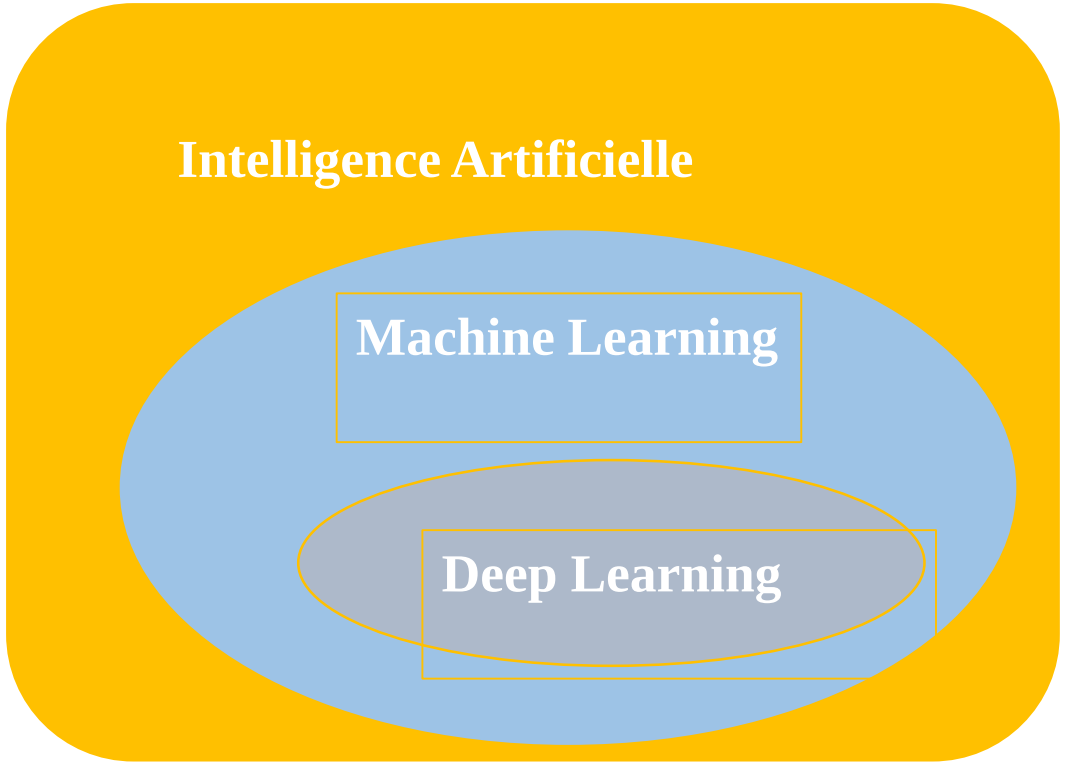
Eléments de data sciences

yncréa

Lidiya YUSHCHENKO

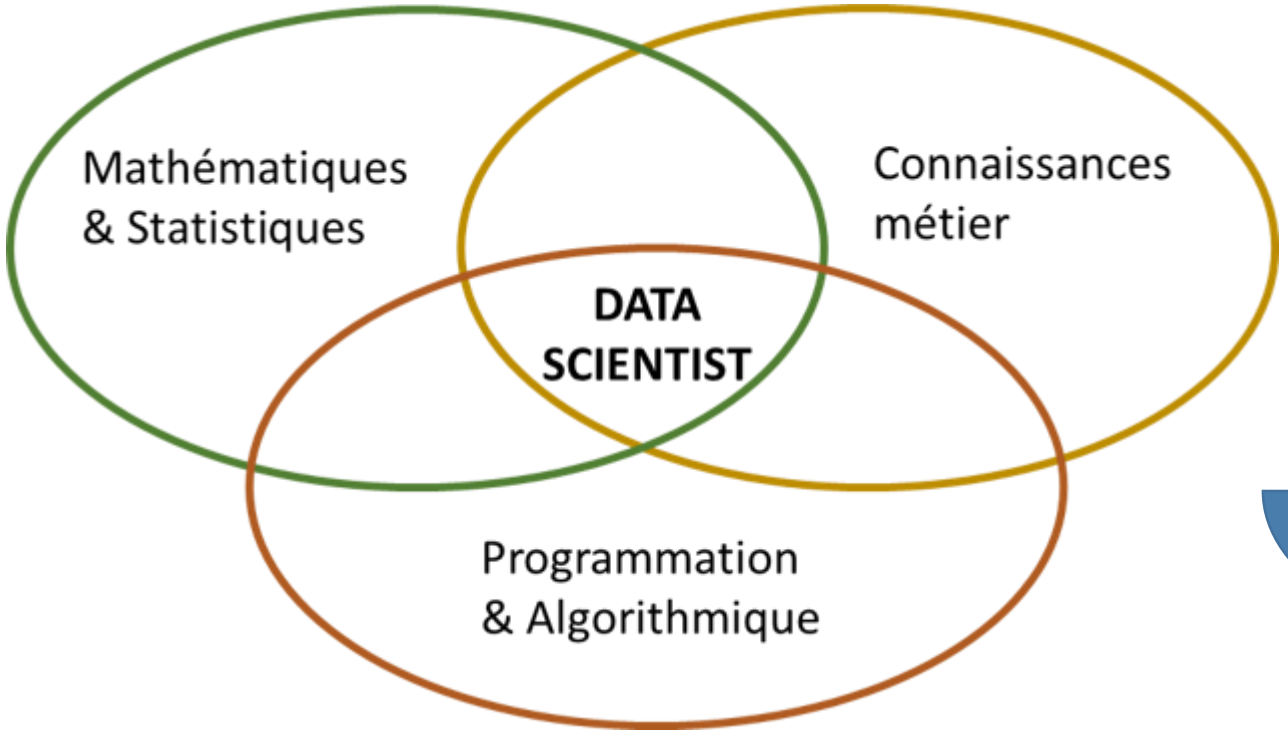
ISEN | école d'ingénieurs
ALL IS DIGITAL!



Data science

Data science

”Démarche empirique qui se base sur des données pour apporter une réponse à des problèmes”



Compétences nécessaires

Fondamentaux et études de cas, E. Biernat, M. Lutz, Eyrolles, 2015

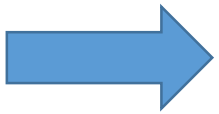
La donnée …

Introduction : Data Science – Science des Données

?

La magie de Data Science ?

Nombre suffisant de données concernant le fonctionnement de votre véhicule



un algorithme pourrait détecter une panne et vous suggérer de changer le démarreur

traduire votre problème humain en éléments qui puissent être analysés par un algorithme de machine learning

vous devez être capable de poser votre problème sous la forme d'un problème de data science

Résolution d'un problème à travers l'analyse de données

Formulation du problème (hypothèses à vérifier, phénomènes/quantités à prédire)

Récolte des données (appropriées)

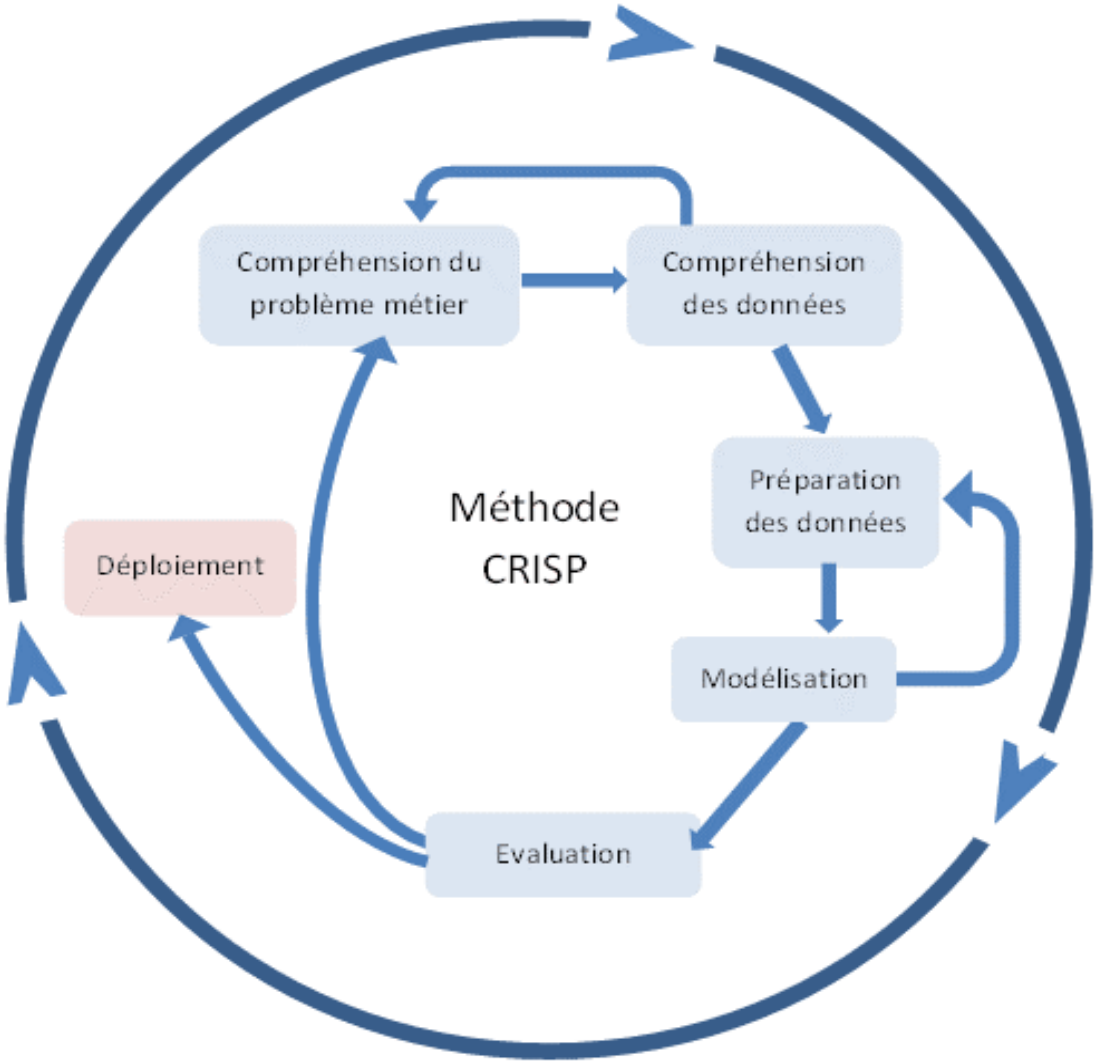
Analyse des données (préparation données, choix algorithmes, prédiction, …)

Utilisation des résultats d'analyse/modèle

Travaille avec les experts du métier (pour les données à analyser)

Travaille avec les services informatiques (récupération des données)

Appliquer une démarche projet spécifique :
méthode CRISP (Cross Industry Standard Process)



Une donnée …

Une donnée c'est « le résultat d'une observation faite sur une population ou sur un échantillon » (Dodge, 2007)

Une donnée est donc un nombre, une caractéristique, qui apporte une information sur un individu, un objet ou une observation.

35 – un nombre → J'ai 35 ans → 35 – une donnée

Types de données

Les données quantitatives

Ce sont des valeurs qui décrivent une quantité mesurable, sous la forme de nombres sur lesquels on peut faire des calculs (moyenne, etc.) et des comparaisons (égalité/différence, infériorité/supériorité, etc.). Elles répondent typiquement à des questions du type **« combien ».**

Données quantitatives continues
(la température, le PIB, le taux de chômage,…)

Données quantitatives discrètes
(le nombre d'enfants par famille,
le nombre de pièces d'un logement,…)

Les données qualitatives

Les données qualitatives décrivent des qualités ou des caractéristiques. Elles répondent à des questions de la forme **« quel type »** ou **« quelle catégorie »**.

Données qualitatives nominales
(la couleur des yeux (bleu, vert, marron, etc.),
le sexe (homme, femme))

Données qualitatives ordinales, dont les modalités sont ordonnées selon un ordre « logique ».
(les tailles de vêtements (S, M, L, XL))

Modélisation : étape centrale de l'analyse de données

Deux principales classes d'algorithmes

Algorithmes supervisés

Extraction de la connaissance à partir des données labélisées (couples entrée-sortie)

Algorithmes non supervisés

Organisation des données non labélisées en groupes homogènes

Les algorithmes

type de problème à traiter

algorithmes de régression

algorithmes de classification

La distinction régression/classification se fait au sujet des algorithmes supervisés

La sortie Y peut prendre une infinité de valeurs dans l'ensemble continu des réels (noté $Y \in \mathbb{R}$) (des températures, des tailles, des PIB, des taux de chômage,…)

La sortie Y prend un nombre fini k de valeurs ($Y = \{1, \ldots, k\}$). On parle alors d'étiquettes attribuées aux valeurs d'entrée. C'est le cas des valeurs de vérité de type OUI/NON ou MALADE/SAIN.

Algorithme	Mode d'apprentissage	Type de problème à traiter
Régression linéaire univariée	Supervisé	Régression
Régression linéaire multivariée	Supervisé	Régression
Régression polynomiale	Supervisé	Régression
Régression régularisée	Supervisé	Régression
Naive Bayes	Supervisé	Classification
Clustering hiérarchique	Non supervisé	-
Arbres de décision	Supervisé	Régression ou classification
Analyse en composantes principales	Non supervisé	-
K-Means	Non supervisé	-

Principales étapes de l’analyse de données

- **Compréhension et formulation du problème**

    - Compréhension du problème et de la donnée

    - Formulation des hypothèses et/ou quantités/phénomènes à prédire

- **Préparation des données**
  - Sélection des variables (colonnes)
  - Réduction de dimension
  - Transformation des variables
  - Recodage
  - …

Une variable, en général c'est une colonne d'un tableau de données.

C'est un objet de type *vecteur*.

Un vecteur est un ensemble d'éléments de même type.

→ Les vecteurs peuvent être de classes différentes, selon le type de données qu'ils contiennent.

La normalisation des données

Redimensionnement des variables numériques pour qu'elles soient comparables sur une échelle commune.

La normalisation

Soit une variable numérique à n observations $(x_1, x_2, \dots, x_n) \in \mathbb{R}$, alors $x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \in [0,1]$

La normalisation standard (standardisation)

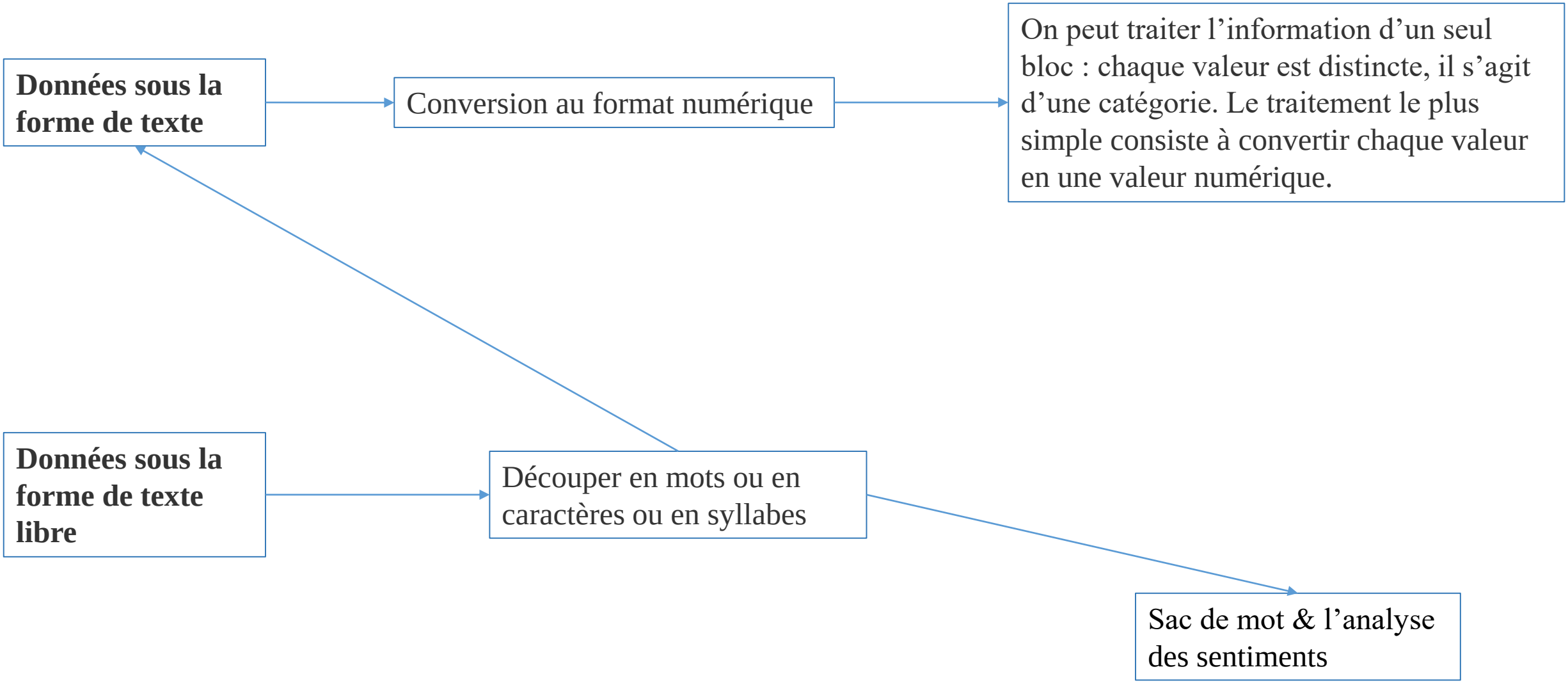
Soit une variable numérique à n observations $(x_1, x_2, \dots, x_n) \in \mathbb{R}$, alors $x_{standard} = \frac{x - x_{min}}{\sigma}$, où σ est l'écart-type.

« + »

Les données sont rarement comparables dans leur état brut

« - »

Attention /!\ l'étape de normalisation constitue une perte d'information dans l'immédiat et peut desservir dans certains cas !

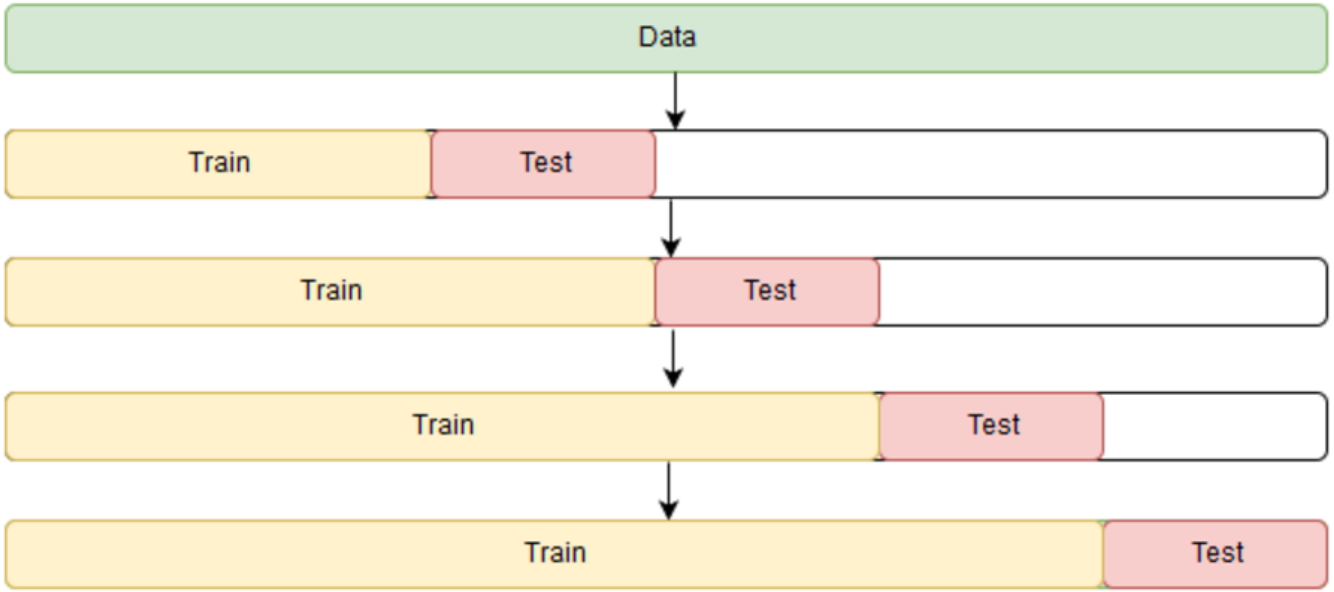

Données sous la forme de texte
Conversion au format numérique
On peut traiter l'information d'un seul bloc : chaque valeur est distincte, il s'agit d'une catégorie. Le traitement le plus simple consiste à convertir chaque valeur en une valeur numérique.
Données sous la forme de texte libre
Découper en mots ou en caractères ou en syllabes
Sac de mot & l'analyse des sentiments

- **Modélisation des données (utilisation d'algorithmes)**
  - Statistique descriptive - Exploration des données

L'objectif des outils de Statistique descriptive élémentaire est de fournir des résumés synthétique de séries de valeurs, adaptés à leur type (qualitatives ou quantitatives), et observées sur une population ou un échantillon.

  - Séparation des données (cross validation)

La Cross-Validation ou validation croisée est une méthode permettant de tester les performances d'un modèle prédictif de Machine Learning.



Principales techniques de validation croisée

La technique du Train-Test Split

Principe : décomposer de manière aléatoire un ensemble de données. Une partie servira à l'entraînement du modèle de Machine Learning, l'autre partie permettra de le tester pour la validation.

/!\
Technique est efficace, sauf si les données sont limitées.

La méthode K-Folds

Principe : séparer l'ensemble de données de manière aléatoire en *K* folds. La procédure a un paramètre unique appelé « *K* » faisant référence au nombre de groupes dans lequel l'échantillon sera divisé.
La valeur de *K* ne doit être ni trop basse ni trop haute, et on choisit généralement une valeur comprise entre 5 et 10 en fonction de l'envergure du dataset. Par exemple, si *K=9*, le dataset sera divisé en *9* parties.
Une valeur K plus élevée mène à un modèle moins biaisé, mais une variance trop large peut conduire à un sur-ajustement.

/!\
Technique est efficace, même si les données sont limitées.

- Evaluation & Amélioration du modèle :
  - Analyse de la qualité d'ajustement (fitting)
  - Analyse du pouvoir prédictif
    - Remarques : Underfitting & Overfitting

L'**Overfitting** (sur-apprentissage), et l'**Underfitting** (sous-apprentissage) sont les causes principales des mauvaises performances des modèles prédictifs générés par les algorithmes de Machine Learning.

Overfitting : un modèle trop spécialisé sur les données du Training Set et qui se généralisera mal

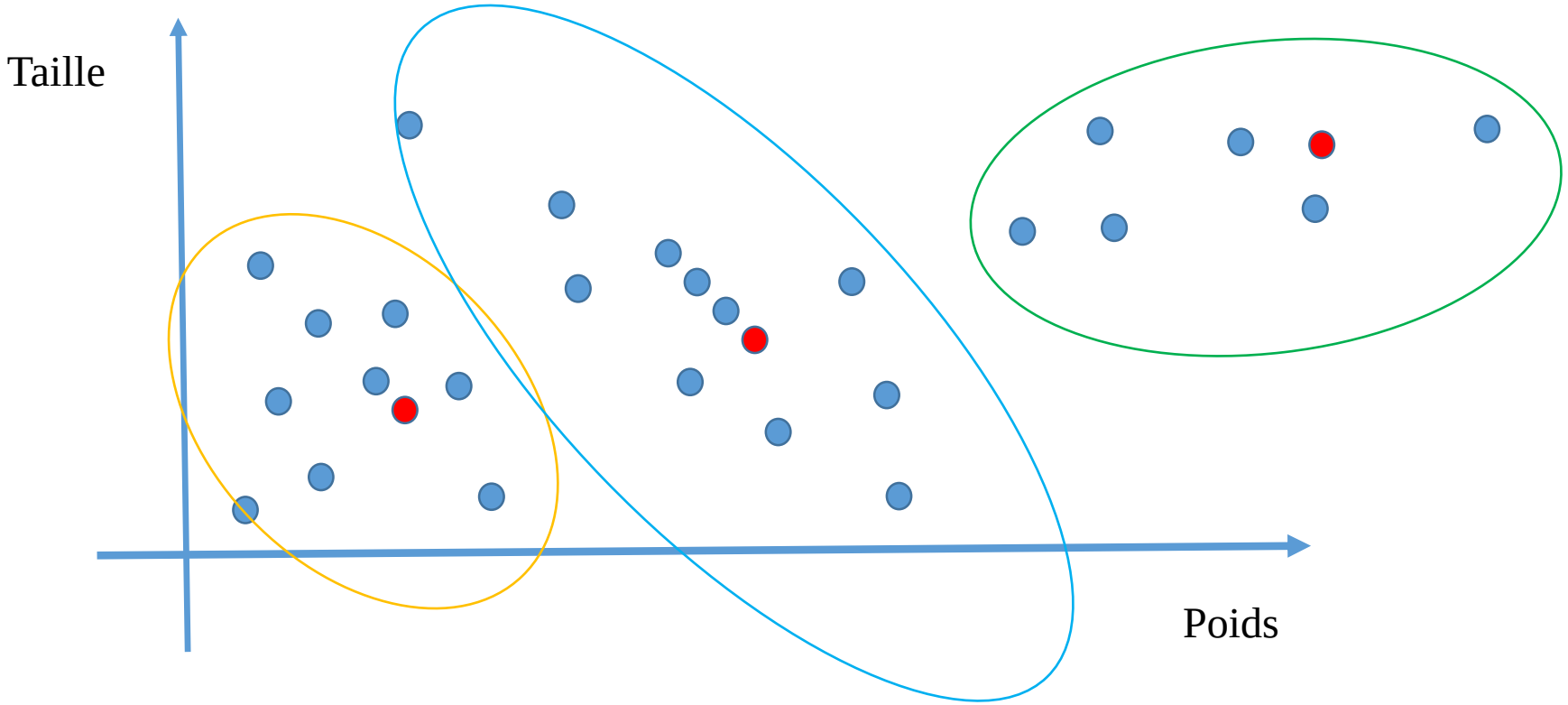
Underfitting, sous entend que le modèle prédictif généré lors de la phase d'apprentissage, s'adapte mal au *Training Set*.

Juste milieu

- Amélioration
  - La sélection du modèle
  - Amélioration des hyperparamètres (cross validation)

K-means (K-moyennes)

Un algorithme de clustering (regroupement, partitionnement de données).
En d'autres termes il permet de réaliser des analyses non supervisées, d'identifier un pattern au sein des données et de regrouper les individus ayant des caractéristiques similaires.



Domaines d’applications

Domaine	Forme des données	Clusters
Segmentation d’images	Images	Zones homogènes dans l’image
Bio informatique	Gènes	Gènes ressemblants
Marketing	Info clients, produits achetés	Segmentation de la clientèle

Remarques/Exemple

K-means. C’est un algorithme qui s’applique sur des variables quantitatives uniquement.

Avant de se lancer dans la classification, il faut déterminer le nombre de cluster que l’on souhaite obtenir.

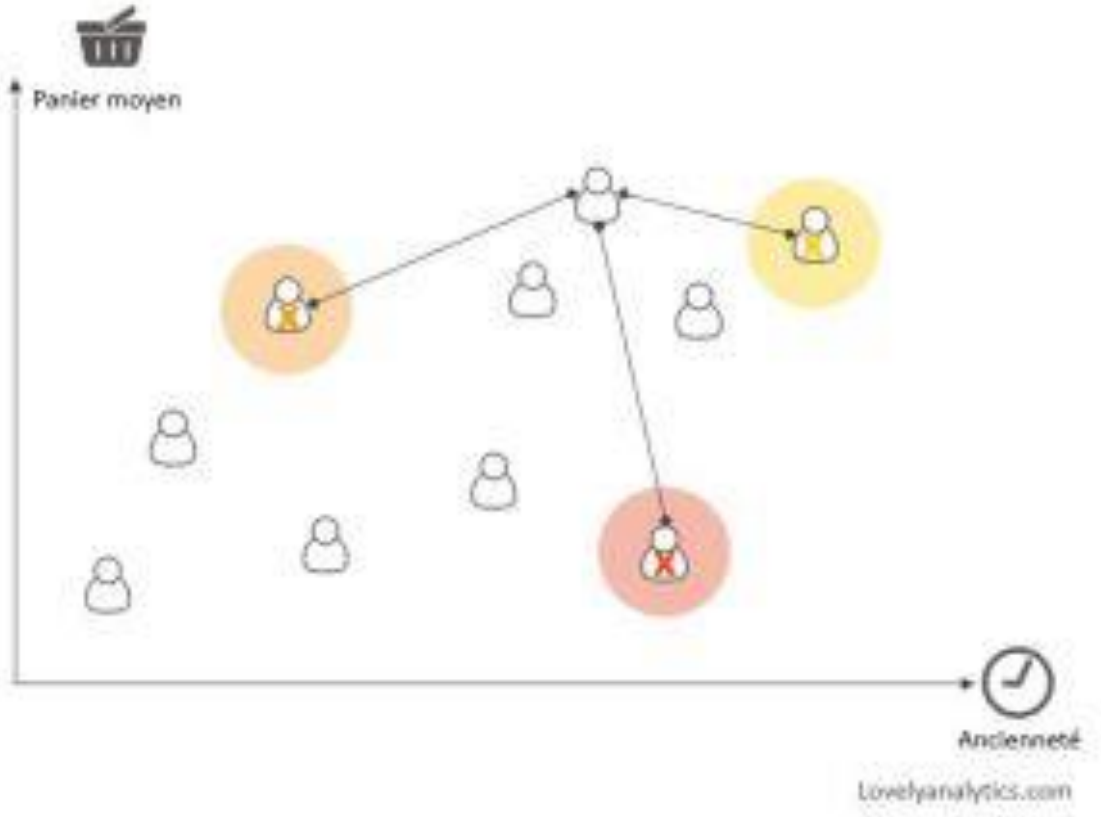
Prenons un exemple simple avec une base de 10 clients pour lesquels on connait l’ancienneté et le panier moyen. On souhaite créer 3 groupes en utilisant la méthode des K-means.

Voilà comment fonctionne l'algorithme :

Etape 0 : Initialisation

On tire aléatoirement 3 individus. Ces 3 individus correspondent aux centres initiaux des 3 classes.

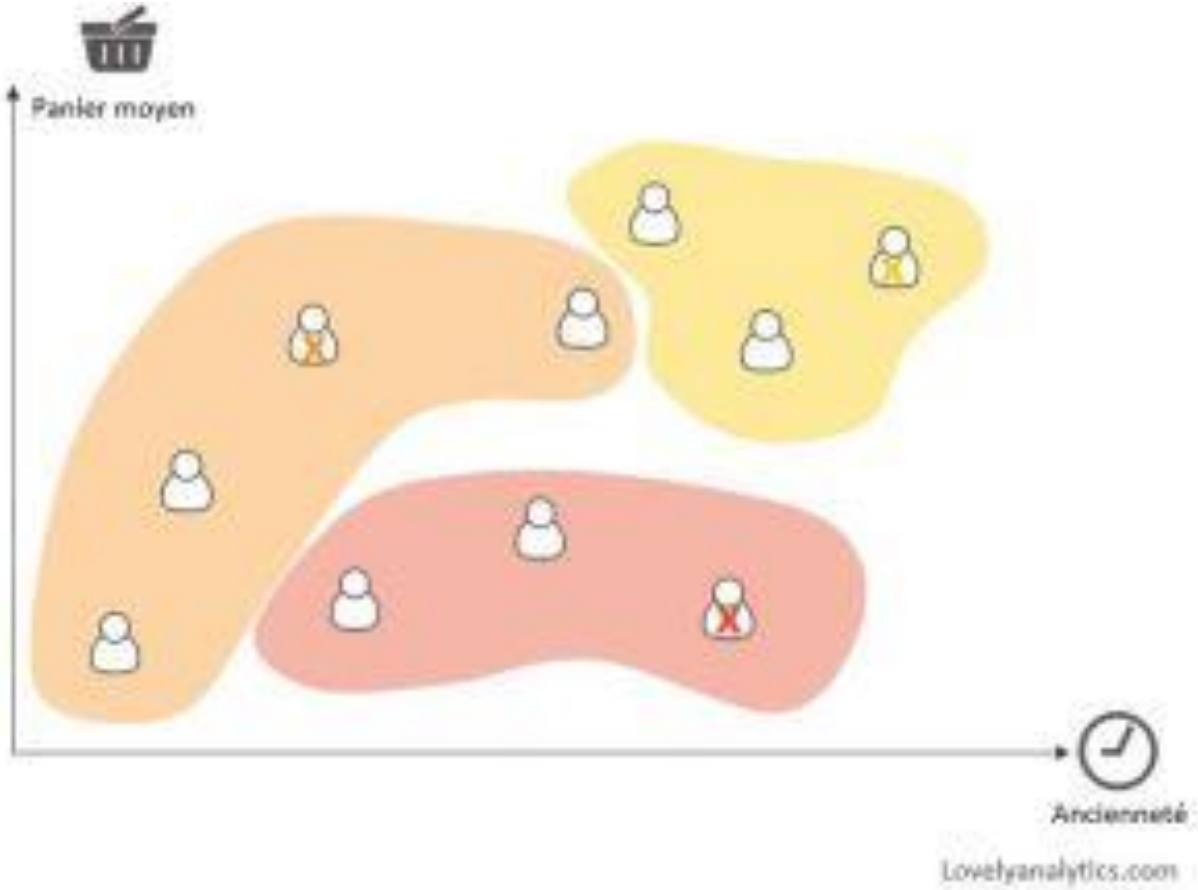
Etape 1 :



On calcule la distance entre les individus et chaque centre.

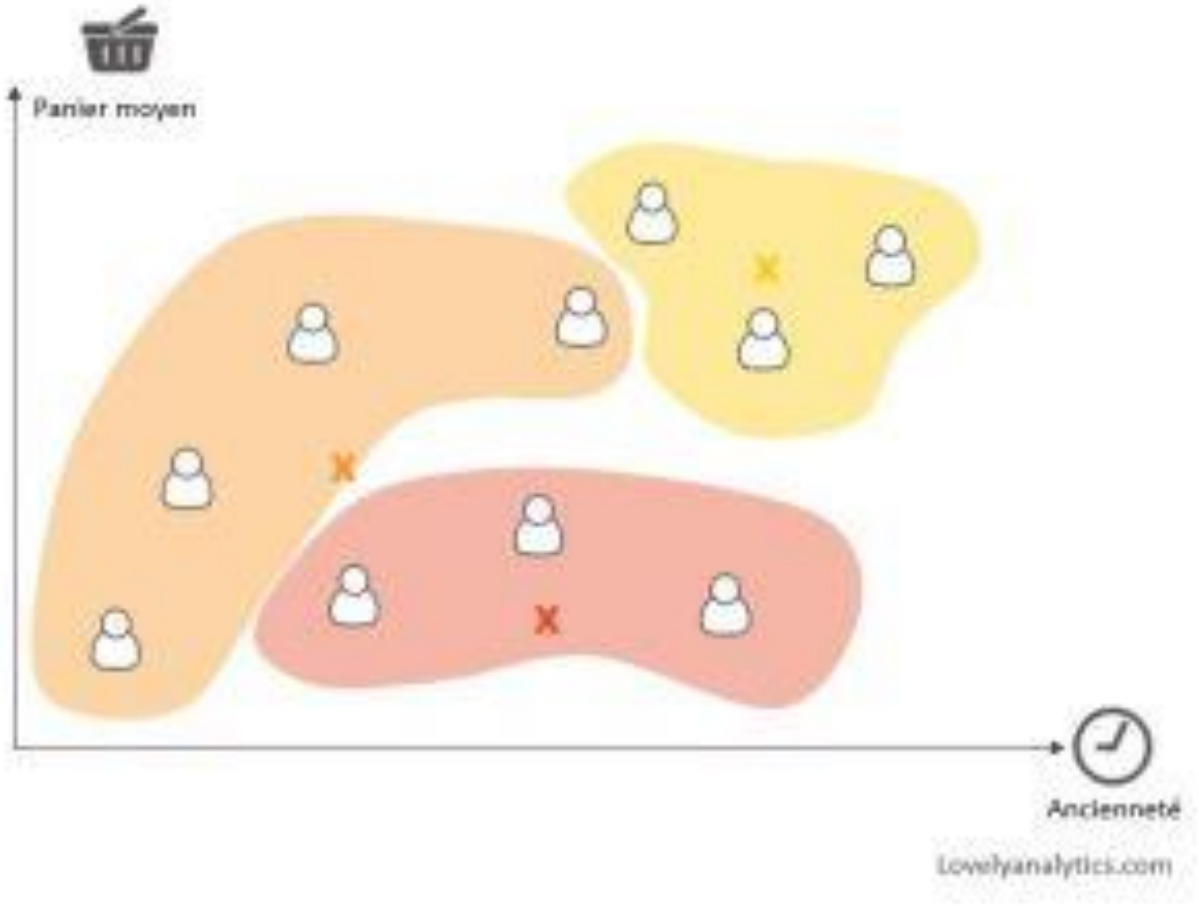
Plusieurs métriques existent pour définir la proximité entre 2 individus.
La méthode “classique” se base sur la distance euclidienne, vous pouvez aussi utiliser la distance Manhattan ou Minkowski.

Etape 2 :



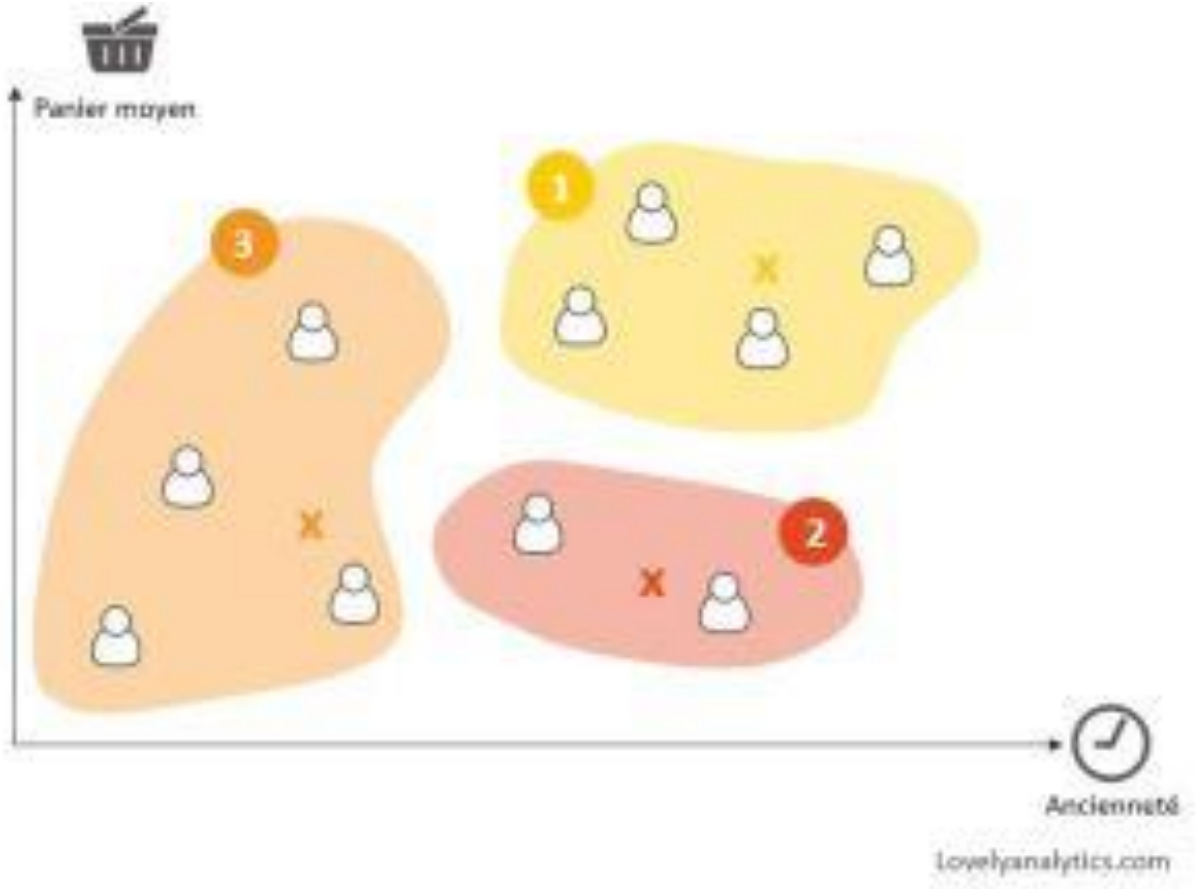
On affecte chaque individu au centre le plus proche.

Etape 3 :



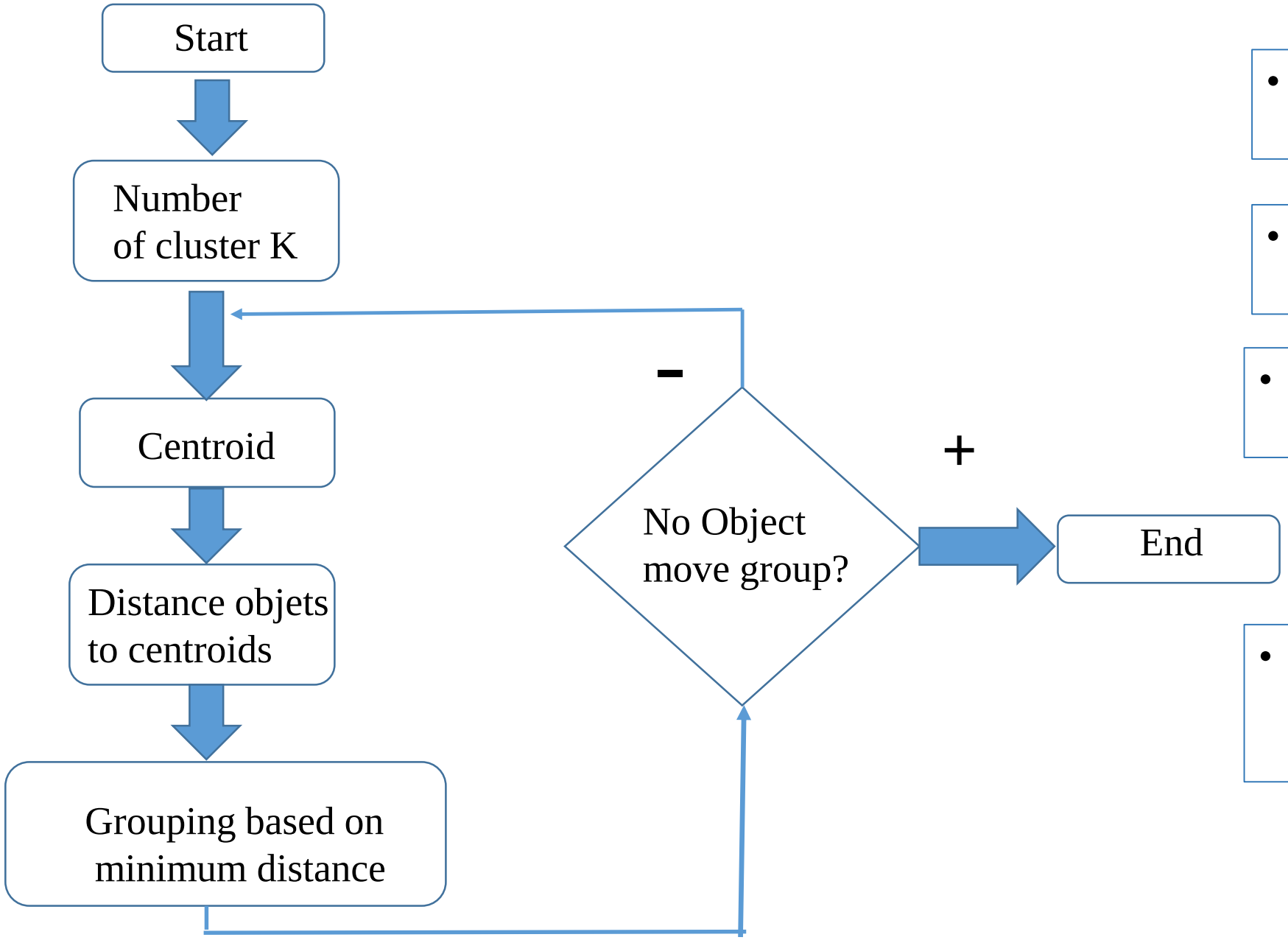
On calcule les centres de gravité des groupes qui deviennent les nouveaux centres

Boucle itérative :



On recommence les étapes 1, 2 et 3 tant que les individus sont réaffectés à de nouveaux groupes après une itération.

Algorithme



- Choisir K éléments initiaux « centres » des K groupes.
- Placer les objets dans le groupe de centre le plus proche.
- Recalculer le centre de gravité de chaque groupe
- Itérer l'algorithme jusqu'à ce que les objets ne changent plus de groupe

Entrée

Ensemble de N données, noté par X

Nombre de groupes souhaité, noté par K

Sortie

Une partition de K groupes $\{C_1, C_2, \dots, C_K\}$

Début

1. Initialisation aléatoire des centres C_K

Répéter

2. Affectation : générer une nouvelle partition en assignant chaque objet au groupe dont le centre est le plus proche :

$$\boldsymbol{x_i \in C_K \;\; si \; \forall \, j \; |x_i - \mu_k| = min|x_i - \mu_j|} \qquad \textbf{(1)}$$

avec μ_k le centre de la classe K.

3. Représentation : Calculer les centres associé à la nouvelle partition

$$\boldsymbol{\mu_k = \frac{1}{N} \sum_{x \in C_K} x_i} \qquad (2)$$

Jusqu'à convergence de l'algorithme vers une partition stable

Fin

Avantages et Inconvénients

- Très facile à comprendre et à mettre en œuvre
- Simple et rapide
- Applicable à des données de grandes tailles et aussi à tout type de données (même textuelles)

- Le nombre de classe doit être fixé au départ
- Le résultat dépend du tirage initiales des centres des classes
- Les cluster sont construits par rapports à des objets inexistants (les milieux)

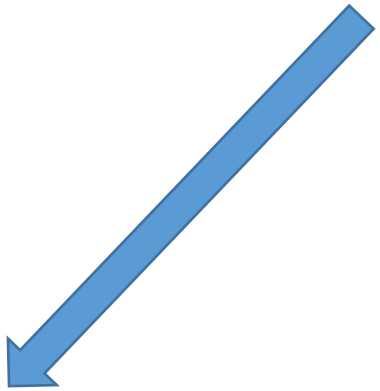
Algorithme d’apprentissage supervisé

Exemple : prédire le prix d’une maison en fonction de ses caractéristiques

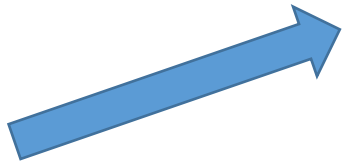
Régression

Permet de prédire des valeurs continues à partir des variables prédictives.

Régression linéaire

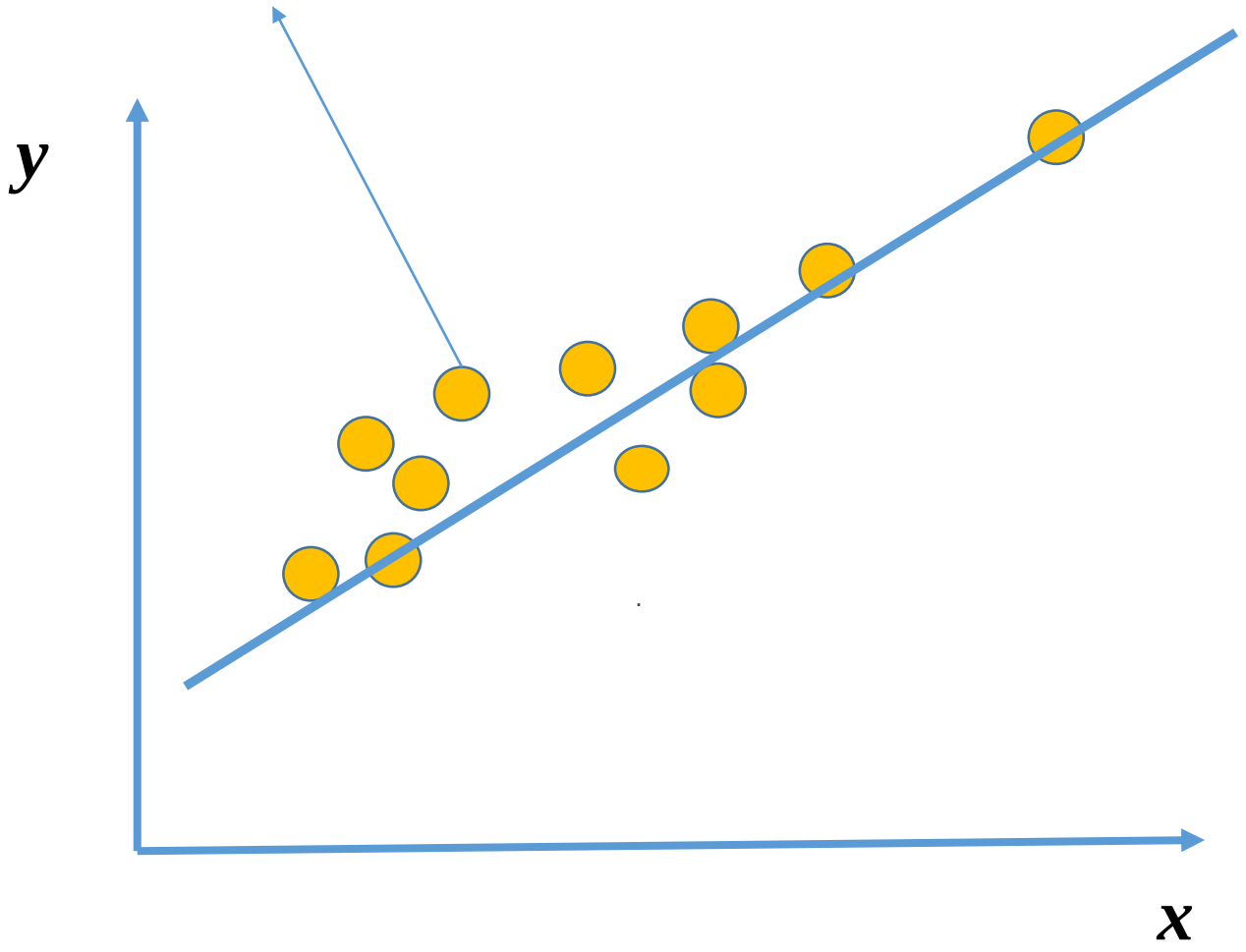


Un algorithme qui va trouver
une droite qui se rapproche
le plus possible d'un ensemble de points.



Les données d'entraînement (*Training Set*)

Les données d'entrée représentées par le couple (x_i, y_i)



x_i - variables prédictives

y_i - variables observées

But : trouver une droite

$F(x) = \alpha x + \beta$

tel que,

$\forall\, x_i,\ F(x_i) \approx y_i$

$$a = \frac{n\sum_{k=1}^{n} x_k y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} y_k}{n\sum_{k=1}^{n} x_k^2 - \left(\sum_{k=1}^{n} x_k\right)^2}; b = \frac{\sum_{k=1}^{n} x_k^2 \sum_{k=1}^{n} y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} x_k y_k}{n\sum_{k=1}^{n} x_k^2 - \left(\sum_{k=1}^{n} x_k\right)^2}$$

Naive Bayes Classifier

Algorithme du Supervised Learning utilisé pour la classification

Particulièrement utile pour les problématiques de **classification de texte**. Un exemple d'utilisation du Naive Bayes est celui du filtre anti-spam.

Le *naive Bayes classifier* se base sur le **théorème de Bayes**. Ce dernier est un classique de la théorie des probabilités. Ce théorème est fondé sur les **probabilités conditionnelles**.

Probabilités conditionnelles

Quelle est la probabilité qu'un événement se produise ***sachant qu'un autre événement s'est déjà produit.***

Exemple :

Supposons que l'on ait un groupe d'extra-terrestres. Soit A et B les deux événements suivants :

- l'événement A : l'extra-terrestre est un Vulcain.
- l'événement B : l'extra-terrestre officie sur l'Enterprise.

☺ « Longue vie et prospérité M. Spock » ☺

Quelle est la probabilité qu'on choisisse au hasard un Vulcain qui officie sur l'Enterprise ?

Le théorème de Bayes permet de calculer ce genre de probabilité.

Notons P la probabilité d'un événement.

$P(A \text{ ET } B) = \\ = P(A) * P(B|A)$

Le terme $P(A|B)$ se lit : **la probabilité que l'événement A se réalise sachant que l'événement B s'est déjà réalisé.**

terme A : l'évidence (faux ami avec le mot anglais **Evidence**)

terme B s'appelle **Outcome.**

Formule du théorème de Bayes

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

	l'extra-terrestre est un **Vulcain (A)**	**$(non\ A)$**	**Total**
l'extra-terrestre officie sur l'Enterprise (B)	10	7	17
(non B)	4	9	13
Totaux	14	16	30

Quelle est la probabilité qu'on choisisse au hasard un extraterrestre qui officie sur l'Enterprise sachant que c'est un Vulcain?

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A|B) * P(B)}{P(A)}$$

P(A) est la probabilité de prendre au hasard un extra-terrestre qui est un Vulcain.
On appelle P(A) **la probabilité antérieure (*prior probability)***.

$$P(A) = \frac{cardinal(A)}{cardinal(\Omega)} = \frac{14}{30} \approx 0,4666$$

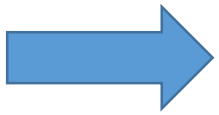
Le nombre d'éléments dans l'ensemble

$$P(B \cap A) = \frac{cardinal(B \cap A)}{cardinal(\Omega)} = \frac{10}{30} \approx 0,3333$$

$$P(B|A) = \frac{0,3333}{0,4666} \approx 0,7143$$

Dans notre exemple :
le *théorème de Bayes* a été appliqué avec une seule variable prédictive (*Evidence)*

Dans les vraies applications du Naive Bayes, on calcule le résultat (*Outcome)* en se basant sur **plusieurs variables**.



Calcul complexe!!!

Pour contourner cela, une approche consiste à **prendre en considération ces variables indépendamment les unes des autres**. Il s'agit d'une **hypothèse forte.** Généralement, les variables prédictives sont liées entre elles. Le terme "**naïve**" vient du fait qu'on suppose cette indépendance des variables.

Exemple d'application du Naive Bayes classification

Supposons qu'on ait un jeu de données sur 1000 fruits.
On dispose de trois types : Banane, Orange, et "autre".
Pour chaque fruit, on a 3 caractéristiques :

- Le fruit est **long** ou non
- Le fruit est **sucré** ou non
- La couleur du fruit est **jaune** ou non

Type	Long	Pas long	Sucré	Non sucré	Jeune	Non jeune	Total
Banane	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Autre fruit	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

L’idée du jeu est de prédire le type d’un fruit (orange, banane ou autre) qu’on n’a pas encore vu en se basant sur ses caractéristiques.
Supposons que quelqu’un nous demande de lui donner le type d’un fruit qu’il a. Ses caractéristiques sont les suivantes :

- Il est jaune
- Il est long
- Il est sucré

Pour savoir s’il s’agit d’une banane, ou d’une orange ou d’un autre fruit, il faut qu’on calcule les trois probabilités suivantes :

- probabilité qu’il s’agisse d’une banane sachant que le fruit est long, jaune et sucré : P(Banane | Long, jeune, sucré)
- probabilité qu’il s’agisse d’une orange sachant que le fruit est long, jaune et sucré : P(Orange | Long, jeune, sucré)
- probabilité qu’il s’agisse d’un autre fruit sachant que le fruit est long, jaune et sucré : P(Autre fruit | Long, jeune, sucré)

Le fruit « inconnu » qu'on cherche à classifier sera celui où on a la plus grande probabilité.

En appliquant la formule de Bayes

$$P(Banane|long, jaune, sucré) = \frac{P(long|\ Banane) * P(sucré|\ Banane) * P(jaune|\ Banane) * P(Banane)}{P(long) * P(sucré) * P(jaune)}$$

$$P(banane) = \frac{cardinal(banane)}{cardinal\ (tout\ les\ fruits)} = \frac{50}{100} = 0,5$$

De la même façon, on a :

$P(orange)$ =0,3
P(Autre fruit)=0,2
P(long)=0,5
P(sucré)= 0,65
P(jaune)=0,8

$$P(long\ |banane) = \frac{cardinal(Banane\ \cap long)}{cardinal\ (banane)} = \frac{400}{500} = 0,8$$

P(sucré | banane)=0,7
P(jaune | banane)=0,9

$$P(\boldsymbol{Banane}|\boldsymbol{long},\boldsymbol{jaune},\boldsymbol{sucr}\text{é}) = \frac{\mathbf{0,8} * \mathbf{0,7} * \mathbf{0,9} * \mathbf{0,5}}{\mathbf{0,5} * \mathbf{0,65} * \mathbf{0,8}} = \mathbf{0,969}$$

$$P(\boldsymbol{Orange}|\boldsymbol{long},\boldsymbol{jaune},\boldsymbol{sucr}\text{é}) = \mathbf{0}$$

$$P(\boldsymbol{Autre\ fruit}|\boldsymbol{long},\boldsymbol{jaune},\boldsymbol{sucr}\text{é}) = \mathbf{0,072}$$

Donc, notre gagnant est bien la banane ☺

Avantages du Naive Bayes Classifier

- *Naive Bayes Classifier* **est très rapide pour la classification** : en effet les calculs de probabilités ne sont pas très coûteux.
- La classification est possible même avec **un petit jeu de données**

Inconvénients du Naive Bayes Classifier

L'algorithme *Naive Bayes Classifier* suppose **l'indépendance des variables** :
C'est une **hypothèse forte** et qui est violée dans la majorité des cas réels.

- https://www.google.fr/search?biw=1244&bih=510&tbm=isch&sa=1&ei=sg8LW_DjG4XyUP3LkYgP&q=animaux+&oq=animaux+&gs_l=img.3..35i39k1l2j0i67k1l4j0l4.70687.73500.0.73872.27.14.0.0.0.0.143.1215.12j2.14.0....0...1c.1.64.img..19.7.634....0.i8n0NWGlZOY#imgrc=3nhGOisrEcmjoM:
- http://www.ieee.ma
- http://www-lisic.univ-littoral.fr/~verel/TEACHING/15-16/data-science/cours01-introduction.pdf
- https://www.antoine-brisset.com/blog/clustering-ruby-k-means/
- http://cedric.cnam.fr/vertigo/Cours/RCP216/tpClassificationAutomatique.html

Le métier de Data scientist

- Data scientist Apparu en 2008, DJ. Patil et Jeff Hammerbacher de Facebook et LinkedIn, ce sont appelés «data scientist »

- Généralisé à partir de 2012 : « Data scientist : The sexiest Job of the 21th Century », T.H. Davenport, DJ. Patil, Harvard Buissiness Review, oct. 2012.

- Rôle du data scientist gagne en importance dans les entreprises : Augmentation (explosion !) du volume des données non structurées (big data)

- Dans les 10 prochaines années, profil data scientist sera très recherché