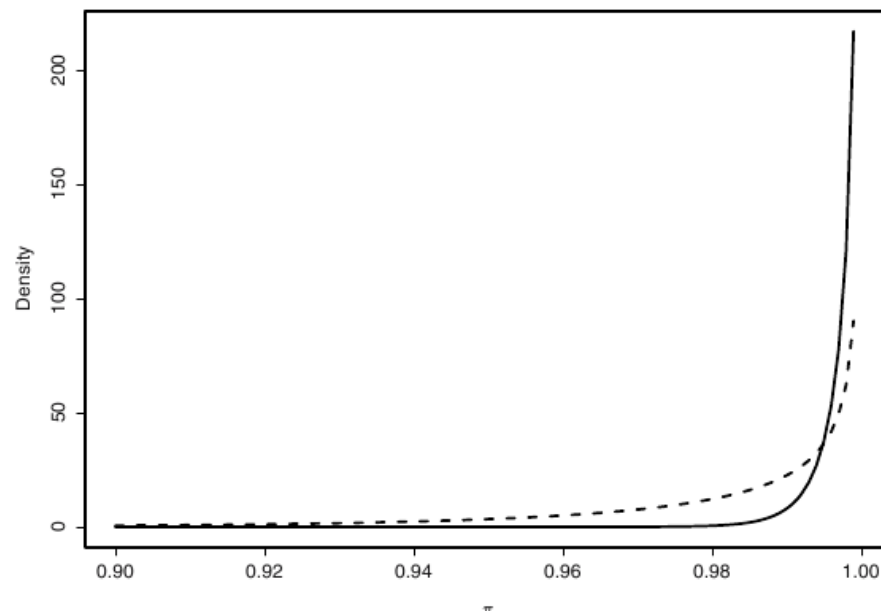
- Les données (*USA nuclear power plants*) sur une centrale nucléaire (de test) sont :
  - Vraisemblance (REX) :  $x = 212$  succès au démarrage sur 212 sollicitations de générateurs
  - A priori : les données de la centrale **de référence** donnent 273 démarrages avec succès sur 278 (donc 5 échecs)
  - MAIS : on ne dispose que seulement de 10% des données de la centrale de référence.
    - Une situation assez fréquente : on pondère les données.

- Dans ce cas, le site de référence génère une distribution *a priori* pour la probabilité de succès  $\pi$  :

$$\text{Beta}(\alpha = (273/278) * 27.8, \quad \beta = (1 - 273/278) * 27.8)) \quad \text{après normalisation/pondération}$$

- De la distribution *a posteriori* précédente, la probabilité  $\pi$  pour la centrale 63 sera :

$$\text{Beta}(\alpha + x, \beta + n - x) = \text{Beta}(239.3 = (273/278) * 27.8 + 212, \quad 0.5 = (1 - 5/278) * 27.8 + 0)$$



La figure :

Les distributions *a priori* et *a posteriori* pour la probabilité de succès  $\pi$

- ligne pointillée : distribution *a priori*
- pleines : la distribution *a posteriori*

→ La distribution *a posteriori* (lignes pleines) est plus piquée que l'*a priori* (pointillée) :

Les données montrent une évidence en faveur d'un taux de succès élevé.

## Résumé de la démarche (Bayésienne) :

- Une connaissance *a priori* ou une expertise permet de se forger une loi de fiabilité (ici loi **Beta**)
  - Cela permet de prédire un comportement du système (sa fiabilité).
- Des observations ont été récupérées sur le terrain (*ReX*).
  - Elles donneront une **vraisemblance** (ici loi **Binomiale**)
- La combinaison de la vraisemblance et de l'expertise permettent de vérifier l'expertise et de la peaufiner.
  - Ce raisonnement permet de calculer une probabilité *a posteriori*  $\propto a\ priori \times vraisemblance$
- L'exemple traité ici sera utilisé pour illustrer dans le chapitre 2 le cas à la sollicitation avec une bonne connaissance des sources (*binomial / beta*).

### 0.24.11 Le réseau Bayésien de l'Exemple météo

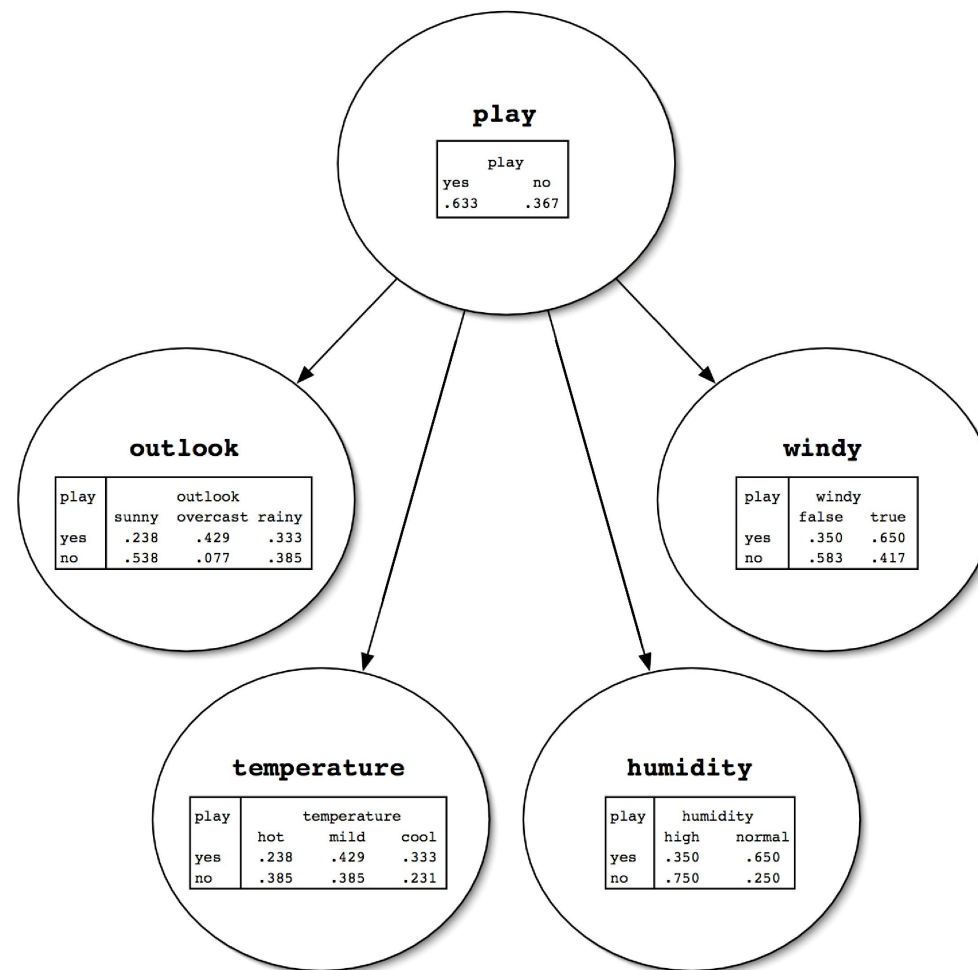


FIGURE 6 – BN pour les données météo avec un seul parent

Et un BN plus complexe :

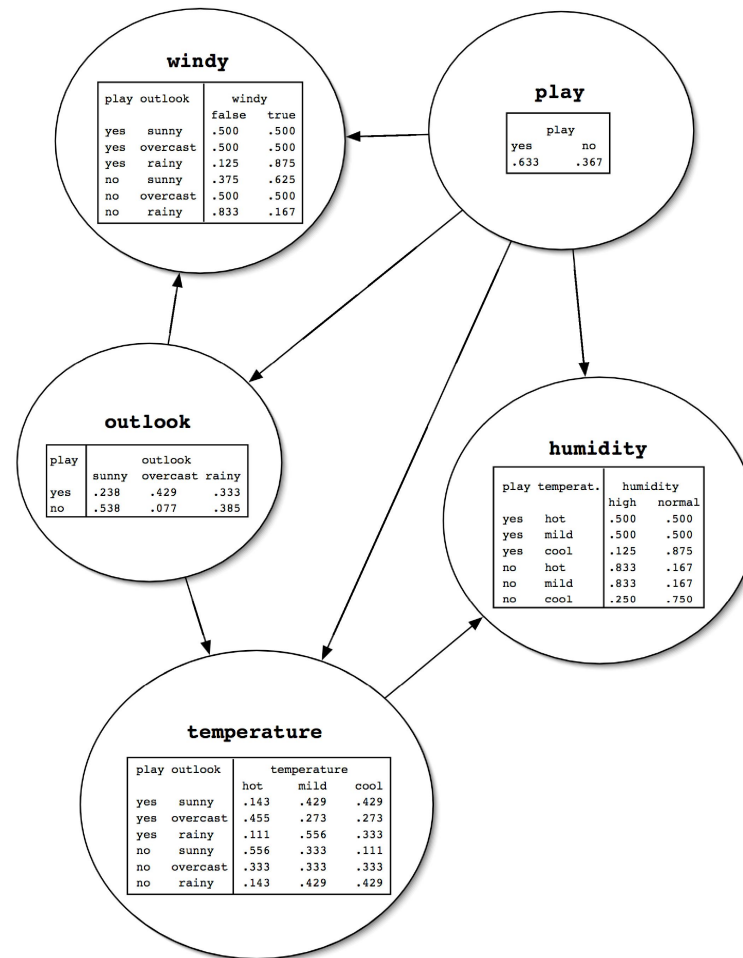


FIGURE 7 – BN pour les données météo avec plus d'un parent

## Exemple de prédiction à l'aide du BN (météo) :

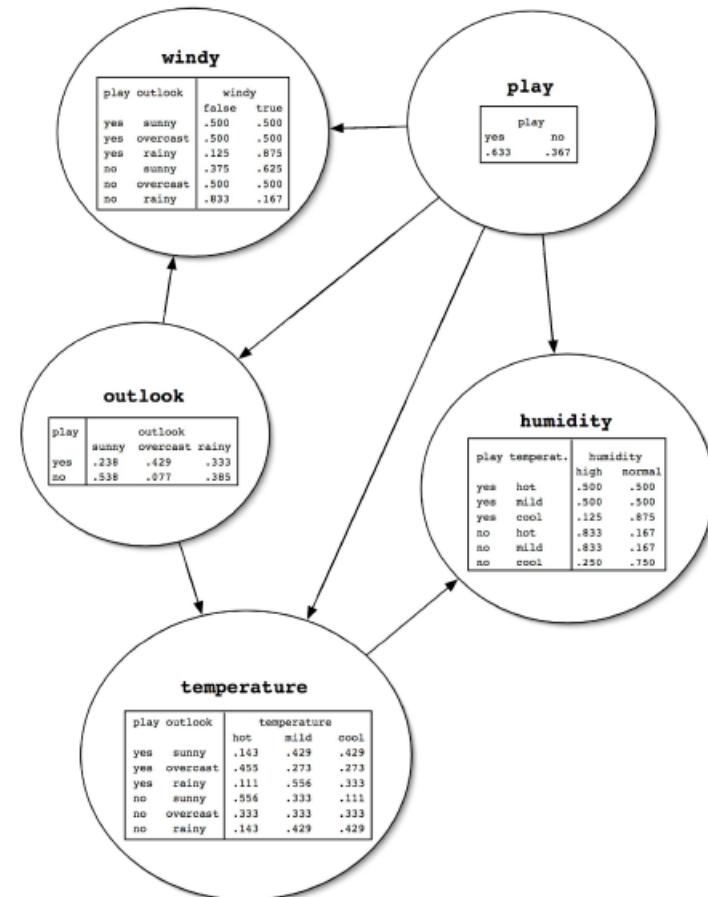
$Pr[play = no \mid E] = ?$  pour la nouvelle instance  $E$  :

$E$  :  $outlook = rainy, temp = cool, humidity = high, windy = true$

- $Pr[play = no] = 0.367$  (proba a priori)
- $Pr[outlook = rainy \mid play = no] = 0.385$
- $Pr[Temp = cool \mid play = no, outlook = rainy] = 0.429$
- $Pr[humidity = high \mid play = no, Temp = cool] = 0.250$
- $Pr[windy = true \mid play = no, outlook = rainy] = 0.167$

$$\Rightarrow = 0.367 * 0.385 * 0.429 * 0.250 * 0.167 = 0.0025$$

➡ A normaliser





### 0.24.12 BN : exemple RATP

*Étude du système de ventilation et de dés-enfumage des lignes souterraines du Métro (RATP)*

[J. Bensoussan & al, 2008, RATP/CGS].

#### **Paramètres de dépendance (expertise) :**

- Qualité de la maintenance (formation ... )
- Performance de détection par l'exploitant
- Saison (température)
- Système technique en soi (ventilateur, démarreur, alimentation ...)

#### **Cahier des charges :**

La disponibilité doit être supérieure à 97,5% pour respecter le niveau de sécurité requis.

☞ Au fait, d'où viennent les expertises ?

Prenons un exemple de mécanique et de démarrage de voiture :

P. Ex. :  $P(\text{démarrage} | \text{batterie faible}) = 0.5$  mais  $P(\text{batterie faible} | \text{démarrage}) = 0.2$

- Il y a deux méthodes pour étudier ce types de problèmes :

1. A partir de dires d'experts : *expertise* → *modélisation* → *quantification* → *analyse*

2. A partir de données de REX (Data Mining)

*récupération de données* → *extraction et analyse* → *modélisation et exploitation*

☞ **Mieux** : utiliser (2) pour peaufiner (1).

- Dans le problème de dés-enfumage :

1. Pour la modélisation experte → **SYSTEME DE DESENFUMAGE**

2. Pour l'extraction de données → **EXPLOITATION GENERALE DU RESEAU**

## Définition

Disponibilité : mesure de performance d'un équipement ou d'un système.

$$d = \frac{\text{Durée totale de fonctionnement}}{\text{Durée totale d'exploitation}}$$

- Prend généralement en compte :
  - Les temps de réglage,
  - La durée des pannes,
  - Le temps d'intervention des mainteneurs,
  - La durée des réparations,
  - Les temps logistiques.

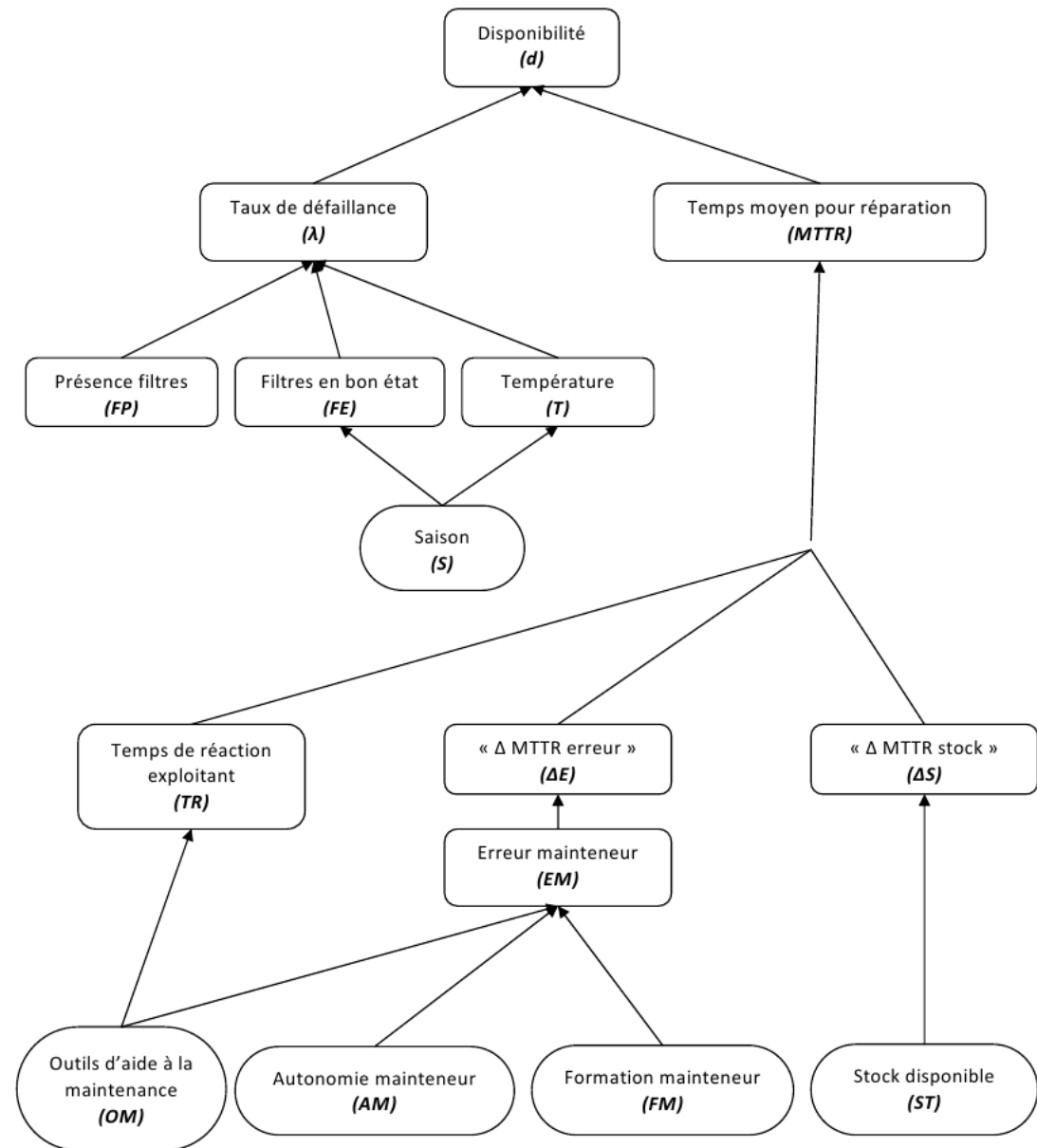
**Problème** : identifier les paramètres pouvant améliorer la **disponibilité** des ventilateurs des tunnels à partir de dires d'experts sur le système de dés-enfumage.

- paramètres techniques
- paramètres liés au comportement humain
- L'étude des paramètres conduit à l'identification des variables et modalités ci-contre ➔ :

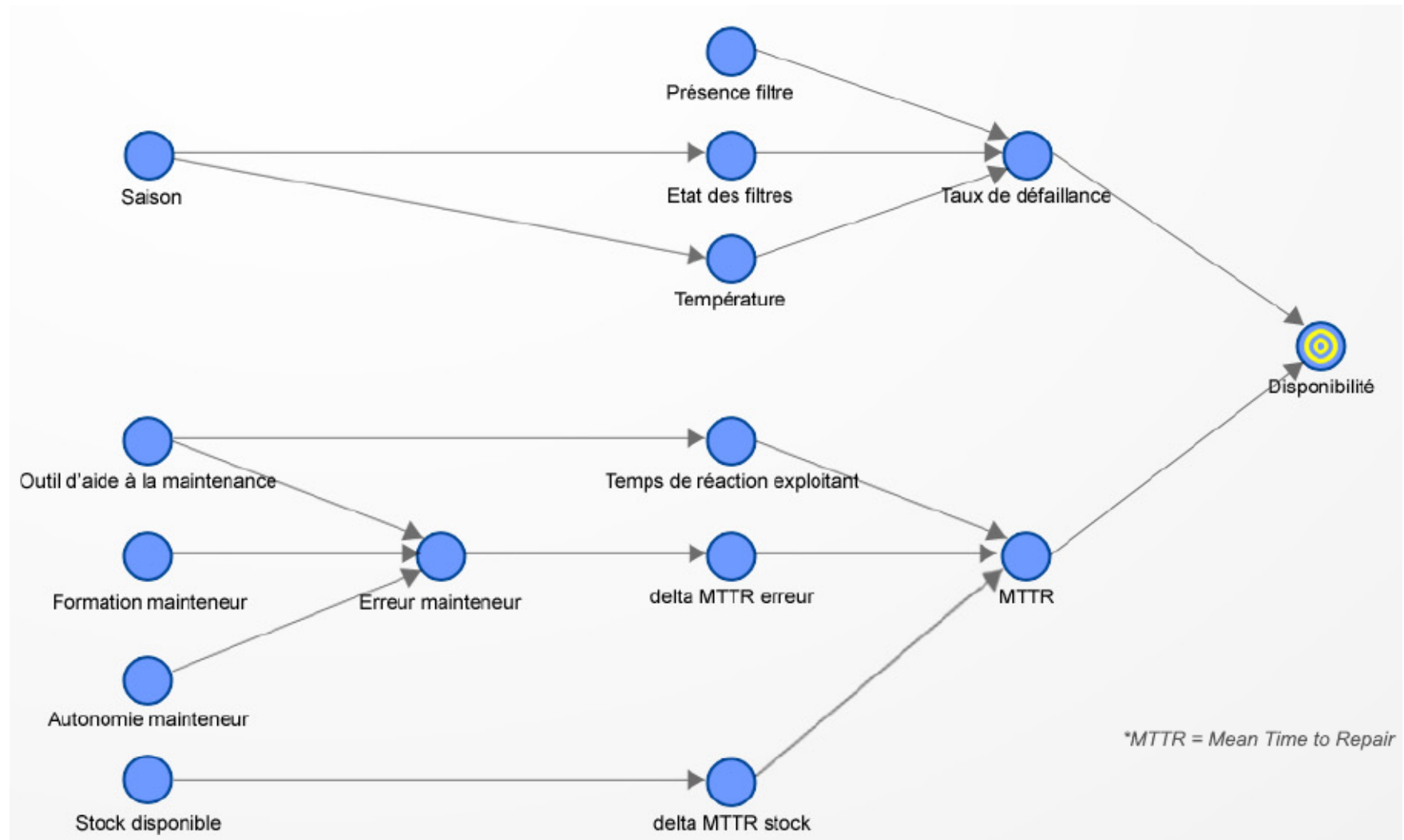
Nom de la variable	Type de variable
Disponibilité (d)	Continue [0 ; 1], discrétisée à 10 intervalles
Mean Time to Repair (MTTR)	Continue [0 ; 3072], discrétisée à 3 intervalles
Taux de défaillance ( $\lambda$ )	Continue [0 ; 0,01], discrétisée à 10 intervalles
Présence filtres (FP)	Discrète : OUI/NON
Filtres en bon état (FE)	Discrète : OUI/NON
Température (T)	Continue [0 ; 35], discrétisée à 3 intervalles
Temps de réaction exploitant (TR)	Discrète : 2 ; 8 ; 24
Outils d'aide à la maintenance (OM)	Discrète : OUI/NON
Autonomie mainteneur (AM)	Discrète : OUI/NON
Erreur mainteneur (FM)	Discrète : OUI/NON
Formation mainteneur (FM)	Discrète : OUI/NON
Stock disponible (ST)	Discrète : OUI/NON
Saison (S)	Discrète : Printemps, Été, Automne, Hiver

**Le réseau Bayésien obtenu grâce aux experts du système.**

- Il est souvent possible d'obtenir un BN à partir des données ReX.
- Une première analyse des données ReX + le réseau Bayésien donne un **taux de disponibilité de 92%**.
- Pour aller plus loin, il faut étudier les données (e.g. analyser les corrélations, vérifier les hypothèses de Bayes, etc).



- Le même BN en plus grand (et à l'horizontal) :



## Exploitation du BN :

- Sur le BN, on remarque une forte corrélation au niveau de

*Saison  $\rightarrow$  Température  $\rightarrow$  Taux de défaillance.*

- Un autre constat (voir page précédente) :

Par l'analyse de la variable cible, on remarque la prépondérance du noeud "stock disponible" en terme d'apport d'information sur la connaissance de la variable cible (= disponibilité).

→ **Importance de la logistique.**

- Une analyse plus détaillée du BN (par BayesiaLab) permet de connaître les variables qui maximisent la valeur moyenne de la variable cible :

→ Taux de défaillance  $\leq 0,001$  et  $\Delta MTTR_{stock} \sim 0$

→ La variable cible monte à 0,95%.

## Autres exploitation des données ReX (hormis le BN) :

- D'autres exploitations des données ReX (rapport d'incident, BD Clients, ...) permettent d'améliorer les résultats

→ P. ex., trouver des corrélations entre les noeuds du BN à l'aide des données ReX.

☞ Une étude détaillée de ces données a montré une importance non négligeable du **vandalisme** (cause de certains incidents relativement fréquents).

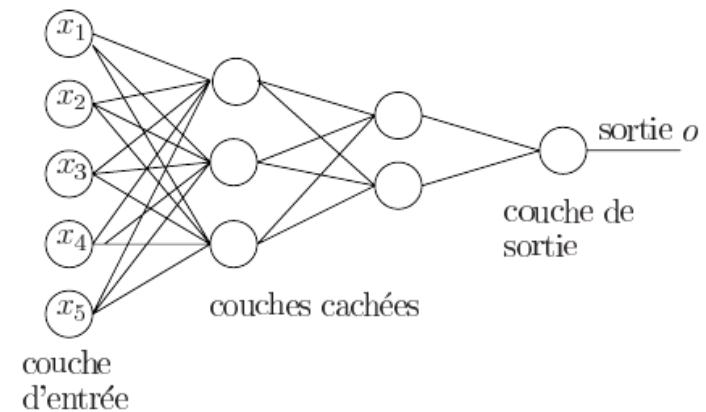
- Étude poursuivie ..... pour atteindre l'objectif.

☞ Secret et résultats confidentiels .... normal.



### 0.24.13 RNs et Perceptron multi couches

- 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.

→ Voir exemple manuscrit plus loin.

### 0.24.14 Méthodes à base de noyaux

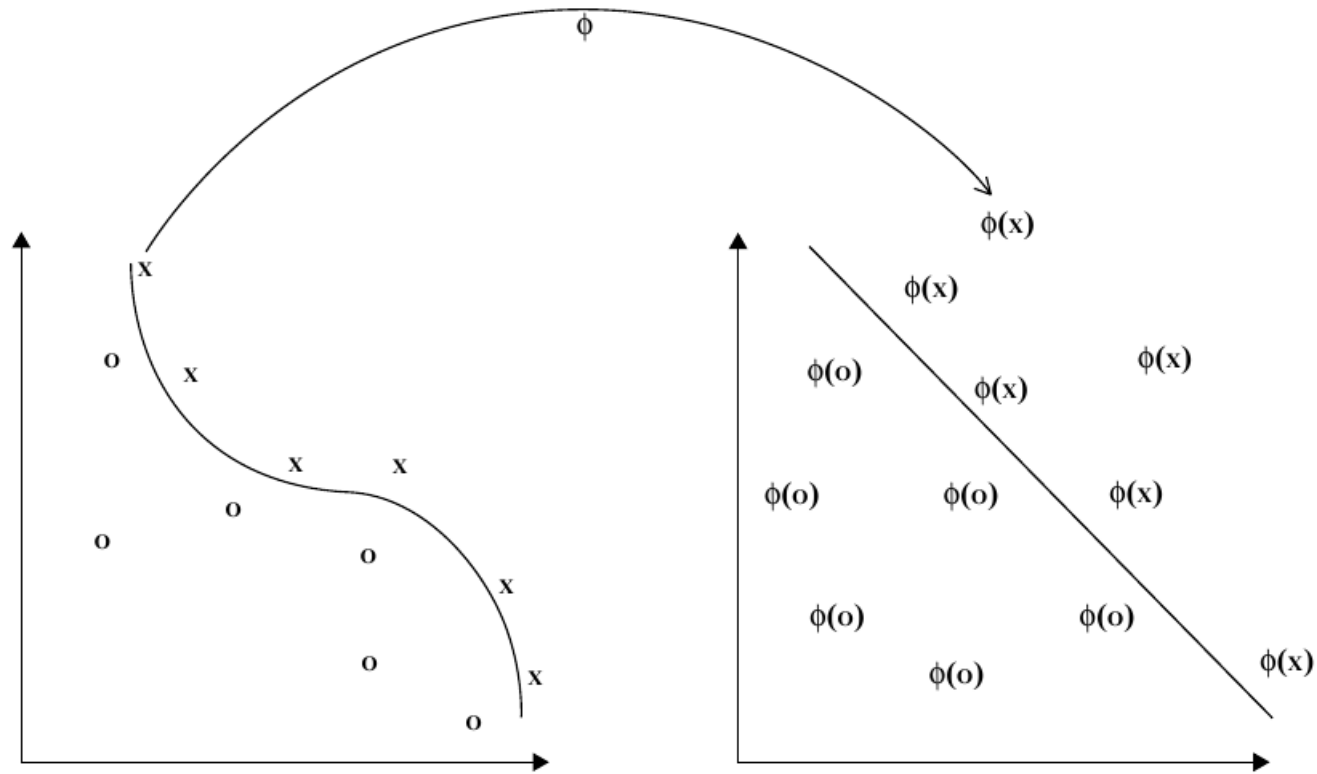
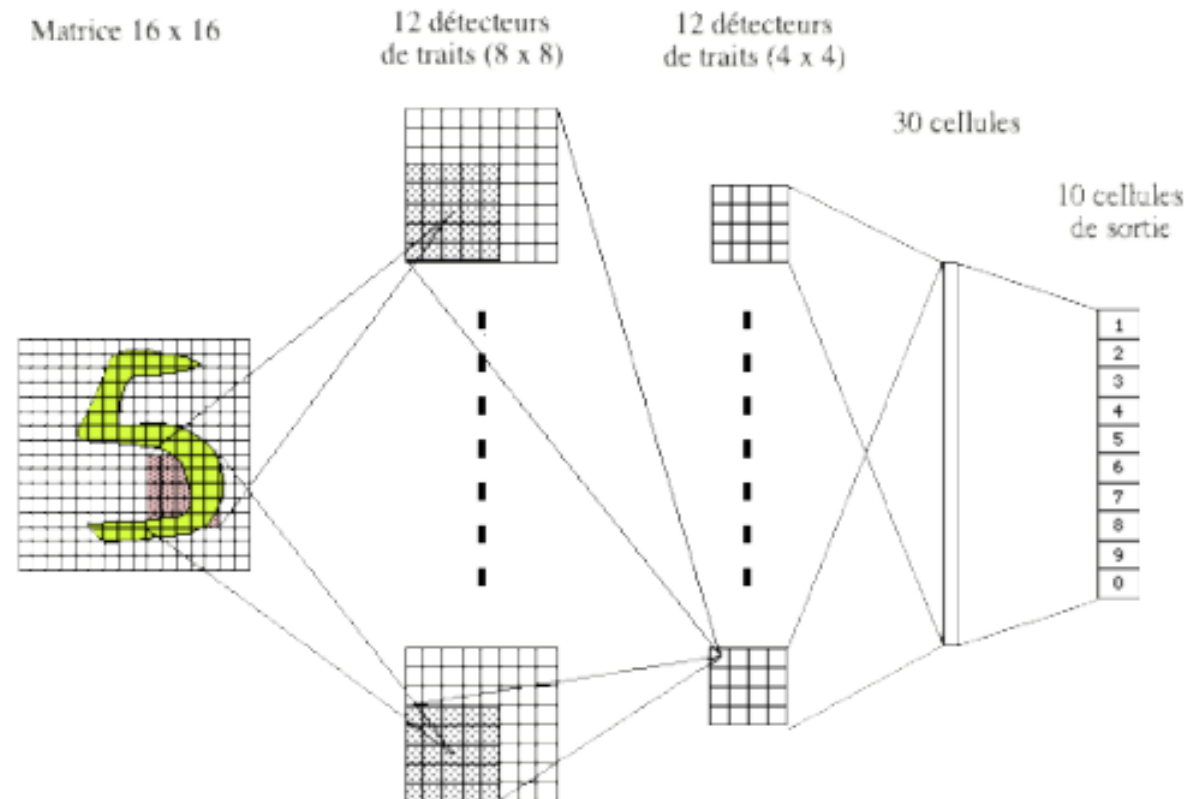


FIGURE 8 – La fonction  $\phi$  envoie les données dans un EdC où les motifs non-linéaires (d'origine) paraissent linéaires. Le noyau calcule ensuite des produits matriciels dans EdC directement àpd des instances d'origine.

Un exemple :

*... La reconnaissance de chiffres manuscrits par réseaux de neurones (ATT Bell labs, 1993)*



## 0.24.15 SVM

**Rappel méthode noyau (kernel) :** soit un ensemble de données non linéairement séparables

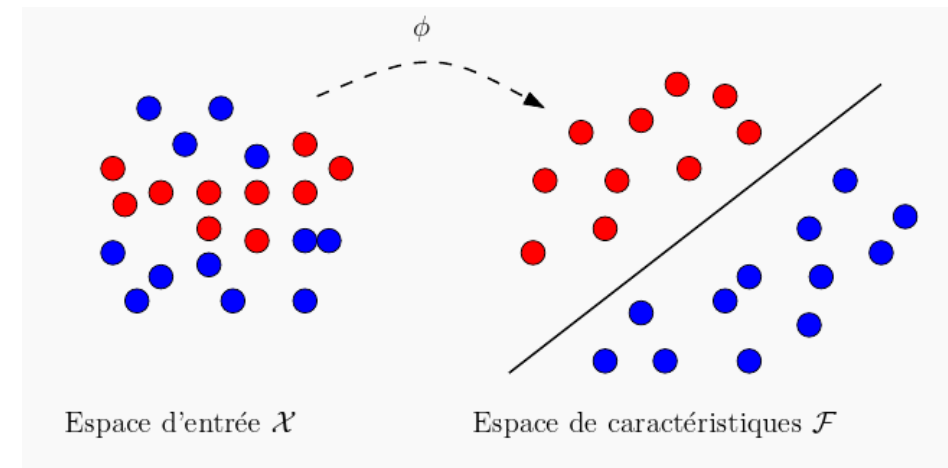
$$S_n : \{(x_1, y_1), \dots, (x_n, y_n)\} \quad y_i : \text{classe de } x_i$$

- Choisir une transformation non linéaire  $\phi$

$$\phi : \mathcal{X} \rightarrow \mathcal{F}$$

$$x \rightarrow \phi(x)$$

où  $\mathcal{F}$  est un espace vectoriel appelé *espace de caractéristiques* (Feature space).



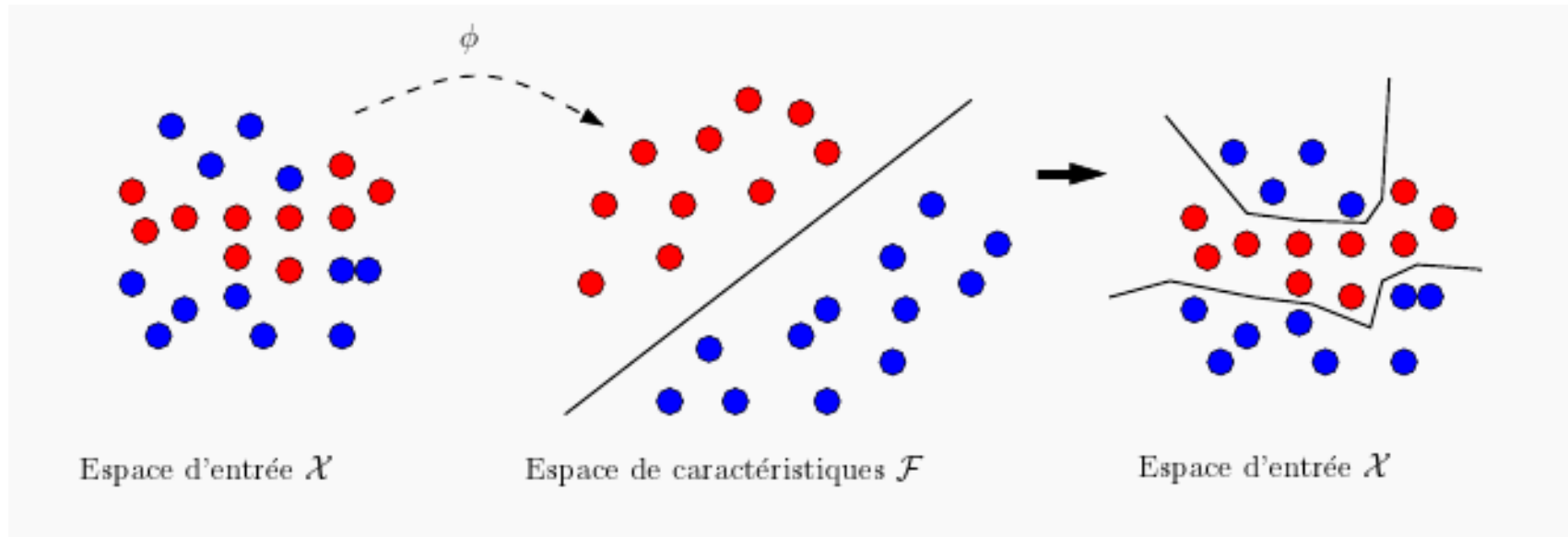
- Trouver un classifieur linéaire (i.e. un hyperplan séparateur) dans  $\mathcal{F}$  pour classifier

$$\{(\phi(x_1), y_1), \dots, (\phi(x_n), y_n)\}$$

- Procéder à une classification linéaire dans l'espace de caractéristiques.

- Implanter ensuite ce classifieur linéaire dans  $H$  (espace des hypothèses) :

$$h(x) = \sum_{i=1, \dots, n} \alpha_i \langle \phi(x_i), \phi(x) \rangle + b$$



## Idée de SVM :

- Il peut exister une infinité de plans de séparation linéaires (selon  $w$ ) :

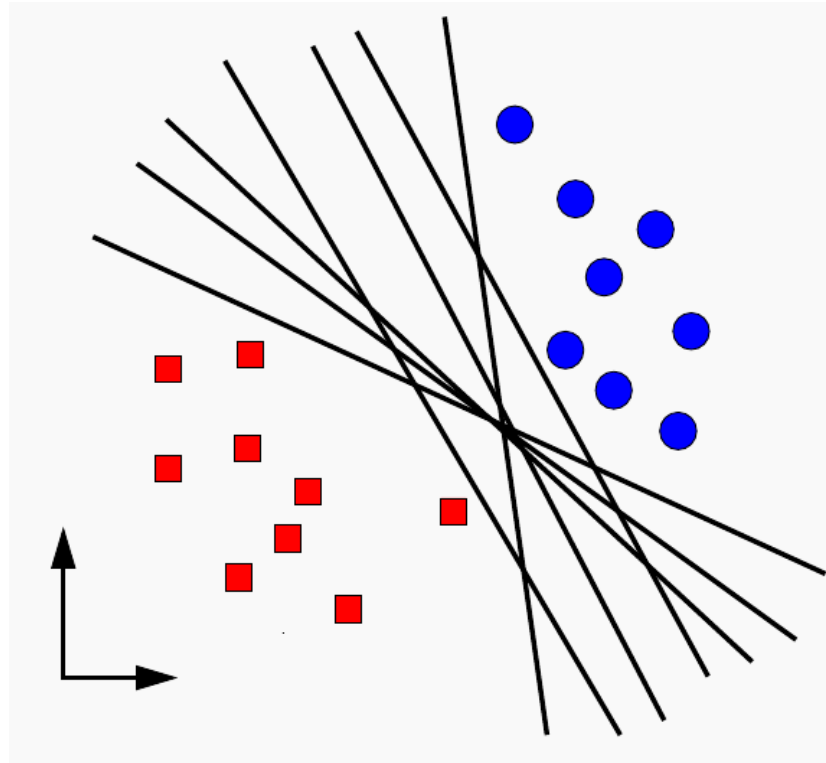


FIGURE 9 – Le problème du choix du plan linéaire

- On choisira celui qui a la plus grande marge

- Exemple : 2 plans de séparations

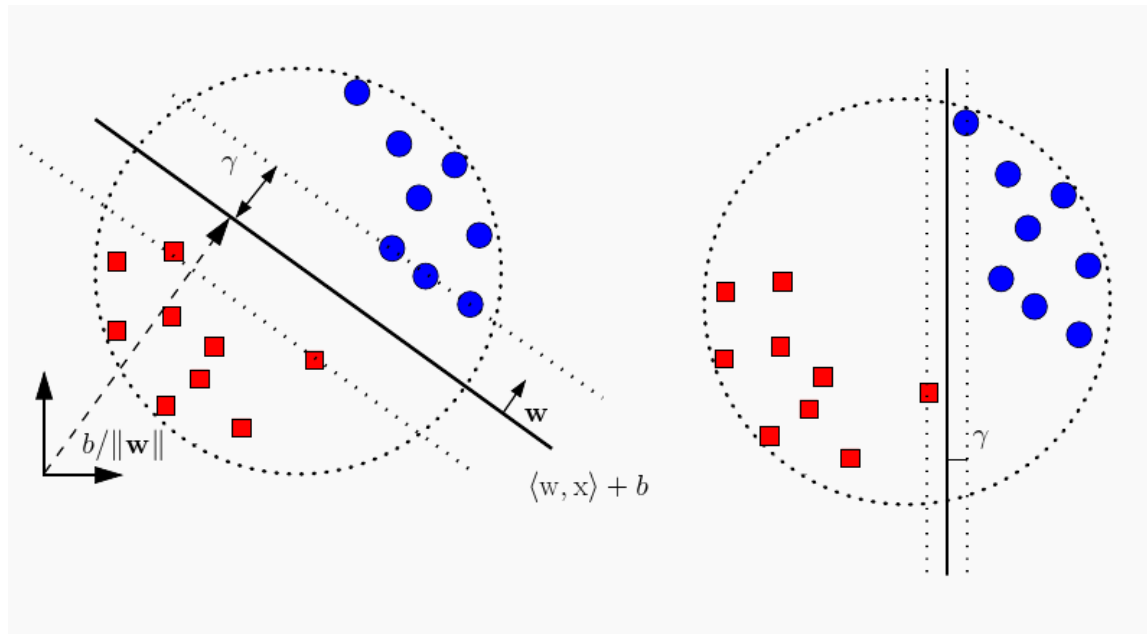
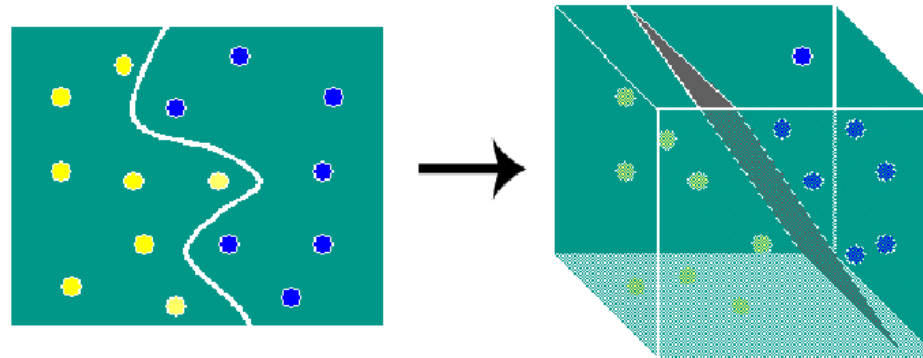


FIGURE 10 – La marge à gauche est plus grande que celle de droite (donnée inchangées)

- La fonction linéaire  $h$  recherchée :  $h(x) = w.x + b$ .
  - La surface de séparation correspondant à la fonction linéaire  $h$  est l'hyperplan  $w.x + b = 0$
  - Elle est valide si  $\forall i, y_i h(x_i) \geq 0$

## SVM non linéaire

- On peut utiliser un noyau pour trouver une projection dans un espace séparable.
  - La projection a lieu dans l'espace des caractéristiques (attribut des données).



→ On constate que la projection (re-description) peut conduire à un espace de séparation *linéaire* mais de **plus grande dimension**.

☞ **On peut utiliser** les noyaux même dans les cas linéaires pour simplifier la classification

→ La projection a souvent lieu dans  $\mathcal{R}$  (facilité des calculs).



## 0.25 Clustering

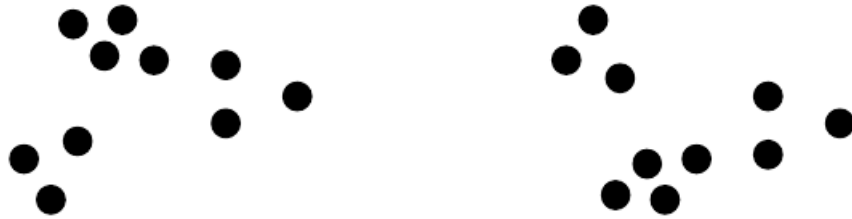
- **Technique non supervisée**

Versus la classification où l'on connaît les classes par avance

↳ Dans Clustering : les clusters (classes) sont inconnus a priori.

- Clustering = Regrouper les données en plusieurs groupes (=clusters)
- Chaque groupe doit être **homogène** et se distinguer des autres groupes.
  - ↳ Minimiser les intra-distances, maximiser les inter-distances
  - ↳ Il arrive d'avoir des groupes qui se recouvrent → probabilités
- Autre variant : Techniques Hiérarchiques → Raffinement par niveau

- La notion de Cluster est ambiguë :



Combien de clusters?



Six clusters



Deux clusters



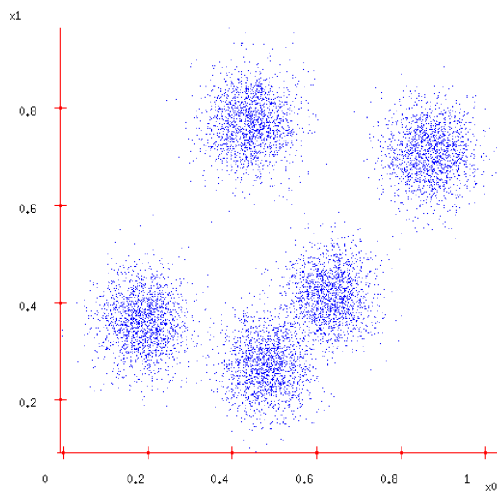
Quatre clusters

### 0.25.1 *Kmeans*

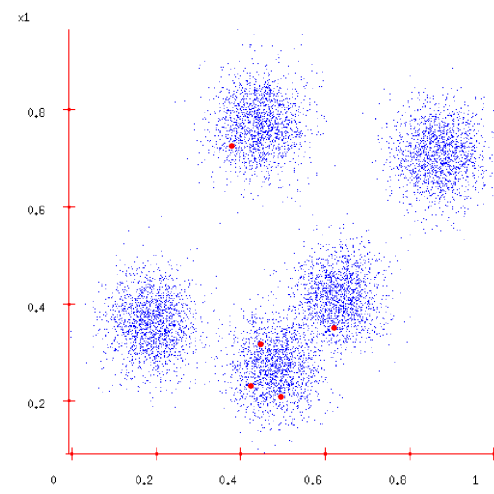
**Principe général** : clustering Itératif à base d'instances

- On souhaite partitionner un ensemble  $S$  en  $k \geq 2$  clusters.
- Soient  $C_1, C_2, \dots, C_k$  des clusters avec les centres  $c_1, c_2, \dots, c_k$  respectivement.
- On définit la **dispersion intra-cluster** par 
$$\sum_{i=1}^k \sum_{x \in C_k} d(c_k, x)$$
- Le but est de trouver un clustering avec une dispersion intra-cluster minimale.

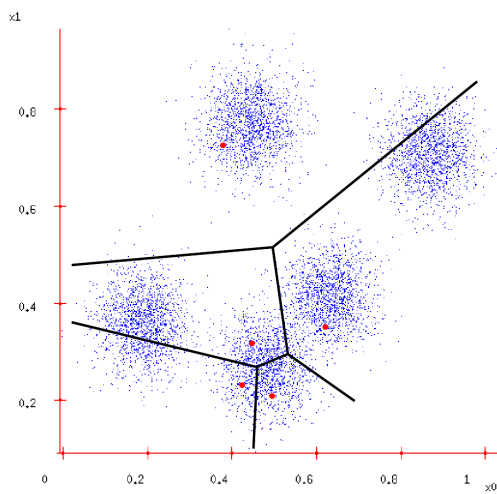
### 0.25.1.1 Illustration de K-means



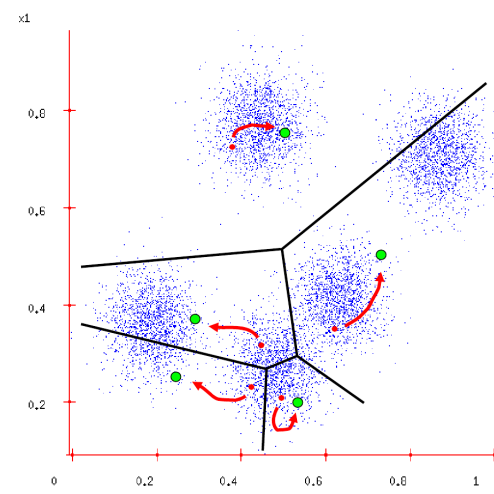
*Points d'origine*



*Initialisation*

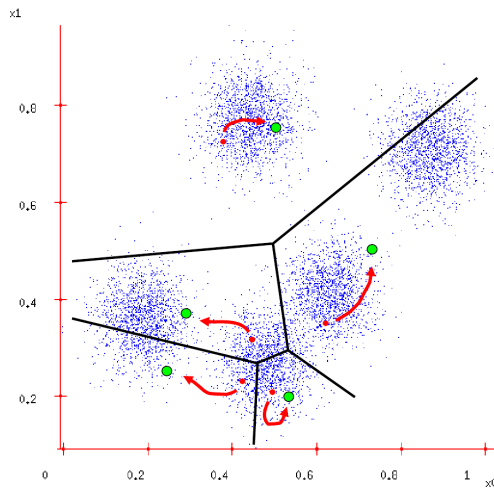


*Etape 1*

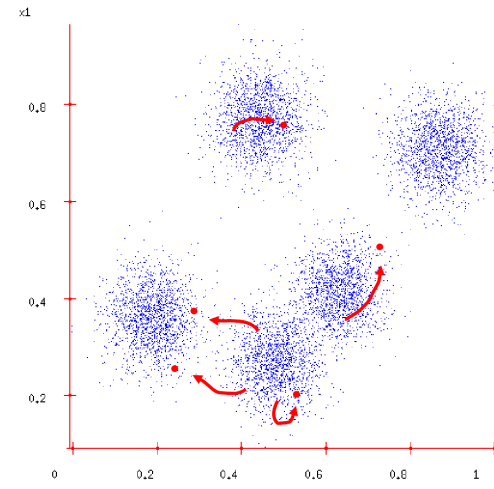


*Etape 2*

## Illustration de K-means (suite) :

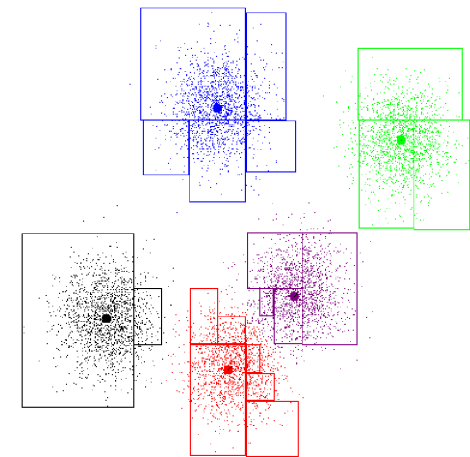


*Etape 3*



*Etape 4 .....*

- On fixe  $k = 5$  et les centroïdes aléatoires initiaux ;
- Chaque instance "trouve" le Centre le plus "proche" ;
- Chaque Centre "trouve" le Centroïde des points qui lui sont devenus "proches"
- Et s'y déplace !
- .... Jusqu'à ne plus se déplacer !



*Etape finale*

## 0.25.2 *Clustering Probabiliste*

### 0.25.3 *La méthode EM*

#### 0.25.3.1 Une application d'EM : données manquantes

- Retrouver des valeurs manquantes d'une BD (avec 4 valeurs  $x_1, x_2, x_3, x_4$ ).
- On connaît  $\frac{1}{2}$  des valeurs d'une BD (uni-variable, attribut continue, loi uniforme).
  - ➡ L'autre moitié ( $x_3, x_4$ ) manquante ou corrompue.
- Calculer les valeurs manquantes par EM (meilleures estimations)
- Hypothèse : la variance est unitaire (=1).
- **Etape E** : estimer les variables manquantes ( $\mu$  puis  $x_3$  et  $x_4$ ).
- **Etape M** : ré estimer les paramètres de la distribution pour maximiser la vraisemblance des données, sachant la BD (complétée)

### Illustration par un exemple (trivial) :

- Initialisation :  $BD = [4, 10, ?, ?]$  (ou  $\sigma = 1$  pour une loi *Normale*),
- On choisit  $\mu = 0$ ,  $\rightarrow$  Nouvelle BD :  $[4, 10, 0, 0]$
- Nouvelle  $\mu = 3.5$ ,  $\rightarrow$  Nouvelle BD :  $[4, 10, 3.5, 3.5]$
- Nouvelle  $\mu = 5.25$ ,  $\rightarrow$  Nouvelle BD :  $[4, 10, 5.25, 5.25]$
- Nouvelle  $\mu = 6.125$ ,  $\rightarrow$  Nouvelle BD :  $[4, 10, 6.125, 6.125]$
- Nouvelle  $\mu = 6.5625$ ,  $\rightarrow$  Nouvelle BD :  $[4, 10, 6.5625, 6.5625]$
- Nouvelle  $\mu = 6.7825$ ,  $\rightarrow$  Nouvelle BD :  $[4, 10, 6.7825, 6.7825]$
- Nouvelle  $\mu = 6.890625$ , ....
- .....
- Nouvelle  $\mu = 7$ , Nouvelle BD :  $[4, 10, 7, 7]$   $\rightarrow$  **Ne change plus.**

☞ Bien entendu, on peut générer les deux valeurs via  $\mu = 7$  directement mais ceci est un exemple trivial de EM!

### 0.25.4 Classification 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)

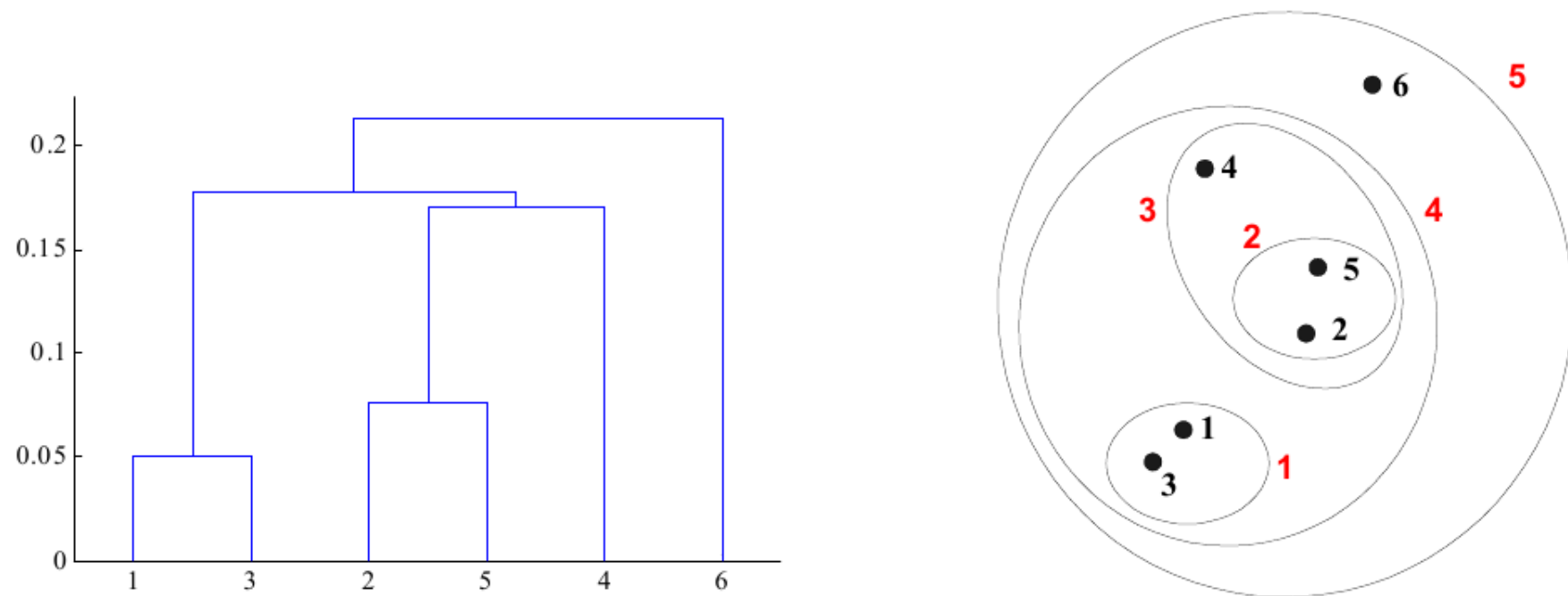


FIGURE 11 – Exemple de classifications hiérarchiques



### 0.25.5 Clustering : un autre exemple

- 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. Transac°	Op°/mois	Sexe	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 (sur quelques clients)

- Pour un compte de type "Bourse", la banque prête de l'argent liquide au client.
  - On veut trouver des "motifs" dans ces données.
  - Supposons qu'un algorithme de classification non supervisé a produit 3 clusters.
- ⇒ Les règles représentatives de ces clusters : ../..

*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)*

→ Couverture = **support**, justesse = **confiance**

- **Cluster 1** : logique : 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 épargne-enfant.

## 0.26 Evaluation

☞ **Tout apprentissage artificiel doit être validé.**

**Quelques éléments importants :**

- On demande à un bon modèle de ne pas être juste bon sur un (seul ?) ensemble d'apprentissage mais sur tous les ensembles concernés (qu'on appellera  $L_s$  : learning sets).

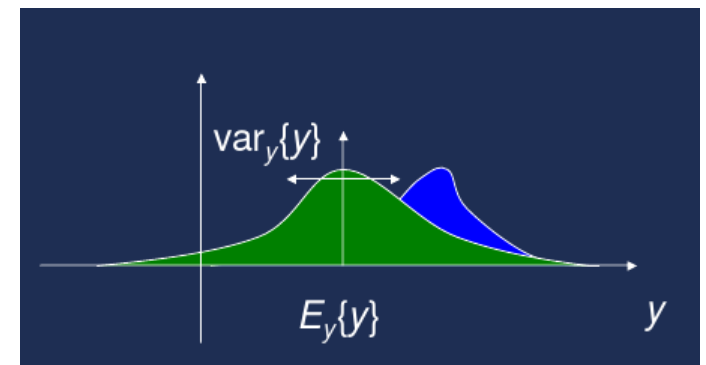
- L'expression de l'erreur :  $E = \mathbb{E}_y(y - \hat{y})^2 = \mathbb{E}_{LS}\{\mathbb{E}_y\{(y - \hat{y})^2\}\}$  que l'on développe :

$$E = \mathbb{E}_y\{(y - \mathbb{E}_y\{y\})^2\} + \mathbb{E}_{LS}\{(\mathbb{E}_y\{y\} - \hat{y})^2\}$$

Le premier terme est l'erreur résiduelle

$$\text{var}_y\{y\} = \varepsilon$$

qui est le minimum atteignable de (toute) l'erreur  $E$



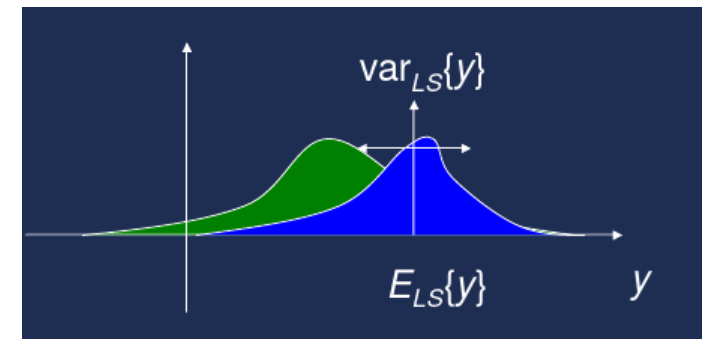
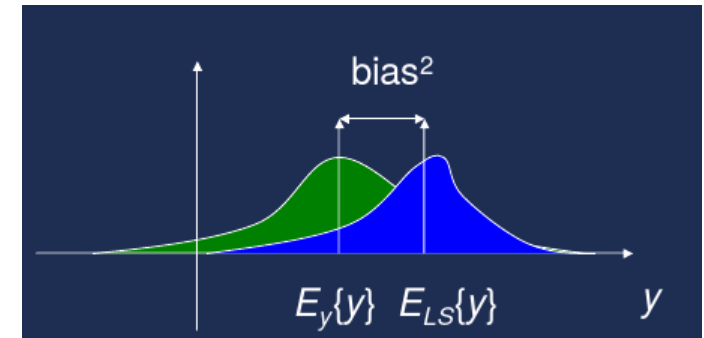
- Si on développe le 2e terme de  $E$  ci-dessus

$(\mathbb{E}_{LS}\{(\mathbb{E}_y\{y\} - \hat{y})^2\})$ , on obtient :

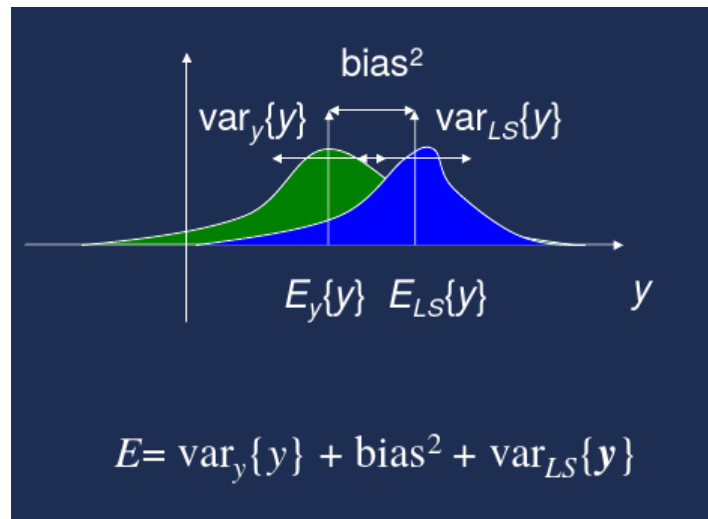
$$(\mathbb{E}_y\{y\} - \mathbb{E}_{LS}\{\hat{y}\})^2 + \mathbb{E}_{LS}\{(\hat{y} - \mathbb{E}_{LS}\{\hat{y}\})^2\}$$

dont la 1e composante est le *biais*<sup>2</sup> = l'erreur entre le modèle obtenu (ML) et le modèle moyen (connu).

- Enfin le terme  $\mathbb{E}_{LS}\{(\hat{y} - \mathbb{E}_{LS}\{\hat{y}\})^2\} = var_{LS}\{\hat{y}\}$   
 $= var_{LS}\{\hat{y}\} = var_{LS}\{y\}$  (sachant  $\mathbb{E}_y\{y\} = \mathbb{E}_y\{\hat{y}\}$ )  
 $=$  une estimation de la variance  
 $=$  une mesure de l'Overfitting éventuel.



- Par conséquent :  $E = \varepsilon + \text{biais}^2 + \text{var}_{LS}(y)$



**Notons** : la variance  $\varepsilon = \text{var}_y(y)$  est le **bruit** (noise, Vars. cachées, ...),

$\text{biais}^2$  est l'erreur entre l'estimation et les observations (la B.D.)

$\text{var}_{LS}(y)$  est la variance de l'estimation-même (app. faits sur  $\neq$  BDs.)

→ Vers les méta méthodes "**Ensemble Learning**".

## 0.27 Annexes

### 0.27.1 *Un exemple de prédiction de fiabilité chez Renault*

- Données de défaillance sur un composant mécanique dans les véhicules Renault.
- **Buts :**
  1. identifier et hiérarchiser l'influence des variables explicatives sur le taux de défaillance moyenne du composant.
    - 2 types de variables :
      - caractéristique technique du véhicule ("type de voiture", "puissance moteur", etc.)
      - Les conditions d'utilisation ("pays", "type utilisateur", etc.)
  2. Déterminer la corrélation entre les variables explicatives.

## **Approches complémentaires utilisées : Régression Logistique & BN.**

- Le modèle permet de de construire une méthodologie d'exploitation des données ReX.

### **0.27.1.1 Les données ReX**

- Récupérées pendant une certaine période de garanti (par Renault)
- Les informations collectées par les garagistes :
  - ID du véhicule défaillant (Id unique, date de fabrication,..),
  - Caractéristiques techniques (puissance, type d'équipement, etc.),
  - Données de son utilisation (pays, kilométrage, etc)
  - Les données de maintenance.
- La variable cible : **Taux de défaillance.**

- Les variables explicatives (qui influencent Taux) :
  - Type du véhicule :  $V = \{V1, V2, \dots V5\}$
  - Pays où l'incident a été enregistré :  $Pays = \{P1, \dots P10\}$
  - Type de boîte de vitesse :  $B = \{BVA(\text{automatique}), BVM(\text{manuel})\}$
  - Présence de climatisation :  $C = \{CA(\text{présent}), DA(\text{absent})\}$
  - Type d'utilisateur :  $U = \{Soc(\text{société}), Par(\text{particulier})\}$
  - Puissance du moteur :  $P = \{pw1, \dots pw4\}$
- Exemple de données :

V	P	C	U	B	Pays	Nb. véhic.	Taux
V1	pw1	DA	Soc	BVA	P1	5857	0,022
V2	pw1	DA	Soc	BVA	P1	1500	0
V3	pw2	CA	Par	BVM	P3	750	0,015
...	...	...	...	...	...	...	



### 0.27.1.2 Application de la régression logistique

- Régression Logistique : méthode intéressante dans le cas de classes binaires.
- Si pour chaque donnée, les variables explicatives sont  $X = x_1 \dots x_p$

et la classe  $Y \in \{0, 1\}$  (défaillant ou non), alors on note

$\pi(x) = Pr(Y = 1|X = x)$  la probabilité d'être défaillant sachant  $X$ .

- Le modèle de régression logistique est estimé par le *log vraisemblance*.

$$\log \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \alpha_0 + \alpha_1 x_1 + \dots + \alpha_p x_p$$

## Interprétation des résultats de la Régression Logistique

- Dans le cas des classes binaires, on peut calculer *Odds Ratio* = *OR* (le rapport des cotes) qui est rapport entre les deux modalités :

$$OR(X = 1 \text{ vs } X = 0) = \frac{\left( \frac{\pi(1)}{1-\pi(1)} \right)}{\left( \frac{\pi(0)}{1-\pi(0)} \right)}$$

Rappel :  $\log(odds) = w_0$  (intercepte)

### Exemple :

- Pour la variable explicative  $X = \text{sexe} \in \{fille, garçon\}$ ,
  - $X = 1$  veut dire sexe=homme et  $X = 0$  veut dire sexe=femme.
- On calcule, pour une **variable cible (une certaine) maladie** (la classe) :
  - le ratio entre  $\pi(1)$  (être homme et être malade) et  $1 - \pi(1)$  (homme et non malade)
  - le ratio entre  $\pi(0)$  (être femme et être malade) et  $1 - \pi(0)$  (femme et non malade)
  - le ratio entre les deux (appelé OR)

### 0.27.1.3 Résultats de la Régression Logistique

- Sur l'ensemble de la BD, on obtient les résultats suivants :
  - 3 variables ont des effets statistiques significatifs sur le Taux (de défaillance) : "Type de Véhicule", "type utilisation" et "Pays" :

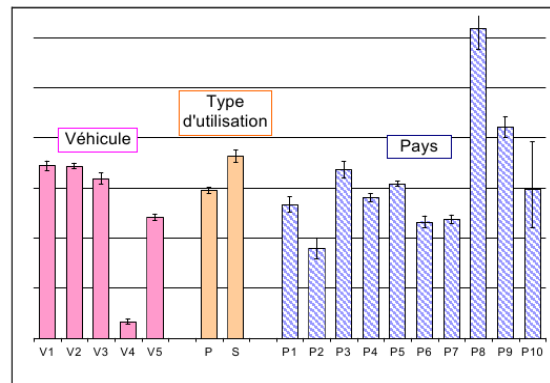


FIGURE 12 – les 3 variables significatives sur la cible

- Pour réaliser cette figure, on a calculé l'effet de chacune des 3 variables

- L'**interaction** entre "Véhicule" et "Type Utilisation" semble significative sur la défaillance.

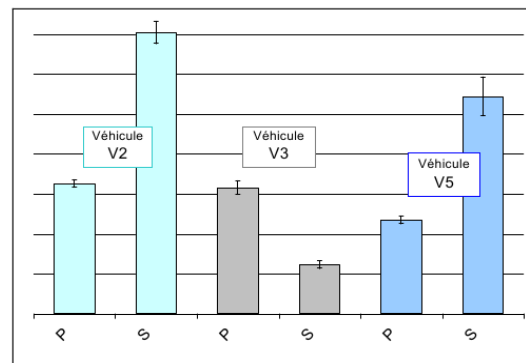
→ Elle montre l'effet du "Type Utilisation" sur la défaillance pour un "Véhicule" donné ( $P$  = "type utilisation= particulier",  $S$  = "type utilisation = société",  $V_i$  : différents types de "Véhicule") :

$$- OR(P \text{ vs. } S | V_3) = 2$$

$$- OR(P \text{ vs. } S | V_2) = 0.5$$

$$- OR(P \text{ vs. } S | V_5) = 0.5$$

→ On remarque (la figure ci-dessous) que  $S$  ("type utilisation = société") est pénalisant pour la défaillance pour les véhicules de type  $V_2$  et  $V_5$  tandis que le type d'utilisation "particulier" est pénalisant pour  $V_3$ .



- N.B. : les calculs ont également montré que le type de boîte (BVA, BVM) a un effet statistique sur la défaillance. Le type BVA n'existe que sur des variants particuliers (type véhicule  $V_5$ ,  $P$  : véhicule de particulier et  $CA$  : avec climatiseur).

→ Si on restreint les calculs à ces 3 sortes de véhicules, alors  $OR(BVA \text{ vs. } BVM) = 1.9$  montre que  $BVA$  est pénalisant pour la défaillance.

### **Avantages / Inconvénients de Régression logistique :**

Identification des variables significatives et leur hiérarchisation,

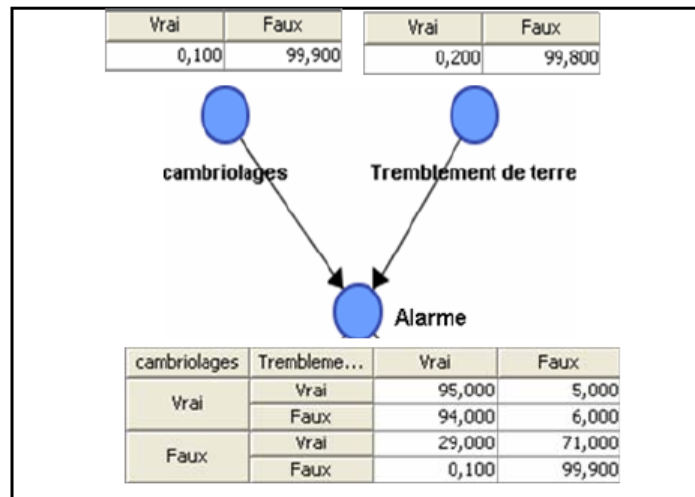
Traitement des variables continues (selon l'outil utilisé).

→ Par contre, l'interprétation des résultats (Odds Ratio par exemple) est quelque peu difficile (post traitement).

→ De même, si les variables sont fortement corrélées, il faut avoir recours aux techniques de sélection des variables (pré traitement).

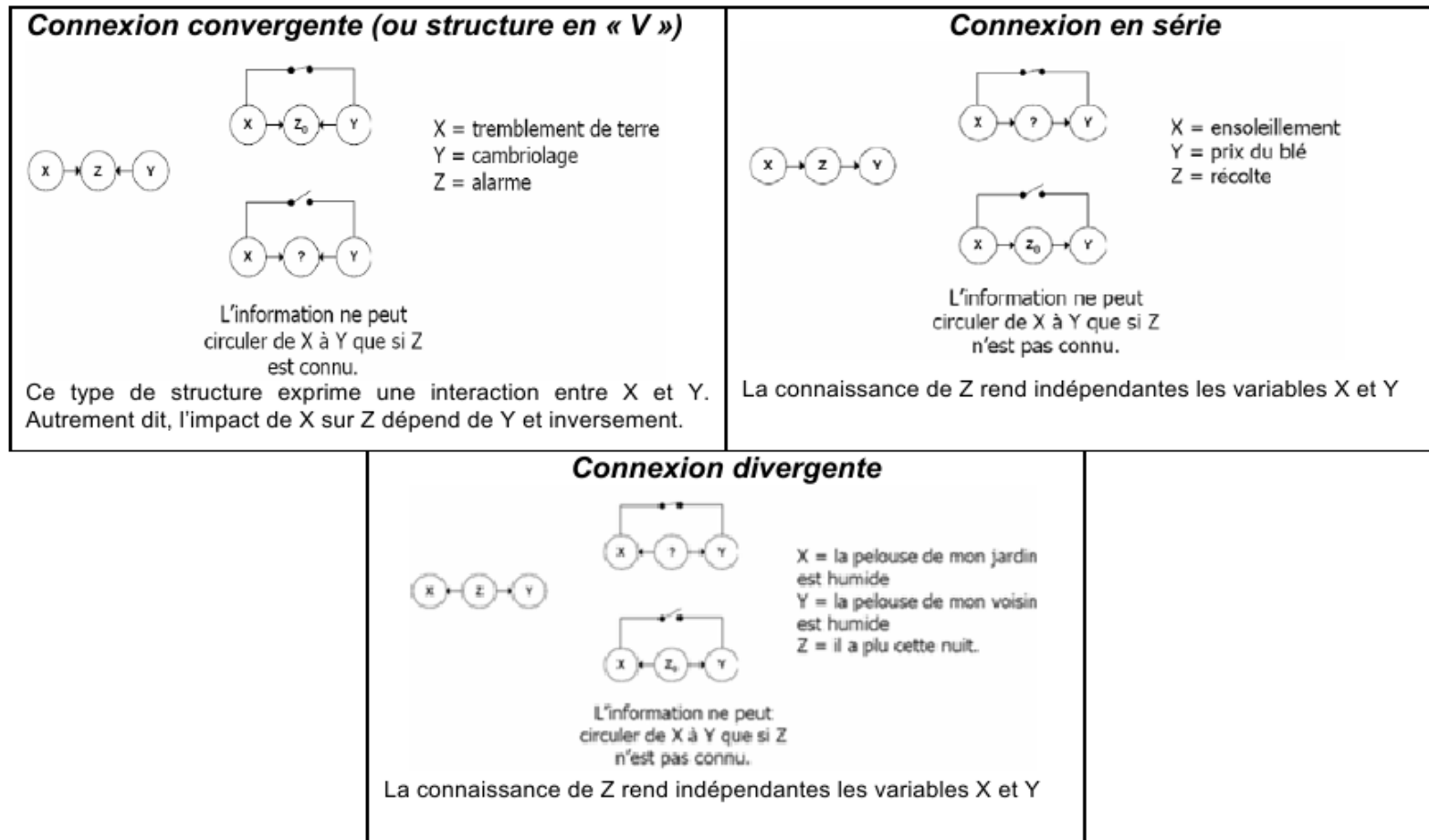
### 0.27.1.4 Application des BN

- Un BN est un graphe orienté qui exprime l'incertitude (proba) de la connaissance des relations entre les variables.
- Un exemple :



- Rappel : règle de Bayes.

## Règles d'interprétation des BN (important) : raisonnement sur un BN



"circulation d'info entre A - B " : A m'apprend quelque chose sur B.



Cette circulation ne suit pas le sens des flèches.

1. **Cas convergent** : si je ne sais pas qu'il y a "Alarme" (Z), je ne sais rien sur le "Tremblement" (X) ou sur le "Cambriolage" (Y).
  - A priori, pas de lien entre "Tremblement" (X) et "Cambriolage" (Y),
  - Mais, si je sais "Alarme" (A), je peux penser qu'il y a eu "Cambriolage" (Y),
  - **Circulation X-Y** : et si j'apprends "Tremblement" (X), **je suis rassuré** sur "Cambriolage" (Y)
2. **Cas Série** : si "Ensoleillement" (X), je sais "Récolte" (Z) et donc baisse du "Prix du blé" (Y).
  - Ce n'est pas une circulation d'info entre X-Y, on a suivi les flèches.
  - **Circulation X-Y** : Mais si je sais l'abondance de la "Récolte" (Z), connaître l'"Ensoleillement" (X) **ne m'apprend rien** sur "Prix du blé" (Y)
3. **Cas divergent** : si "ma pelouse est humide" (X), j'ai tendance à croire que "Pluie" (Z) et donc que "pelouse voisin humide" (Y).
  - Ce n'est pas une circulation d'info entre X-Y, on a raisonné (mais pas suivi les flèches!)
  - **Circulation X-Y** Par contre, si je sais "Pluie" (Z), je sais que "pelouse voisin humide" (Y) et savoir que "ma pelouse est humide" (X) **n'y change rien (n'apporte rien)** .



## Deux type d'apprentissages des BN :

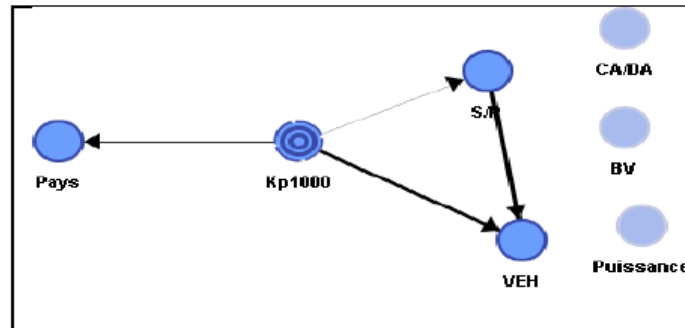
- **Supervisé** : construire un modèle (BN) qui permet d'expliquer la variable Cible par des variables explicatives significatives.

→ En d'autres termes, trouver les variables explicatives dont la connaissance rend la variable cible indépendante des autres variables.

- **Non supervisé** : exprime l'ensemble des relations entre toutes les variables (pas de notion de cible).

## Résultats pour la défaillance mécanique

- Résultat d'un apprentissage supervisé :

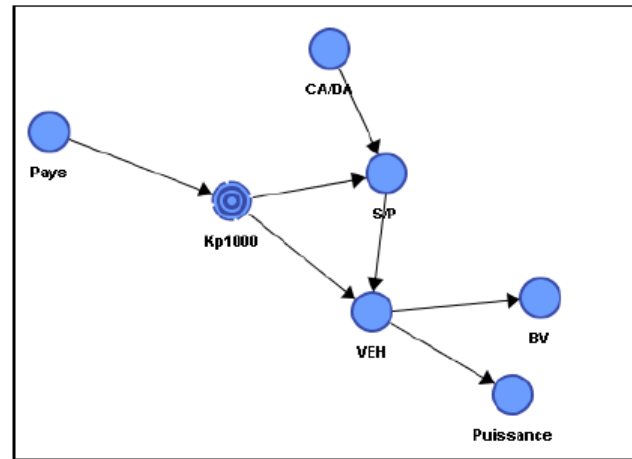


- On constate :

- L'influence des 3 variables sur la défaillance : "type véhicule", "pays", "type utilisation".
- La structure en V qui converge vers "type Véhicule" lie deux variable "défaillance" et "type d'utilisation" (S/P) et montre une interaction entre "type utilisation" et "Véhicule".
- C'est à dire : l'impacte de "type utilisation" sur le taux de "défaillance" dépend du "type Véhicule".

- Ces mêmes résultats ont été trouvés par la régression logistique précédente.
- On remarque que ce modèle ne détecte pas la relation entre la modalité "BVA" de la variable "type Boite" et la modalité "V5" de la variable "type véhicule".
  - La raison : **un BN établie les liens entre deux variables que si la majorité de leurs modalités sont en relation.**

## Cas apprentissage non supervisé :



- Ce modèle montre les relations entre les variables explicatives (et cible).
  - Par exemple, dans la structure divergente Veh/BV/Puissance :
    - ☞ il y a une corrélation entre BV et Puissance (sachant Véhicule).
    - ☞ Cette corrélation vient du "type du véhicule" : le véhicule "V1" est équipé d'une boîte BVA et d'une puissance Pw1.
- ➔ On trouve le même résultat dans l'apprentissage supervisé : les 3 variables explicatives les plus influentes sur la défaillance ("type utilisation", "type véhicule" et "pays") et l'interaction.

- Les BN constituent un outil intéressant pour la modélisation des données. Cependant, l'absence d'arc entre deux variables n'indique pas l'absence de lien entre certaines modalités des variables. En outre, les BNs sont insuffisants dans le cas de multiples variables cibles.

#### 0.27.1.5 Utilisation des deux méthodes

- On peut utiliser les **BN non supervisés pour identifier les corrélations entre les variables explicatives** et ainsi désigner les **moins corrélées**.
- Ensuite, utiliser les **BN supervisés pour trouver les variables les plus influentes** (sur la cible).
- Ensuite, utiliser la **régression logistique pour identifier les relations significatives entre les modalités** (point faible des BNs) et quantifier les effets de ces variables sur la variable cible ainsi que les interactions entre les variables explicatives.

- Cette démarche appliquée à la BD Renault, on obtient :
  - Le modèle non supervisé n'a pas retenu des corrélations significatives et on conserve donc l'ensemble des variables ;
- Le modèle supervisé permet d'identifier les variables "type d'utilisation", "type véhicule" et "pays" comme les plus significatives sur la cible ainsi que l'interaction de "type utilisation" et "type véhicule" ;
- La régression log permet de quantifier les variables significatives sur la cible (figure 12 de la subsection 0.27.1.3). Aussi, elle permet d'identifier et de quantifier les relations entre les modalités "BVA" et "V5".

# Table des matières

0.1	Introduction . . . . .	1
0.1.1	Propos . . . . .	1
0.1.2	Plan . . . . .	3
0.1.3	Une vue synthétique de la problématique "fiabilité" . . . . .	4
0.2	Aperçu de la terminologie fiabiliste . . . . .	6
0.3	Un exemple classique de calcul de fiabilité . . . . .	10
0.4	Fiabilité et Fouille de données . . . . .	12
0.5	Introduction à l'EC . . . . .	15
0.6	Apprendre des concepts . . . . .	19
0.7	Les données : Big Data . . . . .	20
0.7.1	D'où viennent les données . . . . .	23
0.8	Quelques métiers . . . . .	26
0.9	Visualisation . . . . .	29
0.10	Fouille de données : quelques exemples intuitifs . . . . .	33

0.11 Des données brutes à la Connaissance (information) . . . . .	34
0.12 Quelques exemples d'application (Real World) . . . . .	35
0.13 Extraction de Connaissances : Domaine multi disciplinaire . . . . .	36
0.14 Recherche de motifs dans les données . . . . .	37
0.15 Objectifs de l'Extraction de Connaissances . . . . .	38
0.15.1 Induction / Dédution . . . . .	39
0.15.2 Ex. de démarche Dédutive / Inductive . . . . .	40
0.16 Analyses et modélisations en DM . . . . .	41
0.17 Différentes formes de Modèles . . . . .	43
0.17.1 Exemple : Clustering . . . . .	44
0.17.2 Régression Linéaire . . . . .	45
0.17.3 Exemple : réseaux de "causalités" (BN) . . . . .	46
0.17.4 Arbre de décision . . . . .	47
0.17.5 Détection d'anomalies (et d'outsiders) . . . . .	48
0.17.6 Données saisonnières . . . . .	49
0.17.7 Recherche d'associations . . . . .	50
0.17.8 Cas de données séquentielles et Images . . . . .	51
0.17.9 Réseaux de Neurones . . . . .	52



---

0.17.10 Apprentissage de règles . . . . .	53
0.17.11 Text Mining . . . . .	54
0.18 Résumé des principales tâches de l'EC . . . . .	55
0.19 Processus d'Extraction de Connaissances dans les BDs . . . . .	57
0.20 Quelques Outils . . . . .	58
0.21 Qu'est-ce que les ordinateurs peuvent apprendre . . . . .	60
0.22 Recherche de Concepts par Généralisation (Induction) . . . . .	62
0.23 Les questions qui se posent . . . . .	66
0.24 Extraction de Motifs (Patterns) sur une BD. . . . .	69
0.24.1 Exemple d'un jeu in-door . . . . .	69
0.24.2 Extraction de règles . . . . .	70
0.24.3 Arbre de décision . . . . .	72
0.24.4 Le calcul de l'information . . . . .	74
0.24.5 Un autre exemple d'arbre de décision . . . . .	75
0.24.6 Régression . . . . .	76
0.24.7 Extraction de règles d'association . . . . .	80
0.24.8 Apprentissage à base d'instances (IBL) . . . . .	83
0.24.8.1 Un exemple : la cueillette de champignons . . . . .	85

---

0.24.9	Modélisation Statistique . . . . .	86
0.24.9.1	Prédiction Bayésienne sur l'Exemple Météo . . . . .	87
0.24.9.2	Application de de Bayes . . . . .	88
0.24.9.3	Valeurs Numériques dans Bayes . . . . .	90
0.24.10	Un exemple d'analyse Bayésienne (à partir du REX) . . . . .	91
0.24.11	Le réseau Bayésien de l'Exemple météo . . . . .	97
0.24.12	BN : exemple RATP . . . . .	100
0.24.13	RNs et Perceptron multi couches . . . . .	108
0.24.14	Méthodes à base de noyaux . . . . .	109
0.24.15	SVM . . . . .	111
0.25	Clustering . . . . .	116
0.25.1	Kmeans . . . . .	118
0.25.1.1	Illustration de K-means . . . . .	119
0.25.2	Clustering Probabiliste . . . . .	121
0.25.3	La méthode EM . . . . .	121
0.25.3.1	Une application d'EM : données manquantes . . . . .	121
0.25.4	Classification Hiérarchique . . . . .	123
0.25.5	Clustering : un autre exemple . . . . .	124

---

0.26	Evaluation . . . . .	126
0.27	Annexes . . . . .	129
0.27.1	Un exemple de prédiction de fiabilité chez Renault . . . . .	129
0.27.1.1	Les données ReX . . . . .	130
0.27.1.2	Application de la régression logistique . . . . .	132
0.27.1.3	Résultats de la Régression Logistique . . . . .	134
0.27.1.4	Application des BN . . . . .	137
0.27.1.5	Utilisation des deux méthodes . . . . .	144
	Table des matières . . . . .	146