olist

Projet 5
Segmentation des Clients

Eva Bookjans

Objective

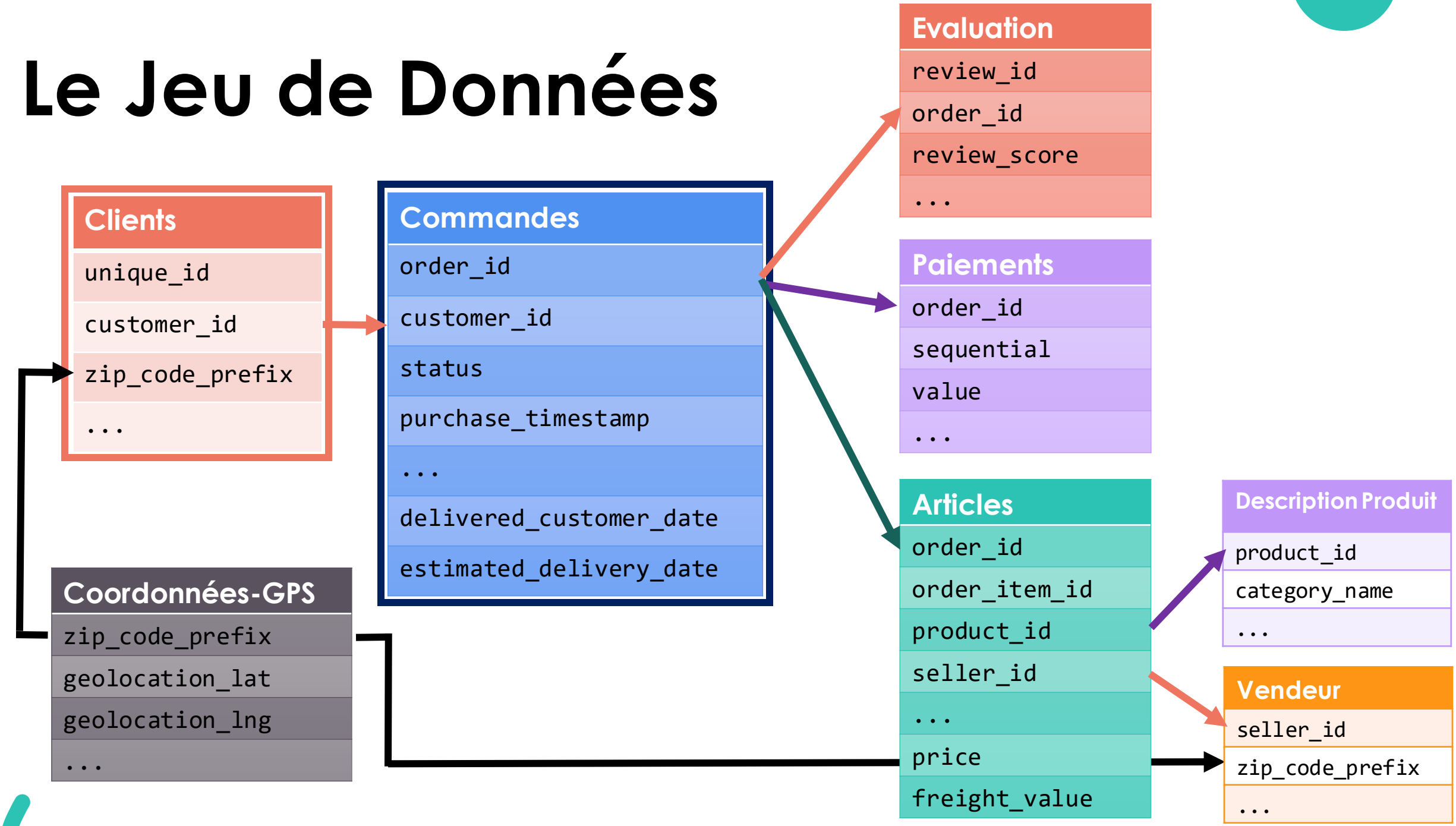
Segmentation des clients du e-commerce

olist

- Identifier les **différents types d'utilisateurs**
- Fournier une **description actionnable** de la segmentation et sa logique sous-jacente
- Proposer un **contrat de maintenance** basée sur une analyse de la stabilité des segments au cours du temps

➢ **Exploitable et facile d'utilisation** par l'équipe marketing

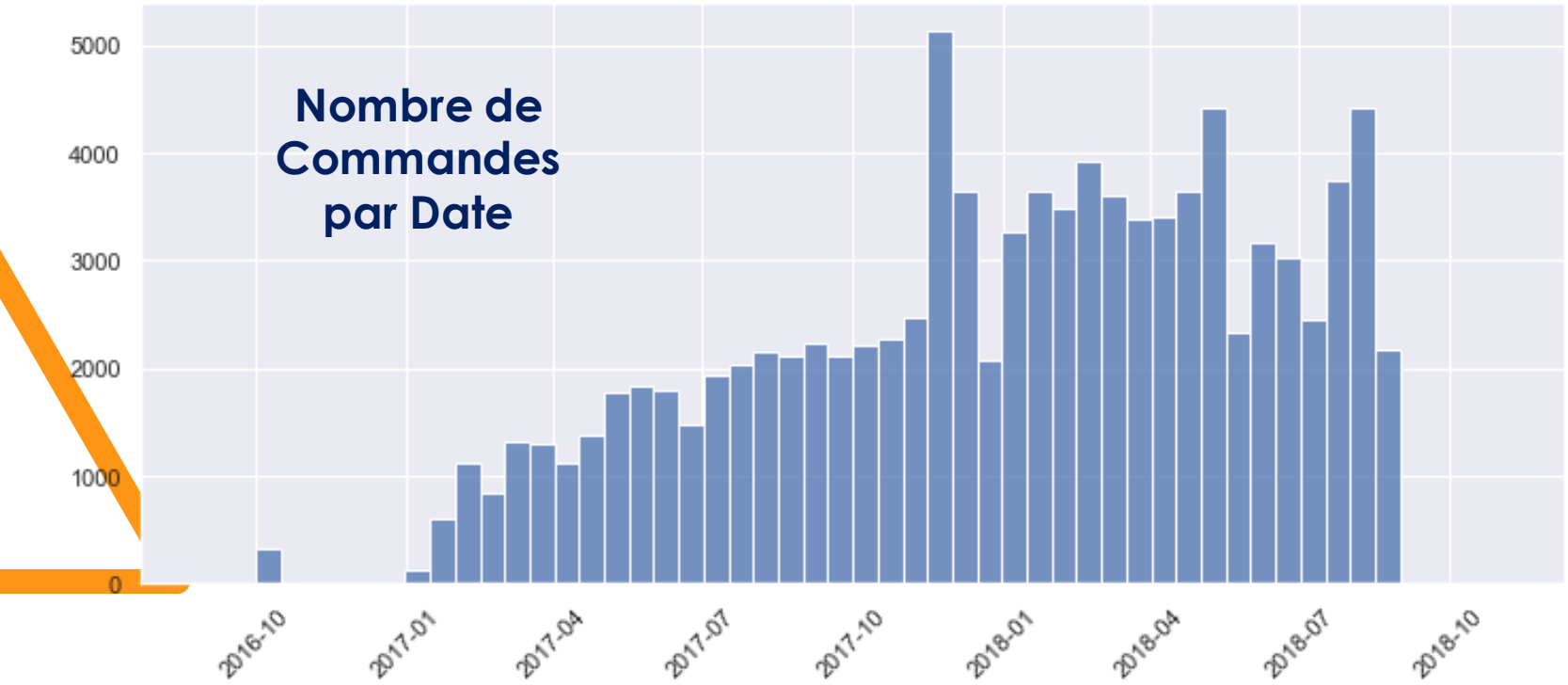
Le Jeu de Données

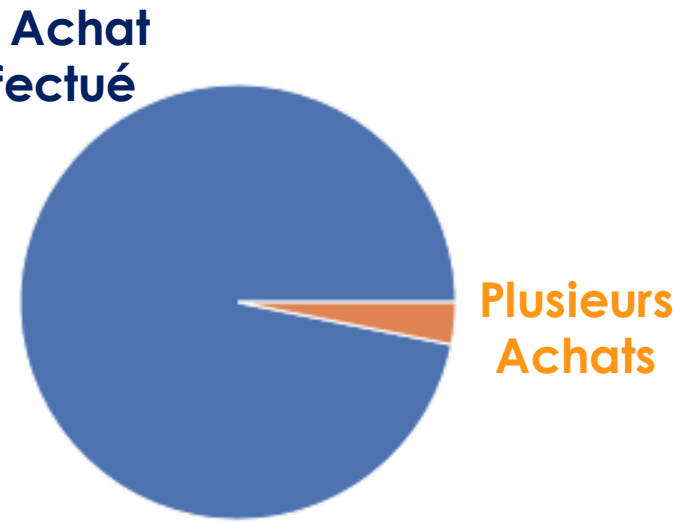


Le Jeu de Données

Dates: 4 Sep. 2016 à 17 Oct. 2018
Nombre de Commandes: 99441
Nombre de Clients: 96089

- **seul ~ 3% de clients réguliers**
  - ➢ profil basé sur un seul achat
- **grand ensemble**
  - ➢ des modèles utilisables limités



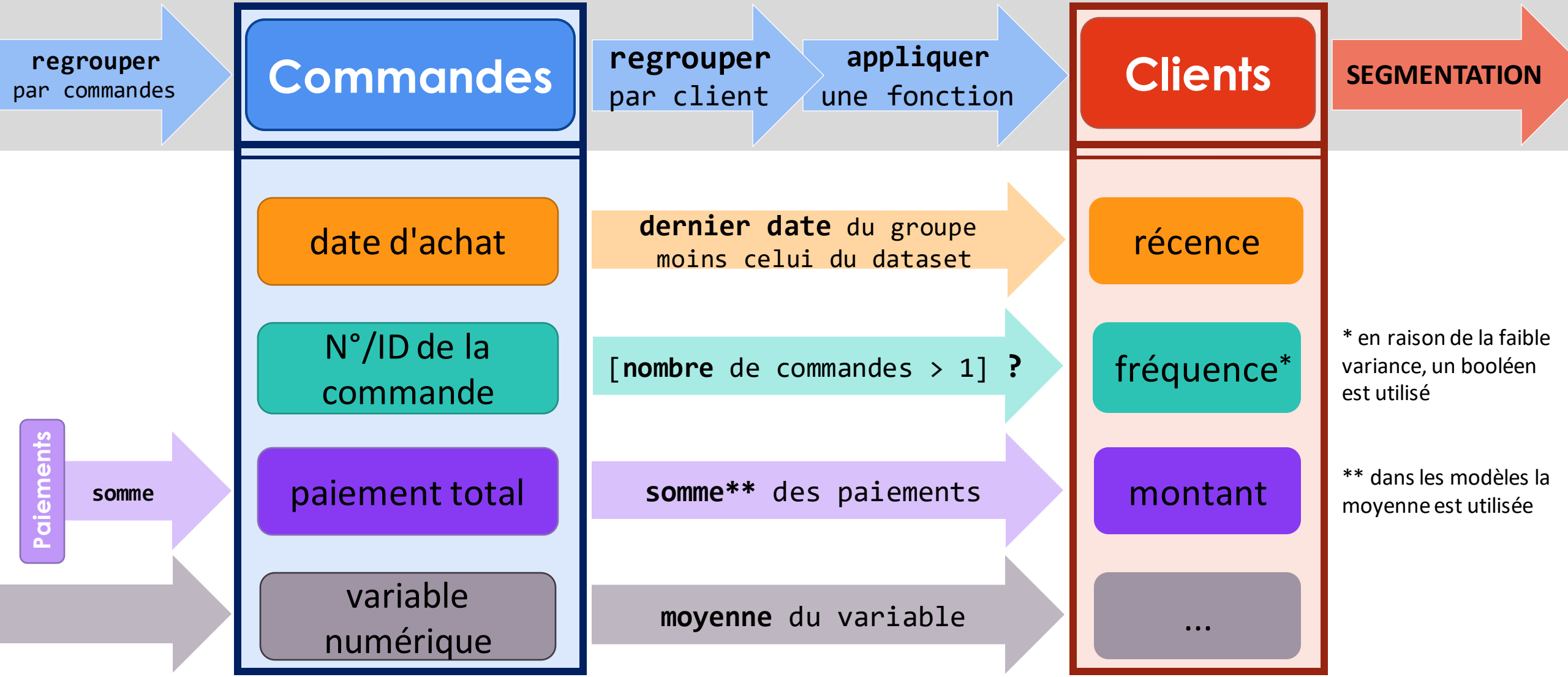


Segmentation

R écence
F réquence
M ontant

- **Récence**
  date de la dernier commande
- **Fréquence**
  des achats / commandes
- **Montant**
  du dernier commande ou sur une période donnée

Des Commandes aux Clients



* en raison de la faible variance, un booléen est utilisé

** dans les modèles la moyenne est utilisée

Features autre que RFM

	Note d'Evaluation	Densité de Population (Estimée)	Heure de la Journée en Ligne
Valeurs	1 à 5	> 0	5 à 30
Engineering		• la **fraction des clients avec le même 3 premier chiffres pour leur code postal** • en prenant la dernière commande effectuée	• Moyenne* entre **l'heure d'achat et l'heure d'évaluation** • 5 = 5h de matin • 25h = 1h de matin
Méthode d'agrégation	Moyenne	Moyenne	Moyenne* * l'heure est projetée sur un cercle
Imputation	3	- pas nécessaire -	- pas nécessaire -
Mise à l'échelle	StandardScaler	Log-scale + Standardscaler	Standardscaler

Features autre que RFM

	Distance de Livraison	Statut du Commande
Valeurs	≥ 0	-1, 0, 1
Engineering	• la **distance entre le client et le vendeur** • donnant leur coordonnées-gps • qui sont déterminer par la moyenne des **coordonnées-gps attribués à un code postal**	**-1 = indisponible, annulée** 0 = commandée, en route, ... 1 = livrée
Méthode d'agrégation	Moyenne	Moyenne
Imputation	• la localisation du code postal suivant • la localisation où il y a le plus de vendeurs	- pas nécessaire -
Mise à l'échelle	Log1p-scale + Standardscaler	--

Features autre que RFM

Temps / Délai de Livraison, Type de Paiement, Nombre d'Articles, Evaluation contient un Commentaire / une Message, ...

ATTENTION:

- **Trop de features** => modèle difficile à interpréter
- **Faibles variances** => tailles de clusters déséquilibrées
- **Fortes corrélations** => augmentation inutile (coûteuse) de la dimensionalité

un Modèle Simple et Facile à Interpréter

➢ modéliser avec un minimum de features

4 Features

RFM
+ note d'éval.

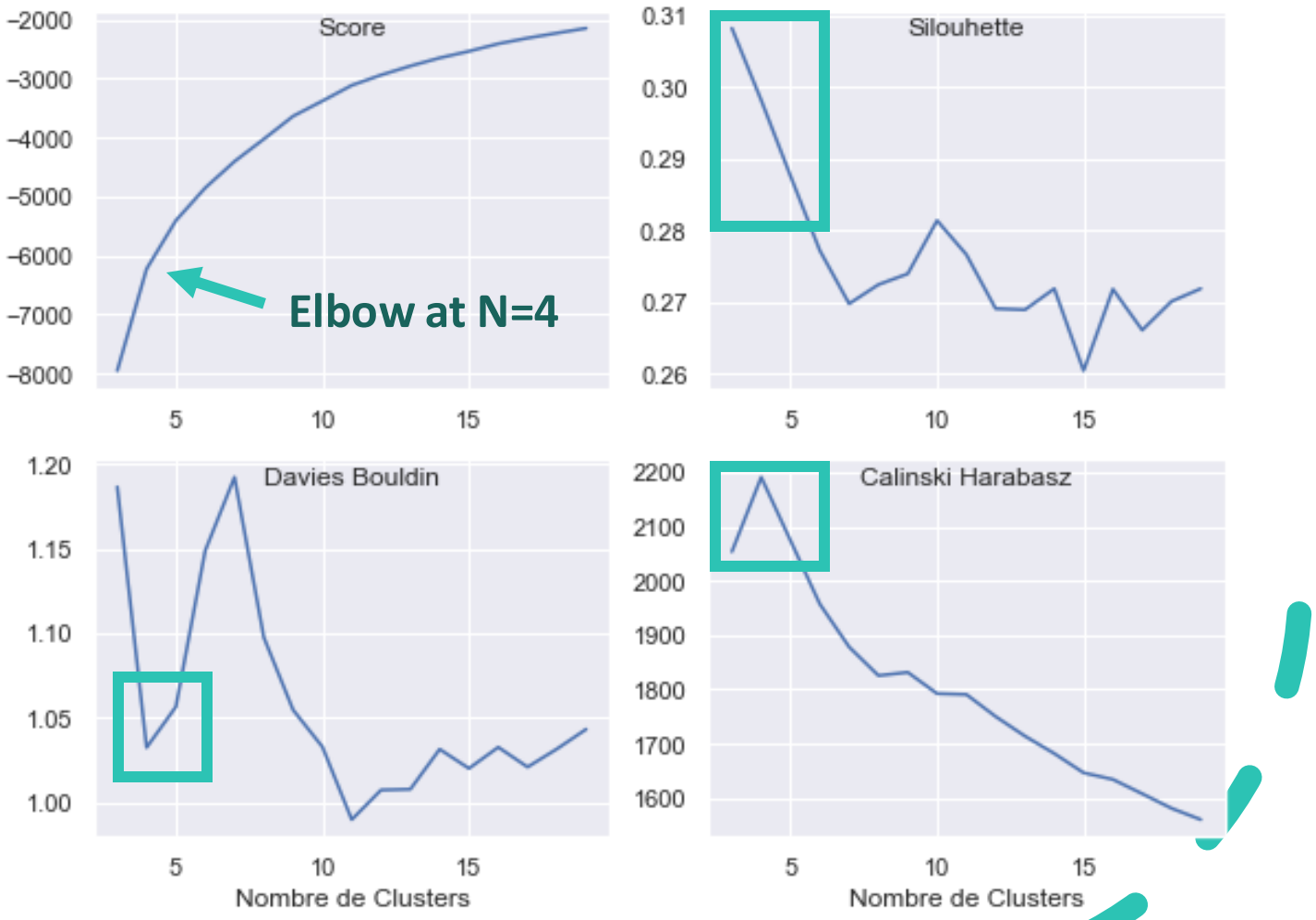
Les Métriques pour évaluer des Clusterings

- **Inertia** – **mesure de "dispersion"** égale au somme de carrées des distances dans un cluster (Kmeans)
  - un score plus bas indique des clusters plus dense
- **Davies-Bouldin** – **mesure de "partition"** qui compare la distance entre les clusters avec la taille des clusters elles-mêmes.
  - scores plus proches de 0 indiquent une meilleure partition, 0 étant le score le plus bas possible.
- **Silhouette** – **mesure de "distinction"** qui compare les distances au sein du cluster avec celles au cluster voisin le plus proche
  - Valeurs entre –1 (clustering incorrect) et +1 (clustering dense); Un valeur autour de 0 indique que les clusters se chevauchent.
- **Calinski-Harabasz** – **mesure de "distinction"** qui compare la dispersion entre les clusters avec la dispersion au sein des clusters pour tous les clusters
  - un score plus élevé indique des clusters mieux définis / séparés

KMeans
n-clusters = ?

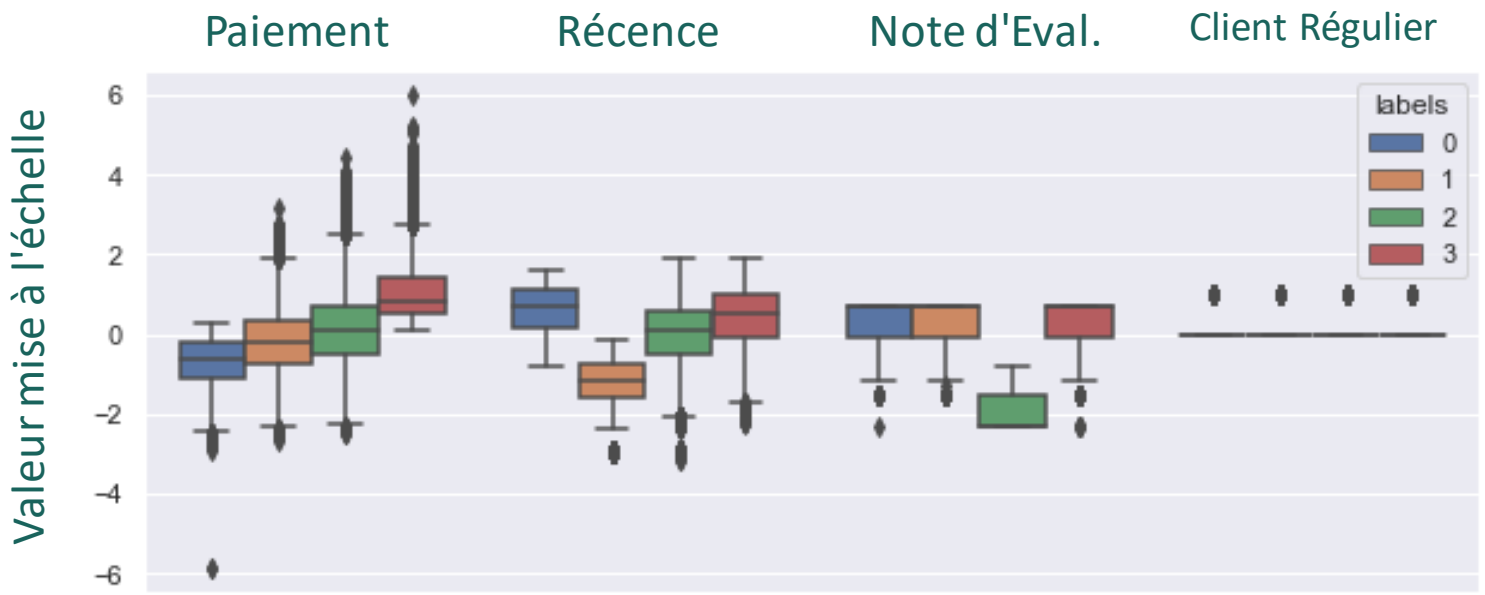
- Récence
- Client Régulier (F)
- Paiement (M)
- Note d'Evaluation

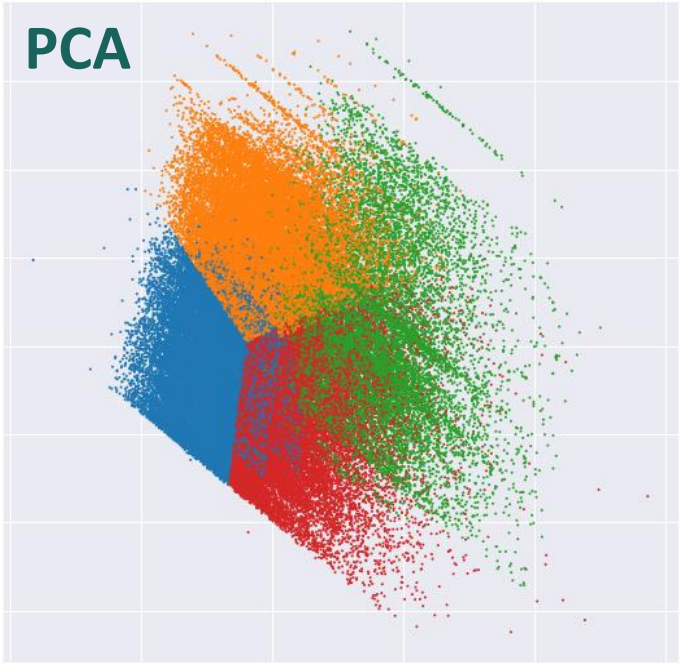
Evaluation des Modèles



KMeans
n-clusters = 4

- Récence
- Client Régulier (F)
- Paiement (M)
- Note d'Evaluation



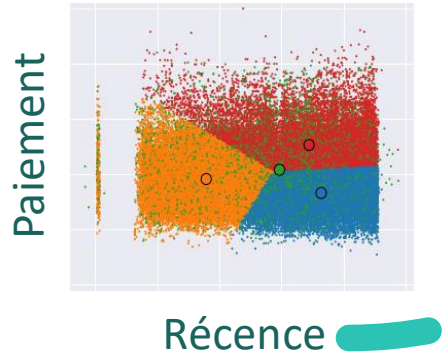


0 – économe (33%)

1 – ancien (26%)

2 – malcontent (16%)

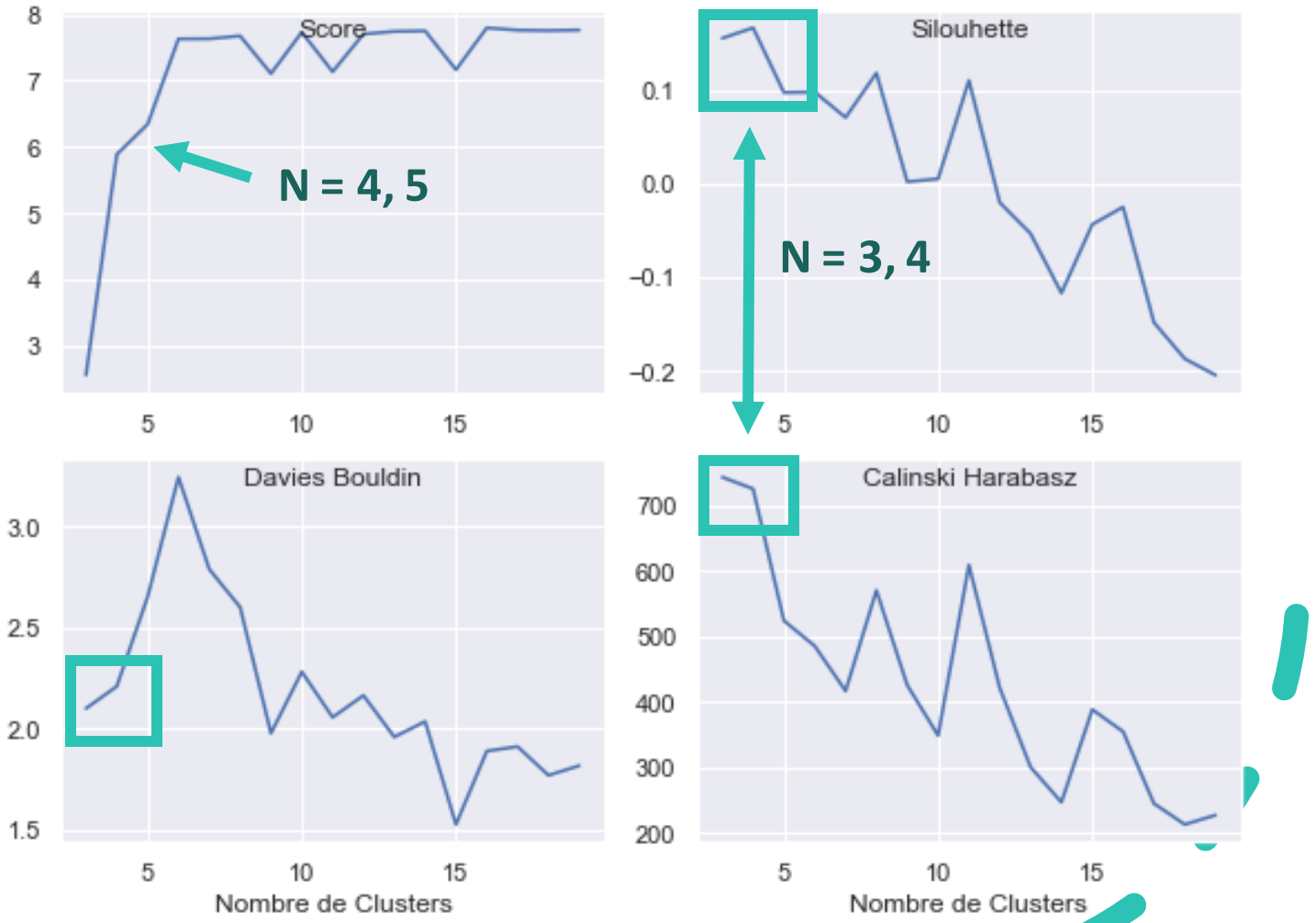
3 – grand dépensier (25%)



GaussianM
n-clusters = ?

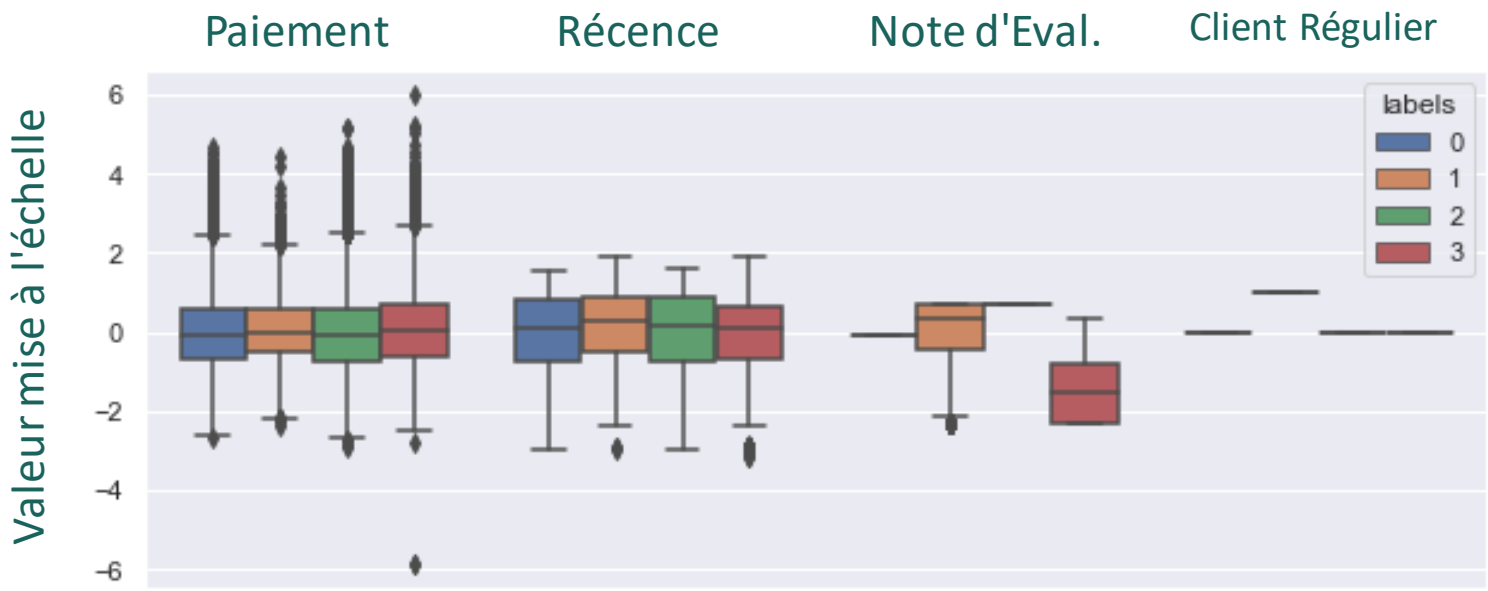
- Récence
- Client Régulier (F)
- Paiement (M)
- Note d'Evaluation

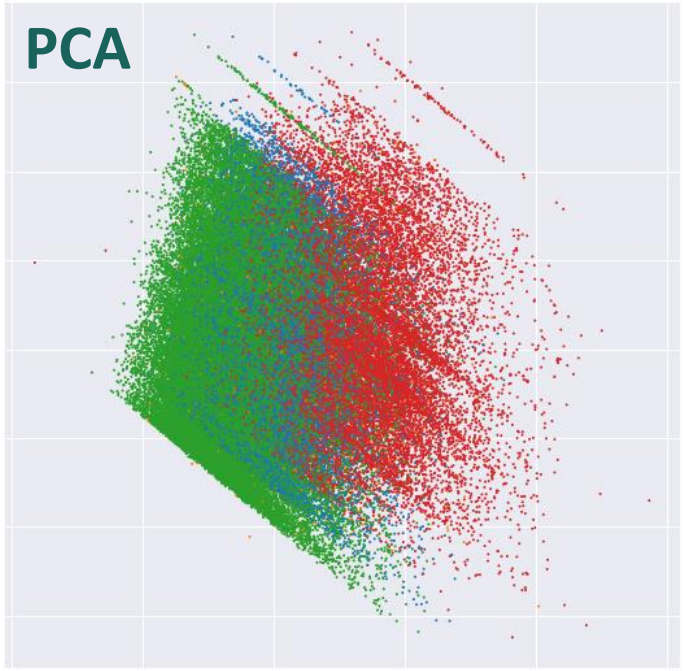
Evaluation des Modèles



GaussianM
n-clusters = 4

- Récence
- Client Régulier (F)
- Paiement (M)
- Note d'Evaluation





0 – indifférent (17%)

1 – client régulier (3%)

2 – très content (55%)

3 – malcontent (23%)

Différents Modèles

- **Choix des Features**
  - Version 0
  - Version 1
  - ...
- **Type de Modèle**
  - KMeans
  - GaussianMixture
  - BIRCH

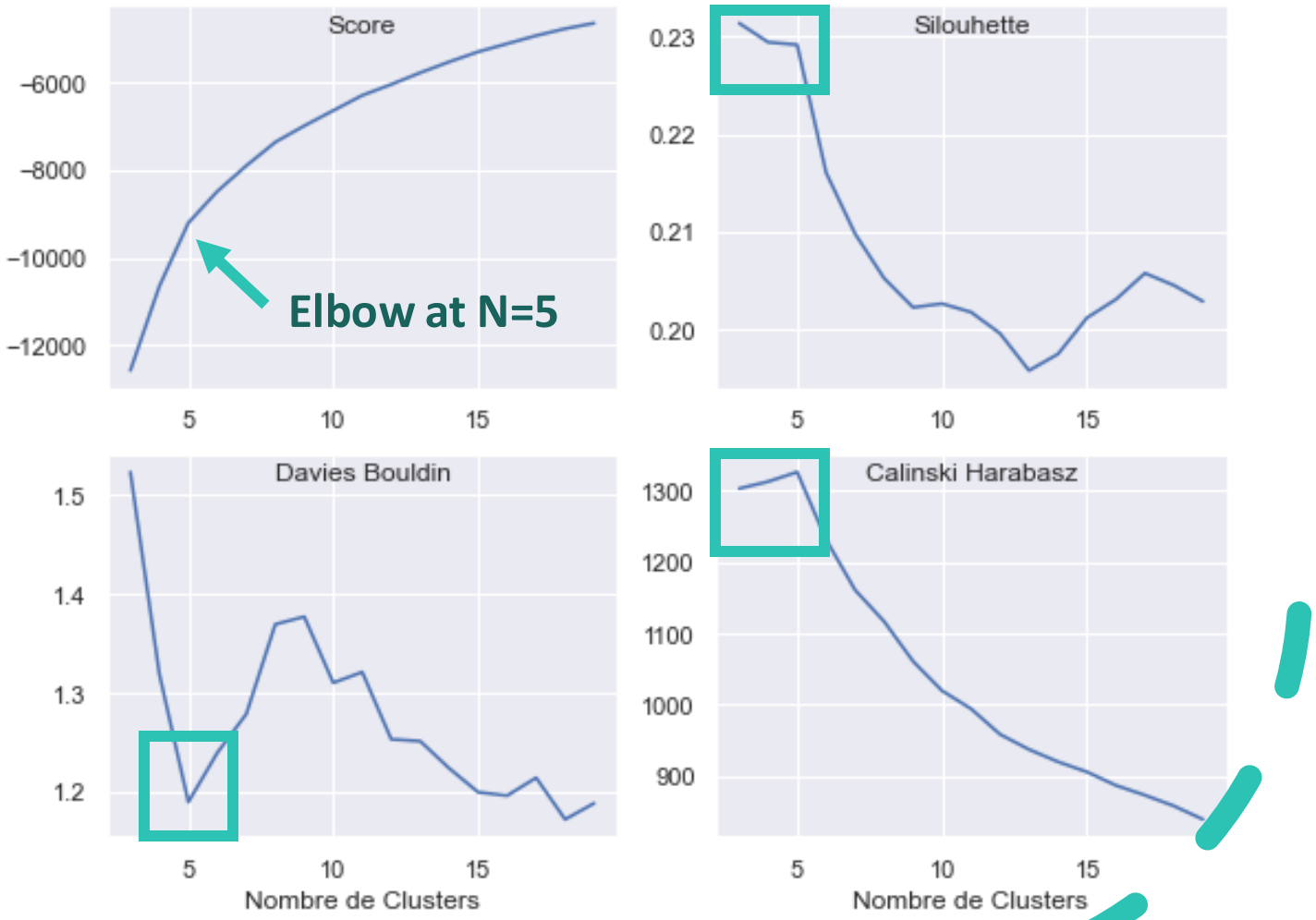
Version 0	Version 1	Version 2
• **Récence** • Client Régulier • **Paiement** • **Note d'Eval.**	• **Récence** • ~~Client Régulier~~ • **Paiement** • **Note d'Eval.**	• **Récence** • ~~Client Régulier~~ • **Paiement** • **Note d'Eval.** • Popul. Densité

Version 3	Version 4	Version 5
• **Récence** • ~~Client Régulier~~ • **Paiement** • **Note d'Eval.** • Popul. Densité • Distance Livr.	• **Récence** • ~~Client Régulier~~ • **Paiement** • **Note d'Eval.** • Popul. Densité • hh:mm en ligne	• **Récence** • ~~Client Régulier~~ • **Paiement** • **Note d'Eva** • Statut Livr. • Distance Livr.

KMeans
n-clusters = ?

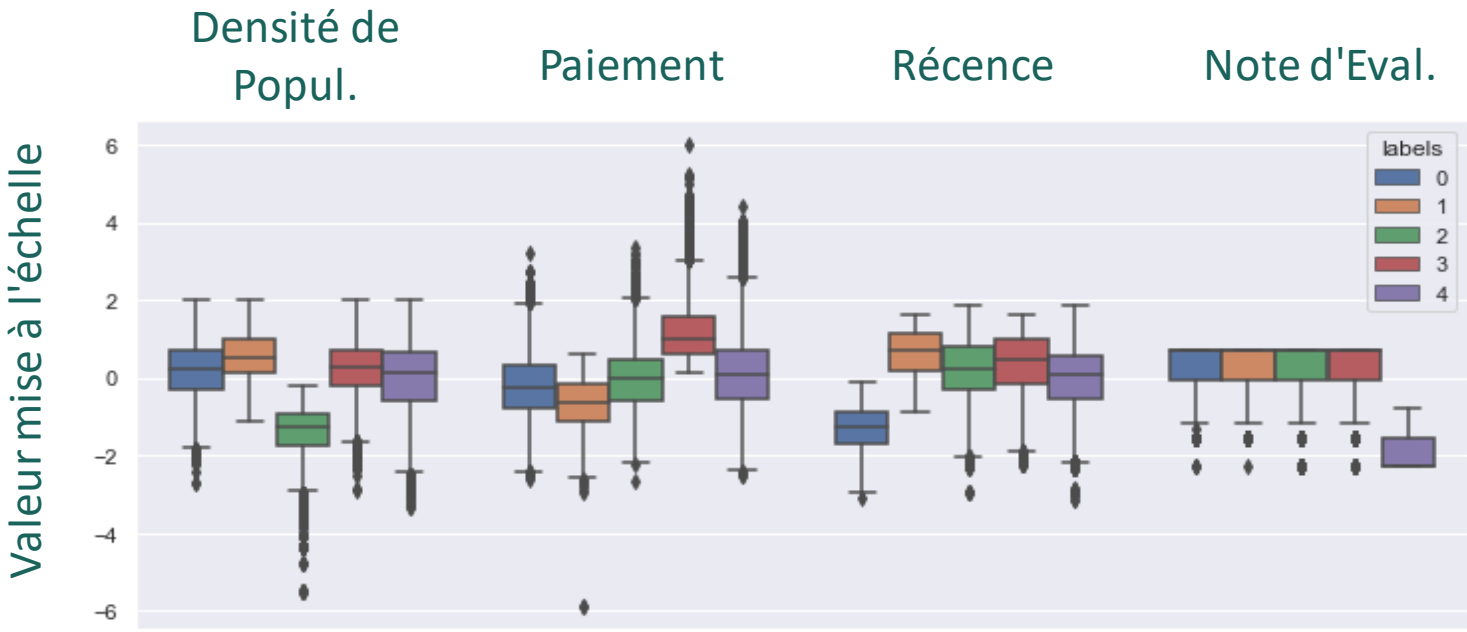
- Récence
- ~~Client Régulier (F)~~
- Paiement (M)
- Note d'Evaluation
- **Densité de Population**

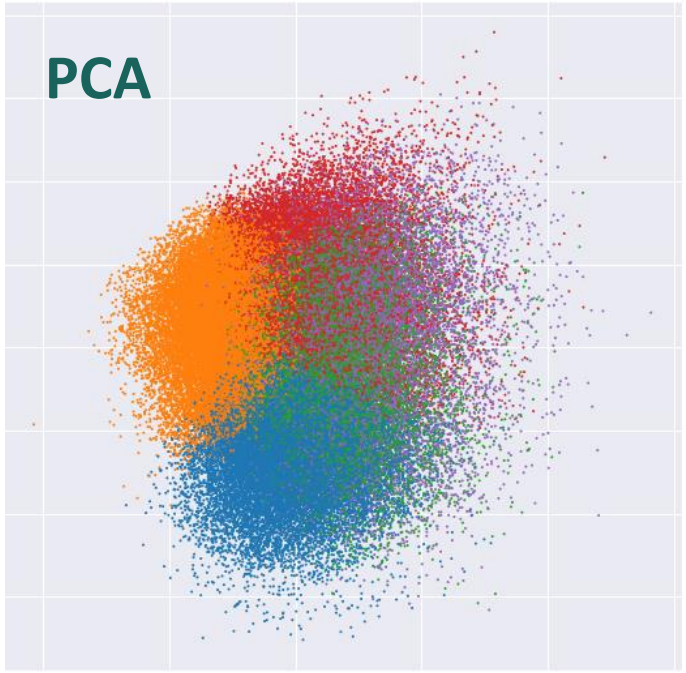
Evaluation des Modèles



KMeans
n-clusters = 5

- Récence
- ~~Client Régulier (F)~~
- Paiement (M)
- Note d'Evaluation
- **Densité de Population**

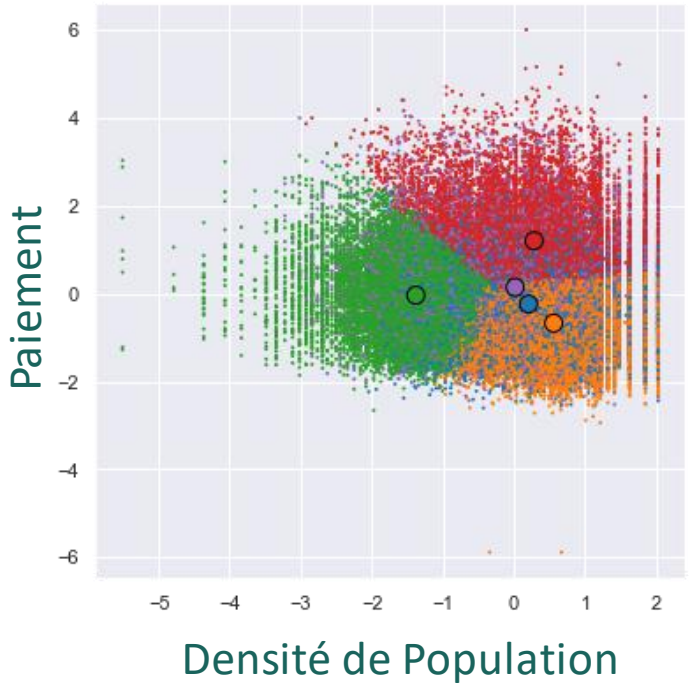


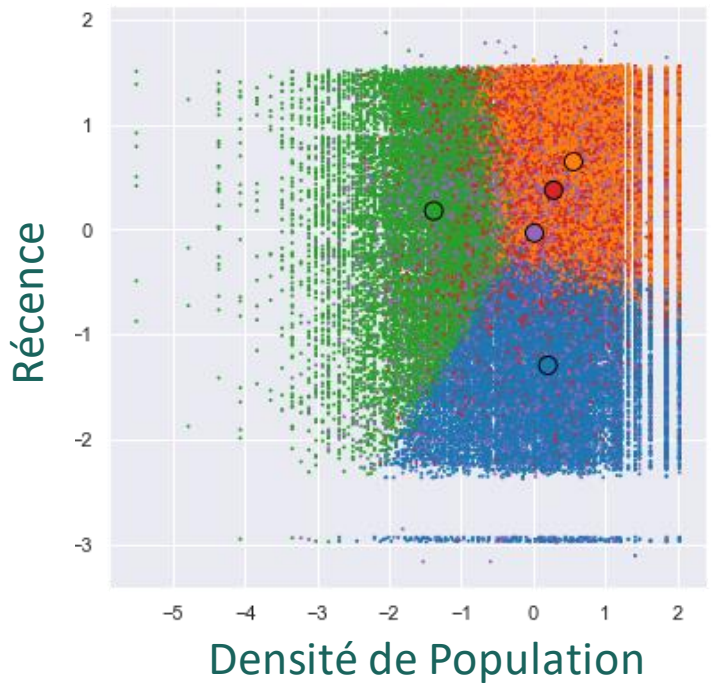


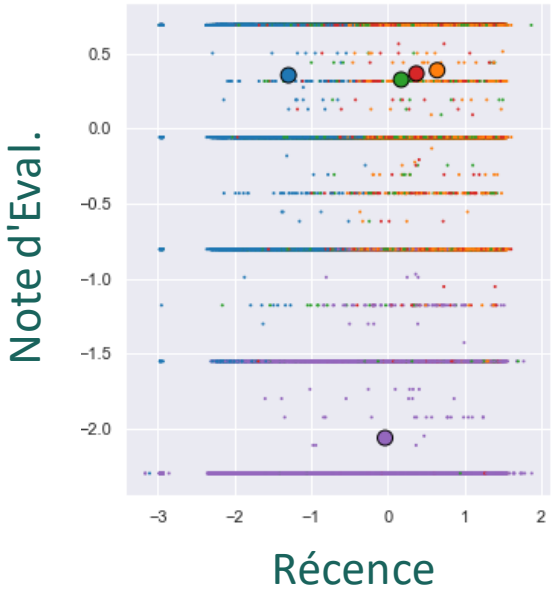
0 – ancien (21%)
1 – urbain / économe (28%)
2 – rural (18%)
3 – urbain / dépensier (18%)
4 – malcontent (15%)

KMeans
n-clusters = 5

- Récence
- ~~Client Régulier (F)~~
- Paiement (M)
- Note d'Evaluation
- **Densité de Population**







0 – ancien (21%)
1 – urbain / économe (28%)
2 – rural (18%)
3 – urbain / dépensier (18%)
4 – malcontent (15%)

Comparions des Modèles

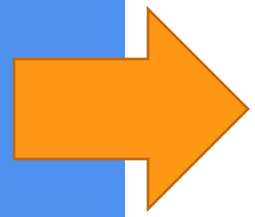
- **Choix des Features**
  - Version 0
  - Version 1
  - ...
- **Type de Modèle**
  - KMeans
  - GaussianMixture
  - BIRCH

Model	# Cl.	Sil. Score	C-H Score	D-B Score	Sil. Rank	C-H Rank	D-B Rank
KMeans_1	**4**	**0.303**	**2237**	**1.01**	**1**	**1**	**1**
KMeans_0	4	0.298	2191	1.03	2	2	2
KMeans_1	5	0.292	2126	1.04	3	3	3
KMeans_0	5	0.288	2074	1.06	4	4	4
KMeans_1	6	0.281	2013	1.15	5	5	7
KMeans_0	6	0.277	1957	1.15	6	6	8
BIRCH_1	5	0.256	1474	1.15	7	8	6
KMeans_5	**4**	**0.256**	**1476**	**1.26**	**8**	**7**	**13**
BIRCH_1	6	0.246	1268	1.09	9	13	5
KMeans_5	**5**	**0.236**	**1446**	**1.20**	**10**	**9**	**10**
KMeans_2	**4**	**0.229**	**1312**	**1.32**	**11**	**12**	**16**
KMeans_2	**5**	**0.229**	**1326**	**1.19**	**12**	**11**	**9**
BIRCH_1	4	0.227	1012	1.27	13	19	14

Fixé entre 4 et cluster 6 pour ce tableau

Différents Modèles

- **Choix des Features**
  - Version 0
  - Version 1
  - ...
- **Type de Modèle**
  - KMeans
  - GaussianMixture
  - BIRCH

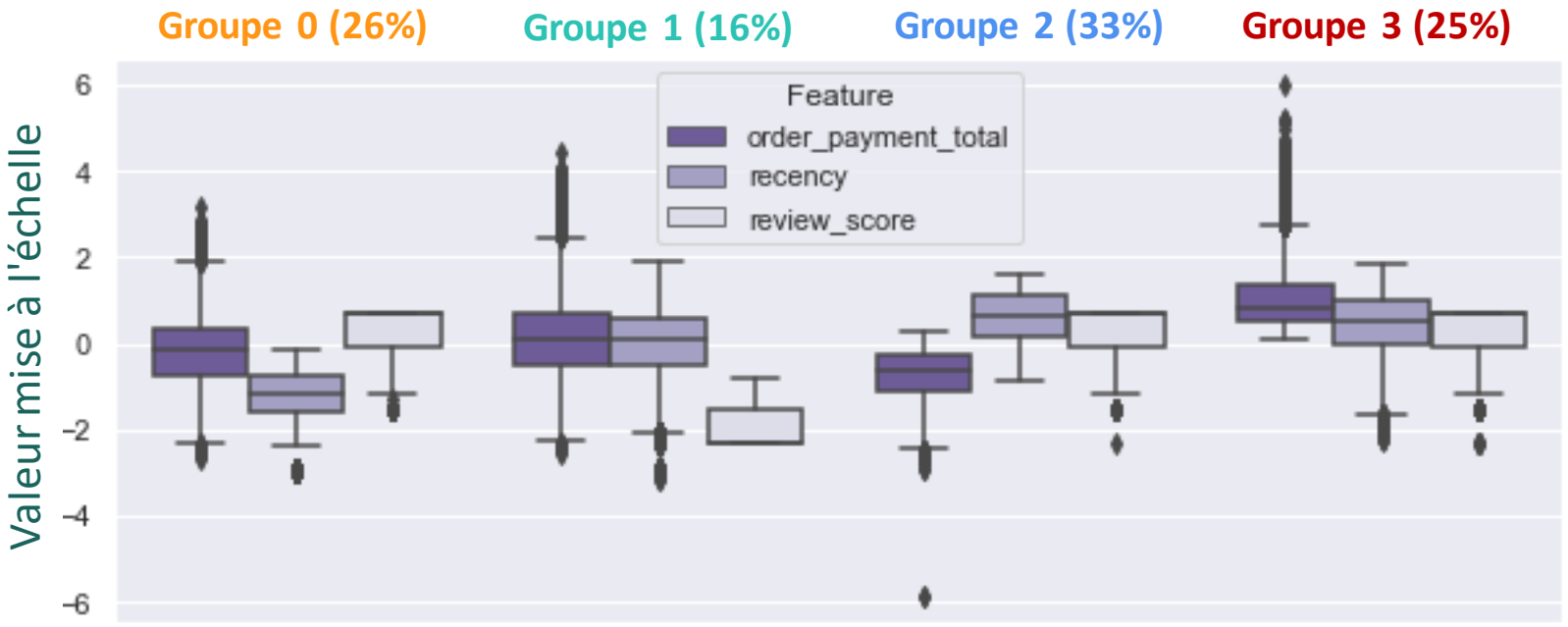


Version 0	Version 1	Version 2
• **Récence** • Client Régulier • **Paiement** • **Note d'Eval.**	• **Récence** • ~~Client Régulier~~ • **Paiement** • **Note d'Eval.**	• **Récence** • ~~Client Régulier~~ • **Paiement** • **Note d'Eval.** • Popul. Densité

Version 3	Version 4	Version 5
• **Récence** • ~~Client Régulier~~ • **Paiement** • **Note d'Eval.** • Popul. Densité • Distance Livr.	• **Récence** • ~~Client Régulier~~ • **Paiement** • **Note d'Eval.** • Popul. Densité • hh:mm en ligne	• **Récence** • ~~Client Régulier~~ • **Paiement** • **Note d'Eva** • Statut Livr. • Distance Livr.

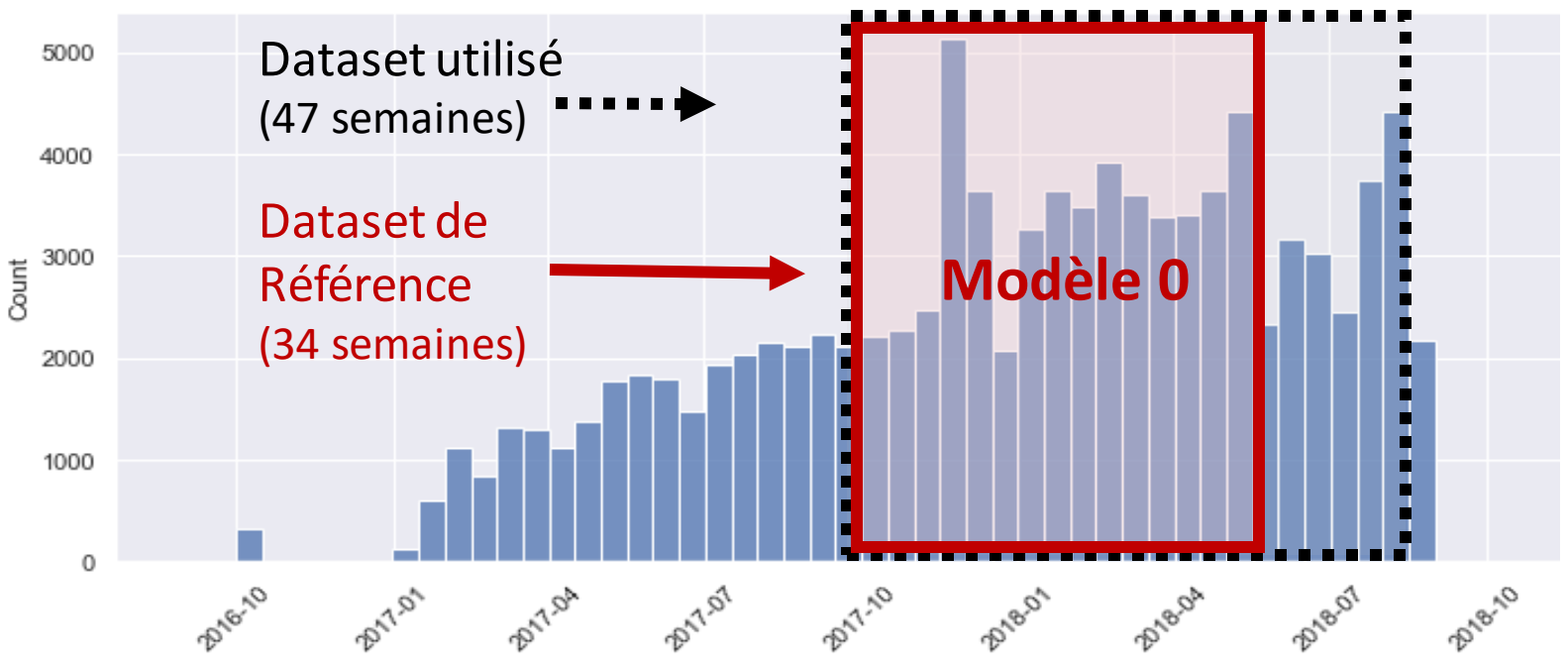
KMeans
n-clusters = 4

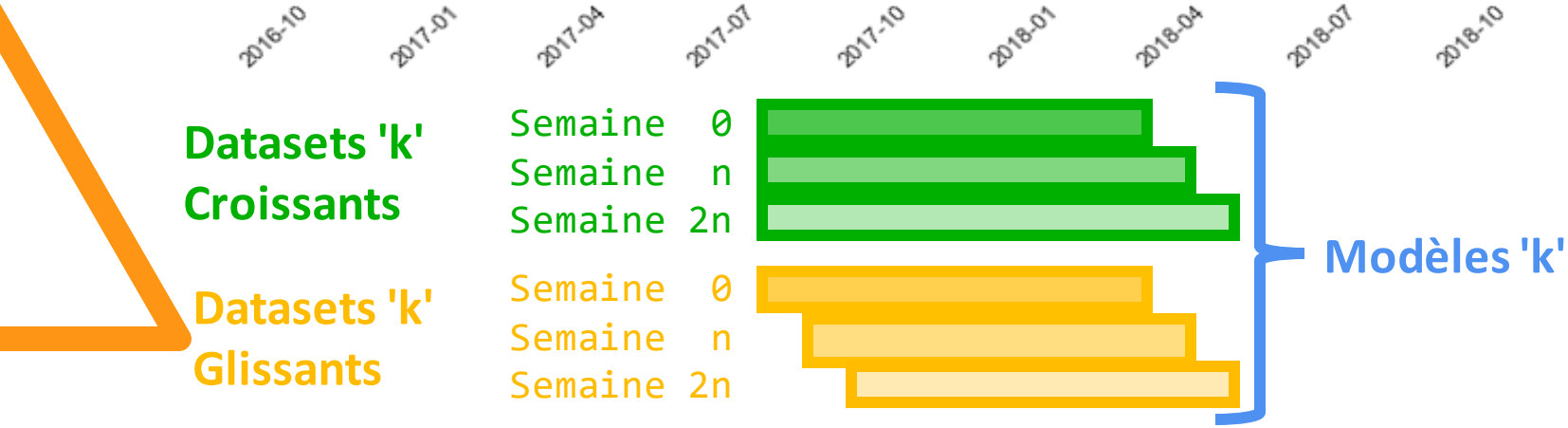
- Récence
- ~~Client Régulier (F)~~
- Paiement (M)
- Note d'Evaluation



	Ancien Content	Malcontent	Économe Content Récent	Dépensier Content Récent
Paiement (Real)	95,56	124,05	62,22	254,84
Récence (jours depuis dernier achat)	474	293	191	218
Note d'évaluation (entre 1 et 5)	4.6	1.4	4.6	4.6

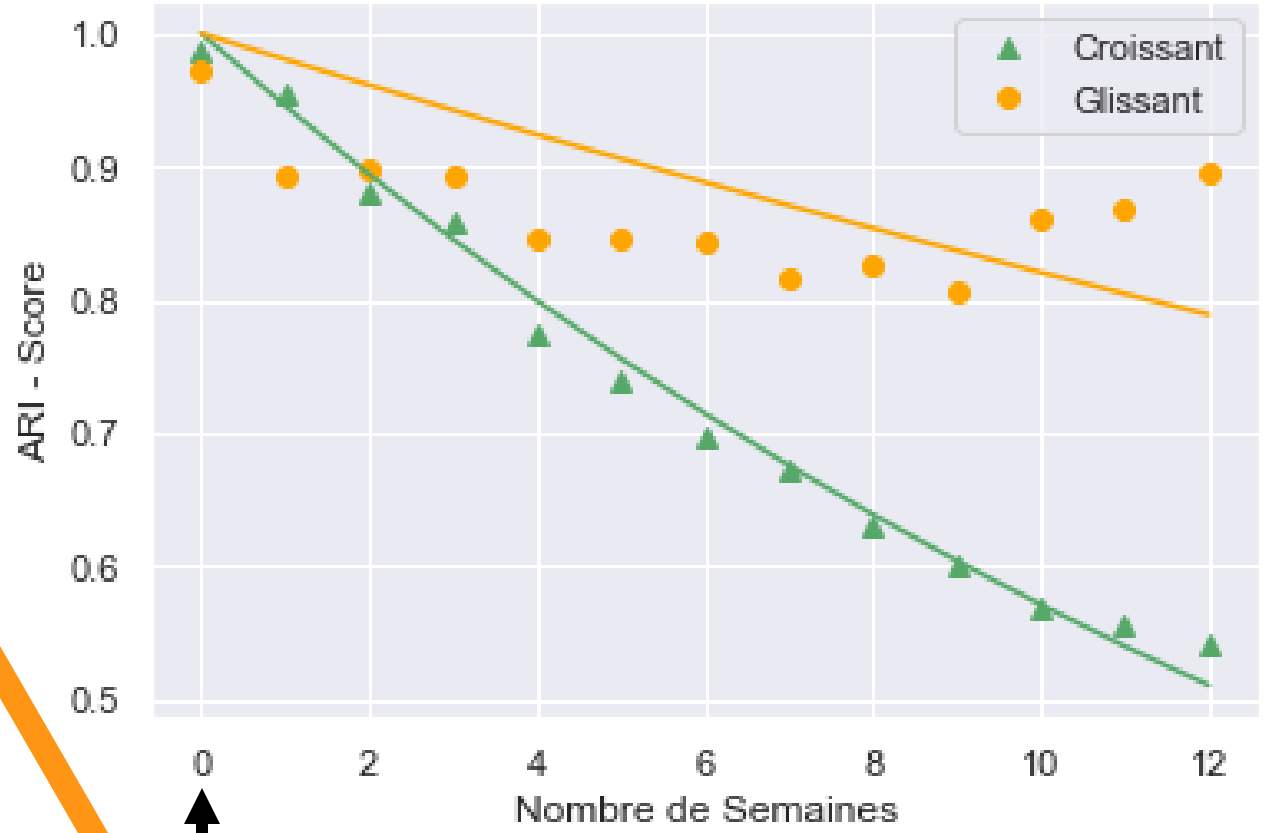
Fréquence de Mise à Jour





- **Avancement** dans le temps **en unités d'une semaine ('k')**
  - à partir du dataset de référence (= Dataset 0)
- **Segmentation** avec :
  1. Modèle 0 (—> Dataset 0)
  2. Modèle 'k' (—> Dataset 'k')
- **Computation du ARI:**
  - entre Modèle 0 et Modèle 'k'
  - mesure de correspondance des deux modèles

Fréquence de Mise à Jour



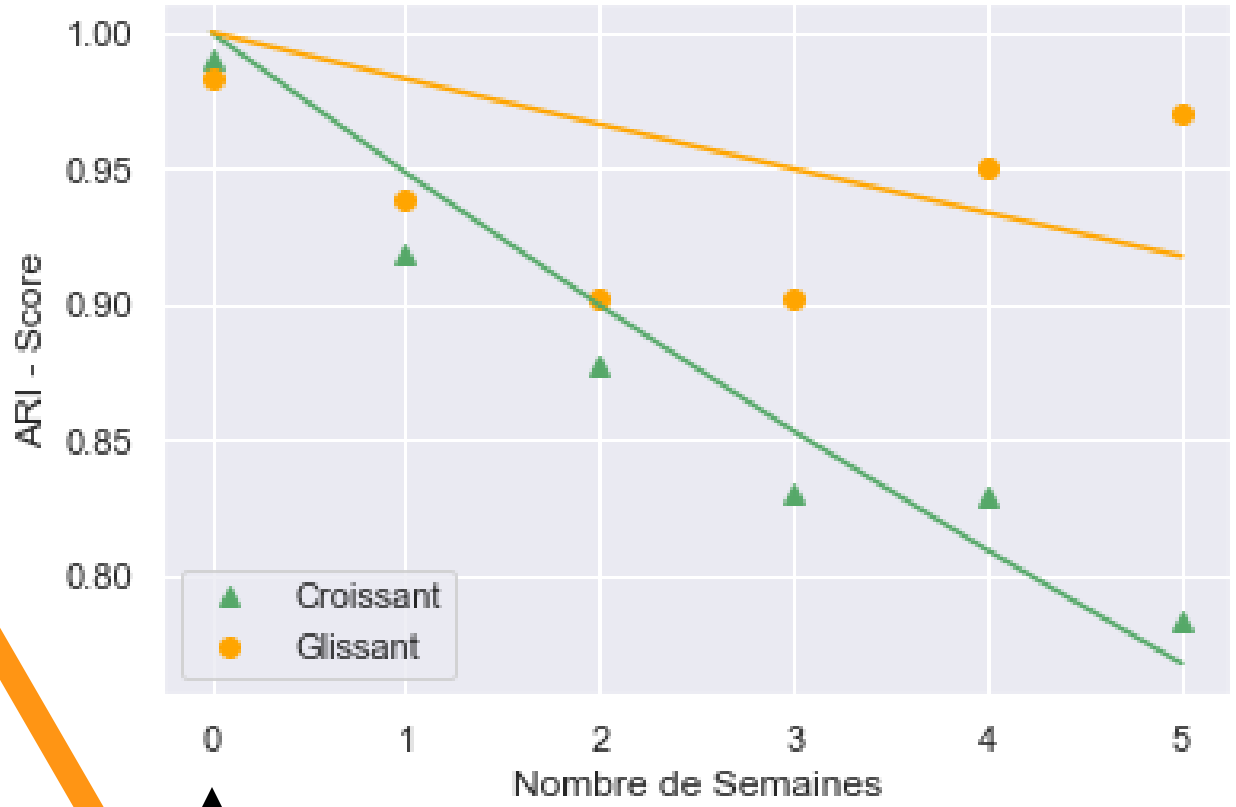
18. Septembre 2017

Décroissance (en semaines)	Datasets Croissants	Datasets Glissants
Constante Ajustée	17.9	50.6
à 0.95	0.9	2.6
à 0.90	1.9	5.3
à 0.85	2.9	8.2
à 0.80	4.0	11.3

Fréquence de mise à jour:

- tous les **4 semaines**
- min. 1 fois par mois (4-5 semaines)

Fréquence de Mise à Jour



6. Novembre 2017

Décroissance (en semaines)	Datasets Croissants	Datasets Glissants
Constante Ajustée	18.9	58.2
à 0.95	1.0	3.0
à 0.90	2.0	6.1
à 0.85	3.1	9.5
à 0.80	4.2	13.0

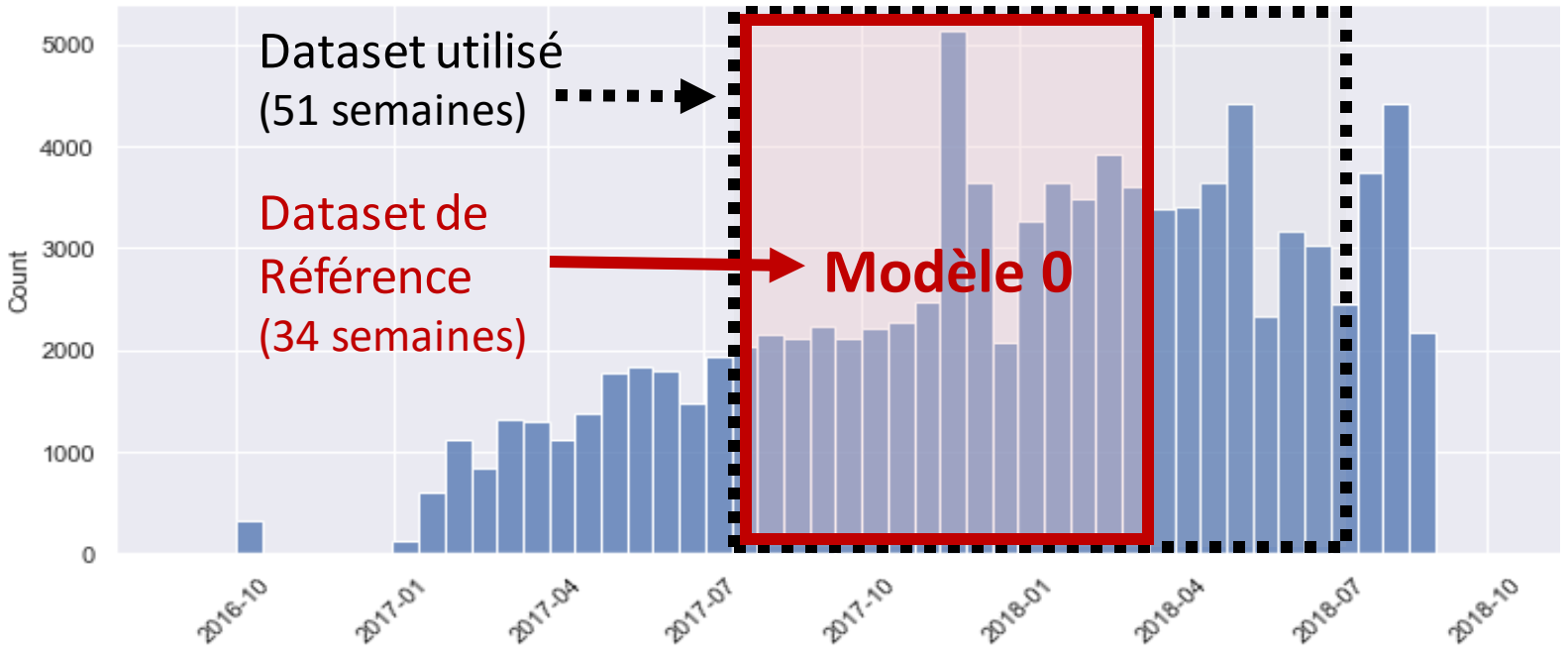
Fréquence de mise à jour:

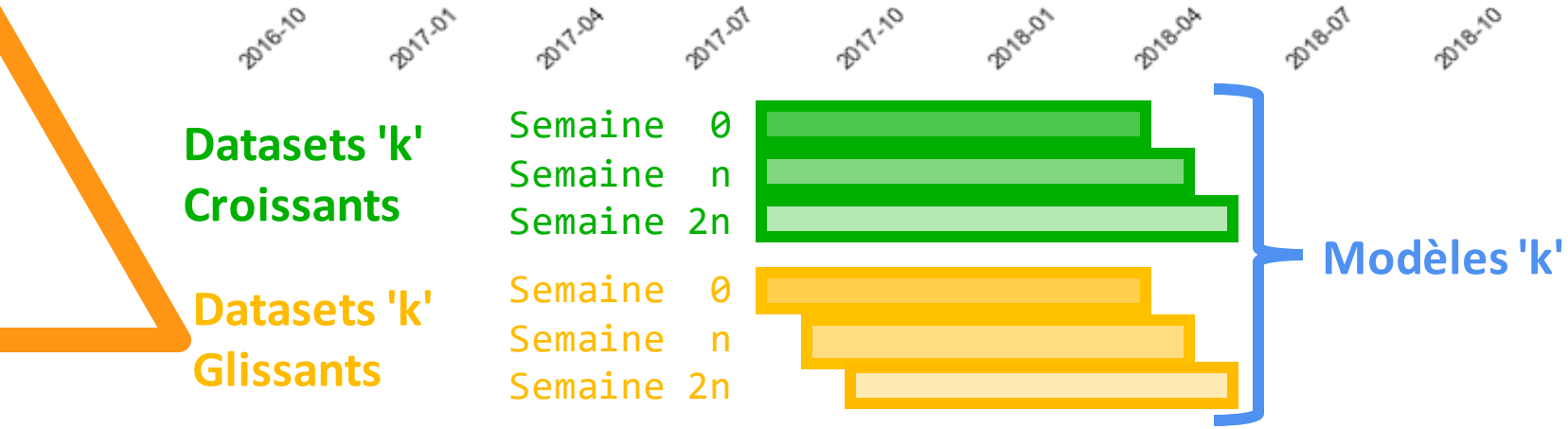
- tous les **4 semaines**
- min. 1 fois par mois (4-5 semaines)

Axes d'Amélioration

- Retour de l'équipe marketing sur les stratégies marketing prévues
  - Choix plus intentionnel de features qui donne une segmentation plus adaptée à la stratégie globale de l'entreprise
- Faire de la sous-segmentation
  - Traitement différente des clients réguliers / occasionnels
    - Avec le temps développer des profils plus approfondis pour les clients réguliers
  - Traitement différente des clients contents / malcontents
- Des informations supplémentaires sur les clients
  - Homme/Femme, Age, ….

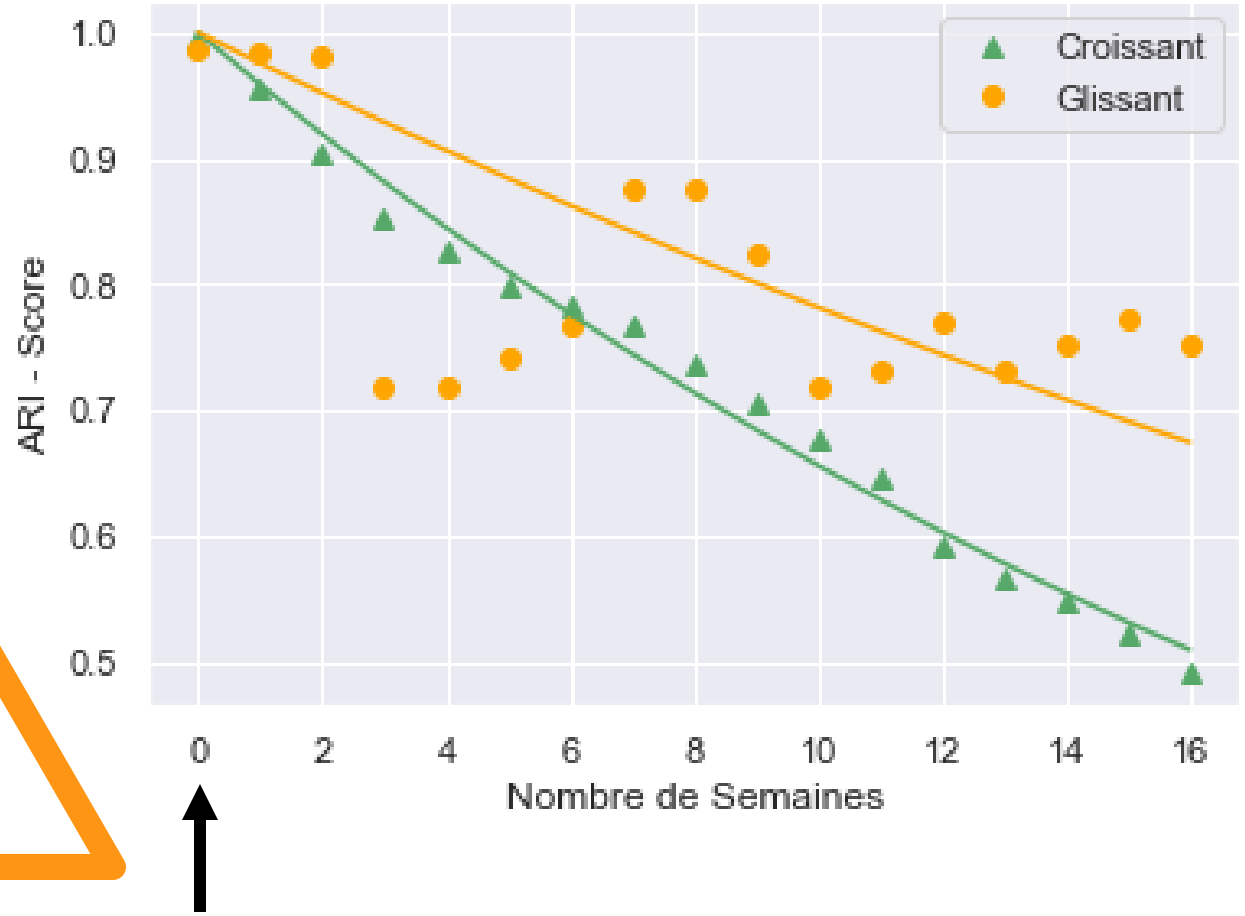
Fréquence de Mise à Jour





- **Avancement** dans le temps **en unités d'une semaine ('k')**
  - ➢ à partir du dataset de référence (= Dataset 0)
- **Segmentation** avec :
  1. Modèle 0 (—> Dataset 0)
  2. Modèle 'k' (—> Dataset 'k')
- **Computation du ARI:**
  - ➢ entre Modèle 0 et Modèle 'k'
  - ➢ mesure de correspondance des deux modèles

Fréquence de Mise à Jour



Décroissance (en semaines)	Datasets Croissants	Datasets Glissants
Constante Ajustée	23.7	40.6
à 0.95	1.2	2.1
à 0.90	2.5	4.3
à 0.85	3.9	6.6
à 0.80	5.3	9.1

Fréquence de mise à jour:

➢ tous les **5 semaines**