Clustering

Presented by **Morgan Gautherot**

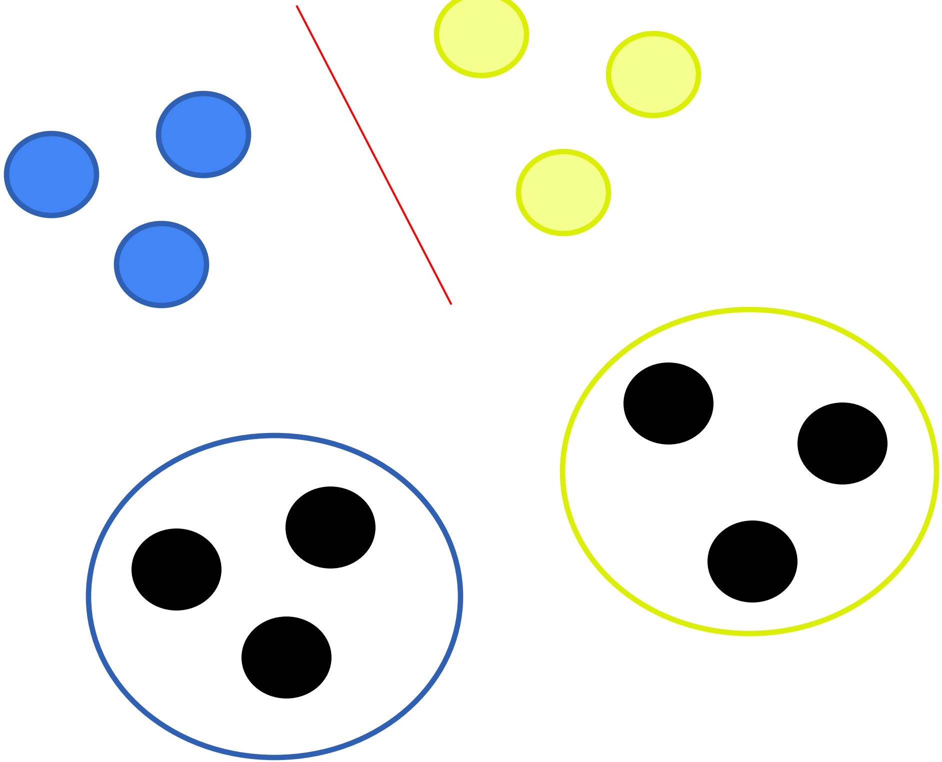
Classification vs clustering

Supervised learning - Classification
- Labeled data (x, y)

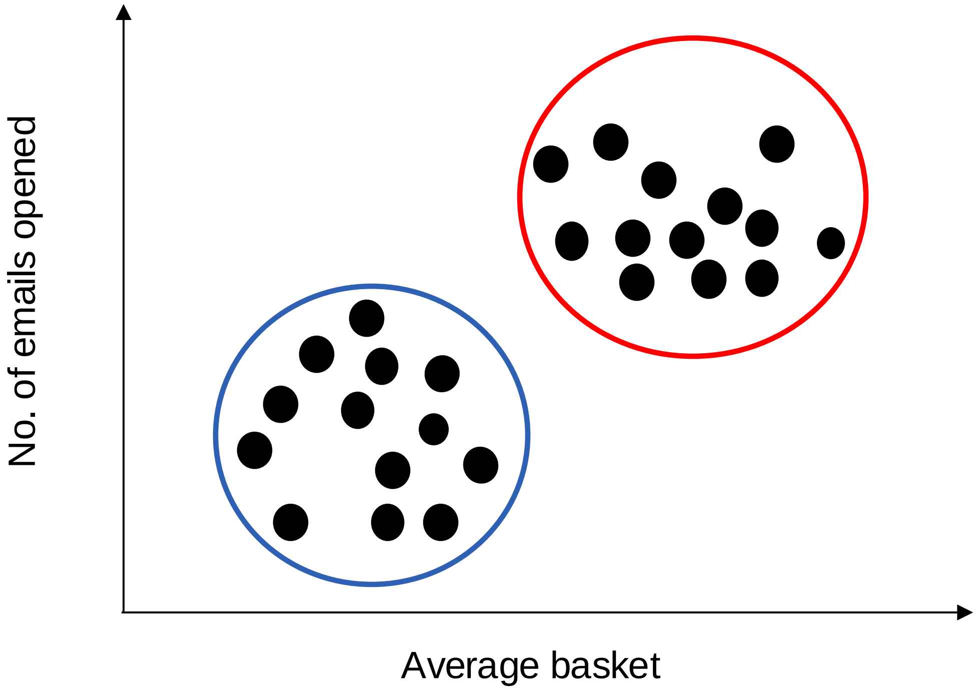
Learn to go from x to y

Unsupervised learning - Clustering
- Unlabeled data (x)

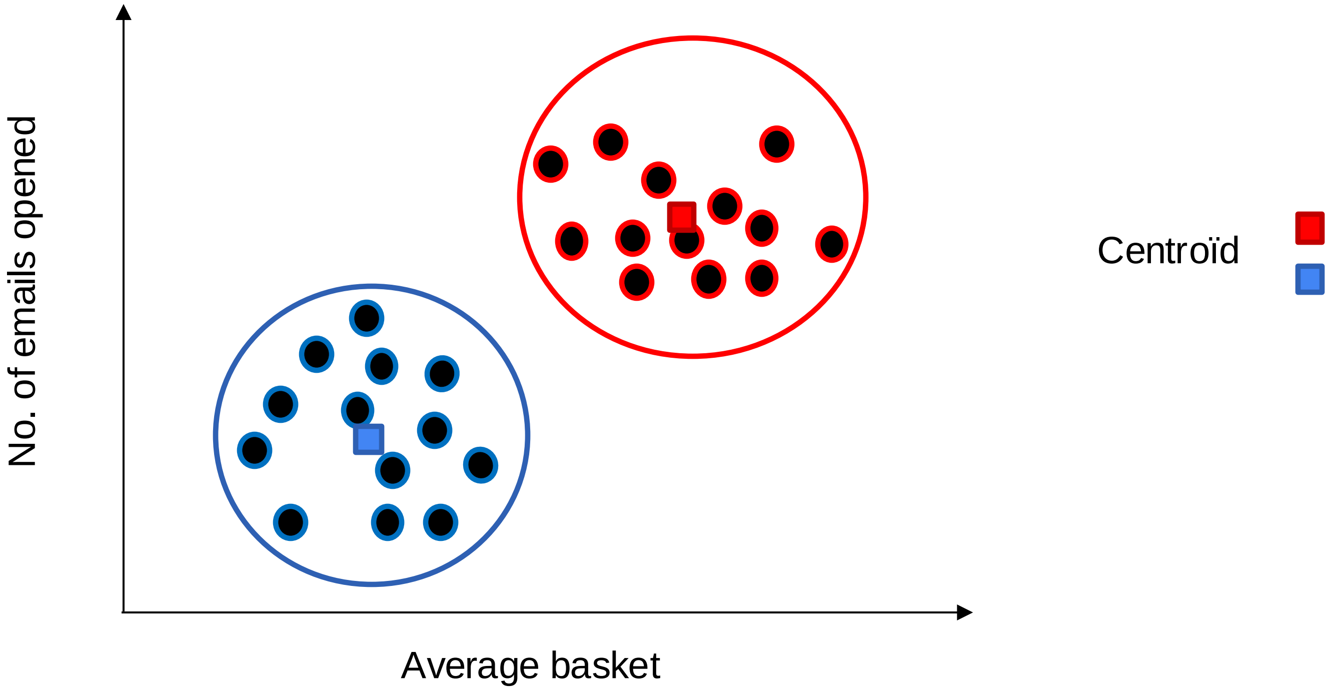
Learning hidden structures



Visualization



Visualization



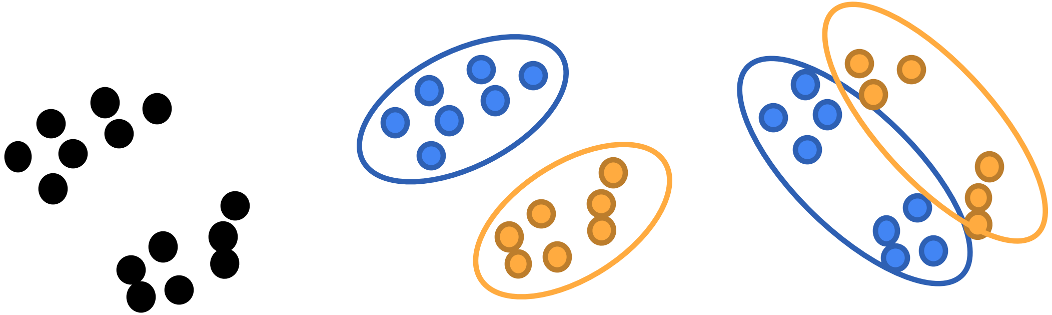
Clustering algorithms

- Hierarchical clustering
- K-means
- Gaussian Mixture
- DB-SCAN

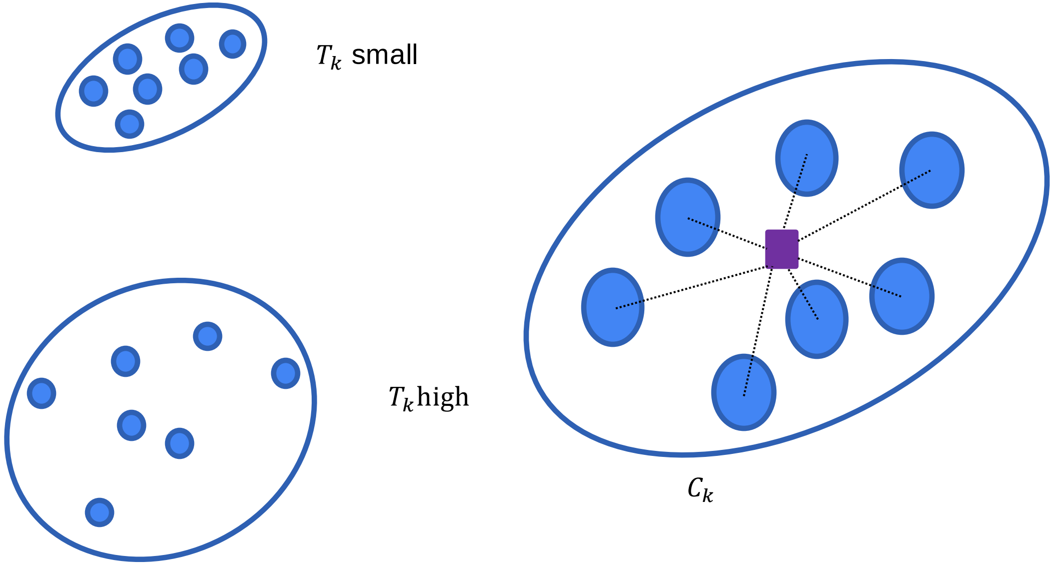
Validating a clustering model

- Shape
- Stability
- Consistency

Shape



Tightness



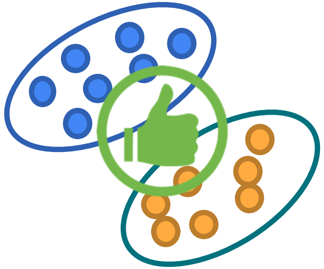
$$n_k = |C_k|$$

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

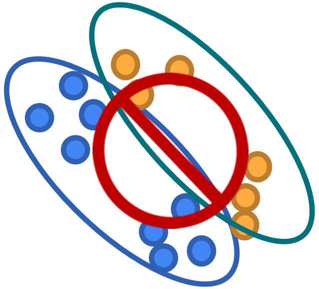
$$T_k = \frac{1}{n_k} \sum_{x \epsilon C_k} d(x, \mu_k)$$

Tightness

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



T small

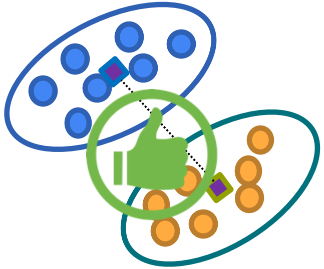


T high

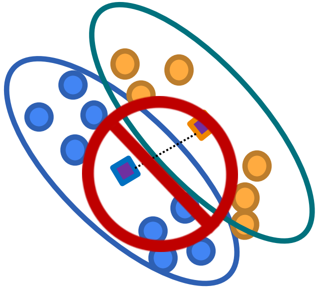
Cluster separation

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

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



S high

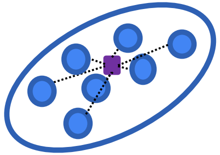


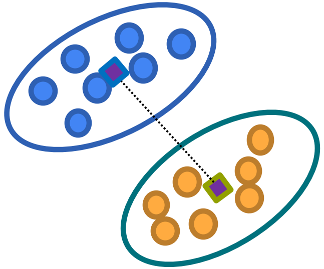
S small

Davies-Bouldin index

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

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





Silhouette score

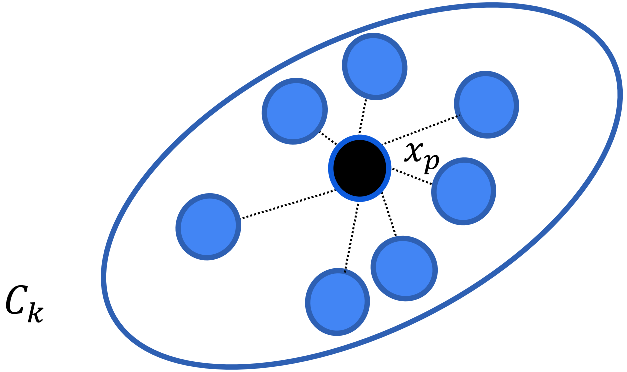
$$s \epsilon [-1, 1]$$

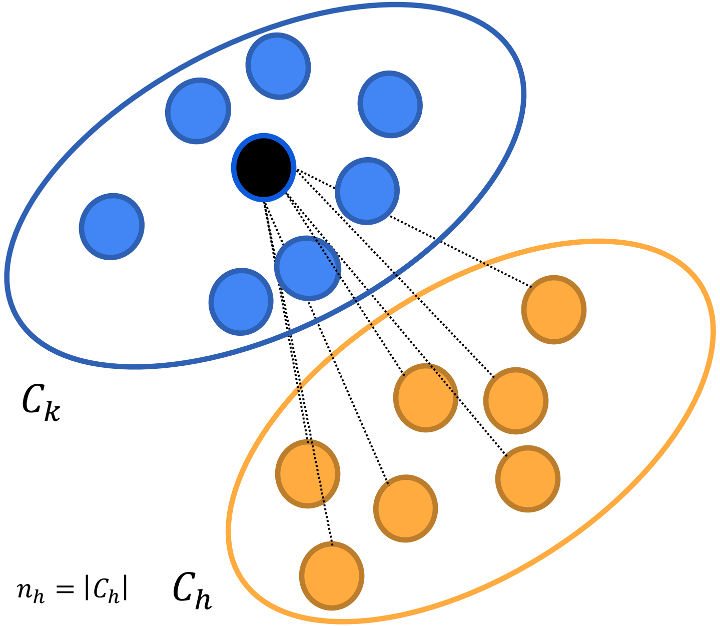
$$a = \frac{1}{n_k} \sum_{i \epsilon C_k} d(x_p, x_i)$$

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

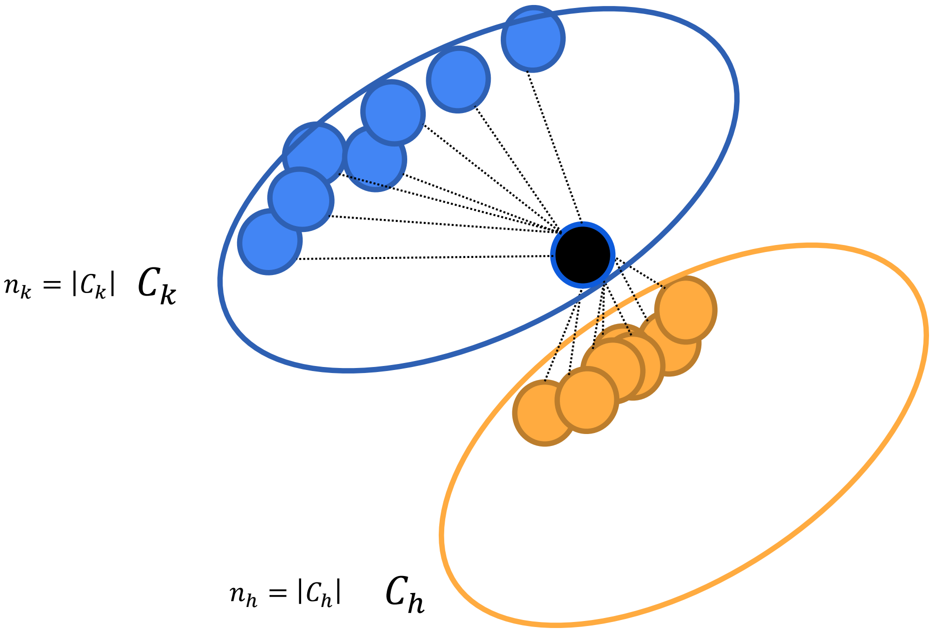
$$s = \frac{b - a}{\max(a, b)}$$

$n_k = |C_k|$





Silhouette score



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

$$a = 10$$

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

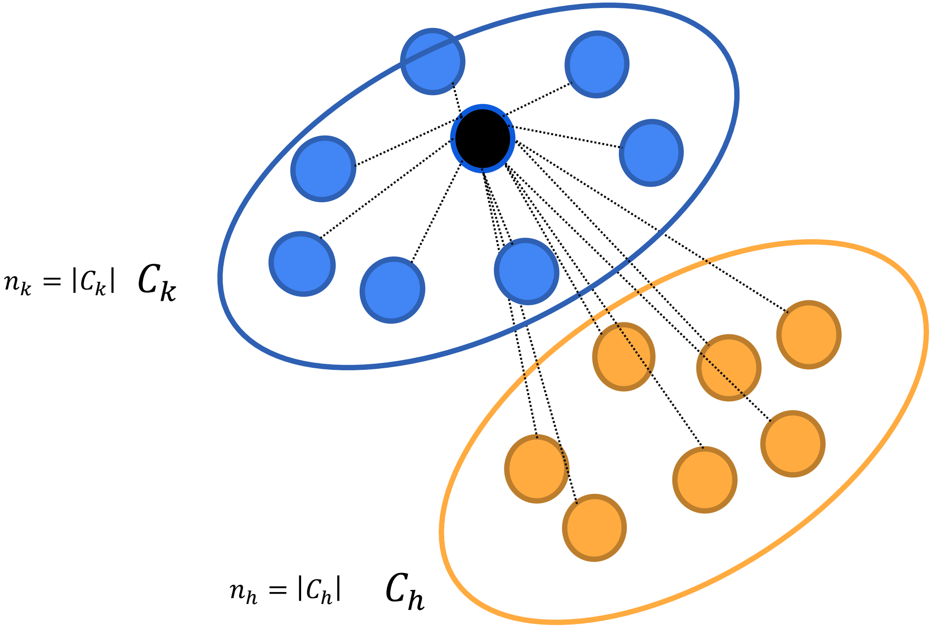
$$b = 3$$

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

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

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

Silhouette score



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

$$a = 3$$

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

$$b = 10$$

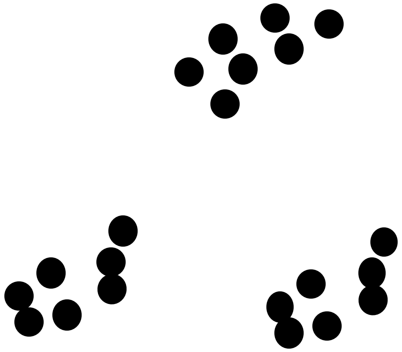
$$s = \frac{b - a}{\max(a, b)}$$

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

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

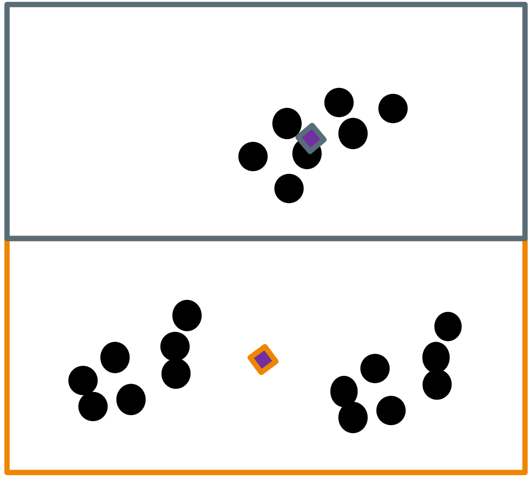
Stability

K = 2



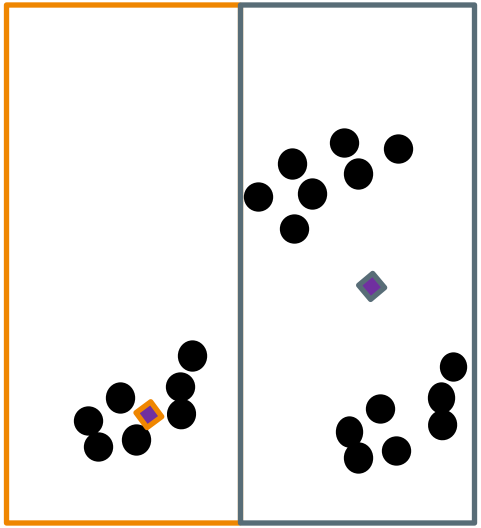
Stability

K = 2



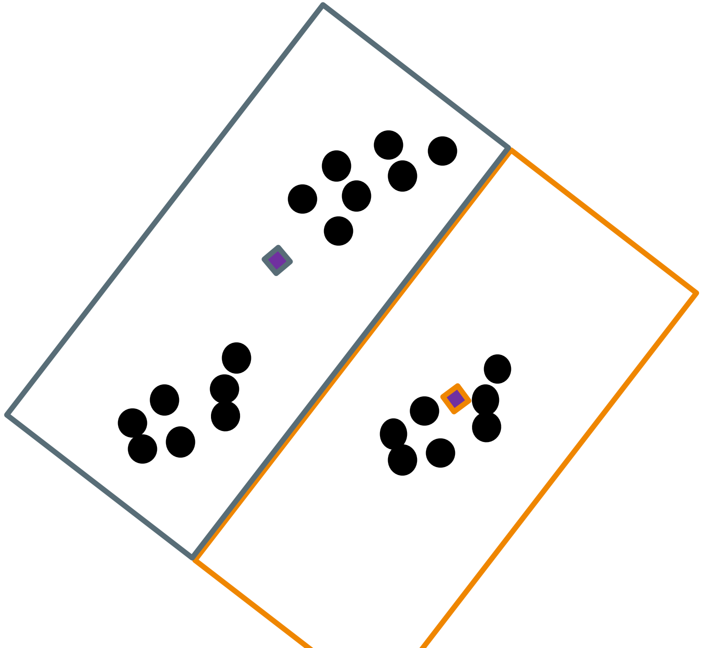
Stability

K = 2



Stability

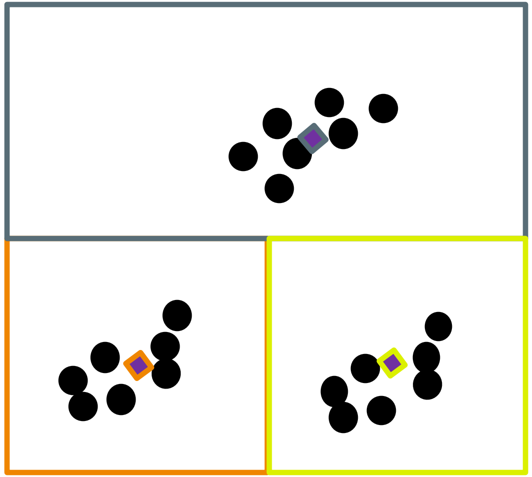
K = 2 Instable



Stability

K = 2 Instable

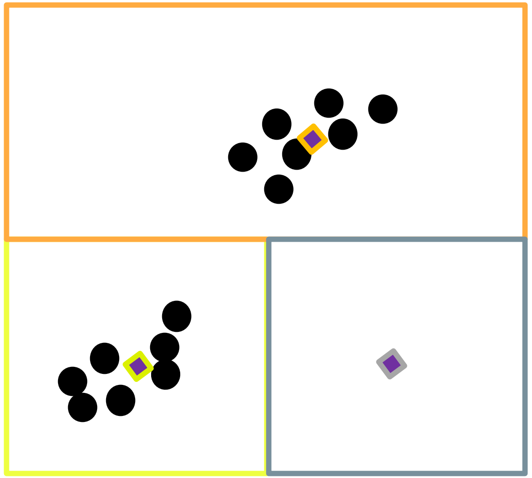
K = 3



Stability

K = 2 Instable

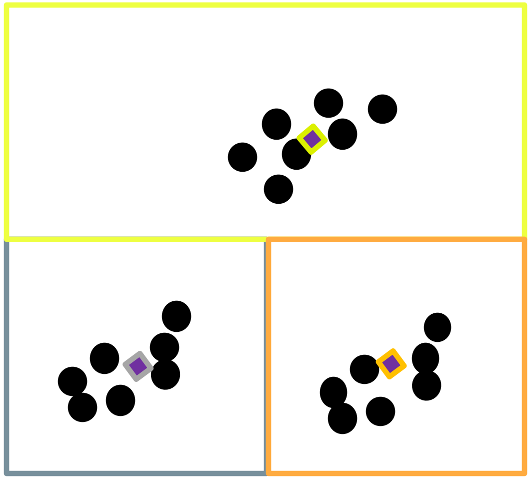
K = 3



Stability

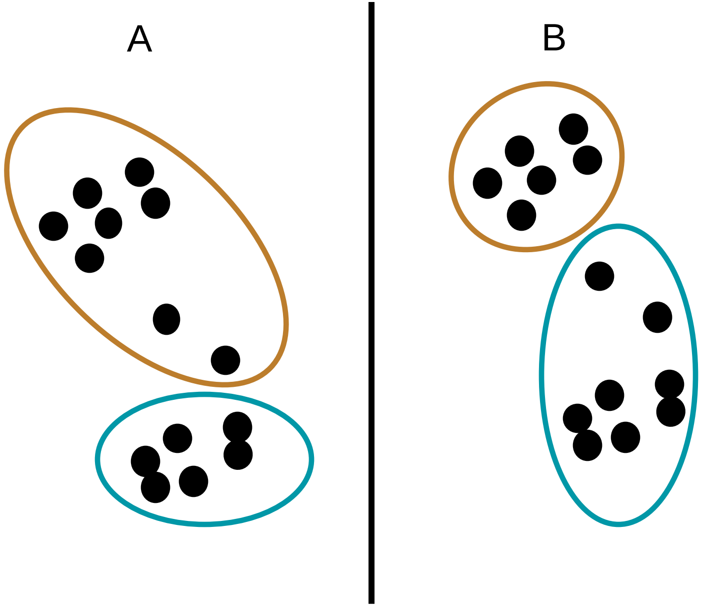
K = 2 Instable

K = 3 Stable

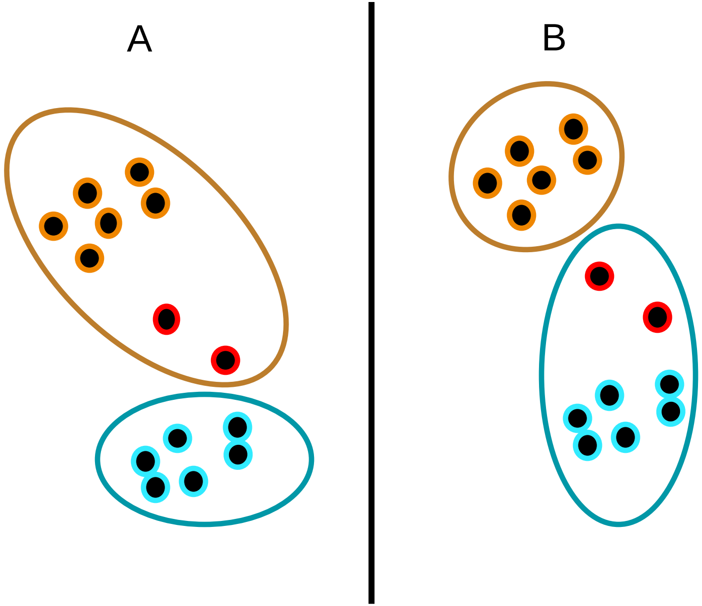


Rand index

A vs B



Rand index



A vs B

$$Rand\ index = \frac{no.\ in\ the\ same\ class}{no.\ total\ of\ observations} = \frac{12}{14}$$

Consistency

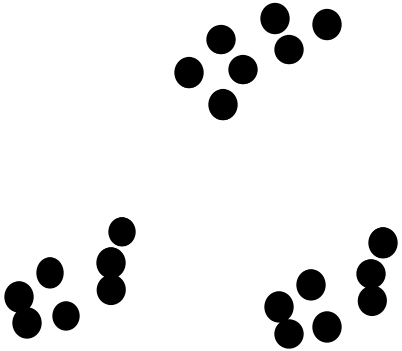
- Use business knowledge to check the cluster's relevance.

Case studies

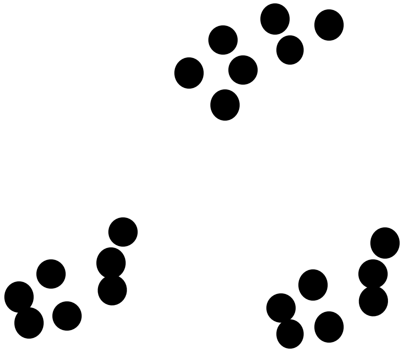
		Profil
🟦	Cluster 1	Over 50, buys little but large amounts
🟥	Cluster 2	Under 20, buys a lot but small amounts
🟨	Cluster 3	Under 30, buys a lot and in large amounts

Determining the number of classes

K = 2



K = 3



Distortion or Sum of Square Error (SSE)

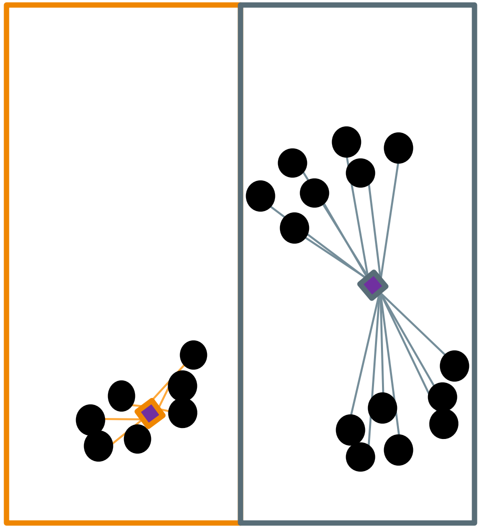
$$SSE = \sum_{j} \sum_{i} D(c_j, x_i)^2$$

With:

- c_j: The cluster center (centroid).
- x_i: the ith observation in the cluster with centroid c_j
- $D(c_j, x_i)$: The distance between the cluster center and the point x_i

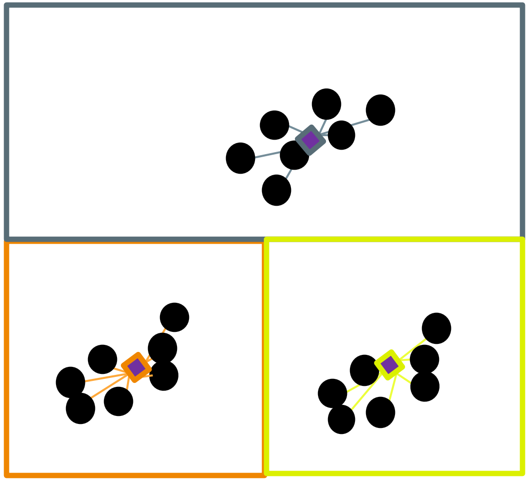
Determining the number of classes

K = 2



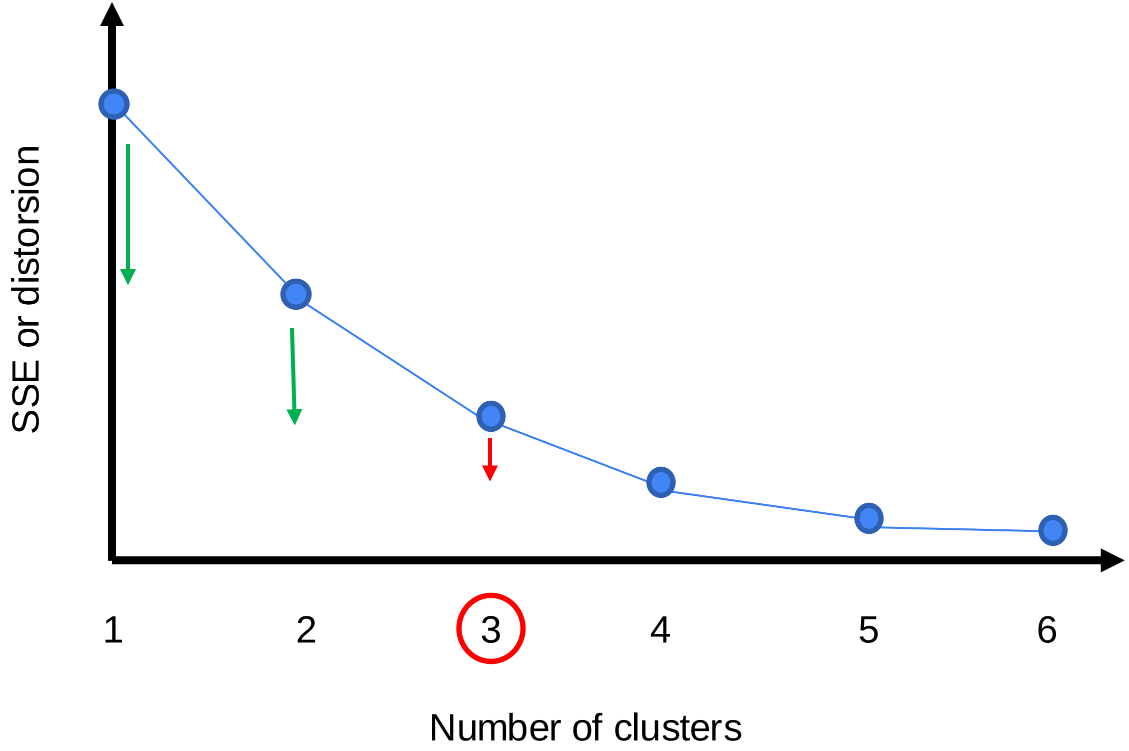
SSE Hgih

K = 3



SSE small

Elbow method


SSE or distorsion
Number of clusters