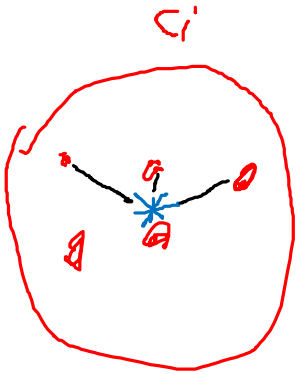
K-Means Clustering Evaluation

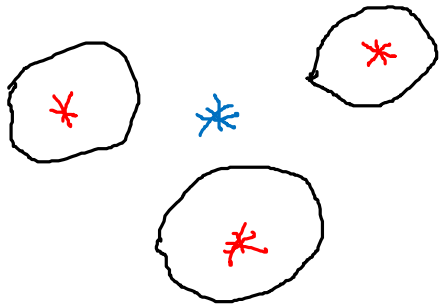
Cluster Cohesion

- Measures how closely related are objects in a cluster.
- An example can be sum of squares (SSE).
- Also known as Within-Cluster Sum of Square (WCSS).
- $SSE = WCSS = \sum_{i \in |C|} \sum_{x \in C_i} (x - m_i)^2$

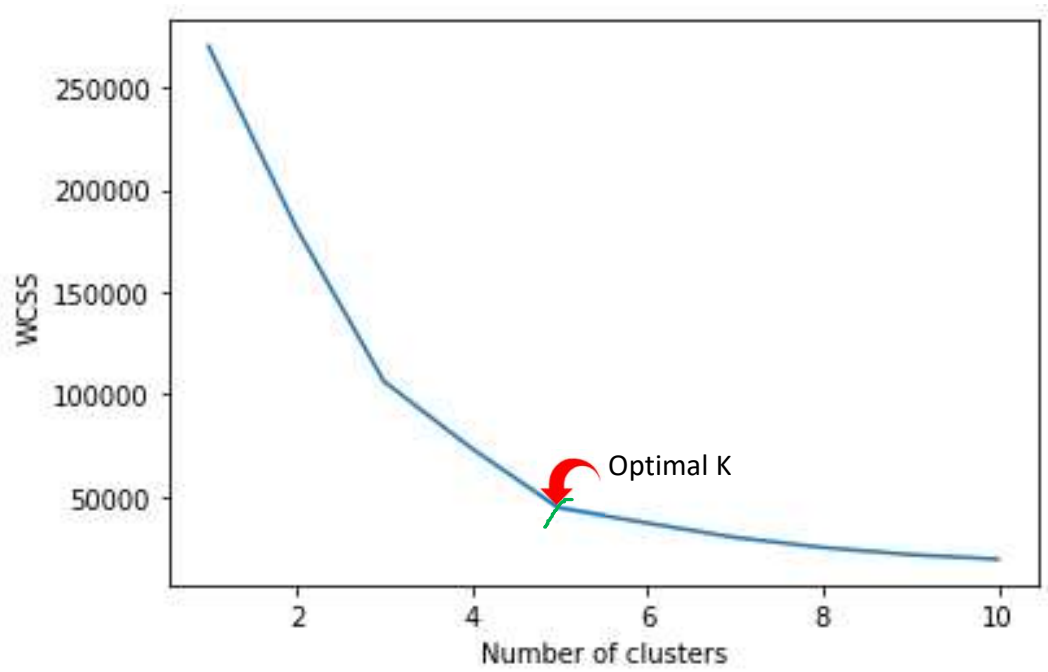


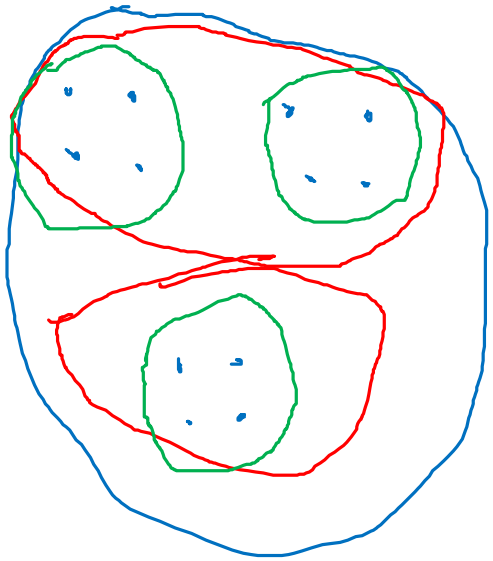
Cluster Separation

- Measures how distinct or well separated a cluster is from other clusters.
- An example can be between cluster sum of squares (BSS).
- $BSS = \sum_{i \in |C|} |C_i|(m - m_i)^2$



Elbow Method



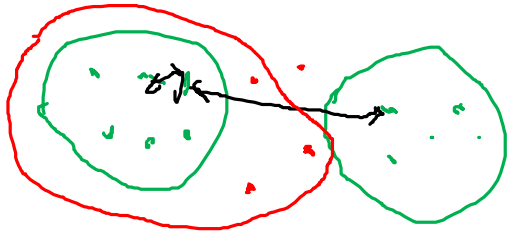


Elbow Method (Scikit-Learn)

```
from sklearn.cluster import KMeans
wcss = [] for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state = 42)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)
```

Silhouette Coefficient

- Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.
- 1: Means clusters are well apart from each other and clearly distinguished.
- 0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.
- -1: Means clusters are assigned in the wrong way.

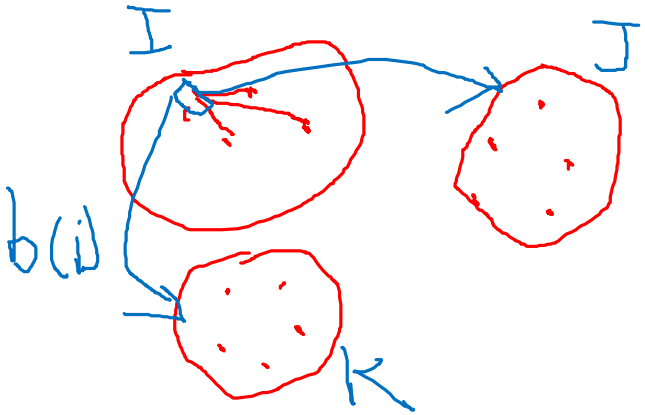


Silhouette Coefficient Formula

- For data point, i in cluster C_I, let a(i) and b(i) be defined as follows:

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i,j)$$

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)$$



- The silhouette coefficient S(i) is:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_I| > 1$$

- Overall silhouette coefficient S is calculated as : *S = mean{S(i)}.*