Machine Learning

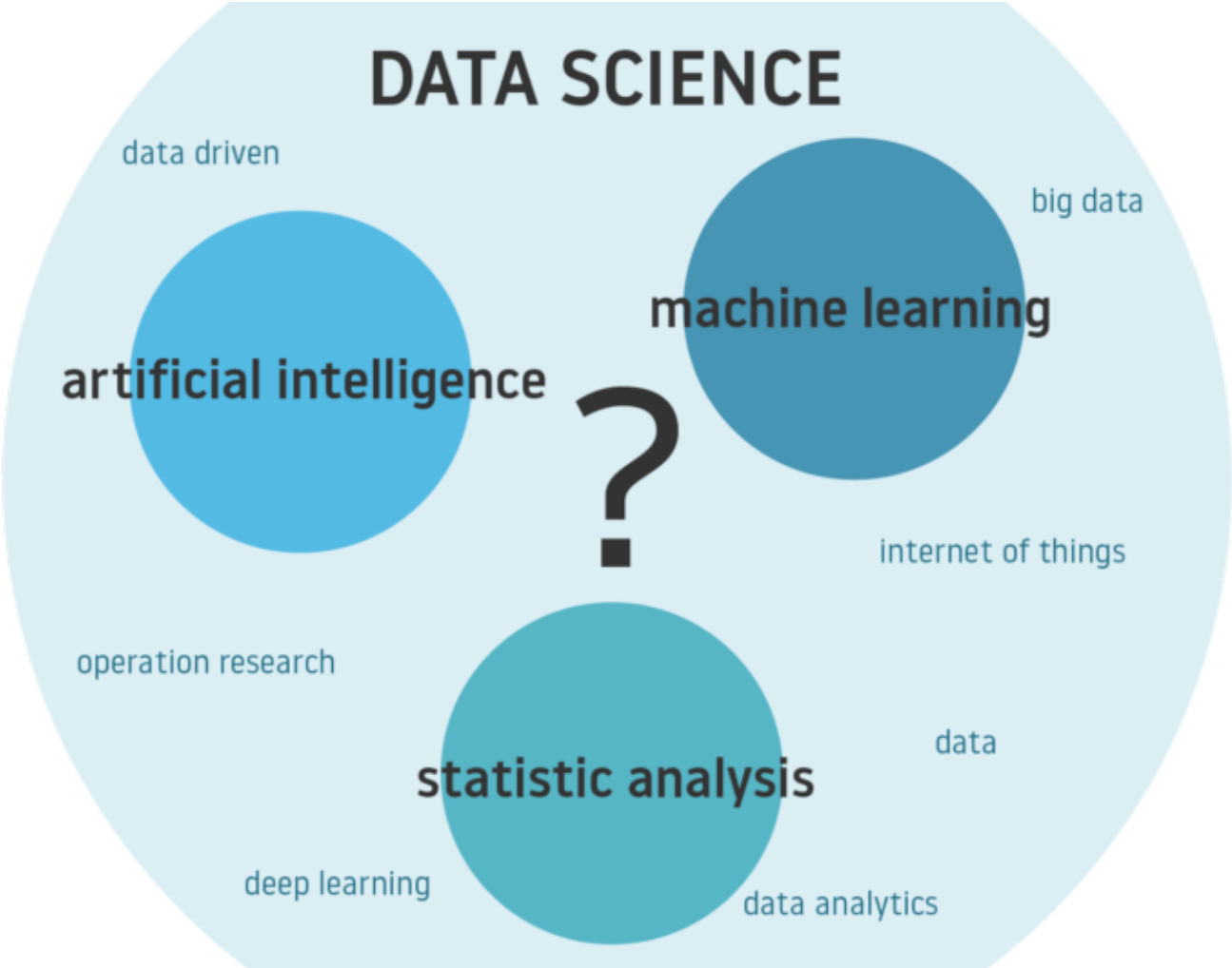
Introduction

marie.szafranski@ensiie.fr

Début : 09h30

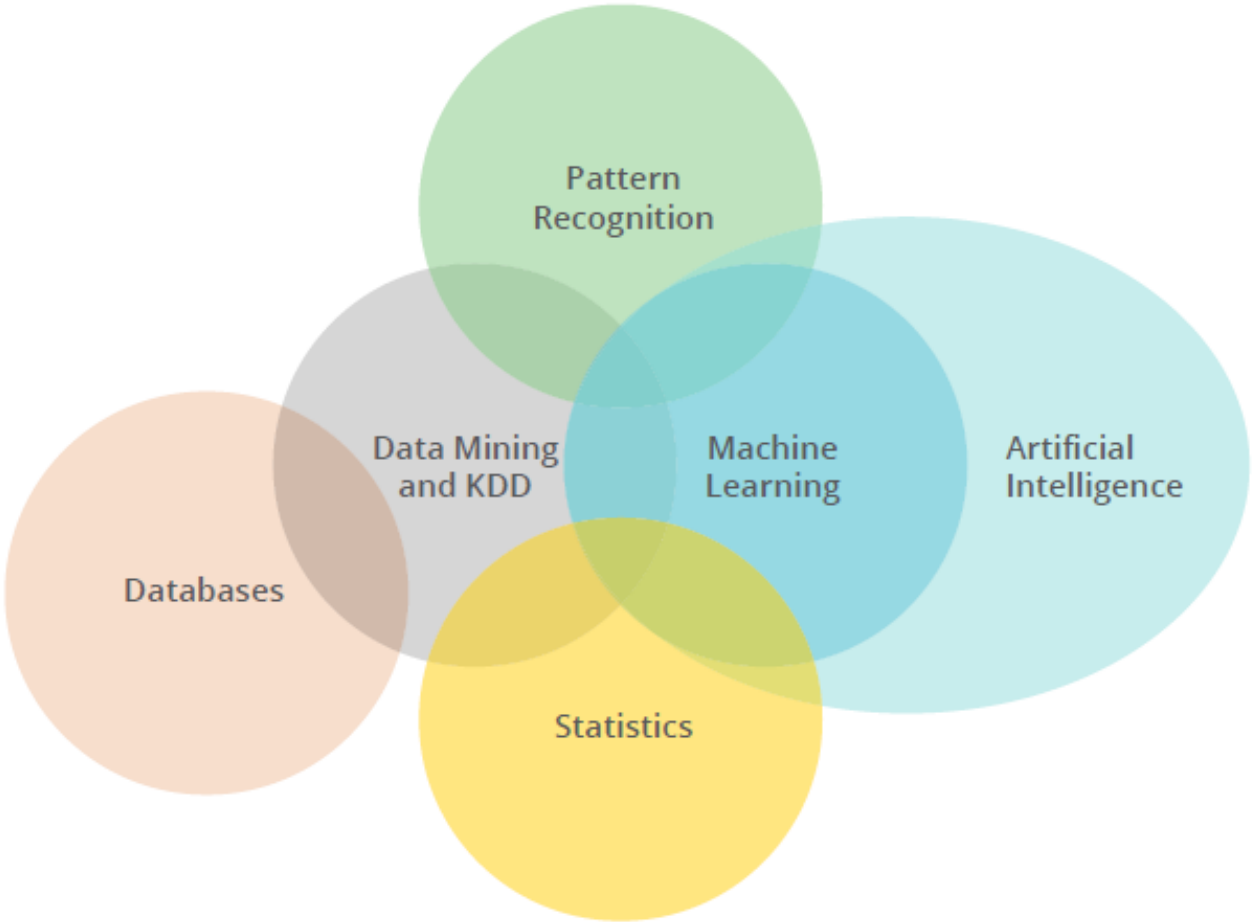
De quoi parle-t-on ?

Buzzwords en vrac



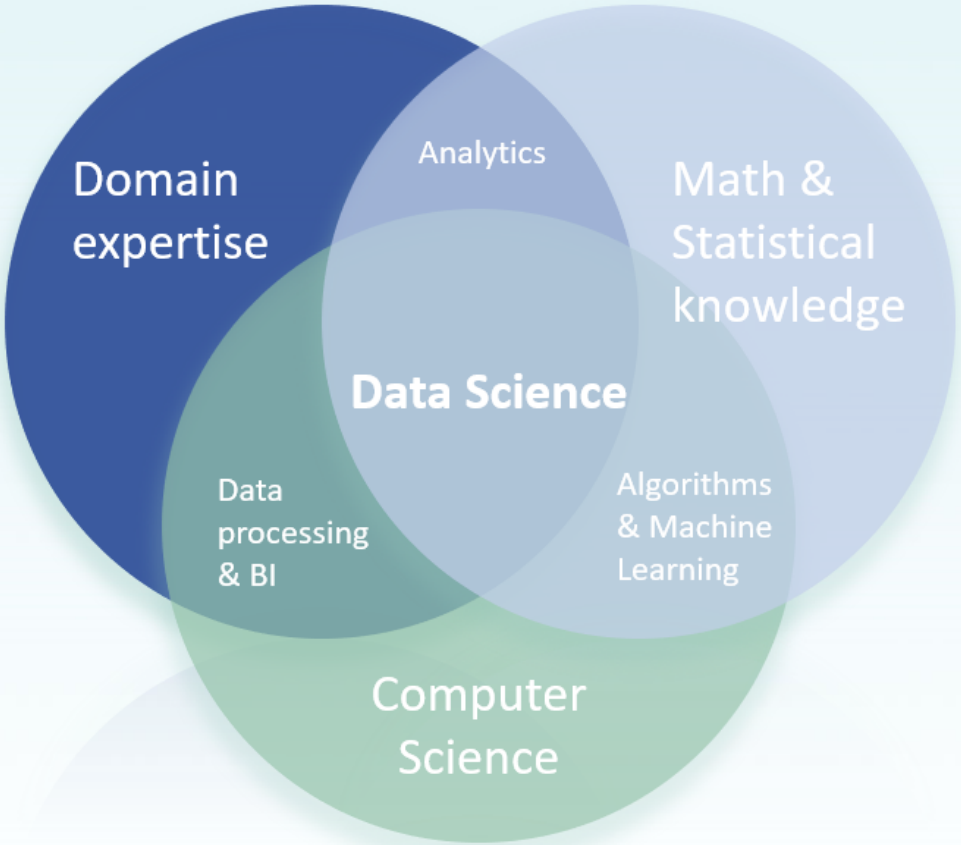
De quoi parle-t-on ?

Une structure



De quoi parle-t-on ?

Une autre structure



Définition (tentative I)

Machine learning : une boîte noire ? NON

https://xkcd.com/1838/

Définition (tentative II)

Traduction de machine learning

- Apprentissage machine — littéralement
- Apprentissage artificiel — référence à l'IA
- Apprentissage automatique — $\rightsquigarrow$ *algorithmes*, connotation informatique
- Apprentissage statistique — $\rightsquigarrow$ *modèles*, connotation statistique

Définition

Wikipedia

Machine learning is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead

Apprentissage automatique vs statistique

Deux points de vue complémentaires [Breiman, 2001]

Scientifique

$\neq$ **magie...**

Quelles données ?

N individus décrits par D variables

Quelques ordres de grandeur

[Wikistat, 2016]

- **kO** 1970s
  Analyse de données
- **MO** 1980s
  Apprentissage automatique $\rightsquigarrow$ réseaux de neurones
- **GO** 1990s
  Exploration et fouille de données [Fayyad et al., 1996]
  Apprentissage statistique $\rightsquigarrow$ SVM [Vapnik, 1995]
- **TO** 2000s
  Bioinformatique $D \gg N$
- **PO** 2010s
  Réseaux sociaux, e-commerce, etc. . . $N \gg D$

Des données à grande échelle

N ou/et D très grands

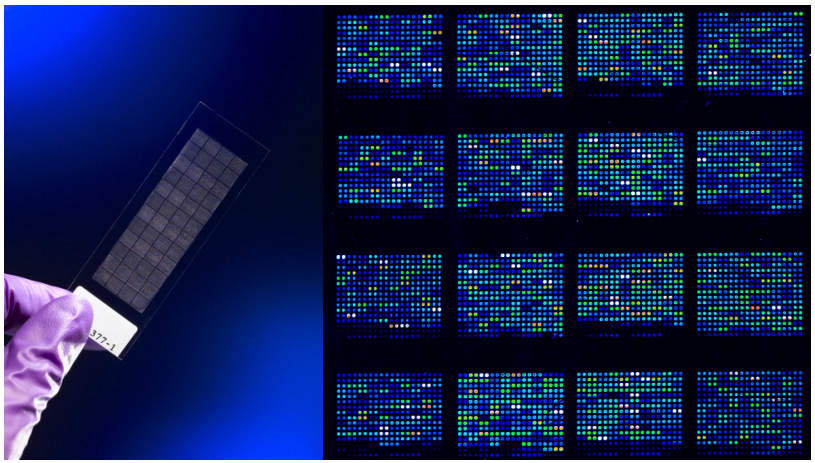
Évolutions technologiques dans les techniques d'acquisition

$+$ Augmentation des capacités de stockage

$\rightsquigarrow$ Explosion de la quantité d'information disponible

Génomique

- N : quelques centaines de patients
- D : mesure de l'expression de plusieurs millions de variants génétiques



Des données à grande échelle

N ou/et D très grands

Évolutions technologiques dans les techniques d'acquisition

$+$ Augmentation des capacités de stockage

$\rightsquigarrow$ Explosion de la quantité d'information disponible

Astronomie

- N : plusieurs millions de corps célestes
- D : quelques centaines de mesures (positions, vitesses, etc.)

Des données à grande échelle

N ou/et D très grands

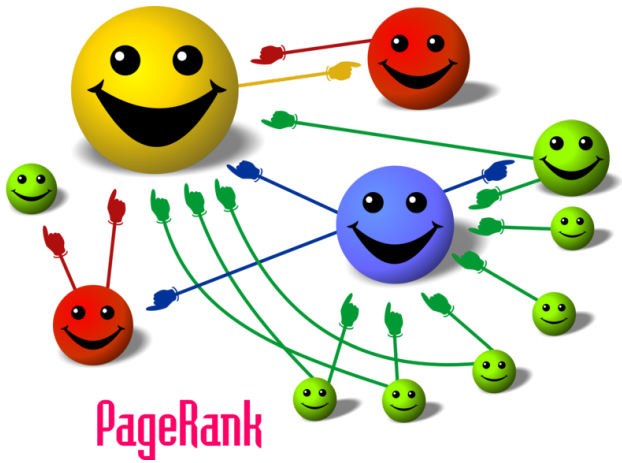
Évolutions technologiques dans les techniques d'acquisition

$+$ Augmentation des capacités de stockage

$\rightsquigarrow$ Explosion de la quantité d'information disponible

Web et text mining

- N : plusieurs millions de sites
- D : liens entrants et sortants, ancres, nom de domaine, hébergeur, etc.



Quelle finalité pour ces données ?

Définir un modèle et réaliser l'algorithme associé dans une perspective prédictive et / ou explicative

Exemples

- **Bioinformatique :** identifier les gènes qui permettent de distinguer des patients sains de patients atteints d'une maladie
- **Astronomie :** étudier des relations liant des paramètres de positions et de vitesses de corps célestes à leur composition chimique
- **Web mining :** analyser le comportement d'internautes pour définir un algorithme de recherche sur les sites web (PageRank de Google)

Schéma global

À la croisée de l'informatique et des statistique

1. Phase de collecte et d'intégration

- Architecture des données
- Web et réseaux distribués

Schéma global

À la croisée de l'informatique et des statistique

2. Phase exploratoire

- Analyse de données
- Recherche d'information et de motifs

Schéma global

À la croisée de l'informatique et des statistique

3. Phase explicative ou décisionnelle

- Machine Learning
- IA interprétable ($\neq$ XAI)

Besoins et défis actuels

À la croisée de l'informatique et des statistique

3. Phase explicative ou décisionnelle

- Machine Learning
- IA interprétable ($\neq$ XAI)

Environnement : vraie R&D

Concurrence : +++ ($>$ doctorat)

Besoins et défis actuels

À la croisée de l'informatique et des statistique

2. Phase exploratoire

- Analyse de données
- Recherche d'information et de motifs

Environnement : partout

Concurrence : ++ (ingé. XP / doct.)

Besoins et défis actuels

À la croisée de l'informatique et des statistique

1. Phase de collecte et d'intégration

- Architectures des données
- Web et réseaux distribués

Environnement : partout

Concurrence : $-/=$ (ingé. / doct.)

DataOps $\rightsquigarrow$ MLOps

Différentes classes de méthodes

Pourquoi différentes classes ?

- Différents types de données
- Différents objectifs

Classes de méthodes

- Méthodes descriptives
- Méthodes factorielles
- Méthodes non supervisées
- Méthodes supervisées

Méthodes descriptives

Objectif

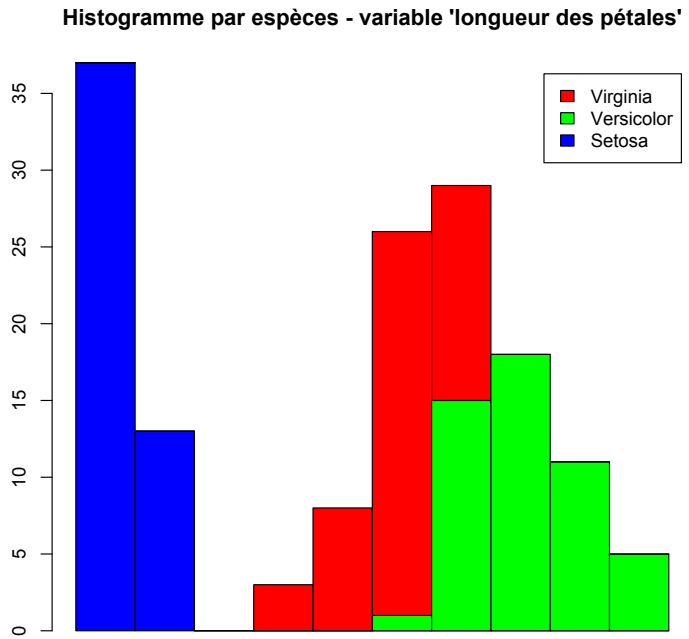
Résumés numériques ou graphiques des données

https://www.rpubs.com/cparoissin/iris

Exemple

Données : les iris de Fisher

- 3 espèces
  - Setosa
  - Virginica
  - Versicolor
- 4 variables
  - longueur des pétales
  - largeur des pétales
  - longueur des sépales
  - largeur des sépales



Méthodes factorielles

Objectif

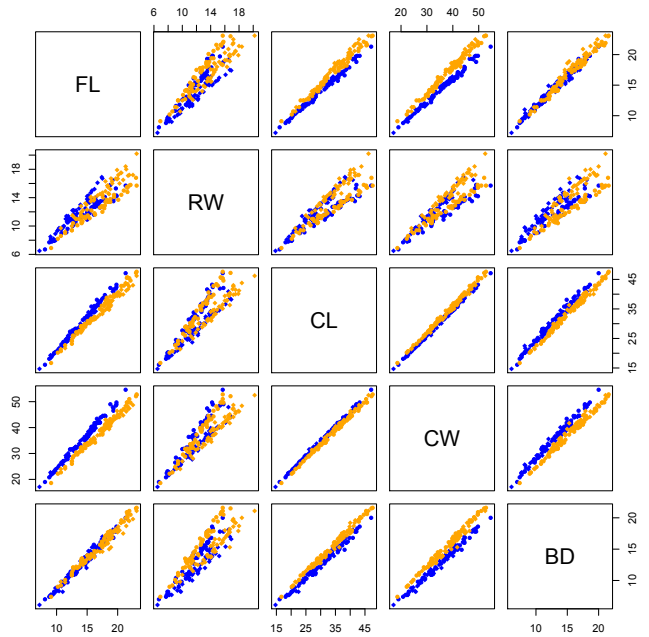
Remplacer D variables (souvent redondantes) par un nombre réduit de nouvelles variables en conservant le maximum d'information

http://www.stats.ox.ac.uk/~sejdinov/teaching/dmml17/PCA_crabs.html

Exemple

Données : 200 crabes

- 5 variables
  - longueur de la carapace
  - hauteur du crabe
  - ...
- 2 espèces
  - orange
  - bleue



Méthodes factorielles

Objectif

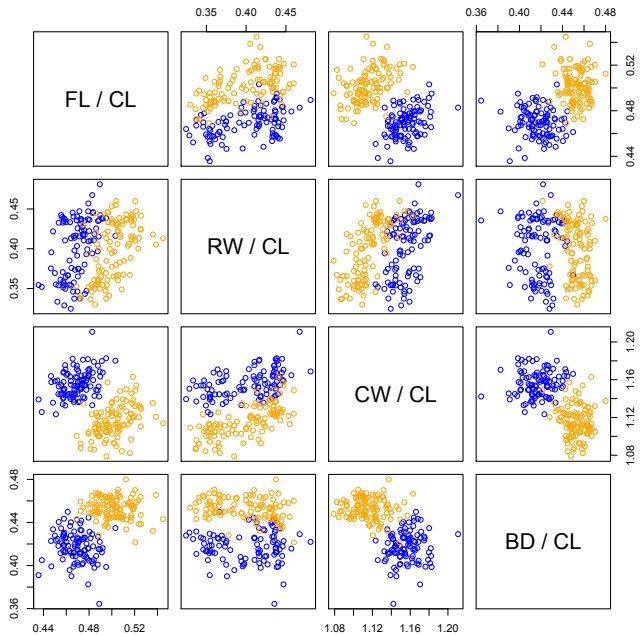
Remplacer D variables (souvent redondantes) par un nombre réduit de nouvelles variables en conservant le maximum d'information

http://www.stats.ox.ac.uk/~sejdinov/teaching/dmml17/PCA_crabs.html

Exemple

Données : 200 crabes

- 5 variables
  - longueur de la carapace
  - hauteur du crabe
  - ...
- 2 espèces
  - orange
  - bleue



Méthodes factorielles

Objectif

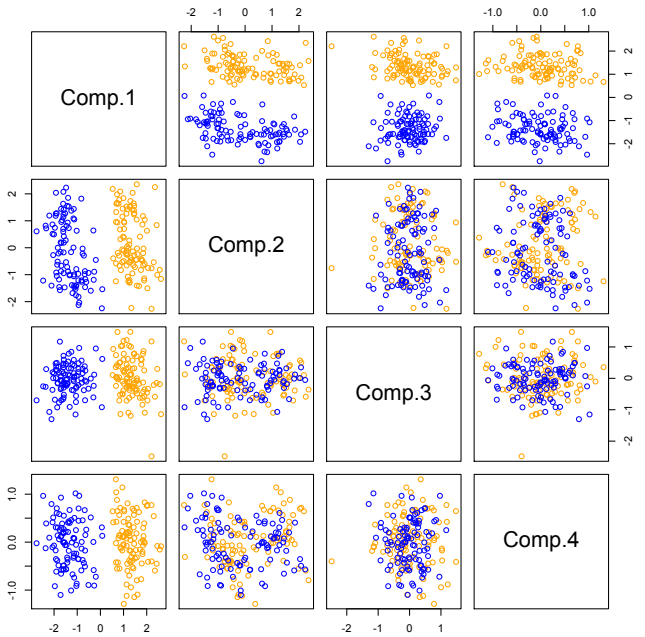
Remplacer D variables (souvent redondantes) par un nombre réduit de nouvelles variables en conservant le maximum d'information

http://www.stats.ox.ac.uk/~sejdinov/teaching/dmml17/PCA_crabs.html

Exemple

Données : 200 crabes

- 5 variables
  - longueur de la carapace
  - hauteur du crabe
  - ...
- 2 espèces
  - orange
  - bleue



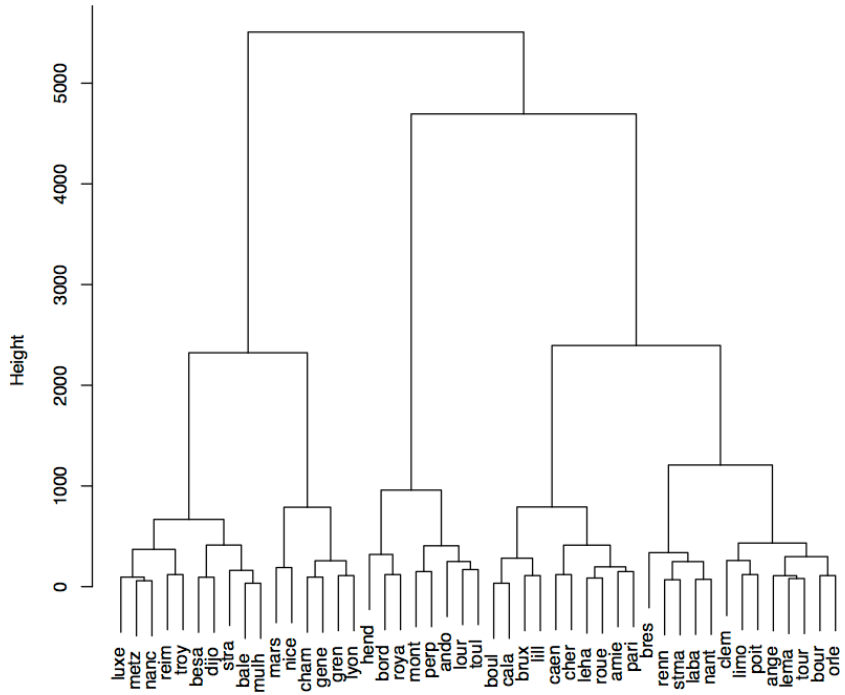
Méthodes de classification (clustering – non supervisé)

Objectif

Identifier par une procédure automatique des classes « naturelles »

Exemple

47 villes identifiées par les distances kilométriques qui les séparent des autres

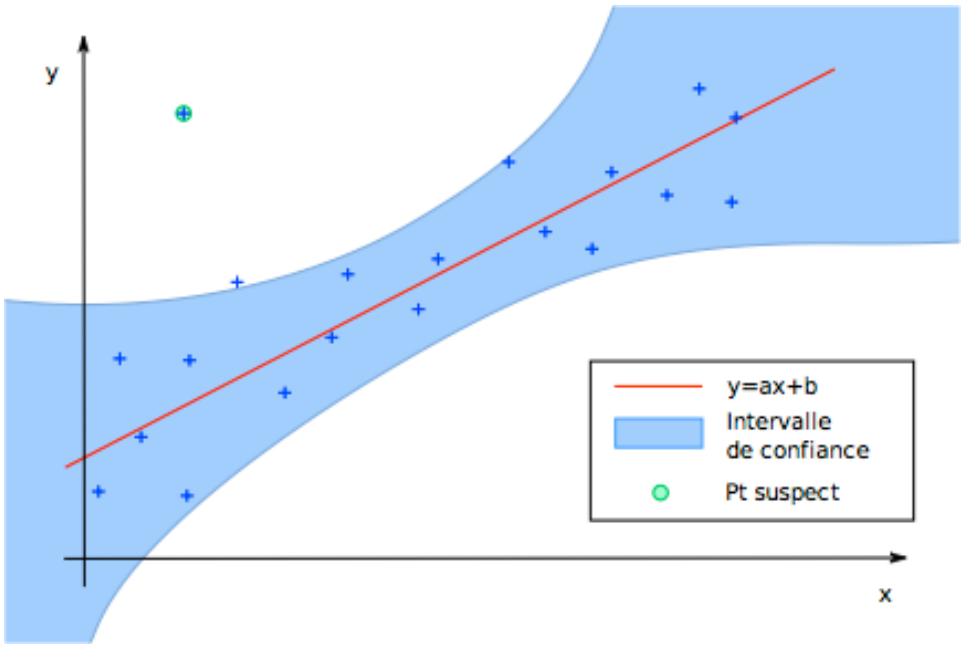


Méthodes de régression (supervisé)

Objectif

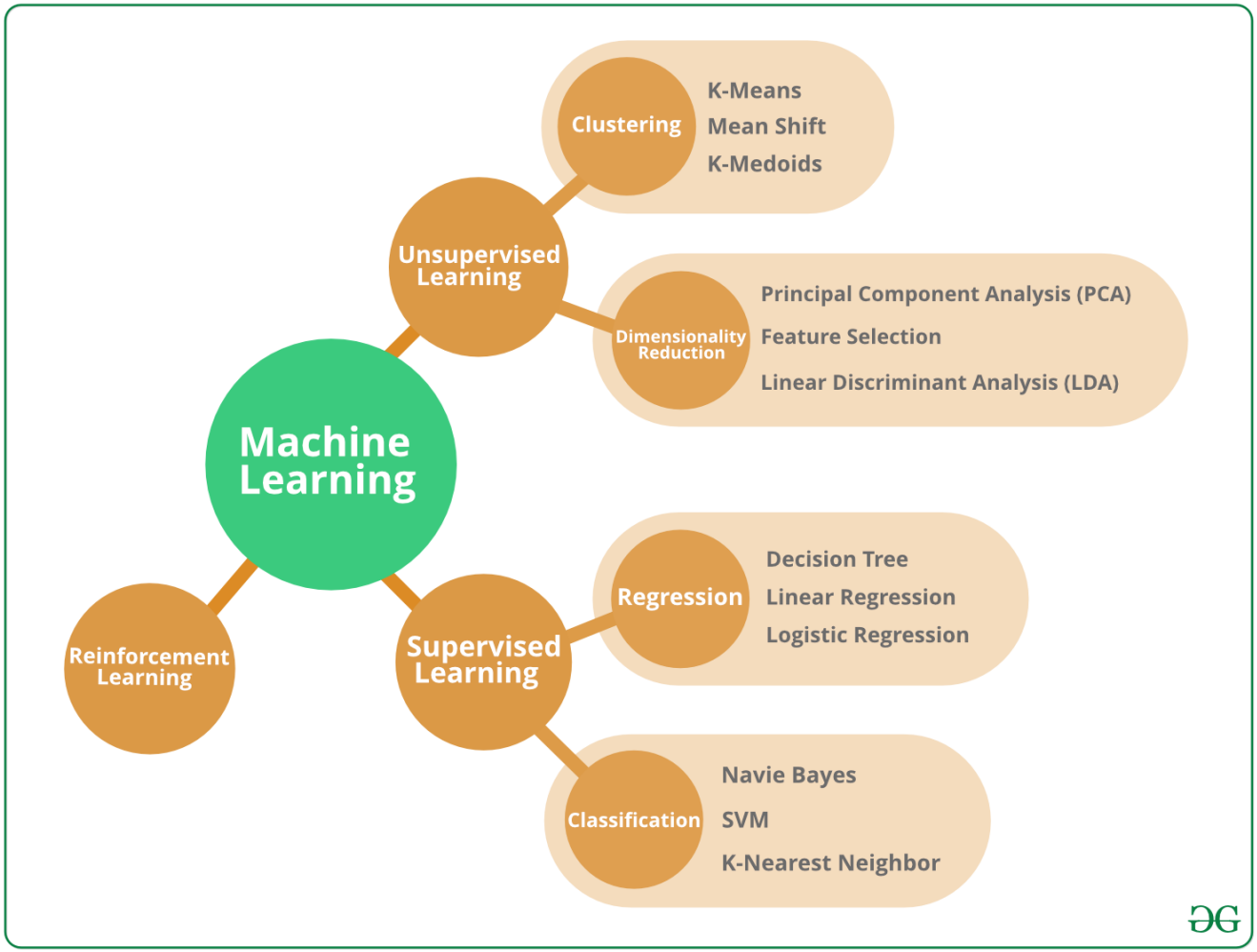
Étudier la dépendance entre une ou plusieurs variables *explicatives* et une variable *à expliquer*

Exemple



Différentes classes de méthodes

Une déclinaison partielle



Quelles classes de méthodes dans ce cours

- Méthodes descriptives

  $\rightsquigarrow$ à connaître !

  `http://wikistat.fr/pdf/st-l-Intro-statElem.pdf`

  + `http://wikistat.fr/pdf/st-l-des-uni.pdf`

  + `http://wikistat.fr/pdf/st-l-des-bi.pdf`

- Méthodes factorielles

  $\rightsquigarrow$ un champ à part entière, à connaître aussi...

  `http://wikistat.fr/pdf/st-m-Intro-ExploMultidim.pdf`

- Méthodes supervisées

  $\rightsquigarrow$ un focus sur la *régression* et les *arbres de décisions*

- Méthodes non supervisées

  $\rightsquigarrow$ un focus sur les K*-means* et la *classification hiérarchique*

Informations pratiques

Logistique

- Créneaux ⇝ lundi de 09h30 à 17h30
- Supports https://pydio.pedago.ensiie.fr/ ⇝ /pub/FISE_PYDS35
- Rendus https://exam.ensiie.fr

Projet : mise en place d'un protocole rigoureux

https://github.com/rfordatascience/tidytuesday/tree/master/data

1. Jeu de données avec problématique — 1 page : 09/10/2023
2. Description du jeu de données — 2 pages : 16/10/2023
3. Exploration du jeu de données — 2 pages : 06/11/2023

\+ Comparaison de méthodes supervisées — 2 pages : 06/11/2023

\+ Notebook `Python` — 06/11/2023

Informations pratiques

Planning prévisionnel

Séances	Matin	∼ 3h30	Après-midi	∼ 3h30
02/10	J1. Méthodologie	// TP	J1. Méthodologie	// TP + projets
09/10	J2. Supervisé	// TP	J2. Supervisé	// TP + projets
16/10	J3. Non supervisé	// TP	J3. Non supervisé	// TP + projets
06/11	Soutenances projets		Soutenances projets	

Intervenants

- Kylliann De Santiago
- Marie Szafranski

Informations pratiques

Language pour les TP et le projet

`Python`

Ceci n'est pas un cours de `Python`

Pratique avec python

`https://github.com/paris-saclay-cds/data-science-workshop-2019`

- `Numpy`, `Pandas`, `Matplolib`, `Seaborn` — maniplation : day 1
- `Scikit-learn` — ML : day 2

Références I

Leo Breiman. Statistical modeling : The two cultures. *Statistical Science*, 16 :199–231, 2001. URL `https://doi.org/10.1214/ss/1009213726`.

Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence, 1996.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

Team Wikistat. Wikistat, 2016. URL `http://wikistat.fr/`.