

k-Means Clustering

Prof. Dr. Jan Kirenz
HdM Stuttgart

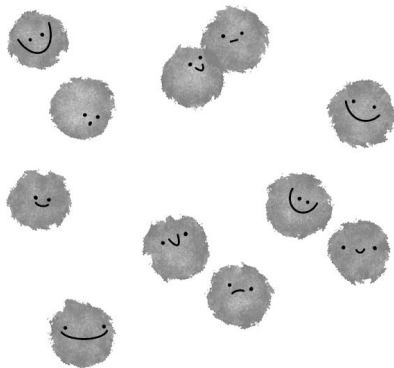
k-Means algorithm example

k-means

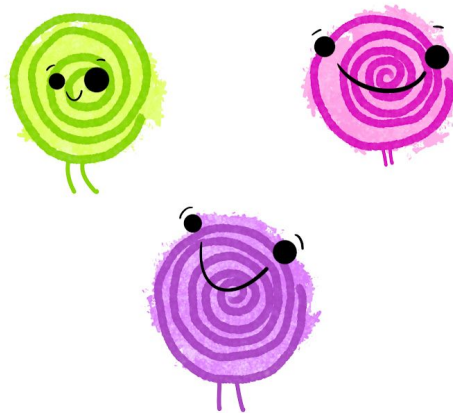
clustering

- assign each observation to one of k clusters based on the nearest cluster centroid.

OBSERVATIONS



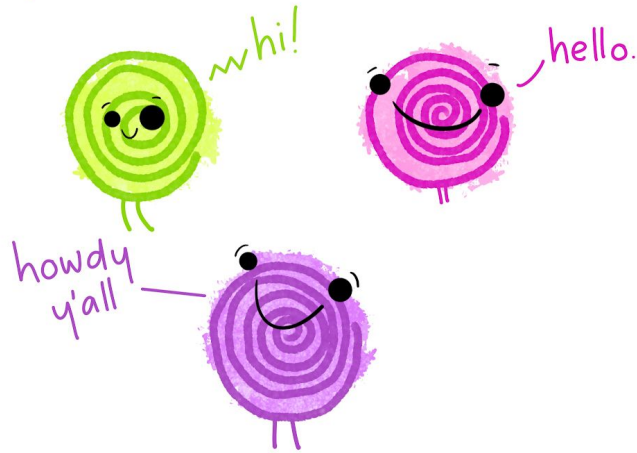
cluster
CENTROIDS



①

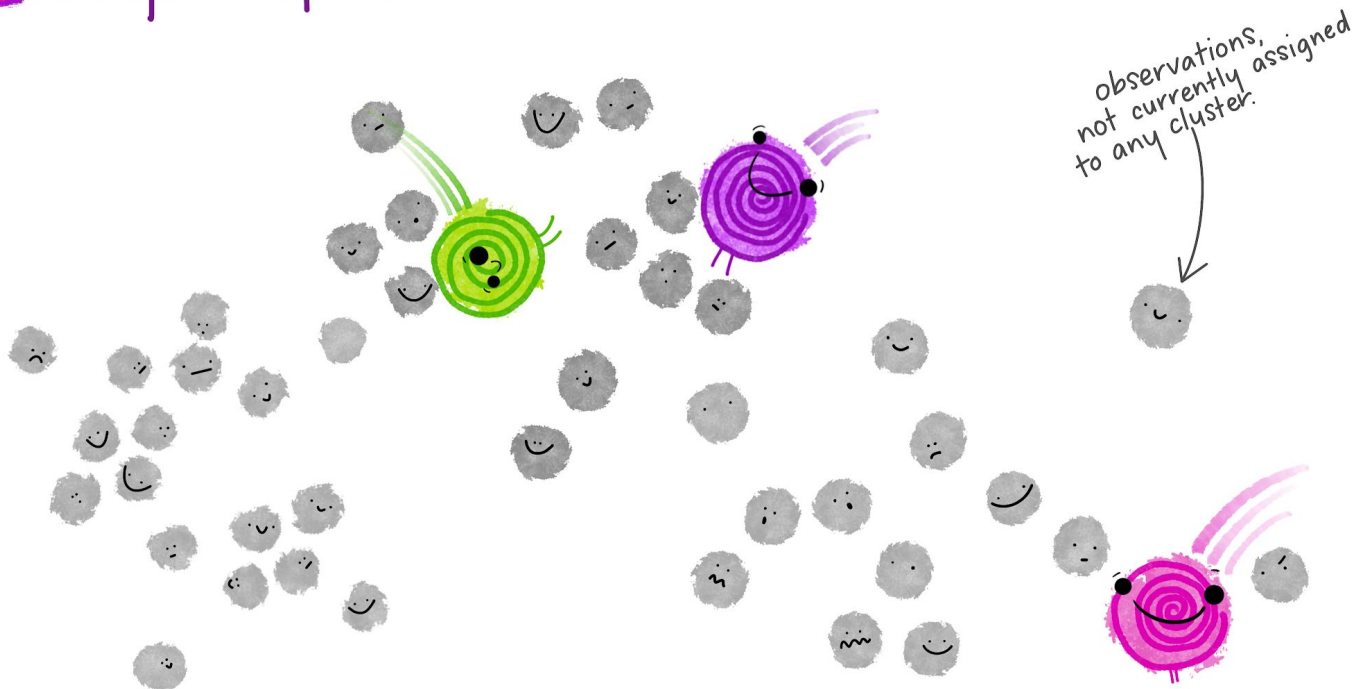
Specify the number of clusters (in this example, $k=3$).

Then imagine k cluster centroids are created.



2

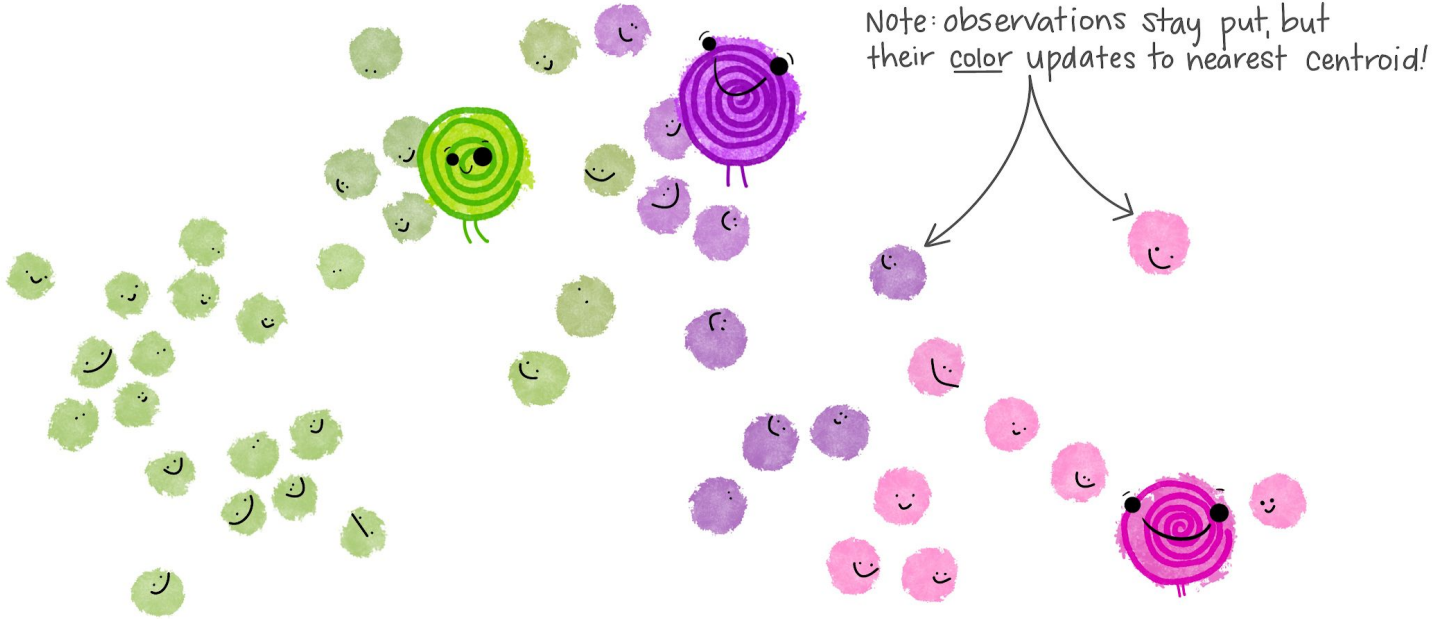
Those k centroids get randomly placed in your space.



3

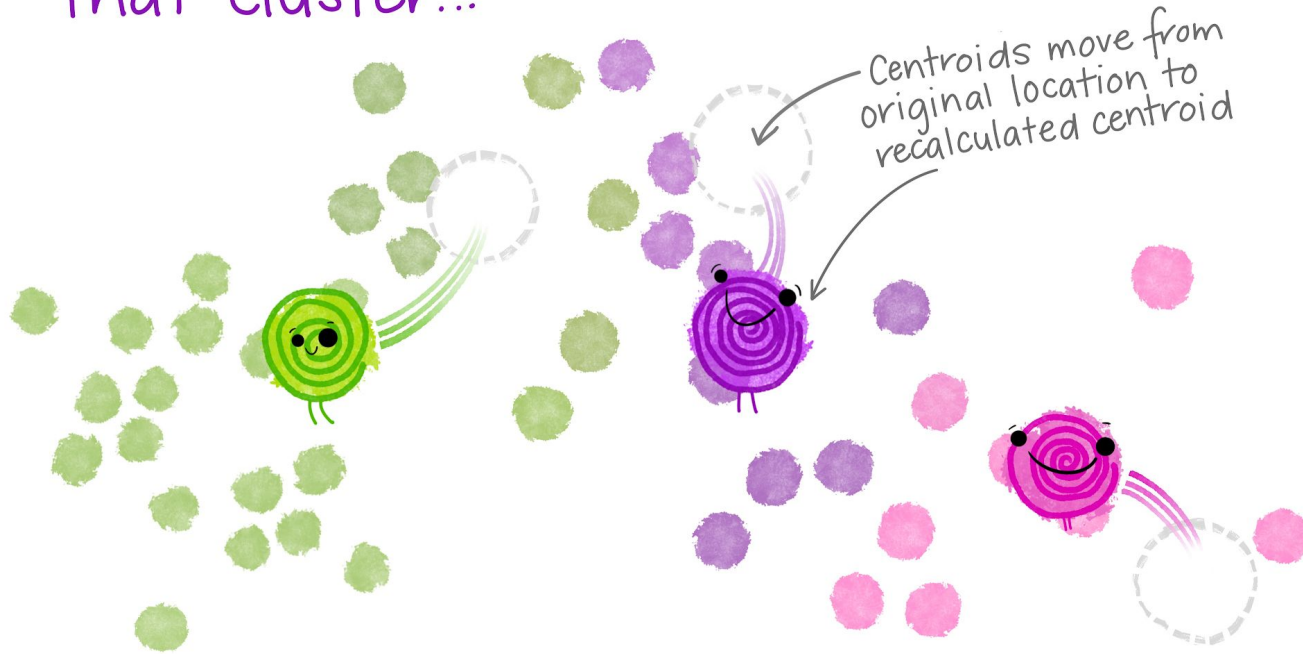
Each observation gets temporarily "assigned" to its closest centroid.

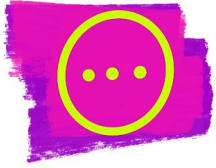
↖ (e.g. by Euclidean distance)



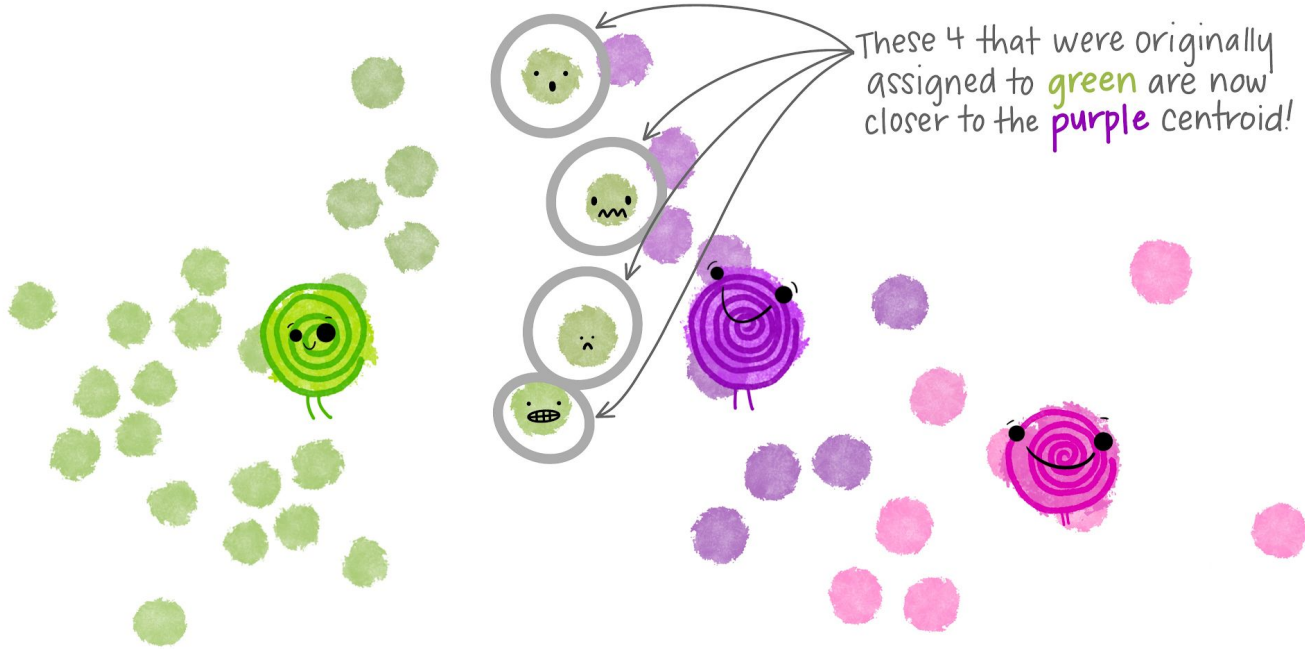
4

Then the centroid of each cluster is calculated based on all observations assigned to that cluster...





UH OH. Now that the cluster centroids have moved, some of the observations are now closer to a different centroid!

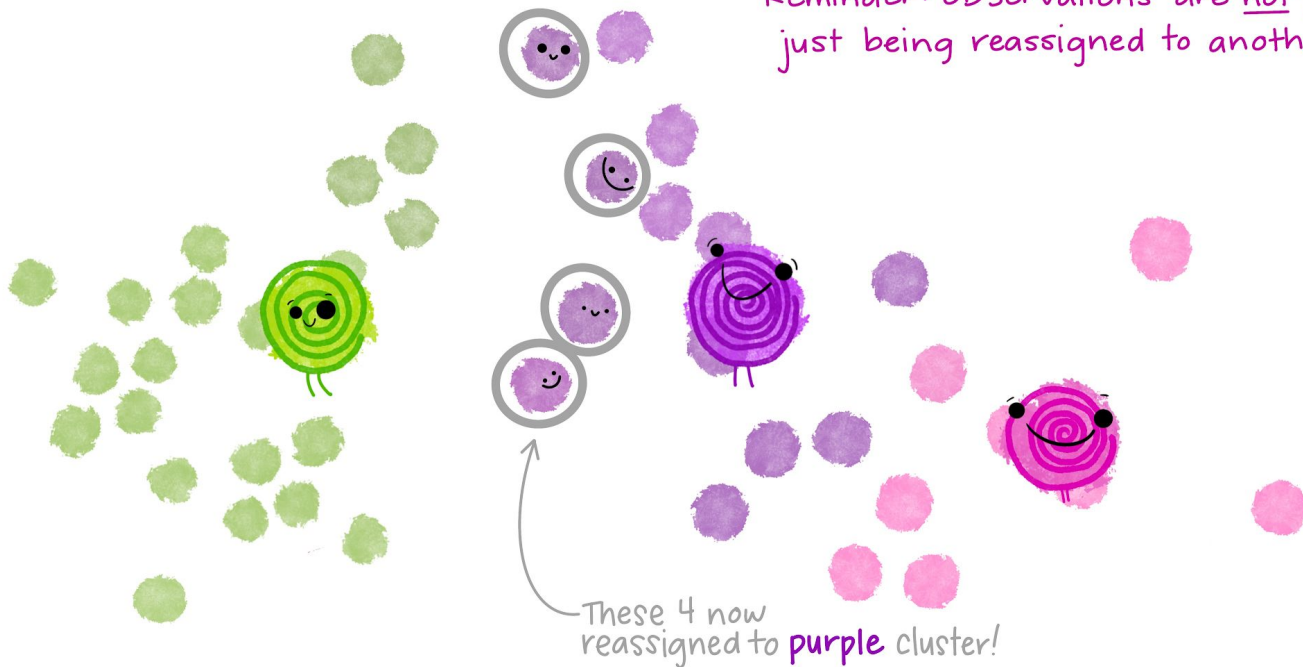


5

NO PROBLEM!

Observations get reassigned* to a different cluster based on the recalculated centroid.

*Reminder: observations are not moving, just being reassigned to another cluster.



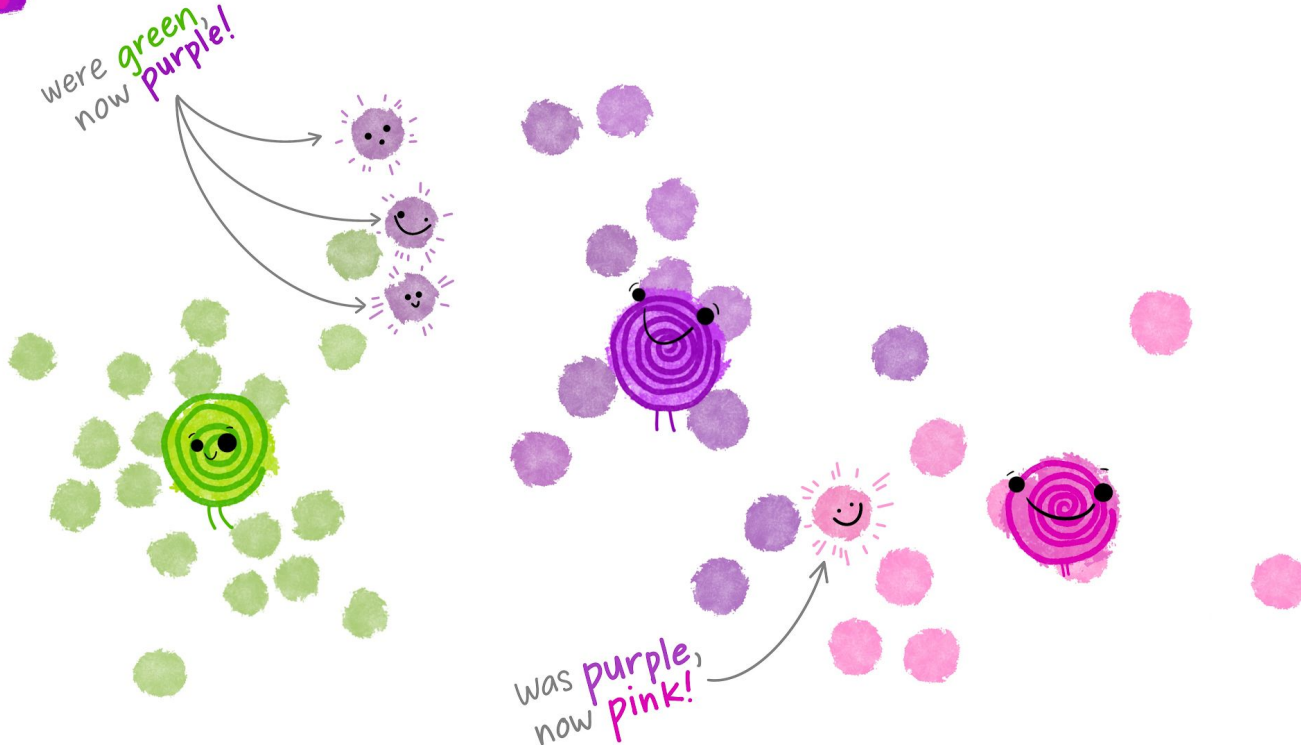
6

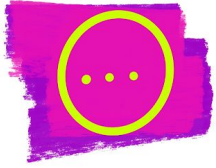
But now that observations have been reassigned,
the centroids need to move again [recalculate
centroids from updated clusters]





Again, now observations are reassigned as needed to the closest centroid.

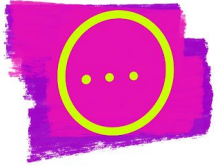




Then the centroid for each cluster
is recalculated...



...which means observations will be reassigned..



That iterative process of

Recalculate cluster centroids

→ Reassign observations to nearest centroid

→ Recalculate cluster centroids

→ Reassign observations to nearest centroid

→ Recalculate cluster centroids

→ Reassign observations to nearest centroid



Continues until nothing is moving
or being reassigned anymore!



Which means the iteration is done and each observation is assigned to its final cluster.

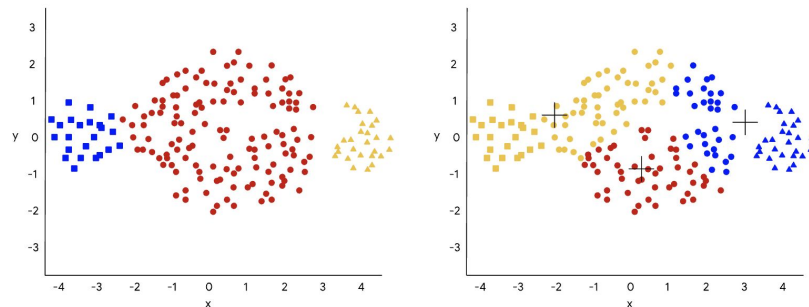


Advantages of k-means

- Relatively simple to implement.
- Scales to large data sets.
- Guarantees convergence.
- Can warm-start the positions of centroids.
- Easily adapts to new examples.
- Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

Disadvantages of k-means

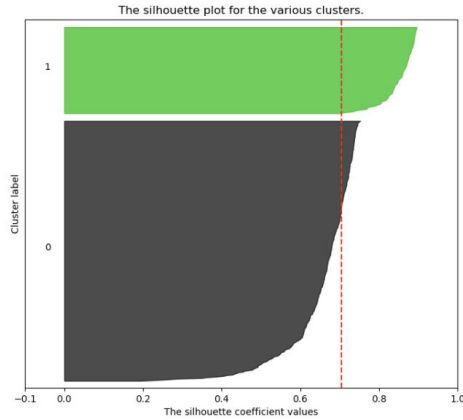
- Choosing k manually.
- Being dependent on initial values
- Clustering data of varying sizes and density.
- Clustering outliers.
- Scaling with number of dimensions.



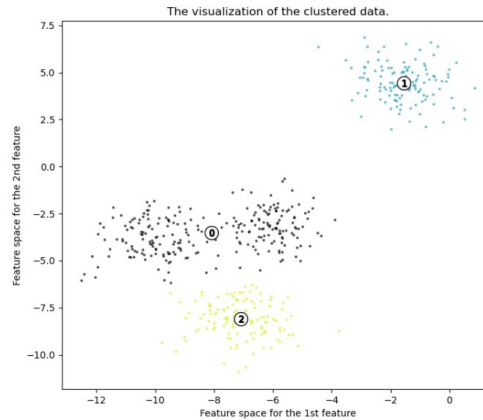
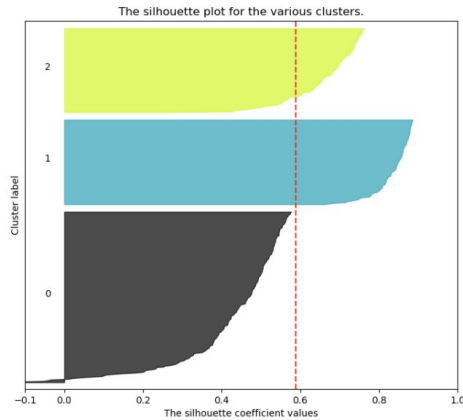
Selecting the number of clusters with silhouette analysis on KMeans clustering

- Can be used to study the **separation distance** between the resulting clusters.
- Silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters
- Provides a way to assess parameters like number of clusters visually.
- Measure has a range of $[-1, 1]$.
- **Silhouette coefficients** (as these values are referred to as) near **+1** indicate that the sample is far away from the neighboring clusters.
- A value of **0** indicates that the sample is on or very close to the decision boundary between two neighboring clusters
- **Negative values** indicate that those samples might have been assigned to the wrong cluster.

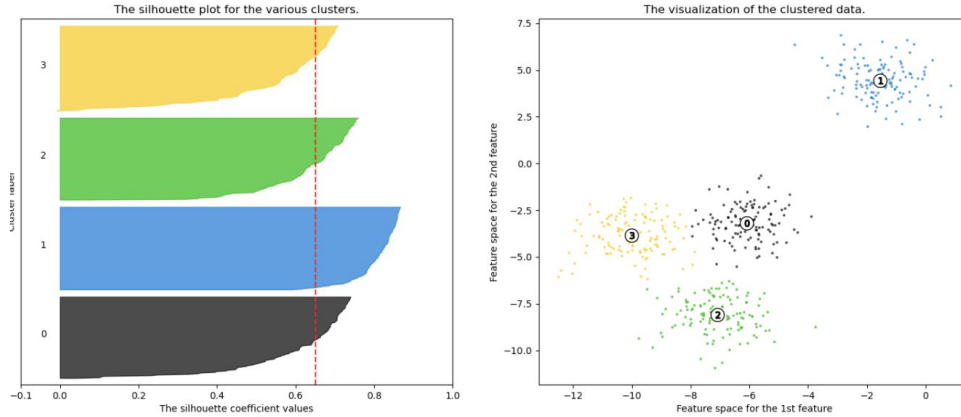
Silhouette analysis for KMeans clustering on sample data with $n_clusters = 2$



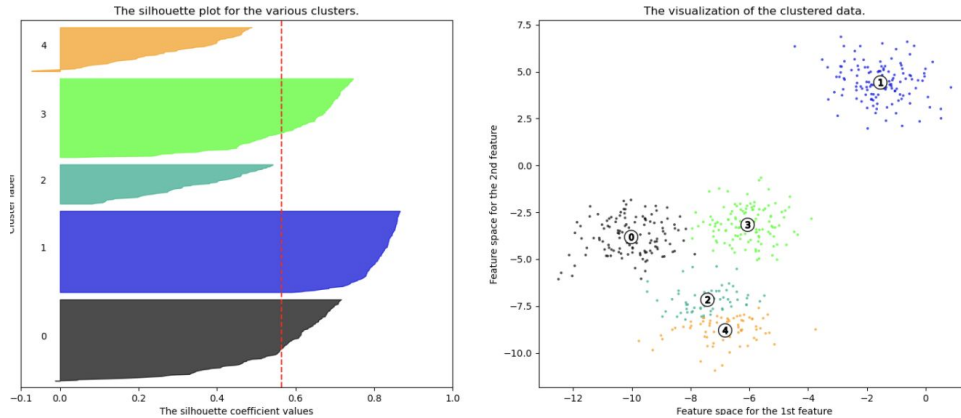
Silhouette analysis for KMeans clustering on sample data with $n_clusters = 3$



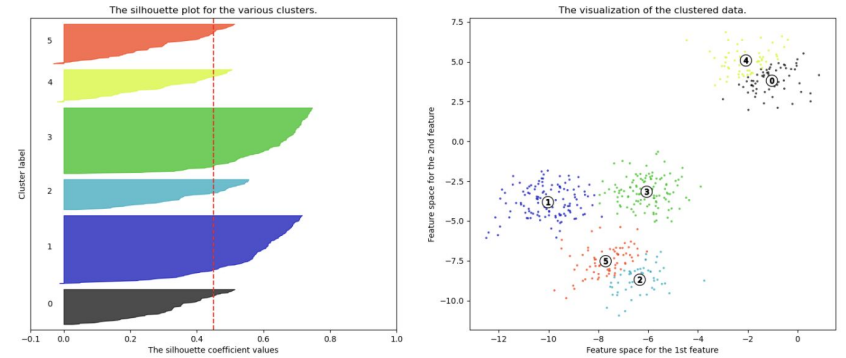
Silhouette analysis for KMeans clustering on sample data with $n_clusters = 4$



Silhouette analysis for KMeans clustering on sample data with $n_clusters = 5$



Silhouette analysis for KMeans clustering on sample data with $n_clusters = 6$



: For n_clusters = 2 The average silhouette_score is : 0.7049787496083262
For n_clusters = 3 The average silhouette_score is : 0.5882004012129721
For n_clusters = 4 The average silhouette_score is : 0.6505186632729437
For n_clusters = 5 The average silhouette_score is : 0.56376469026194
For n_clusters = 6 The average silhouette_score is : 0.4504666294372765