Le clustering

Partie 1 : La théorie

Présenté par **Morgan Gautherot**

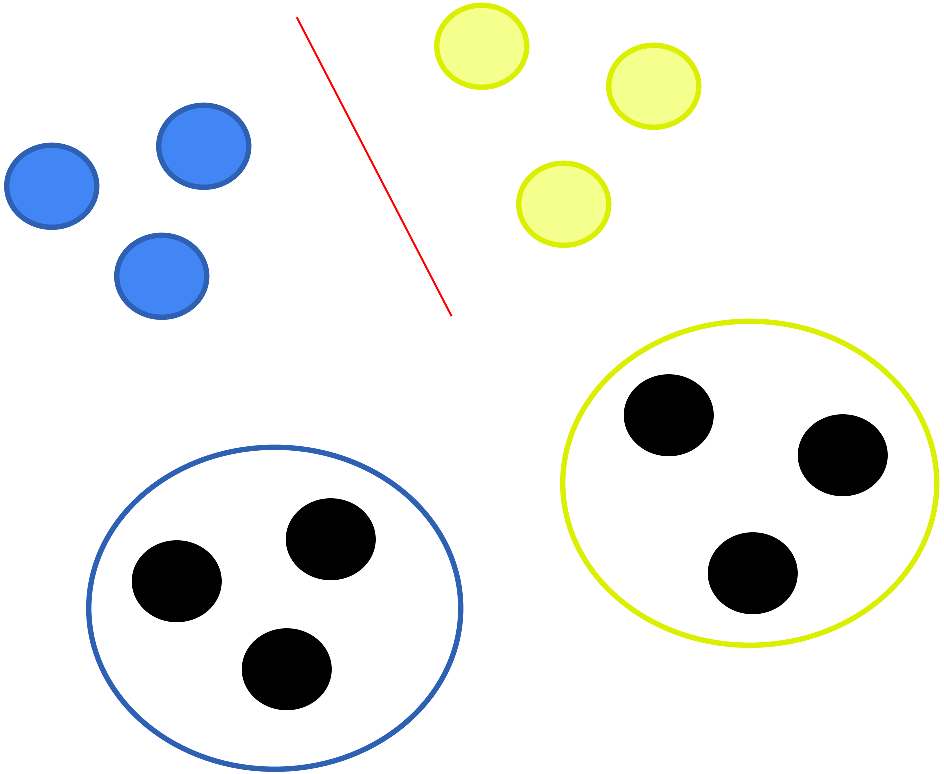
Classification vs clustering

Apprentissage supervisé - Classification
- Données labélisées (x, y)

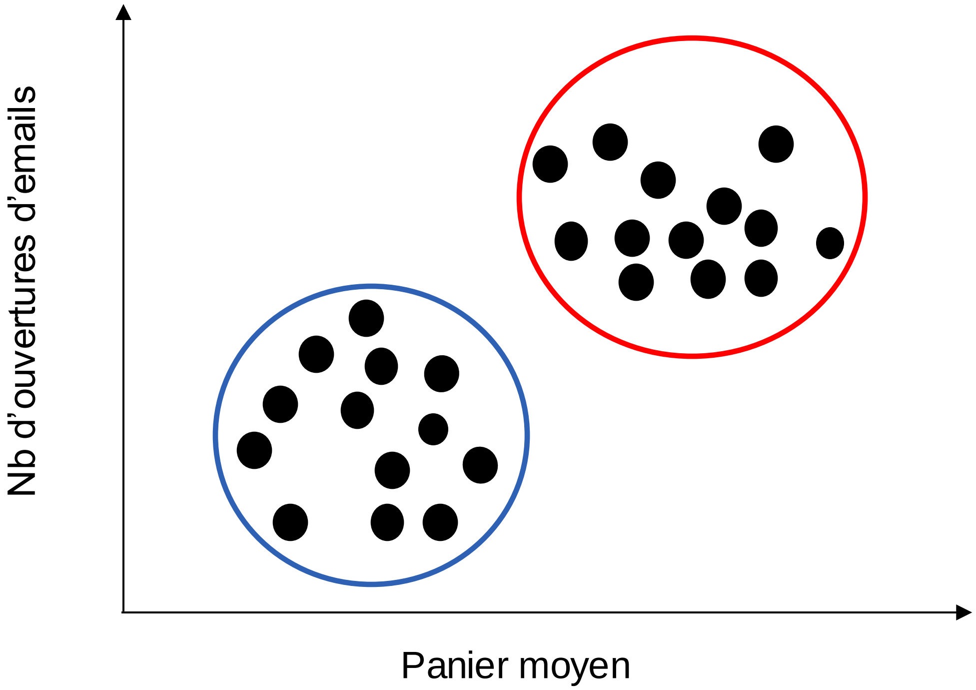
Apprendre à passer de x à y

Apprentissage non supervisé - Clustering
- Données non labélisées (x)

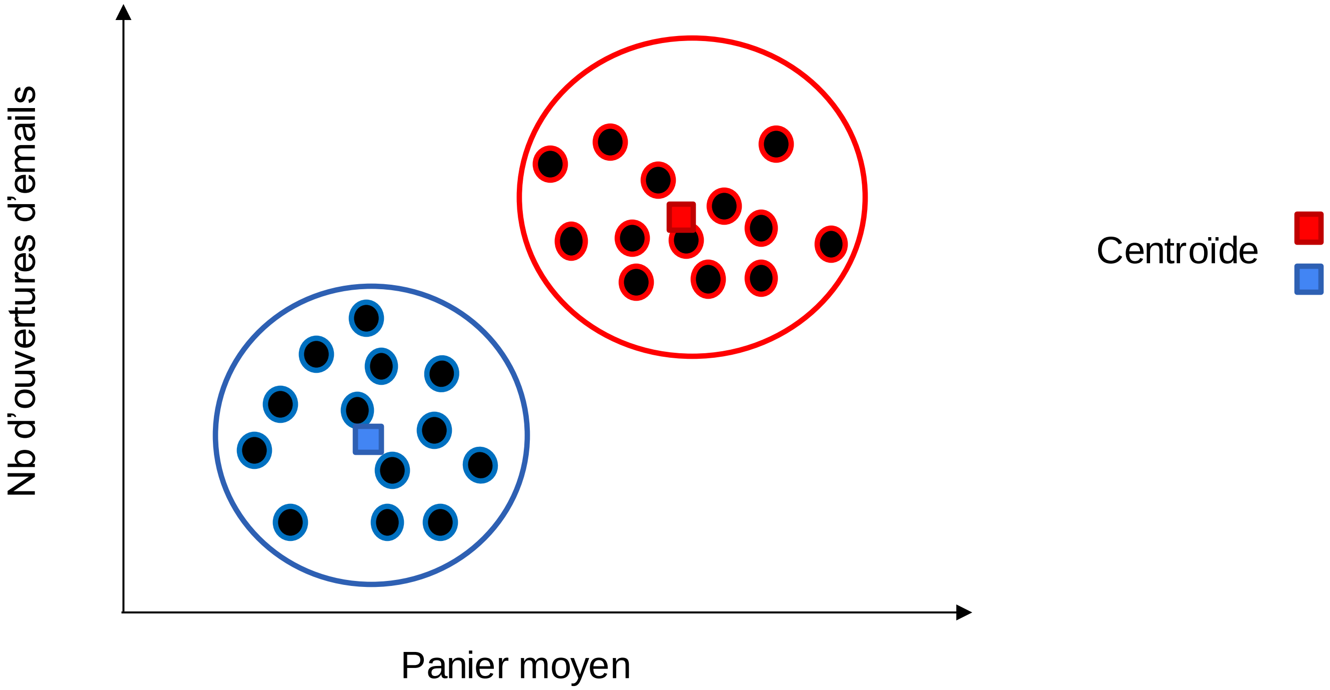
Apprendre les structures cachées



Visualisation



Visualisation



Les algorithmes de clustering

- Le hierarchical clustering
- Le K-means
- Gaussian Mixture
- DB-SCAN

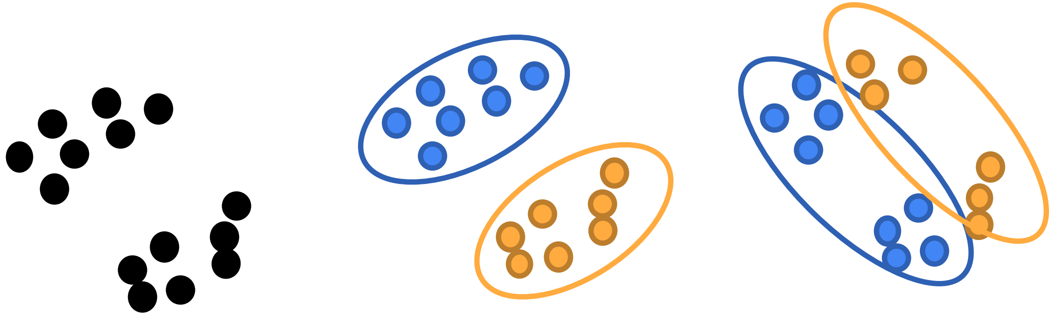
Découpe du cours

- Tout savoir sur la théorie
- Code l'algorithme from scratch
- Utilisation des sklearn

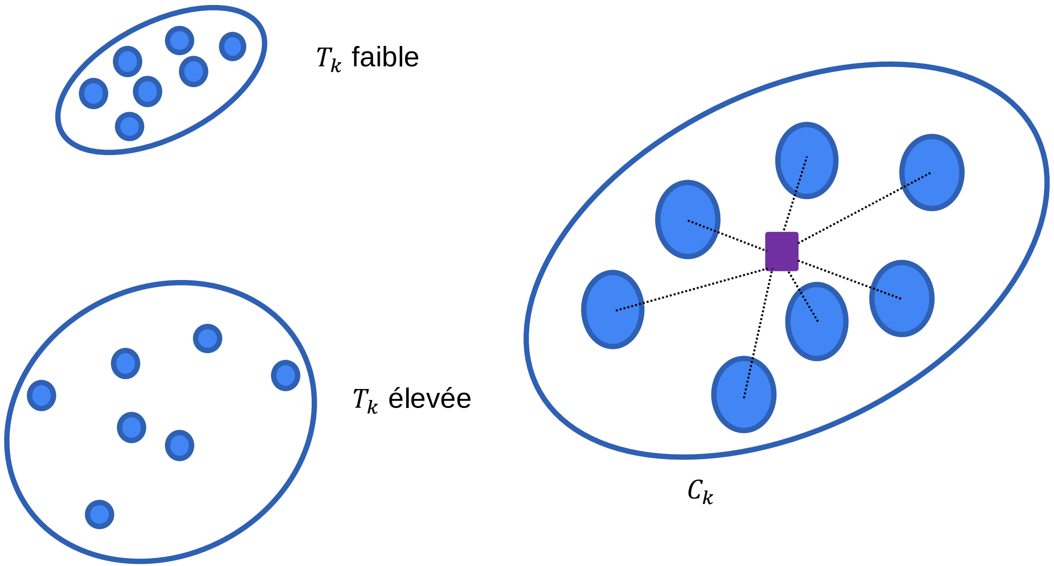
Valider un modèle de clustering

- La forme
- La stabilité
- Le cohérence

La forme



Tightness ou tension



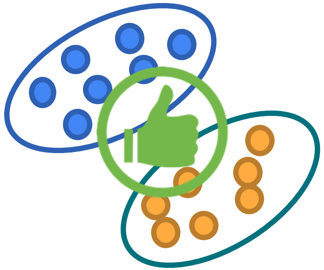
$$n_k = |C_k|$$

$$\mu_k = \frac{1}{n_k} \sum_{x_i \epsilon C_k} x_i$$

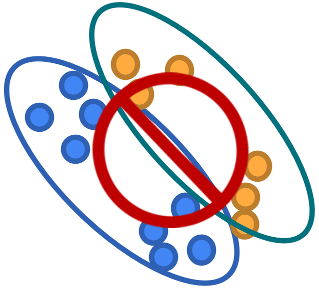
$$T_k = \frac{1}{n_k} \sum_{x \epsilon C_k} d(x, \mu_k)$$

Tightness ou tension

$$T = \frac{1}{K}\sum_{k=1}^{K} T_k$$



T faible

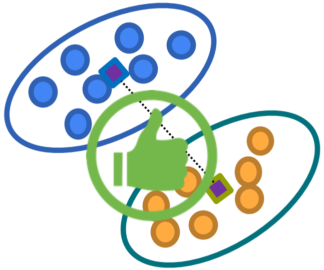


T élevée

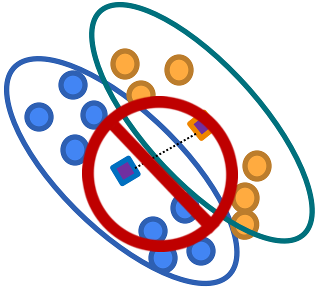
Séparation des clusters

$$S_{kl} = d(\mu_k, \mu_l)$$

$$S = \frac{2}{K(K-1)} \sum_{k=1}^{K} \sum_{l=k+1}^{K} S_{kl}$$



S élevée

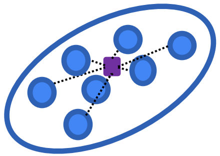


S faible

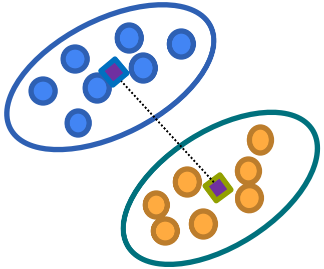
Davies-Bouldin index

$$D_k = \max_{l:l\neq k}\frac{T_k + T_l}{S_{kl}}$$

$$DB = \frac{1}{K}\sum_{k=1}^{K} D_k$$



T



S

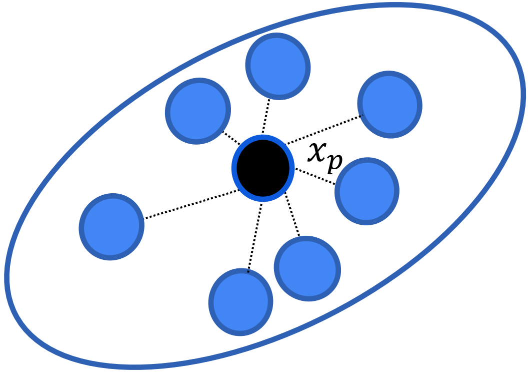
Le coefficient de silhouette

$s \epsilon [-1, 1]$

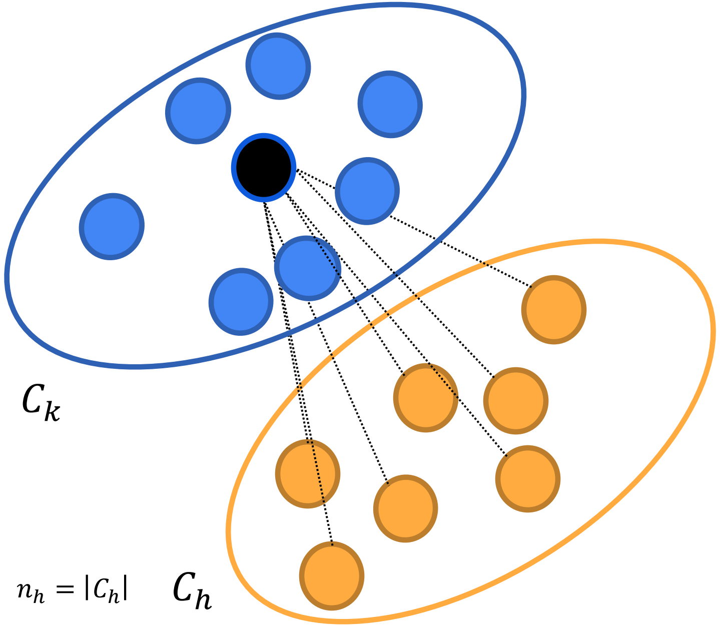
$$a = \frac{1}{n_k} \sum_{i \epsilon C_k} d(x_p, x_i)$$

$$b = \frac{1}{n_h} \sum_{i \epsilon C_h} d(x_p, x_i)$$

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



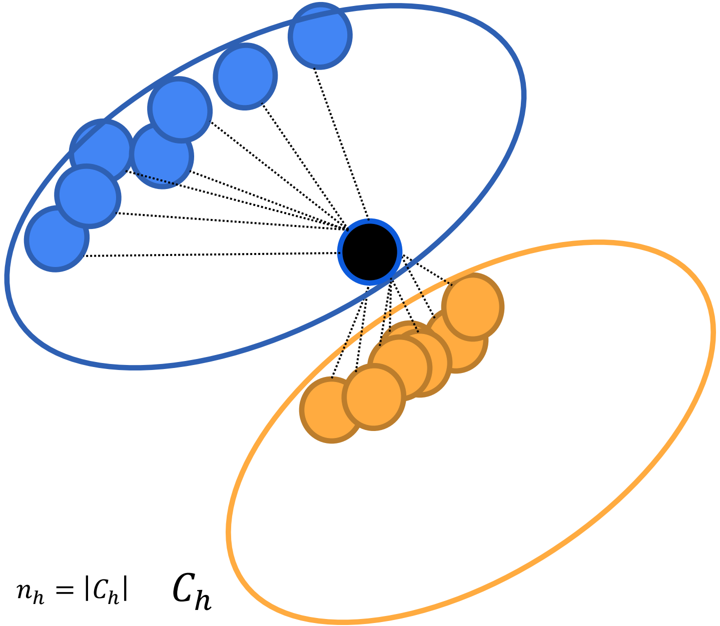
$n_k = |C_k|$ C_k



C_k

$n_h = |C_h|$ C_h

Le coefficient de silhouette



$n_k = |C_k|$ C_k

$n_h = |C_h|$ C_h

$$a = \frac{1}{n_k} \sum_{i \epsilon C_k} d(x_p, x_i)$$

$$a = 10$$

$$b = \frac{1}{n_h} \sum_{i \epsilon C_h} d(x_p, x_i)$$

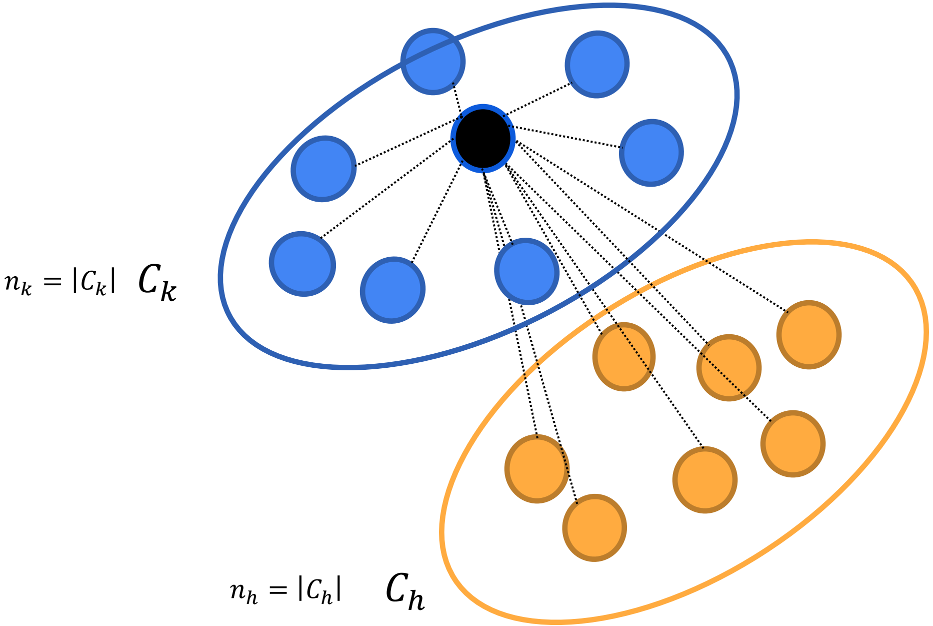
$$b = 3$$

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

$$s = \frac{3 - 10}{10}$$

$$s = \frac{-7}{10} = -0{,}7$$

Le coefficient de silhouette



$$a = \frac{1}{n_k} \sum_{i \epsilon C_k} d(x_p, x_i)$$

$$a = 3$$

$$b = \frac{1}{n_h} \sum_{i \epsilon C_h} d(x_p, x_i)$$

$$b = 10$$

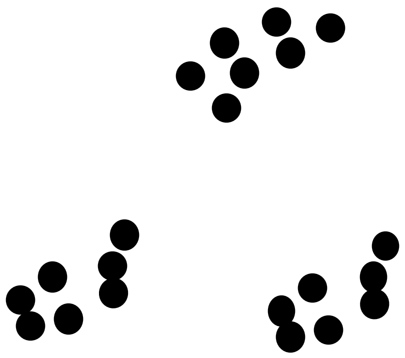
$$s = \frac{b - a}{\max(a, b)}$$

$$s = \frac{10 - 3}{10}$$

$$s = \frac{7}{10} = 0{,}7$$

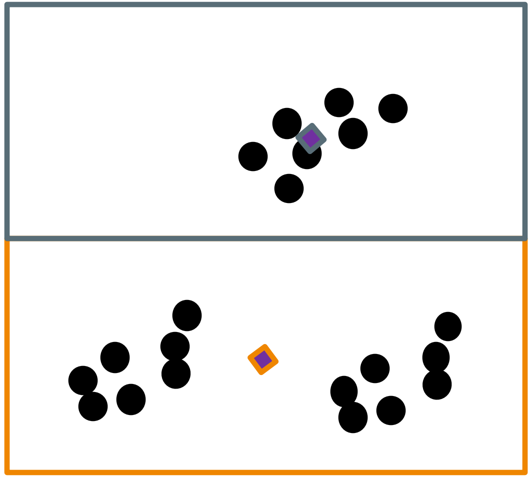
La stabilité

K = 2



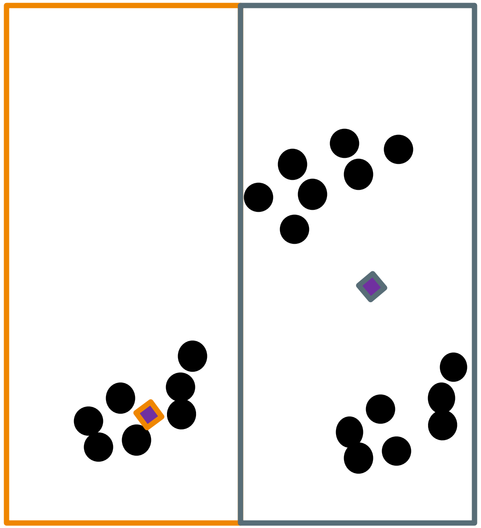
La stabilité

K = 2



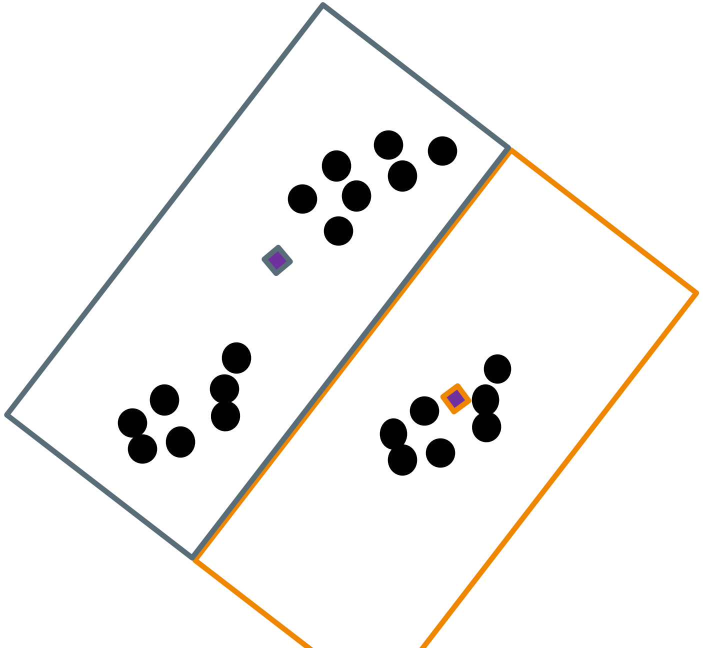
La stabilité

K = 2



La stabilité

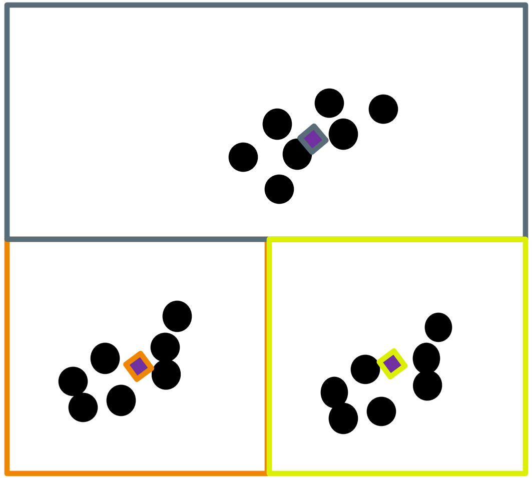
K = 2 Instable



La stabilité

K = 2 Instable

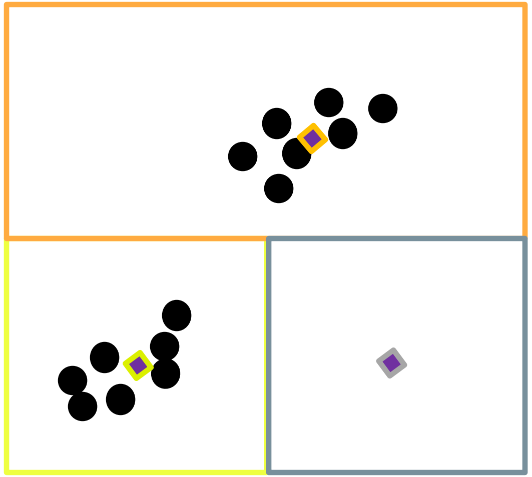
K = 3



La stabilité

K = 2 Instable

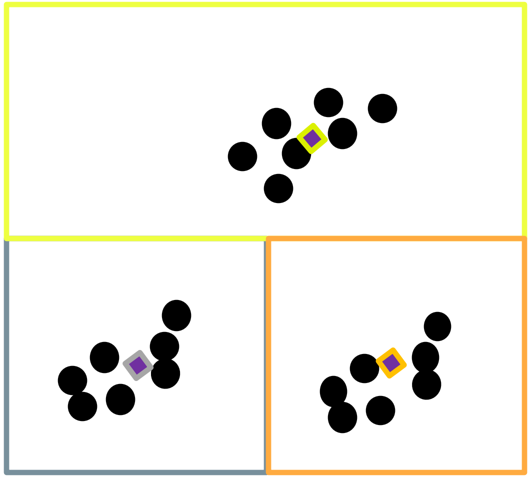
K = 3



La stabilité

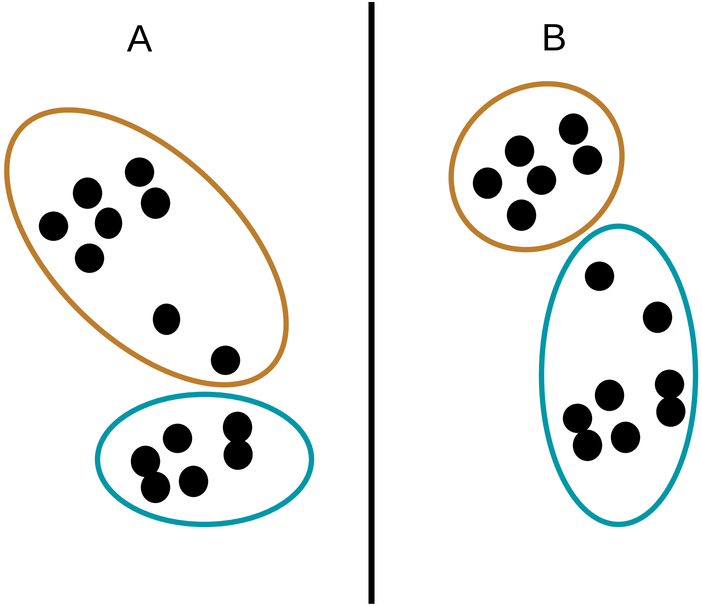
K = 2 Instable

K = 3 Stable

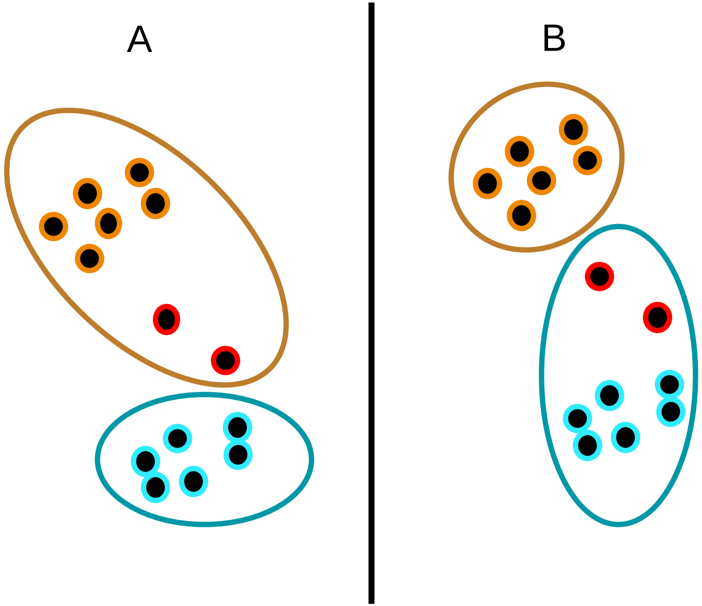


Rand index

A vs B



Rand index



A vs B

$$Rand\ index = \frac{nb\ dans\ la\ même\ classe}{nb\ total\ d'observations} = \frac{12}{14}$$

Cohérence

- Utilisez les connaissances métiers de vos collaborateurs pour vérifier la pertinence du cluster.

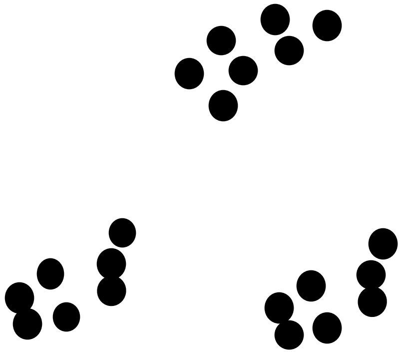
Cas d’application

Personas

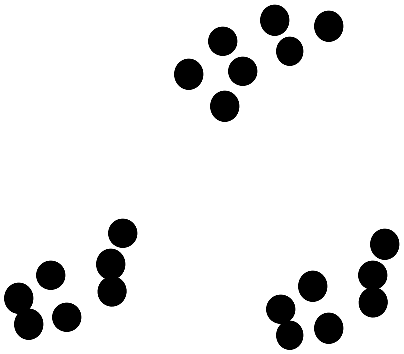
Cluster 1	Agé de plus de 50 ans, achète peu mais des gros montants
Cluster 2	Agé de moins de 20 ans, achète beaucoup mais des petits montants
Cluster 3	Agé de moins de 30 ans, achète beaucoup et des gros montants

Détermination du nombre de classes

K = 2



K = 3

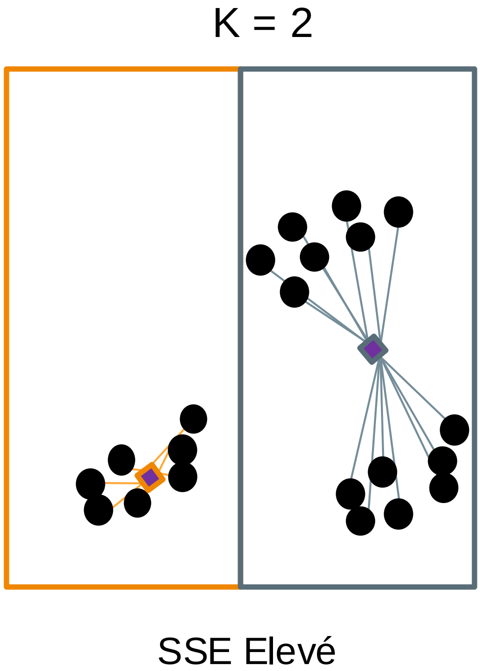


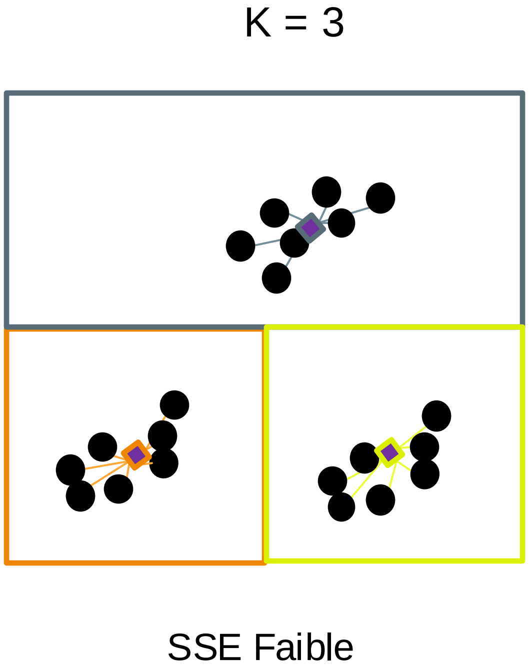
Distortion ou Sum of Square Error (SSE)

$$SSE = \sum_j \sum_i D(c_j, x_i)^2$$

Avec :
- c_j : Le centre du cluster (centroïd)
- x_i : la ième observation dans le cluster ayant pour centroïd c_j
- $D(c_j, x_i)$: La distance entre le centre du cluster et le point x_i

Détermination du nombre de classes





Méthode du coude