Analyse statistique de données

K. BELATTAR,
belattar.alger1@gmail.com

Département Informatique - Université d'Alger 1

Présentation de la matière

Semestre : 01

Intitulé de l'UE : UEF2 (Unité Enseignement Fondamentale)

Intitulé de la matière: Analyse statistique de données

Nombre de crédits : 04

Coefficient de la matière : 02

Site : https://canvas.instructure.com/enroll/MJE6GE

Résumé

Prérequis

- Statistiques et probabilités
- Algèbre linéaire (espace vectoriel)

Objectifs du cours

- Se familiariser avec le domaine de l'analyse de données,
- Organiser n'importe quelle masse de données,
- Préparer et nettoyer les données,
- Résumer les données,
- Trouver les relations qui peuvent exister entre les données,
- Savoir maitriser des nouvelles techniques et outils permettant d'analyser les données.

Evaluation du cours (interrogation écrite + examen final)

Contenu du cours

Cours 1: Introduction à l'analyse de données

Cours 2: Rappel des concepts de base de statistiques et probabilités

Cours 3: Analyse prédictive (régression multiple)

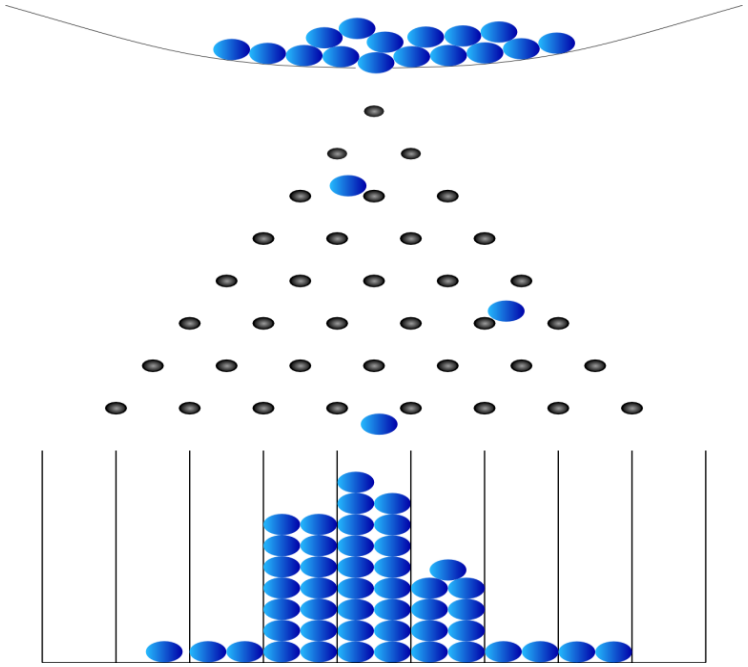
Cours 4: Analyse descriptive (réduction de dimensionnalité (ACP))

Introduction

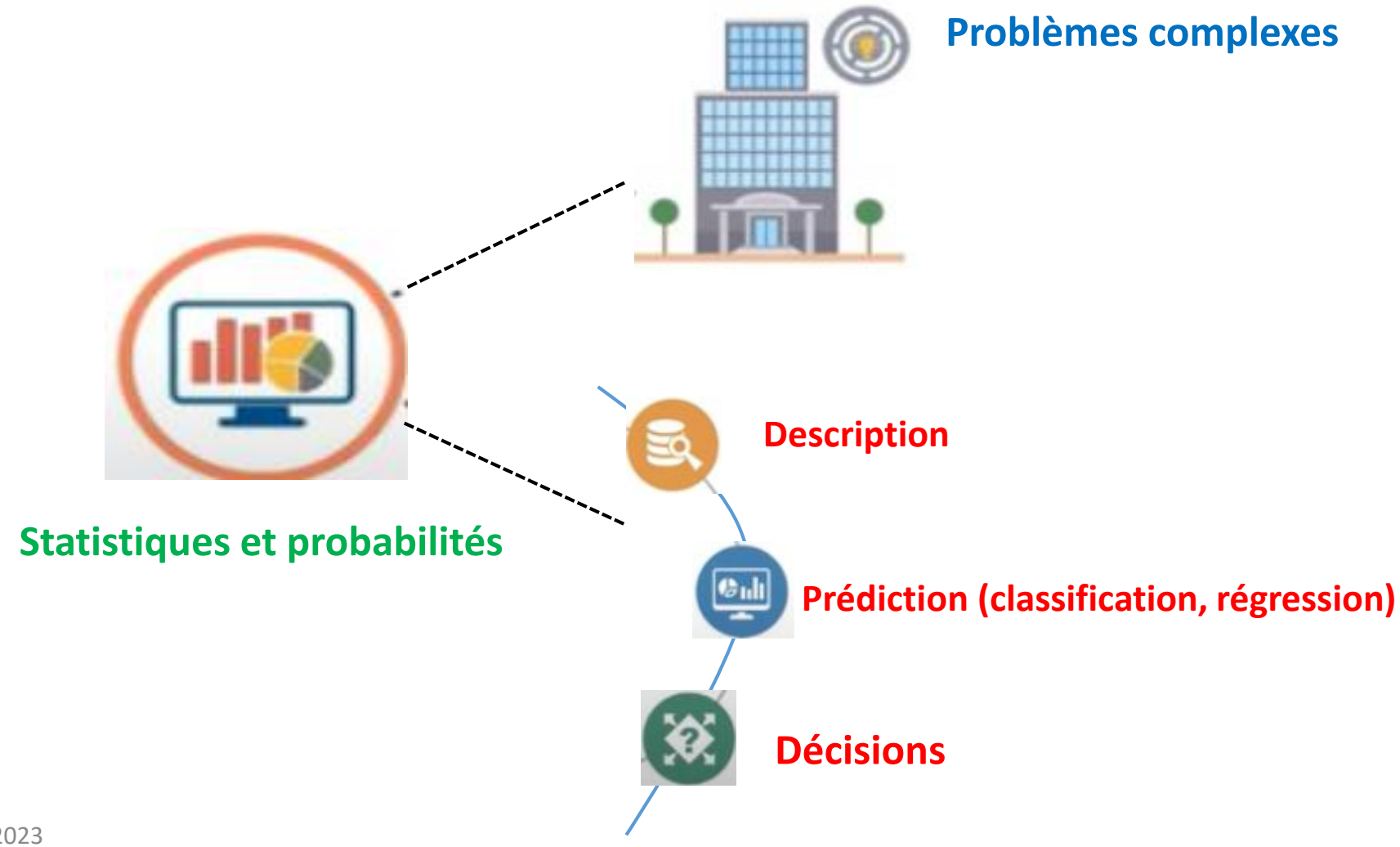
- Dans notre vie, la majorité phénomènes naturels sont caractérisés par **le hasard et l'incertitude**.

- Phénomènes observés **non prévisibles à l'avance avec certitude**.

→ Pouvoir présenter le résultat de déroulement d'un phénomène sous forme **des estimations de probabilités** en assumant des échantillons de données assez **larges et représentatifs**.

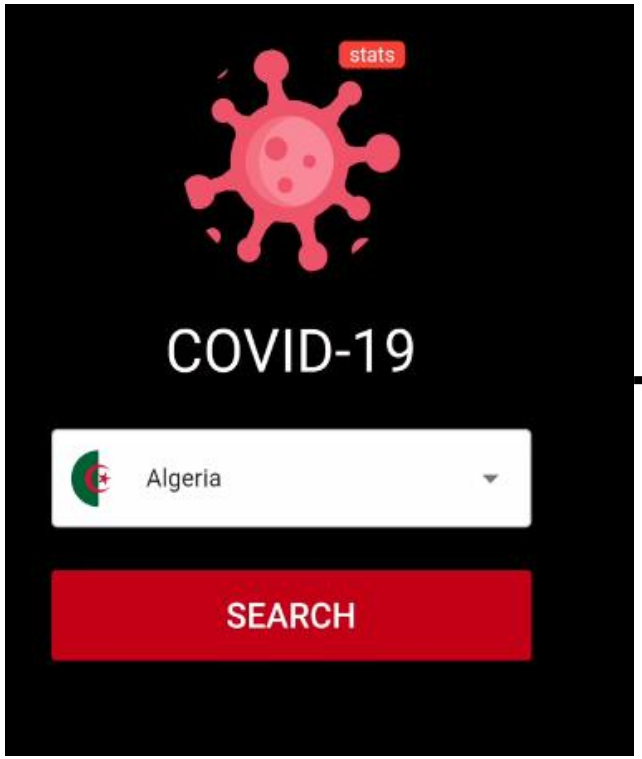


Introduction

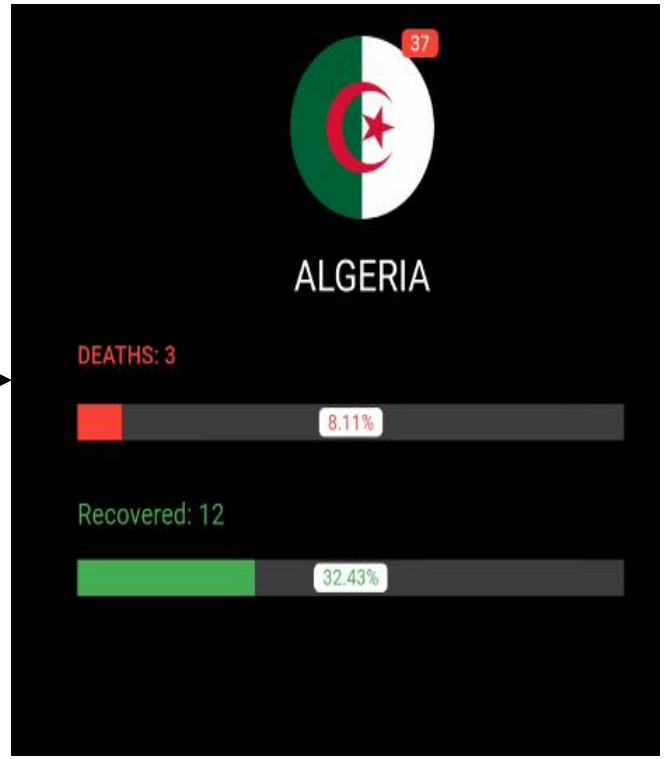


Statistiques

Exemple 1 d'un problème réel :



Analyse statistique de données



- Taux d'infection au Covid19
- Taux de propagation du Covid19

Introduction

<u>Exemple 2 d'un problème réel</u> :

Un fabricant d'ampoules produit environ un demi million (soit 500 000) d'ampoules par jour.
Le service de contrôle qualité doit estimer le taux de non-conformité (défaut) des ampoules.
C'est la statistique qui va permettre de résoudre ce type de problème. Pour cela, il faut expérimenter, recueillir, traiter et analyser des données [Gaudoin].

Introduction

Exemple 2 d'un problème réel

(1) Tester chaque ampoule → coûteux.

(2) Sélectionner un échantillon de 1000 ampoules de la production quotidienne de 500000 ampoules et à tester chacune des 1000 ampoules.

Calculer la fraction d'ampoules défectueuses (non conformes) dans les 1000 ampoules testées,

Utiliser ce taux comme une estimation de la fraction défectueuse dans la production de la journée entière, à condition que les 1000 ampoules sélectionnées soient représentatives.

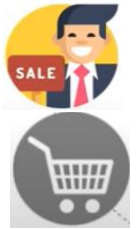
➡ Déterminer des ajustements dans le processus de production des ampoules.

Application des statistiques (et des probabilités)

Biologie, médecine, sciences humaines

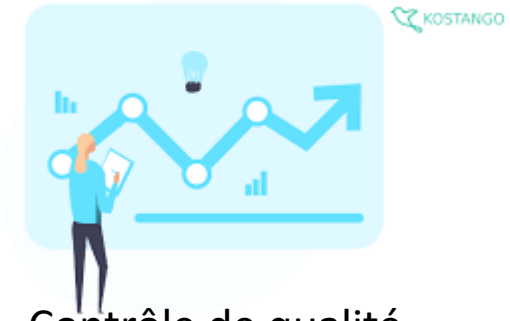
- Analyse de génomes,
- Dynamique d'une population,
- Etude durée de vie des individus,
- etc.

Sciences de l'ingénieur



- Prévision de ventes d'un produit,
- Analyse du comportement des clients,
- etc.

Industrie



Contrôle de qualité

Economie, assurance, finance

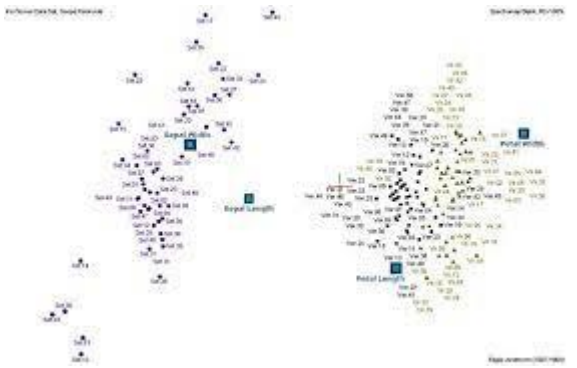
- Analyse de la consommation des ménages,
- Prévisions économétriques,
- Etudes quantitatives de marchés,
- etc.

Application des statistiques (et des probabilités)

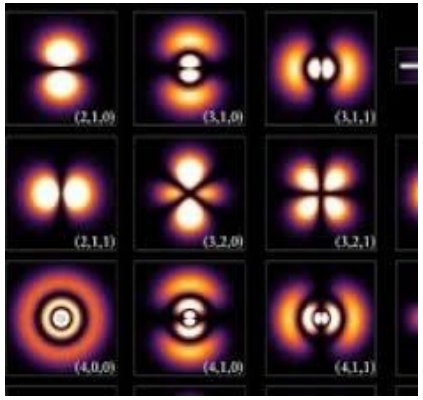
Géosciences

- Prévisions météorologiques,
- Exploration pétrolière,
- etc.

Informatique (Analyse exploratoire de données)



Physique (mécanique quantique)



Jeux de hasard

Statistiques

Statistiques fait référence à « un ensemble de principes et des techniques mathématiques utiles pour tirer (extrapoler) des conclusions relatives à l'étude d'un phénomène aléatoire, sur des données variables et entachées d'incertitudes (erreur) » [Ott].

Statistiques

Ces données numériques peuvent etre de toute nature, sous la forme :

- chiffres de ventes trimestriels,
- pourcentage d'augmentation de la criminalité,
- niveaux de contamination dans les échantillons d'eau,
- taux de survie des patients sous thérapie médical,
- chiffres de recensement,
- Des données permettant de déterminer la marque de voiture à acheter.
- Etc.

Statistiques

4) Présentation de données

1) Collection de données

Statistiques

2) Analyse de données (outils)

- **Inspecter les données**
- **Nettoyer les données**
- **Transformer les données**
- **Modéliser les données**

3) Interprétation de données

Analyse de données multidimensionnelles

Analyse statistique

- Appliquer des méthodes statistiques sur des échantillons ou ensemble de données (historiques) afin de faire des inférences et comprendre le caractère de la population étudiée.
- **Analyse quantitative** (généralement, des **données numériques** ayant une **distribution normale**)

Analyse non statistique

- Extraire des informations pertinentes à partir du texte, du son, des images et des vidéos (données volumineuses).
- **Analyse qualitative** (**données non numériques**)

Analyse de données multidimensionnelles

Analyse statistique

(1) Méthodes d'analyse descriptive (exploratoire)

- Méthodes d'analyse factorielles (Analyse en Composante Principale et Analyse Factorielle Multiple)
- Apprentissage non supervisé (Clustering)

(2) Méthodes d'analyse inférentielle
(3) Méthodes d'analyse prédictive

Objectif: Vérifier des patterns dans un ensemble de données

Analyse non statistique

Techniques de fouilles de données (text mining, Web mining, Graph mining, temporal mining, mulitmédia mining, etc.)

Objectif: Extraire des patterns (motifs) dans un ensemble de données afin de distinguer des relations imprévues et inconnues.

Analyse statistique descriptive de données

- Résumer les données (contenue dans des jeux de données) en leur assignant une nouvelle représentation,
- Synthétiser les données en faisant ressortir ce qui est dissimulé par le grand volume de données.
- Trouver les individus les plus proches ou les plus éloignés entre eux ;
- mais aussi trouver les exceptions ou les cas atypiques.
- Egalement détecter les variables liées (corrélées),
- Expliquer une variable en fonction des autres,
- Repérer les variables les plus influentes,
- ou encore regrouper les individus dans des catégories.

Dégager les caractéristiques essentielles du phénomène étudié sous une forme **simple, claire et efficace**.

- Les **représentations graphiques** de données et
- Les **indicateurs statistiques** numériques

Analyse statistique descriptive de données

Exemple 1:

- Etudier la teneur en huile et le niveau de coloration (faible ou élevé) des olives.

https://delladata.fr/les-3-principaux-types-danalyses-statistiques/

- Prélever un échantillon de 500 olives sur la production d'une oliveraie donnée.

→ Faire une analyse statistique descriptive consiste à calculer des paramètres statistiques (max, min, moyenne, etc.) qui vont résumer des valeurs observées.

Analyse statistique inférentielle de données

Généraliser, à l'échelle de la population, des conclusions tirées à partir des données (large dataset) **d'un échantillon représentatif.**

→ Estimer des paramètres de la population en se basant sur **des modèles probabilistes du phénomène aléatoire étudié**
→ **Tests d'hypothèses sur les données**
→ **Comparer la différence entre deux échantillons**

Dataset large

Théorie de probabilité

Analyse statistique inférentielle de données

Exemple 2: comparer deux échantillons

- Disposer deux échantillons (chacun de taille 500) composé d'olives de deux espèces différents.
- **Les olives de l'espère 1 ne contient pas plus d'huile que les olives de l'espèce 2?**

Espèce 1

Espèce 2

Analyse statistique prédictive de données

Analyser les données actuelles afin de faire des hypothèses sur des comportements futurs des individus déjà présents mais aussi de nouveaux individus.

- Disposer, au préalable, jeu de données (échantillonné en un ensemble d'apprentissage et du test) et des algorithmes d'apprentissage automatique.

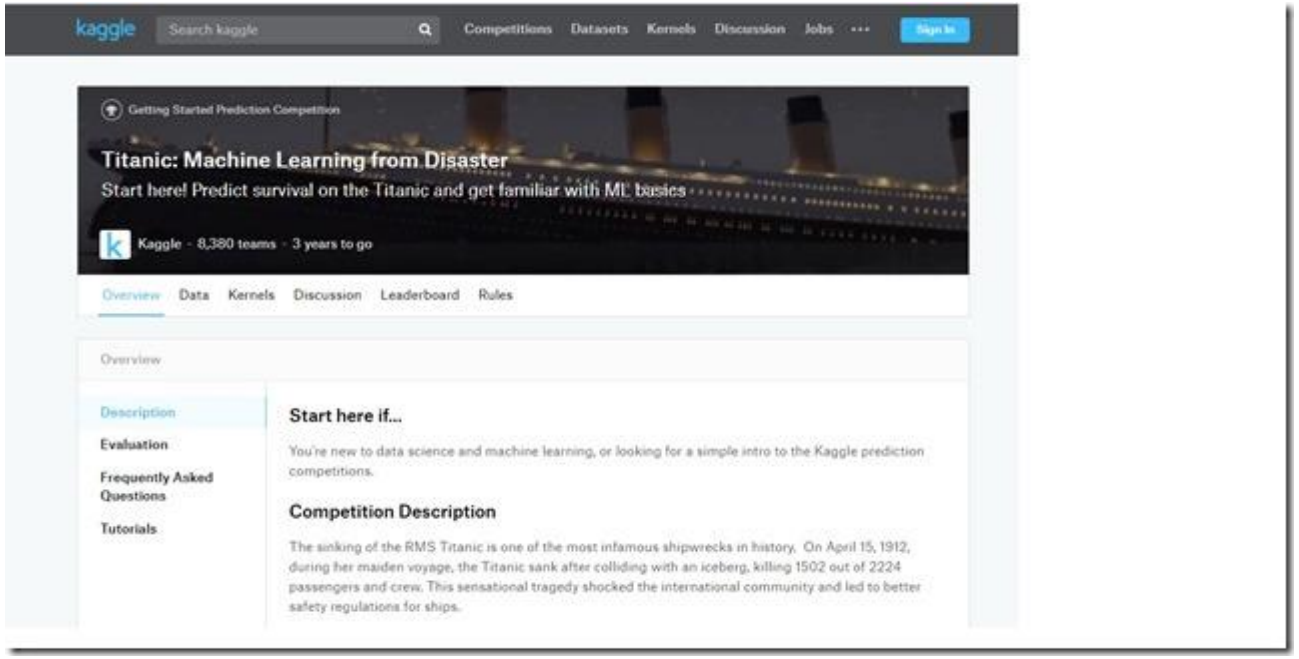
L'apprentissage automatique est un ensemble de techniques puissantes permettant de créer des modèles **prédictifs et/ou descriptifs** à partir de données, **sans avoir été explicitement programmées**.

(1) Algorithmes de **régression**
(2) Algorithmes de **classification**

→ **Prédire** des données de type numérique ou catégorique.

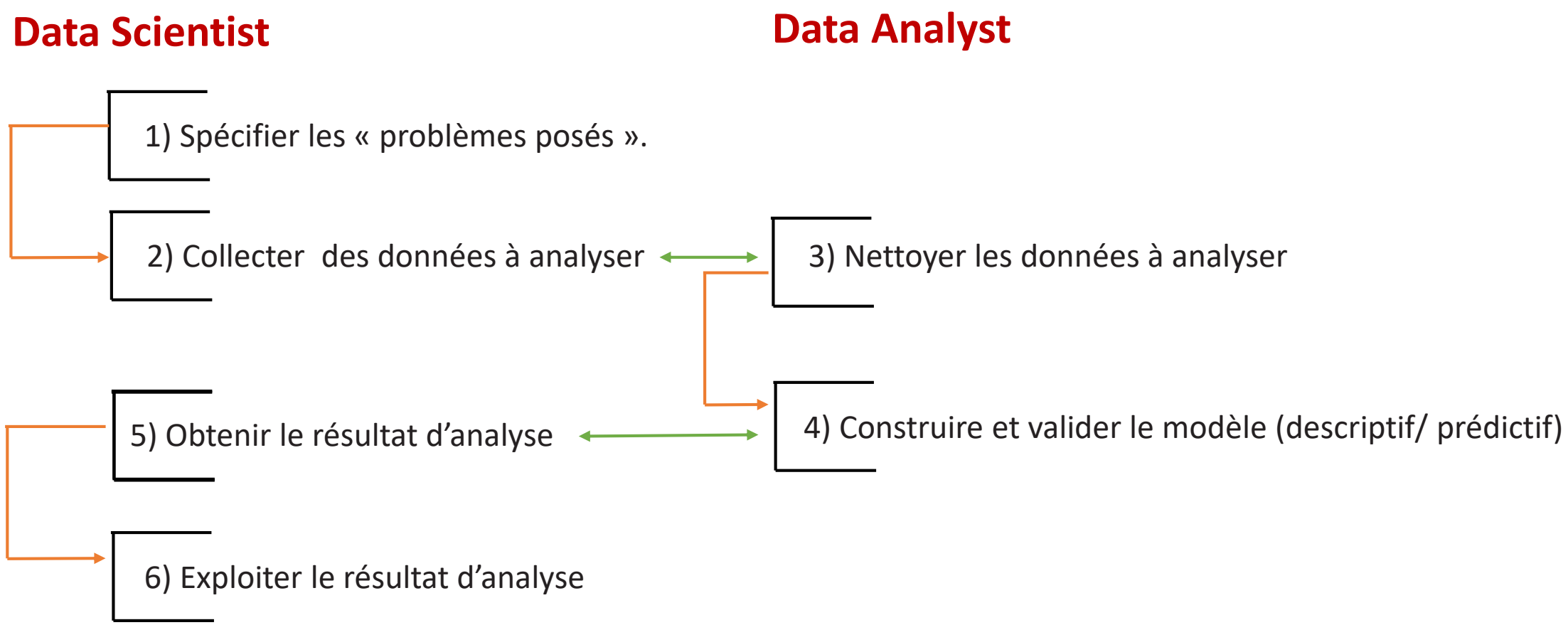
Analyse statistique prédictive de données

Exemple 3: prédire la consommation en électricité d'une famille en fonction du jour de la semaine, de la température, du vent, de la pression atmosphérique, et de la quantité de pluie des 5 jours précédents.



https://www.kaggle.com/competitions

Démarche d'analyse de données



Enjeu actuel

Est d’avoir des outils qui :

1 - Analysent plus de données,

2 - Plus vite et

3 - Surtout qui intègrent facilement des données plus hétérogènes et moins structurées (cas réel),

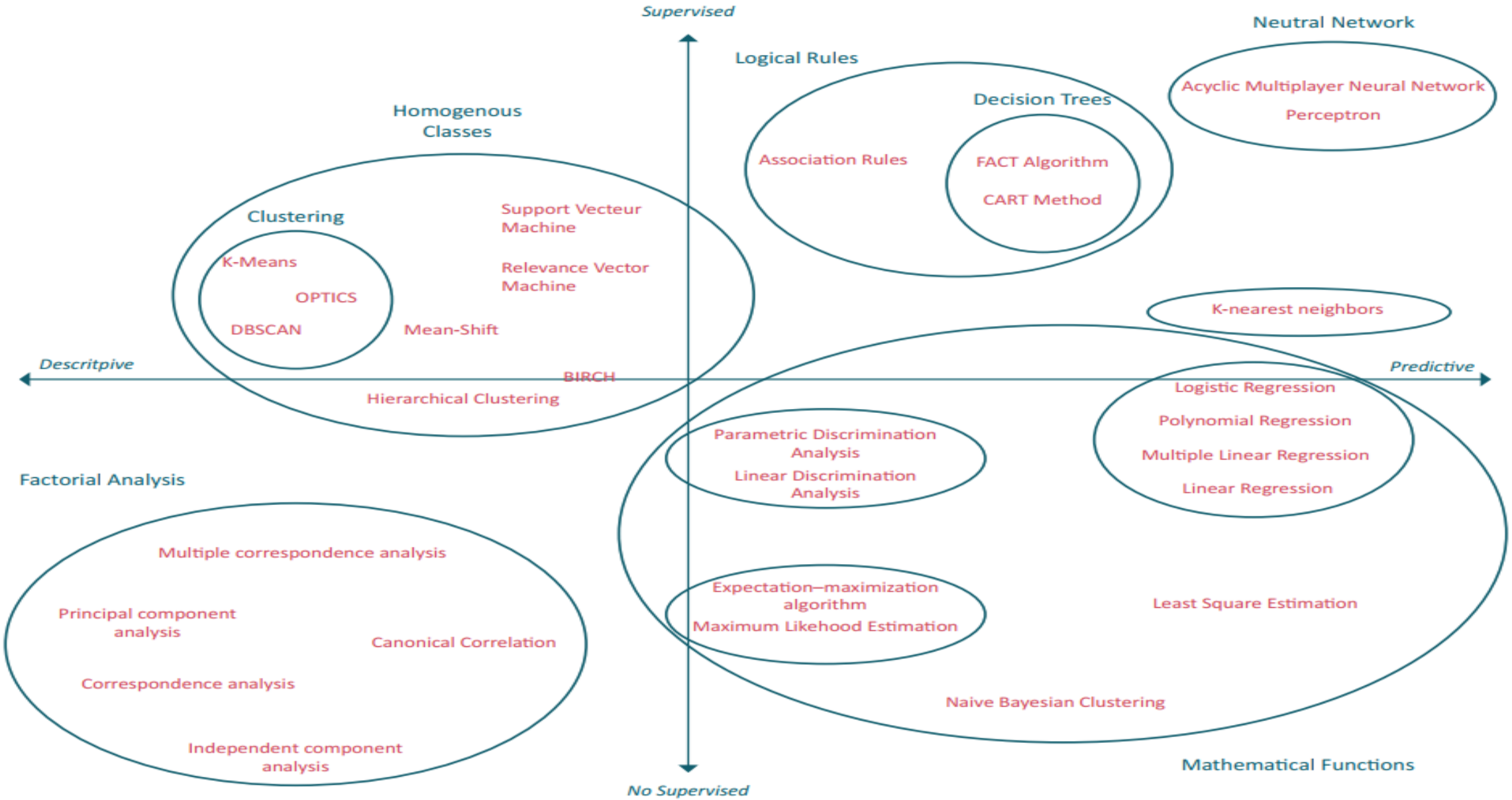
4- Analysent en direct un flux continu de données afin de pouvoir exploiter les résultats en temps réel. (Certains commencent à apparaître).

Outils d’analyse de données

✓ **Langages interprétés d’analyse de données** (R, Python ou encore Matlab)

(+) Offrir une très large bibliothèque de fonctions statistiques extensibles avec des packages grâce auxquels on peut faire appel à des algorithmes déjà implémentés.
(-) Aucun contrôle sur les paramètres de l’algorithme utilisé
(-) Non validation du modèle d’analyse de données

✓ **Logiciels d’analyse statistique** avec une interface simple
Oracle E- Business Suite, Watson (édité par IBM), SAS BI, KNIME, RevolutionR de Revolution Analytics, Tableau ou encore DataIKU, Visio, Minitab and Stata.
(+) Générer le modèle d’analyse de données (avec moins de cout)


Supervised
Neutral Network
Logical Rules
Acyclic Multiplayer Neural Network
Perceptron
Decision Trees
Homogenous Classes
Association Rules
FACT Algorithm
CART Method
Clustering
Support Vecteur Machine
K-Means
Relevance Vector Machine
OPTICS
DBSCAN
Mean-Shift
K-nearest neighbors
Descritpive
Predictive
BIRCH
Hierarchical Clustering
Logistic Regression
Polynomial Regression
Multiple Linear Regression
Linear Regression
Parametric Discrimination Analysis
Linear Discrimination Analysis
Factorial Analysis
Multiple correspondence analysis
Expectation–maximization algorithm
Maximum Likehood Estimation
Least Square Estimation
Principal component analysis
Canonical Correlation
Correspondence analysis
Naive Bayesian Clustering
Independent component analysis
Mathematical Functions
No Supervised

Références bibliographiques

[1] Olivier Gaudoin. Principes et Méthodes Statistiques, notes de cours.

[2] R. Lyman Ott, Michael Longnecker. An introduction to statistical methods and data analysis (fifth edition), 2001.

[3] M. Maumy-Bertrand et F. Bertrand : Initiation à la statistique avec R. Dunod, 2e édition, 2014.

[4] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. No. 10. New York: Springer series in statistics, 2001.

[5] F. Grosjean, J.-Y. Dommergues et G. Macagno : La Statistique en clair. Ellipses, 2011.

[6] Hamon et N. Jégou : Statistique descriptive. Presses universitaires de Rennes, 2008.
[7] J.-J. Daudin, S. Robin et C. Vuillet : Statistique inférentielle : idées, démarches et exemples. Presses Universitaires de Rennes, 1999.

[8] Saporta Gilbert. Probabilités, analyse des données et statistique » 3e édition révisée (1990-2006) . Date de publication : Juillet 2011.

Références bibliographiques

https://www.youtube.com/watch?v=2rSEkwmYs4Q&list=PLCILw_sLfhbPo5-xfYrbo2-Cp0meZu-cN

https://www.youtube.com/watch?v=6Cl-gaNtZWA&list=PLCILw_sLfhbOvZRLDHiI6idPQwZJltAgH

https://www.youtube.com/watch?v=AN3UkzE3HMg&list=PLqzoL9-eJTNBZDG8jaNuhap1C9q6VHyVa

https://www.youtube.com/watch?v=xxpc-HPKN28