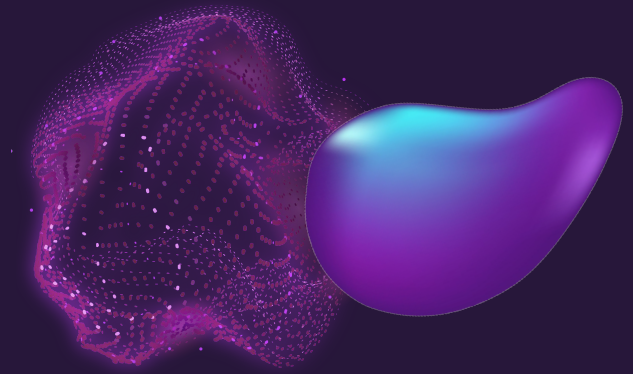


ARTIFICIAL INTELLIGENCE (AI)

Le Clustering

Présenté par:

Aoua SOW
Adama KEÏTA
Issa CISSE
Ousmane SANOGO

SOMMAIRE

01	**Qu’est-ce que le clustering**	-Clustering -Algorithme non-supervisé -Clustering Vs Classification
02	**Les types de clustering**	-Quelques catégories
03	**Quelques Exemples d’algo. De Clustering**	
	A) K-Means Clustering	-Qu’est-ce que c’est? -Comment ça marche? -Illustration
	B) Agglomerative Technique	
05	**Quelques Applications**	Dans différents domaines

01

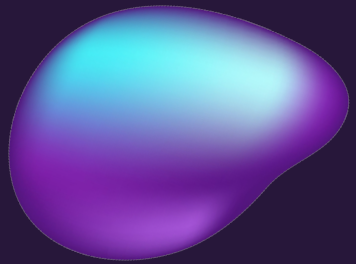
Le clustering?

01 -a) Qu’est-ce que le clustering?

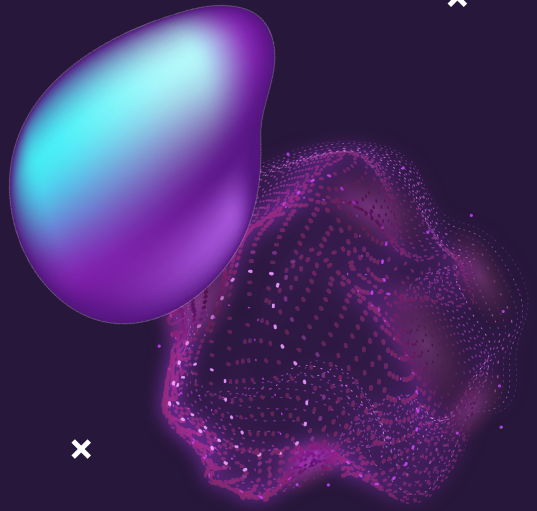
Le clustering est une technique d’apprentissage automatique non-supervisée permettant de regrouper une ensemble de données par distance ou par similarité

Pourquoi Non-Supervisée??

Chaque groupe de points de données similaires est appelé *cluster*.



01 -b) Algorithme non-supervisée

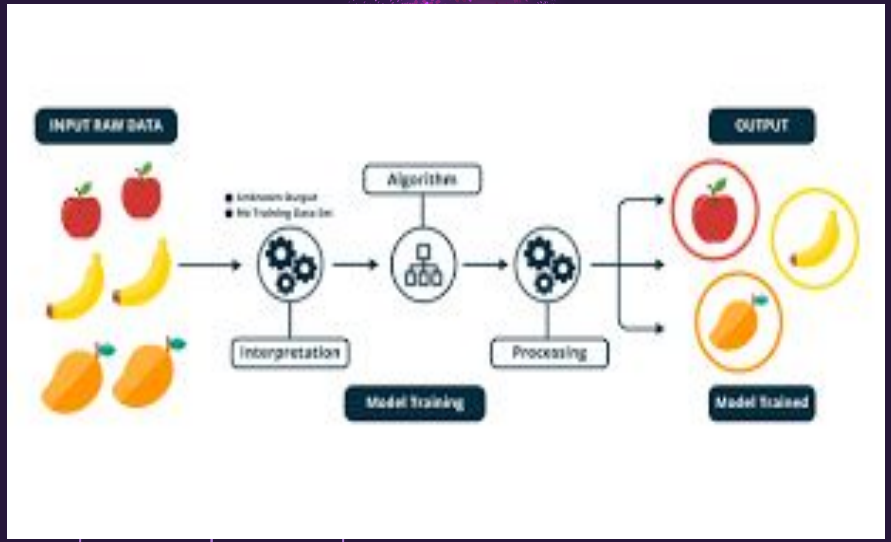


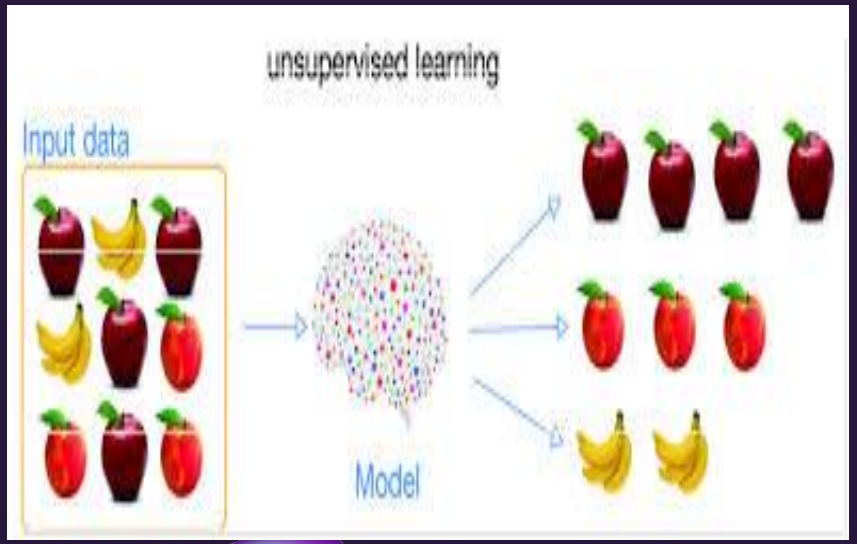
En machine learning, la technique de l'apprentissage non supervisé (ou unsupervised learning) consiste à entraîner des modèles, sans réaliser d'étiquetage manuel ou automatique des données au préalable. Les algorithmes regroupent les données en fonction de leur similitude, sans aucune intervention humaine.

Quelle différence avec le supervisé que je connais déjà?

01 -c) Unsupervised Vs Supervised

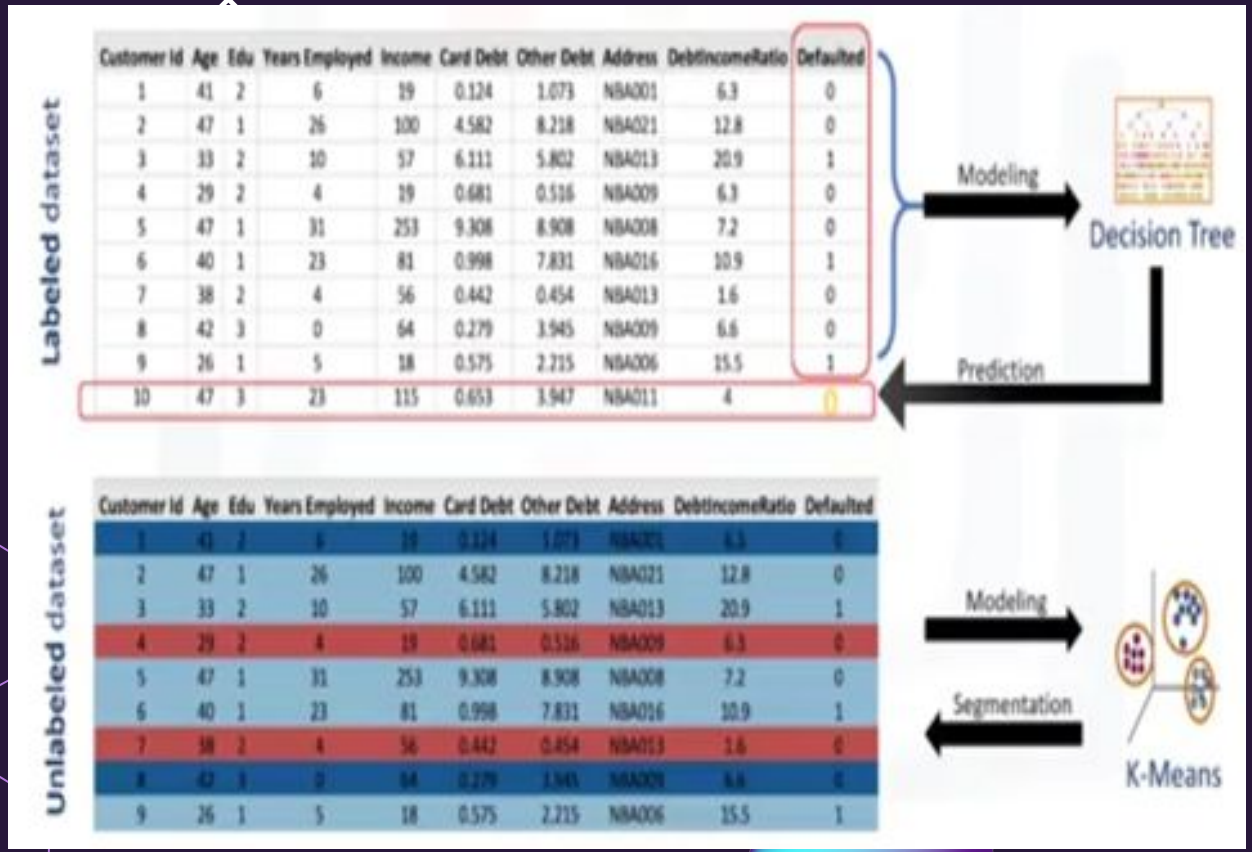
Exemple1





01 -c) Unsupervised Vs Supervised

Exemple2:
Précisément:
(Clustering
Vs
Classification)

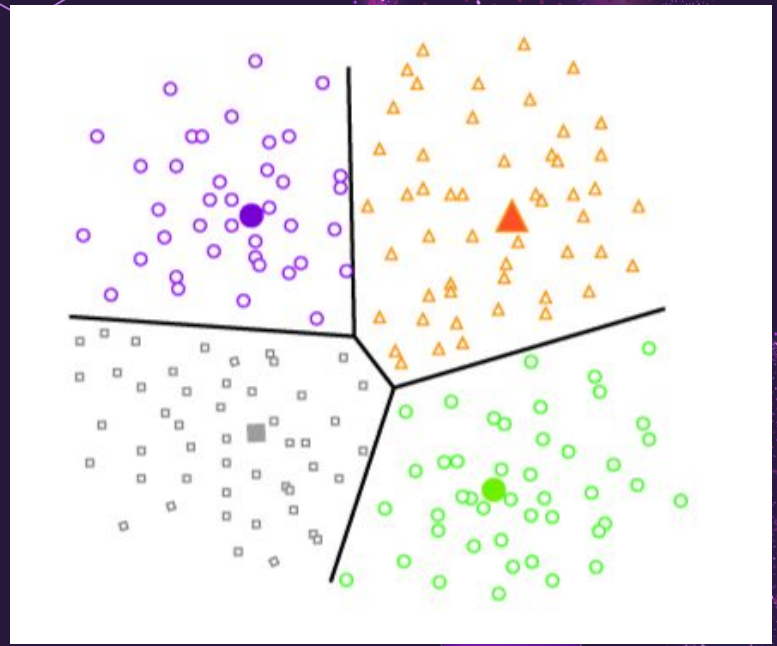


02

Les types De Clustering

02 a) Partitioned-based clustering

Le clustering de partitionnement est également appelé clustering basé sur la notion de point central appelé “centroïde” et de regroupement basé sur ce centroïde. Dans le clustering de partitionnement, le point de données est divisé en groupes non hiérarchiques



Exemples: K-Means, fuzzy c-Means

02 b) Hierarchical clustering

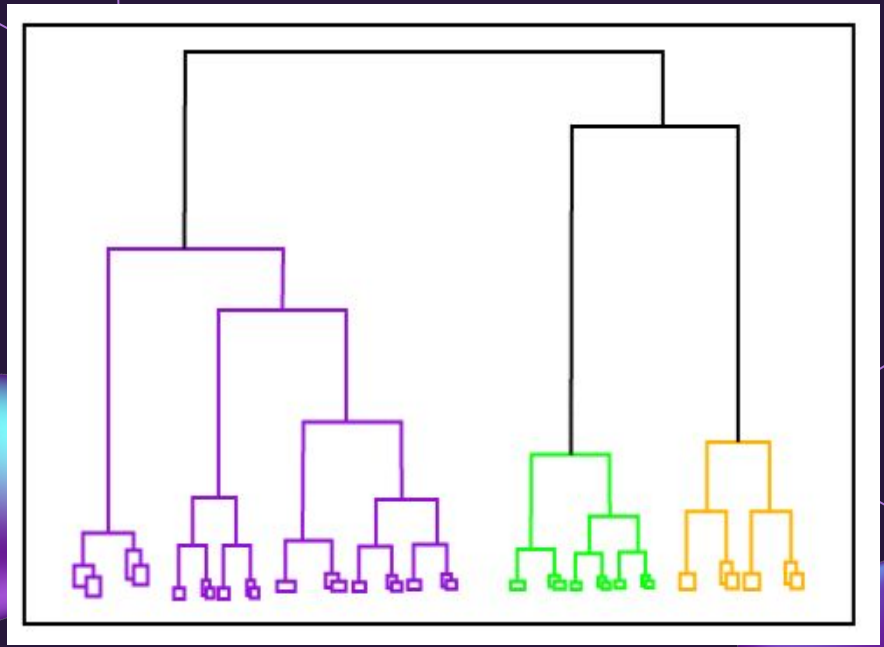
-Regroupement hiérarchique qui consiste à créer une arborescence de cluster pour représenter les données.

-Deux types de hiérarchisation:

*__Divisive:__ Consiste à diviser un cluster en sous ensembles de clusters.

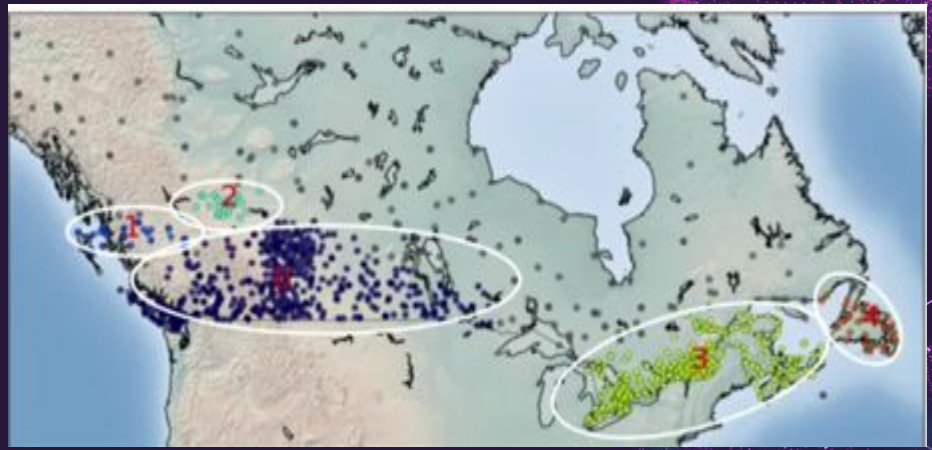
***Agglomerative :** Regrouper des sous ensembles de clusters en cluster généralisé.

-Les groupes sont imbriqués entre eux et organisés sous la forme d'un arbre.



02 c) Density-based clustering

Le clustering basé sur la densité est un moyen de regrouper un ensemble de données en fonction de la densité des points de données et de connecter ces points de données plus denses dans un cluster.



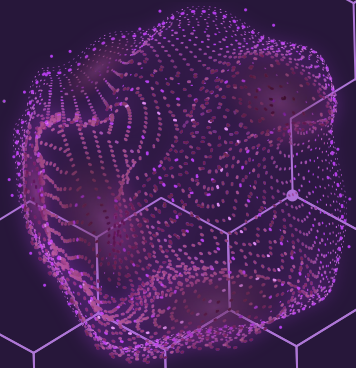
Produit un cluster de forme arbitraire

Exemples: DBSCAN

03
Quelques exemples d’algo. De Clustering

03 a) K-Means

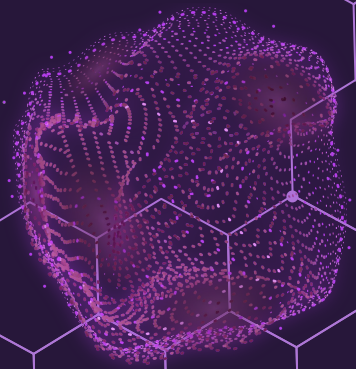
- C’est une méthode de clustering basée sur la division par groupes similaires.
- Le k-means divise les données ou data points en sous-ensembles distincts.
- Cette division est basée sur une structure interne des données mais sans target (technique non-supervisée).
- Chaque groupe de clusters est suffisamment différent.



03 a) K-Means

Fonctionnement

- Choisir aléatoirement le nombre de centroïdes.
- Calculer la distance de chaque point par rapport aux centroïdes.
- Affecter les points aux centroïdes les plus proches
- Recalculer les coordonnées des centroïdes par rapport aux points dans le but de les faire déplacer
- Répéter les étapes du 2 à 4 jusqu'à ce qu'on ne puisse plus bouger les centroïdes



03 a) K-Means

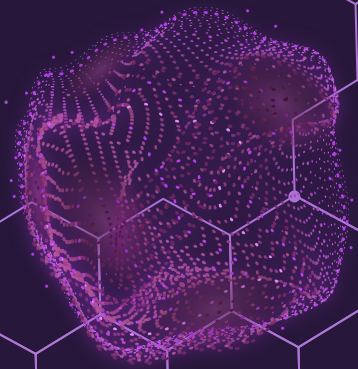
Illustration

03 a) K-Means

Illustration

1.Choisir aléatoirement le nombre de centroïdes.

- Le choix du nombre de clusters
- Dépend fortement de la taille et de la distribution des données.
- C’est une opération très délicate. Heureusement, il existe des techniques d’observation ou de visualisation qui proposent des alternatives.
- Un constat lié à l’expérience, nous dit qu’un nombre élevé de k produit toujours une erreur minimum.
- La méthode du coude de la courbe ou Elbow méthode aide beaucoup à choisir le nombre de clusters.



03 a) K-Means

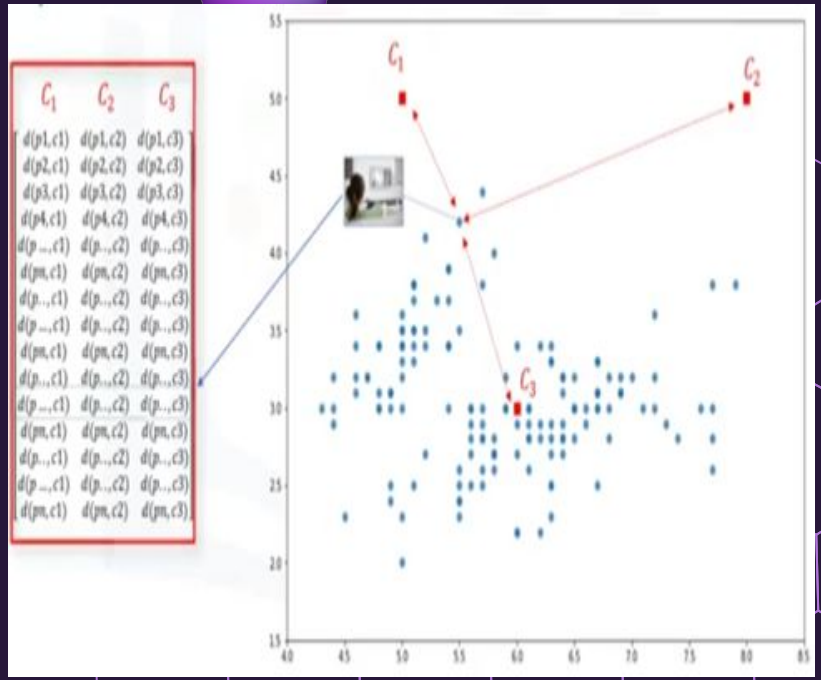
Illustration

- Génère la matrix ou le tableau des distances entre les données.
  C'est une Matrice de Distance entre data points ou entre les exemples et le centroïde.

- Dans cet exemple, il s'agit de la distance entre chaque client de l'entreprise.

- *La méthode du coude de la courbe ou Elbow méthod aide beaucoup à choisir le nombre de clusters.*

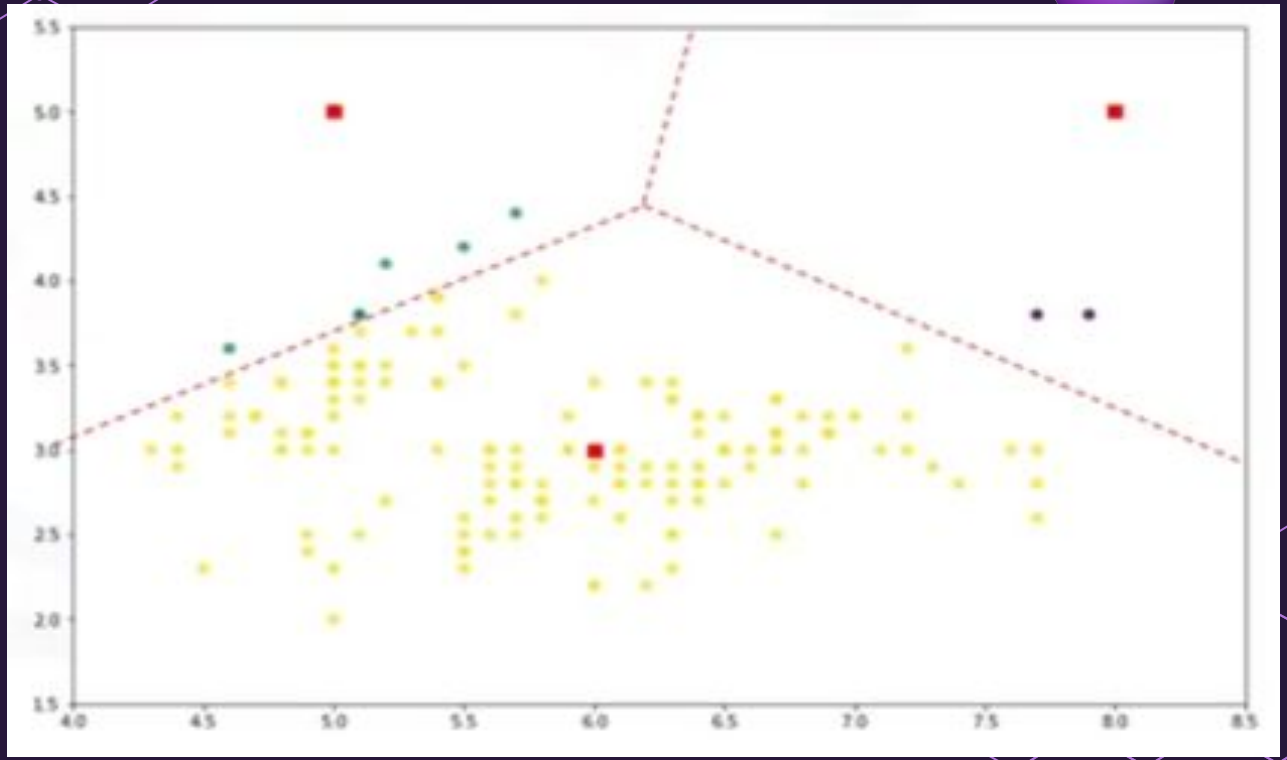
2. Calculer la distance de chaque point par rapport aux centroïdes.



03 a) K-Means

Illustration

3.Affecter les points aux centroïdes les plus proches

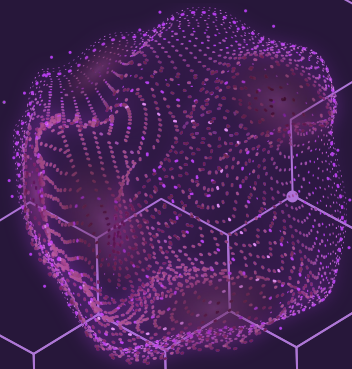


03 a) K-Means

Illustration

4.Recalculer les coordonnées des centroïdes par rapport aux points dans le but de les faire déplacer (l'Erreur)

- C'est la distance entre chaque point et le centroïde.
- Cette erreur nous renseigne par rapport à l'efficacité des centroïdes choisis.
- Si elle est grande (l'erreur), alors déplace les centroïdes.
- Le choix du déplacement des centroïdes est effectué par le calcul des moyennes des coordonnées de chaque point.
- Les coordonnées des nouveaux centroïdes seront les moyennes des points les plus proches



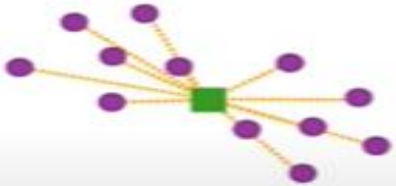
03 a) K-Means

Illustration

- 4.Recalculer les coordonnées des centroïdes par rapport aux points dans le but de les faire déplacer (l'Erreur)

En résumé :

K-Mean cherche la position des centres qui **minimise** la **distance** entre les **points** d'un cluster (x_i) et le **centre** (μ_j) de ce dernier:

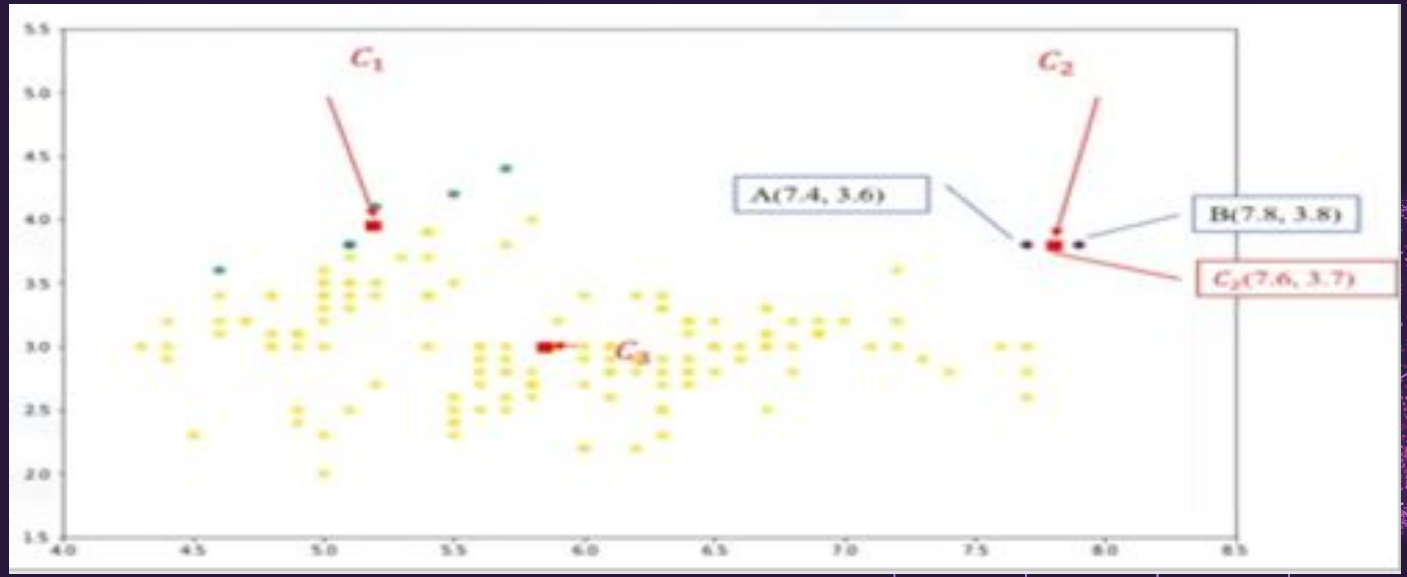


$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

03 a) K-Means

Illustration

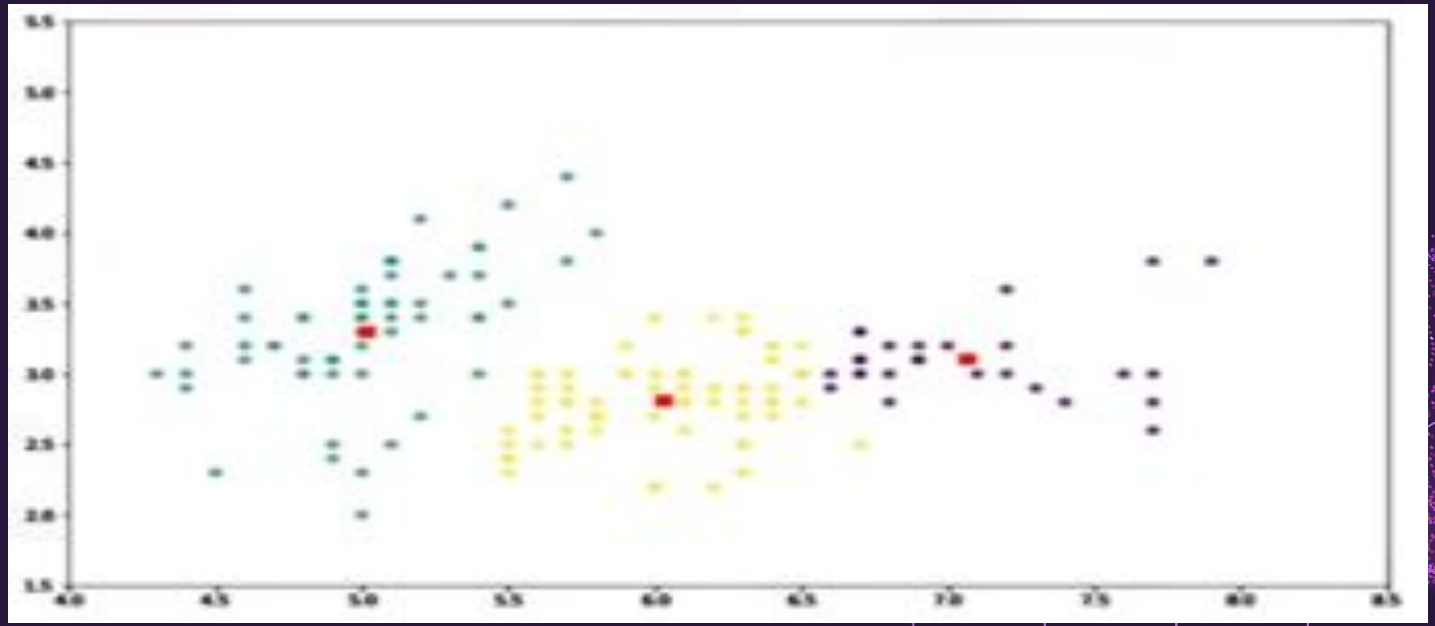
- 4. Recalculer les coordonnées des centroïdes par rapport aux points dans le but de les faire déplacer (l'Erreur)



03 a) K-Means

Illustration

- 5. Répéter le déplacement des centroïdes jusqu'à ce qu'on ne puisse les déplacer.

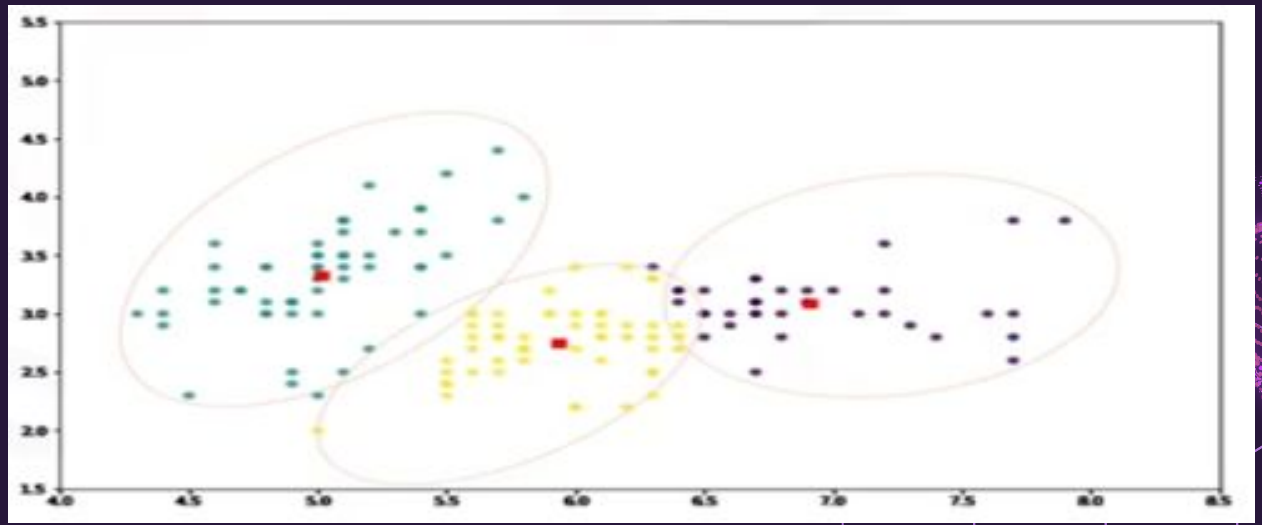


03 a) K-Means

Illustration

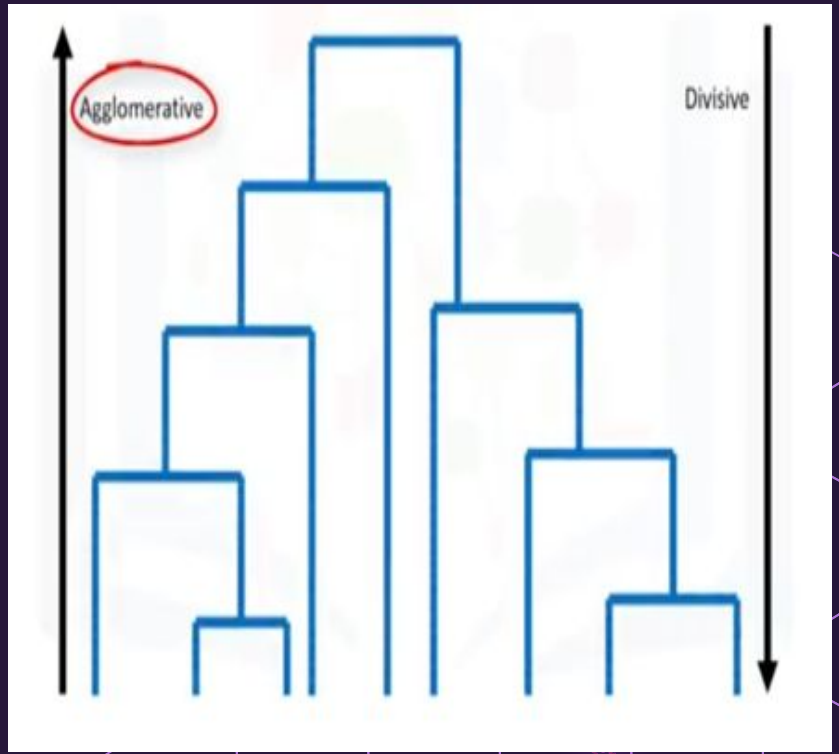
- 5. Répéter le déplacement des centroïdes jusqu'à ce qu'on ne puisse les déplacer.

Les clusters résultants de ce procédé sont ceux générant l'erreur la plus basse – The minimum Error not the best.



03 b) Agglomerative

C’est un clustering hiérarchique qui fait une structure arborescente. Dans l'algorithme Agglomerative, le clustering sera de type ascendant où chaque point de données sera considéré comme un cluster au début, puis il sera fusionné dans l'arborescence.



03 b) Agglomerative

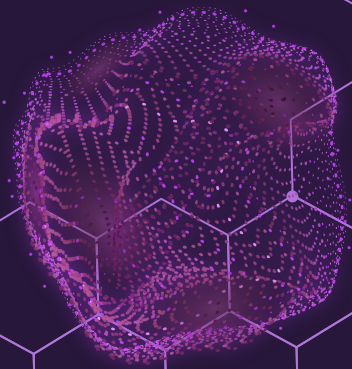
Fonctionnement

1. Créer n clusters. C’est-à-dire chaque data point devient un cluster.

2. Calculer la matrice de proximité selon la distance entre les clusters.

3. Répéter les opérations suivantes:
   Fusionner les deux clusters les plus proches.
   Corriger ou faire la mise à jour de la matrice.

4. Jusqu’à ce qu’il ne reste qu’un seul cluster.



03 b) Agglomerative

Illustration

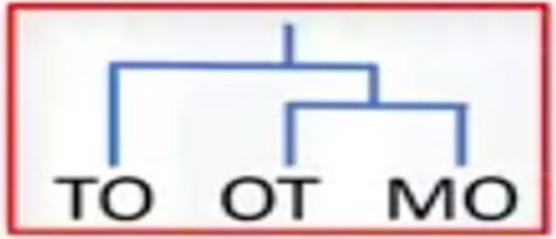
Agglomerative clustering

TO OT MO VA ED WI

	TO	OT	VA	MO	WI	ED
TO		351	3363	505	1510	2699
OT			3543	167	1676	2840
VA				3690	1867	819
MO					1824	2976
WI						1195
ED						

03 b) Agglomerative

Illustration

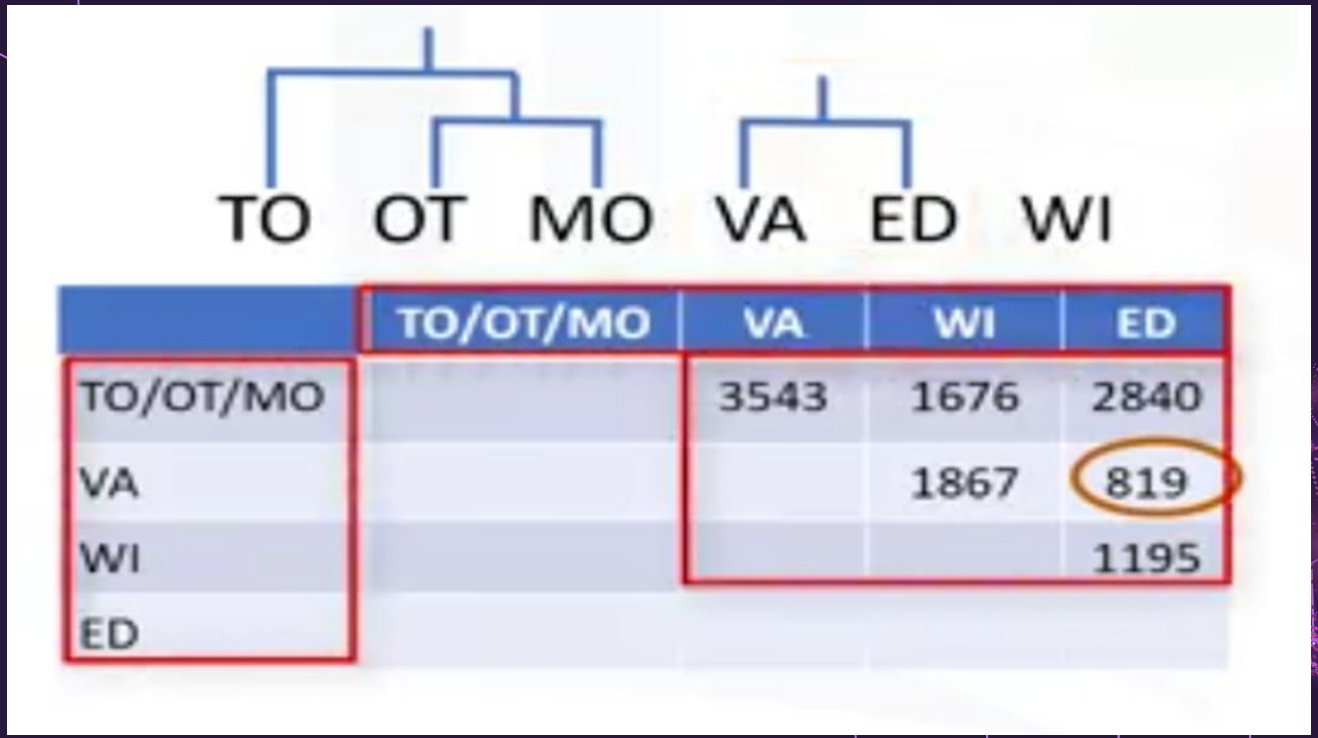


VA ED WI

	TO	OT/MO	VA	WI	ED
TO		351	3363	1510	2699
OT/MO			3543	1676	2840
VA				1867	819
WI					1195
ED					

03 b) Agglomerative

Illustration

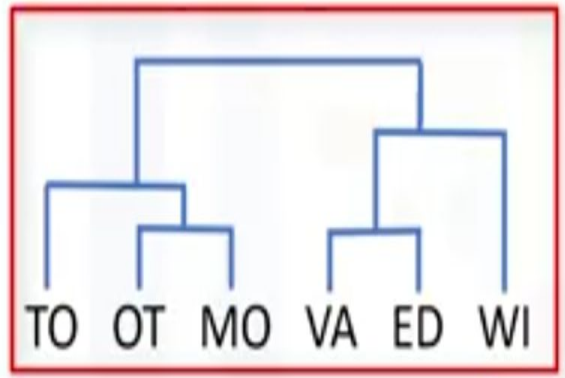


	TO/OT/MO	VA	WI	ED
TO/OT/MO		3543	1676	2840
VA			1867	819
WI				1195
ED				

03 b) Agglomerative

Illustration

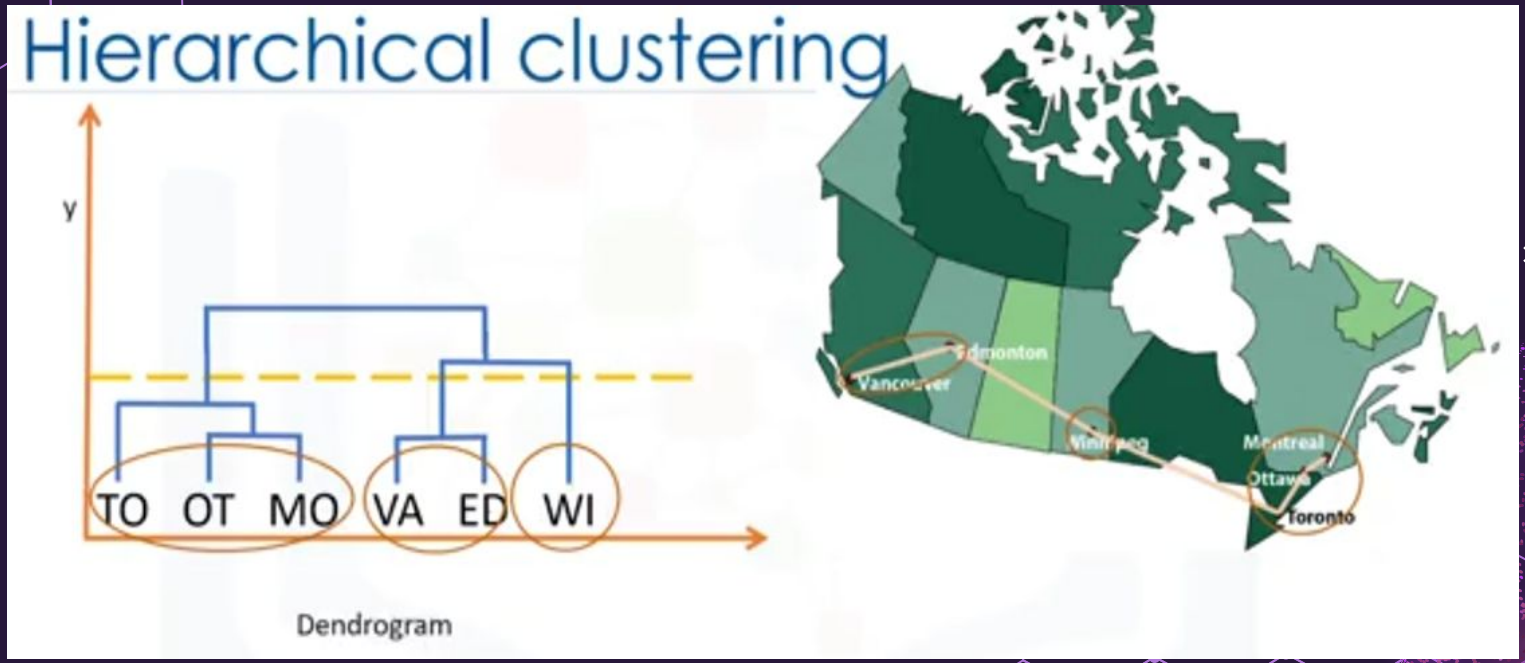
	TO/OT/MO	VA/ED	WI
TO/OT/MO		2840	1676
VA/ED			1667
WI			



	TO/OT/MO	VA/ED/WI
TO/OT/MO		1676
VA/ED/WI		

03 b) Agglomerative

Illustration



04
Quelques Applications

04

- **Marketing**
  - ·Pour identifier les profils d’achat des clients.
  - ·Pour faire des recommandations à des clients.
- **Banque**
  - ·Pour détecter les fraudes.
  - ·Pour identifier les clients fidèles.
- **Assurance**
  - ·Pour la détection des fraudes.
  - ·Pour calculer les risques.
- **Presse / Media**
  - ·Pour classifier les articles selon leurs contenus.
  - ·Pour recommander des articles.
- **Médicine**

  Pour trouver des indices selon le comportement des clients.
- **Biologie**

  Pour faire des regroupements génétiques dans le but d’identifier des familles.

Merci pour votre attention!