

The background of the slide is a collage. On the left, there's a yellow highlighter and a blue pen on a white notepad with horizontal lines. On the right, there's a cluster of pink and orange balloons against a light blue sky with white clouds. Some of the balloons have simple smiley faces drawn on them. In the center, a blue-outlined rectangle contains a yellow sticky note with the main title. A small yellow sticky note with two diagonal lines is also attached to the top of the blue rectangle.

Segmentez des clients d'un site e-commerce

Soutenance Projet 5

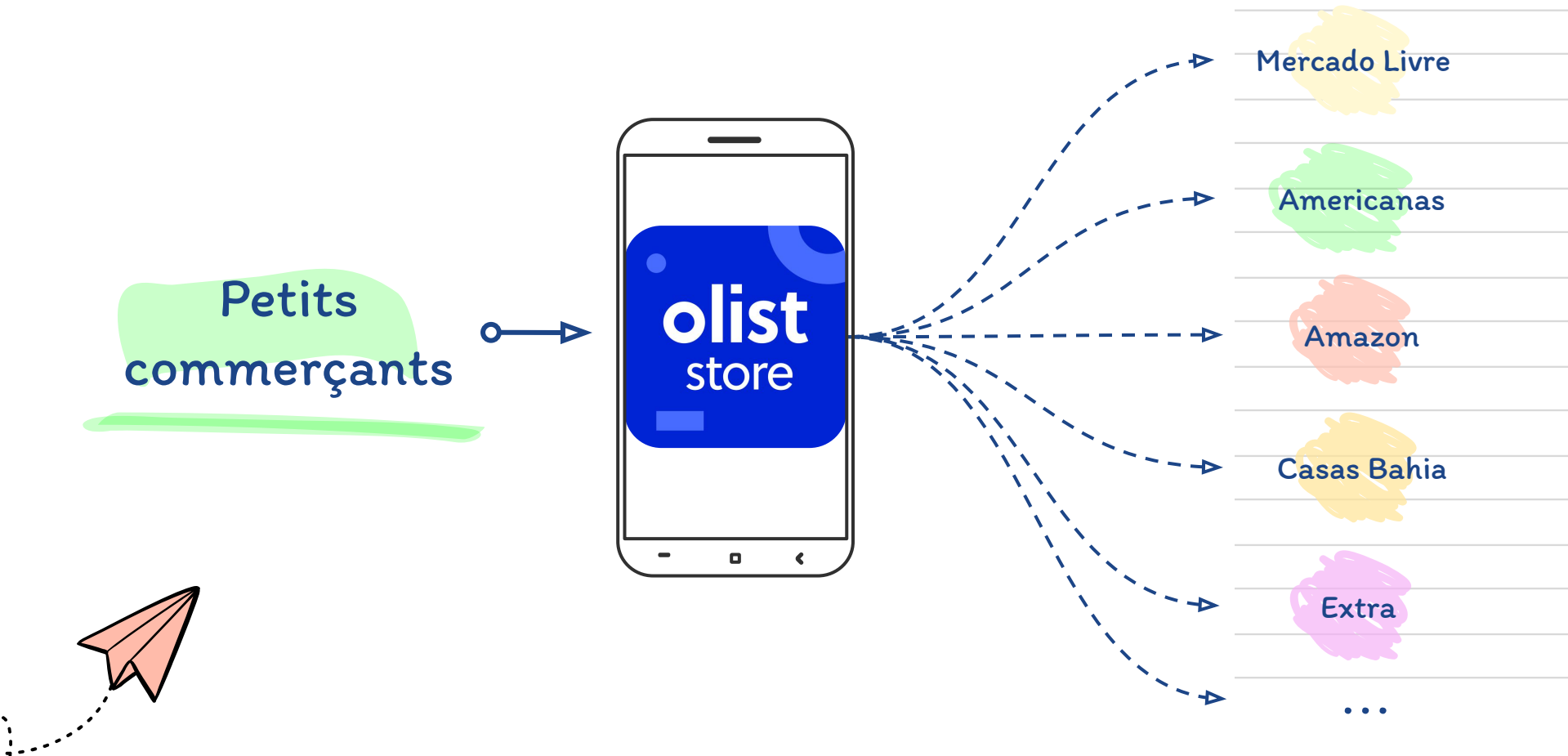
OpenClassrooms

Formation Data Scientist

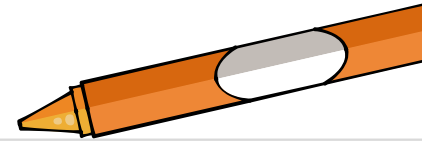
15 Mars 2023

Photo de Madison Oren sur Unsplash

Marché e-commerce brésilien fractionné



Missions :



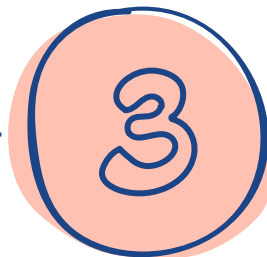
Plan soutenance :



Préparation données



Modélisation



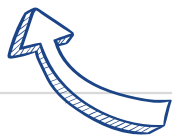
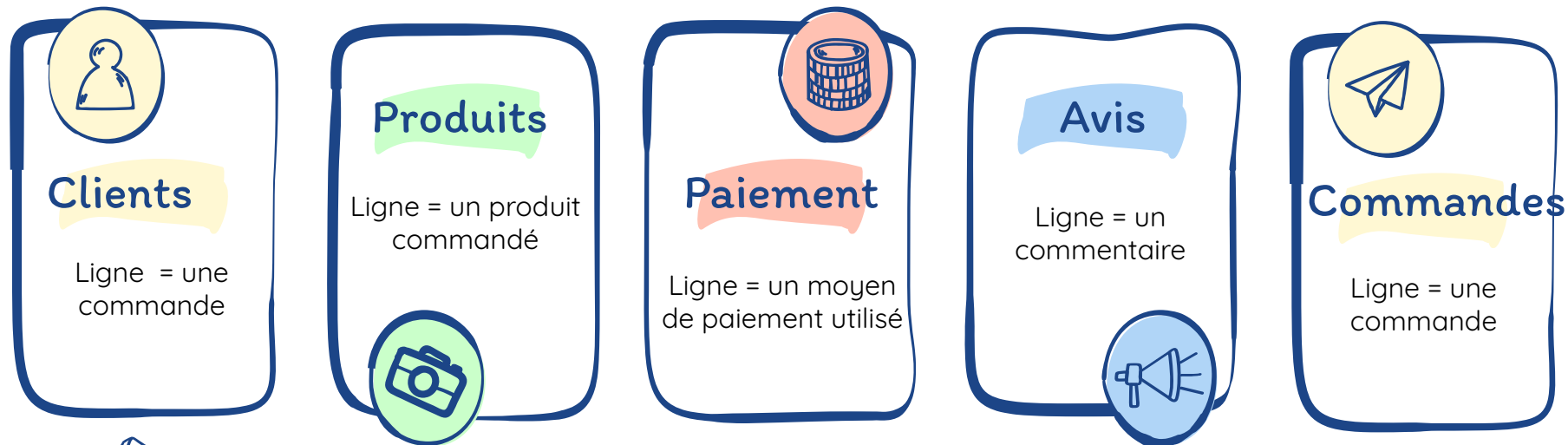
Maintenance





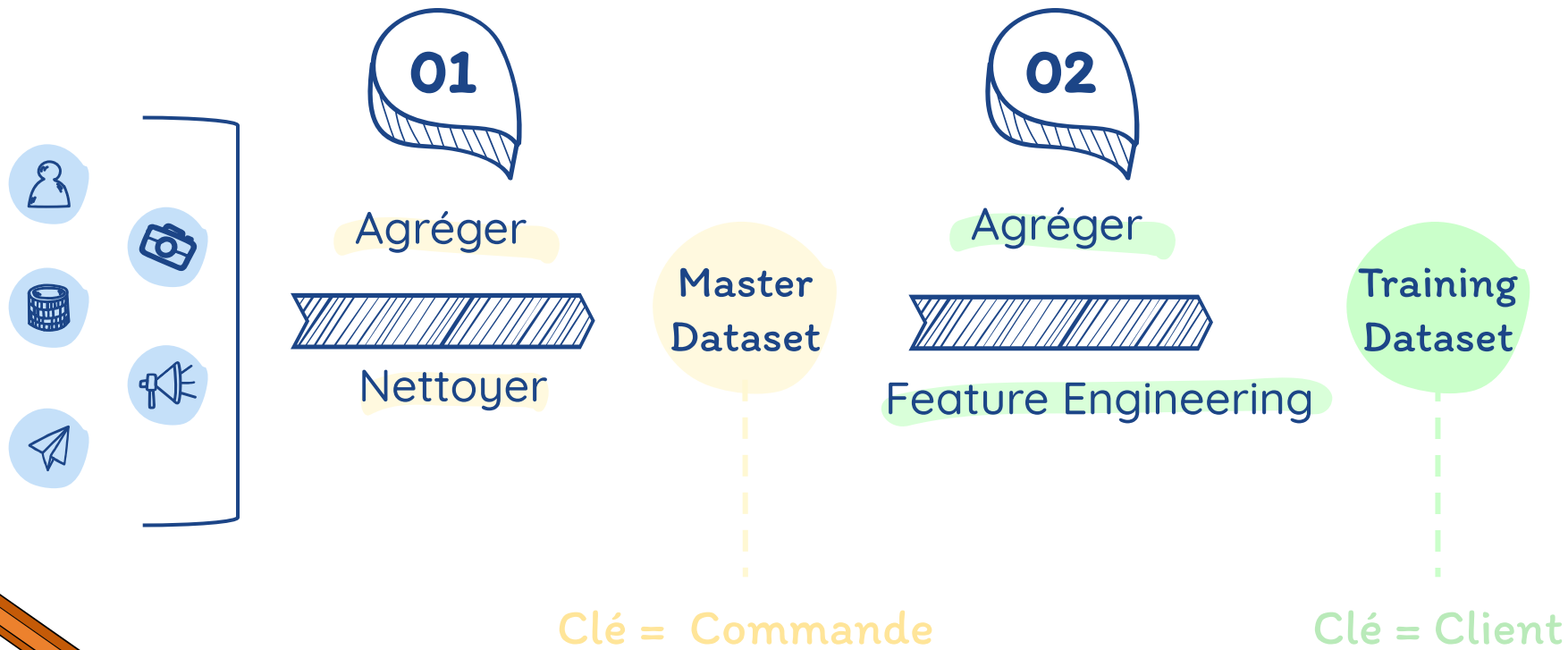
Préparation des données

Extraits de 9 tables SQL



96_096 clients

Construction jeu de données





Traitement des valeurs manquantes

○ Note satisfaction client



Moyenne

○ Montant commande



Prix + Frais de port

○ Nombre de produits



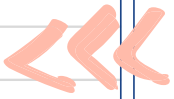
Montant commande

Coût moyen d'un produit

****Conserve les outliers**



5 Features



Récence

Nombre de jours depuis la dernière commande

Fréquence

Nombre de commandes

Montant

Montant total des achats

Note de satisfaction moyenne

Nombre moyen de produits par commande



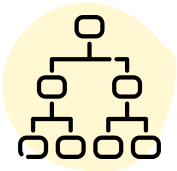
****StandardScaler**



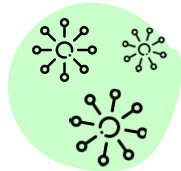
Modélisation

Classification non supervisée = Clustering

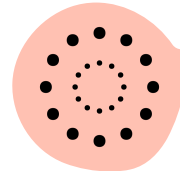
3 algorithmes d'apprentissage :



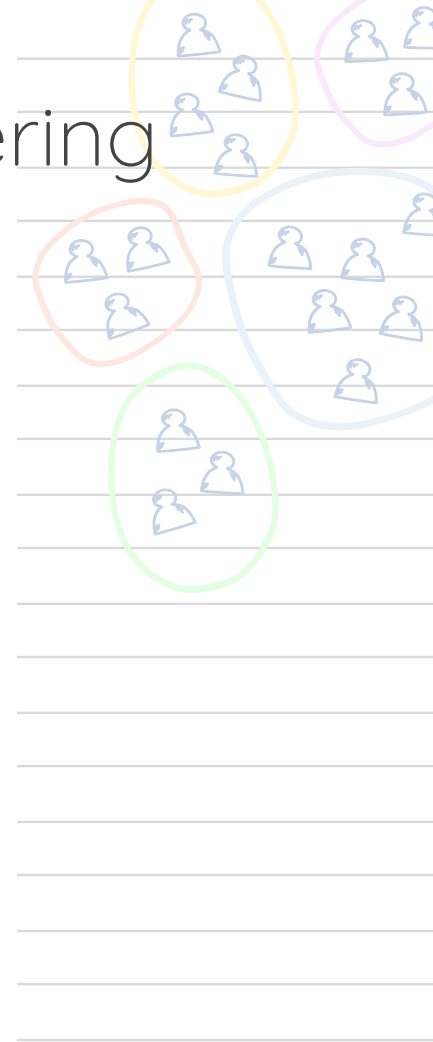
Classification ascendante
hiérarchique

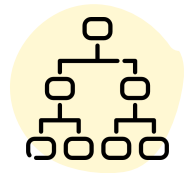


KMeans

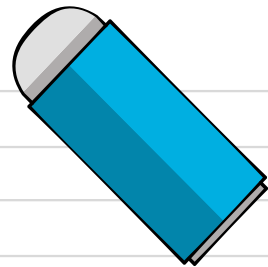
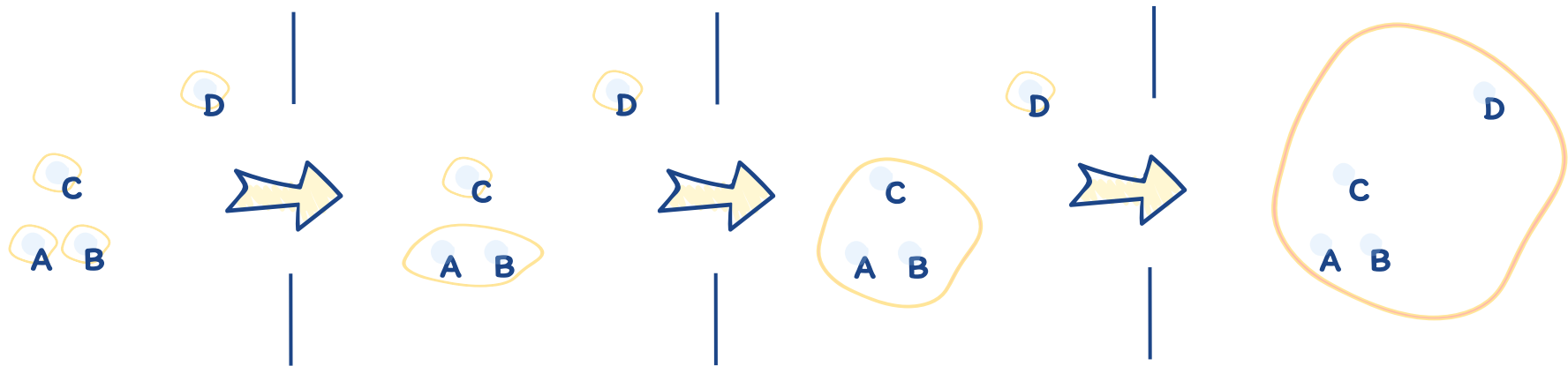


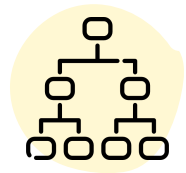
DBSCAN





Clustering Agglomératif

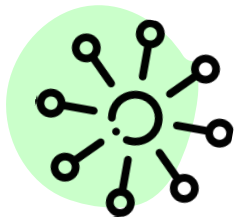




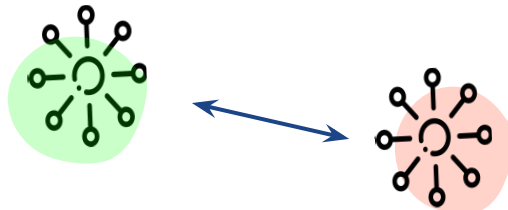
Méthode de Ward :

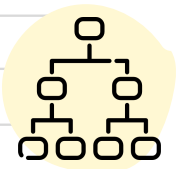
Agrège les clusters dont l'agrégation minimise l'augmentation de l'inertie intraclasse.

Inertie intra-classe

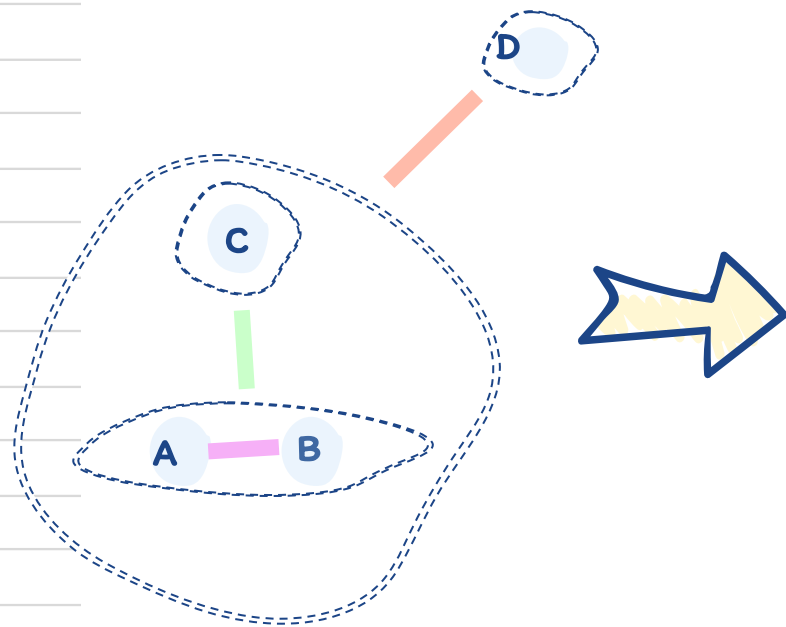


Inertie inter-classe

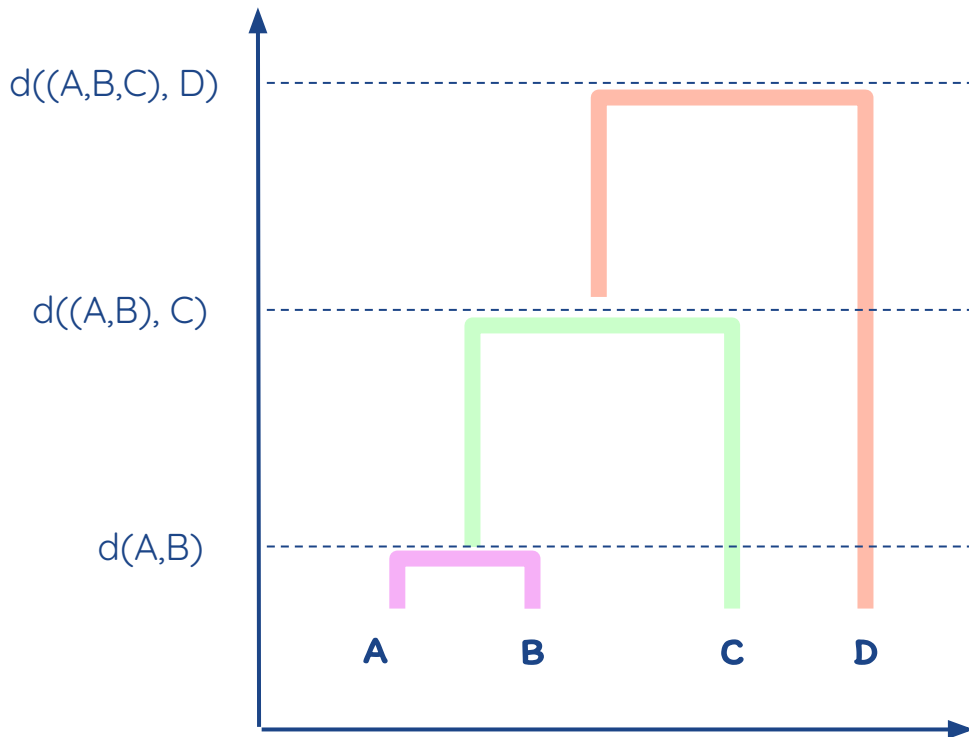




Choix du nombre de clusters

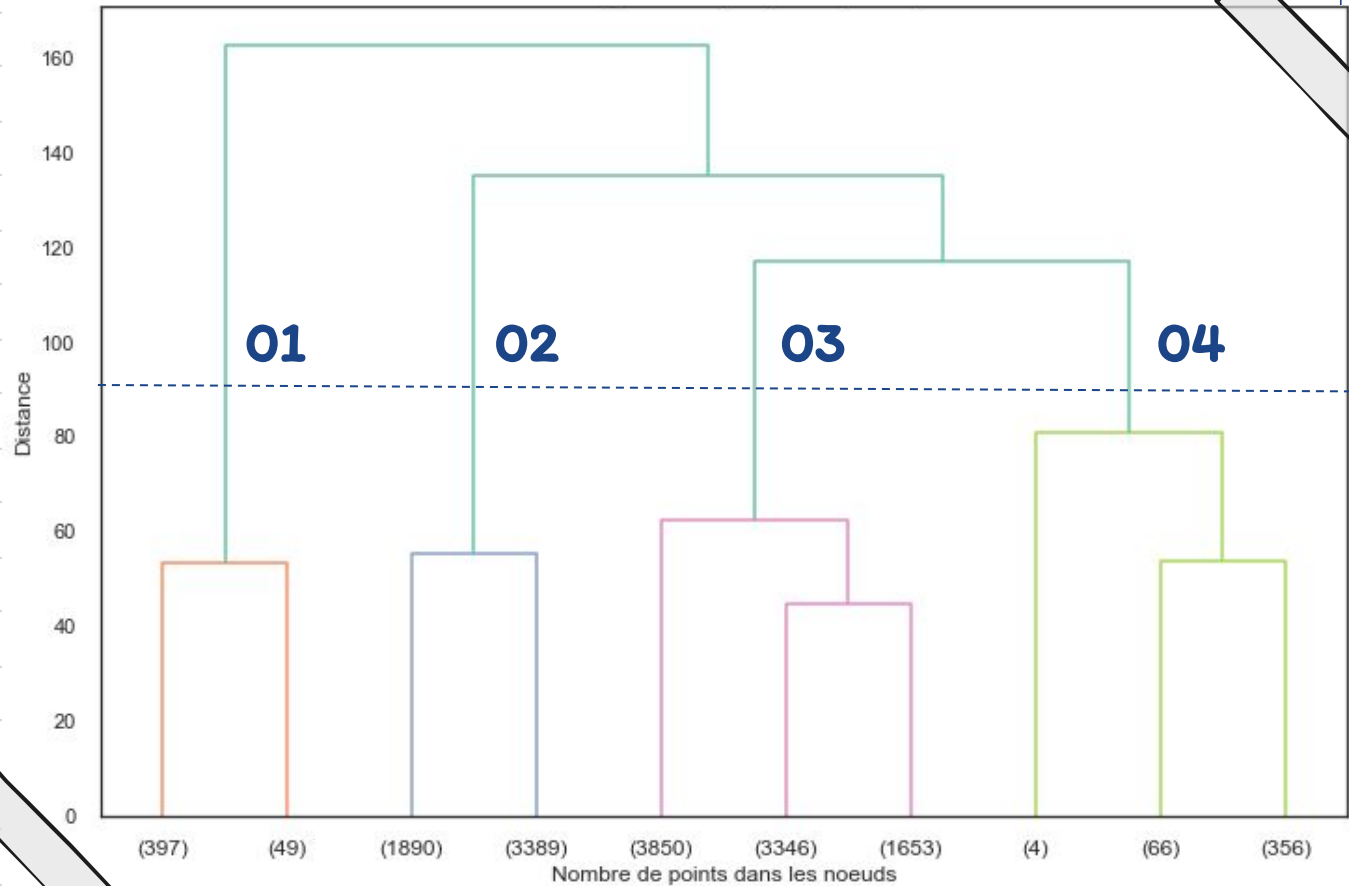


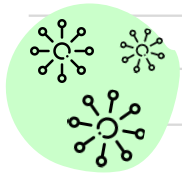
Dendrogramme



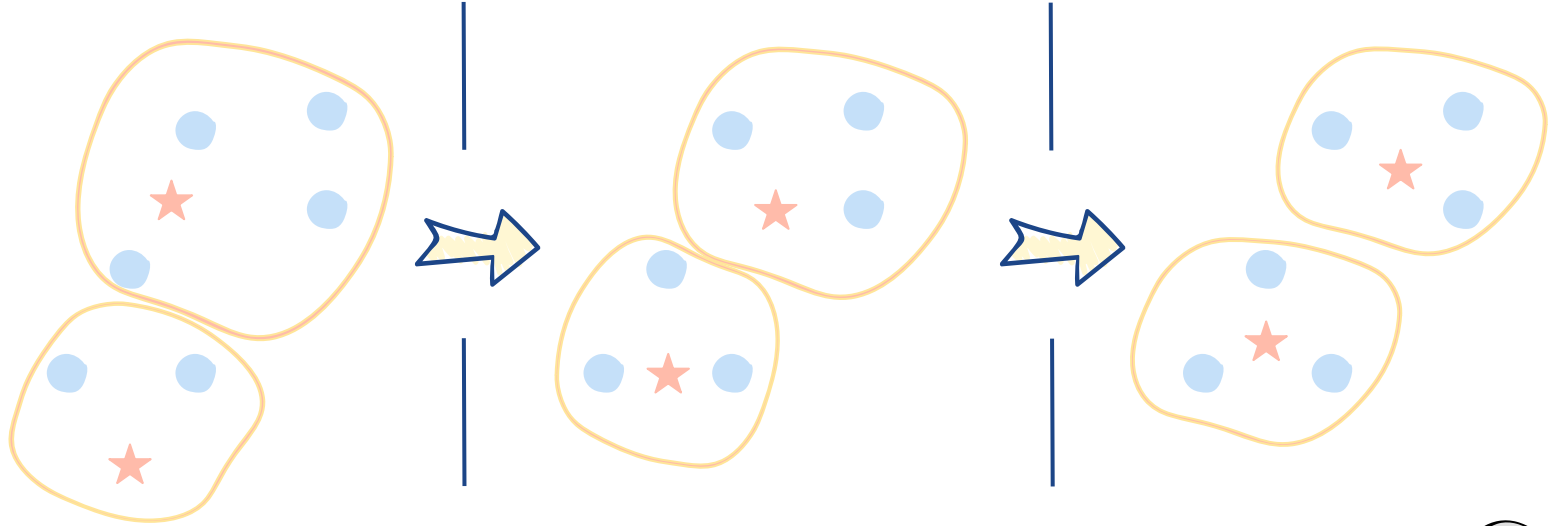
Dendrogramme

4
Clusters

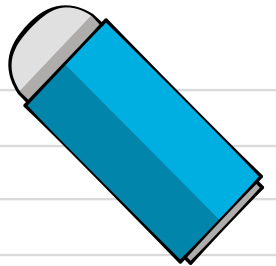


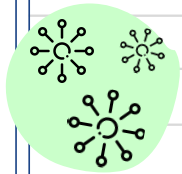


K-Means



Centre de gravité du cluster





K-Means

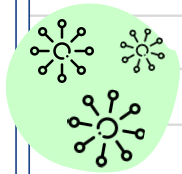
Limites

- Algorithme non-déterministe
- Minimum local

Solutions

- Relancer + choisir le résultat qui minimise l'inertie intra-classe
- K-Means ++

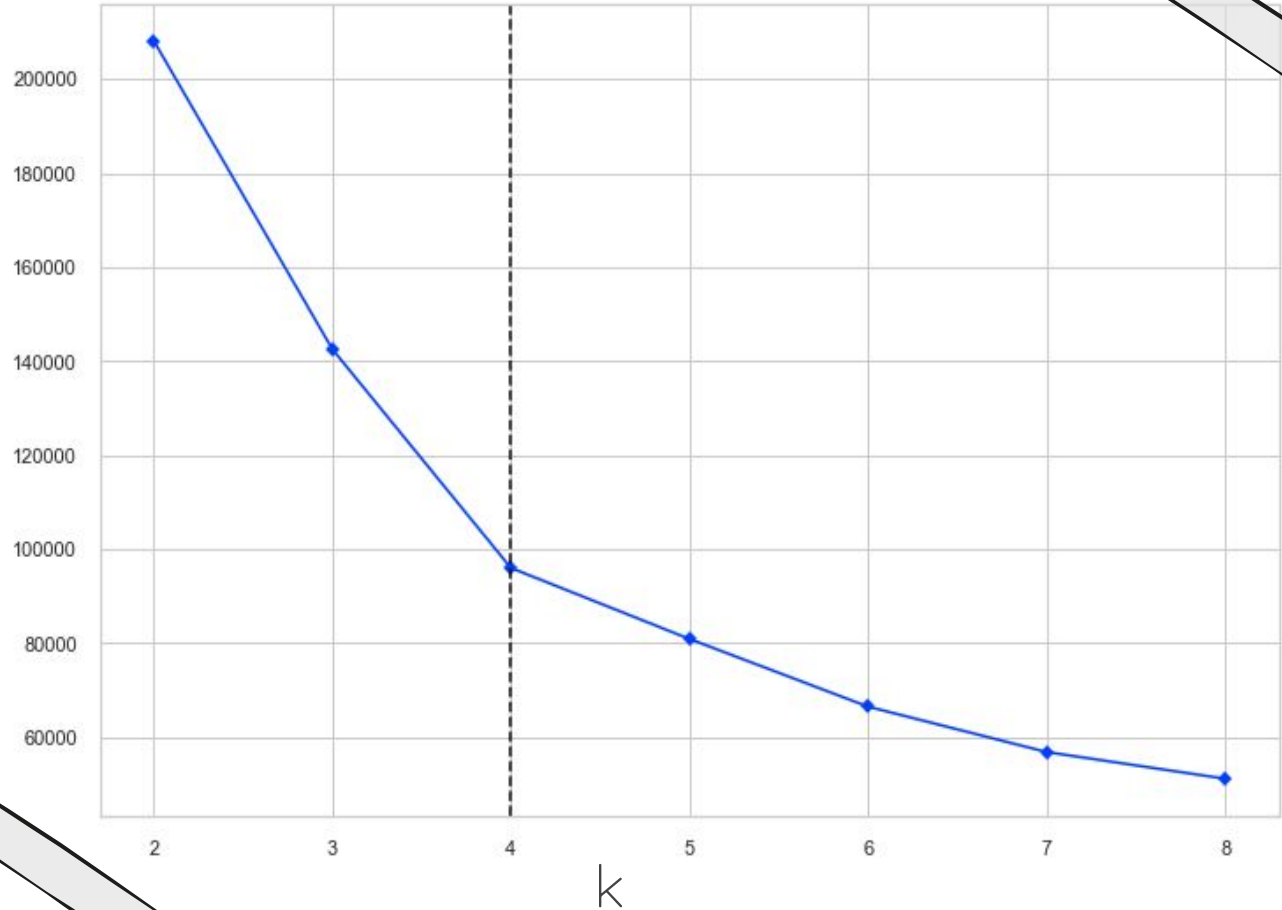
****par défaut dans sklearn**

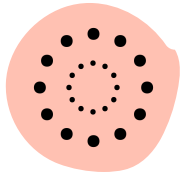


Choix du nombre de clusters

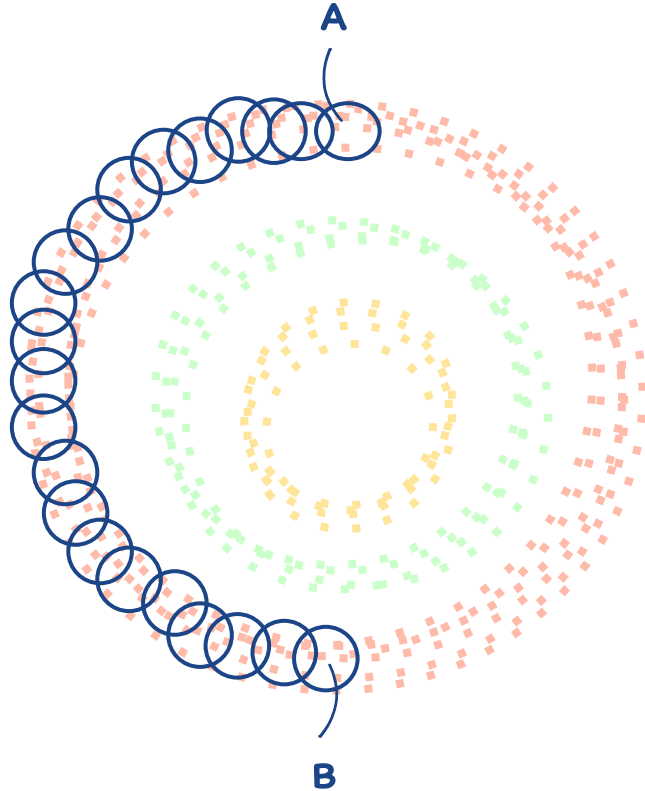
Méthode du “coude”

Distortion score





DBSCAN

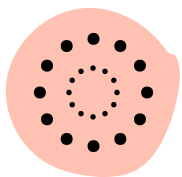


Clusters = points atteignables
par densité les uns depuis les
autres



Epsilon voisinage



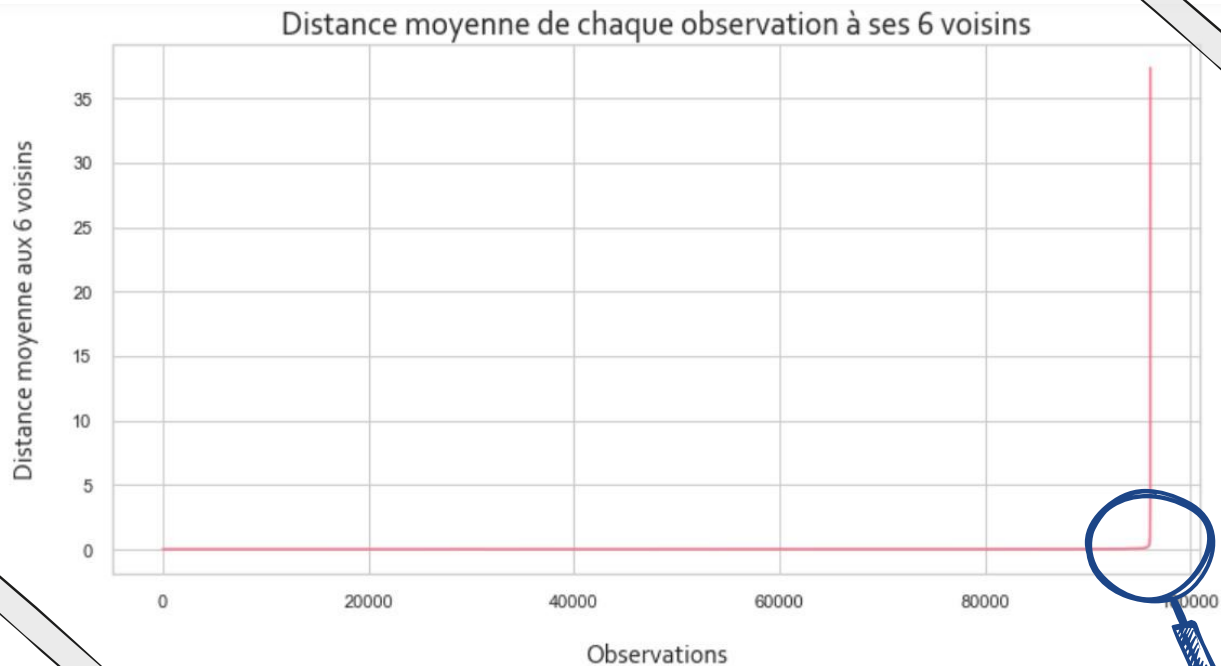


Choix des hyperparamètres

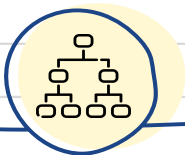
- Nombre de voisins minimum :

$2 * \text{nombre de dimensions}$

- Epsilon voisinage :
méthode du coude

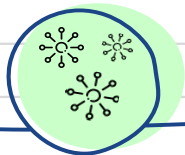


Algorithme le plus adapté



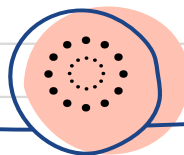
CAH

- Clusters interprétables
- Complexité algorithmique



K-Means

- Clusters interprétables
- Plus léger
- Facile à paramétrer



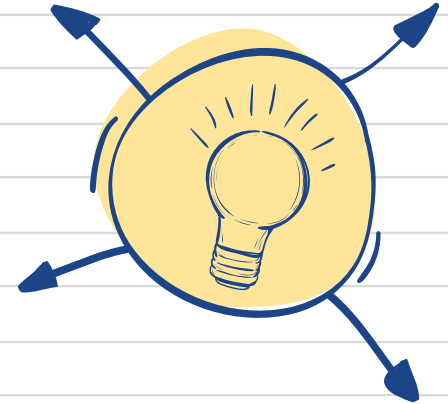
DBSCAN

- Clusters difficiles à interpréter
- Clusters déséquilibrés



Evaluer la performance des modèles

- Clusters vs connaissances du domaine : **pairplot, boxplot, radarplot**
- Clusters équilibrés : **diagramme circulaire**
- Forme des clusters : **coefficient de silhouette**



Coefficient de silhouette

Évalue à quel point l'individu x appartient au “bon” cluster

Homogénéité : “ x ” proche des points du cluster auquel il appartient ?

01

$$a(x) = \frac{1}{|C_k| - 1} \sum_{u \in C_k, u \neq x} d(u, x)$$

Séparation : “ x ” loin des points des autres clusters ?

02

$$b(x) = \min_{l \neq k} \frac{1}{|C_l|} \sum_{u \in C_l} d(u, x)$$

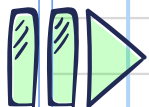
Coefficient de silhouette

Évalue à quel point l'individu x appartient au “bon” cluster

$$-1 \sim s(x) = \frac{b(x) - a(x)}{\max(a(x), b(x))} \sim 1$$



Comparaison modèles



	Nombre de Clusters	Coefficient de Silhouette	Répartition
RFM	4	0.489	40% <u>54%</u> 3% 3%
RFM + Produits	5	0.442	39% <u>53%</u> 3% 2% 2%
RFM + Note	5	0.418	33% 44% 17% 3% 2%
RFM + Note + Produits	5	0.368	33% 44% 17% 3% 3%

5 profiles clients :

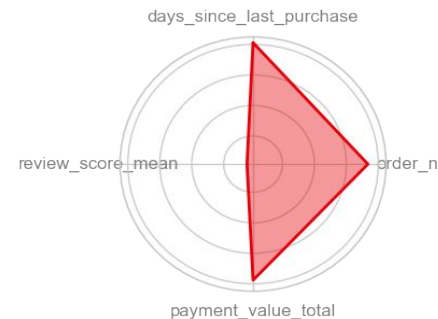
one time shoppers



new customers



dissatisfied customers



loyal customers

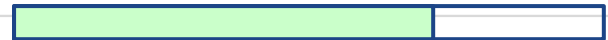
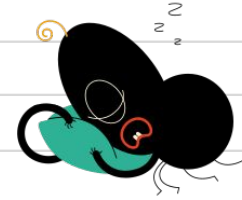


spendthrifts



Profils clients :

one time shoppers



Profils clients :

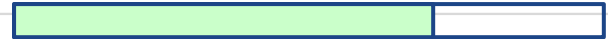
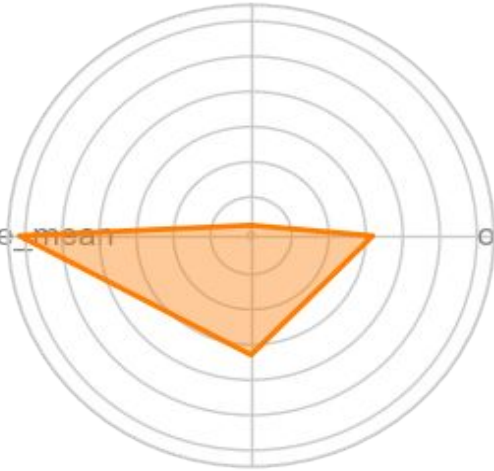
new customers

days_since_last_purchase

review_score_mean

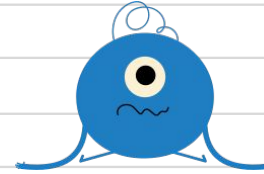
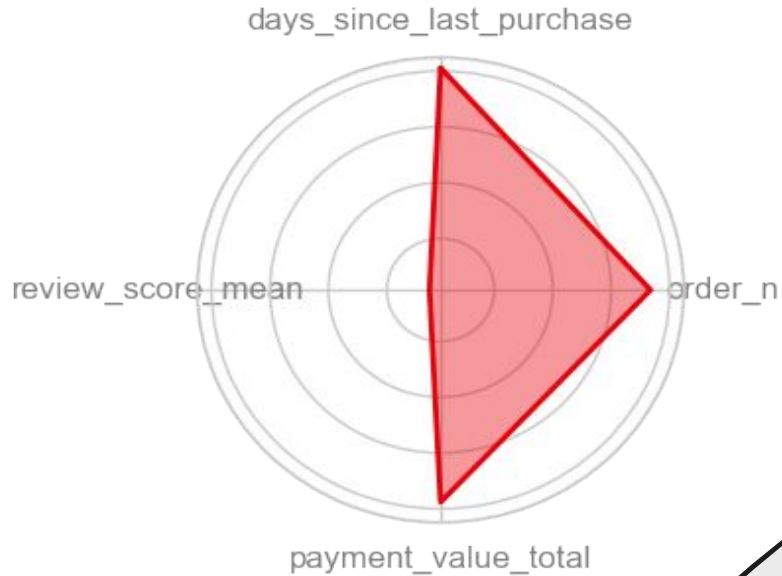
order_n

payment_value_total



Profils clients :

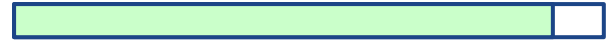
dissatisfied customers



Profils clients :

loyal customers

days_since_last_purchase



Profiles clients :



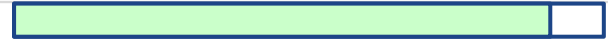
spendthrifts

days_since_last_purchase

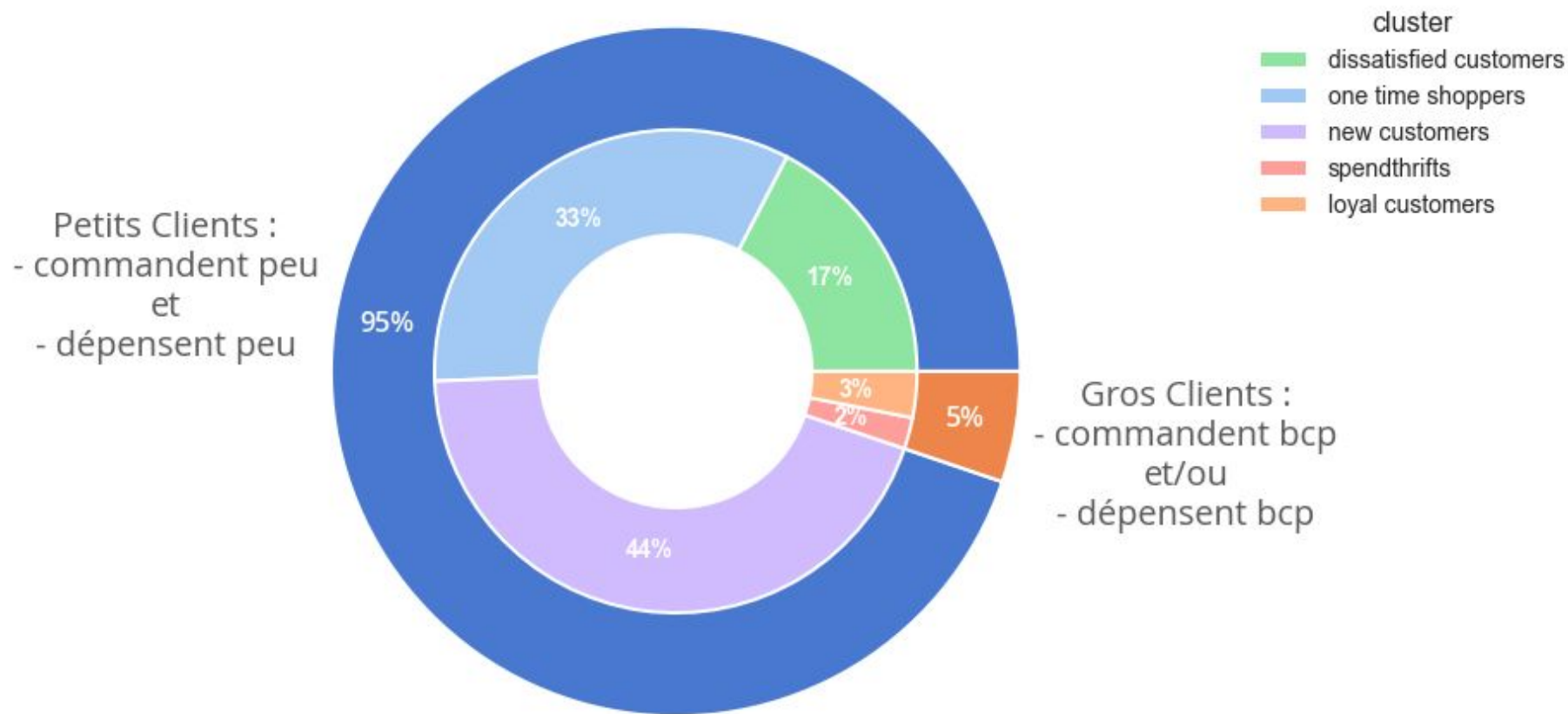
review_score_mean

order_n

payment_value_total

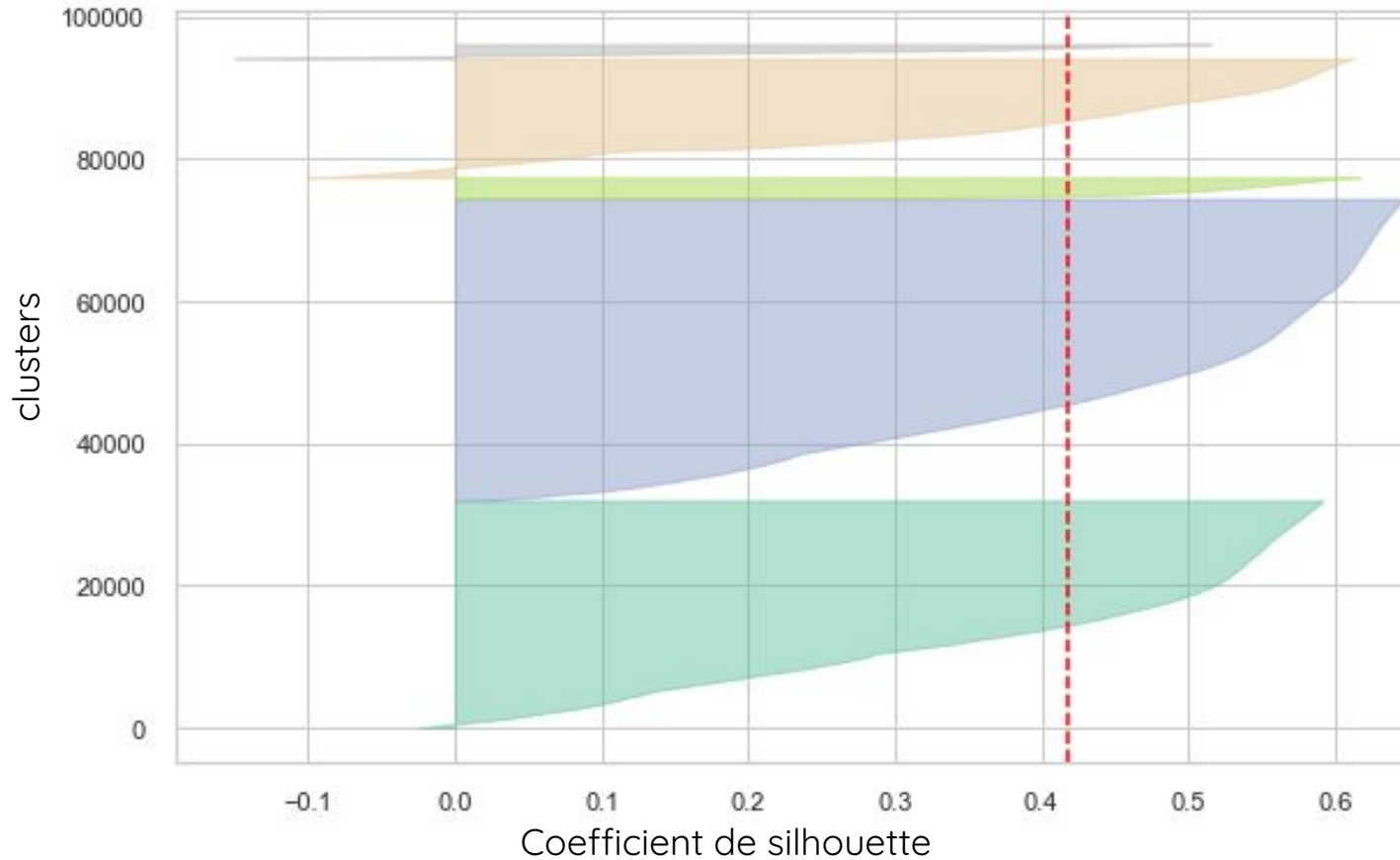


Nombre d'individus par cluster



Coefficients de silhouette

--- Coefficient de silhouette moyen





Maintenance

Indice de Rand ajusté

Évalue la concordance de deux partitions du jeu de données

0



$$ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)}$$



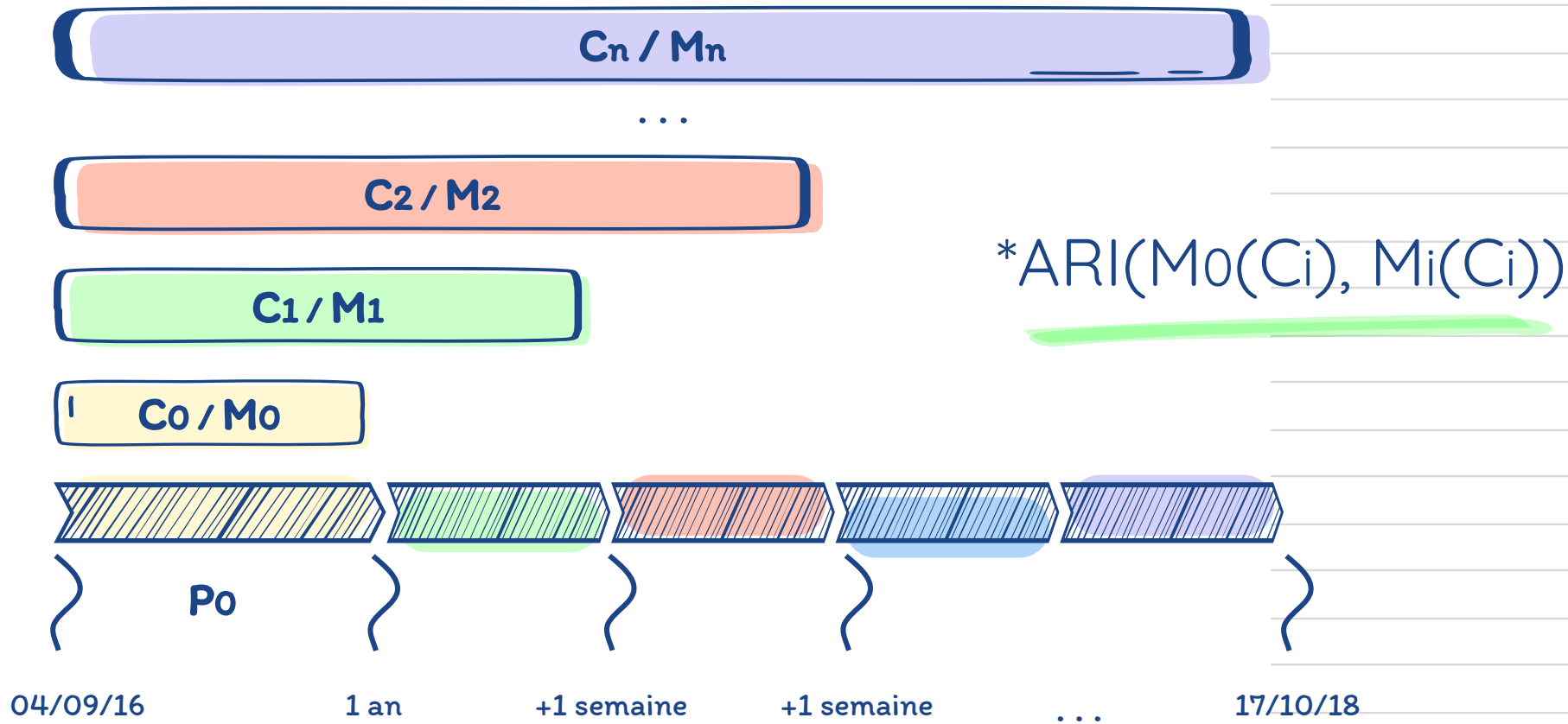
1

Clustering aléatoire

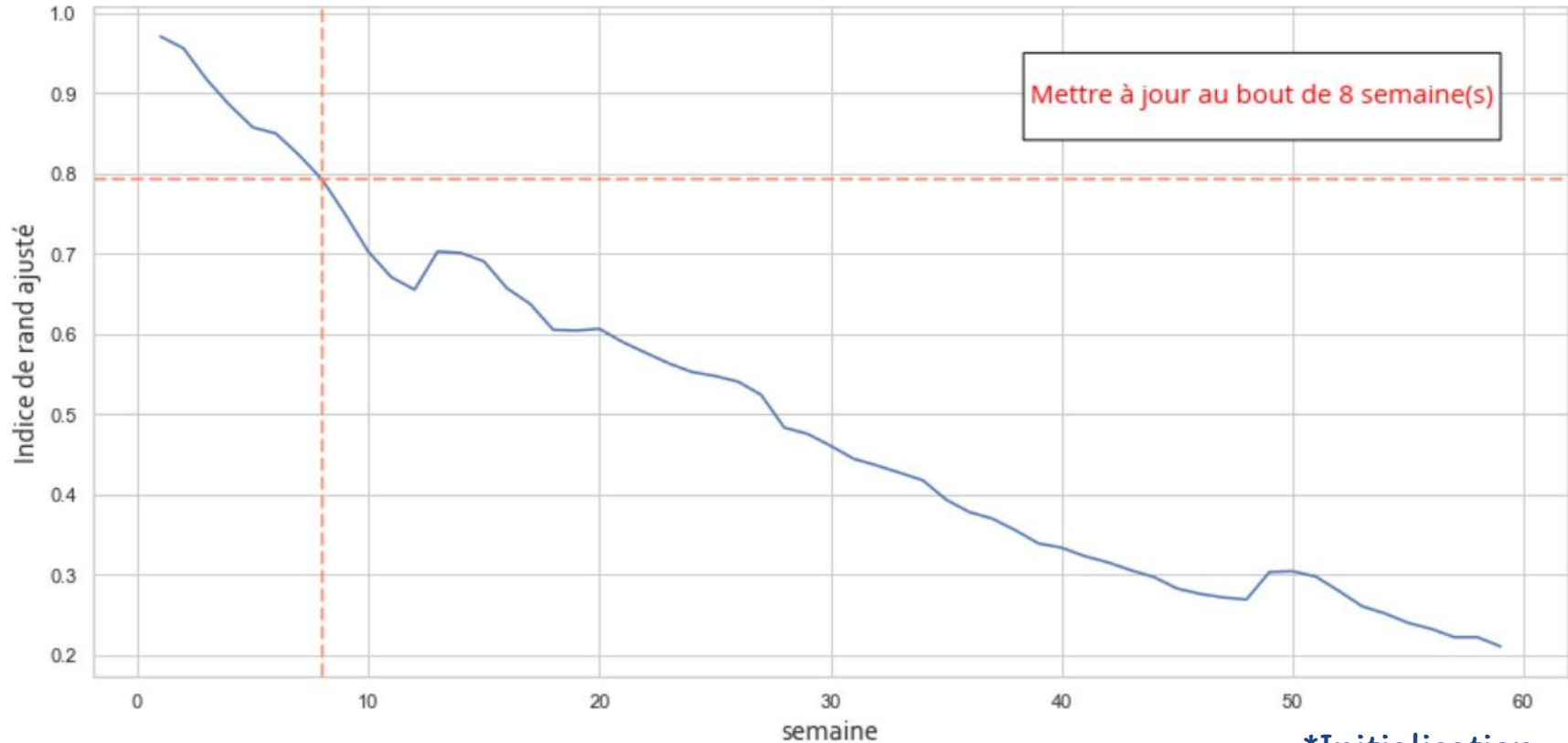
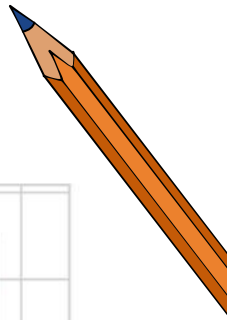
Clustering correspond
exactement à la partition
initiale

*RI : indice de rand

Fréquence de mise à jour du modèle :

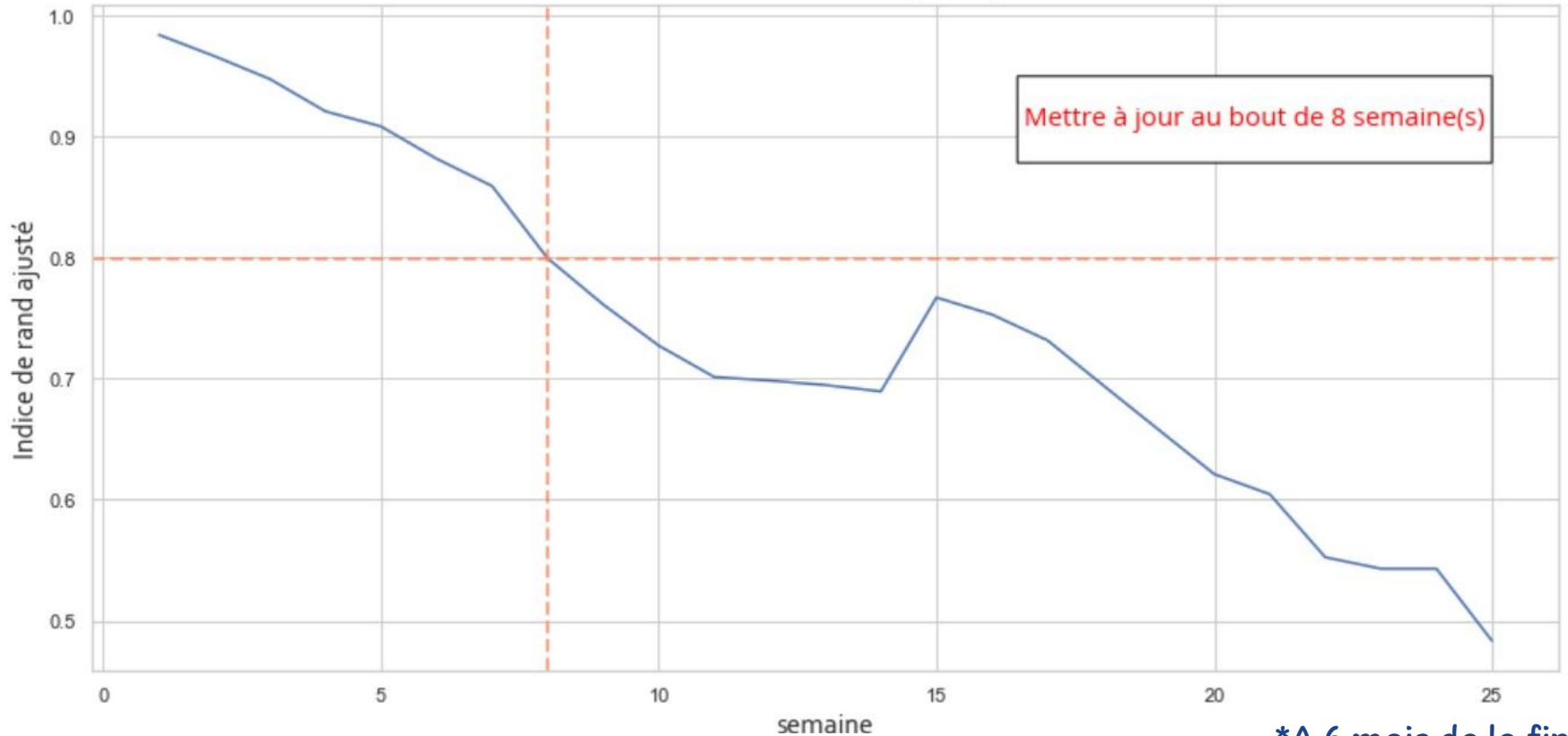
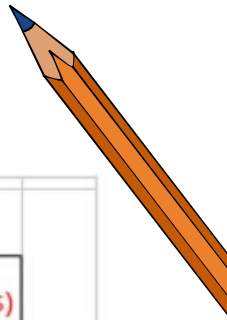


Evolution de l'indice de rand ajusté :



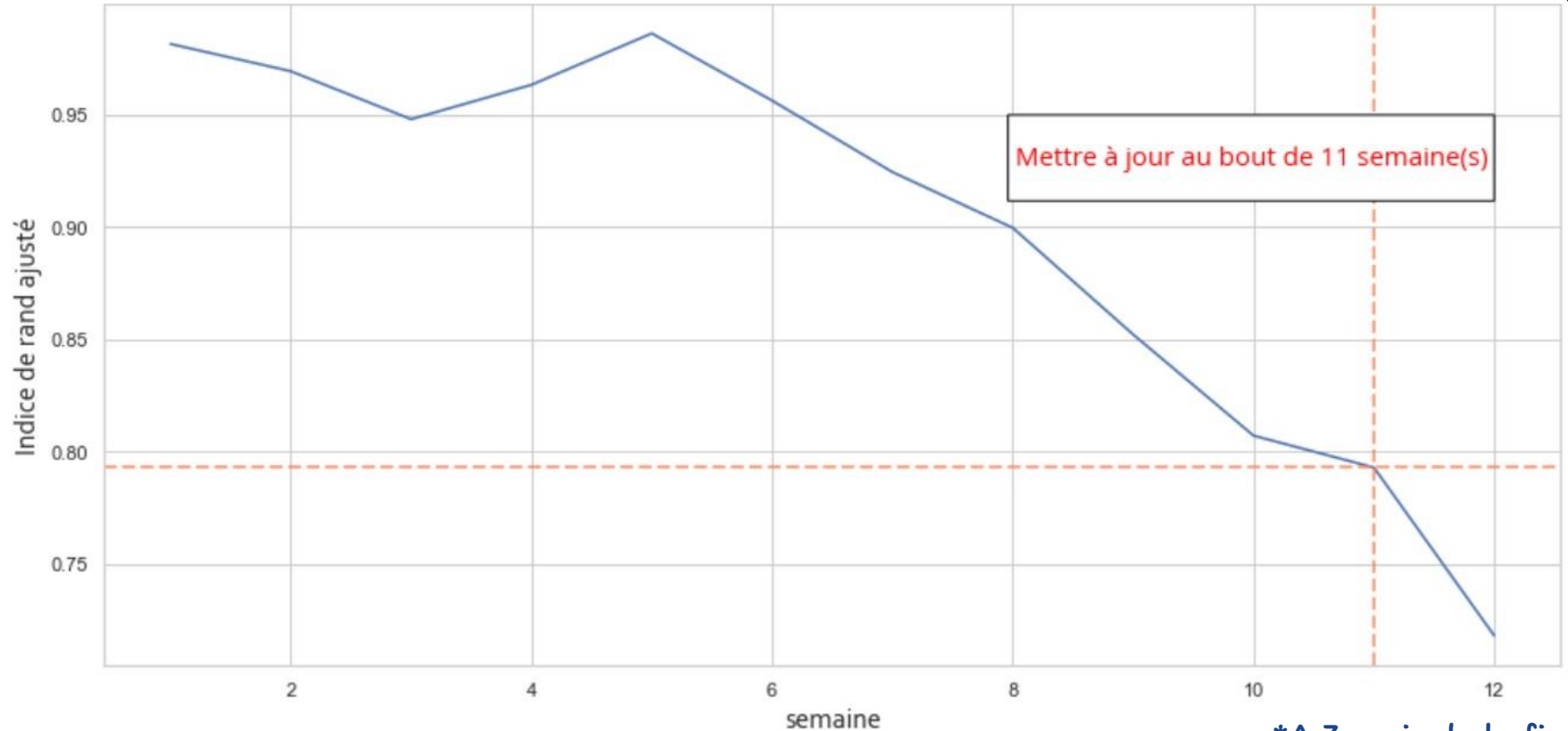
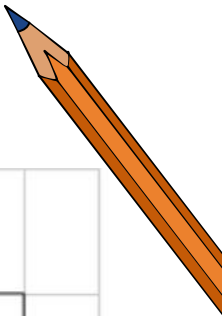
*Initialisation = 1 an

Evolution de l'indice de rand ajusté :



*A 6 mois de la fin

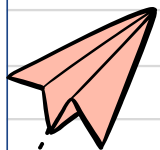
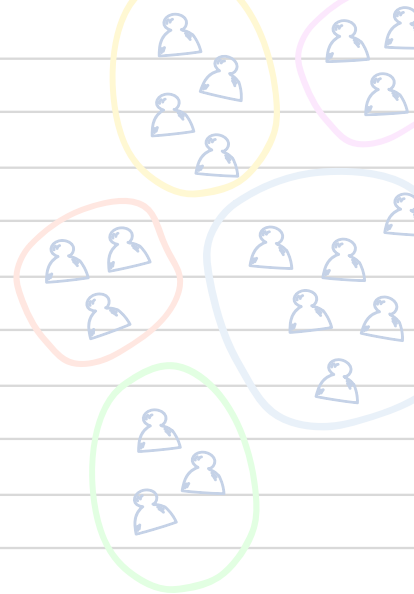
Evolution de l'indice de rand ajusté :



*A 3 mois de la fin

Conclusion :

- Algo d'apprentissage : KMeans
- Features : RFM + note moyenne
- Profils Clients : 5



Mise à jour : tous les 2 mois ou automatiser



Annexes

Inertie :

$$\text{Inertie interclasse} = 1/n * \sum n_c d(G_c, G)^2$$

$$\text{Inertie intra-classe} = 1/n * \sum \sum d(M_i, G_c)^2$$

d : distance euclidienne

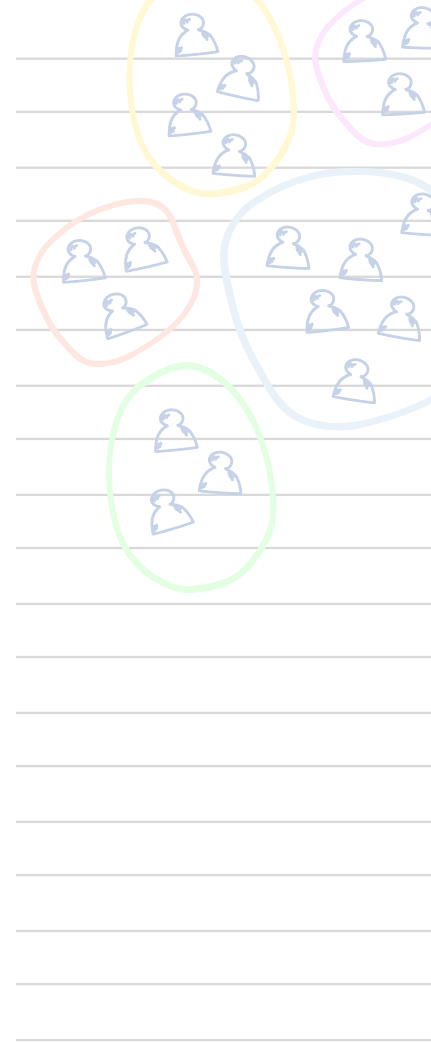
n : nombre d'individus dans le nuage


n_c : nombre d'individus dans la classe "c"

G : centre de gravité du nuage d'individus

G_c : centre de gravité de la classe "c"

M_i : point correspondant à l'individu i





CREDITS: This presentation template was
created by Slidesgo,
including icons by Flaticon and
infographics & images by Freepik

