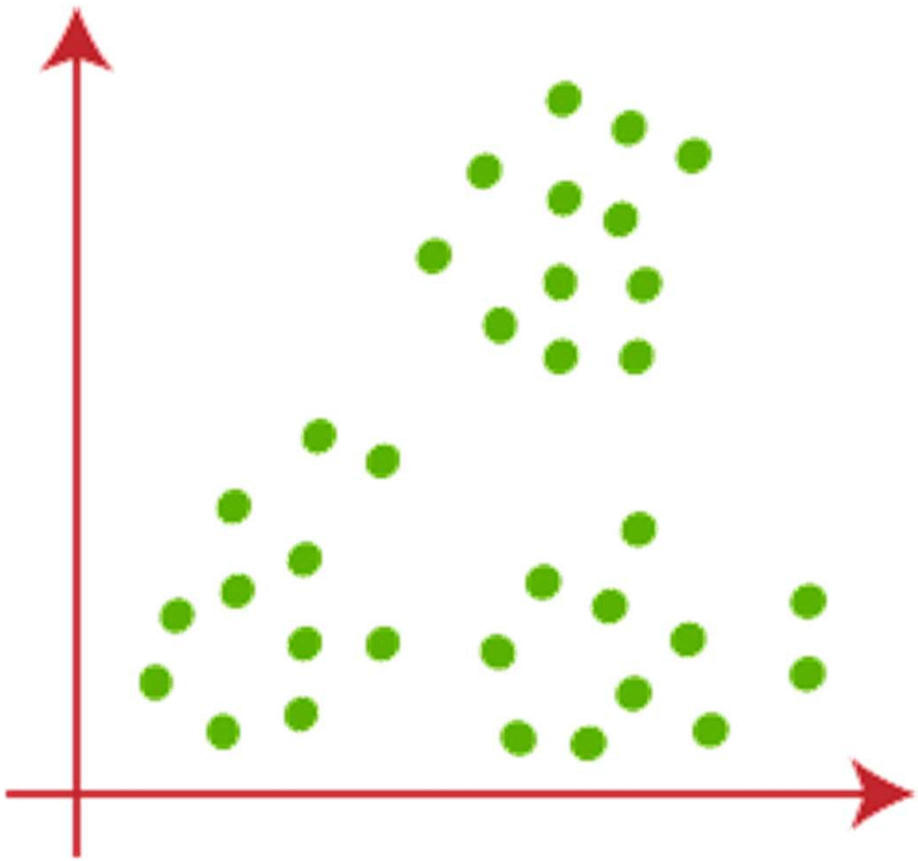
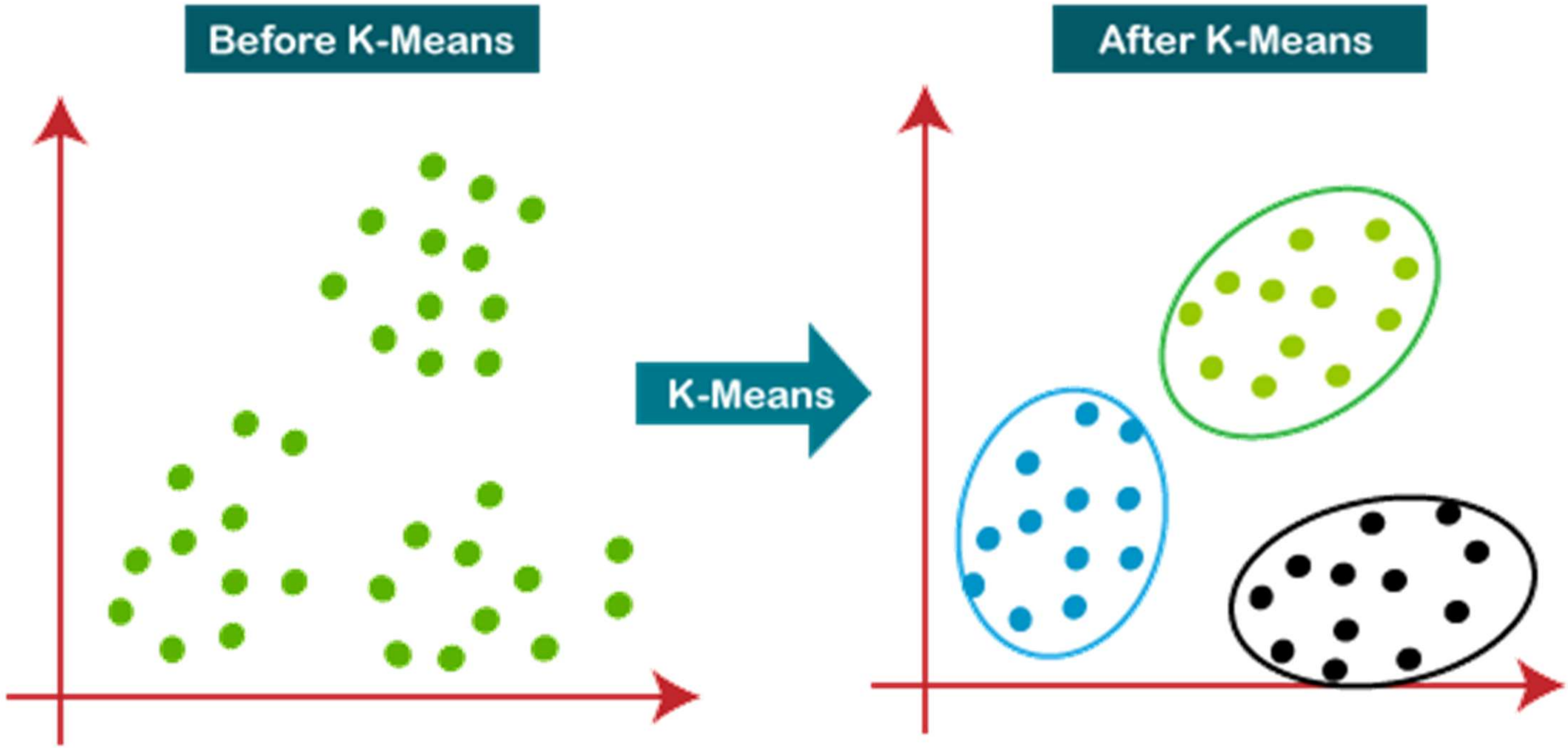
K-means clustering

Python

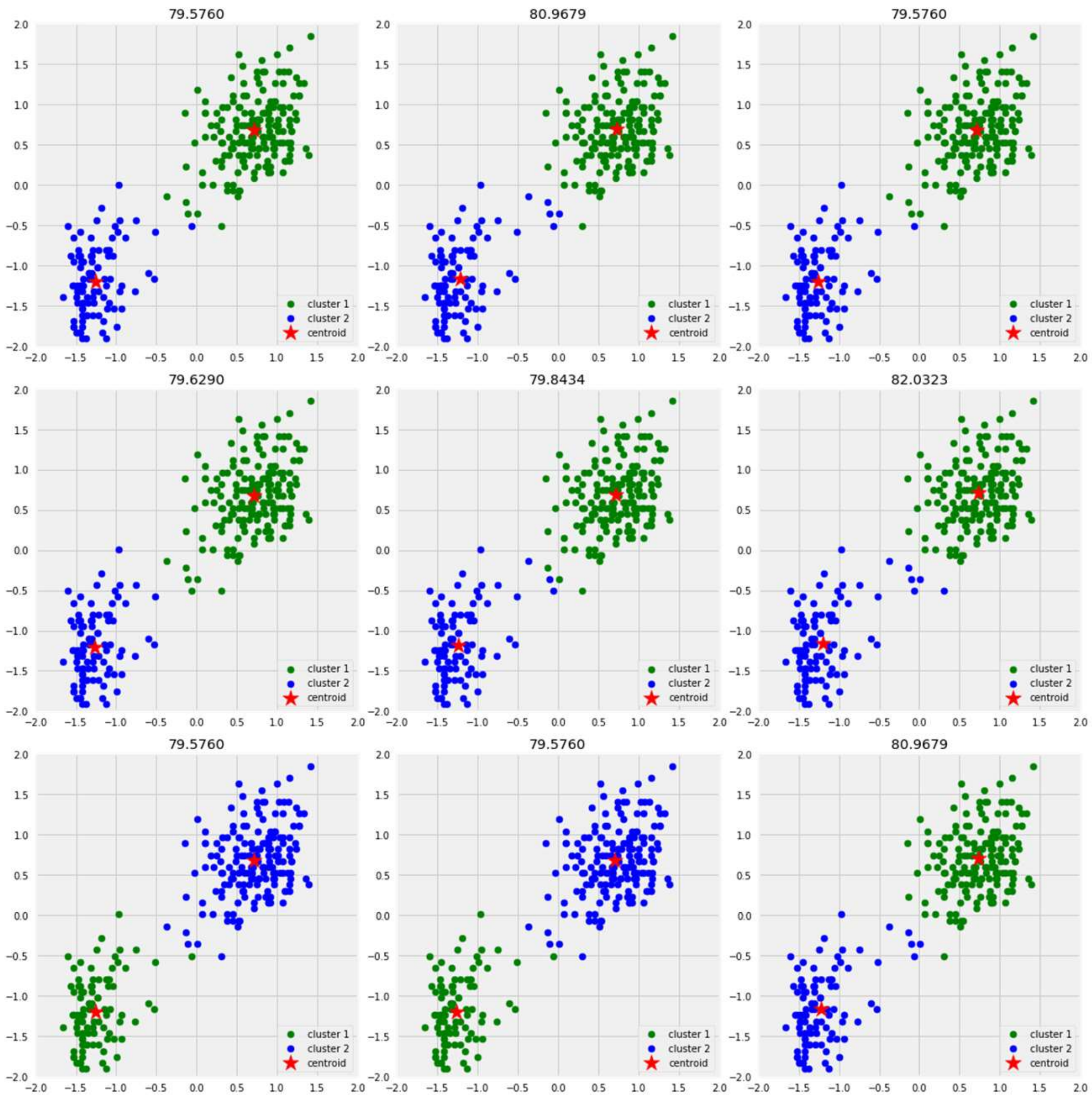
Wrocław, 2022




Before K-Means
K-Means
After K-Means

Clustering

Centroid

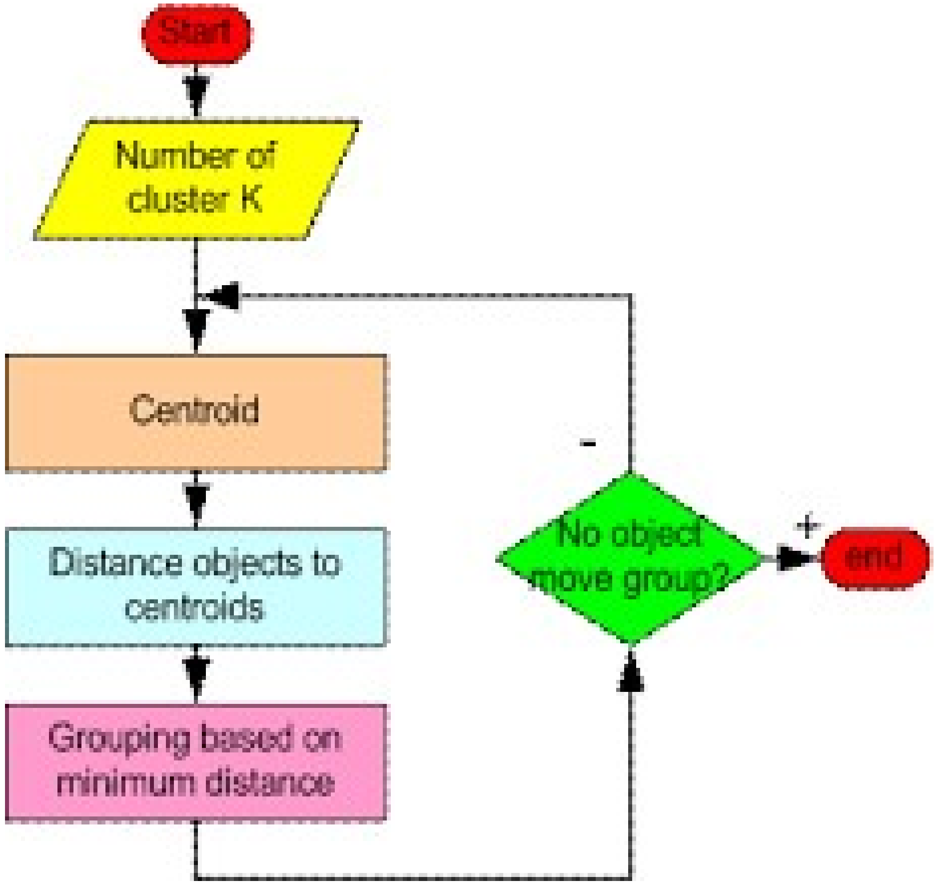

79.5760
80.9679
79.5760
79.6290
79.8434
82.0323
79.5760
79.5760
80.9679
cluster 1
cluster 2
centroid

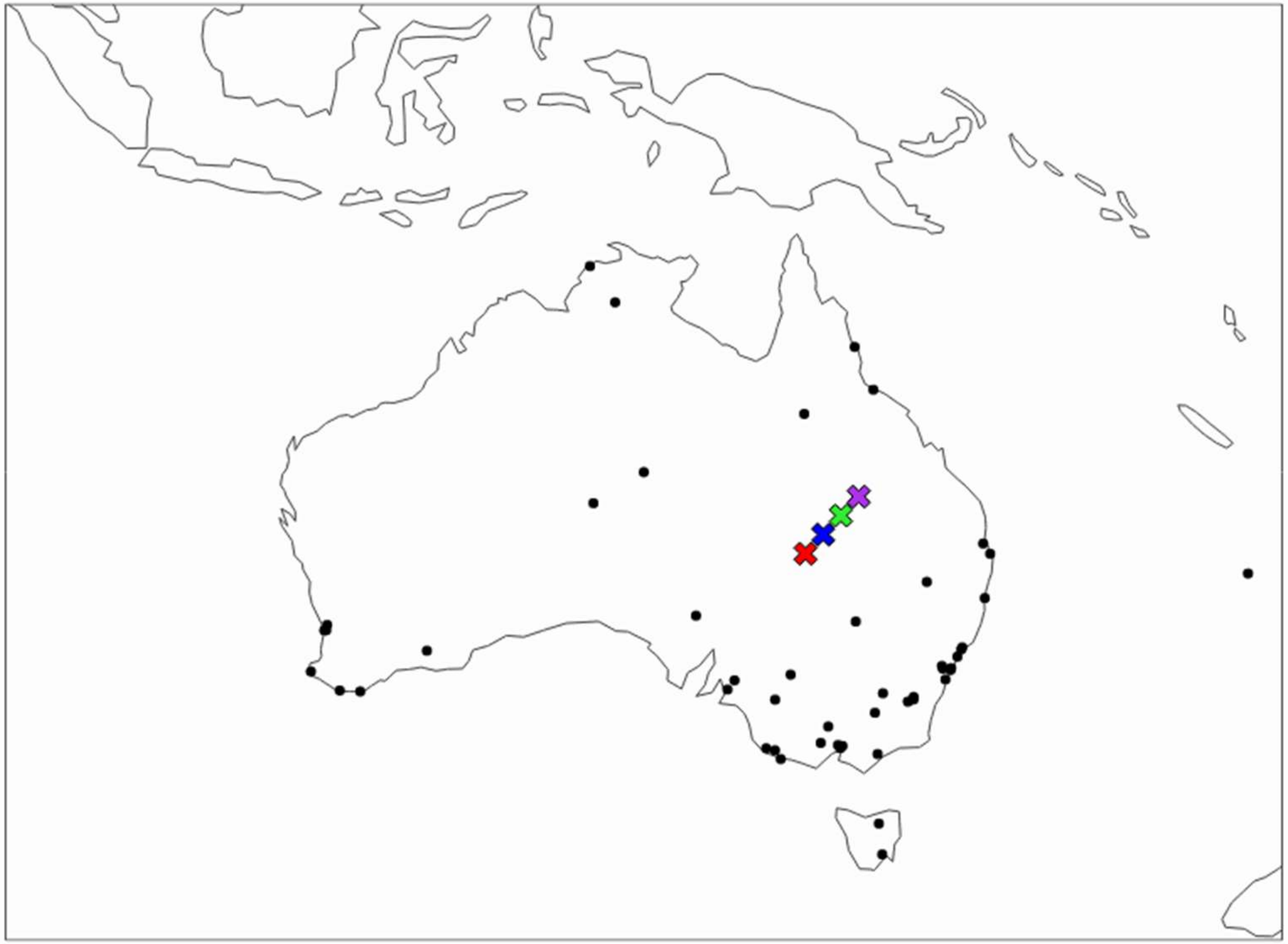
K-means clustering

Algorithm 1 k-means algorithm

```
Specify the number k of clusters to assign.
Randomly initialize k centroids.
repeat
    expectation: Assign each point to its closest centroid.
    maximization: Compute the new centroid (mean) of each cluster.
until The centroid positions do not change.
```

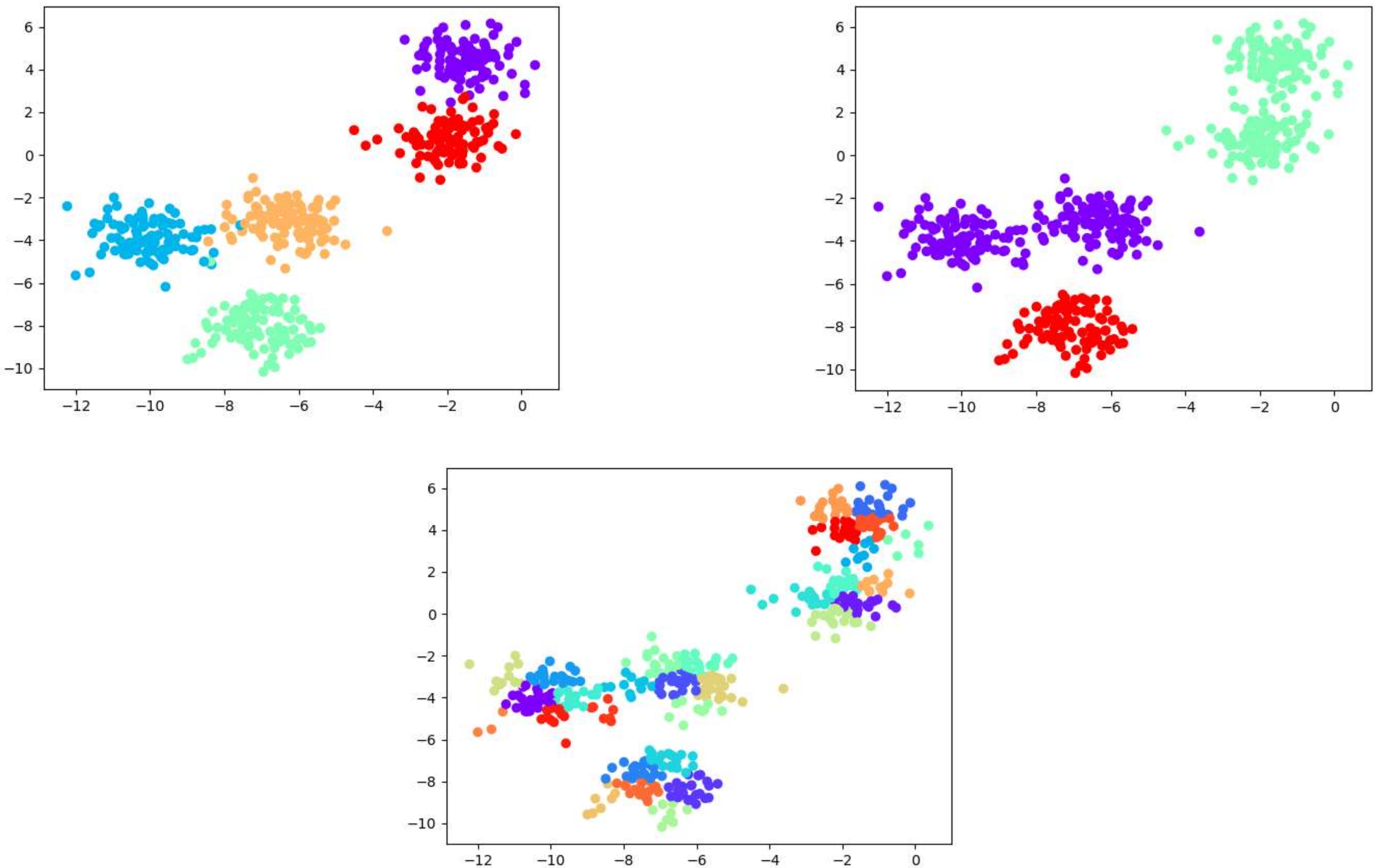
K-means clustering





How good is clustering?

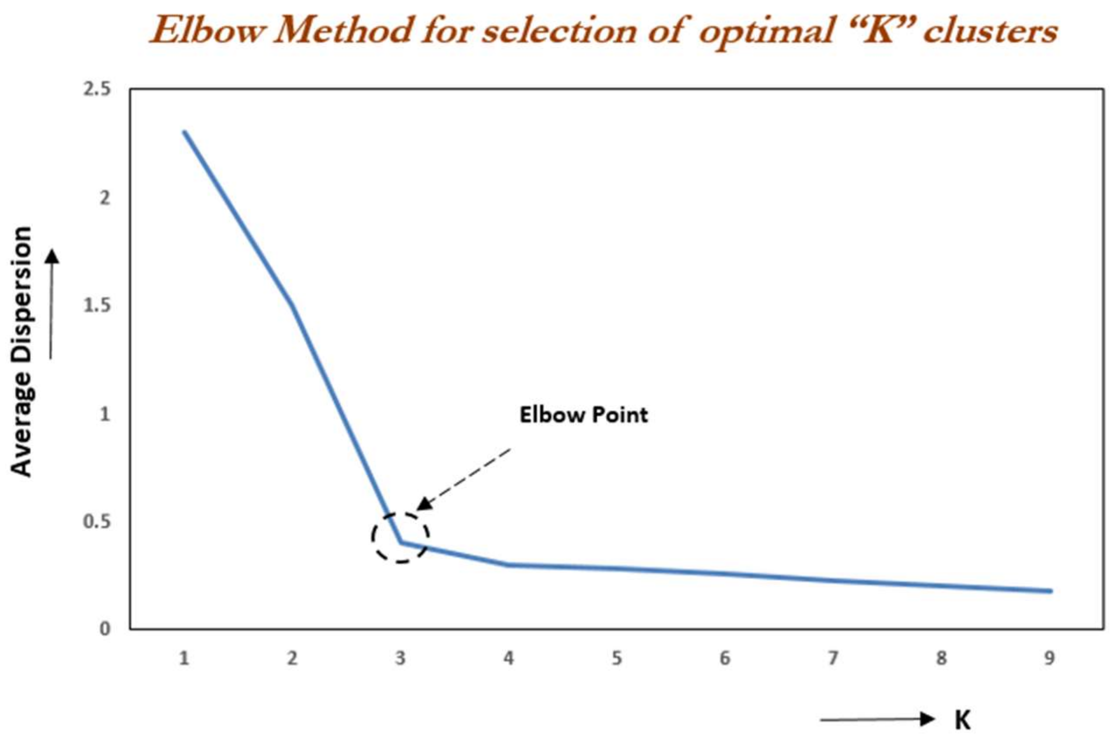
How to choose number of clusters


6
4
2
0
−2
−4
−6
−8
−10
−12
−10
−8
−6
−4
−2
0
6
4
2
0
−2
−4
−6
−8
−10
−12
−10
−8
−6
−4
−2
0
6
4
2
0
−2
−4
−6
−8
−10
−12
−10
−8
−6
−4
−2
0

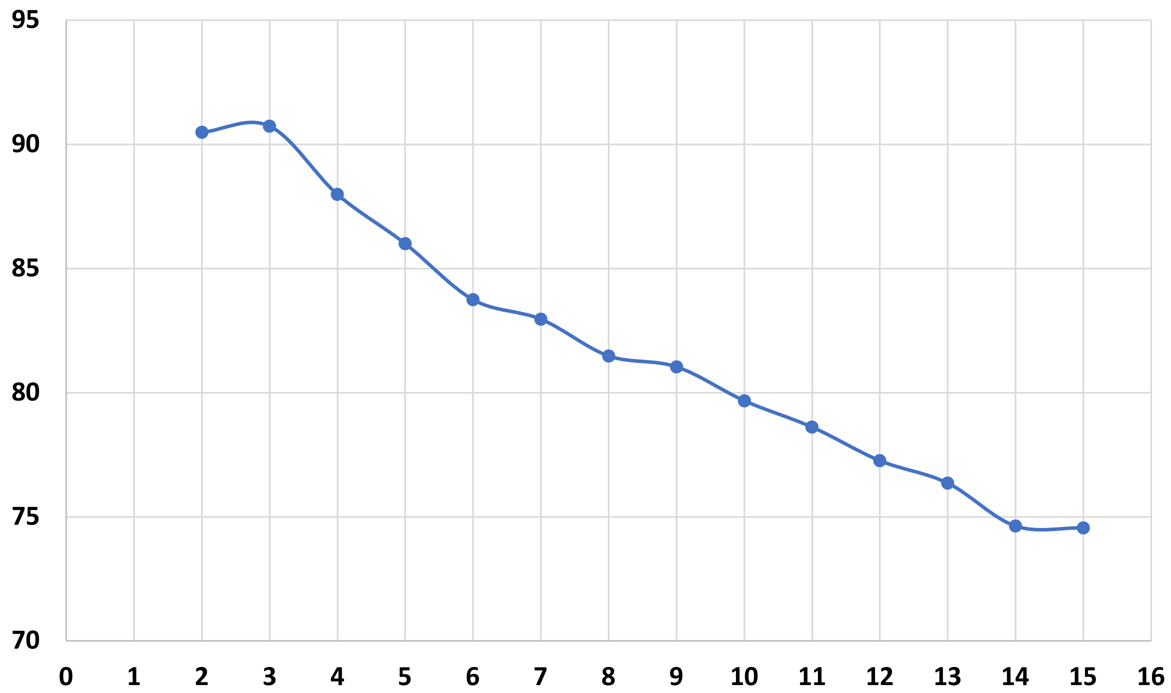
Elbow method

$$\mathrm{WCSS}(k) = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in \text{cluster } j} \|\mathbf{x}_i - \bar{\mathbf{x}}_j\|^2,$$

where $\bar{\mathbf{x}}_j$ is the sample mean in cluster j

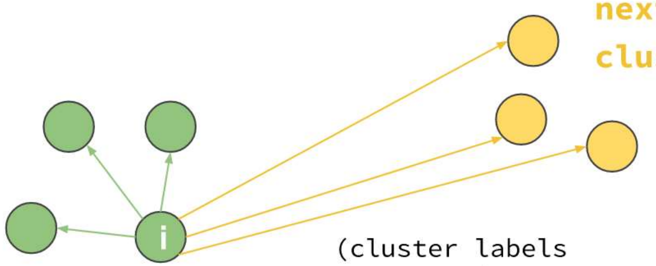


Elbow method – possible problems



Silhoutte

b(i): distance between i and its **next nearest cluster centroid**



(cluster labels assigned from ground-truth classes)

a(i): average distance between i and all of the other points in **its own cluster**

For a single point, i

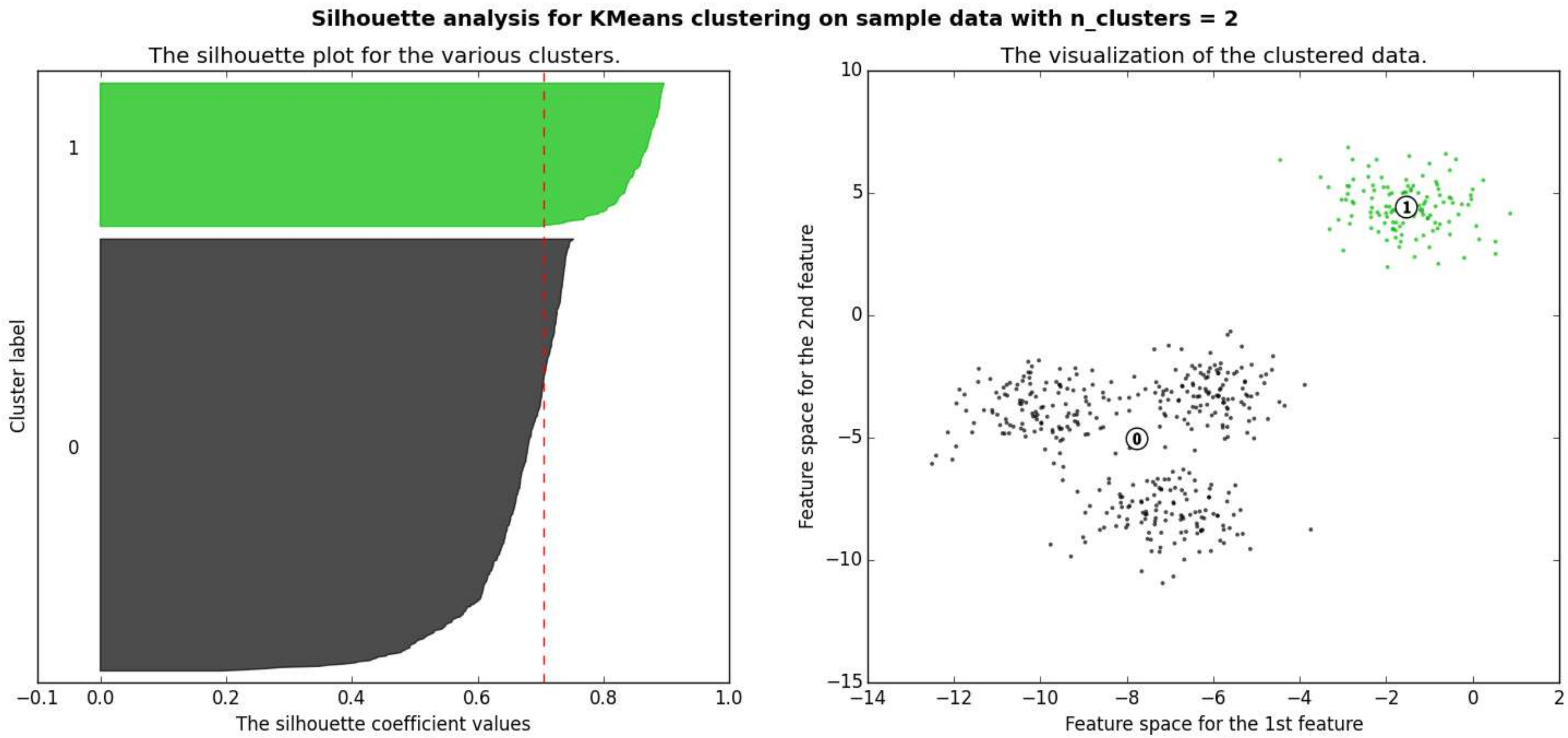
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$-1 \leq s(i) \leq 1$$

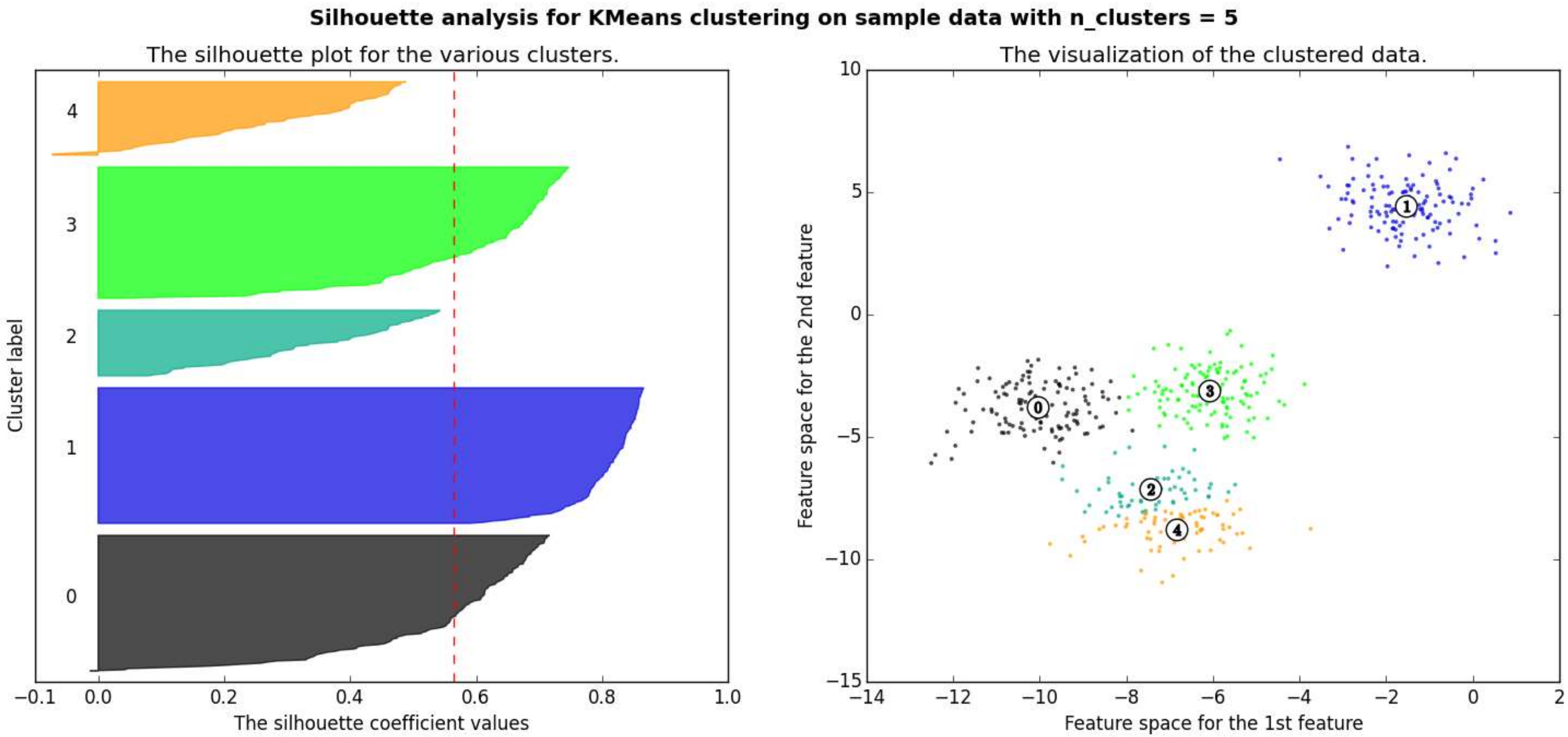
Sprawling, overlapped clusters

Tight, well-separated clusters

Silhoutte

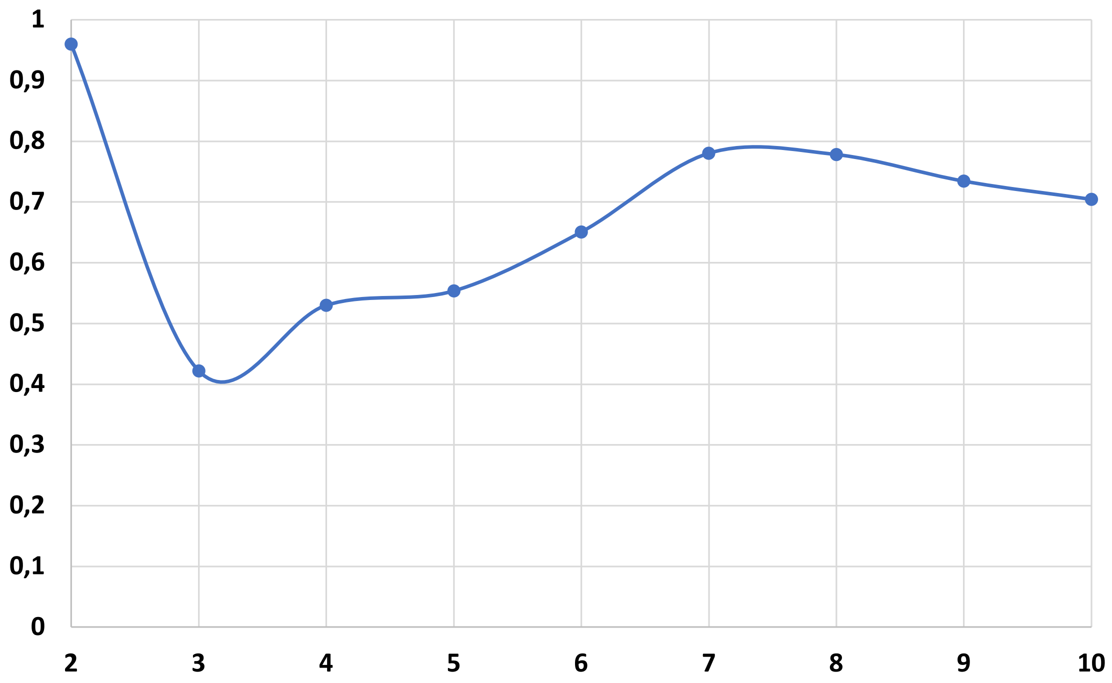

Silhouette analysis for KMeans clustering on sample data with n_clusters = 2
The silhouette plot for the various clusters.
Cluster label
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
10
5
0
−5
−10
−15
−14
−12
−10
−8
−6
−4
−2
0
2
Feature space for the 1st feature

Silhoutte



Silhouette analysis for KMeans clustering on sample data with n_clusters = 5
The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature
10
5
0
−5
−10
−15
−14
−12
−10
−8
−6
−4
−2
0
2

Silhoutte



Other methods…

Współczynnik Calińskiego-Harabasza

$$\frac{SS_B}{SS_w} \cdot \frac{N-k}{k-1}$$

K – liczba klastrów,

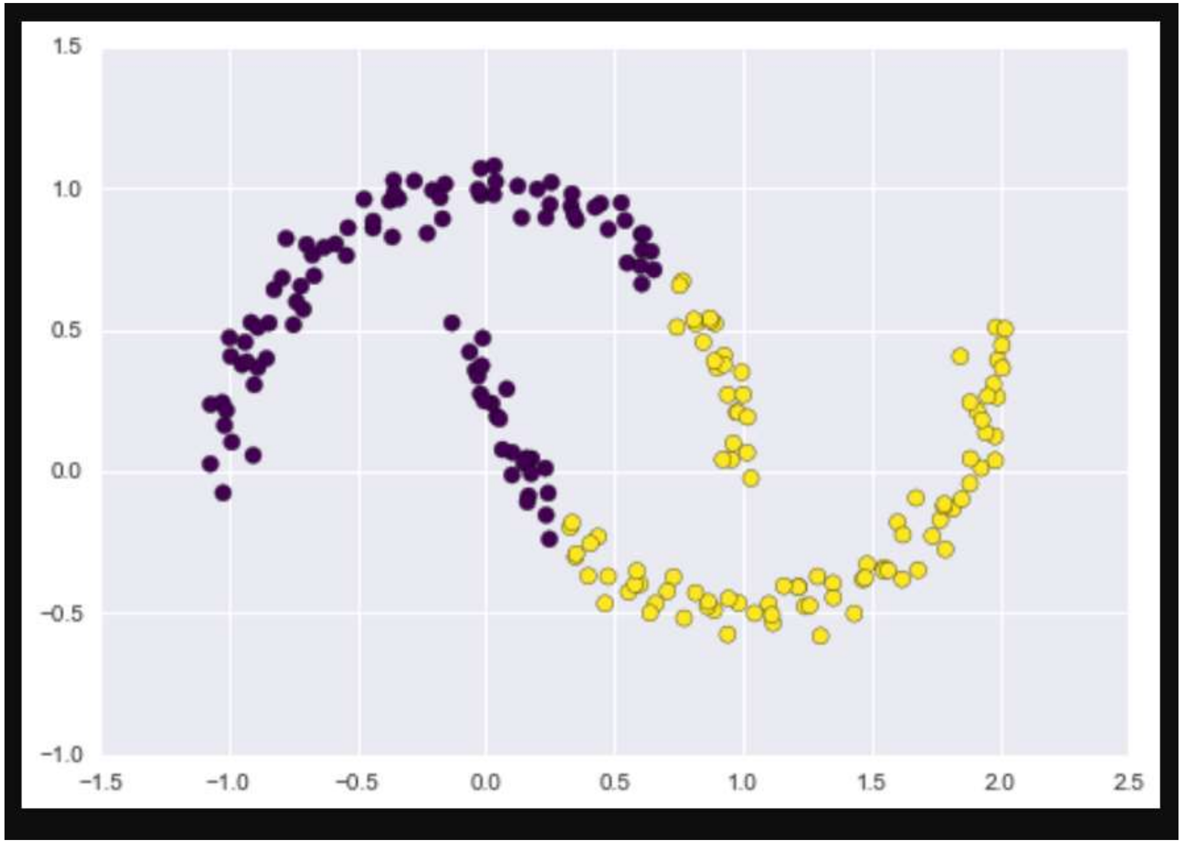
N – liczba obserwacji,

SS_w – wariancja wewnątrz-klastrowa

SS_B – wariancja między-klastrowa

- Im wyższa wartość tym lepiej

K-means clustering - problems



Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large `n_samples`, medium `n_clusters` with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with `n_samples`	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	number of clusters	Medium `n_samples`, small `n_clusters`	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	number of clusters, linkage type, distance	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large `n_samples`, medium `n_clusters`	Non-flat geometry, uneven cluster sizes	Distances between nearest points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers
Birch	branching factor, threshold, optional global clusterer.	Large `n_clusters` and `n_samples`	Large dataset, outlier removal, data reduction.	Euclidean distance between points