MACHINE LEARNING

Guido.Bologna@hesge.ch

Introduction

Diapositives adaptées du cours de Anne-Claire Haury

Rendre les ordinateurs intelligents

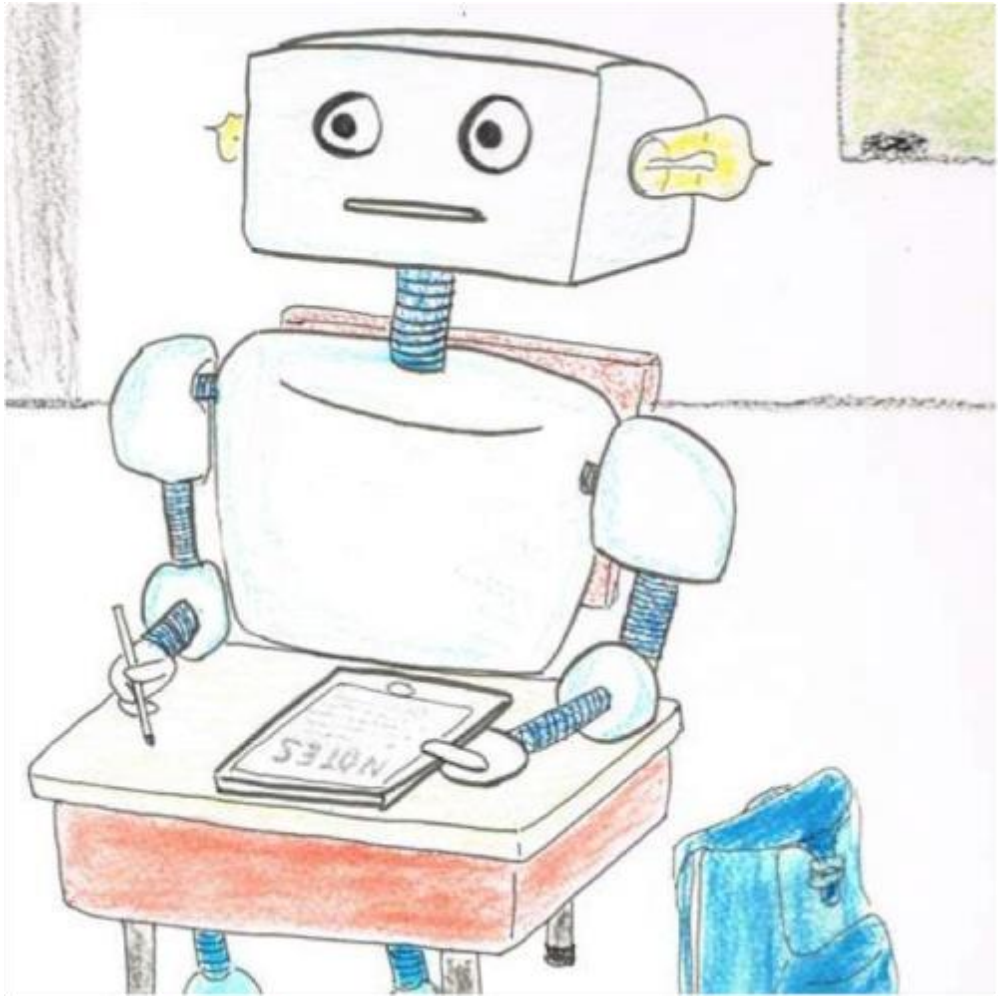


Figure: Tiré du blog du laboratoire "Computer and Cognition", NYU.

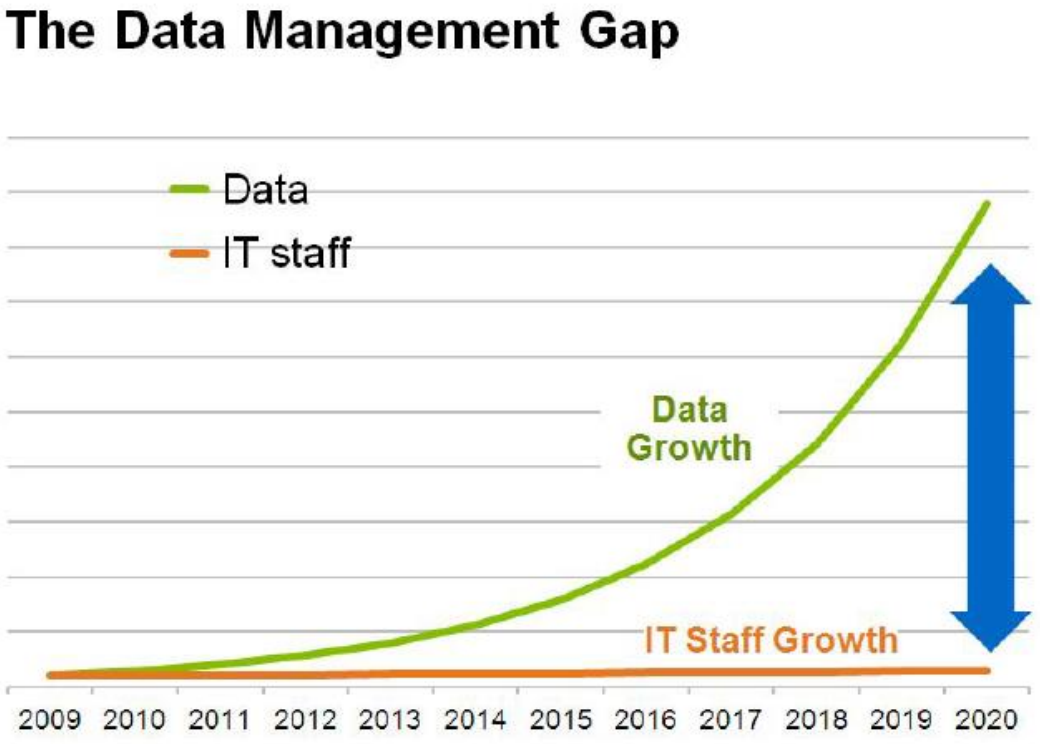
Une science à la mode

Pourquoi ?

- Stockage et traitement des données : de moins en moins cher.
- Impossible de les comprendre "à la main". Exemples : génétique, finance, réseaux sociaux, publicité...
- Dépendance d'un grand nombre de facteurs.
- Big Data: le mot magique (qui n'a pas toujours de sens).
- Compétences recherchées par les entreprises (mots-clés) : datamining, analyse de données, big data, traitement automatique de texte, d'images, machine learning...

Le fossé des données

Le fossé des données (**data gap**)



Une grosse quantité de données qui n'est jamais analysée
⇒ mettre en place des mécanismes d'analyse automatique.

Plan

- **Description**
- Faire parler les données de manière simple
- Composants de base
- Données et typologie des méthodes

DATA MINING

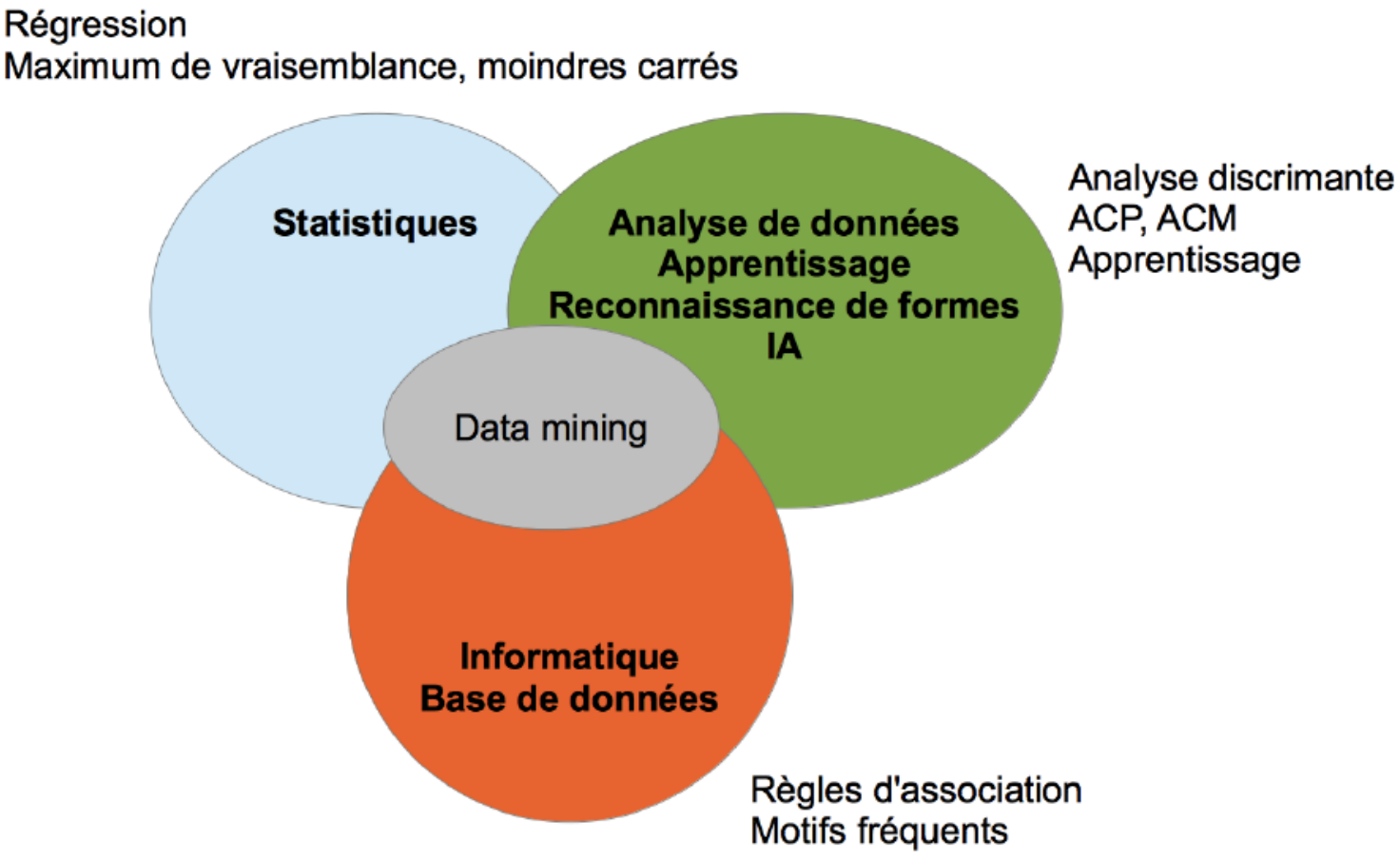
C’est un domaine récent, dont le vocabulaire n’est pas fixé.

C’est un domaine en pleine évolution qui est le résultat de la rencontre de plusieurs disciplines.

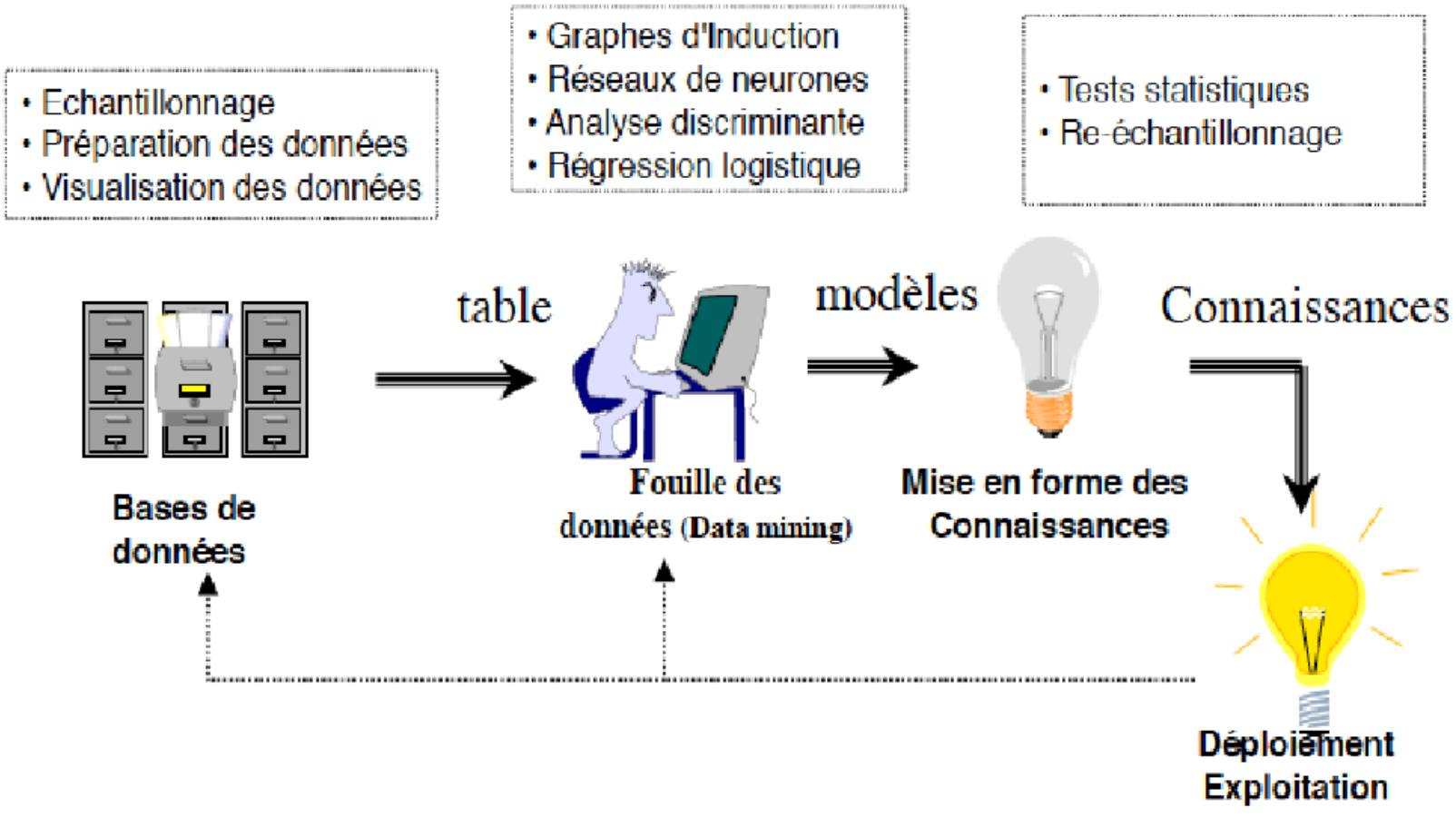
Définition : ensemble des techniques d’exploration de données permettant d’extraire des connaissances sous la forme de **modèles de description**, afin de :

- **Décrire** le comportement actuel des données.
- Et/ou **prédire** le comportement futur des données.

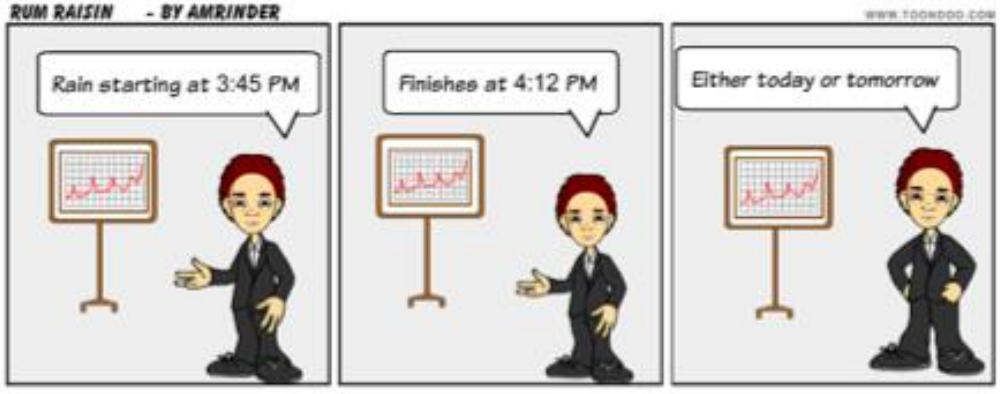
La rencontre de plusieurs disciplines



DATA MINING : une démarche plus qu'une théorie



Quelques applications



Applications Web

Plan

- Description
- **Faire parler les données de manière simple**
- Composants de base
- Données et typologie des méthodes

Esprit statistiquement critique

Les absurdités et manipulations à base de chiffres sont partout : politique, presse, et même recherche.

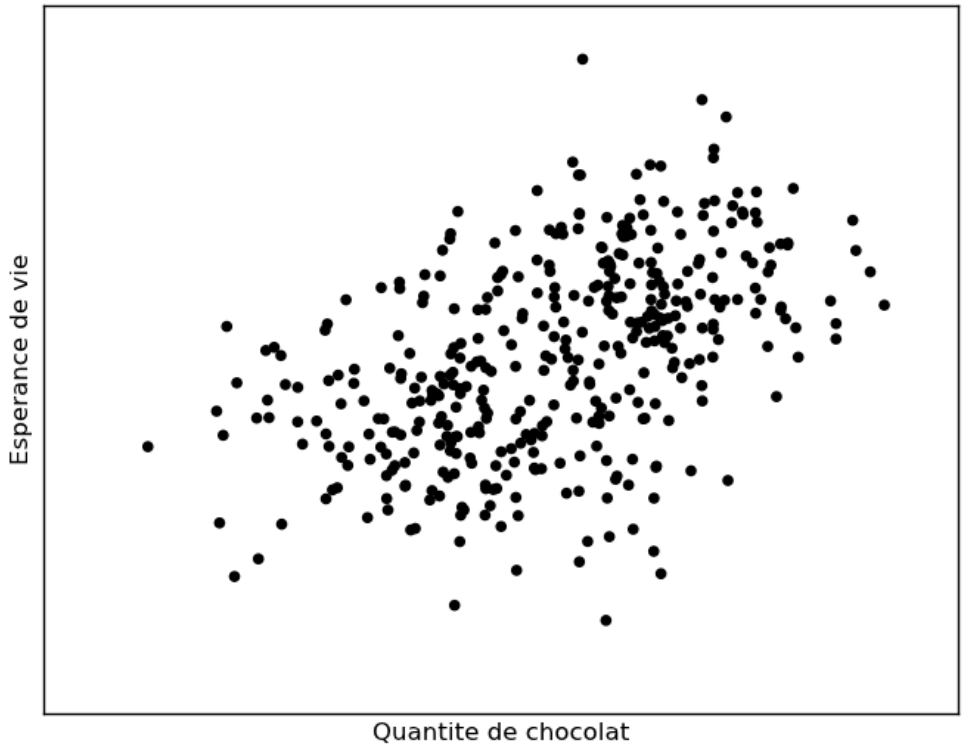
Les chiffres ont, pour la plupart des gens, une autorité intrinsèque ("c'est scientifique").

Les conclusions ne sont que le fruit de **l'interprétation**. Il faut dissocier résultats et conclusion.

On ne fait rien dire du tout aux chiffres, mais on peut les utiliser pour faire passer ses opinions.

Chocolat et espérance de vie

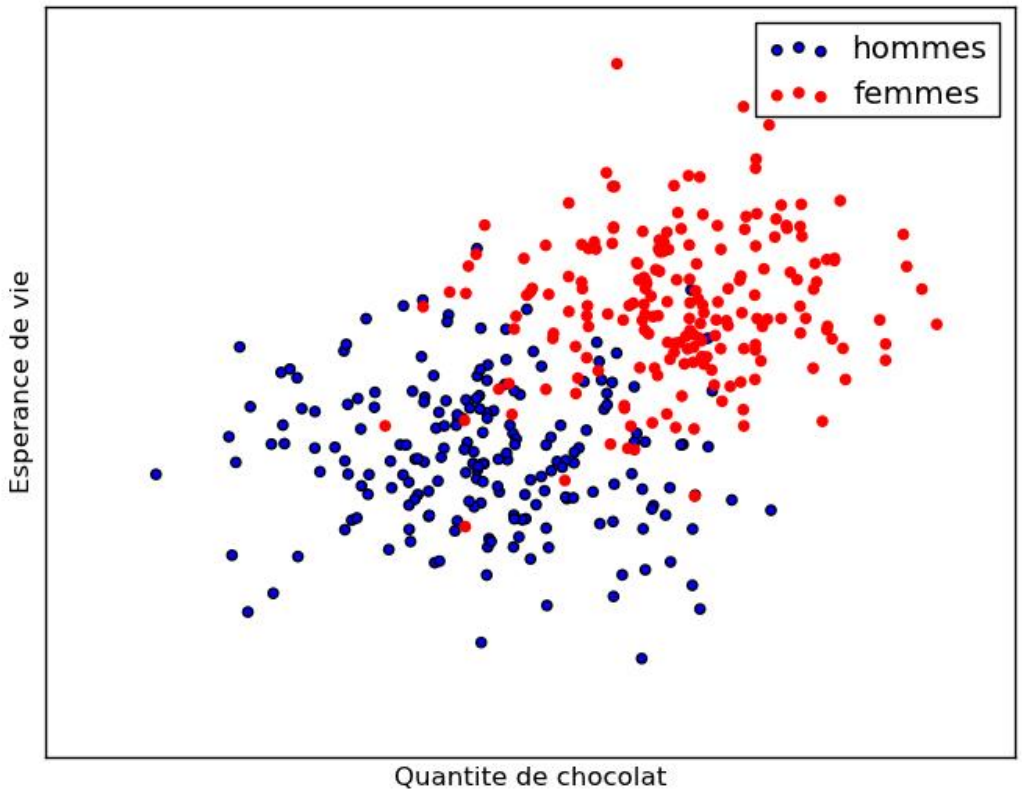
Exemple emprunté à Isabelle Guyon



Manger du chocolat augmente l'espérance de vie.

Chocolat et espérance de vie

Exemple emprunté à Isabelle Guyon



Manger du chocolat n'augmente pas l'espérance de vie.

Ce qu’on en conclut

Avoir de l’information change drastiquement la donne !

Paradoxe des anniversaires

Quelle est la probabilité que deux personnes parmi vous aient la même date d'anniversaire ?

> 50% si vous êtes plus de 23

> 80% si vous êtes plus de 35

> 90% si vous êtes plus de 41

> 95% si vous êtes plus de 47

> 99% si vous êtes plus de 58

Paradoxe des anniversaires : explication

Il serait **très improbable** que vous ayez tous une date différente d'anniversaire. Itérons:

- La première personne choisit sa date parmi 365 dates. Il reste 364 choix pour la seconde.
- La seconde choisit sa date. Il reste 363 choix.
- La n-ème personne a (365 − n + 1) choix.

Si on transforme cela en probabilités, on obtient :

$$p = \frac{365}{365} \times \frac{364}{365} \times \cdots \times \frac{365 - n + 1}{365}$$

p est la probabilité que les n personnes aient des anniversaires différents. Très rapidement, cette probabilité devient **infime** (on ne multiplie que des nombres < 1). La probabilité que deux personnes **au moins** partage la même date est donc 1 − p.

Exemple avec 50 personnes

Probabilité d'avoir des anniversaires différents:

$$
\begin{aligned}
p &= \frac{365}{365} \times \frac{364}{365} \times \cdots \times \frac{365 - 50 + 1}{365} \\
&= \frac{365 \times 364 \times \cdots \times 316}{365^{50}} \\
&= 0.0296
\end{aligned}
$$

Il y a donc 97% de chances qu'au moins 2 personnes aient le même anniversaire.

Plan

- Description
- Faire parler les données de manière simple
- **Composants de base**
- Données et typologie des méthodes

Composants de base pour le Data Mining

Grande quantité de données + algorithmes efficaces

Disponibilité de grandes quantités de données

- Si un ensemble de données est trop petit, les structures émergentes peuvent résulter du hasard. En général on peut espérer qu'un gros volume de données représente bien le domaine ciblé.

Des algorithmes sûrs et efficaces

- Algorithmes sûrs : fondés théoriquement, corrects.
- Efficaces en temps et en espace.
- Paramètres ajustables facilement et rapidement.

Un exemple

Issu du livre de Adriaans and Zantige (d'après B. Espinasse)

- Un éditeur vend 5 sortes de magazines : sport, voiture, maison, musique, cinéma.
- Il veut étudier ses clients pour découvrir de nouveaux marchés ou vendre plus à ses clients habituels.

Quelques questions

1. Combien de personnes ont pris un abonnement à un magazine de cinéma cette année ?
2. A-t-on vendu plus d'abonnement de magazines de sport cette année que l'année dernière ?
3. Est-ce que les acheteurs de magazines de musique sont aussi amateurs de cinéma ?
4. Quelles sont les caractéristiques principales des lecteurs de magazine de cinéma ?
5. Peut-on prévoir les pertes de clients et prévoir des mesures pour les diminuer ?

Un exemple

1 : Combien de personnes ont pris un abonnement à un magazine de cinéma cette année ?

Requête SQL à partir des données opérationnelles suffit si les tables concernées ont été suffisamment indexèes.

2 : A-t-on vendu plus d'abonnement de magazines de sport cette année que l'année dernière ?

- Nécessite de garder toutes les dates de souscription, même pour les abonnements résiliés.
- Requêtes multidimensionnelles de type OLAP.

Un exemple

3 : Est-ce que les acheteurs de magazine de musique sont aussi amateurs de cinéma ?

- Exemple simplifié de problème où l'on demande si les données vérifient une règle.
- Réponse formulée par une valeur estimant la probabilité que la règle soit vraie.
- Utilisation d'outils statistiques.

Un exemple

4. Quelles sont les caractéristiques principales des lecteurs de magazine de cinéma ?

Question plus ouverte, il s'agit de trouver une règle et non plus de la vérifier ou de l'utiliser.

5 : Peut-on prévoir les pertes de client et prévoir des mesures pour les diminuer ?

Question ouverte : il faut disposer d'indicateurs comme durée d'abonnement, délai de paiement, ...

C'est pour ce type de questions que sont mis en oeuvre les outils d'analyse et de fouille de données

Plan

- Description
- Faire parler les données de manière simple
- Composants de base
- **Données et typologie des méthodes**

Les données

Les données peuvent être vues comme une collection d'objets (enregistrements) et leurs attributs.

Un attribut est une propriété et/ou une caractéristique de l'objet. Un ensemble d'attributs décrit un objet.

Attributs

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribut : valeurs

- La valeur d'un attribut est un nombre ou un symbole.
- Ne pas confondre attribut et valeur

Types

- Quantitative (numérique, exprime une quantité)
  - Discrète (ex : nombre d'étudiants dans un cours) ou continue (ex : longueur)
  - Echelle proportionnelle (chiffre d'affaires, taille), ou échelle d'intervalle (température, QI)
- Qualititative
  - Variable ordinale (classement à un concours, échelle de satisfaction client)
  - Variable nominale (couleur de yeux, diplôme obtenu, CSP, sexe)
- Les modalités d'une variable sont l'ensemble des valeurs qu'elle prend dans les données
  ex : les modalités de notes sont $\{0, 1, 2, \cdots, 20\}$ les modalités de couleur sont {bleu,vert,noir,...}

Typologie des méthodes

Typologie selon l'objectif

- **Classification** : examiner les caractéristiques d'un objet et lui attribuer une classe. Exemple : diagnostic ou décision d'attribution de prêt à un client.

- **Prédiction** : prédire la valeur future d'un attribut en fonction d'autres attributs. Exemple : prédire la qualité d'un client.

- **Association** : déterminer les attributs qui sont corrélés. Exemple : analyse du panier de la ménagère.

- **Segmentation** : former des groupes homogènes à l'intérieur d'une population.

Typologie selon le type de modèle obtenu

Modèles prédictifs.

- Utilisent les données existantes et des résultats connus sur ces données pour développer des modèles capables de prédire les valeurs d'autres données.
- e.g. *Prédire les clients qui ne rembourseront pas leur crédit.*
- Utilisés principalement en classification et prédiction.

Modèles descriptifs.

- Proposent des descriptions de données pour aider à la prise de décision.
- Souvent en amont de la construction de modèles prédictifs.
- Utilisés principalement en segmentation et association.

Typologie selon le type d'apprentissage utilisé

Apprentissage supervisé : fouille supervisée

- Processus qui prend en entrée des exemples d'apprentissage contenant à la fois des données d'entrée et de sortie.
- Les exemples d'apprentissage sont fournis avec leur classe.
- But : classer correctement un nouvel exemple.
- Utilisés principalement en classification et prédiction.

Typologie selon le type d'apprentissage utilisé

Apprentissage non supervisé : fouille non supervisée

- Processus qui prend en entrée des exemples d'apprentissage ne contenant que des données d'entrée.
- Pas de notion de classe.
- But : regrouper les exemples en groupes (clusters) d'exemples similaires.
- Utilisés principalement en segmentation et association.

Classification

Examiner les caractéristiques d'un objet et lui attribuer une classe (un champ particulier à valeurs discrètes).

- Etant donnée une collection d'exemples (**ensemble d'apprentissage**), trouver un modèle pour l'attribut classe comme une fonction de la valeur des autres attributs.

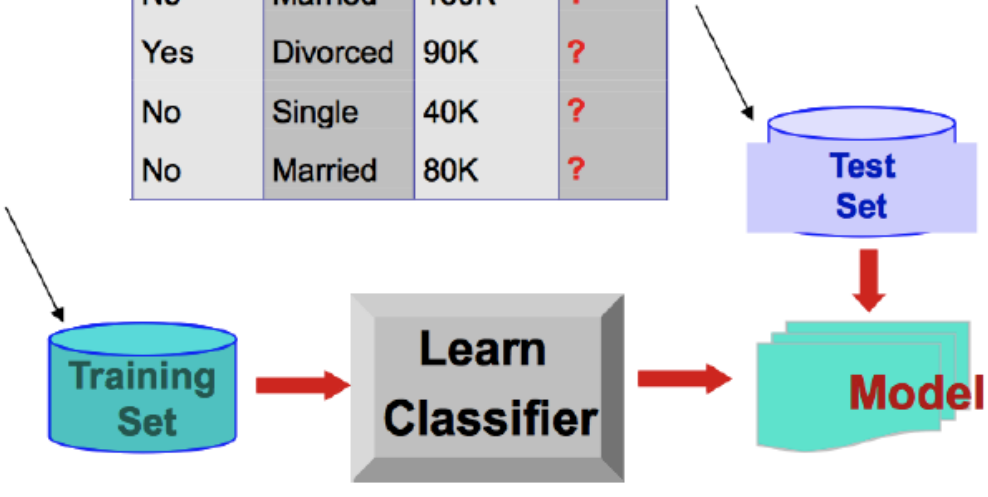
- But : permettre d'assigner une classe à des enregistrements inconnus de manière aussi précise que possible.

- Un **ensemble de test** est utilisé pour déterminer la précision du modèle.

Classification : exemple

Tid	Refund	Marital Status	Taxable Income	Cheat
		categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification : exemple de Marketing direct

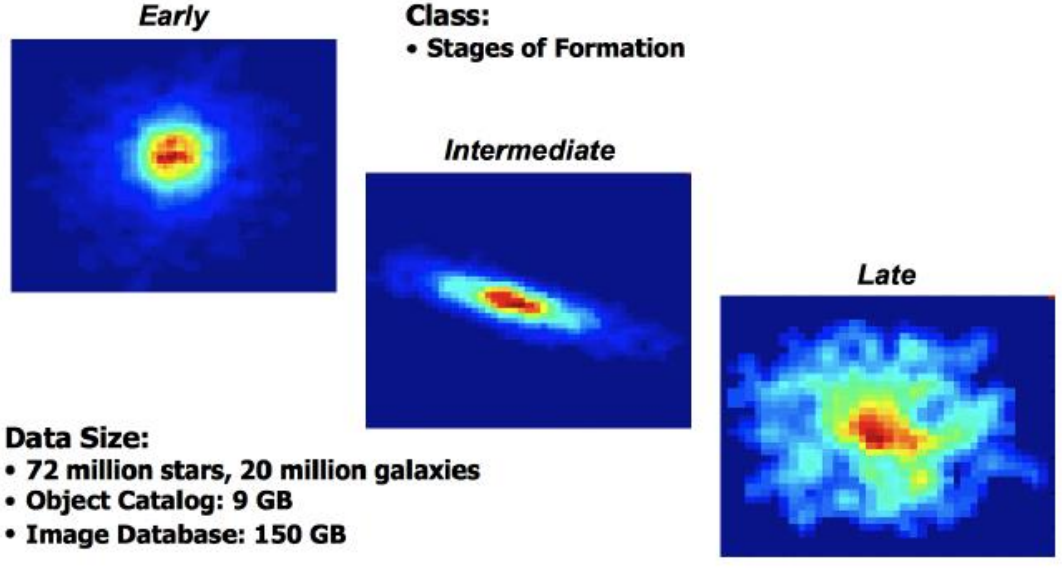
But : réduire le coût du mailing en ciblant un ensemble de consommateurs qui achèteront vraisemblablement un nouveau téléphone portable.

Approche :

- Utiliser des données pour un produit similaire.
- On sait quels consommateurs ont acheté. La décision (*Achat* - *Pas achat*) est l'attribut classe.
- Collecter diverses informations sur ce type de
- consommateurs.
- Ces informations représentent les entrées du classifieur.

Classification : exemples d'applications

- **Détection de fraudes** (carte bancaire) à l'aide des transactions et d'informations sur le porteur du compte.
- **Détection de désabonnement** à l'aide des données sur d'autres consommateurs présents ou passés.
- **Catalogage du ciel** : classification des objets du ciel à l'aide d'images.



Segmentation

Former des groupes homogènes à l'intérieur d'une population.

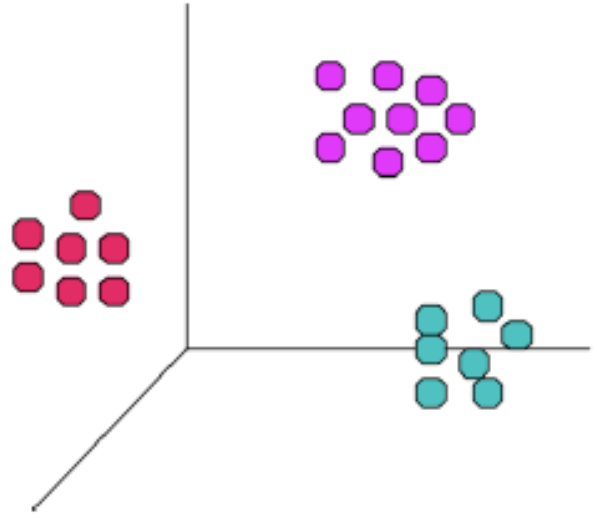
- Etant donné un ensemble de points, chacun ayant un ensemble d'attributs et une mesure de similarité définie sur eux, trouver des groupes tels que :
  - les points à l'intérieur d'un même groupe sont très similaires entre eux ;
  - les points appartenant à des groupes différents sont très dissimilaires.
- Le choix de la mesure de similarité est important

Segmentation : illustration

- Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Segmentation : exemples

- Segmentation de marchés.
- Segmentation de documents.
- Segmentation de clients.
- …

Association

Entrée : Un ensemble de tickets de caisse

- Une observation = un caddie, un ticket de caisse.
- Non prise en compte de la fréquence des produits.
- Un grand nombre de produits, un grand nombre de caddies (petit sous ensemble de l'ensemble de produits).

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Sortie : Des règles

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Association : exemples

- Marketing et promotions sur des produits.
- Gestion du supermarchés : rayonnage.
- Inventaire.
- ...