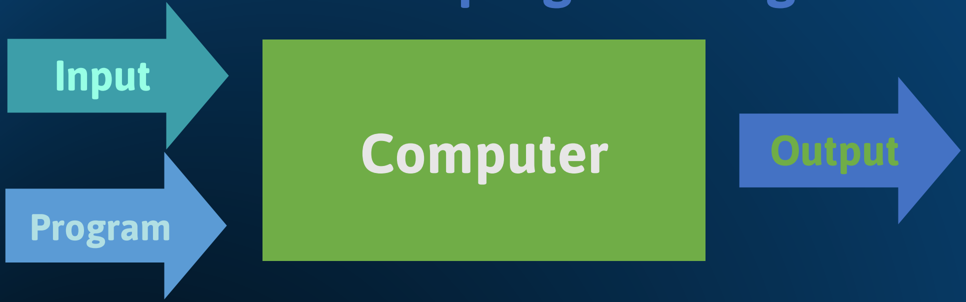
Clustering

Cláudia Antunes

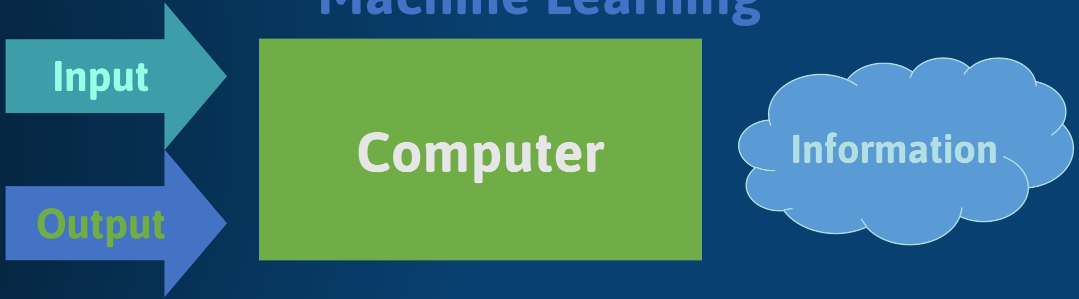
Instituto Superior Técnico – Universidade de Lisboa

Unsupervised Learning

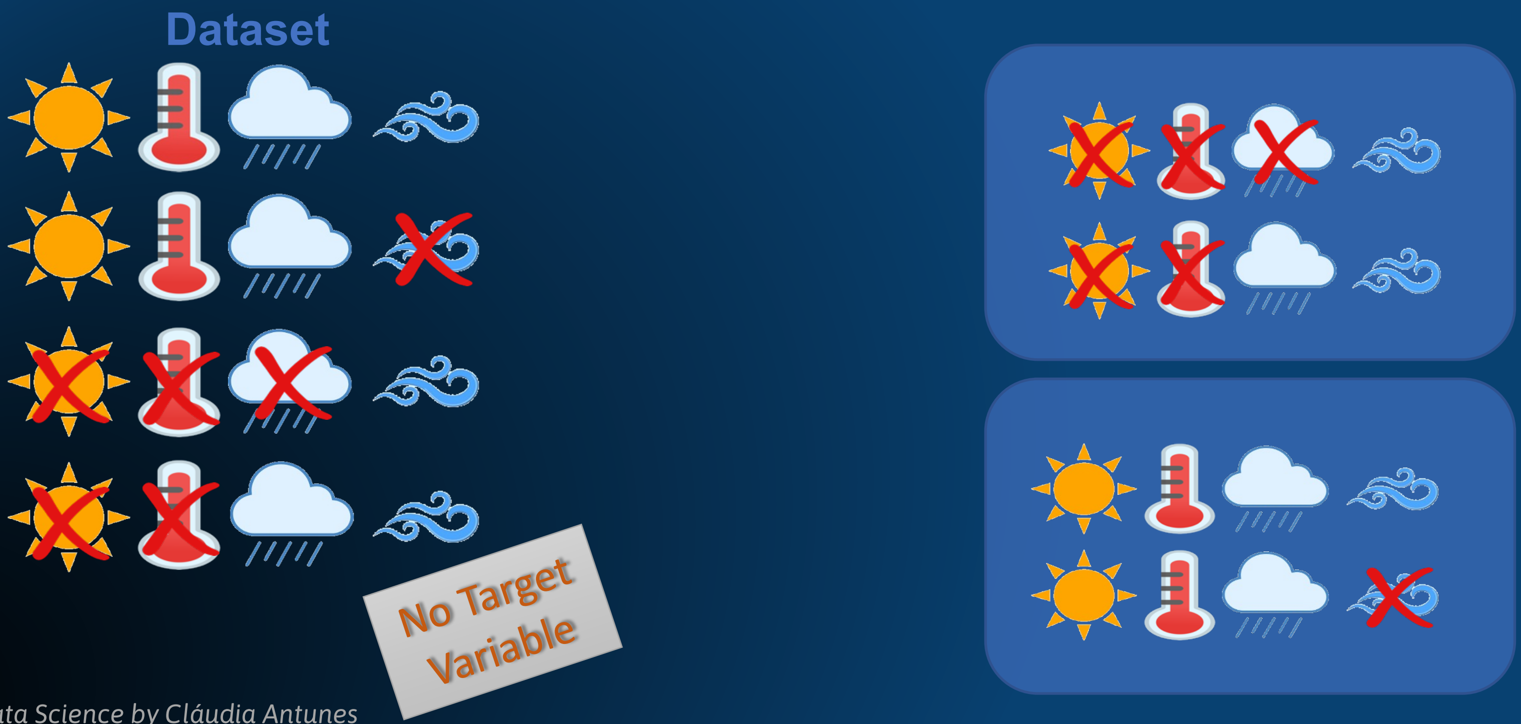
Traditional programming



Machine Learning



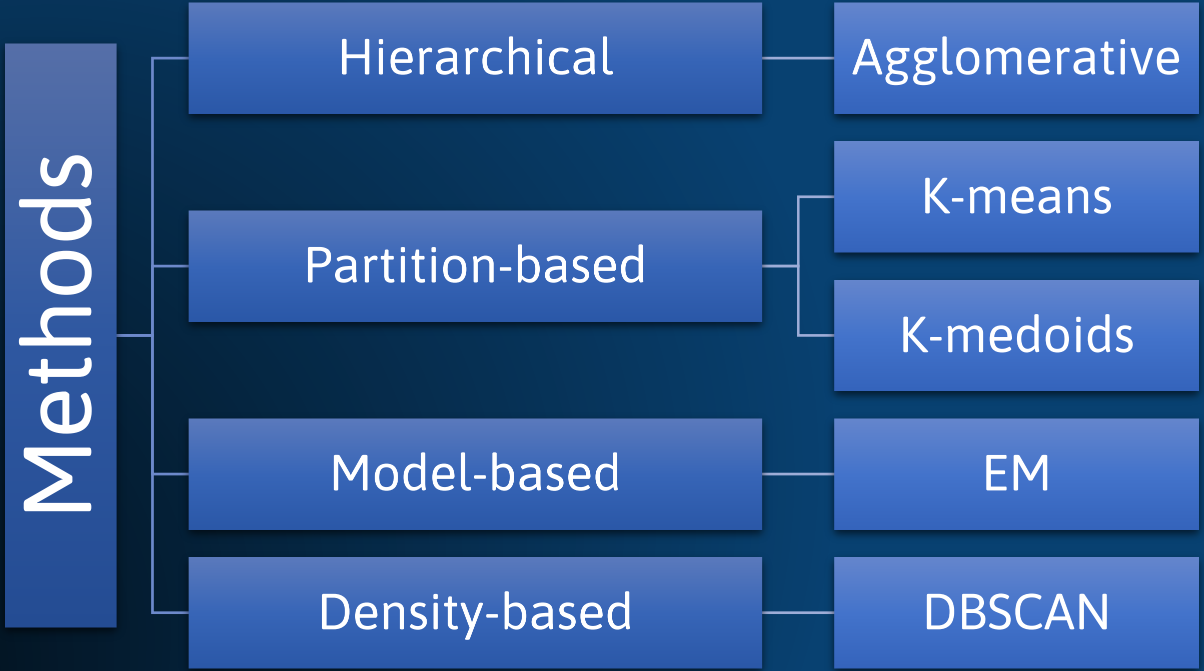
Clustering



Dataset

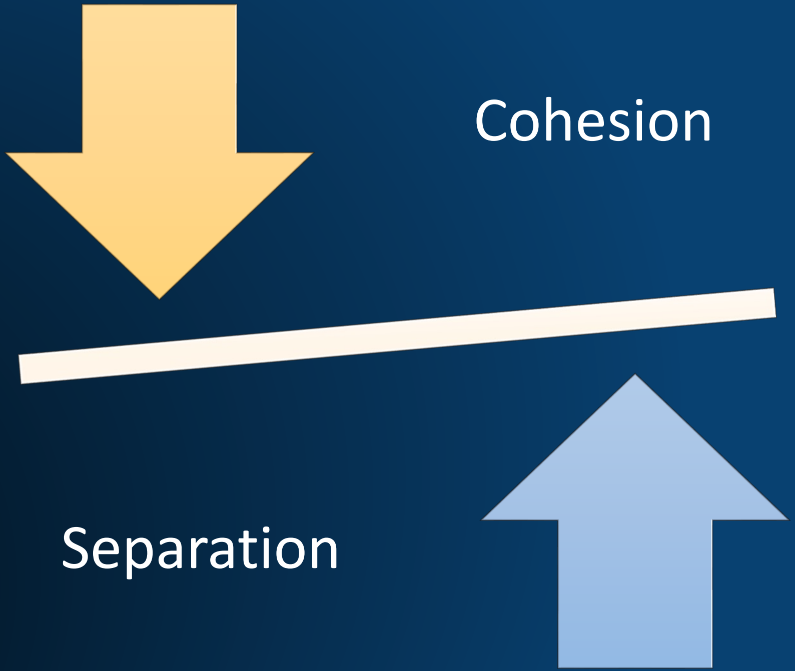
No Target Variable

Approaches


Methods
Hierarchical
Agglomerative
Partition-based
K-means
K-medoids
Model-based
EM
Density-based
DBSCAN

Assessment

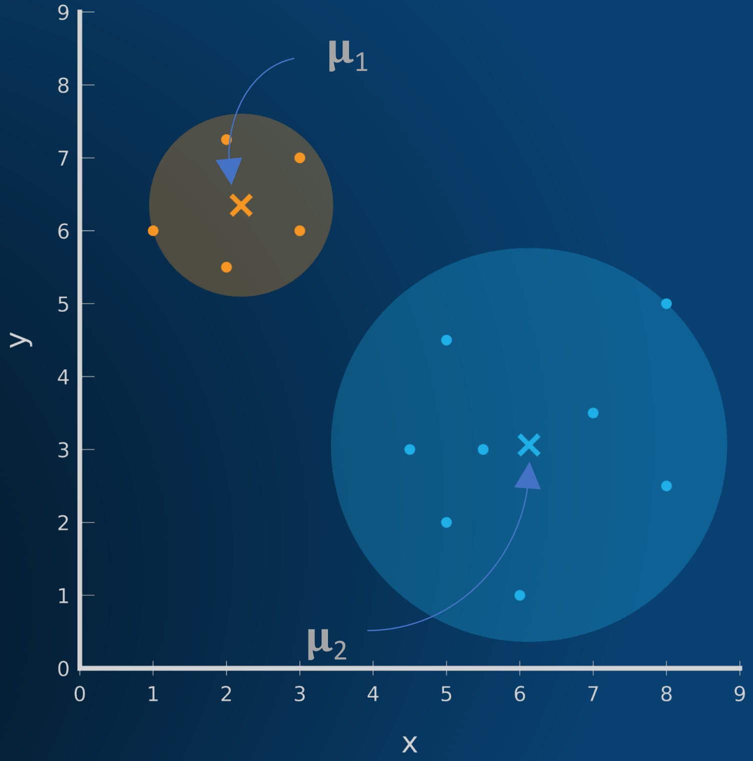
Assessment



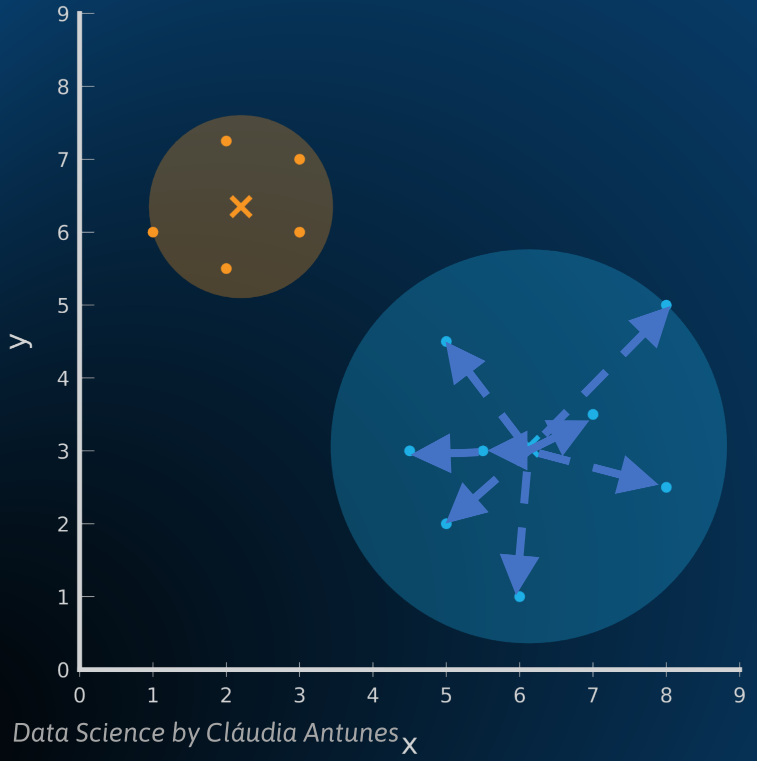
Cohesion

Separation

CLUSTERING



Assessment – Cohesion



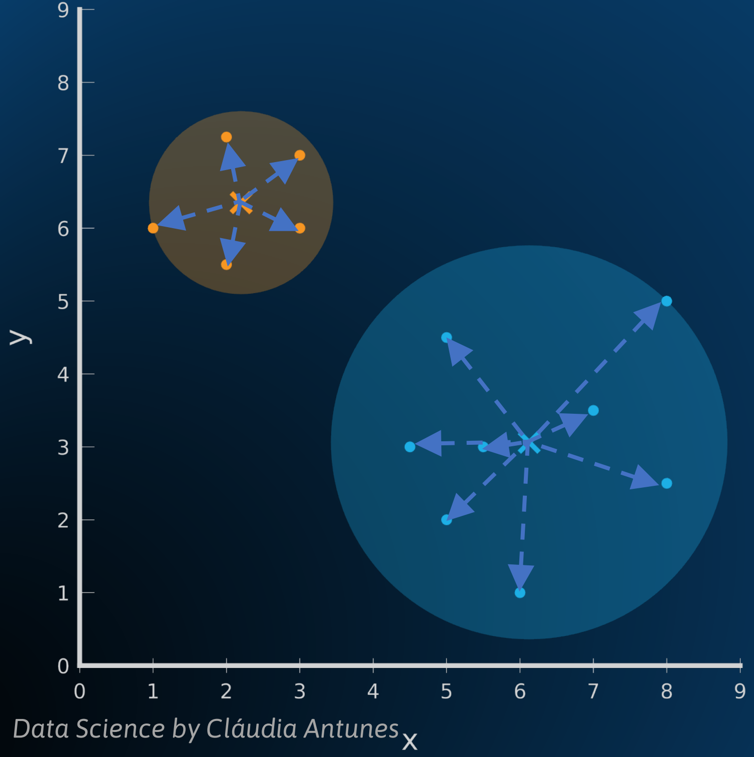
$$radius(C_i) = \\ \qquad max_x\{d(x,\mu_i): x \in C_i\}$$

$$max(C_i) = \\ \qquad max_{x,y}\{d(x,y): x,y \in C_i\}$$

$$avg\ dist(C_i\ ) = \frac{1}{|C_i|} \sum_{x \in C_i} d(x,\mu_i)$$

$$avg\ dist(C_i) = \\ \qquad \frac{1}{|C_i||C_i - 1|} \sum_{\substack{x,y \in C_i \\ x \neq y}} d(x,y)$$

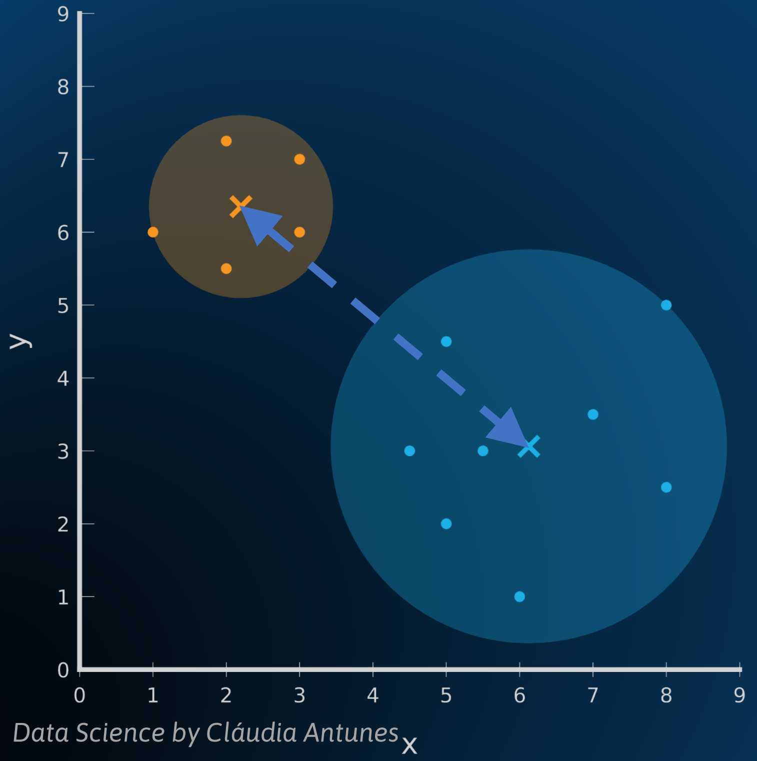
ASSESSMENT – COHESION



$$\text{MSE} = \frac{1}{N}\sum_{i}\sum_{x\in C_i} d(\mu_i, x)^2$$

$$\text{MAE} = \frac{1}{N}\sum_{i}\sum_{x\in C_i} d(\mu_i, x)$$

ASSESSMENT – SEPARATION

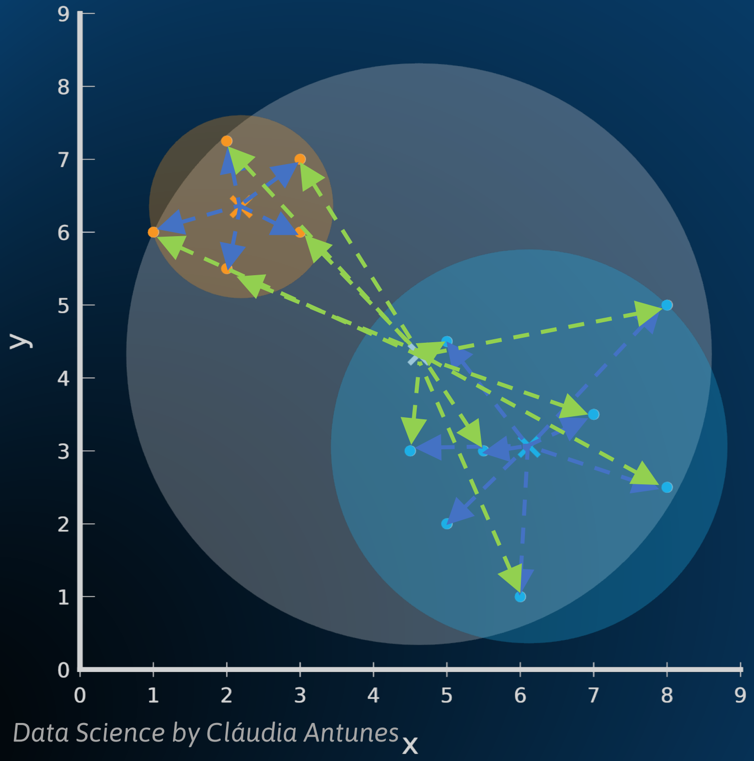


$$d(C_i, C_j) = d(\mu_i, \mu_j)$$

$$slink(C_i, C_j) = \\ min_{x,y}\{d(x, y): x \in C_i, y \in C_j\}$$

$$clink(C_i, C_j) = \\ max_{x,y}\{d(x, y): x \in C_i, y \in C_j\}$$

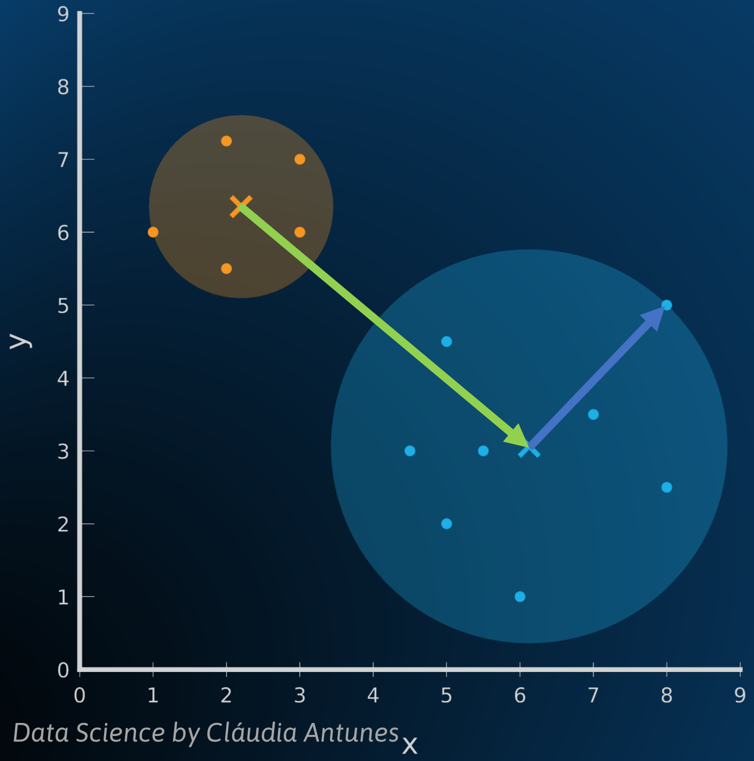
Assessment – Separability



Ward's distance

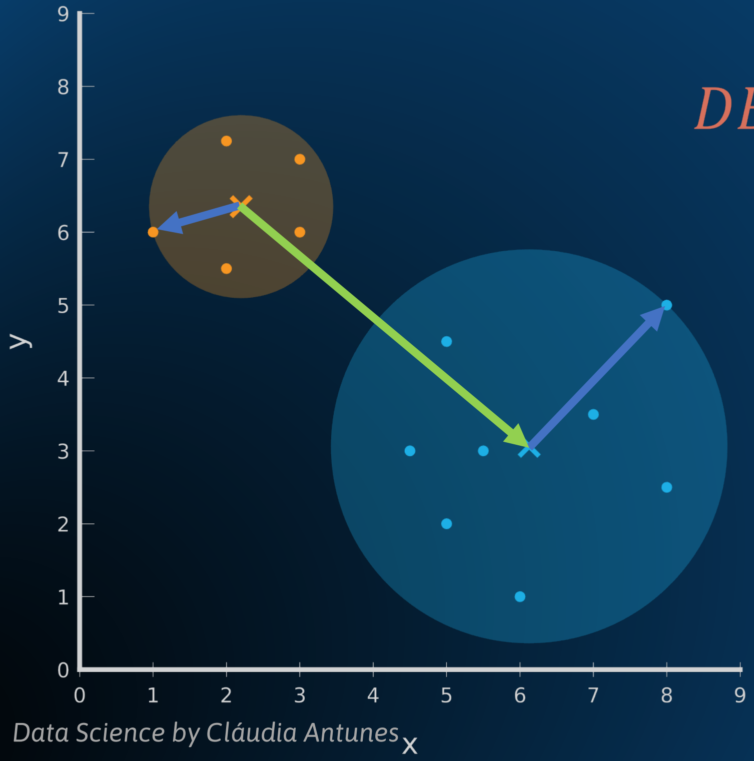
$$d(C_i, C_j) = \sum_{x \in C_i} d(x, \mu_i)^2 + \sum_{x \in C_j} d(x, \mu)^2 - \sum_{x \in C_{ij}} d(x, \mu_{ij})^2$$

Assessment – Dunn index



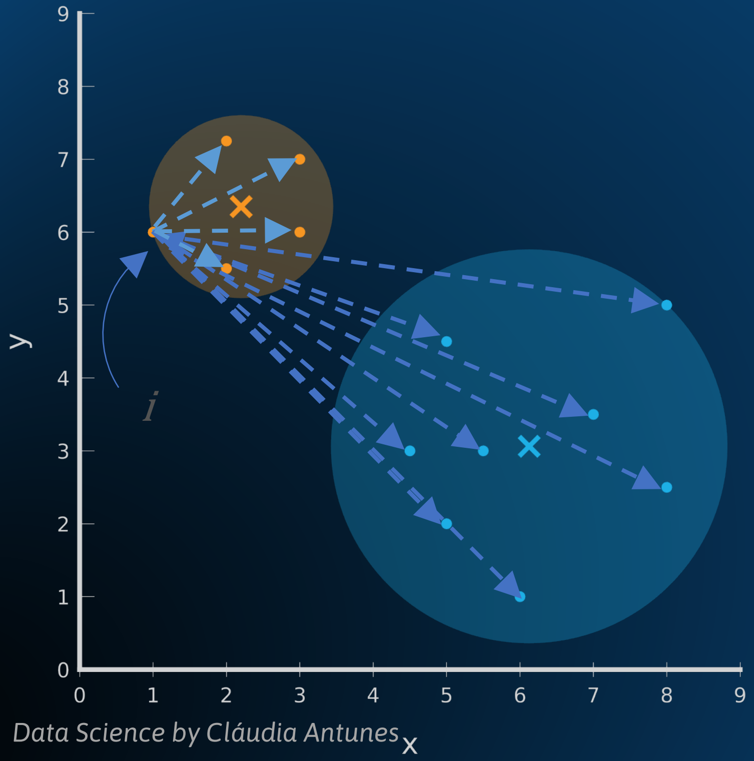
$$DI_C = \frac{min_{i,j}\{d(C_i, C_j): 1 \le i,j \le k\}}{max_i\{diam(C_i): 1 \le i \le k\}}$$

Assessment – Davies-Bouldin index



$$DB_C = \frac{1}{k} \times \sum_{\substack{i=1 \\ i \neq j}}^{k} \max\left\{\frac{diam(C_i) + diam(C_j)}{d(C_i, C_j)}\right\}$$

Assessment – Silhouette

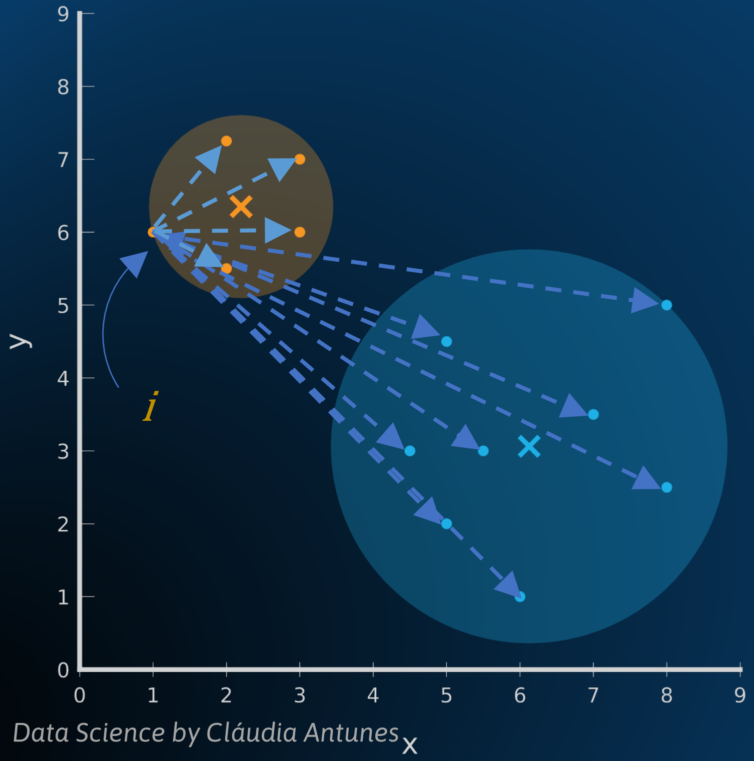


$$\mu_{in}(x_i) = \frac{1}{\left|C_{x_i}\right| - 1} \sum_{x_j \in C_{x_i}, i \neq j} d(x_i, x_j)$$

$$\mu_{out}(x_i) = min_k \frac{1}{|C_k|} \sum_{\substack{x_j \in C_k \\ C_k \neq C_{x_i}}} d(x_i, x_j)$$

$$s(x_i) = \frac{\mu_{out}(x_i) - \mu_{min}(x_i)}{\max\{\mu_{out}(x_i), \mu_{in}(x_i)\}}$$

Assessment – Silhouette



$$s(x_i) = \frac{\mu_{out}(x_i) - \mu_{min}(x_i)}{\max\{\mu_{out}(x_i), \mu_{in}(x_i)\}}$$

$$SC = \frac{1}{|D|}\sum_{x_i \in D} s(x_i)$$

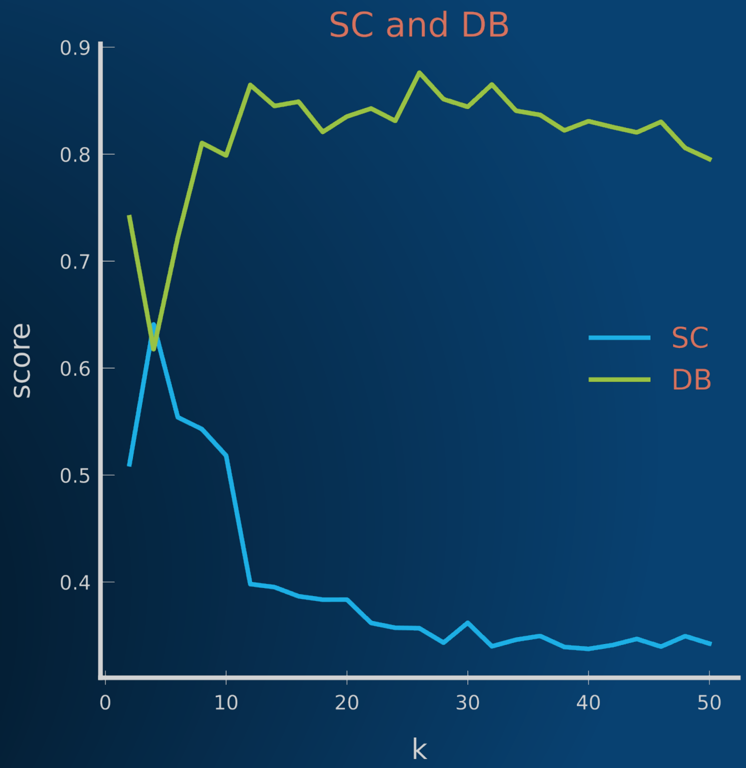
$0.75 \leq SC \leq 1.00$ → excellent

$0.50 \leq SC < 0.75$ → good

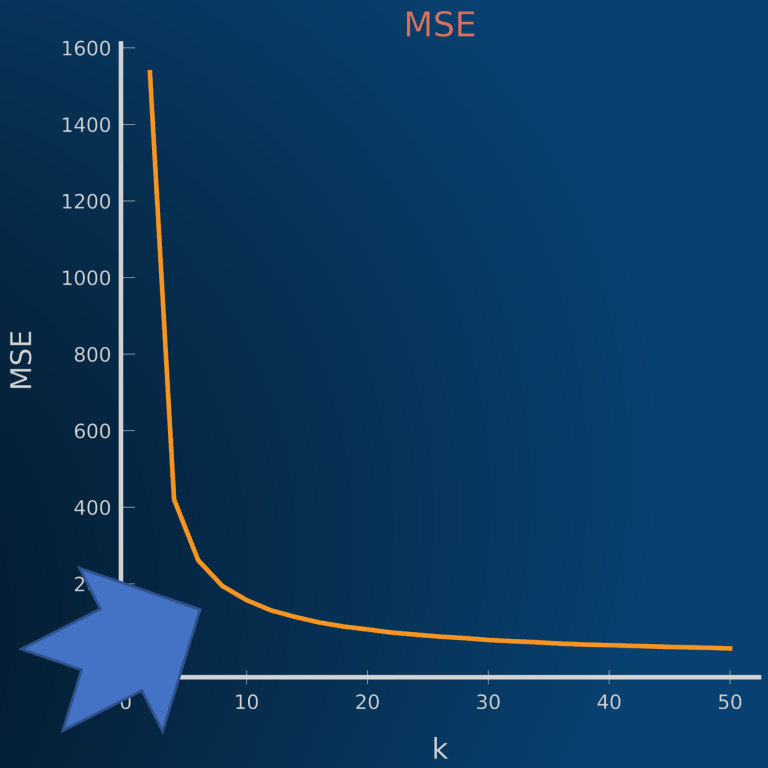
$0.25 \leq SC < 0.50$ → weak

$SC < 0.25$ → no structure

Choosing K



Choosing K – the elbow method



Thank you!

claudia.antunes@tecnico.ulisboa.pt