Introduction au Data Mining

Julien JACQUES

Université de Lyon, Lyon 2 & ERIC EA 3083

Plan

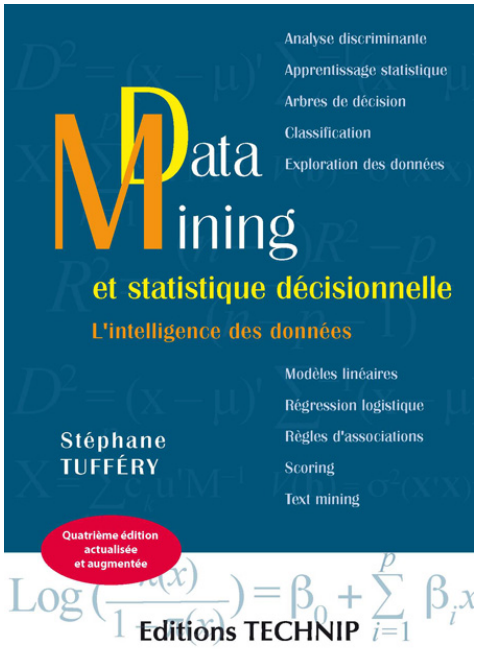
Le data mining : qu'est-ce donc ?

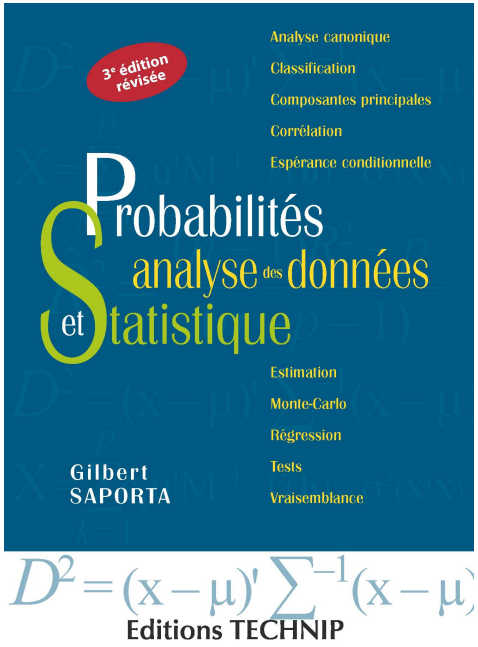
- data mining = fouille de données
- Définition de Wikipédia (2017) :
  - *L'exploration de données, connue aussi sous l'expression de fouille de données, forage de données, prospection de données, data mining, ou encore extraction de connaissances à partir de données, a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques.*
- définition de P. Besse[1]
  - *Ensemble d'approches statistiques permettant d'extraire de l'information de grands jeux de données dans une perspectives d'aide à la décision.*

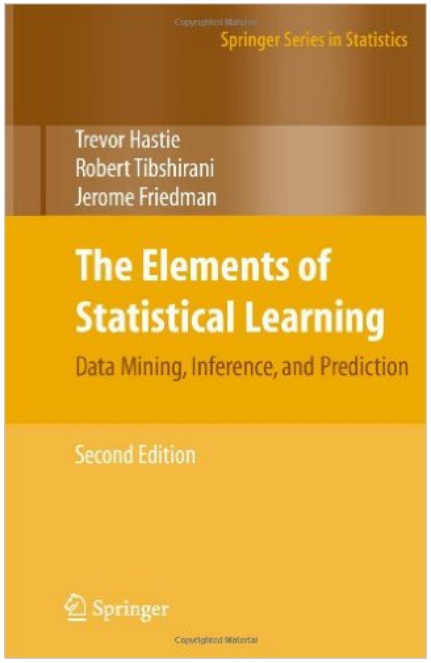
[1] *P. Besse et al., Data Mining et Statistique, Journal de la Société Française de Statistique, 142[1], 2001.*

Le data mining : quelques références

Quelques ouvrages : d'abordable à plus technique...







Le data mining : quelques références

Sites internet

- nombreux cours, e-books, tutoriels (R, tanagra, Excel) :
  `http://eric.univ-lyon2.fr/~ricco/data-mining/`
- site de l'équipe de Rennes : `http://www.sthda.com/french/`
- site de l'équipe de Toulouse : `http://wikistat.fr`
- site de S. Tufféry : `http://data.mining.free.fr`

Plan

Le data mining : rétrospective historique

- Historiquement, les premières approches statistiques étudient un petit nombre *n* d'individus décrits par un petit nombre *p* de variables. Ces données sont issues de plans d'expériences.

Le data mining : rétrospective historique

- Historiquement, les premières approches statistiques étudient un petit nombre *n* d'individus décrits par un petit nombre *p* de variables. Ces données sont issues de plans d'expériences.
- 1990s (MO) : les entreprises commencent à stocker de plus en plus de données concernants leur clients, sans planification expérimentale. Les méthodes statistiques classiques sont massivement utilisées pour extraire de la connaissance de ces données (CRM, gestion de la relation client). C'est la **naissance du data mining**.

Le data mining : rétrospective historique

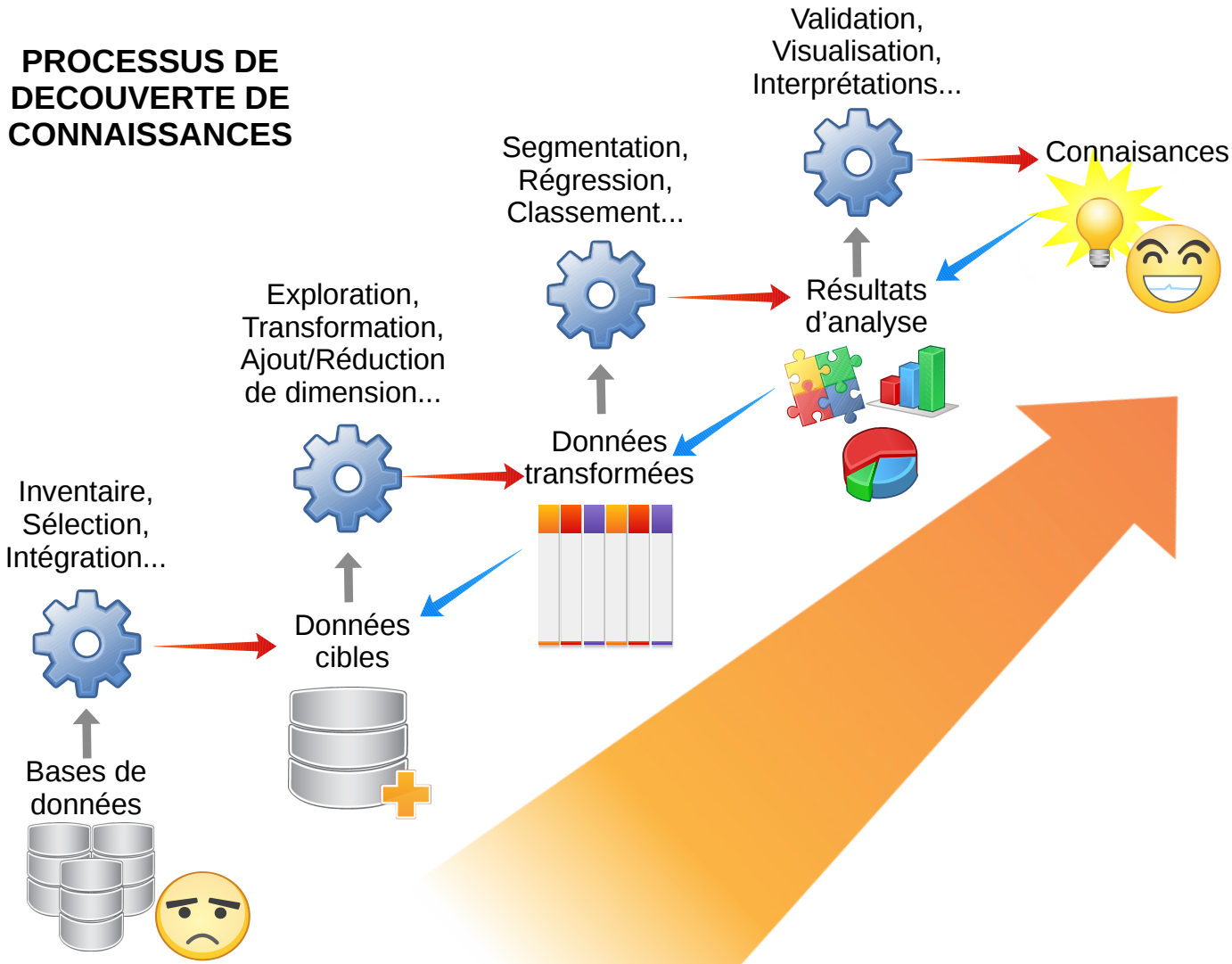
- 2000s (GO) : **première révolution** du data mining avec l'avénement de la bioinformatique et des données omiques : on observe beaucoup de variables sur peu d'individus ($n << p$). On parle du fléau de la dimension et on doit développer de nouvelles méthodes parcimonieuses.

Le data mining : rétrospective historique

- 2000s (GO) : **première révolution** du data mining avec l'avénement de la bioinformatique et des données omiques : on observe beaucoup de variables sur peu d'individus ($n << p$). On parle du fléau de la dimension et on doit développer de nouvelles méthodes parcimonieuses.
- 2010s (TO) : **seconde révolution** due au développement d'internet (commerce en ligne, réseau sociaux). On parle de big data (volume variété vélocité...) et de science des données.

Plan

Les étapes du data mining



Inventaire, sélection et intégration des données

1. Inventaire, sélection et intégration des données
2. Exploration, transformation des données
3. Analyse statistique : segmentation, régression, classement
4. Validation, visualisation et interprétations des résultats

Inventaire, sélection et intégration des données

- définir et identifier
  - ce que l'on veut étudier
  - ce que je veux faire (prédire, découvrir...)
  - quelles observations on va utiliser
- rassembler les informations sur le phénomène
  - quelles sont les variables (descripteurs) existantes
  - sélectionner les variables en lien avec ce que l'on veut étudier
  - intégrer ces variable dans un même jeu de données
- cette partie demande de compétences en base de données, non abordées dans ce cours.

Exploration, transformation des données

- Il faut commencer par *faire connaissance* avec les données, à l'aide de statistiques descriptives
  - statistiques univariées (tendance centrale, dispersion)
  - graphiques (historgrammes, boxplot, ...)
  - statistiques bivariées (corrélation...) pour identifier des variables corrélées
  - méthodes exploratoires (ACP, AFC, ACM...) pour visualiser les grandes tendances

Exploration, transformation des données

- Il faut commencer par *faire connaissance* avec les données, à l'aide de statistiques descriptives
  - statistiques univariées (tendance centrale, dispersion)
  - graphiques (historgrammes, boxplot, ...)
  - statistiques bivariées (corrélation...) pour identifier des variables corrélées
  - méthodes exploratoires (ACP, AFC, ACM...) pour visualiser les grandes tendances
- identifier et gérer les **données manquantes**
  - on les supprime si elles sont peu nombreuses et si le jeu de données est grand
  - on les impute par la moyenne, médiane, le mode, ou par une méthode plus élaborée (packages R `mice`, `missMDA`...)

Exploration, transformation des données

- Il faut commencer par *faire connaissance* avec les données, à l'aide de statistiques descriptives
  - statistiques univariées (tendance centrale, dispersion)
  - graphiques (historgrammes, boxplot, ...)
  - statistiques bivariées (corrélation...) pour identifier des variables corrélées
  - méthodes exploratoires (ACP, AFC, ACM...) pour visualiser les grandes tendances
- identifier et gérer les **données manquantes**
  - on les supprime si elles sont peu nombreuses et si le jeu de données est grand
  - on les impute par la moyenne, médiane, le mode, ou par une méthode plus élaborée (packages R `mice`, `missMDA`...)
- identifier et traiter les **observations atypiques**
  - demande à l'expert métier si c'est une erreur de mesure, une observation hors norme...

Exploration, transformation des données

- Normalisation de variables quantitatives
  - lorsque les variables ont des échelles différentes
  - on centre (soustrait la moyenne) et réduit (divise par l'écart-type) chaque variable
- Transformation de variables qualitatives en quantitatives ou vice-versa
  - d'un point de vue général on évitera un maximum de faire cela, et on cherchera à utiliser des méthodes permettant d'utiliser des données *mixtes*
  - quand on a pas le choix, on peut discrétiser les variables quanti. en variables quali. (mais on perd énormément d'information)
  - on peut faire le contraire en utilisant une ACM par exemple, mais on perd en interprétabilité

Analyse statistique : segmentation, régression, classement

- c'est la partie analyse qui va permettre d'extraire de l'information.
- on verra plus loin la distinction entre les méthodes exploratoires, prédictives, ...

Validation, visualisation et interprétations des résultats

- Validation : on cherchera à valider les résultats à l'aide de données indépendantes, d'avis humain...
- Visualisation : les résultats des analyses seront illustrées graphiquement afin de faciliter leur interprétation.
- Interprétation : c'est ici que, grâce à l'expert métier, on tire de l'information et de la connaissance sur le phénomène étudié.

Plan

Le data mining : à quoi cela sert ?

- Vente, marketing
  - gestion de la relation client (scoring, score d'appétence)
  - segmentation de la clientèle
- Banque, finance, assurance
  - détection de fraude (comportements atypiques)
  - score de risque (attribution ou non d'un crédit)
- Technologie
  - reconnaissance faciale dans une image
  - reconnaissance de la parole
- Médecine, industrie pharmaceutique
  - réponse d'un patient vis-à-vis d'un traitement
  - identification des facteurs de risques
- Energie, transport...
  - prévision de consommation d'électricité
  - prévision de traffic routier

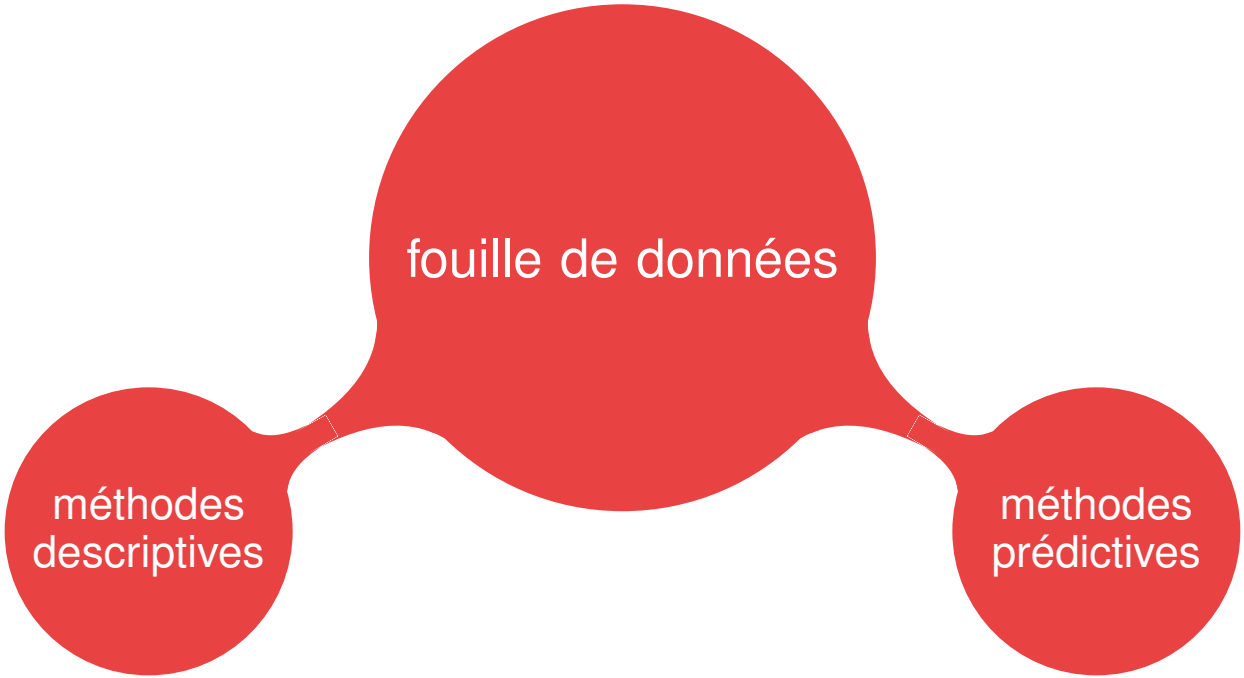
Le data mining : à quoi cela sert ?

- Vente, marketing
  - gestion de la relation client (scoring, score d'appétence)
  - segmentation de la clientèle
- Banque, finance, assurance
  - détection de fraude (comportements atypiques)
  - score de risque (attribution ou non d'un crédit)
- Technologie
  - reconnaissance faciale dans une image
  - reconnaissance de la parole
- Médecine, industrie pharmaceutique
  - réponse d'un patient vis-à-vis d'un traitement
  - identification des facteurs de risques
- Energie, transport...
  - prévision de consommation d'électricité
  - prévision de traffic routier

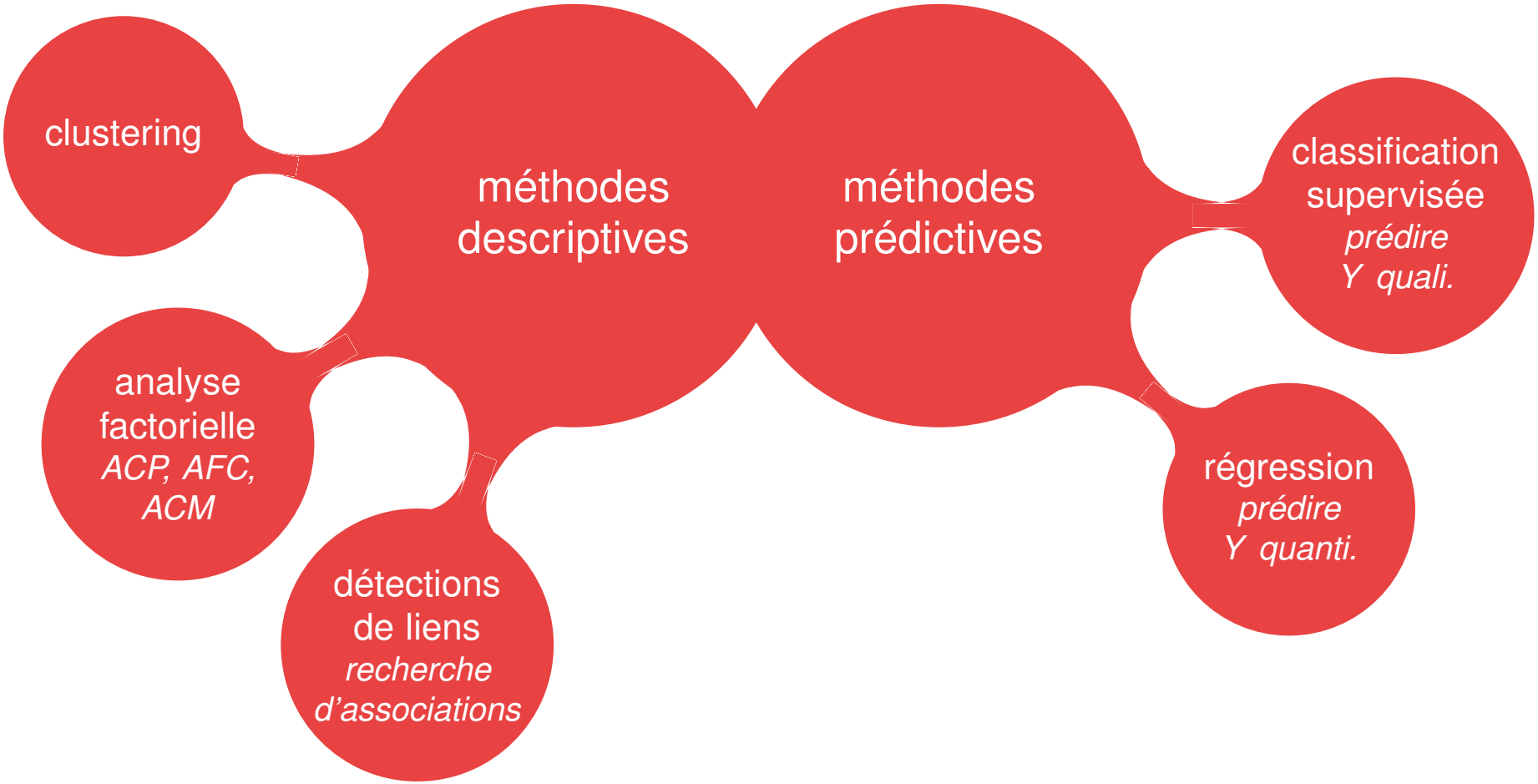
Le Data Mining peut s'appliquer à tout phénomène dont on peut mesurer des observations et dont on souhaite appréhender les caractéristiques et / ou prévoir le comportement

Plan

Le data mining : panorama des méthodes



Le data mining : panorama des méthodes

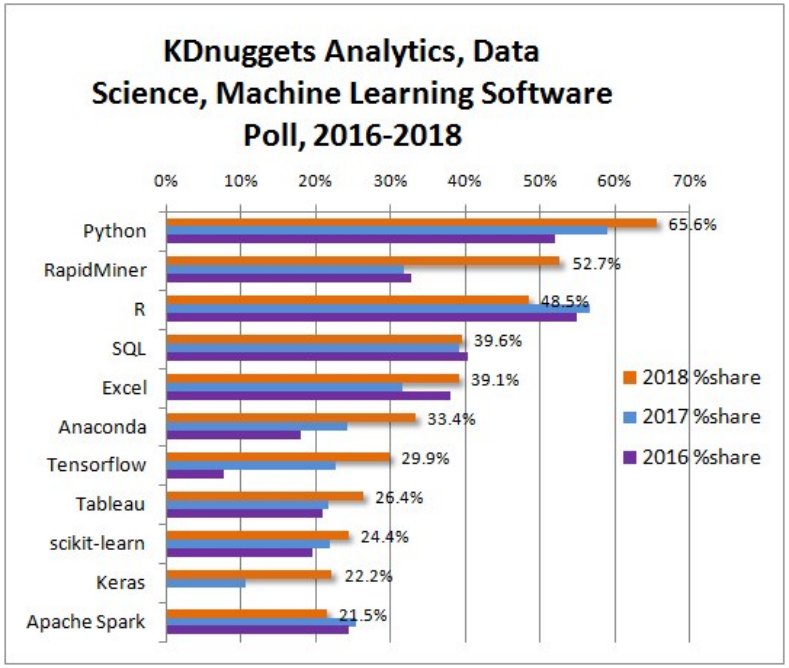


Plan

Les outils du data mining

- propriétaires et payants

- open source et/ou gratuits



Plan

Les données

Les data

On stocke généralement les données sous forme d'un tableau (matrice)

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

où :

- une ligne = une observation, un individu, parfois notée $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$
- une colonne = une variable
- x_{ij} : valeur de la j-ème variable pour le i-ème individu
- n est le nombre d'observations et p le nombre de variables

Le data mining : des data...

Les données X_{ij} peuvent être de différent types

- quantitatif (mesurables)
- catégoriel (nominales, ordinales)
- mais également
  - textes, images, réseaux...



Le data mining : des data...

Les données X_{ij} peuvent être de différent types

- quantitatif (mesurables)
- catégoriel (nominales, ordinales)
- mais également
  - textes, images, réseaux...



  - tout en même temps