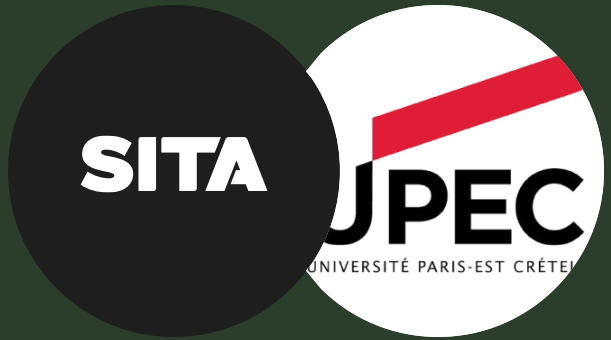


EPISEN – ING3. SI

Machine Learning

Abdallah EL HIDALI
Tech Lead Sita For Aircraft
abdallah.el-hidali@sita.aero

EPISEN
2024/2025

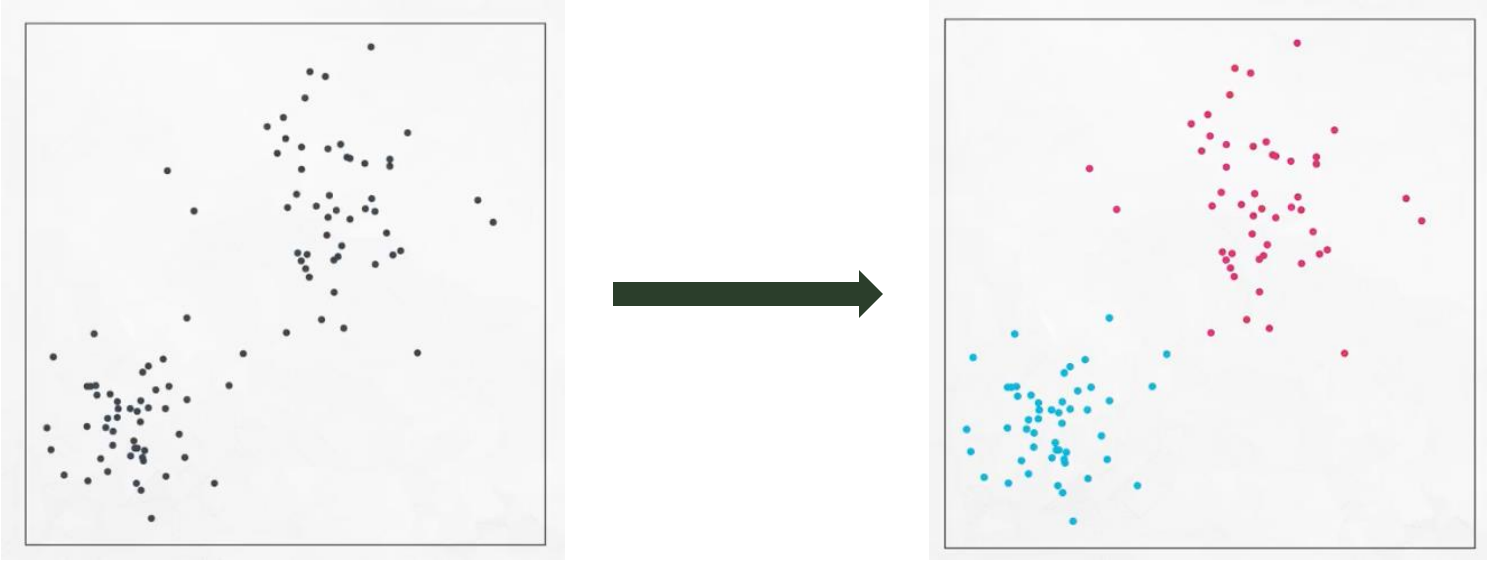
SITA

X. Clustering

Types d'apprentissage non supervisé

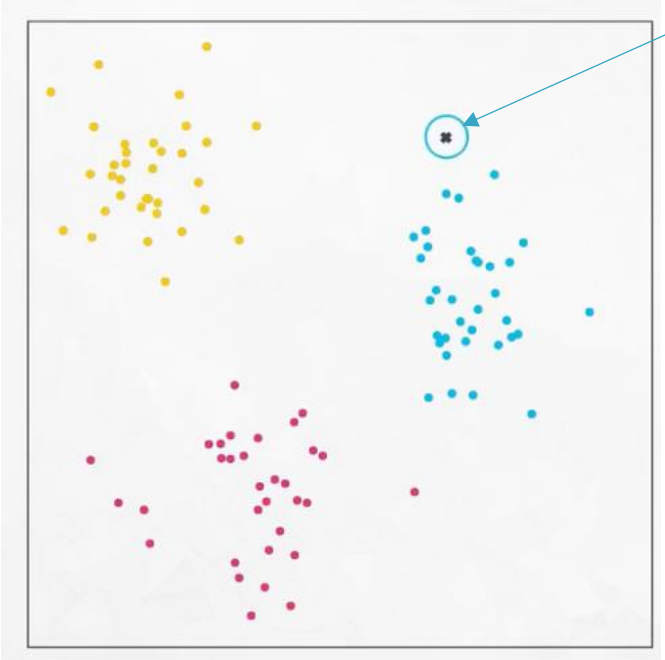
Que peut-on déduire de la distribution de ces points ?

Types d'apprentissage non supervisé



Nous pouvons diviser le nuage de points dont nous disposons en deux catégories.

Types d'apprentissage non supervisé



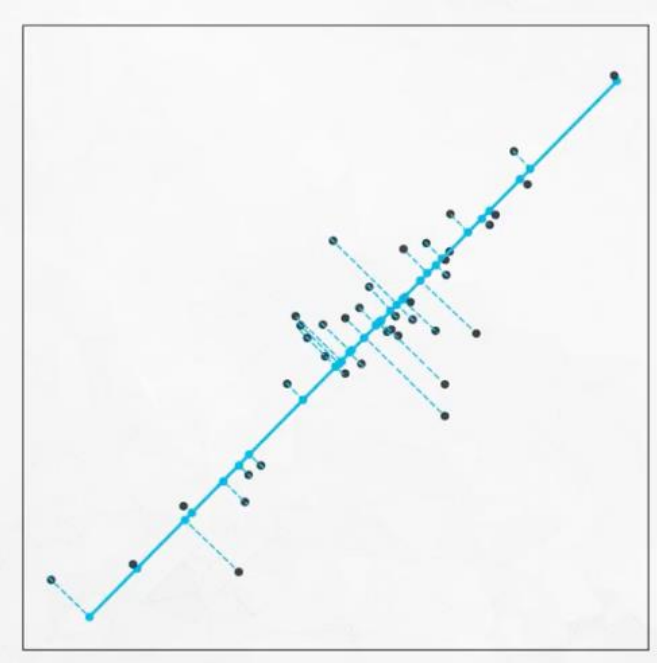
On peut attribuer ce point spécifique au nuage de points bleus

Clustering: regroupement des données en fonction des similarités.

Types d'apprentissage non supervisé

Que peut-on déduire de la distribution de ces points ?

Types d'apprentissage non supervisé



Une projection de tous les points sur la droite bleue permettrait de **réduire la dimensionnalité** du problème, facilitant ainsi son analyse

K-means

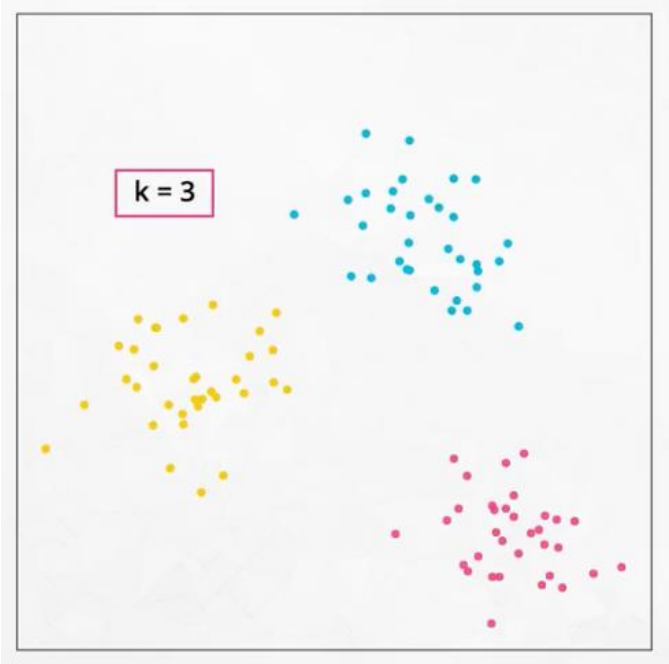
L'algorithme K-Means est utilisé pour regrouper toutes sortes de données.

Il peut regrouper ensemble :

1. Des livres de genres similaires ou écrits par les mêmes auteurs.
2. Des films similaires.
3. De la musique similaire.
4. Des groupes de clients similaires.

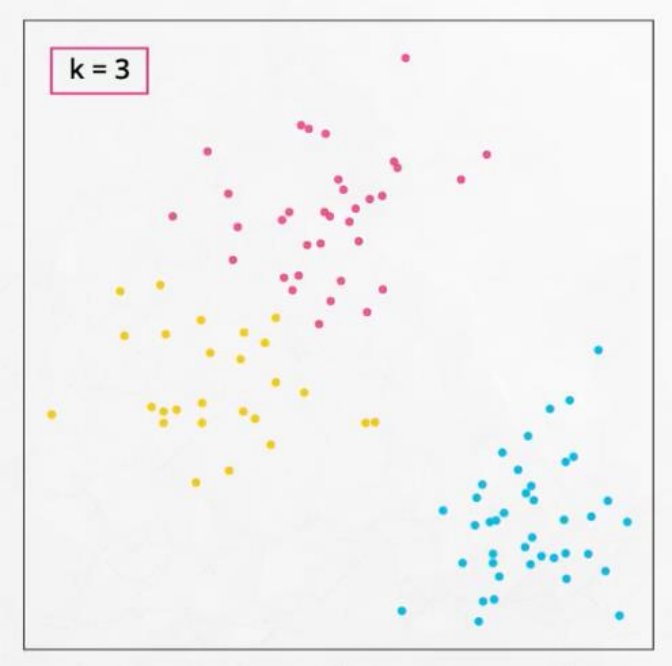
Ce regroupement peut conduire à des recommandations de produits, de films, de musique et à d'autres types de recommandations.

K-means



Le K dans K-means, représente le nombre de clusters

K-means

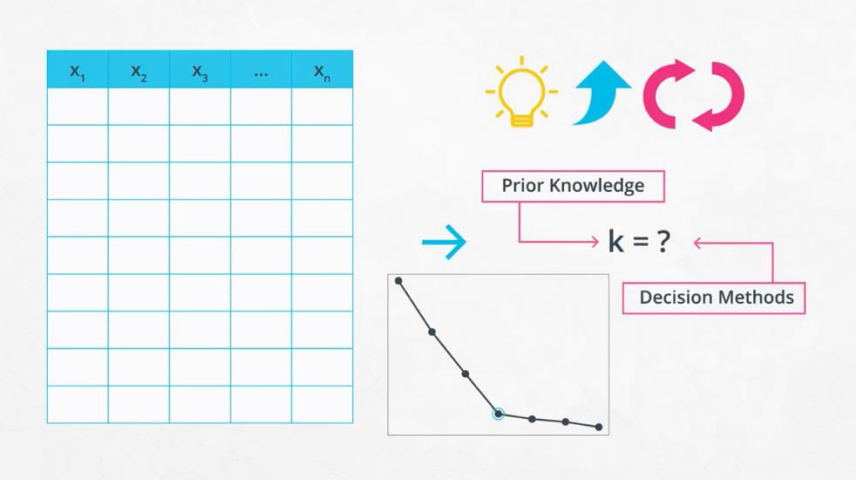


Que peut-on dire du choix du nombre de clusters k=3 dans cet exemple ?

K-means

Dans cet exemple, il est judicieux de choisir k=2 plutôt que k=3

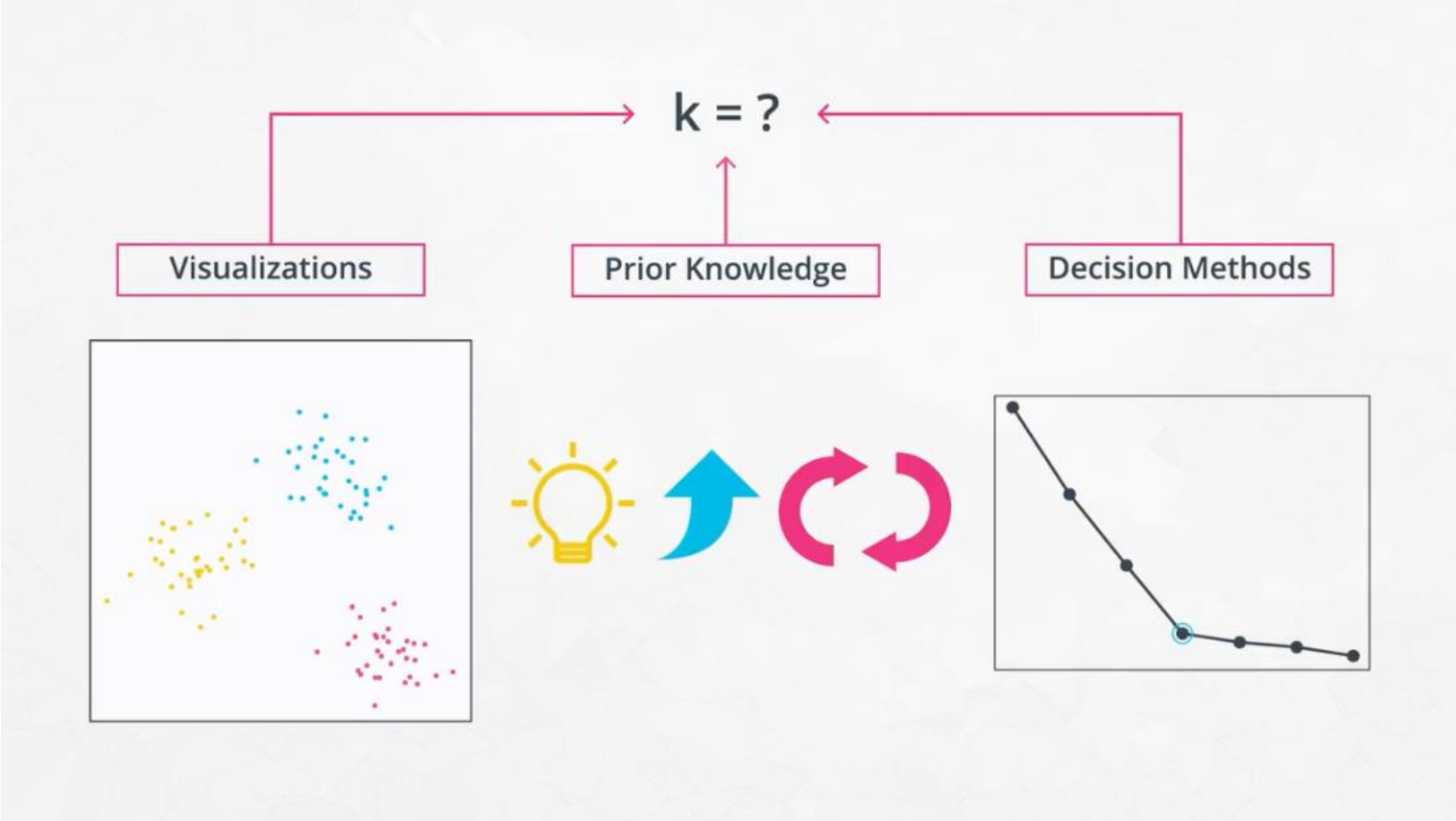
K-means



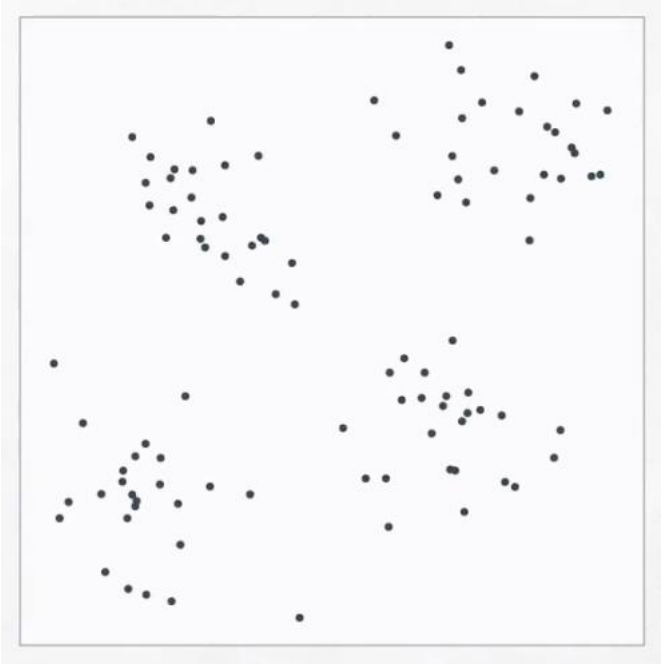
Choisir K

Jusqu'à présent, on a identifié K lorsqu'on inspecte visuellement les données pour identifier le nombre de clusters. Cependant, en pratique, on a souvent une quantité importante de données avec de nombreuses caractéristiques. Cela peut rendre la visualisation des clusters impossible.

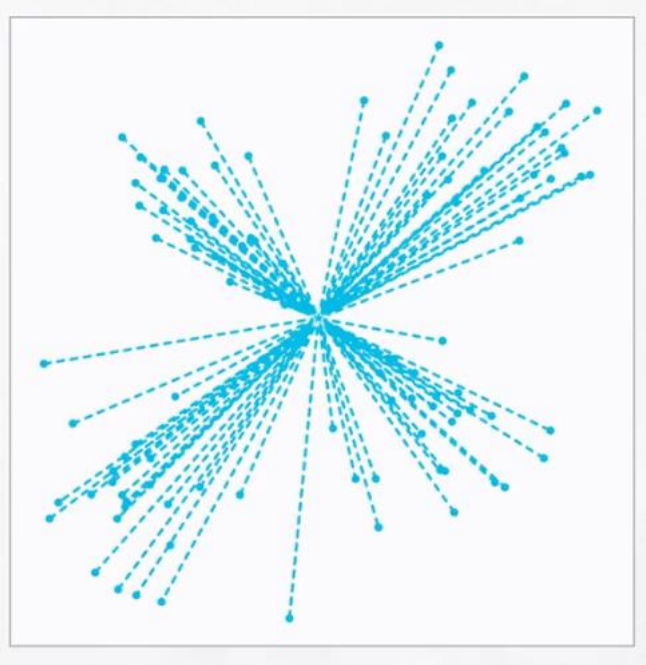
K-means: la méthode Elbow pour choisir K



K-means: la méthode Elbow pour choisir K



K-means: la méthode Elbow pour choisir K



On calcule la distance moyenne entre le nuage de points et le centroïde du cluster.

k = 1: avg. dist = 1.261

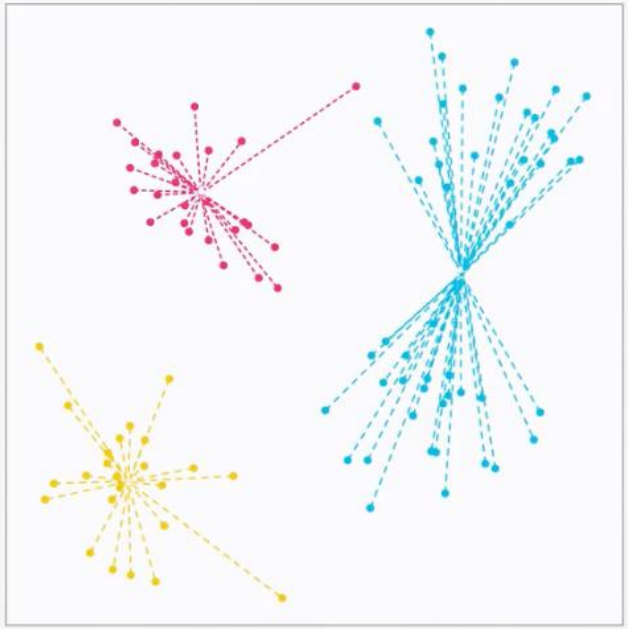
K-means: la méthode Elbow pour choisir K

Elbow Method

k = 1: avg. dist = 1.261

k = 2: avg. dist = 0.923

k = 3: avg. dist = 0.639



K-means: la méthode Elbow pour choisir K

Elbow Method

k = 1: avg. dist = 1.261

k = 2: avg. dist = 0.923

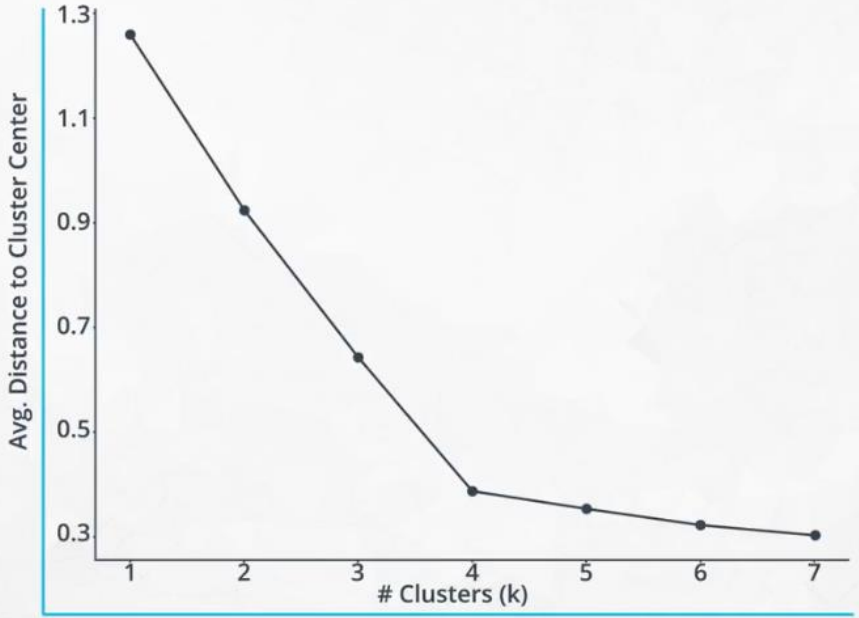
k = 3: avg. dist = 0.639

k = 4: avg. dist = 0.382

k = 5: avg. dist = 0.348

k = 6: avg. dist = 0.318

k = 7: avg. dist = 0.298



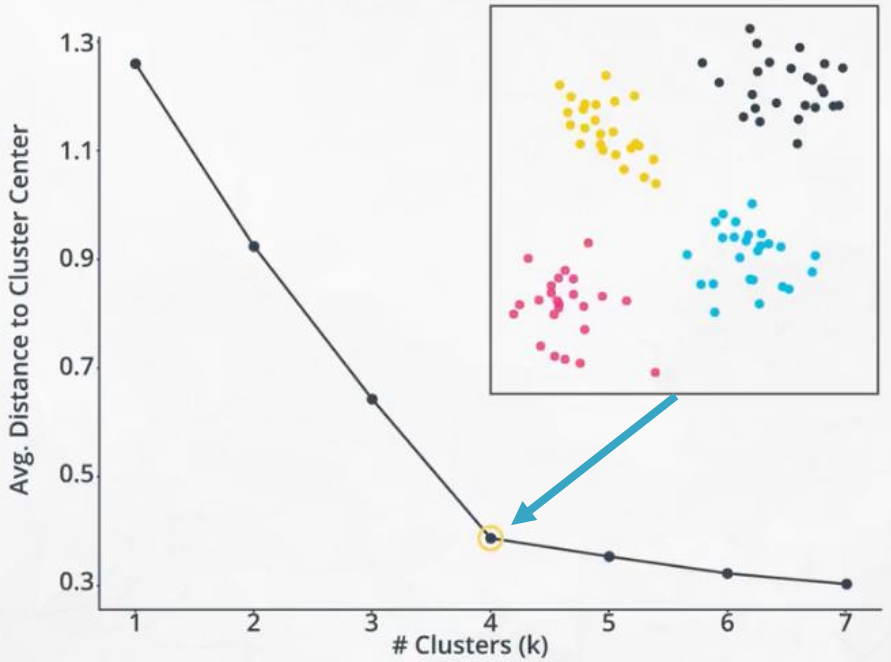
K-means: la méthode Elbow pour choisir K

Elbow Method

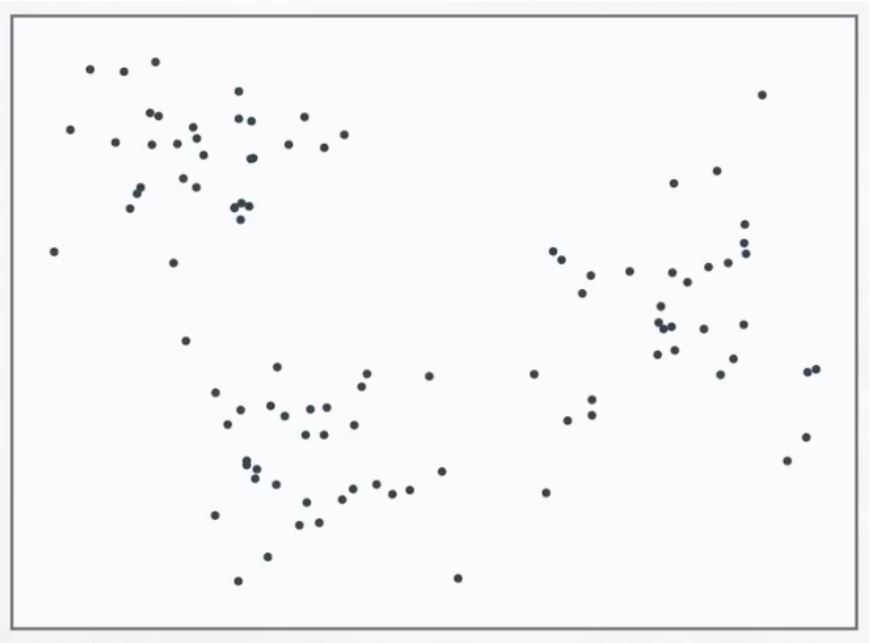
Large decreases at small k

Small decreases at large k

Best choice for k around the 'elbow' of the curve



K-means: comment ça marche



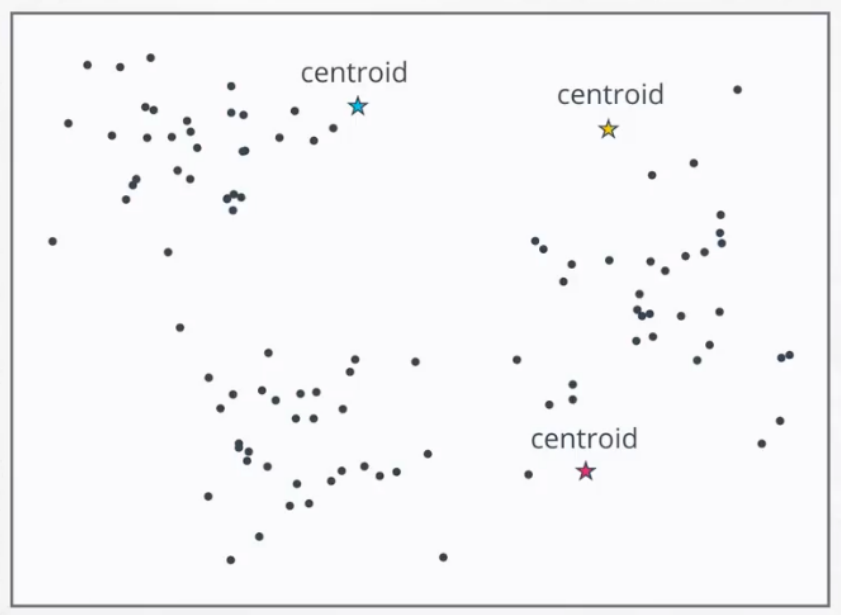
Placez aléatoirement k centroïdes parmi vos données.

Ensuite, dans une boucle jusqu'à convergence, effectuez les deux étapes suivantes :

- Assignez chaque point au centroïde le plus proche.
- Déplacez le centroïde au centre des points qui lui sont assignés.

À la fin de ce processus, vous devriez avoir k clusters de points.

K-means: comment ça marche



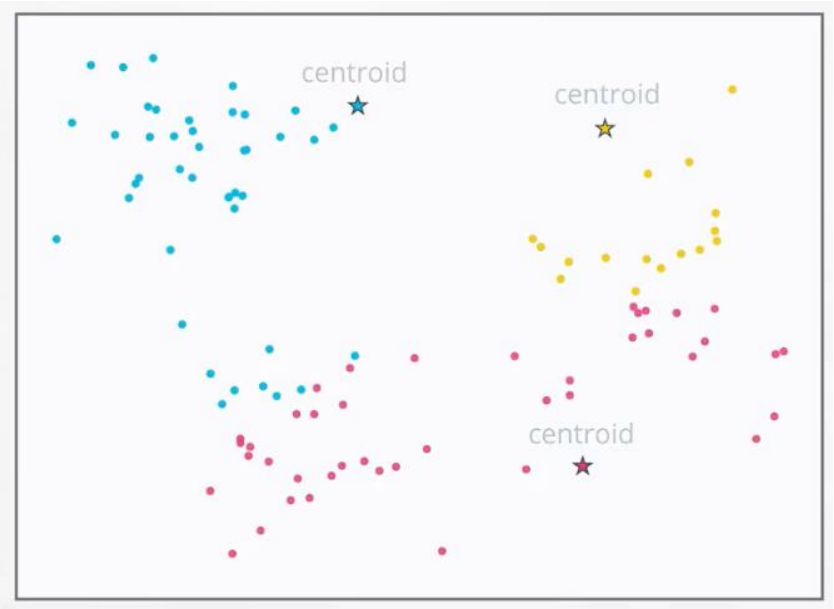
Placez aléatoirement k centroïdes parmi vos données.

Ensuite, dans une boucle jusqu'à convergence, effectuez les deux étapes suivantes :

- Assignez chaque point au centroïde le plus proche.
- Déplacez le centroïde au centre des points qui lui sont assignés.

À la fin de ce processus, vous devriez avoir k clusters de points.

K-means: comment ça marche



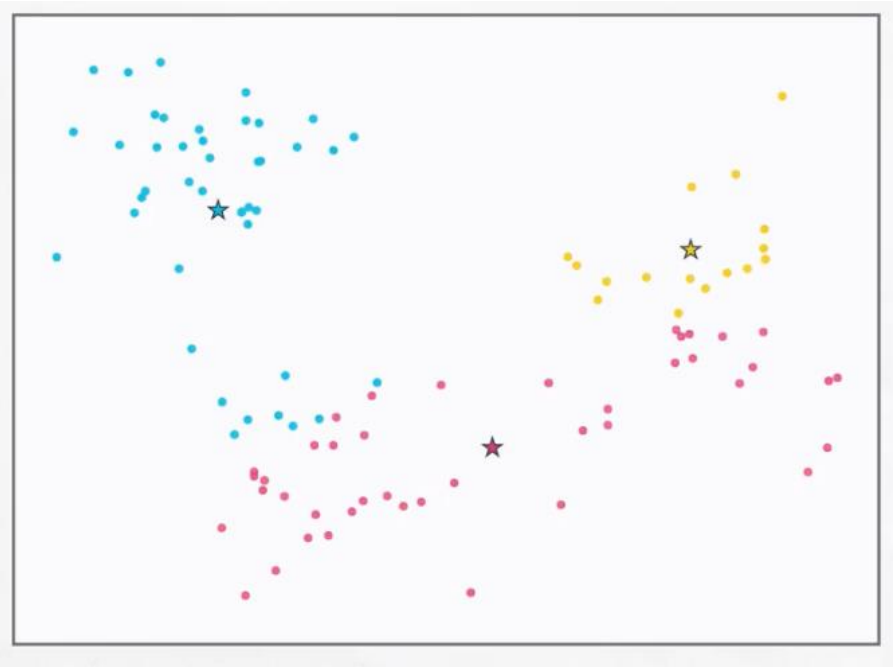
Placez aléatoirement k centroïdes parmi vos données.

Ensuite, dans une boucle jusqu'à convergence, effectuez les deux étapes suivantes :

- Assignez chaque point au centroïde le plus proche.
- Déplacez le centroïde au centre des points qui lui sont assignés.

À la fin de ce processus, vous devriez avoir k clusters de points.

K-means: comment ça marche



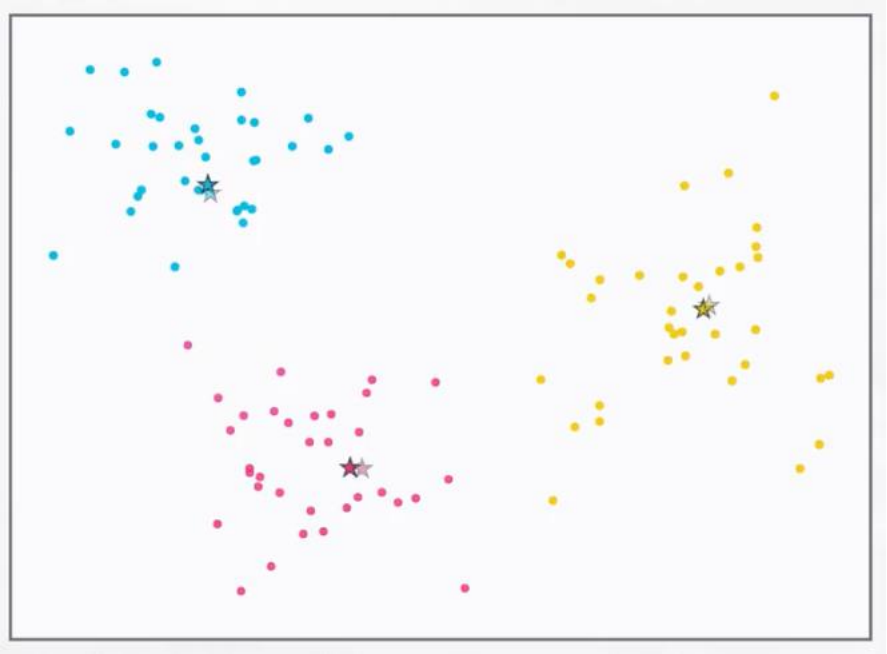
Placez aléatoirement k centroïdes parmi vos données.

Ensuite, dans une boucle jusqu'à convergence, effectuez les deux étapes suivantes :

- Assignez chaque point au centroïde le plus proche.
- Déplacez le centroïde au centre des points qui lui sont assignés.

À la fin de ce processus, vous devriez avoir k clusters de points.

K-means: comment ça marche



Placez aléatoirement k centroïdes parmi vos données.

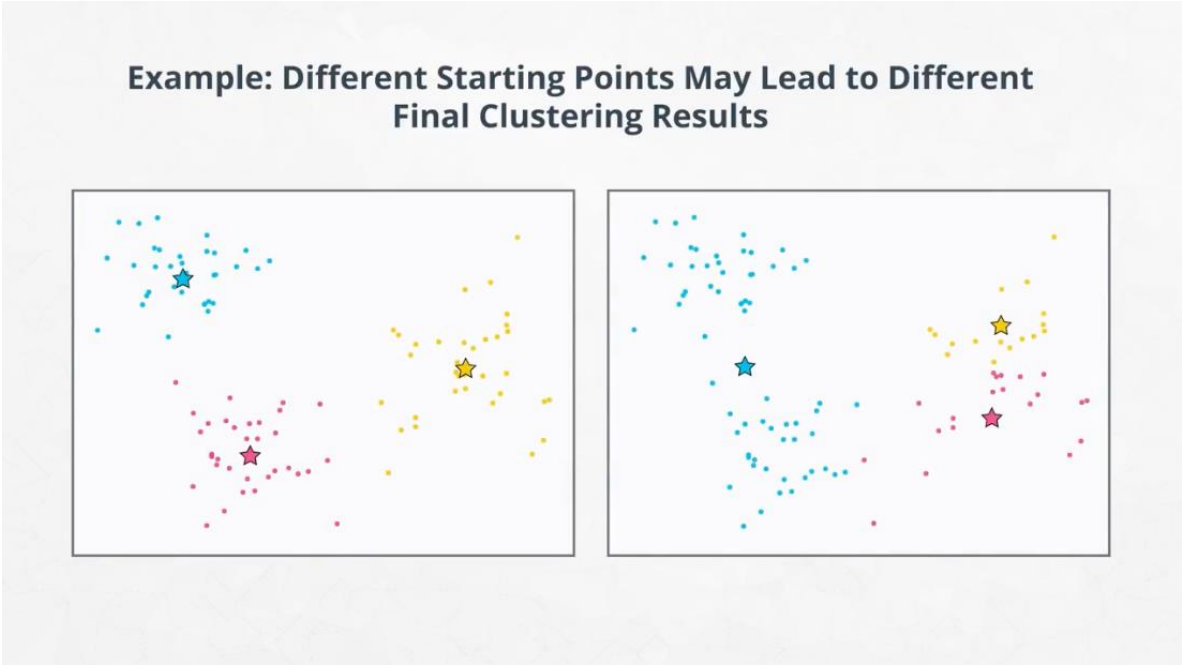
Ensuite, dans une boucle jusqu'à convergence, effectuez les deux étapes suivantes :

- Assignez chaque point au centroïde le plus proche.
- Déplacez le centroïde au centre des points qui lui sont assignés.

À la fin de ce processus, vous devriez avoir k clusters de points.

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

K-means: remarque importante

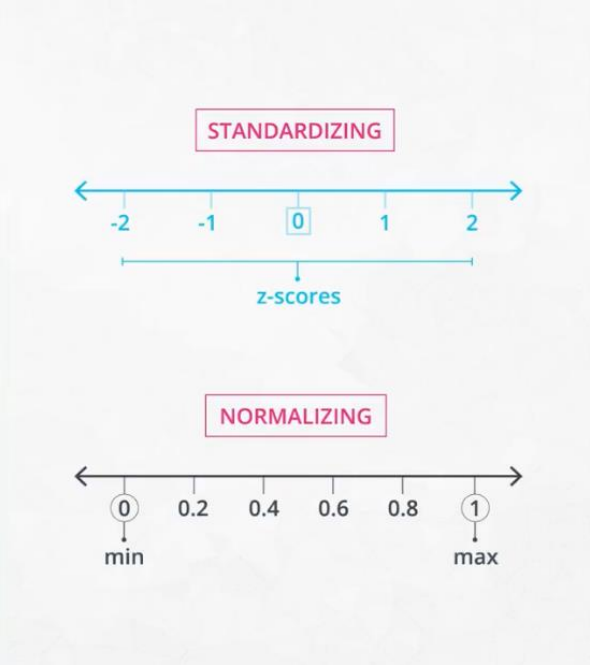


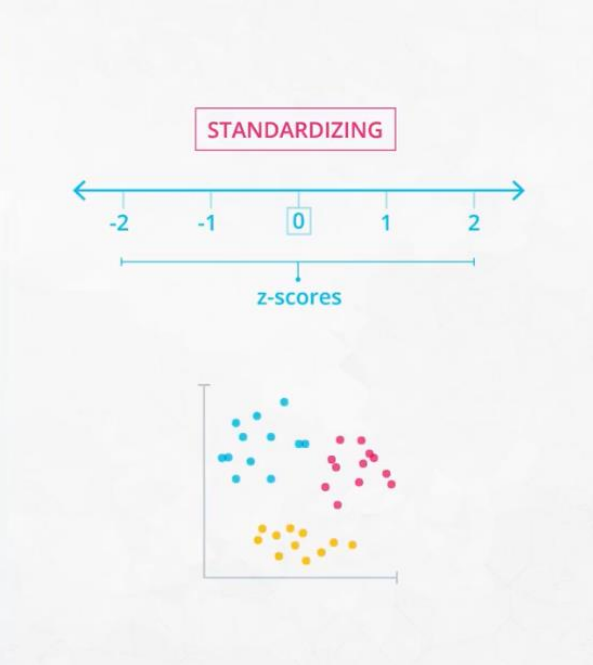
On peut avoir des clusters différents en fonction de l'initialisation aléatoire des centroïdes.

K-means: remarque importante

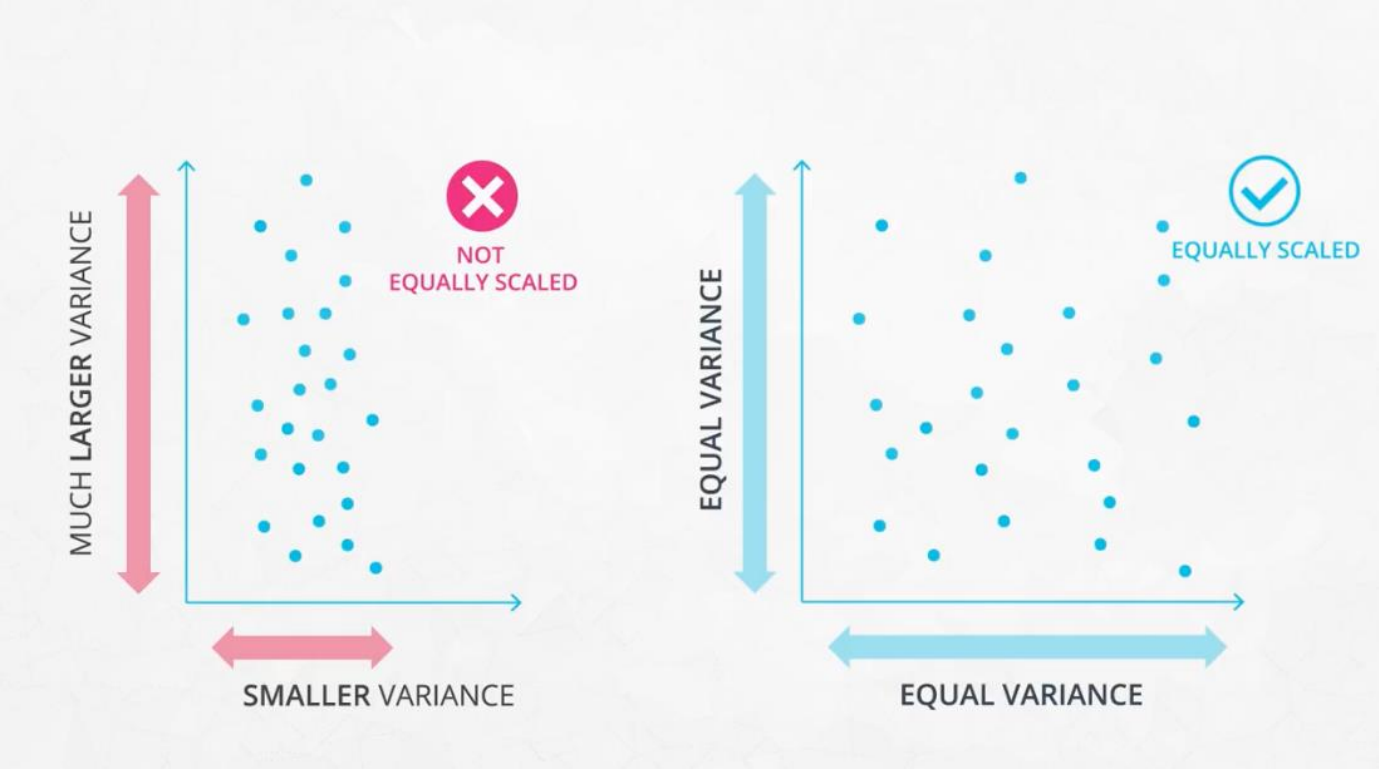
On peut minimiser ce risque en répétant l'initialisation et le clustering plusieurs fois

Scaling des variables



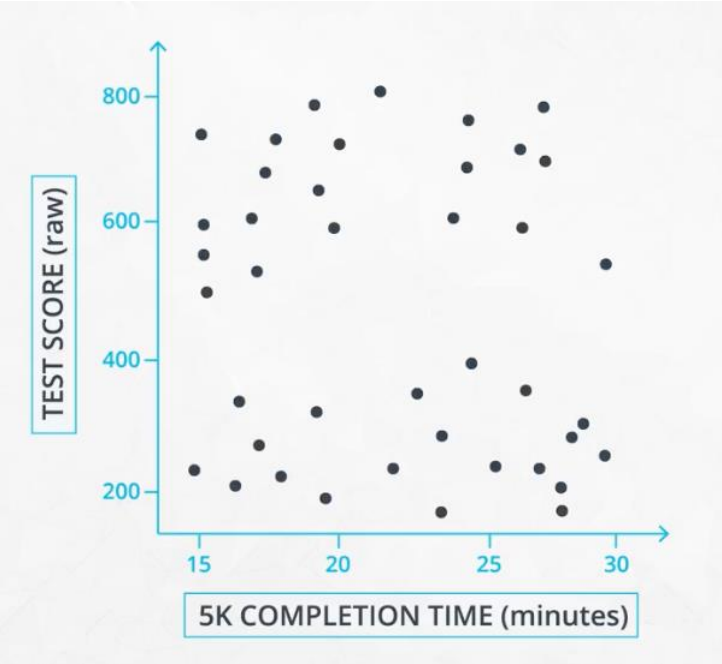


Scaling des variables

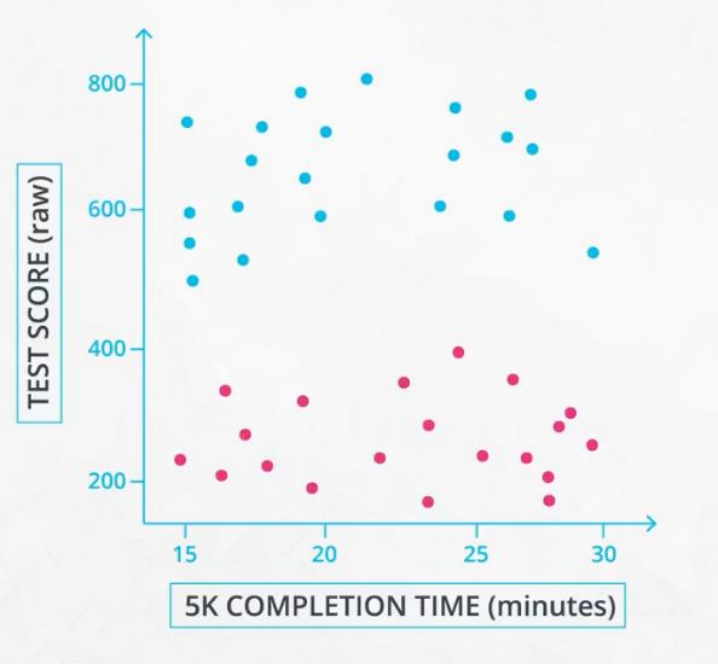


Les variables avec une variance importante vont dominer les variables avec une petite variance.

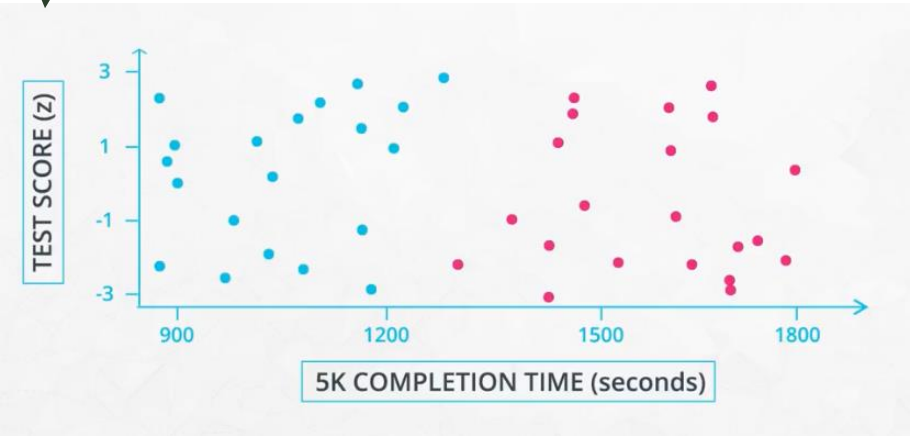
Scaling des variables



Scaling des variables



Standardisation de type z-score



Conversion de minutes en secondes

K-means: exercice

Exercice: https://github.com/elhidali/EPISEN-2024/tree/main/exercice_session_5/k-means