



OPENCLASSROOMS

Projet 5

-

Segmentez des clients d'un site e-commerce



CentraleSupélec

Pierrick BERTHE

Formation Expert en Data Science
Openclassrooms – CentraleSupélec

août 2023 → avril 2024



I – Problématique



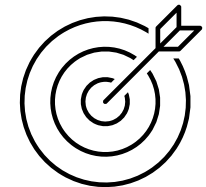
Problématique

Le service marketing veut réaliser la segmentation de ses clients pour différencier les bons et moins bons clients en termes de commandes et de satisfaction.



Missions :

1. Réaliser une courte analyse exploratoire.
2. Tester différents modèles de segmentation pour trouver la meilleure segmentation (>RFM)
3. Réaliser une analyse de la stabilité des segments au cours du temps.



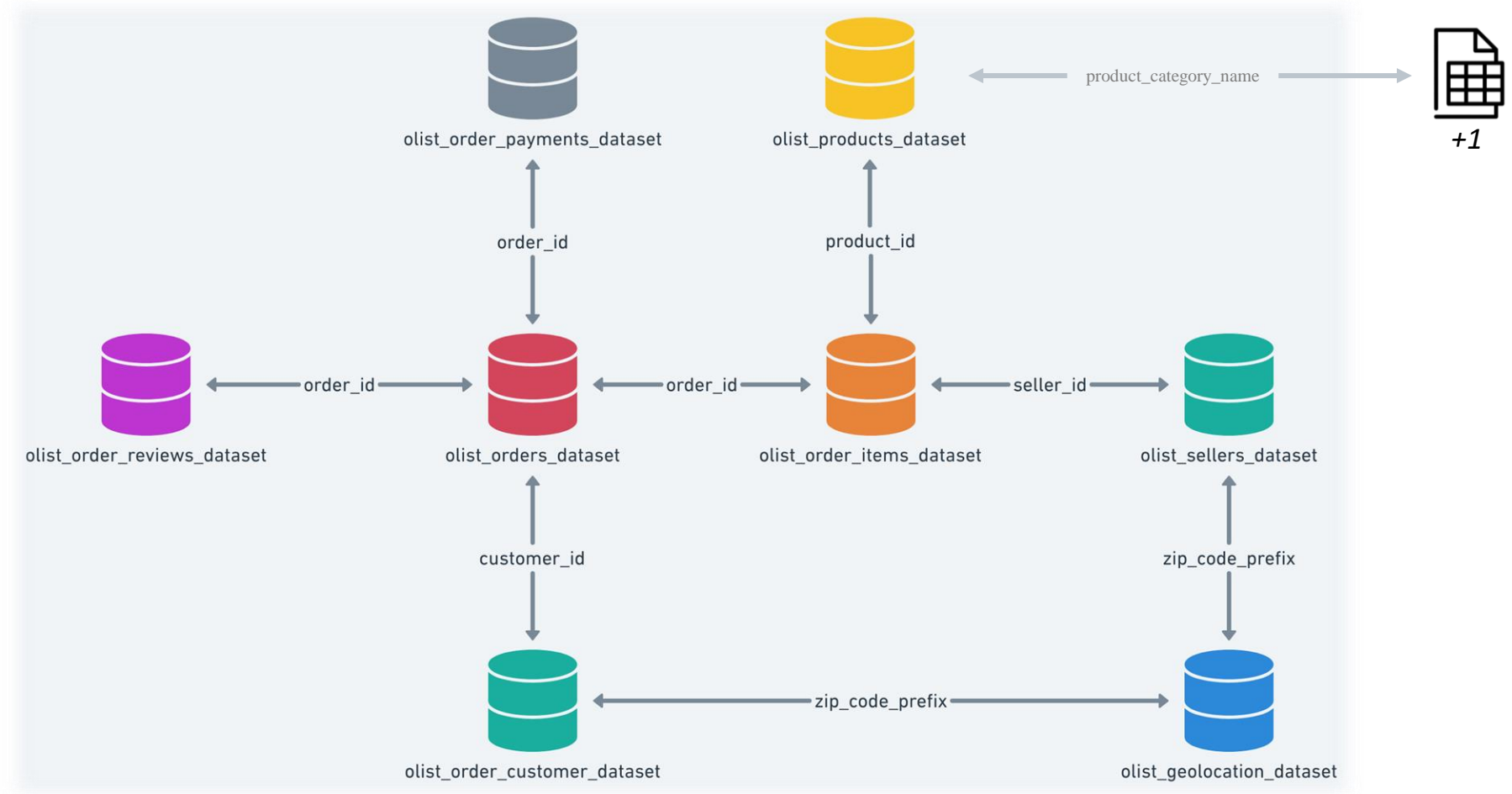


I – Problématique

II – Présentation du jeu de données

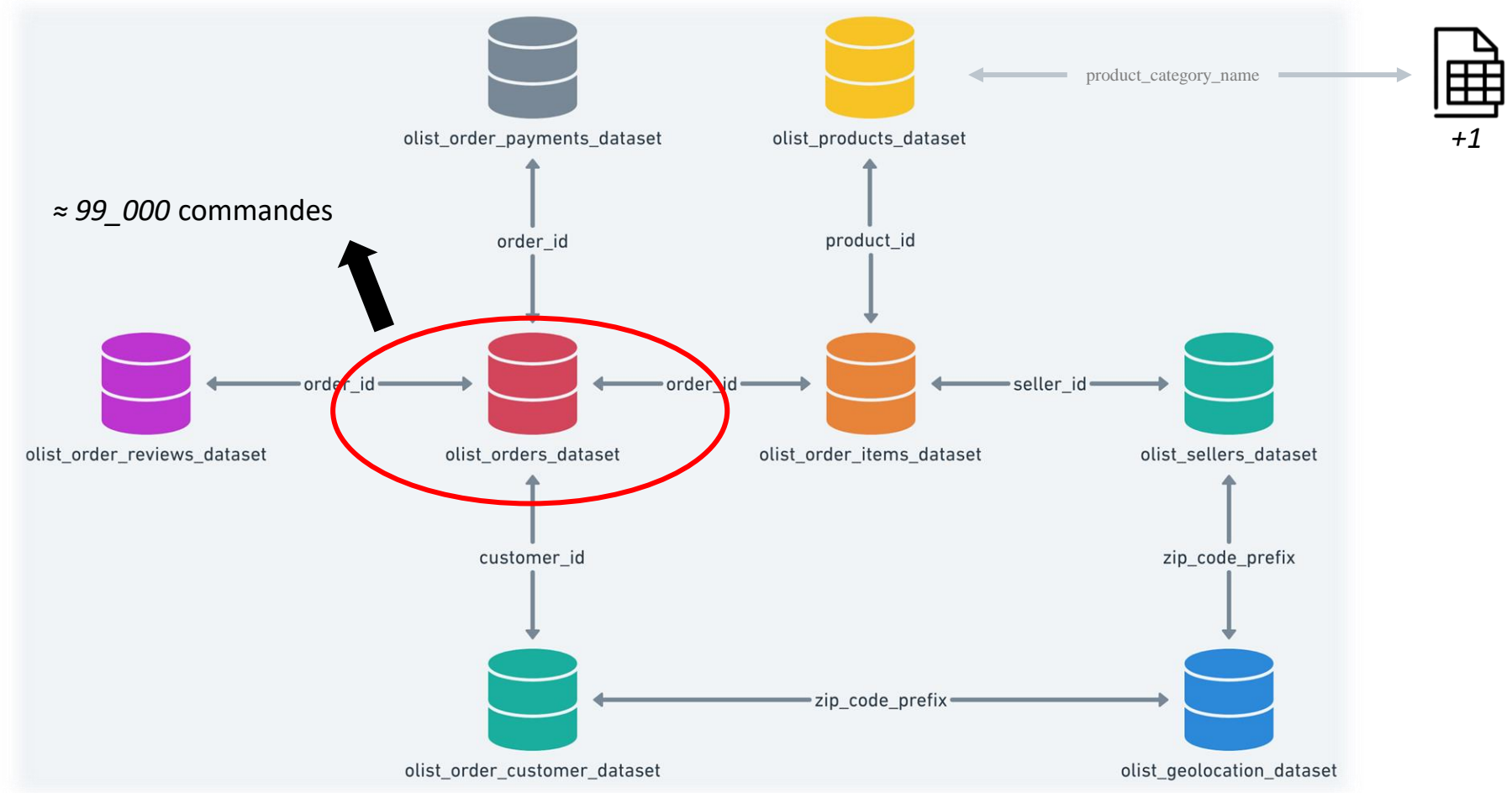


Présentation du jeu de données





Présentation du jeu de données

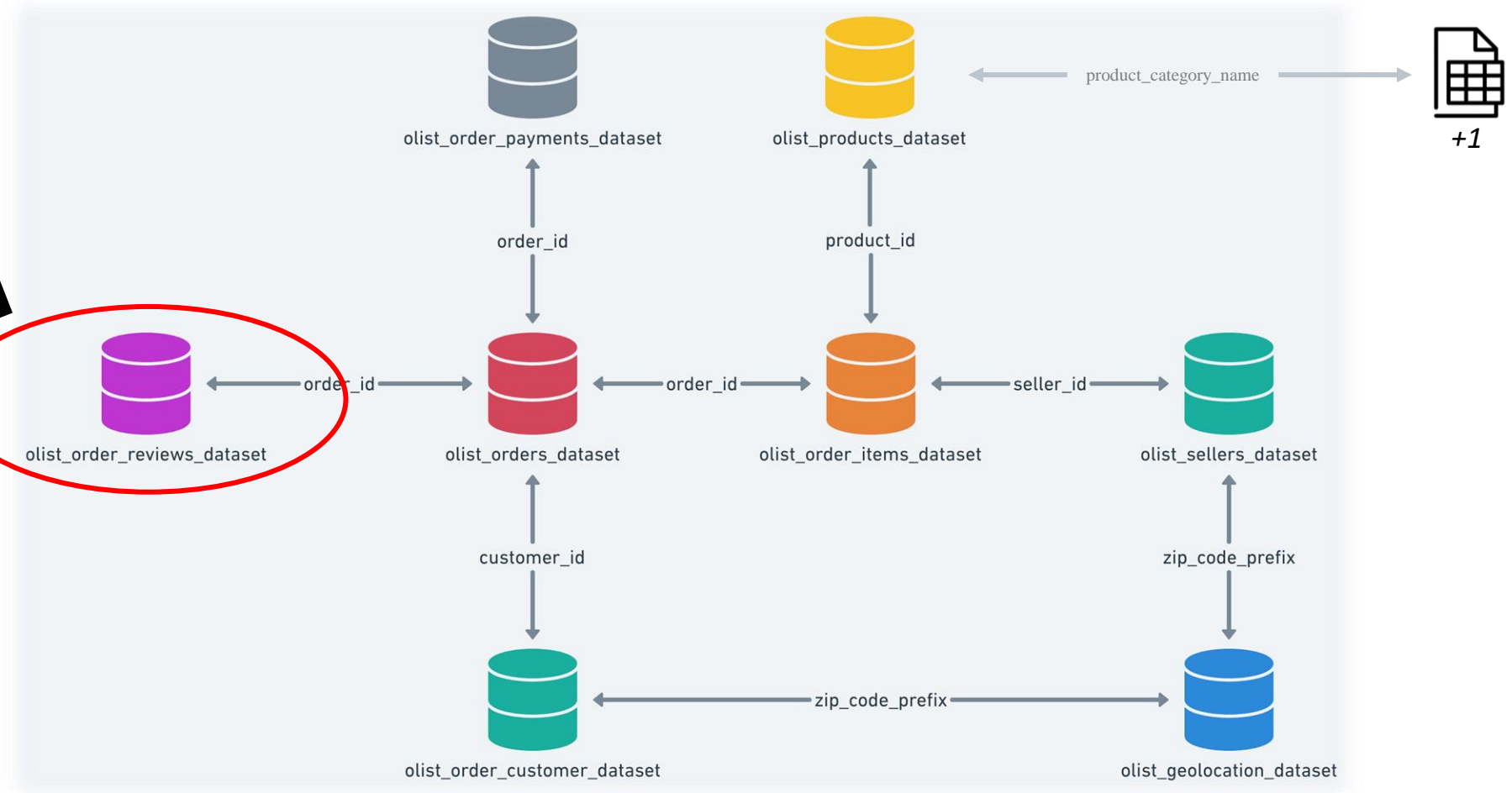




Présentation du jeu de données



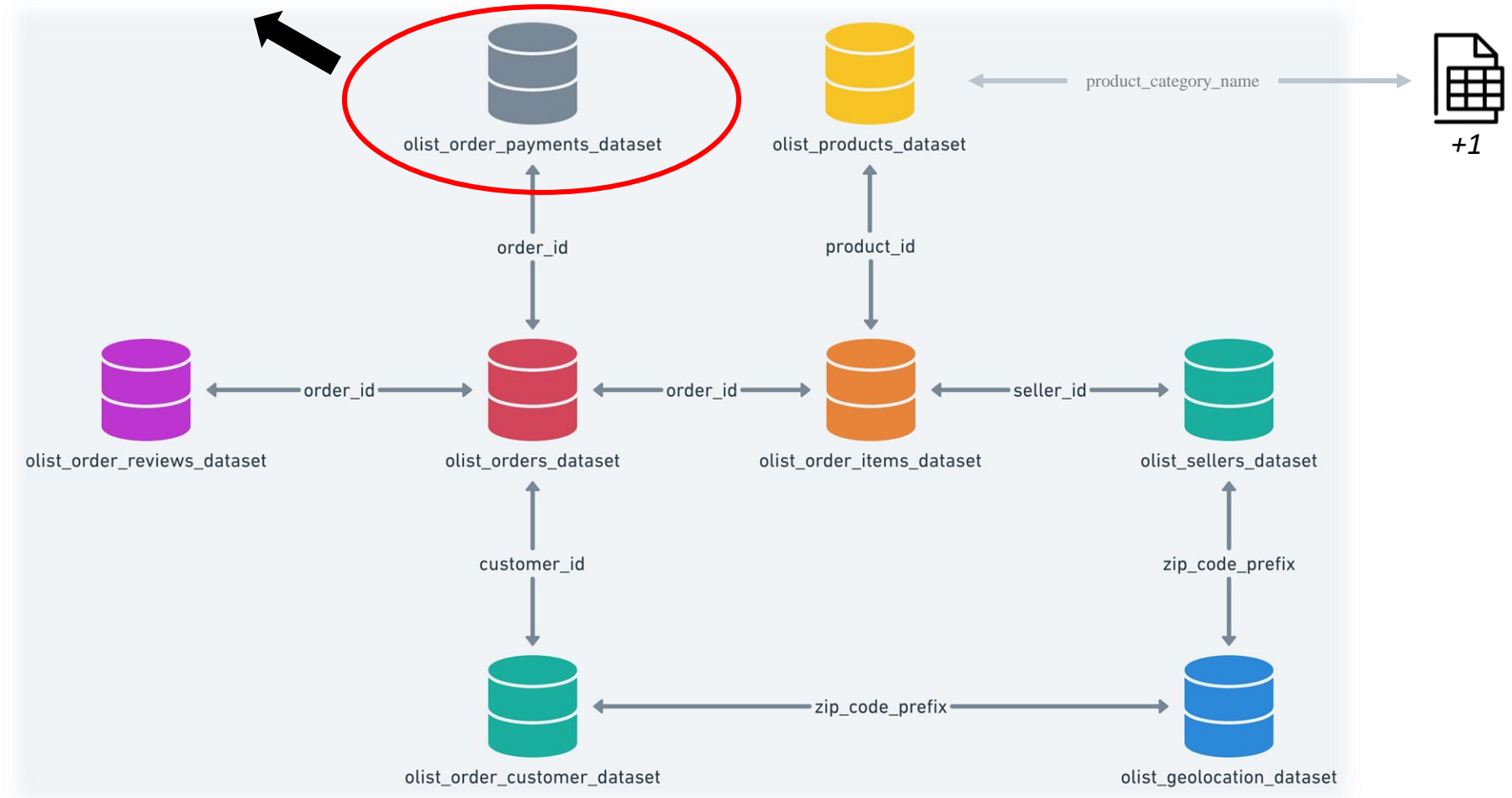
≈ 99_000 avis





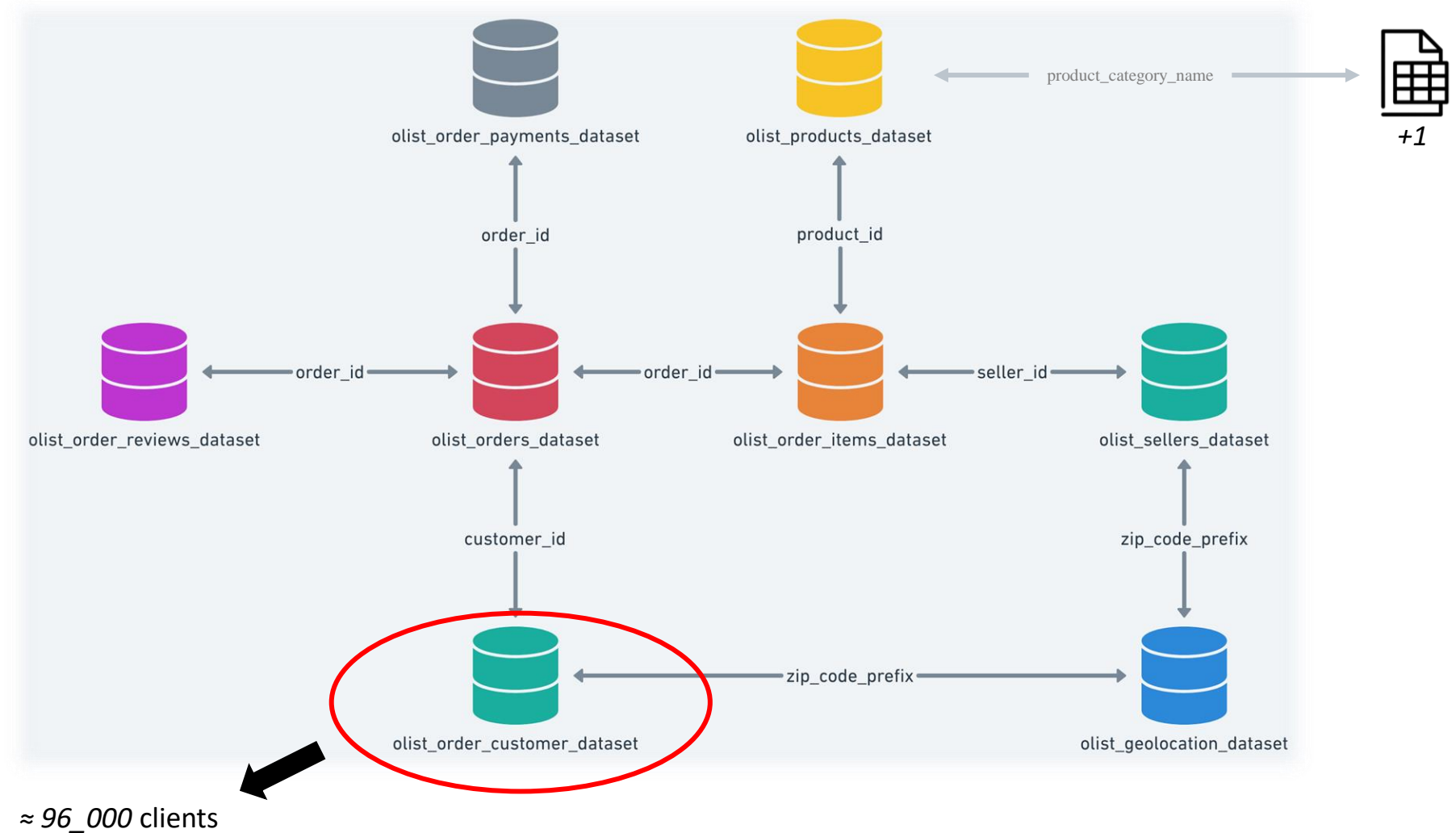
Présentation du jeu de données

≈ 103_000 paiements



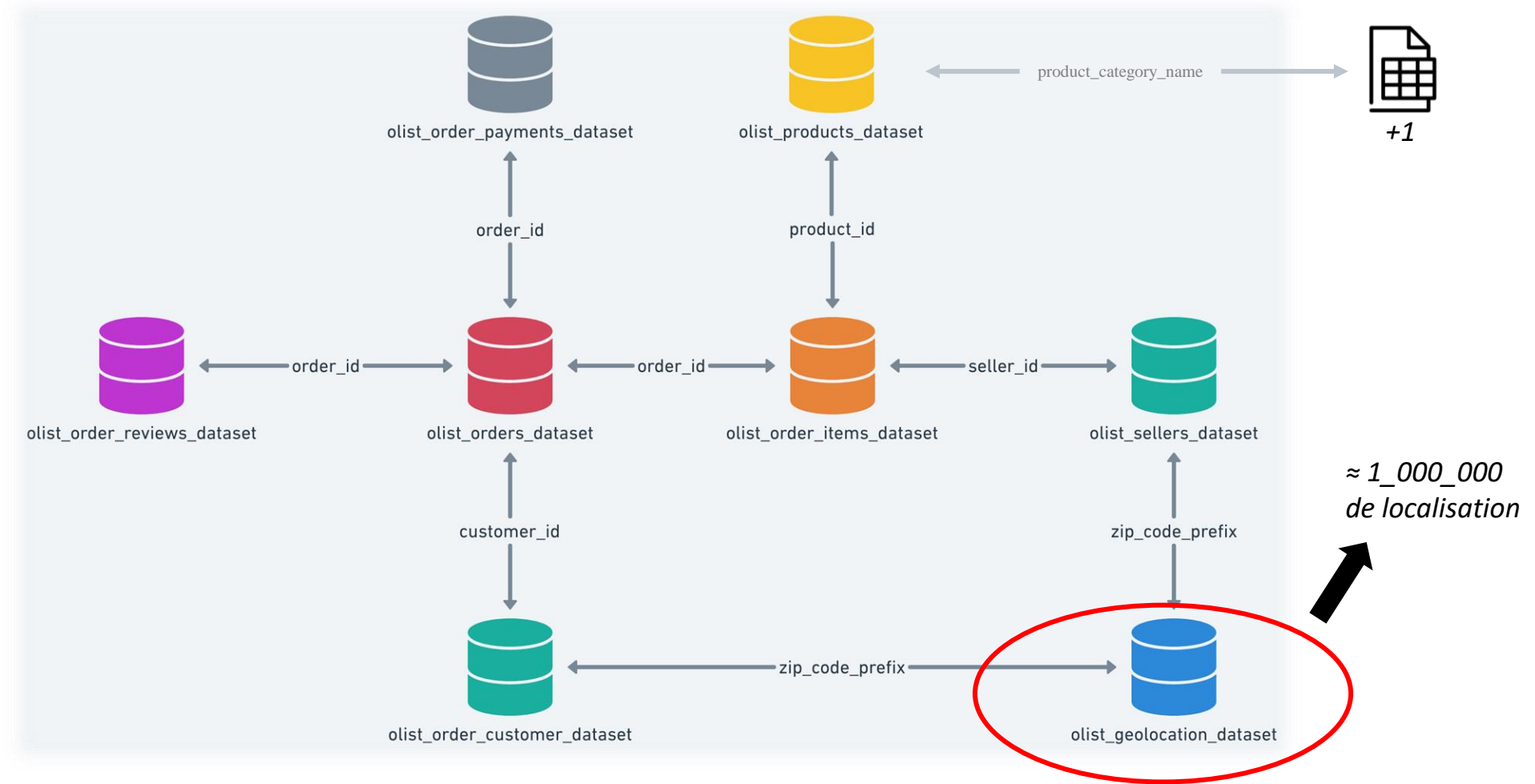


Présentation du jeu de données



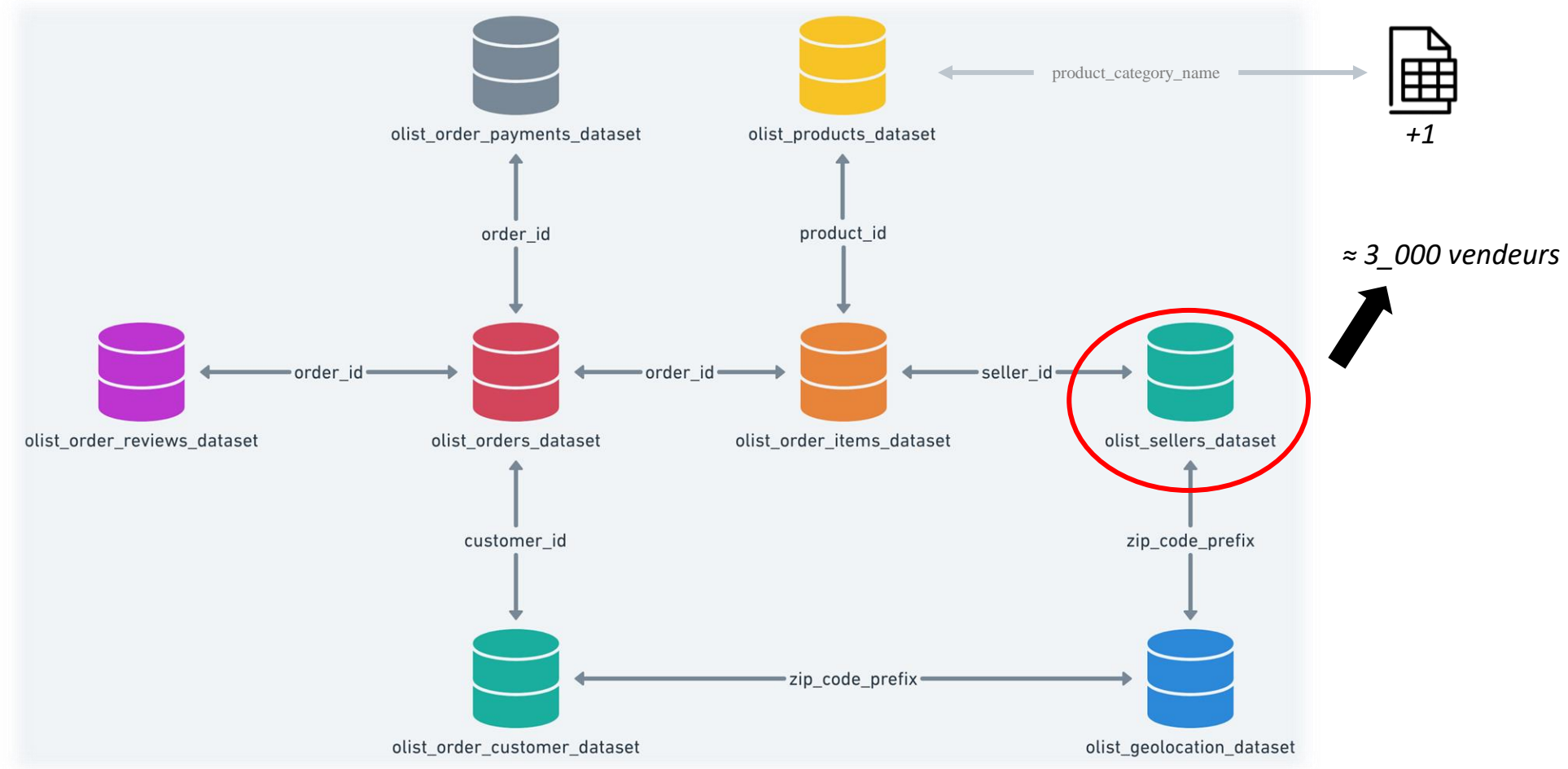


Présentation du jeu de données



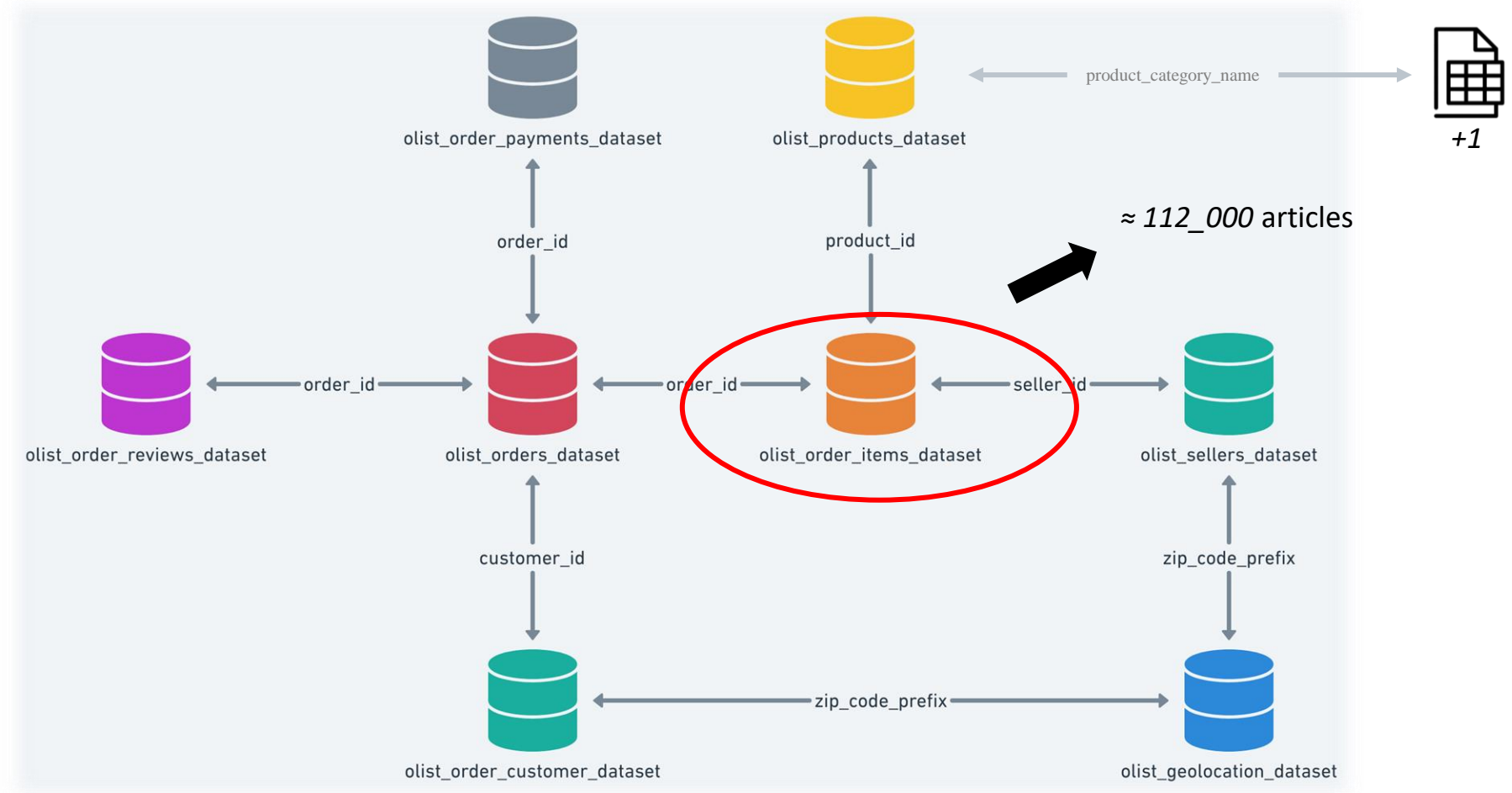


Présentation du jeu de données



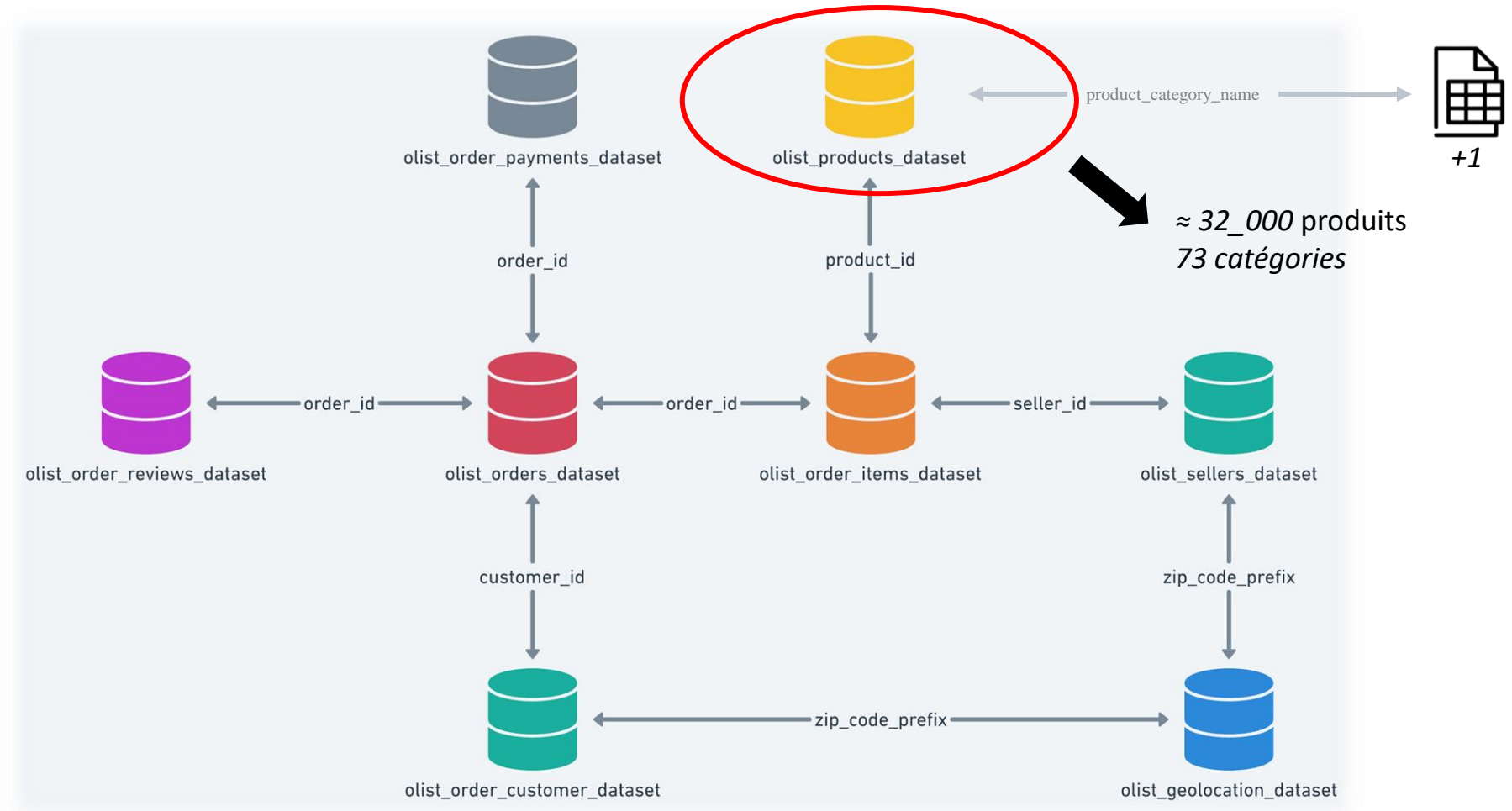


Présentation du jeu de données



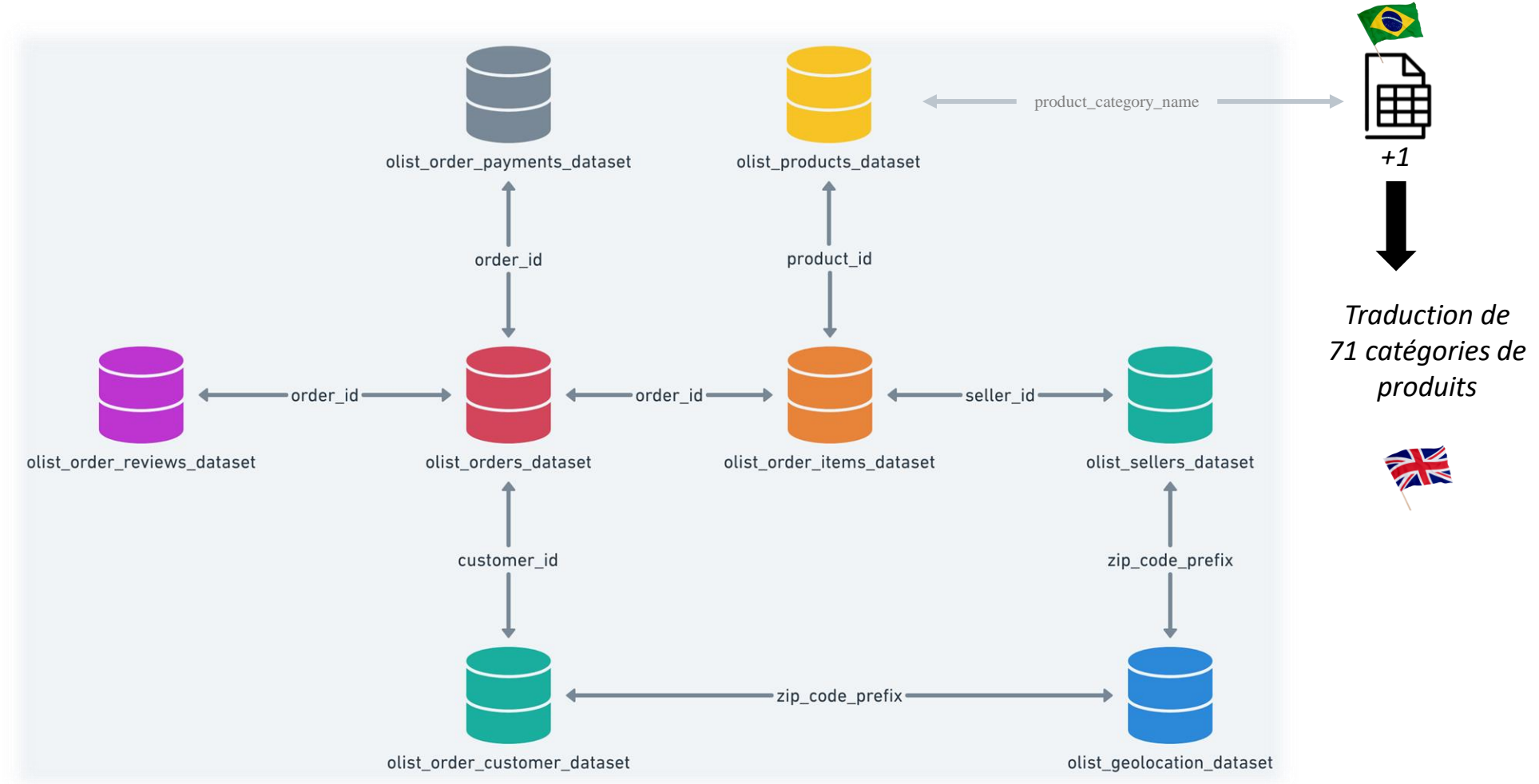


Présentation du jeu de données





Présentation du jeu de données





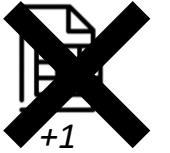
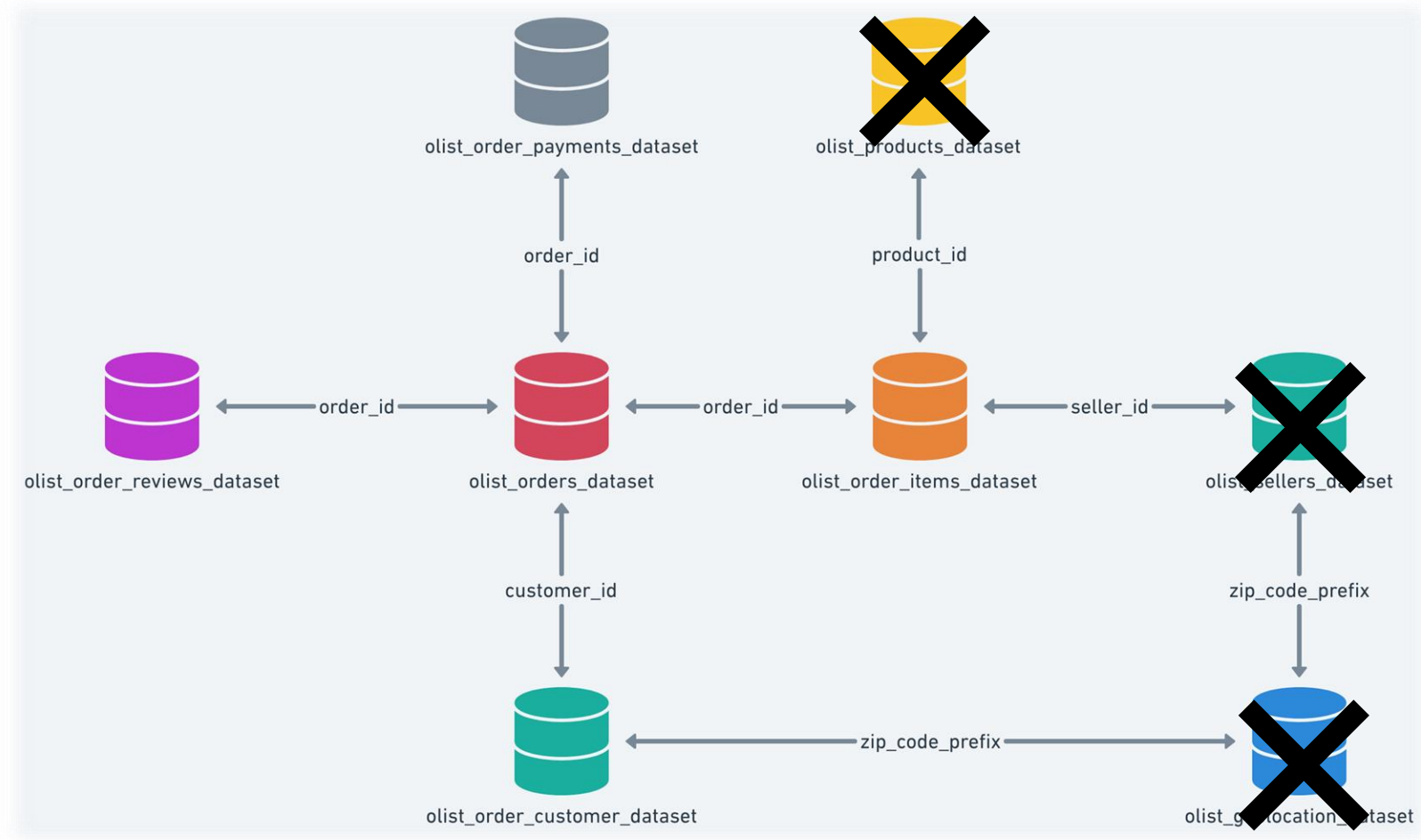
I – Problématique

II – Présentation du jeu de données

III - Nettoyage des données

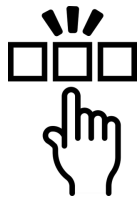
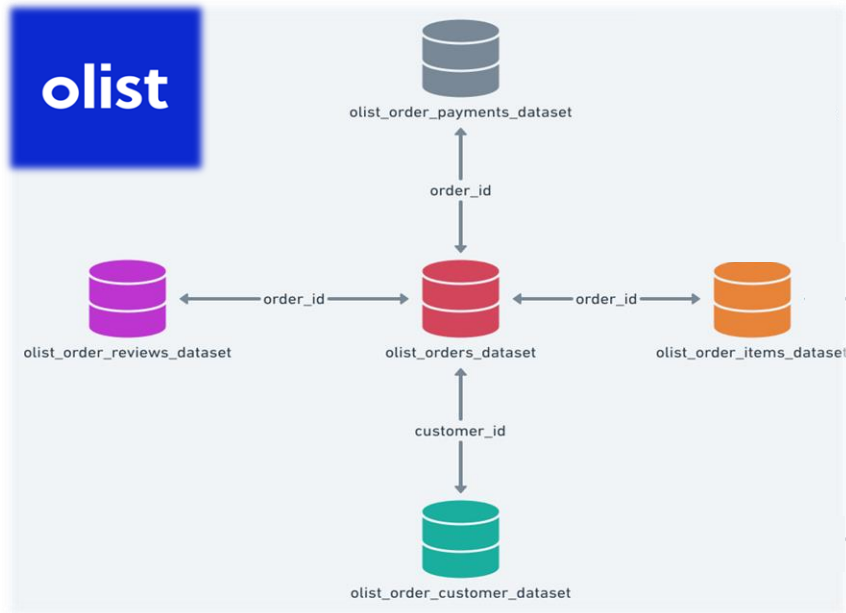


Nettoyage des données



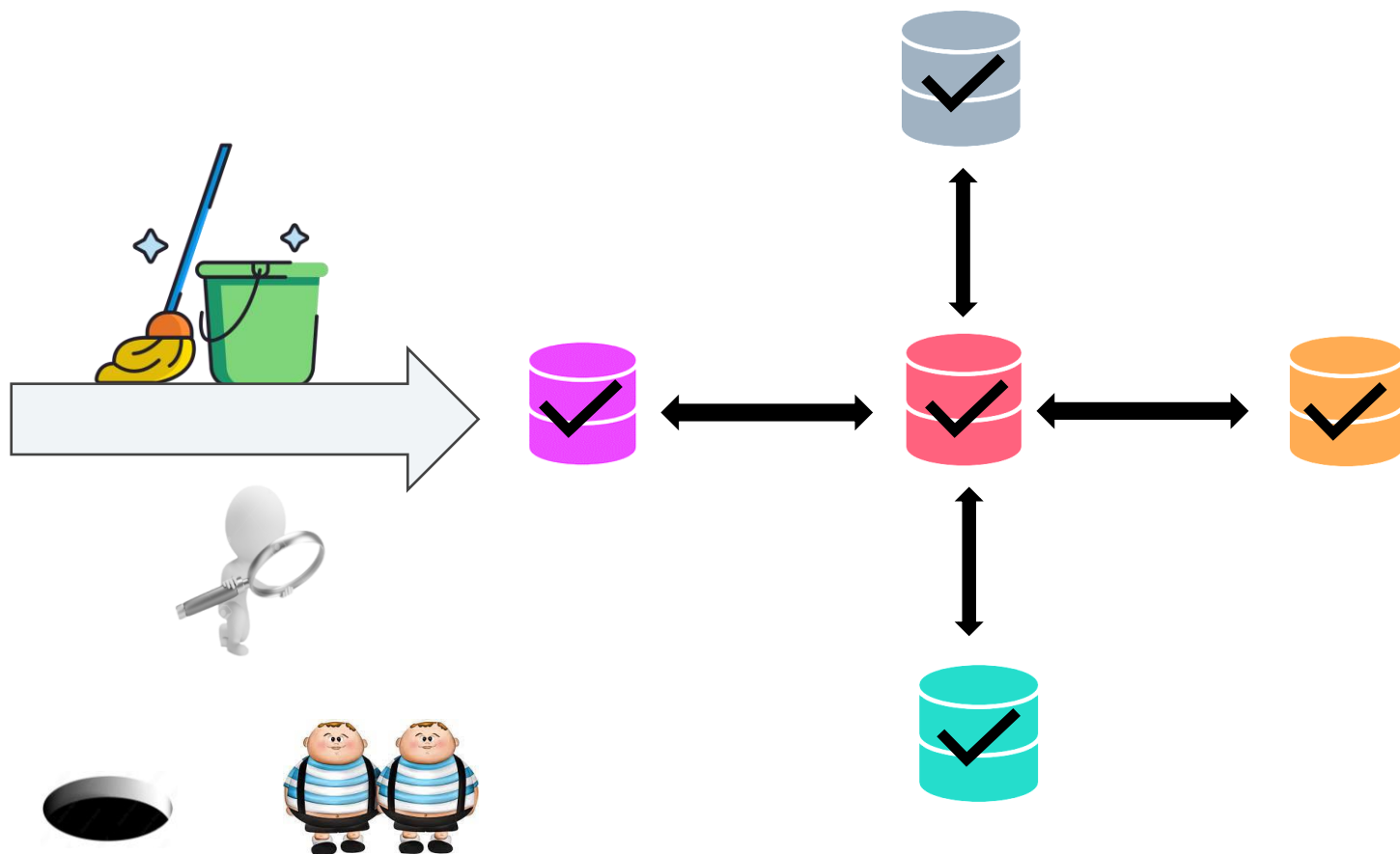
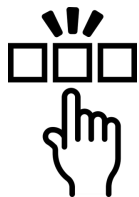
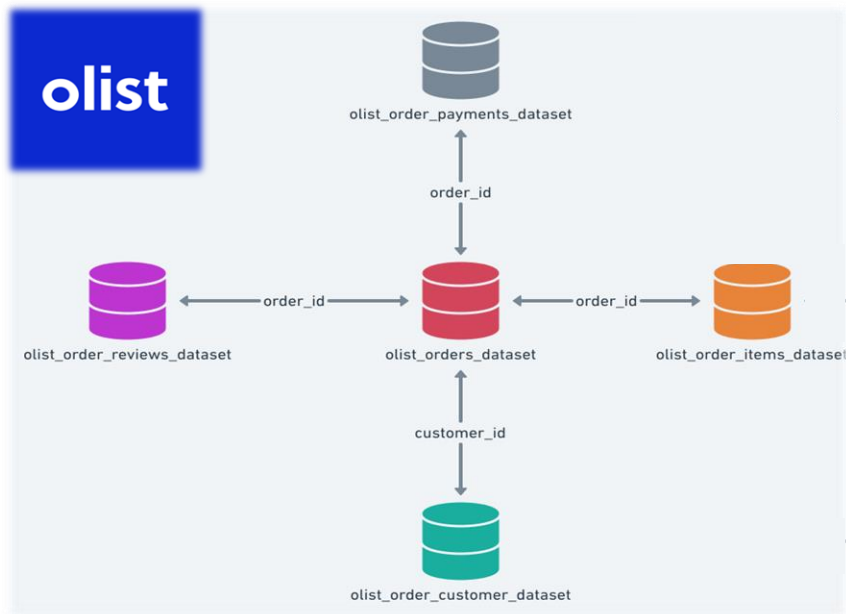


Nettoyage des données



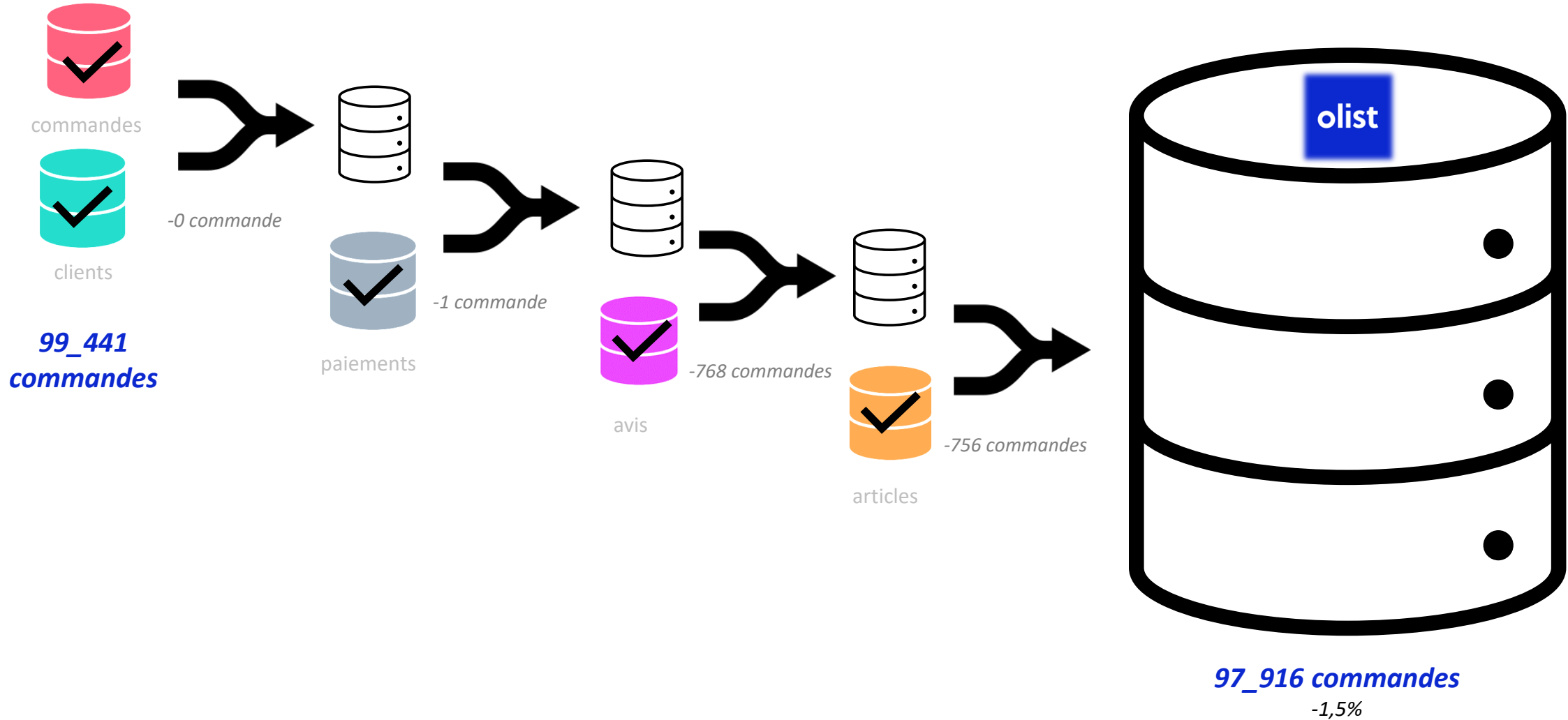


Nettoyage des données





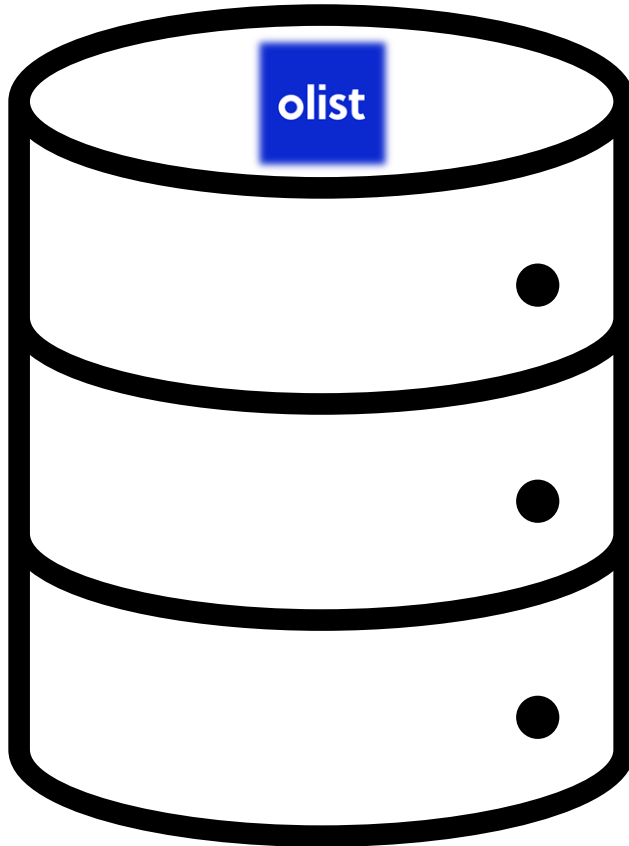
Nettoyage des données





Nettoyage des données

116_226 lignes



9 colonnes



97_916 commandes



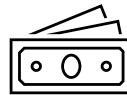
94_720 clients



04/09/2016



03/09/2018



0 R\$



14_000 R\$



I – Problématique

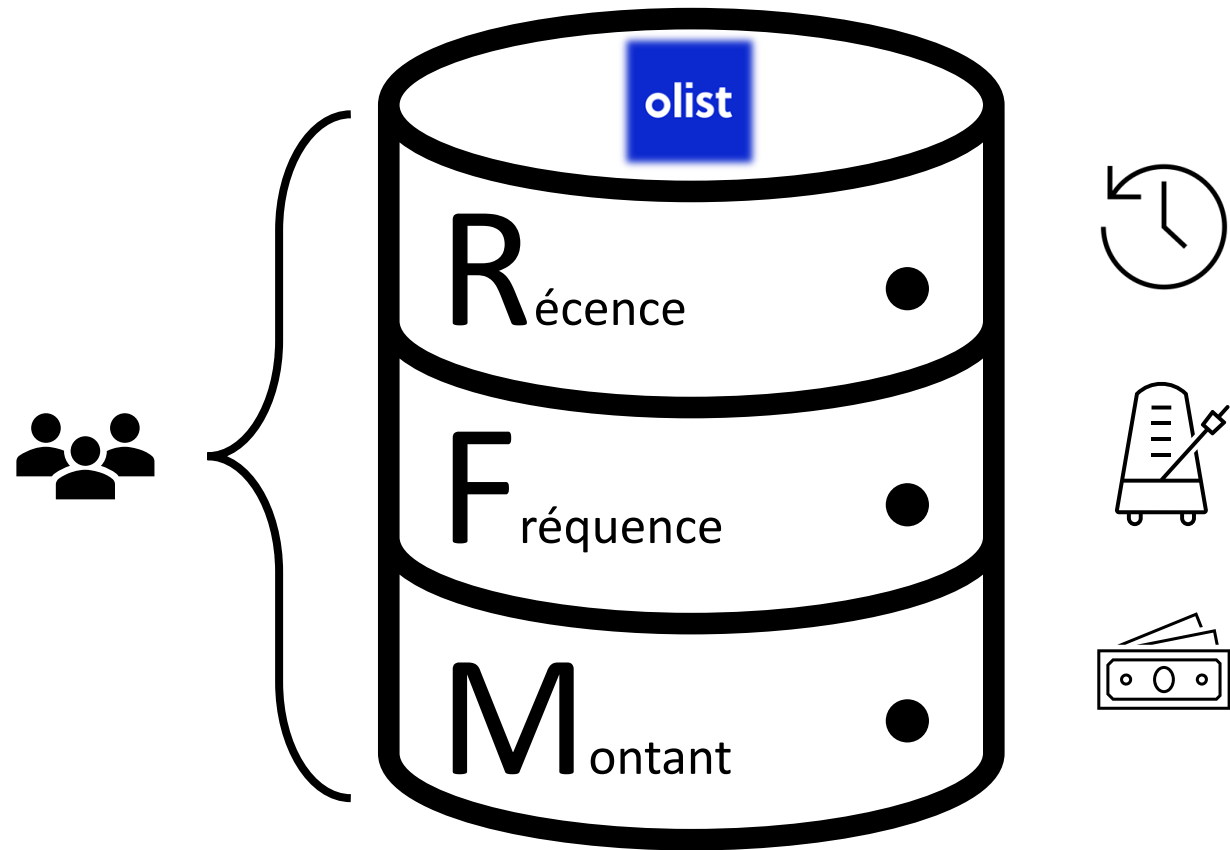
II – Présentation du jeu de données

III - Nettoyage des données

IV – Feature engineering

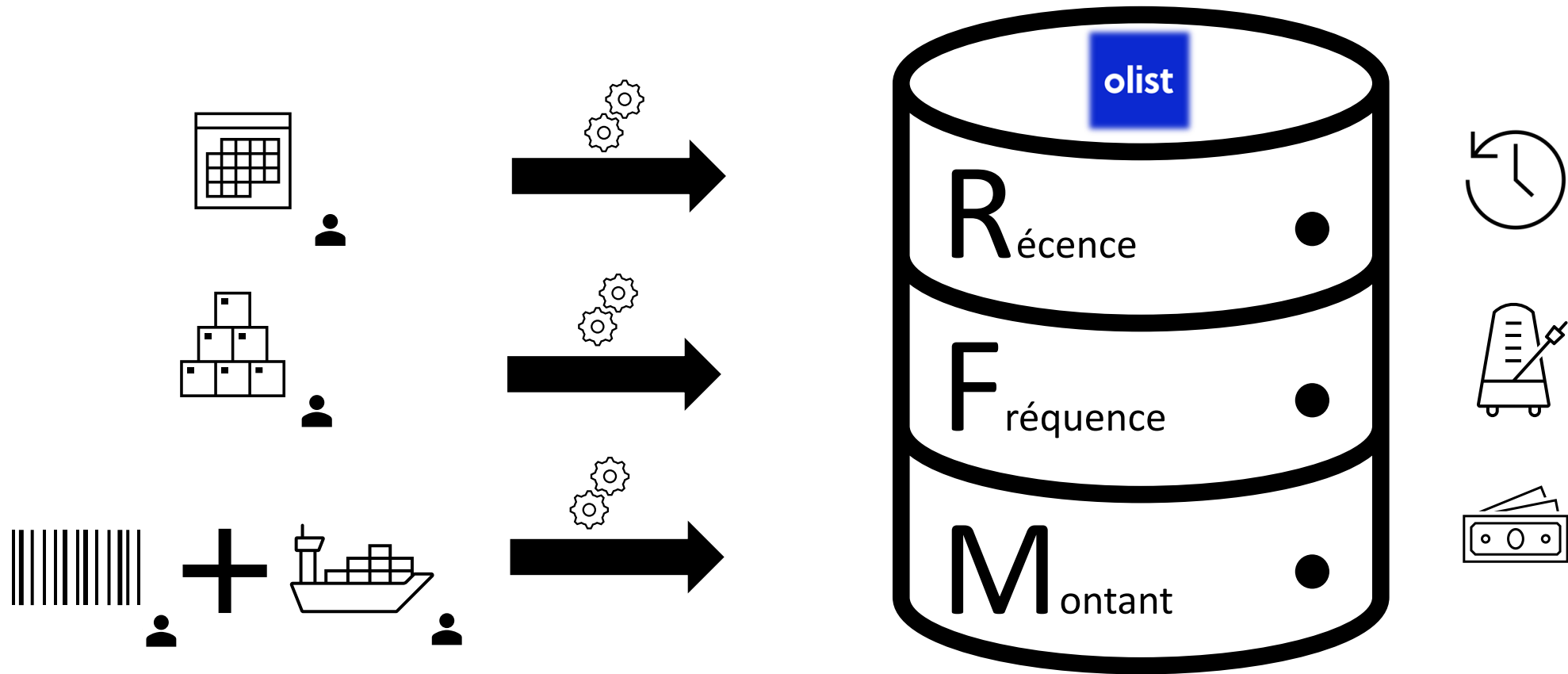


Feature engineering



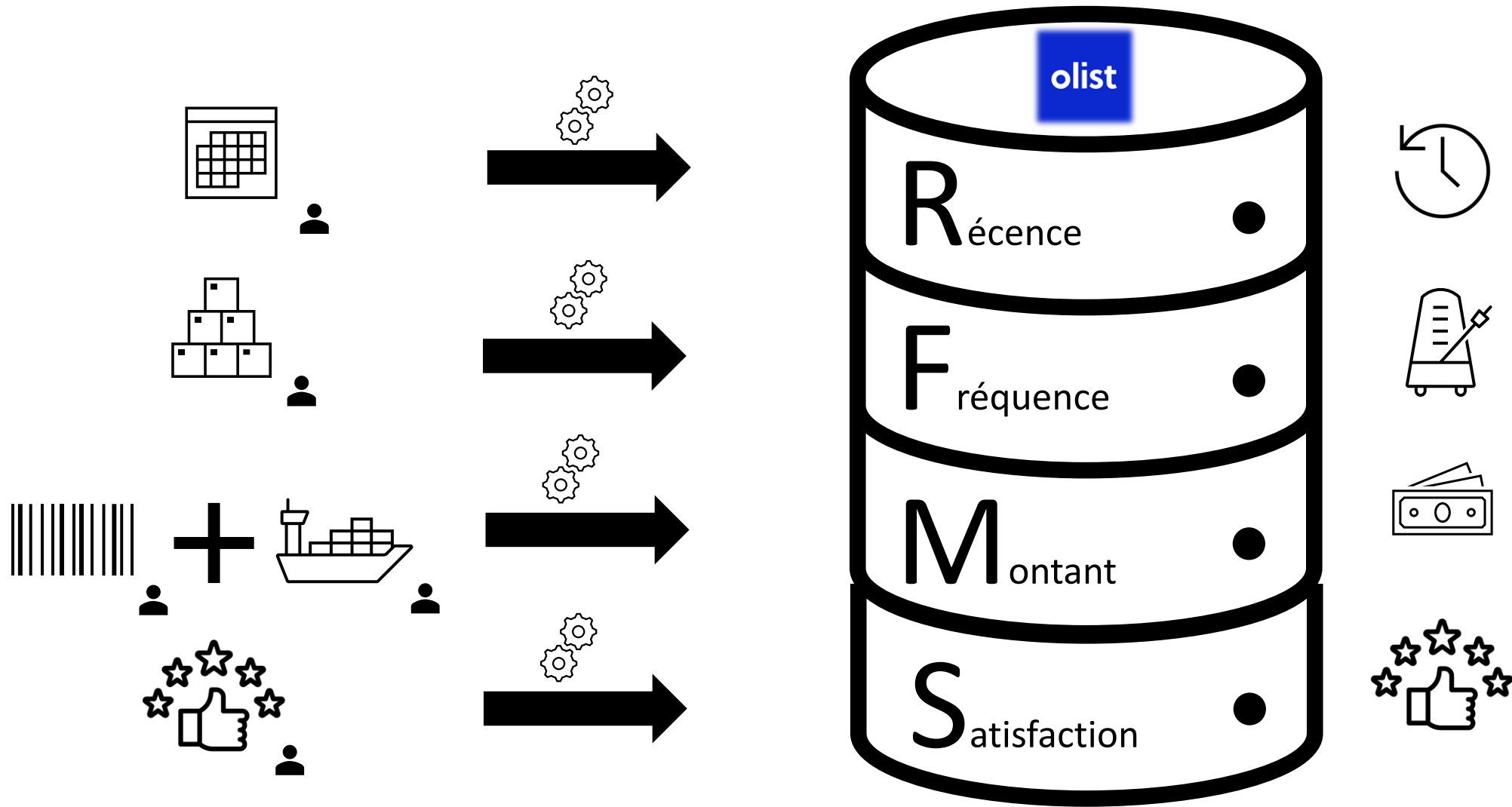


Feature engineering





Feature engineering





I – Problématique

II – Présentation du jeu de données

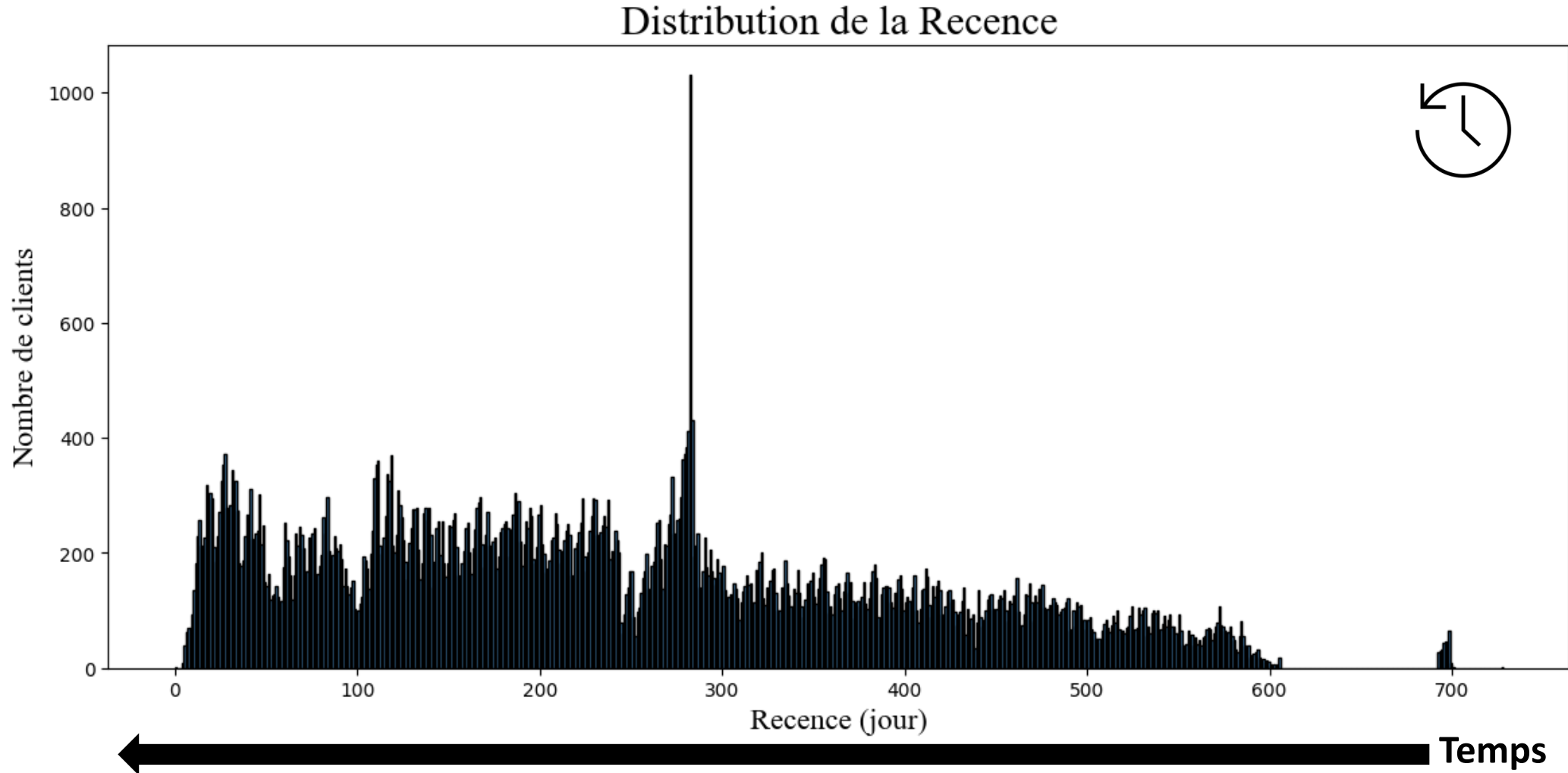
III - Nettoyage des données

IV – Feature engineering

V – Analyses exploratoires

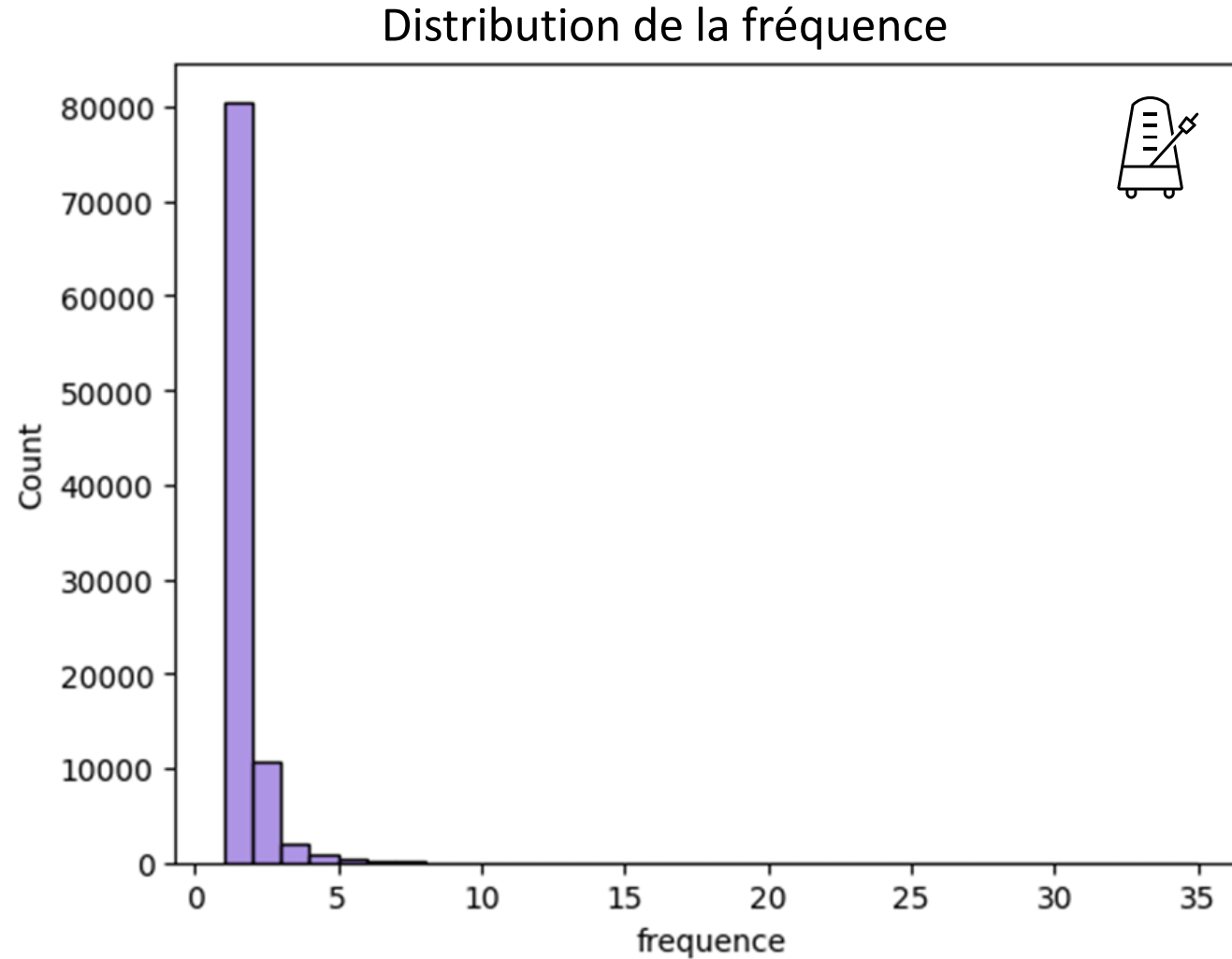


Analyses exploratoires





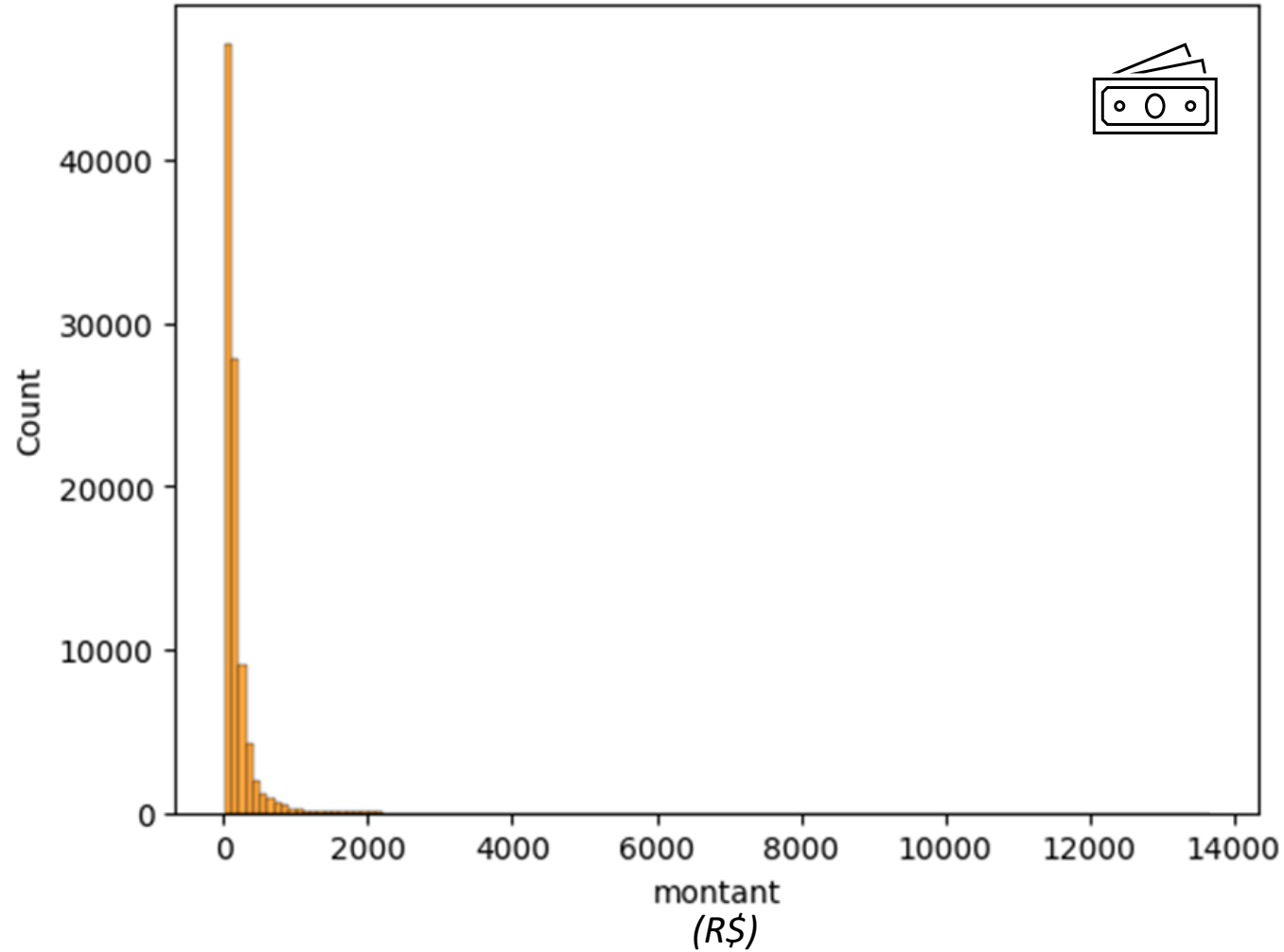
Analyses exploratoires





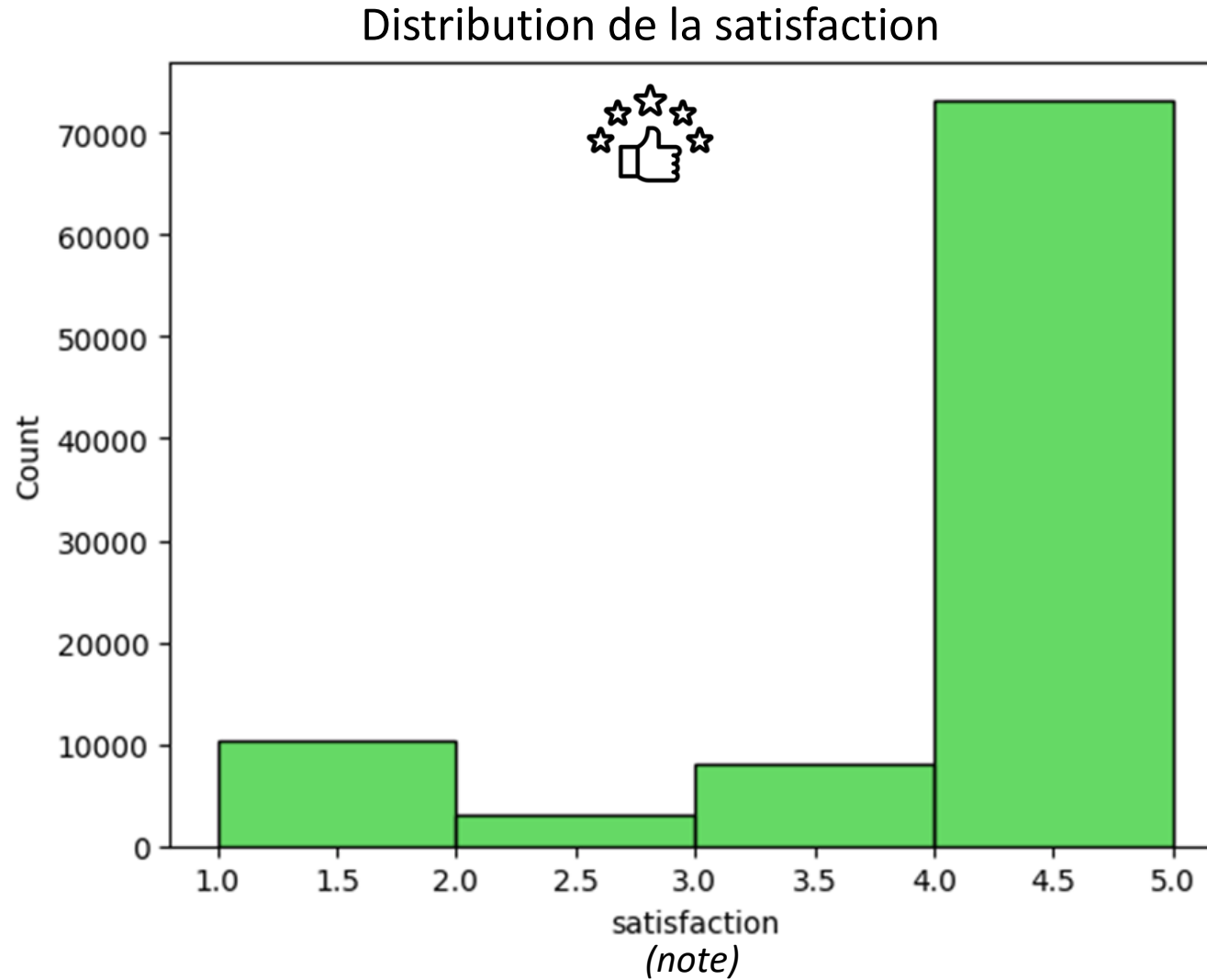
Analyses exploratoires

Distribution du montant



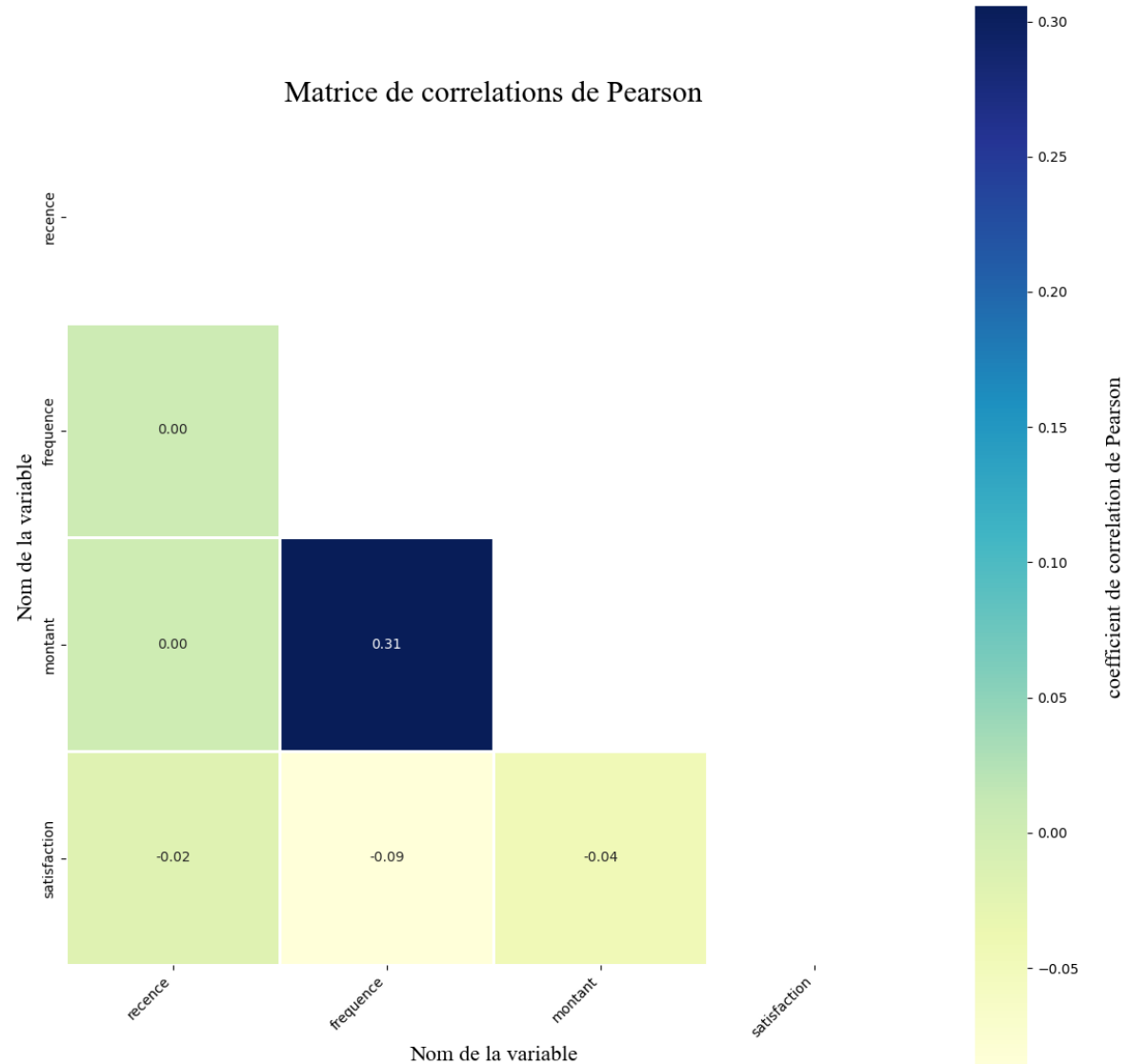


Analyses exploratoires





Analyses exploratoires





I – Problématique

II – Présentation du jeu de données

III - Nettoyage des données

IV – Feature engineering

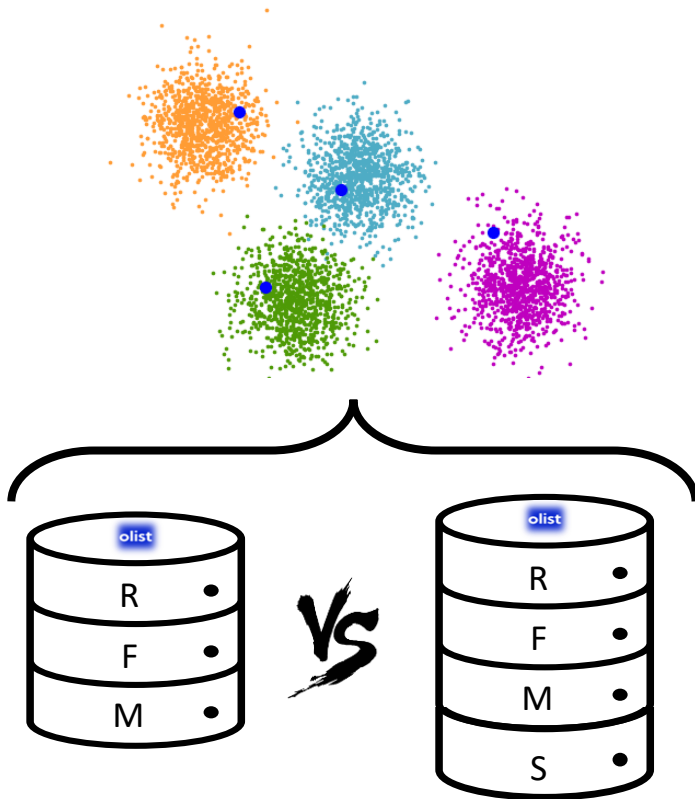
V – Analyses exploratoires

VI – Modèle de segmentation

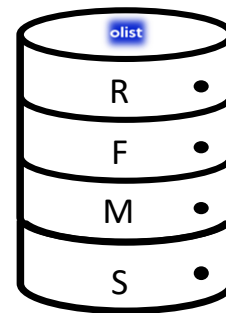
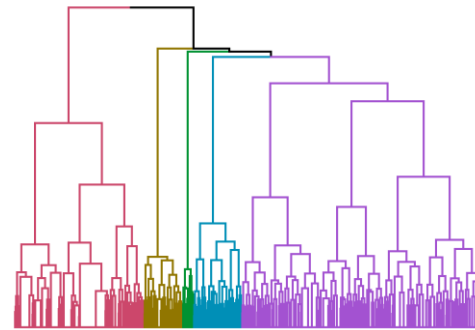


Modèle de segmentation

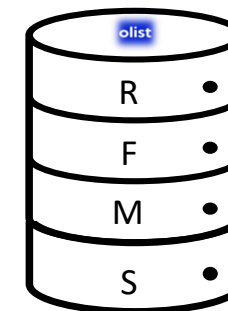
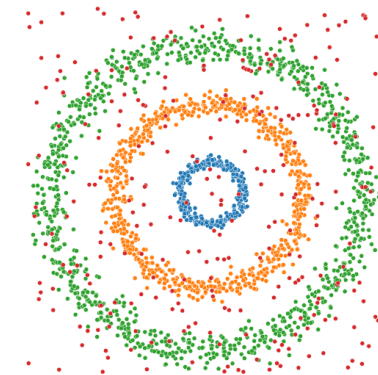
K-means



CAH



DBSCAN

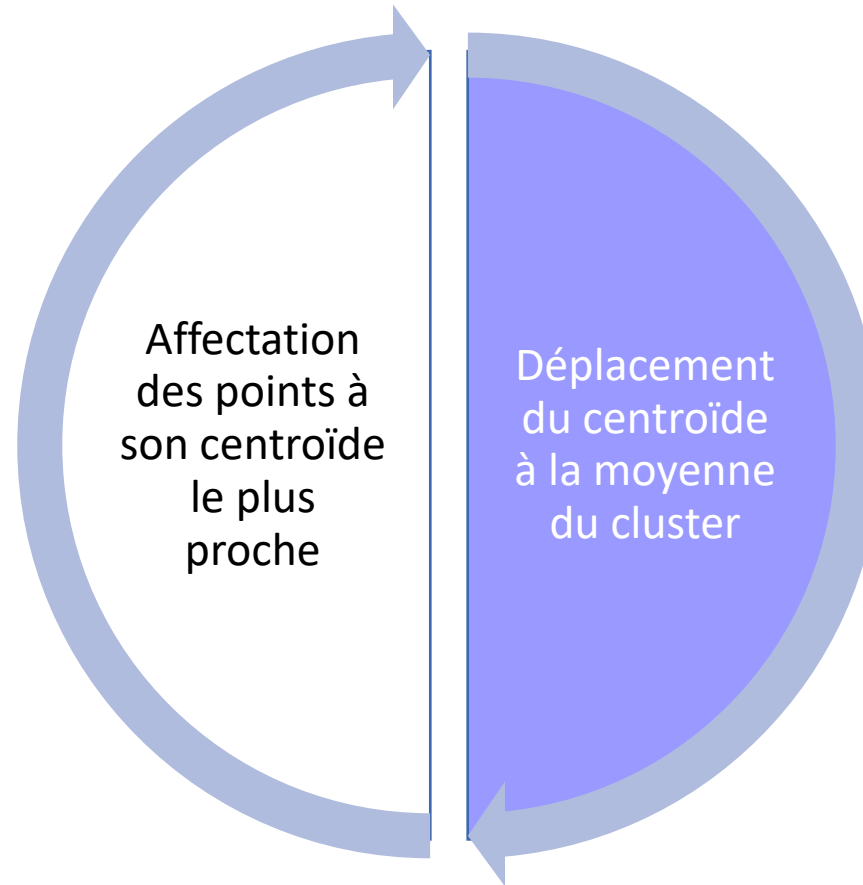
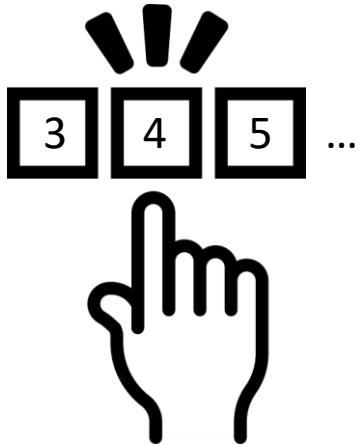




Modèle de segmentation

K-means

Nombre cluster?

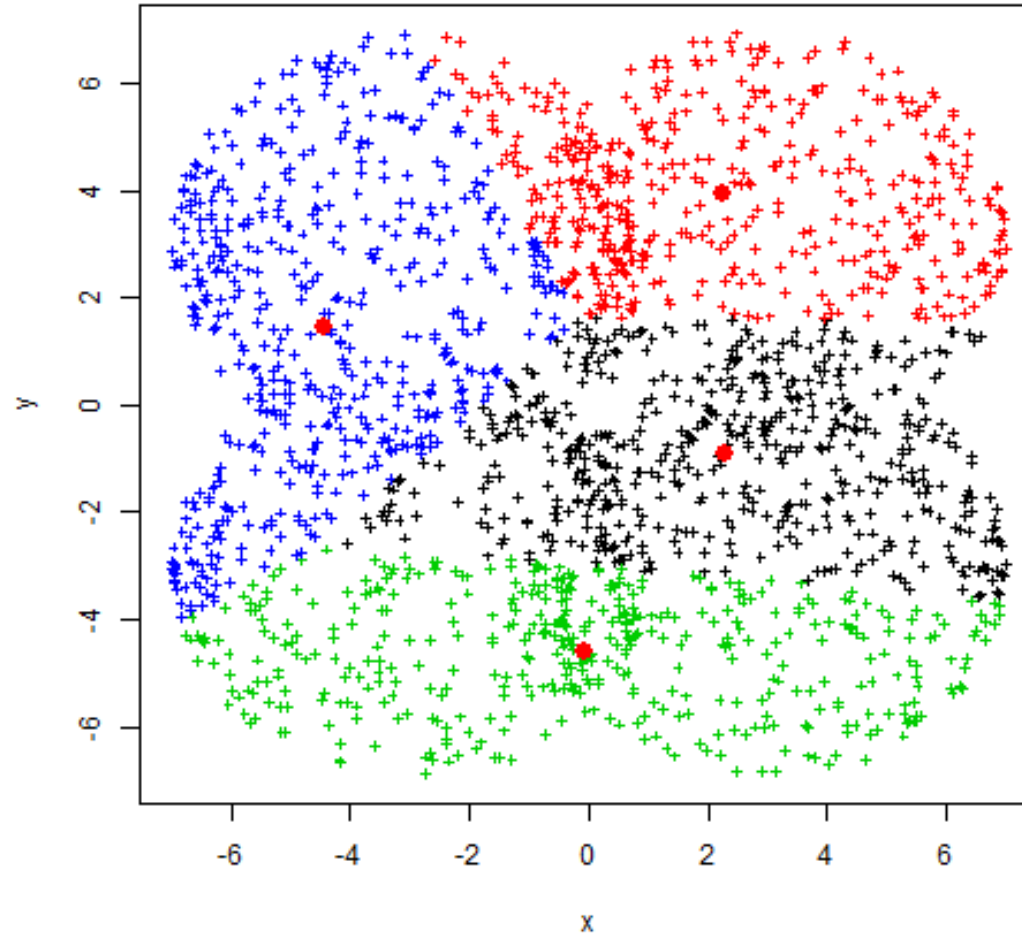


**1 cluster = 1 segment*



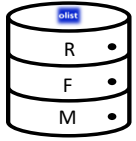
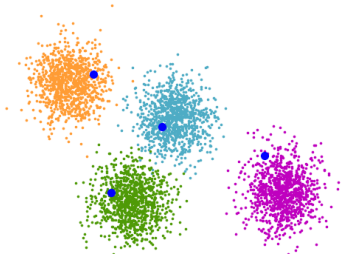
Modèle de segmentation

K-means



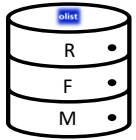
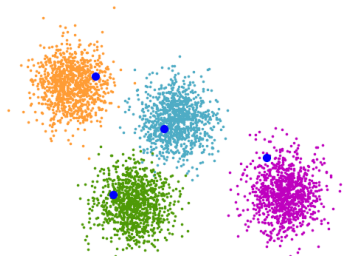


Modèle de segmentation

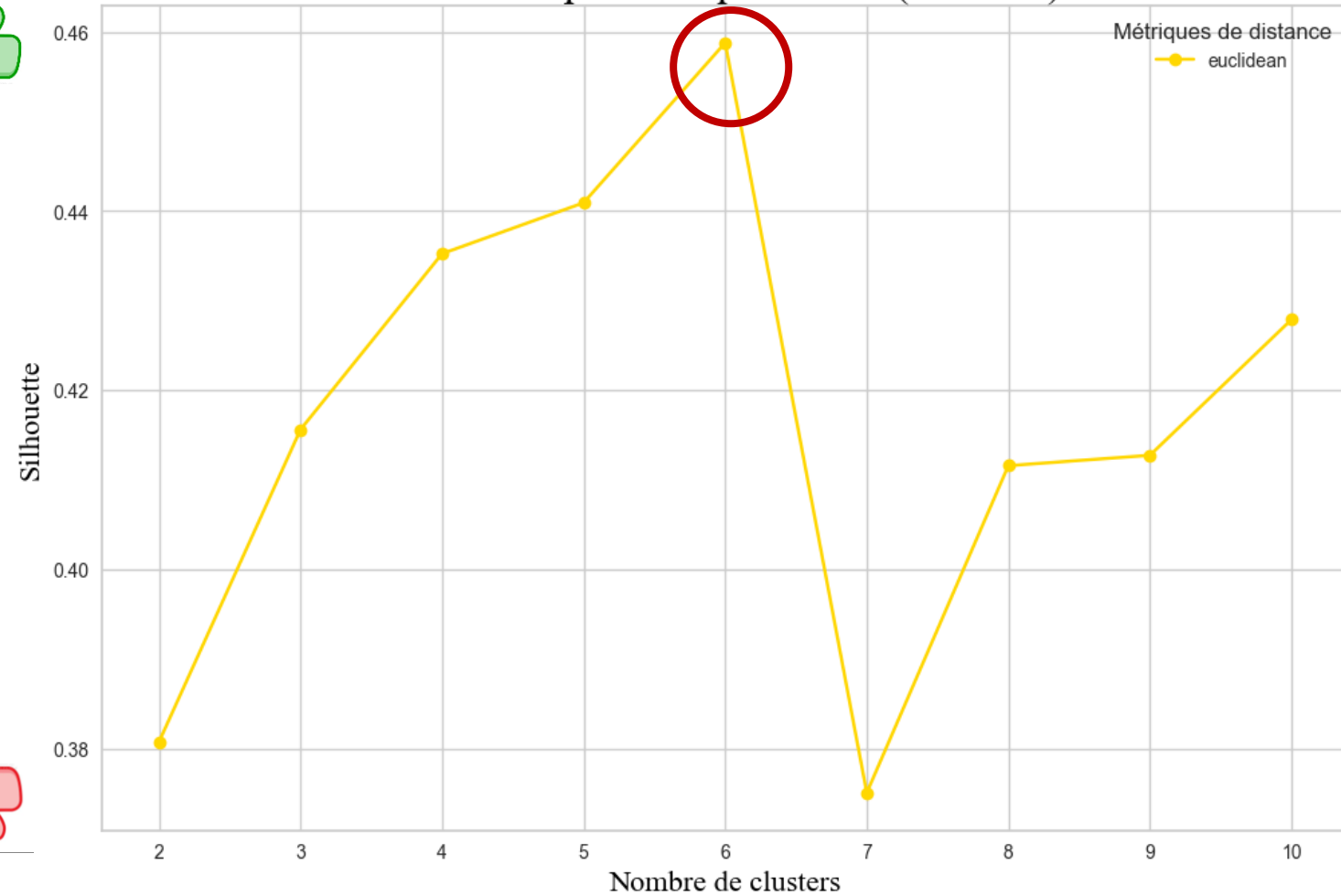




Modèle de segmentation

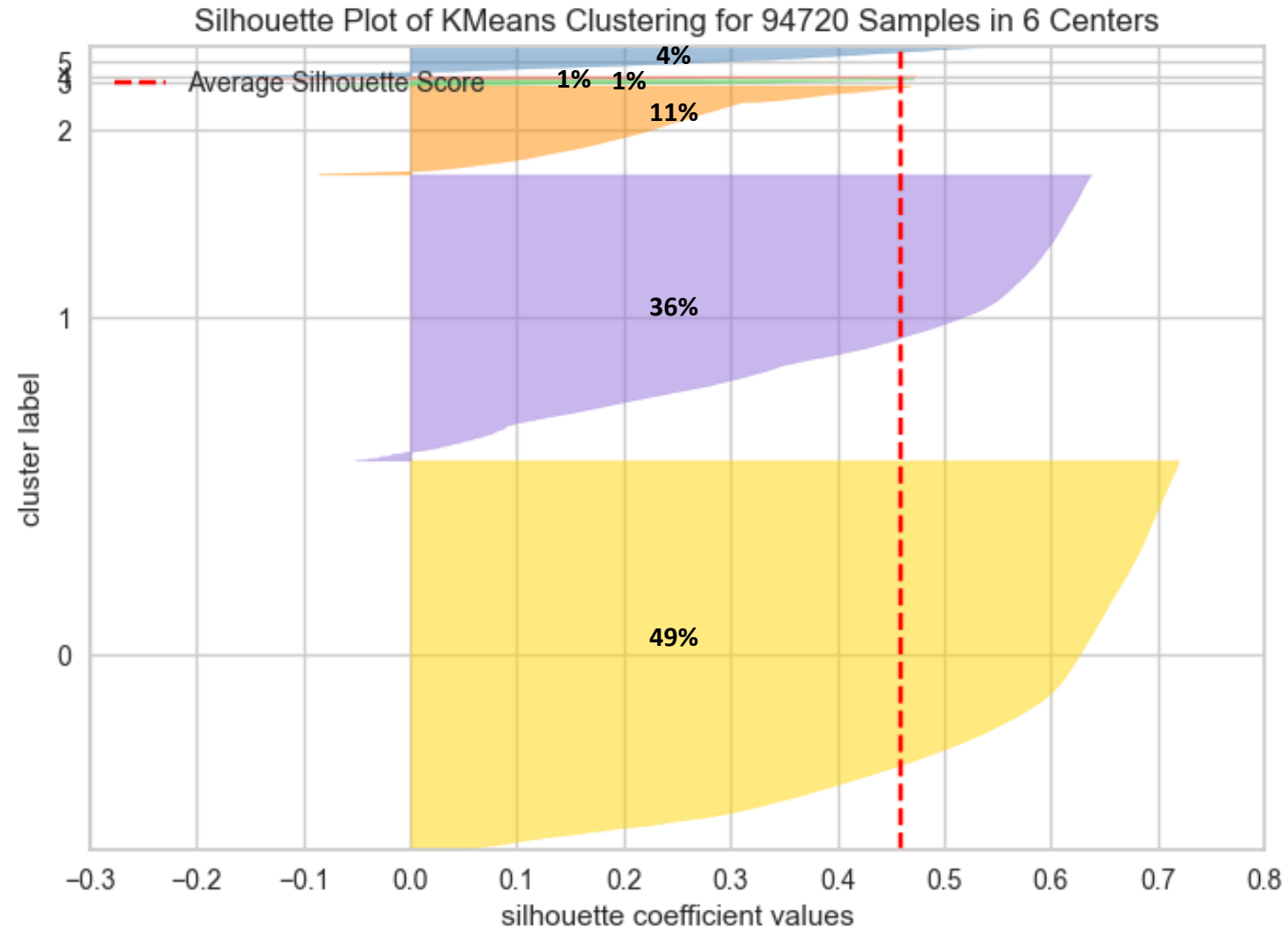
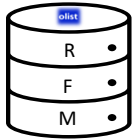
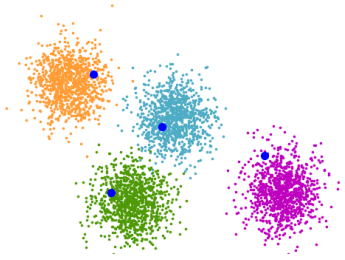


Silhouette pour chaque cluster (KMeans)



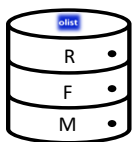
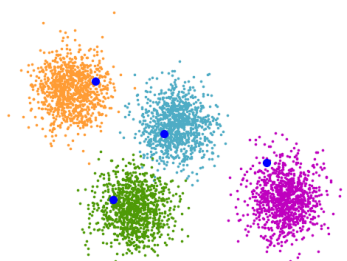


Modèle de segmentation

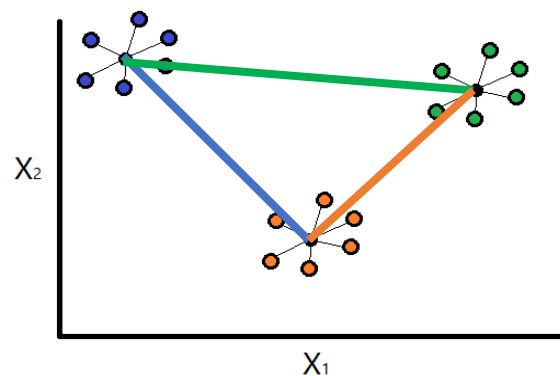




Modèle de segmentation



Indice Davies-Bouldin



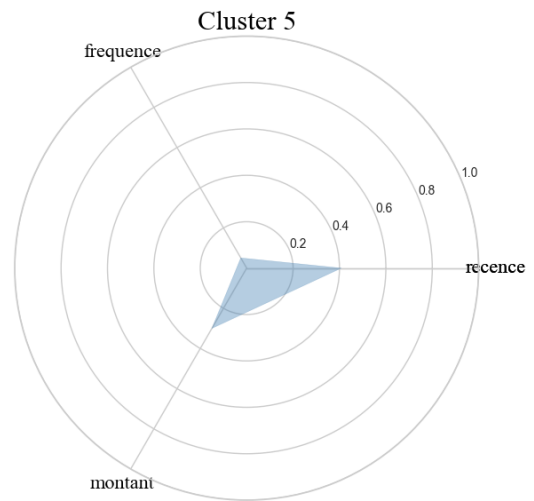
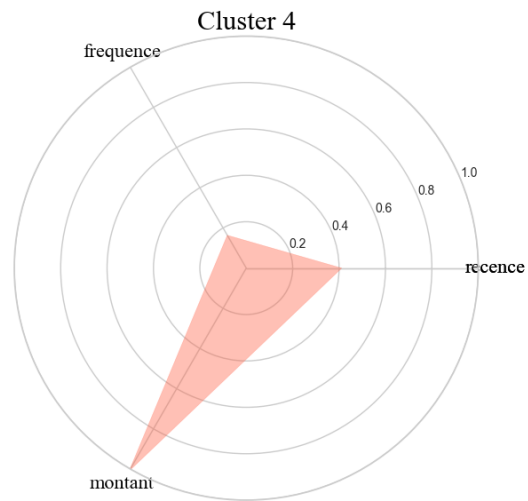
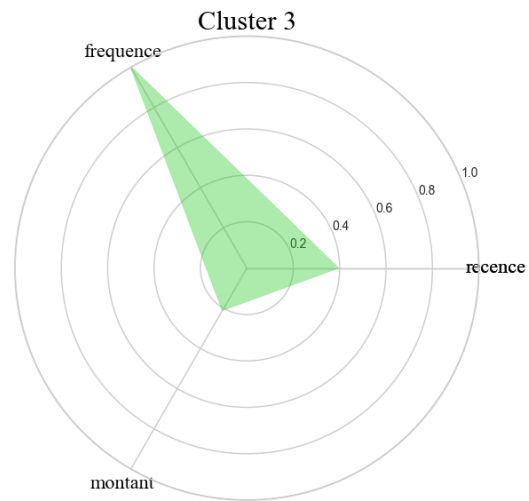
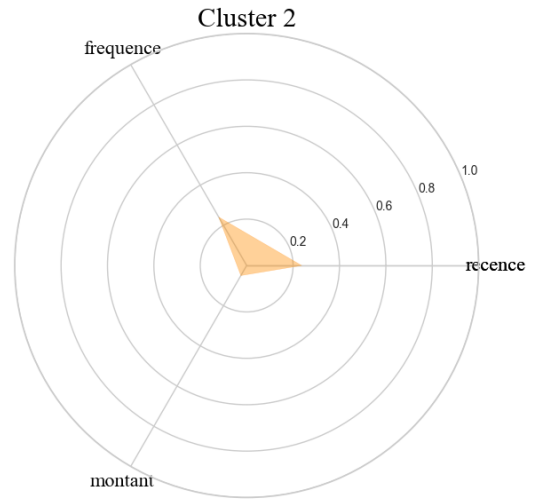
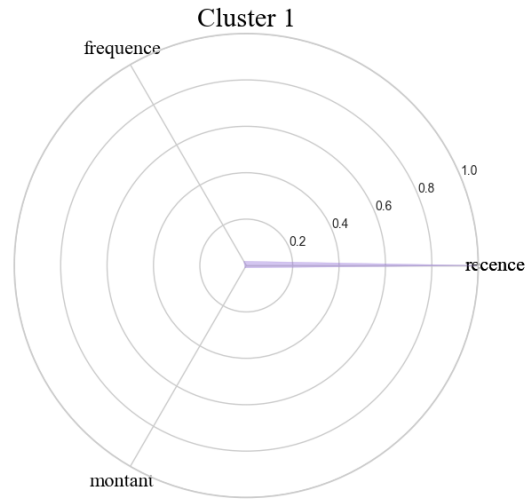
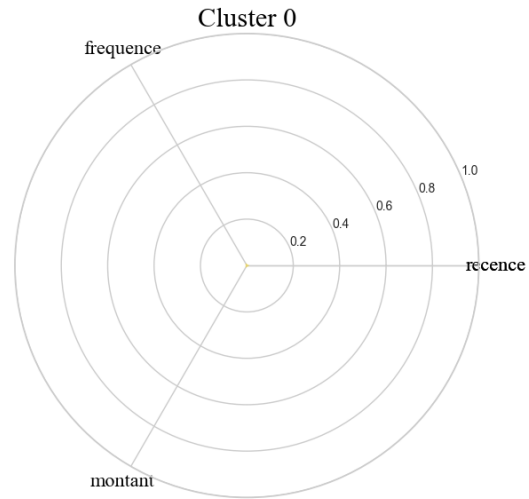
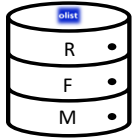
=

0,84



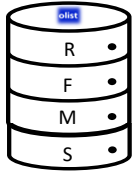
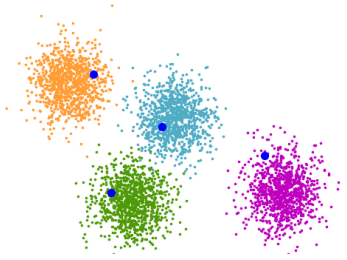


Modèle de segmentation



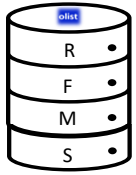
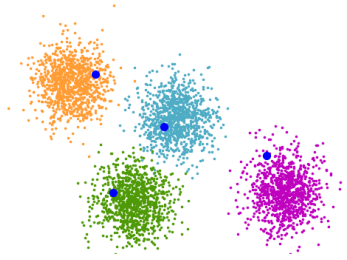


Modèle de segmentation

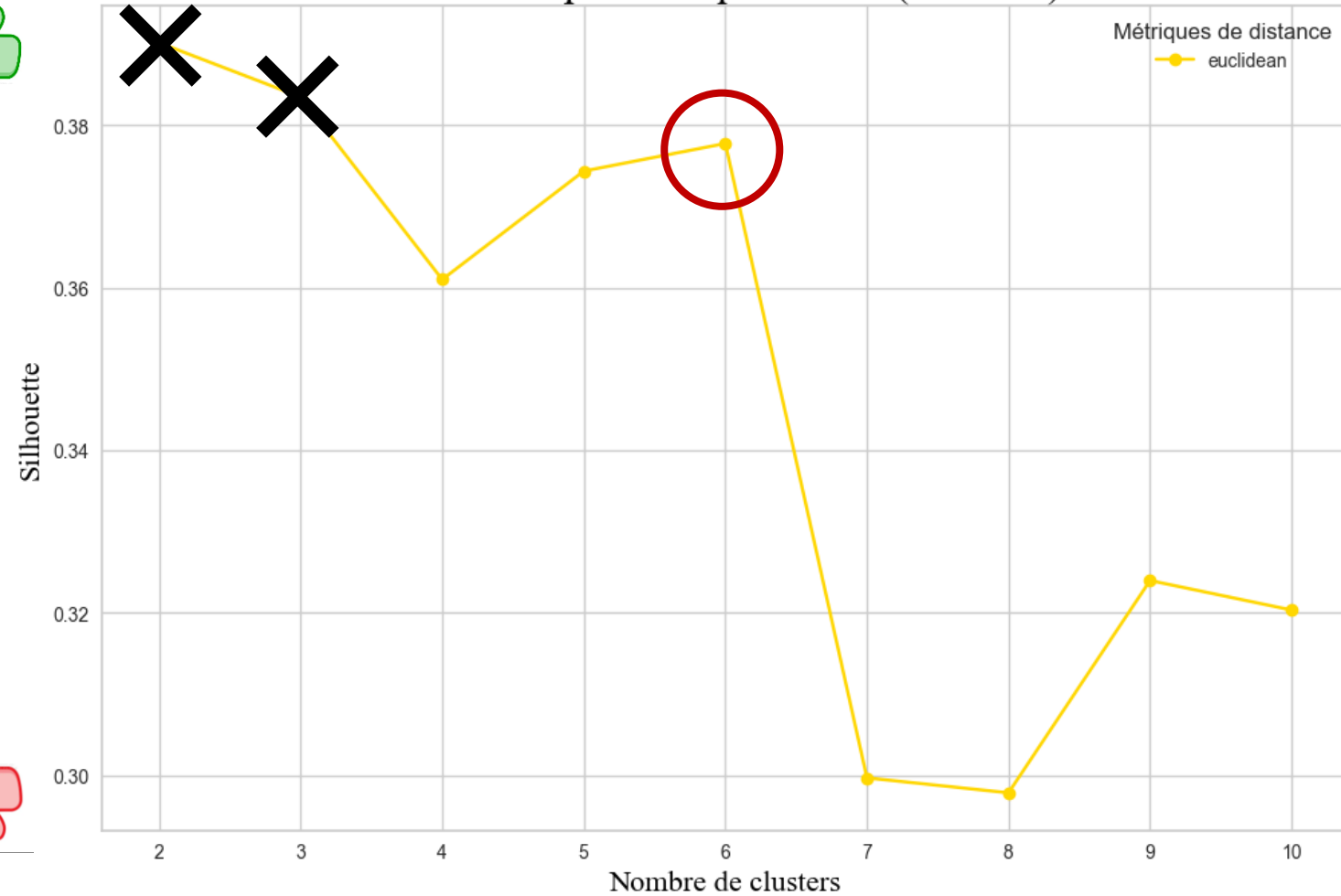




Modèle de segmentation

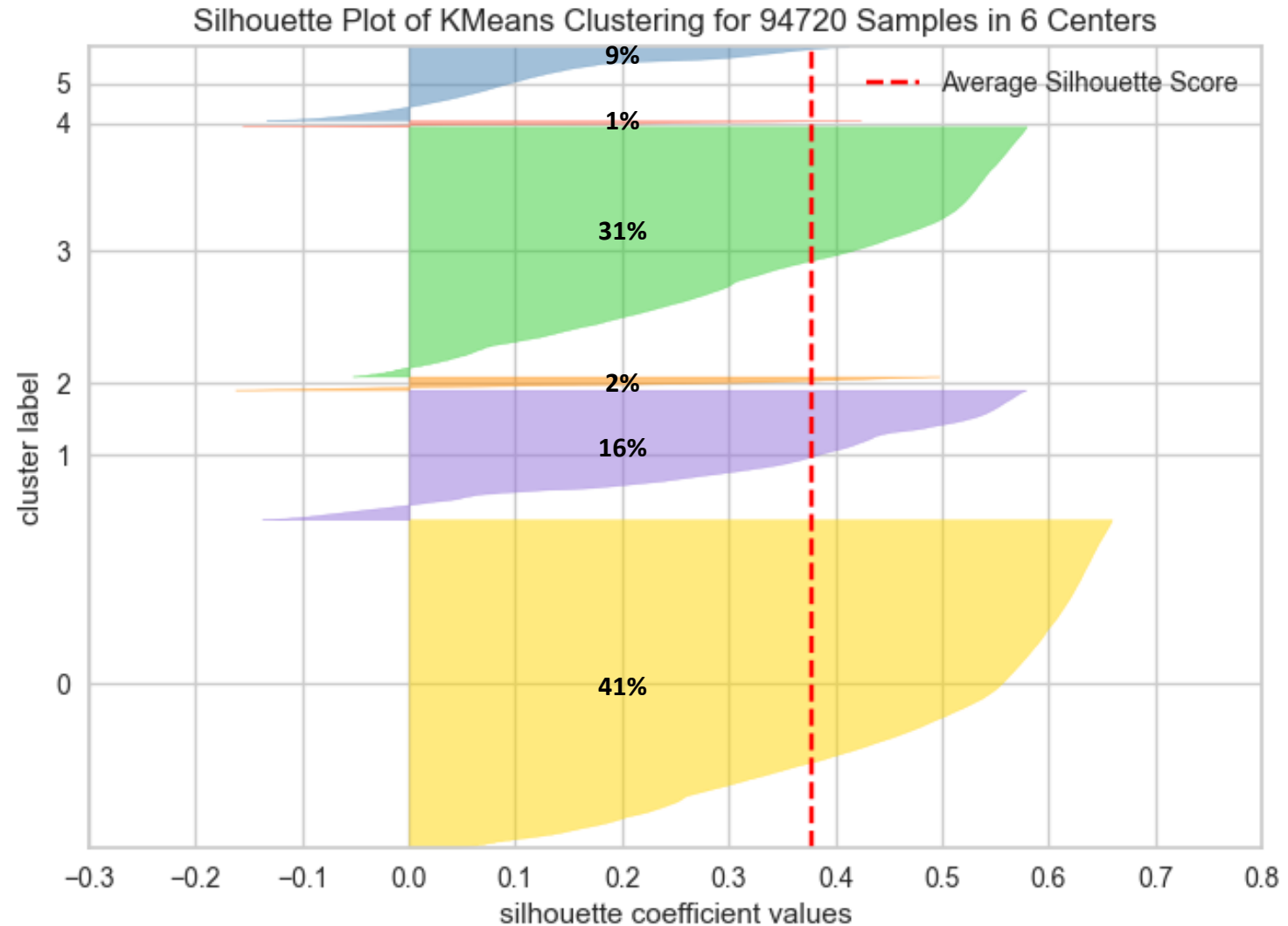
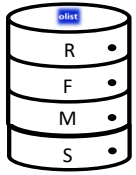
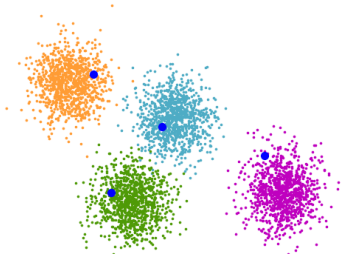


Silhouette pour chaque cluster (KMeans)



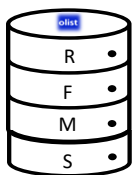
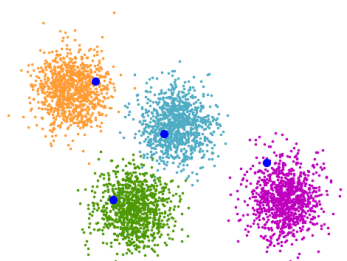


Modèle de segmentation

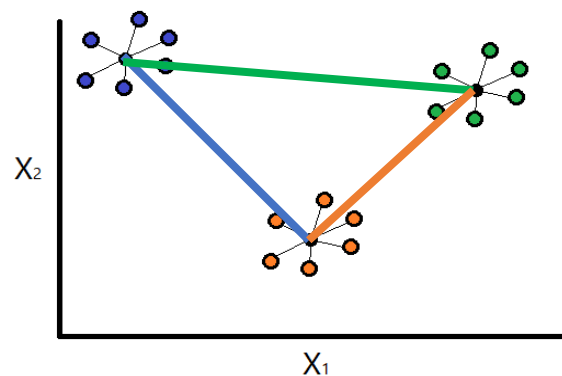




Modèle de segmentation



Indice Davies-Bouldin



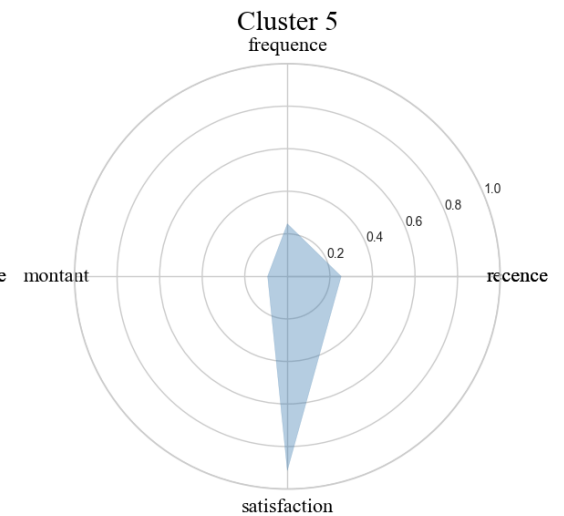
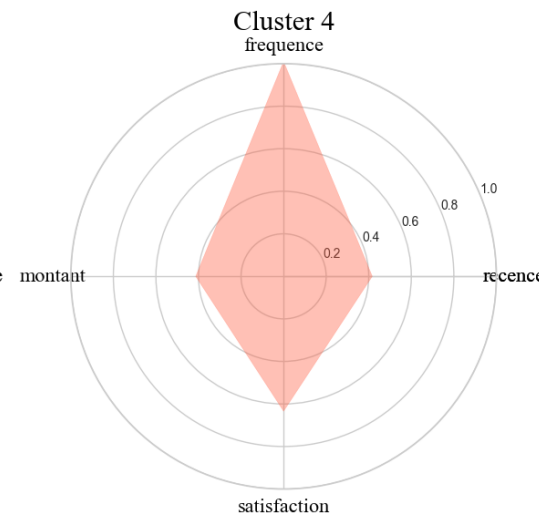
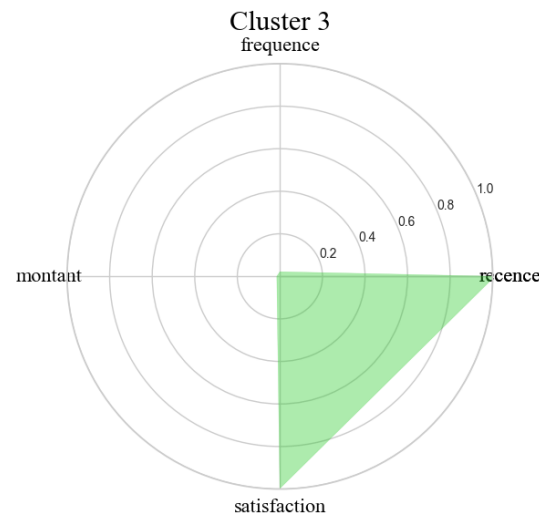
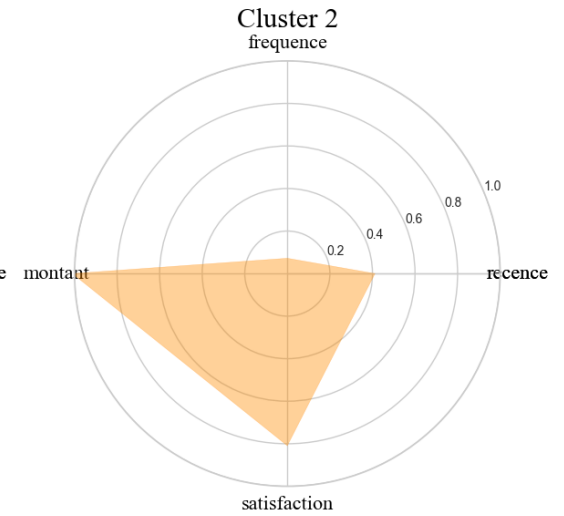
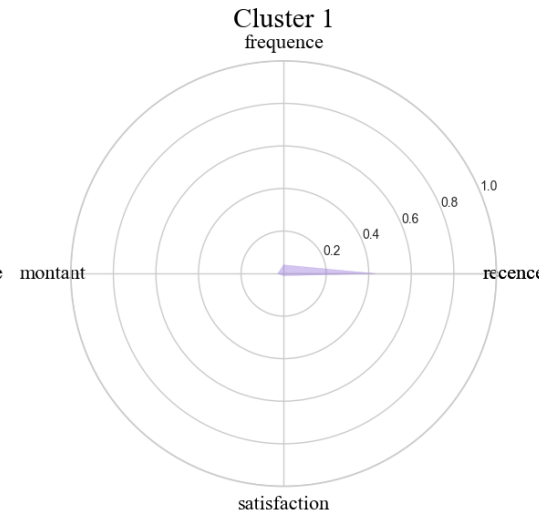
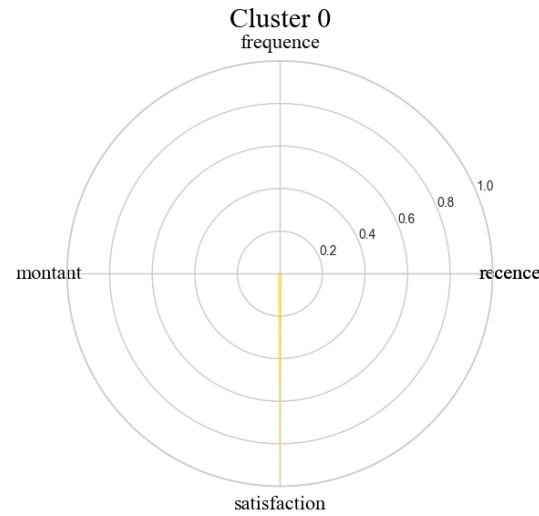
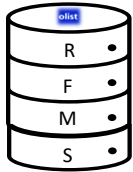
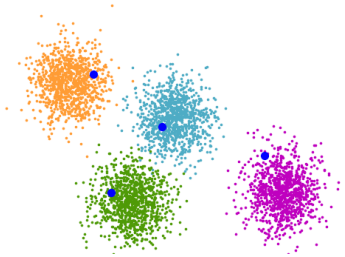
=

0,97





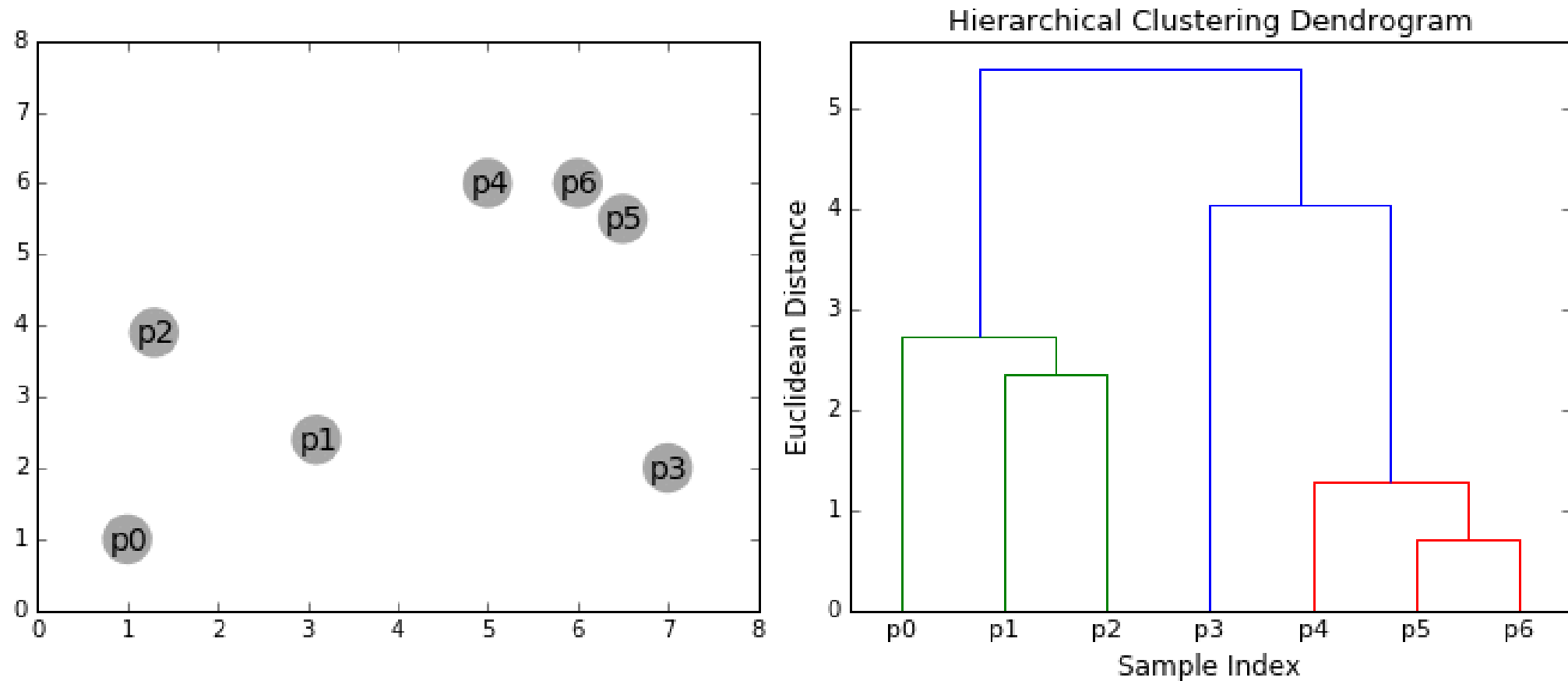
Modèle de segmentation





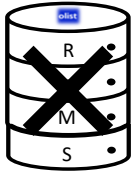
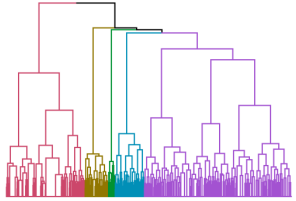
Modèle de segmentation

Clustering Ascendant Hiérarchique (CAH)



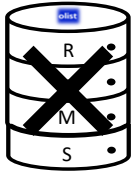
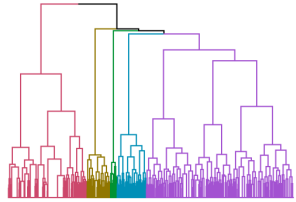


Modèle de segmentation

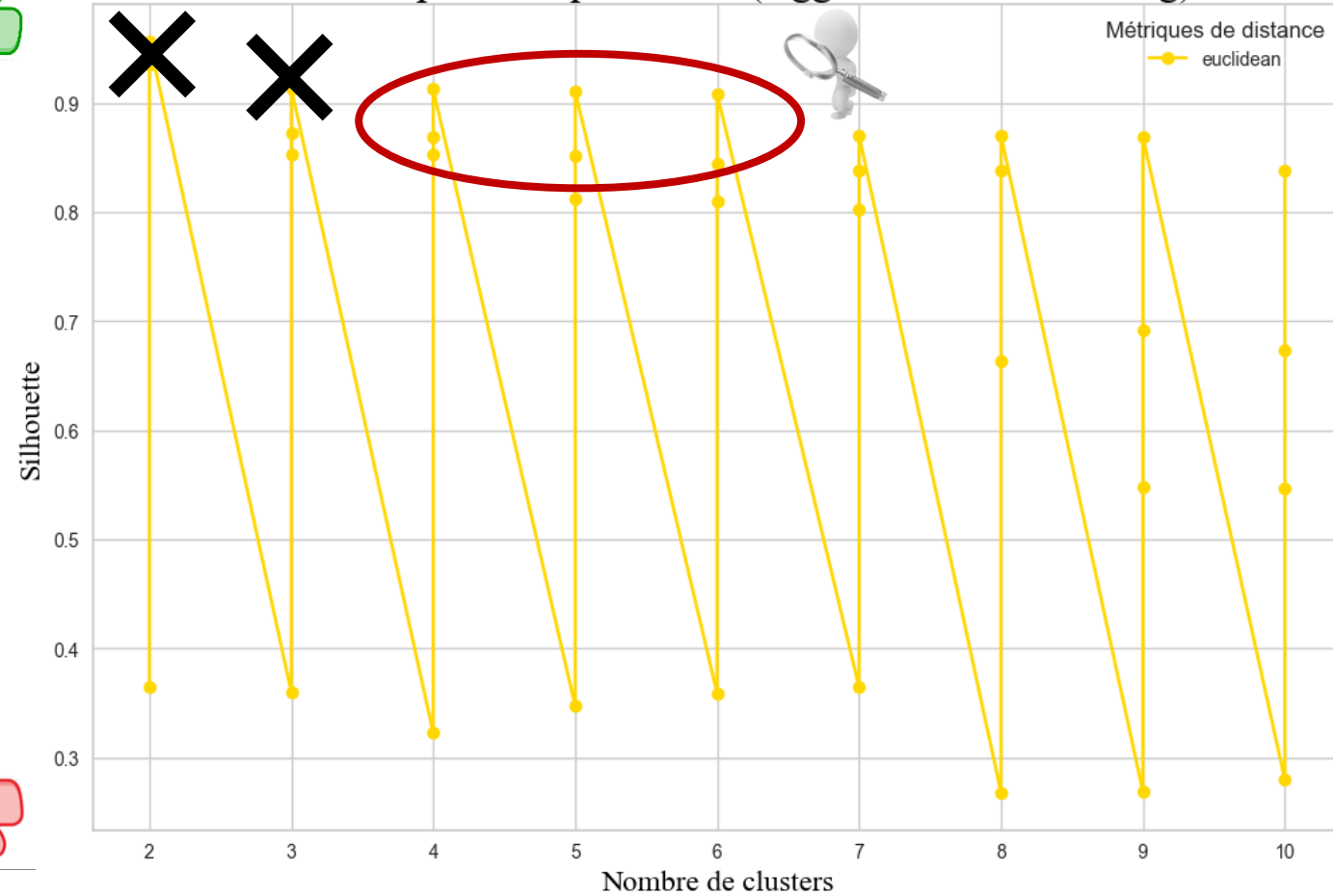




Modèle de segmentation

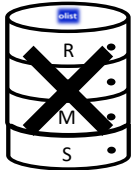
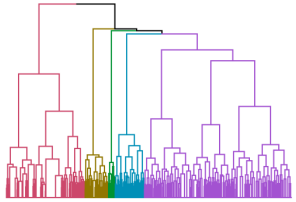


Silhouette pour chaque cluster (AgglomerativeClustering)





Modèle de segmentation



4 clusters

Cluster	
1	12
2	18921
3	10
4	1



5 clusters

Cluster	
1	12
2	18921
3	2
4	8
5	1



6 clusters

Cluster	
1	9
2	3
3	18921
4	2
5	8
6	1



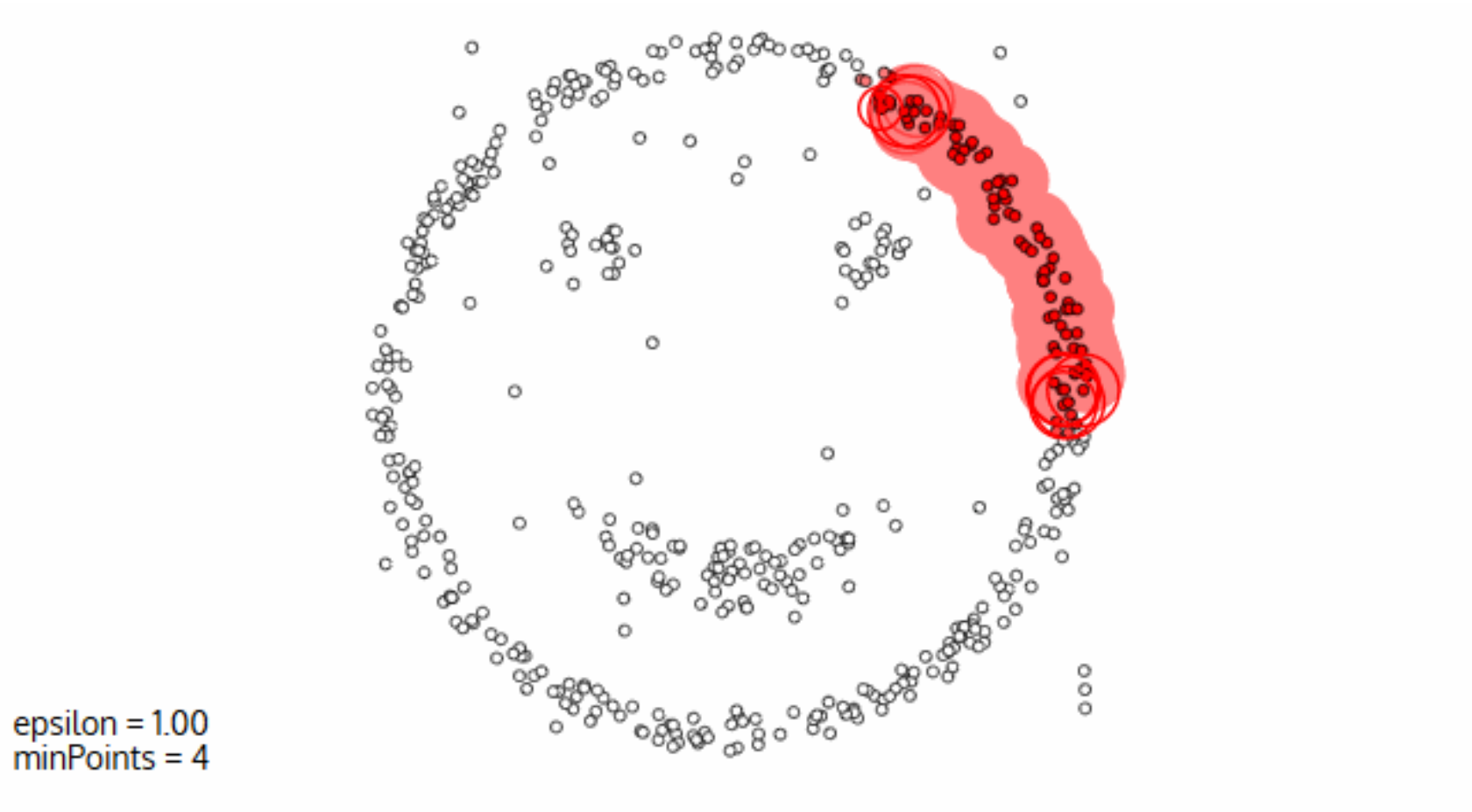
1 cluster > 99% des clients





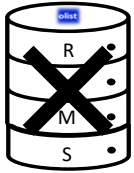
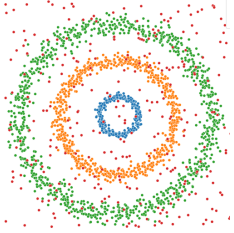
Modèle de segmentation

DBSCAN



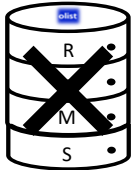
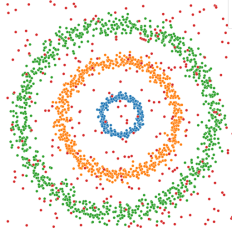


Modèle de segmentation

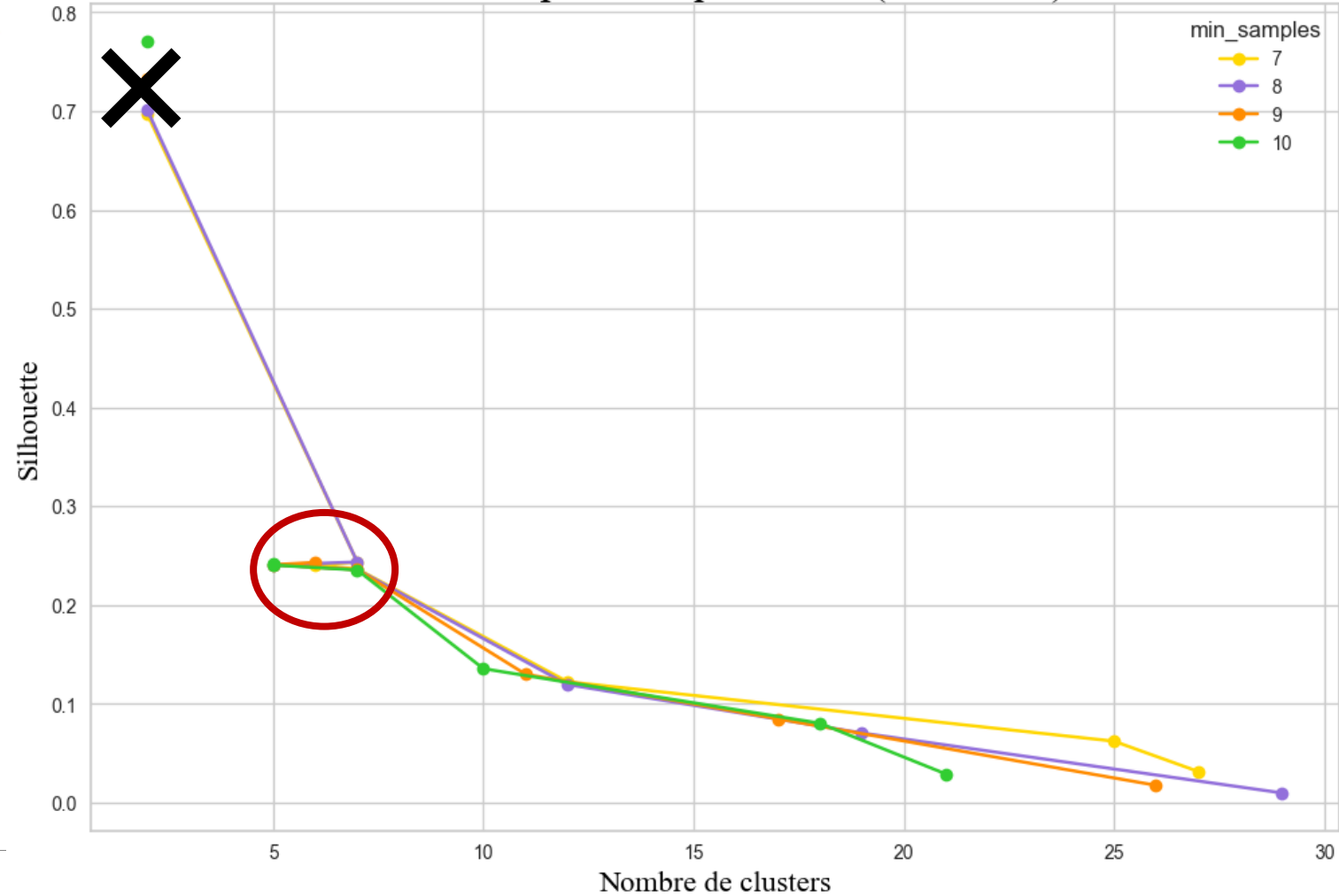




Modèle de segmentation

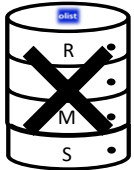
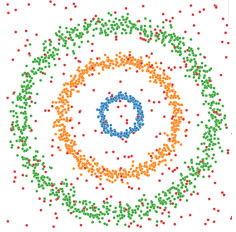


Silhouette pour chaque cluster (DBSCAN)





Modèle de segmentation



6 clusters

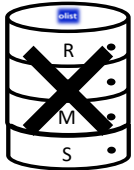
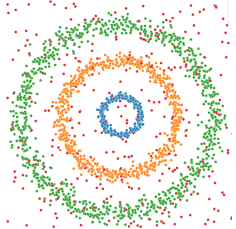
Cluster	
-1	171
0	16119
1	2101
2	355
3	151
4	38
5	9



1 cluster > 85% des clients



Modèle de segmentation



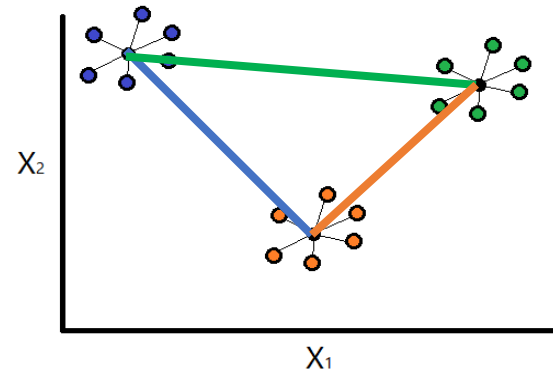
6 clusters


Cluster	
-1	171
0	16119
1	2101
2	355
3	151
4	38
5	9



1 cluster > 85% des clients + qualité de clustering moyen

Indice Davies-Bouldin

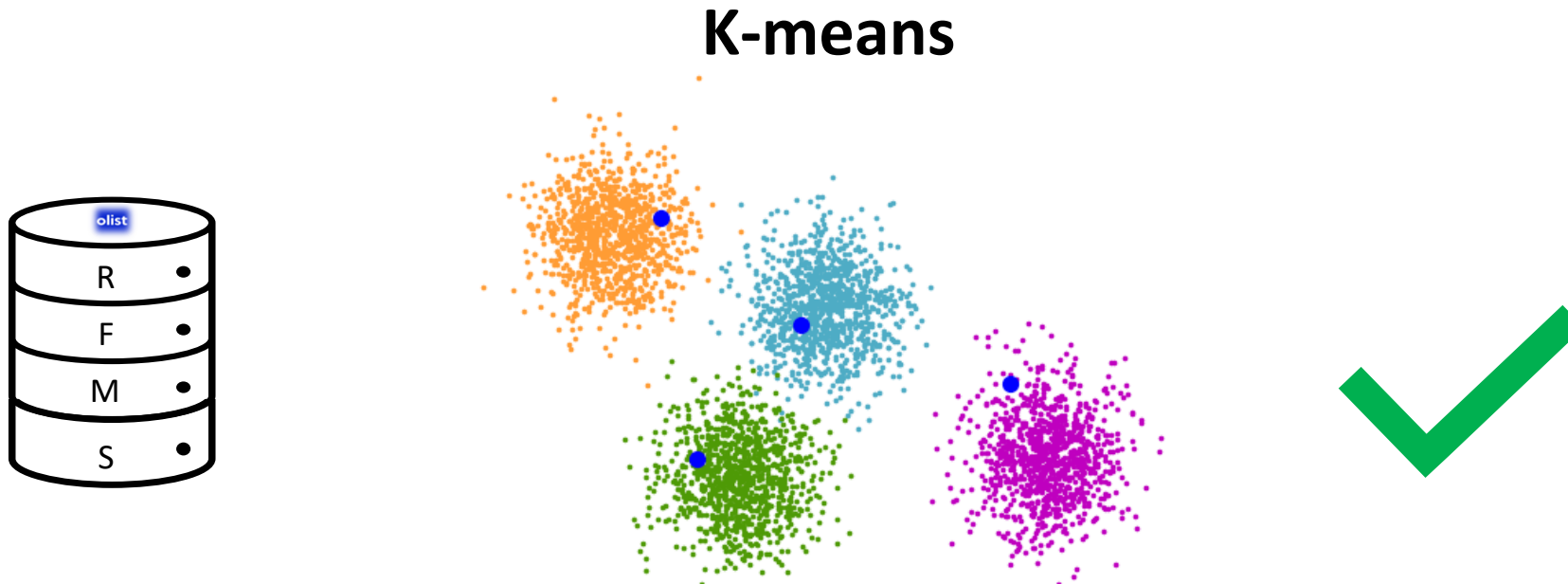


= 1,90 



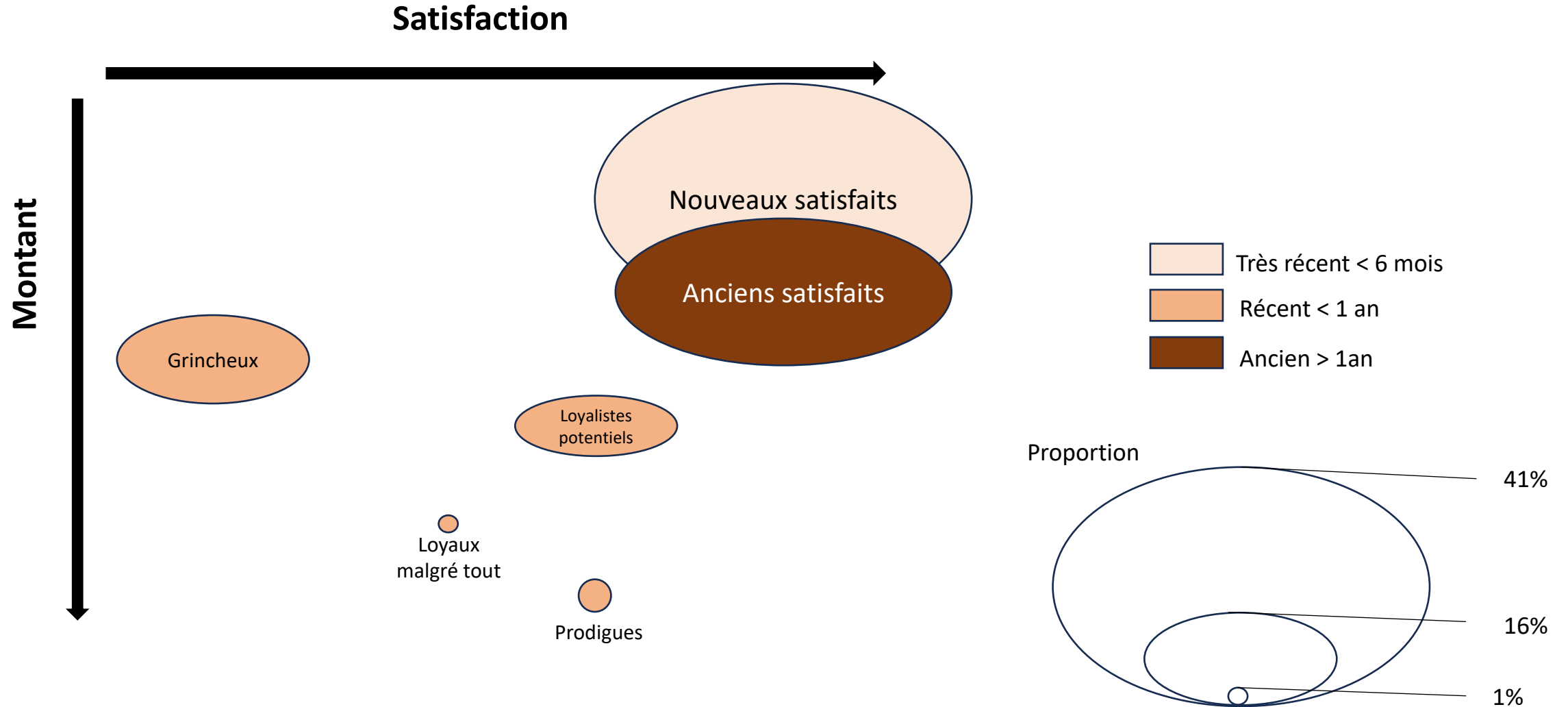


Modèle de segmentation





Modèle de segmentation





I – Problématique

II – Présentation du jeu de données

III - Nettoyage des données

IV – Feature engineering

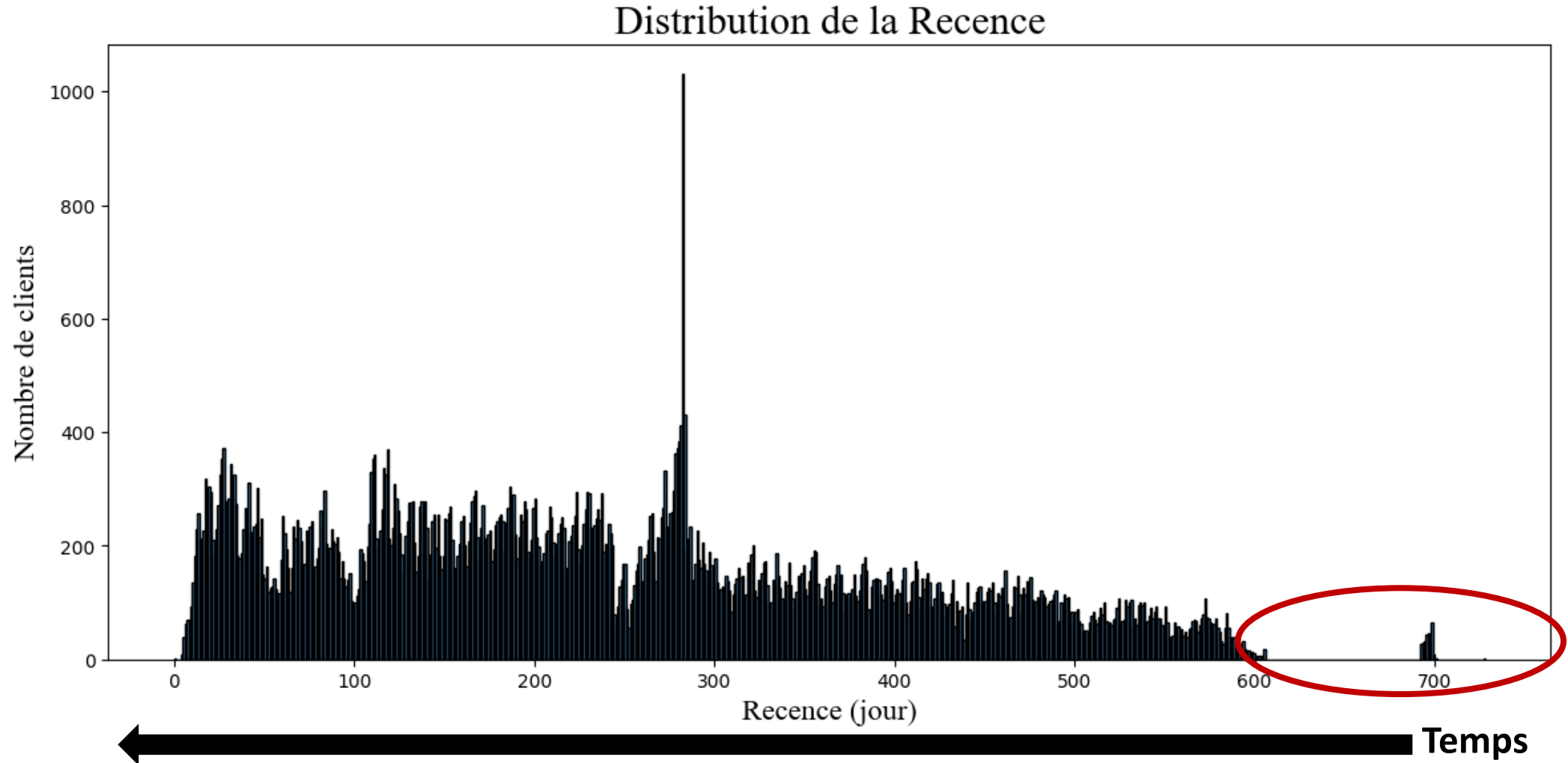
V – Analyses exploratoires

VI – Modèle de segmentation

VII – Simulation obsolescence segmentation

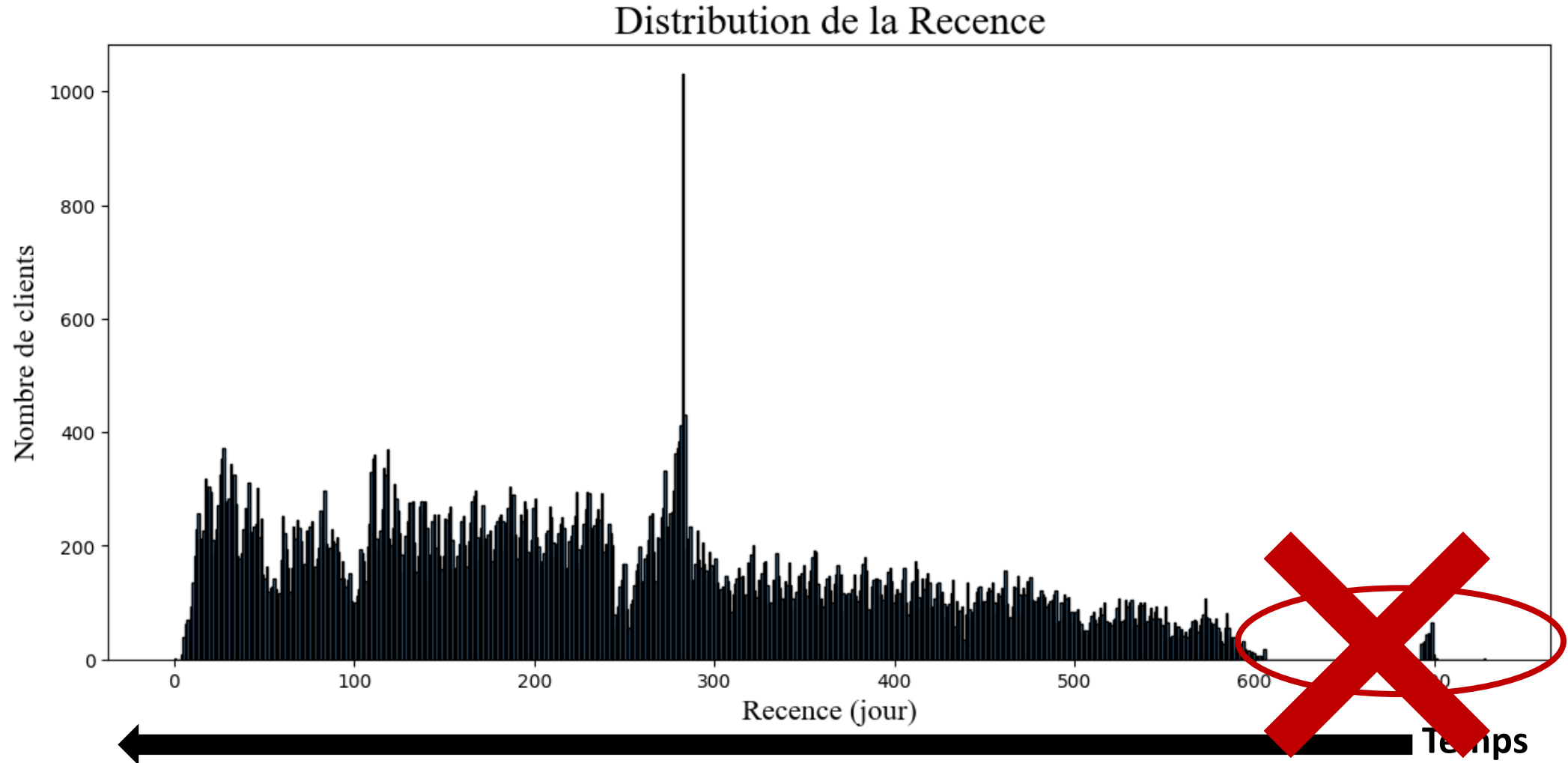


Simulation obsolescence segmentation



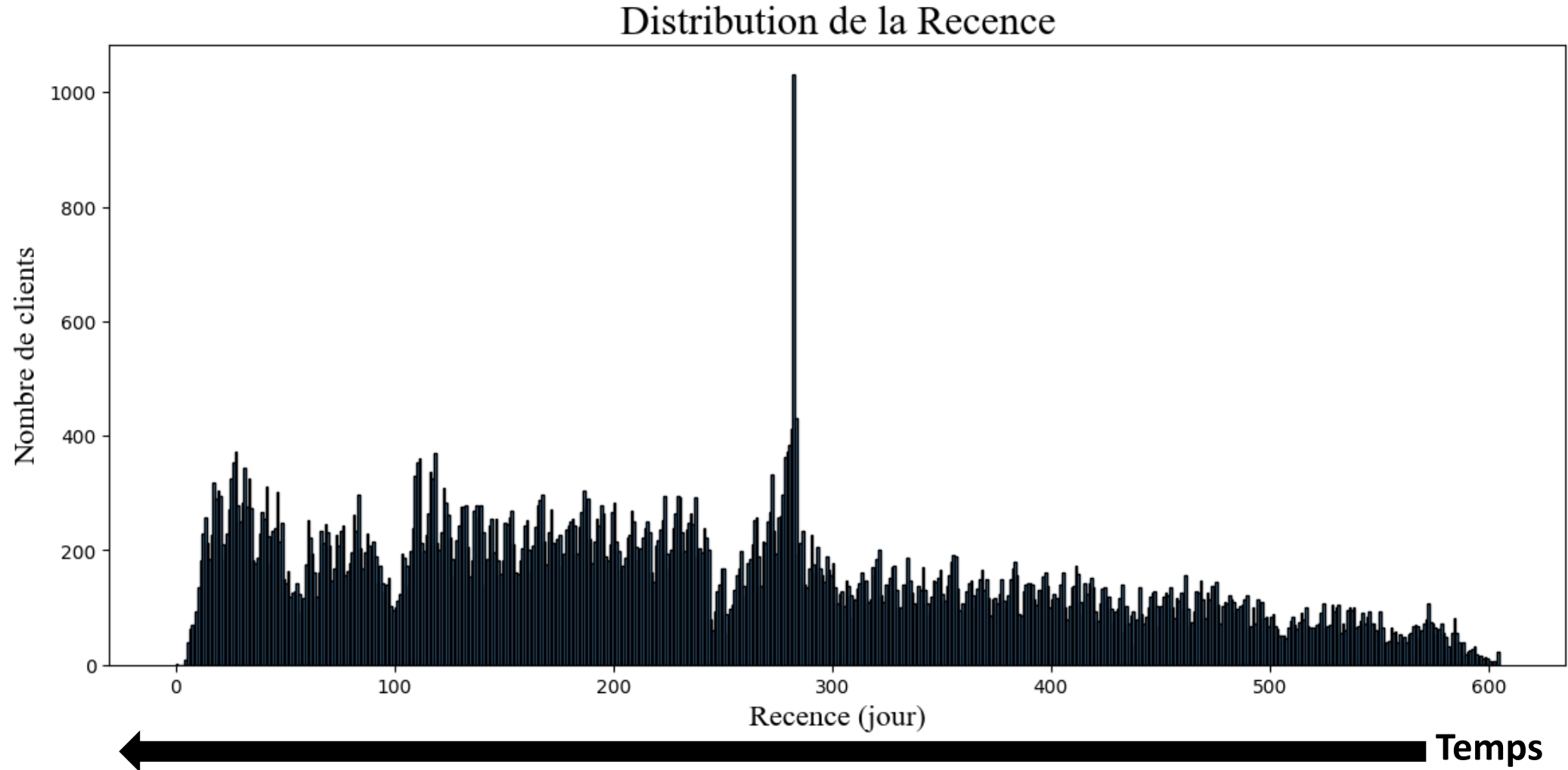


Simulation obsolescence segmentation



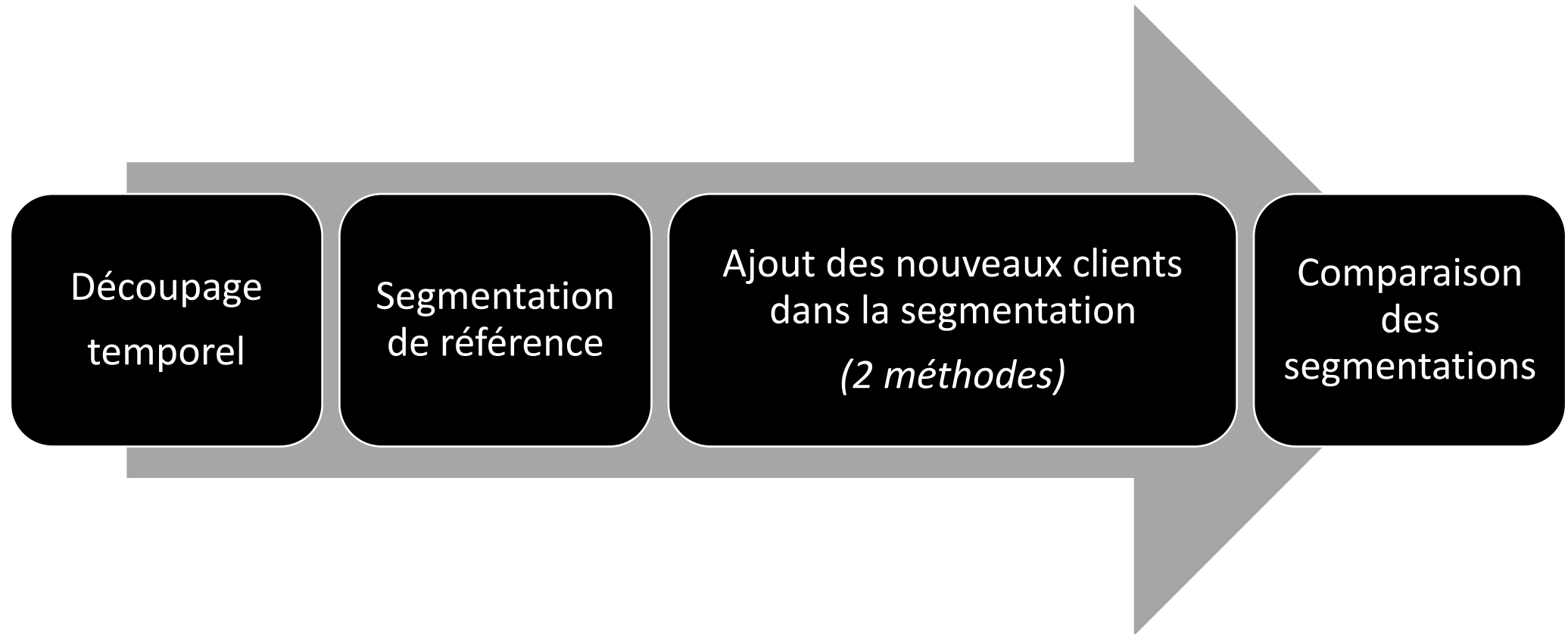


Simulation obsolescence segmentation





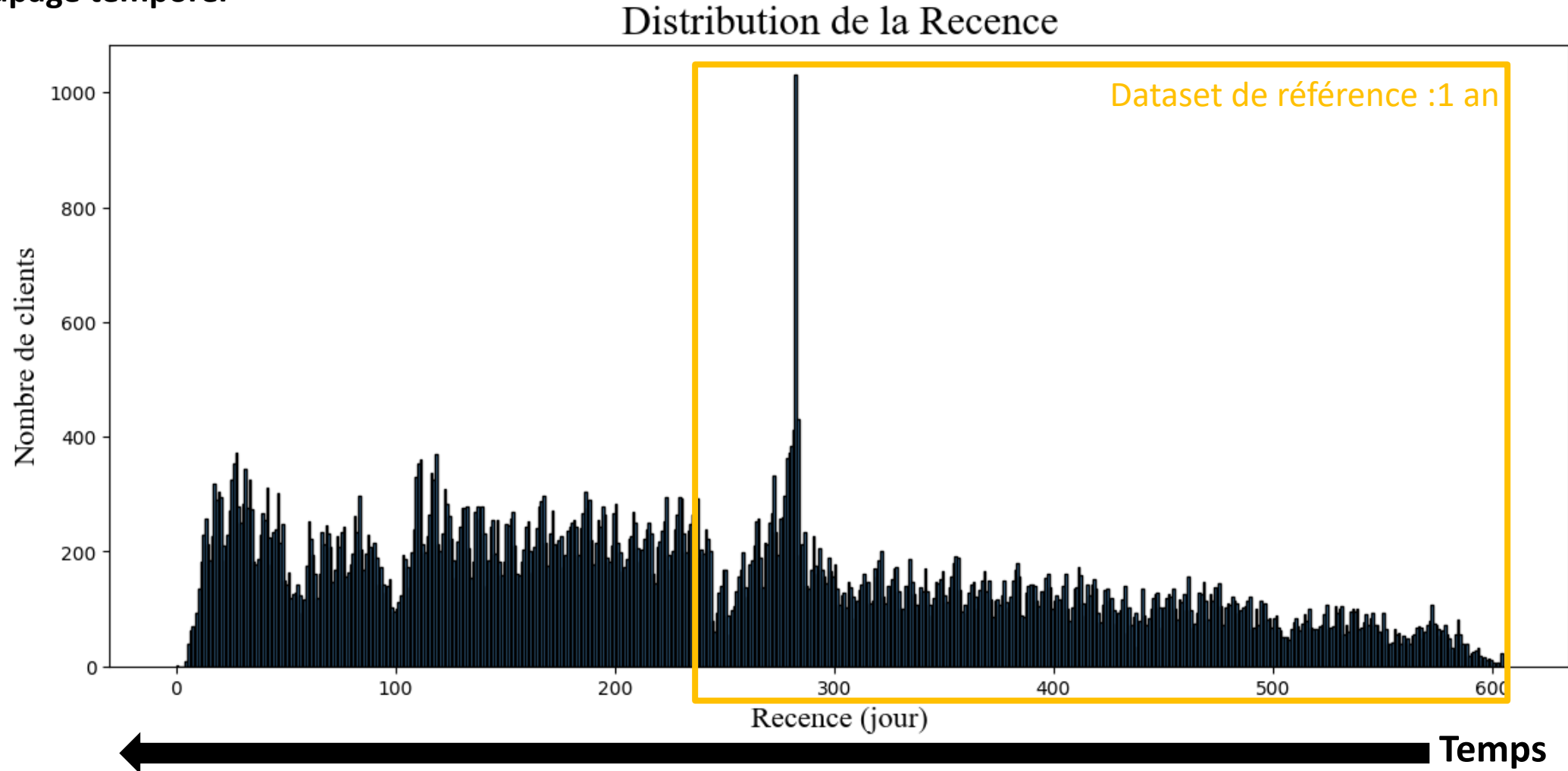
Simulation obsolescence segmentation





Simulation obsolescence segmentation

Découpage temporel



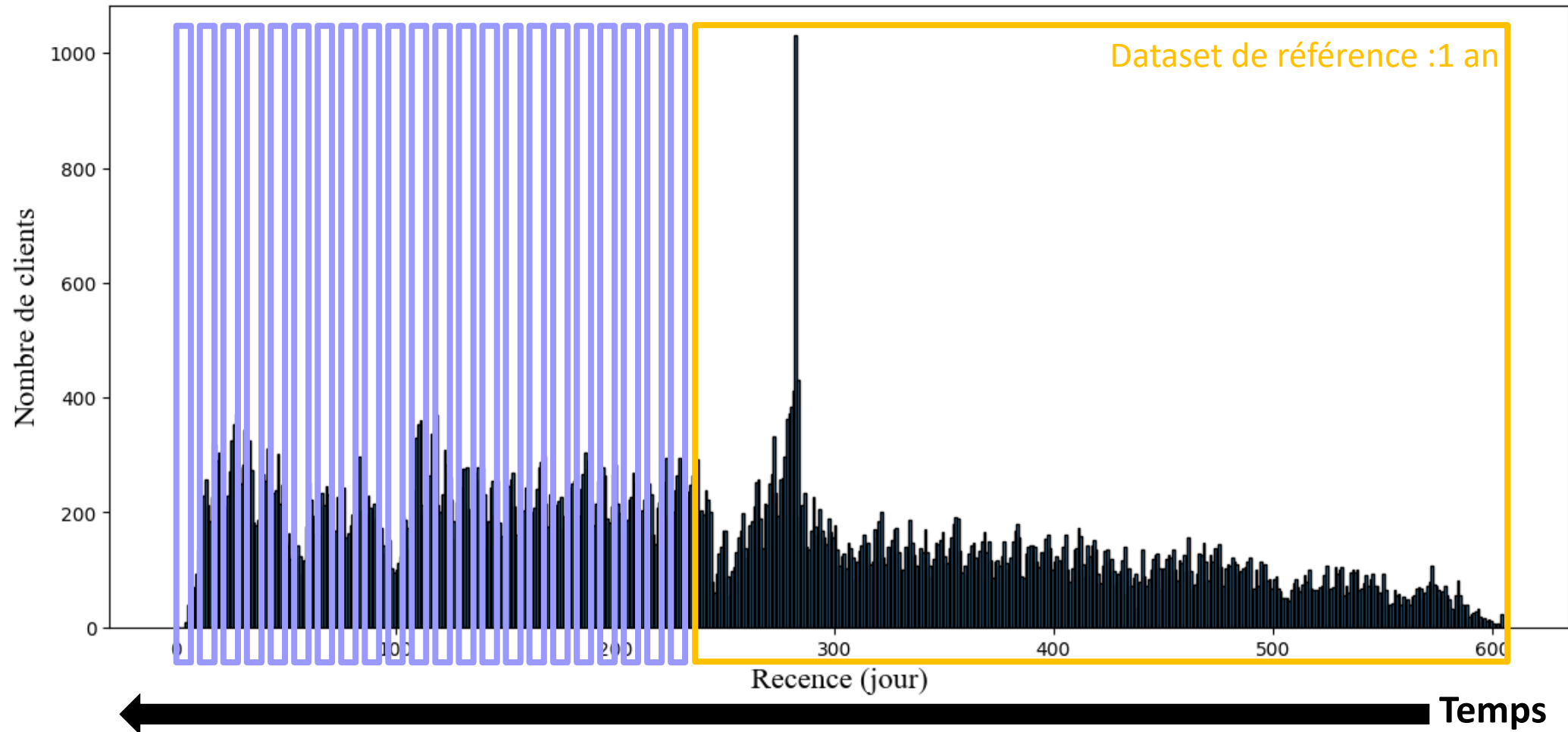


Simulation obsolescence segmentation

Découpage temporel

34 datasets de 7 jours

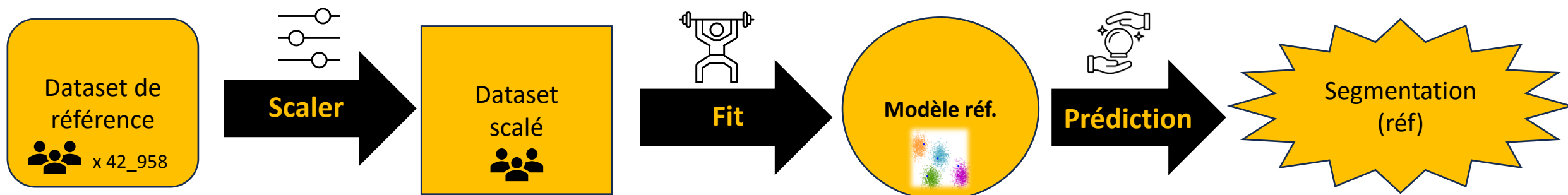
Distribution de la Recence





Simulation obsolescence segmentation

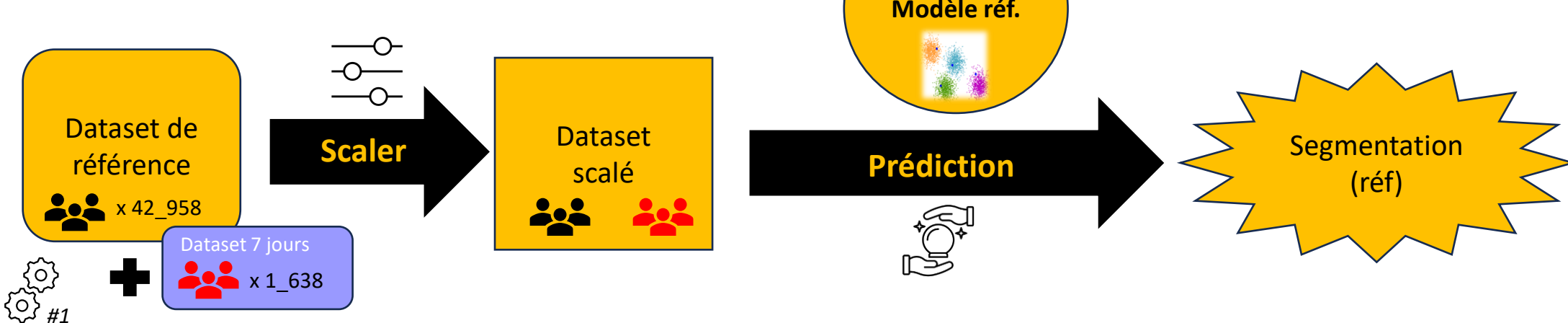
Segmentation de référence





Simulation obsolescence segmentation

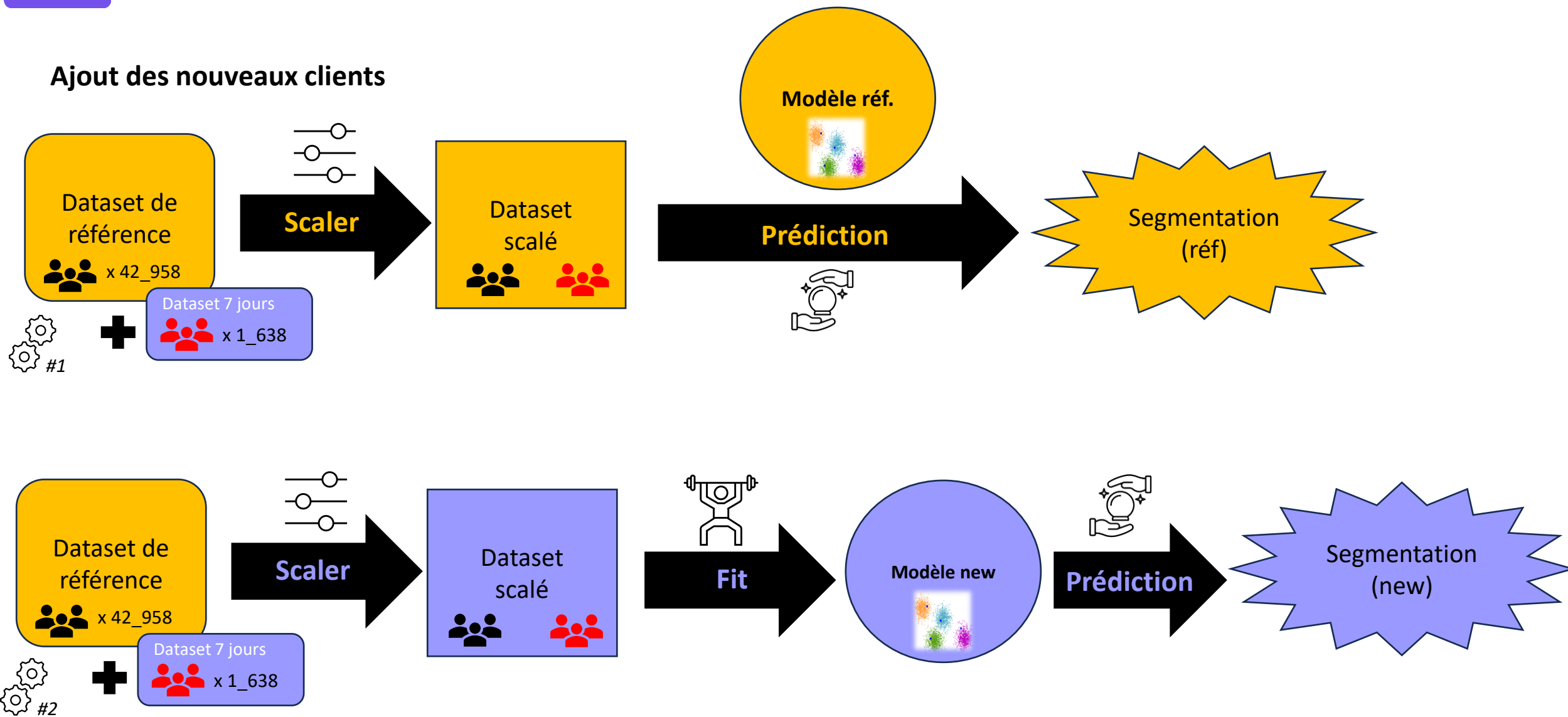
Ajout des nouveaux clients





Simulation obsolescence segmentation

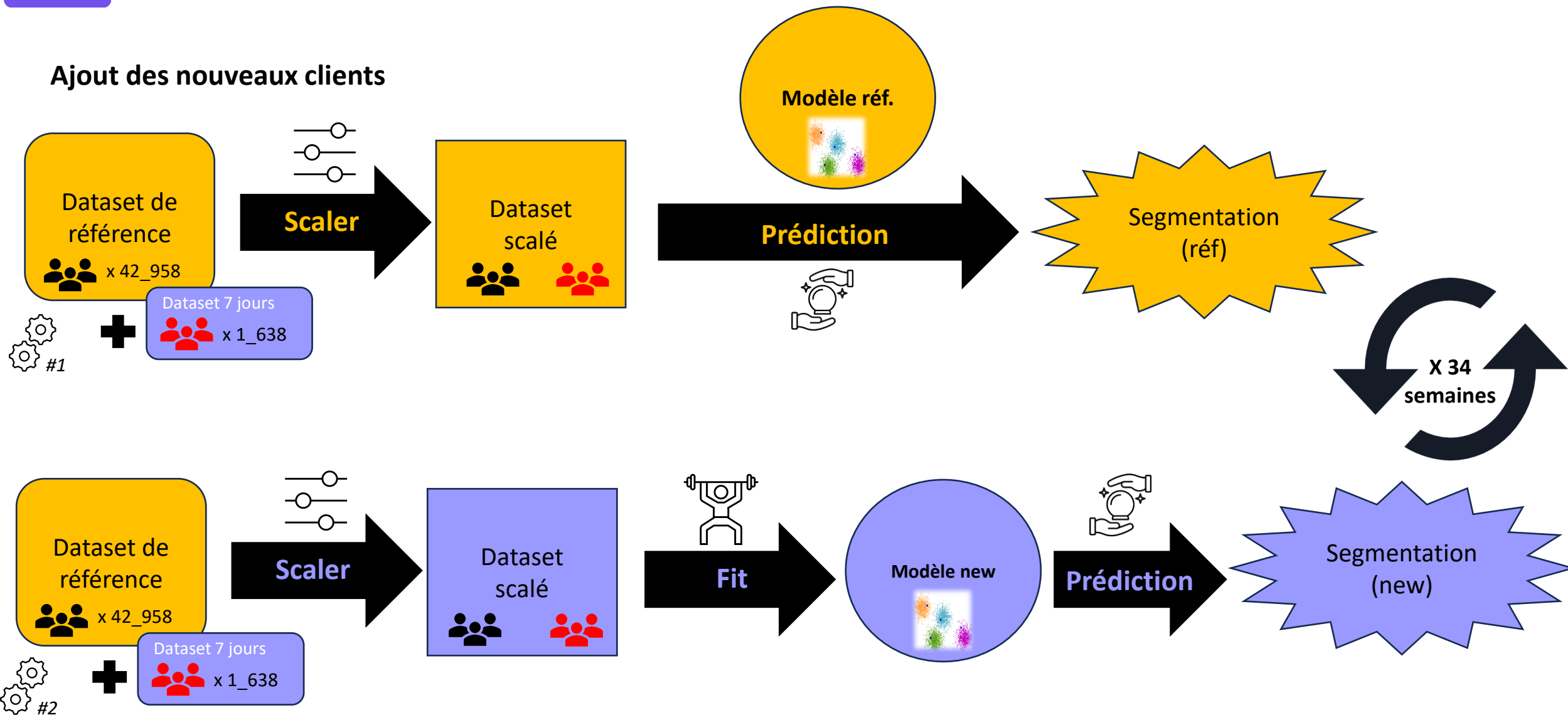
Ajout des nouveaux clients





Simulation obsolescence segmentation

Ajout des nouveaux clients



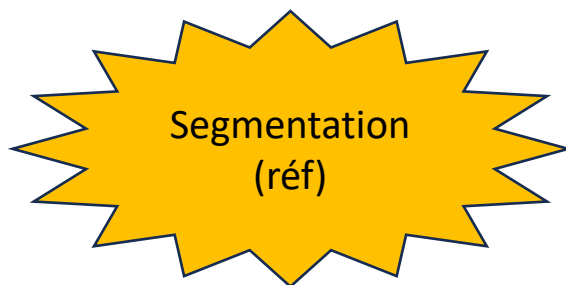


Simulation obsolescence segmentation

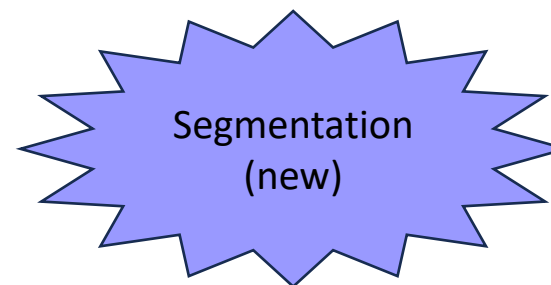
Comparaison des segmentations



ARI
(*Ajusted Rand Score*)

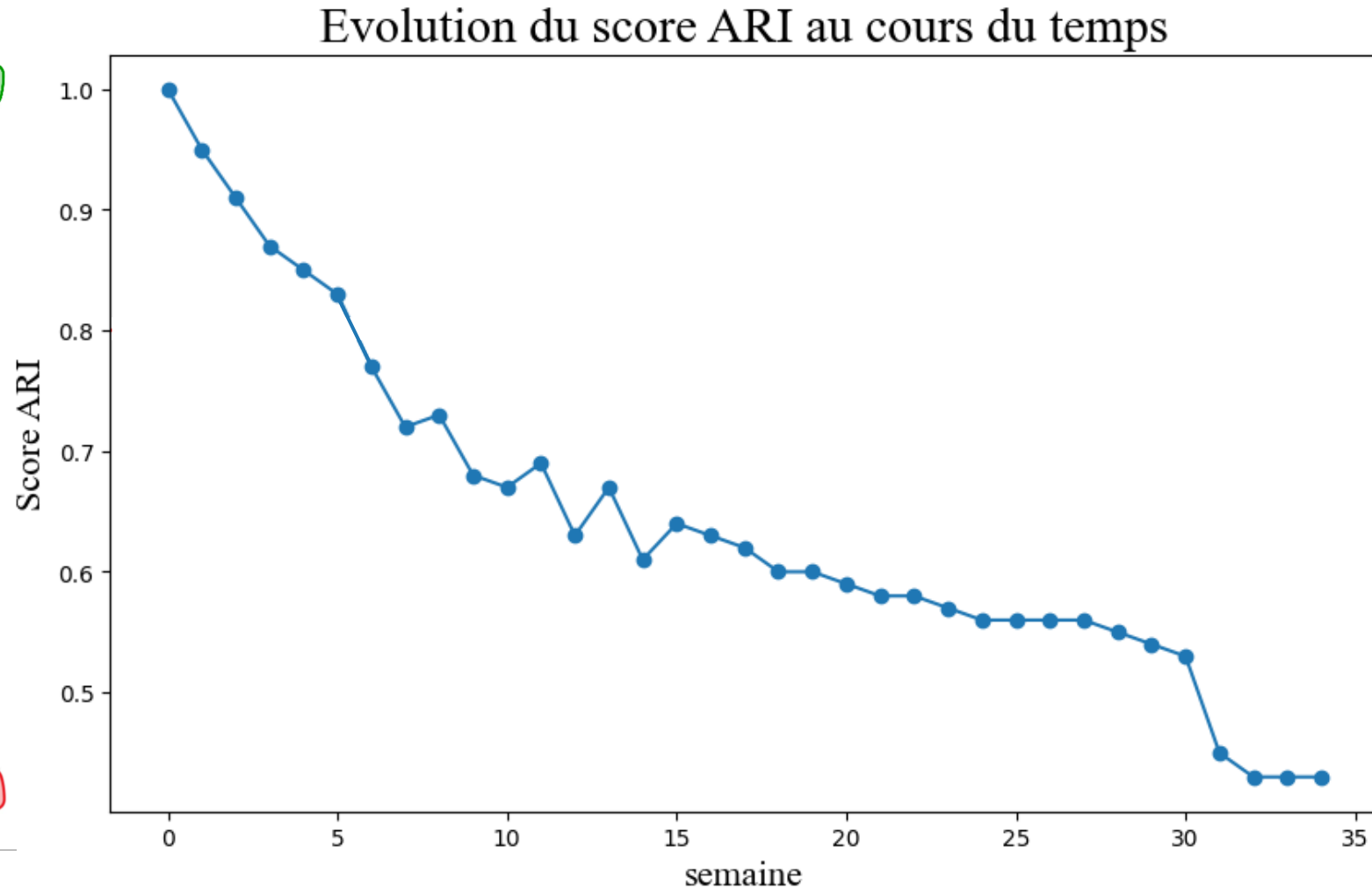


VS



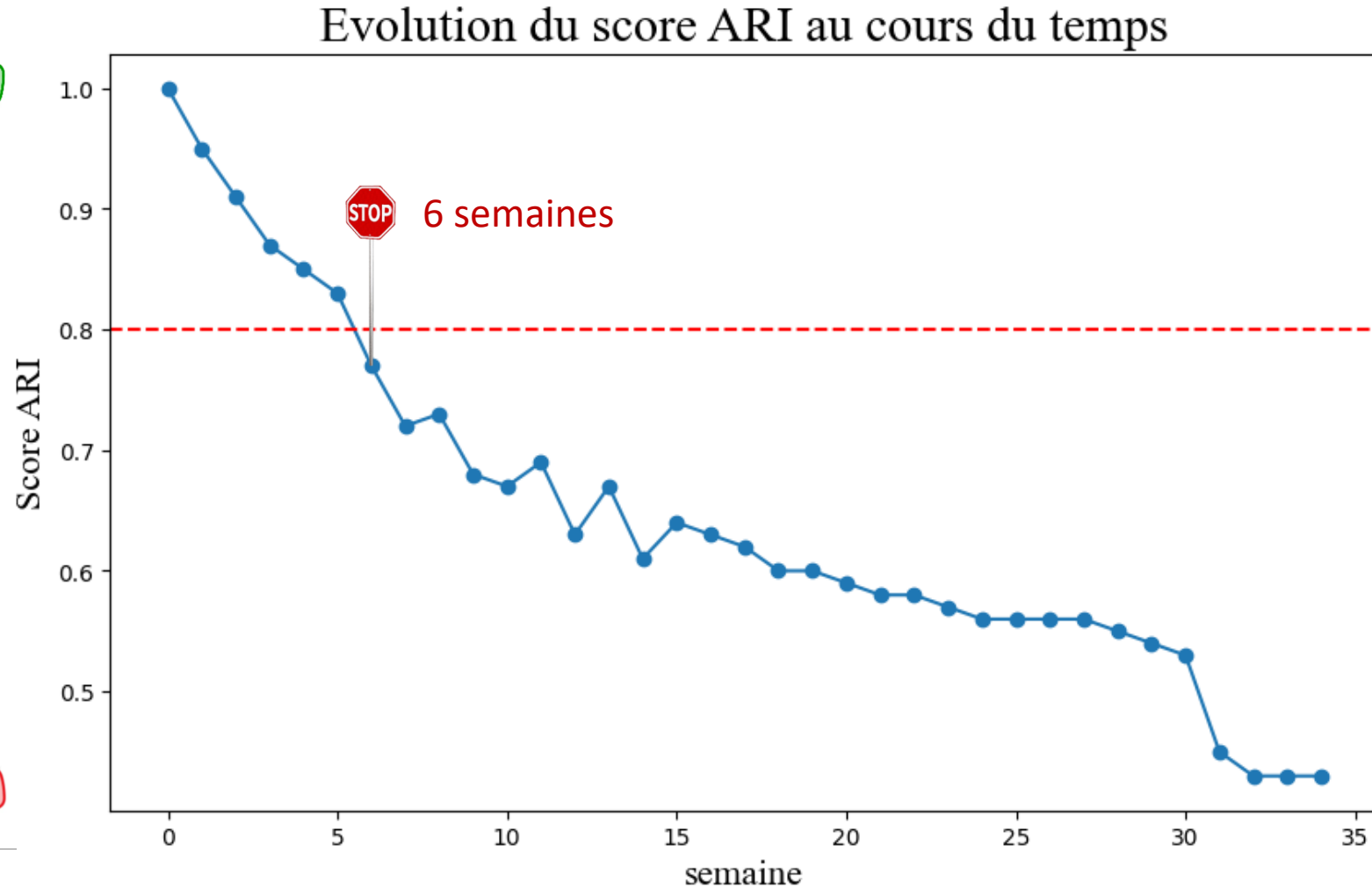


Simulation obsolescence segmentation





Simulation obsolescence segmentation





I – Problématique

II – Présentation du jeu de données

III - Nettoyage des données

IV – Feature engineering

V – Analyses exploratoires

VI – Modèle de segmentation

VII – Simulation obsolescence segmentation

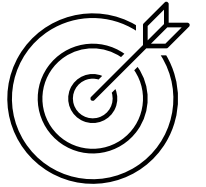
VIII - Conclusion



Conclusion

Missions :

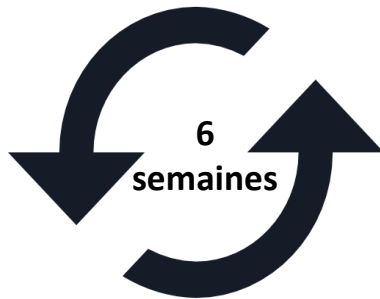
1. Réaliser une courte analyse exploratoire. ✓
2. Tester différents modèles de segmentation pour trouver la meilleure segmentation (>RFM). ✓
3. Réaliser une analyse de la stabilité des segments au cours du temps. ✓





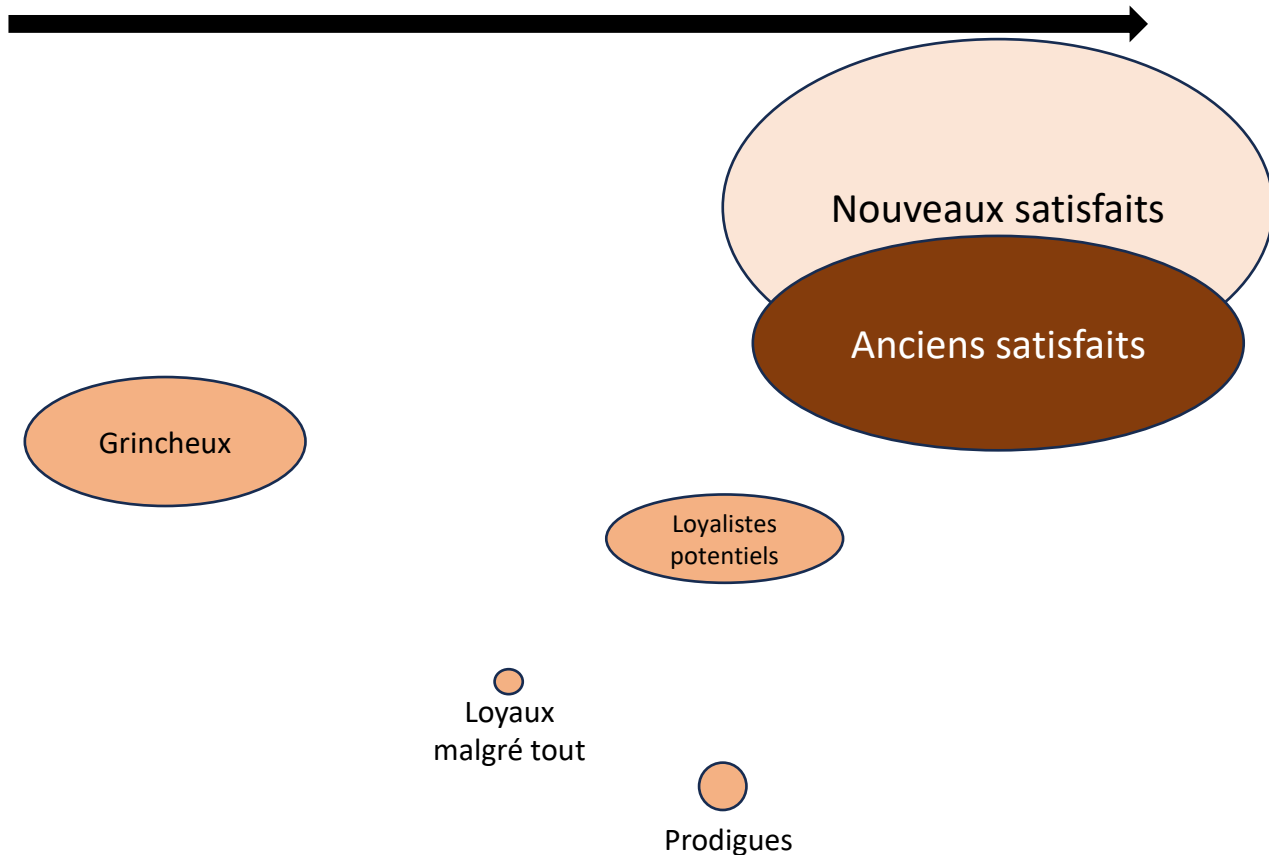
K-means

Conclusion



Satisfaction

Montant



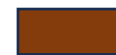
94_720 clients



Très récent < 6 mois

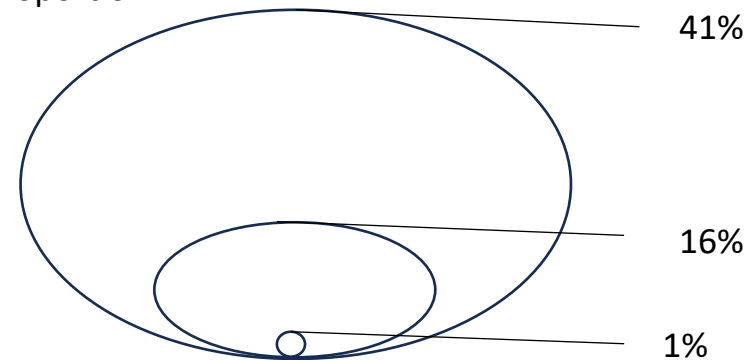


Récent < 1 an



Ancien > 1an

Proportion





Conclusion



Limites :

- Problèmes de puissance de calcul pour tester CAH et DBSCAN
- D'autres feature exploitables (date de livraison estimée et réelle, localisation, ...)

OPENCLASSROOMS

Merci pour votre attention



CentraleSupélec

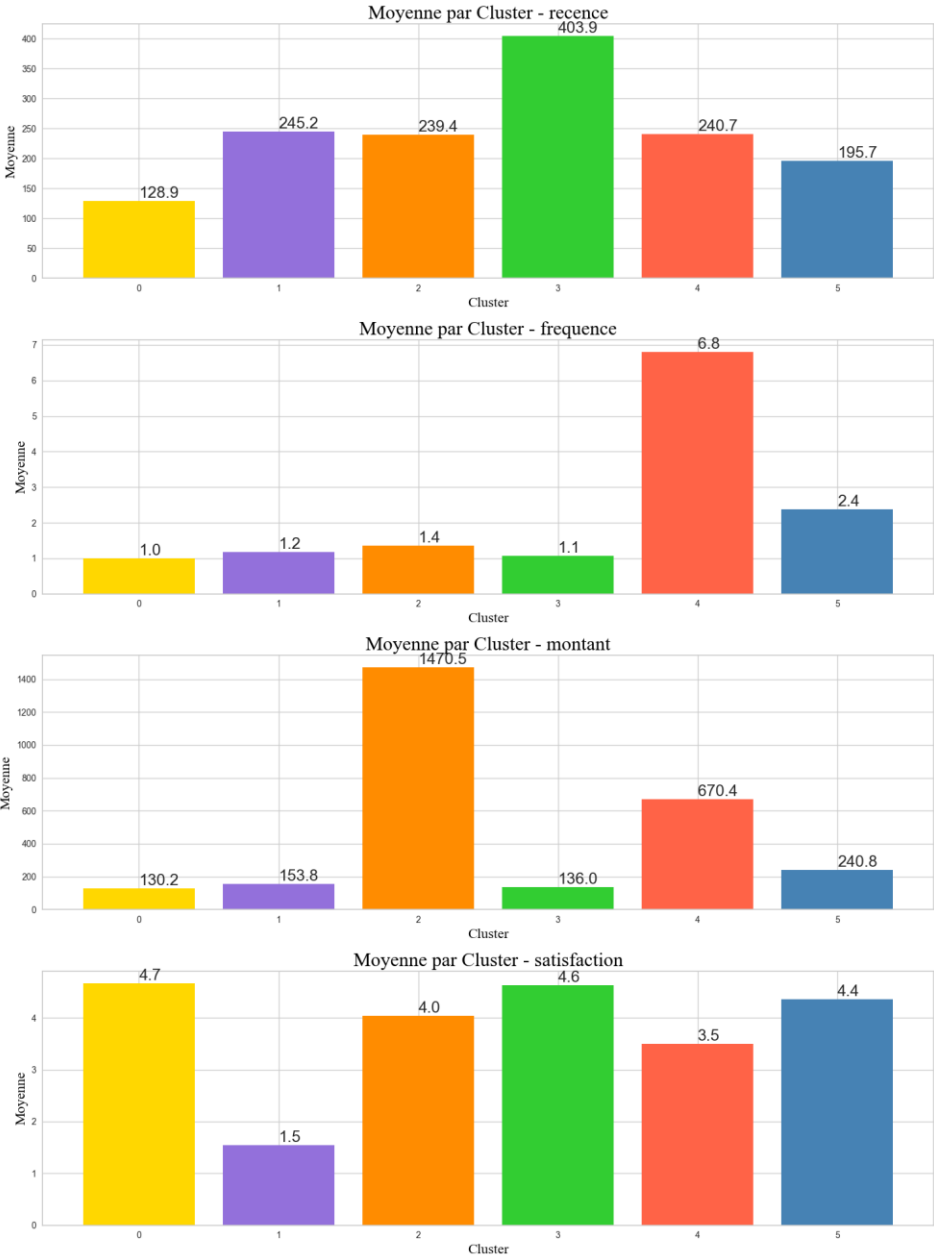
Pierrick BERTHE

Formation Expert en Data Science
Openclassrooms – CentraleSupélec

août 2023 → avril 2024



Extra
RFMS





Extra
RFMS

Plot 3D - K-means - Nombre de clusters: 6

