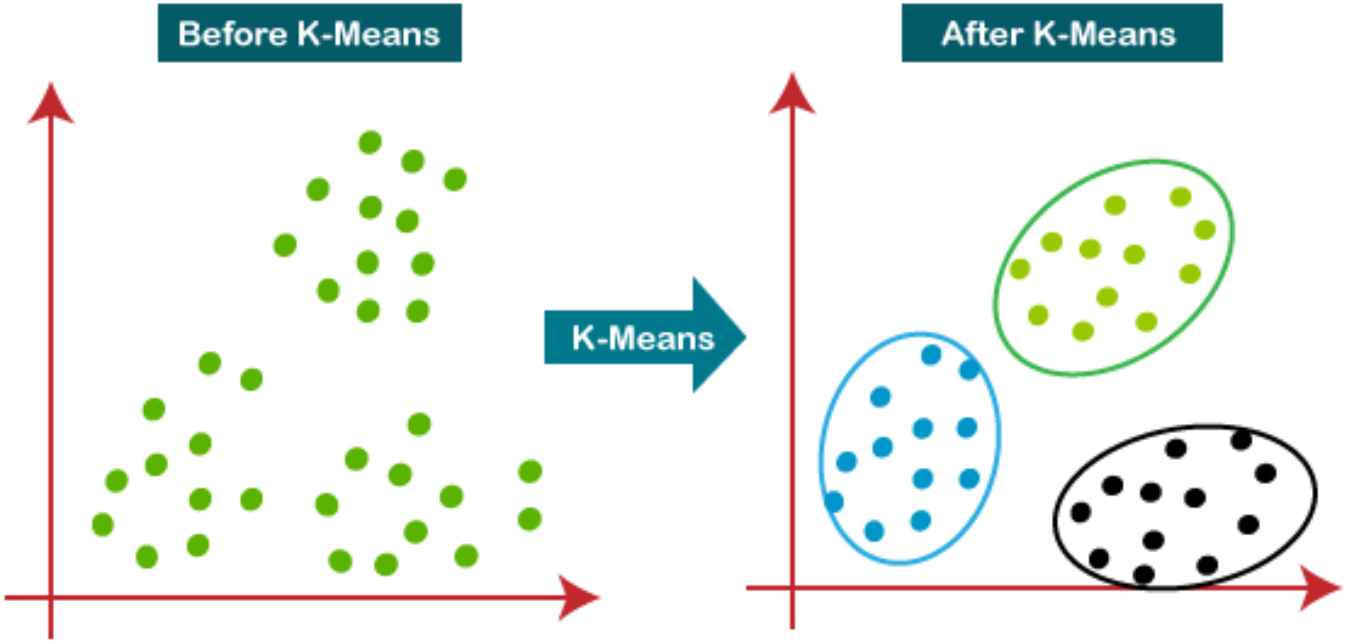
Machine Learning Clustering Algorithms

Introduction to Clustering

- Clustering is a fundamental task in machine learning used to group similar data points together.
- It's an essential tool for data exploration, pattern recognition, and information retrieval.
- This presentation covers 10 important clustering algorithms.

K-Means Clustering



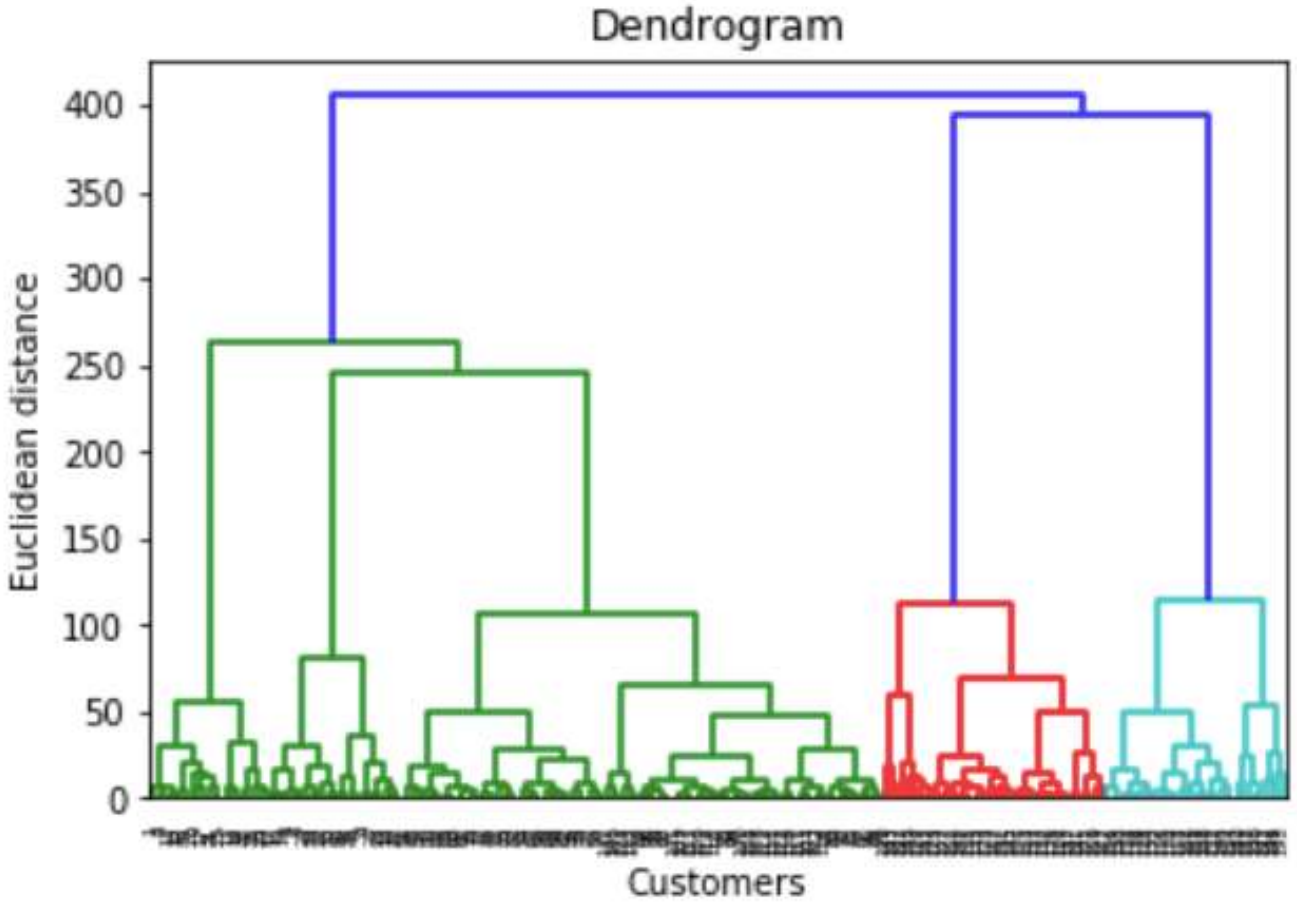
K-Means Clustering

- Working Principle:
- Assigns data points to the nearest cluster center (centroid). Iteratively adjusts the centroids and reassigns points until convergence.

- Advantages:
- Simple and fast. Works well with large datasets.

- Disadvantages:
- Sensitive to initial centroids. May converge to local minima.

Hierarchical Clustering



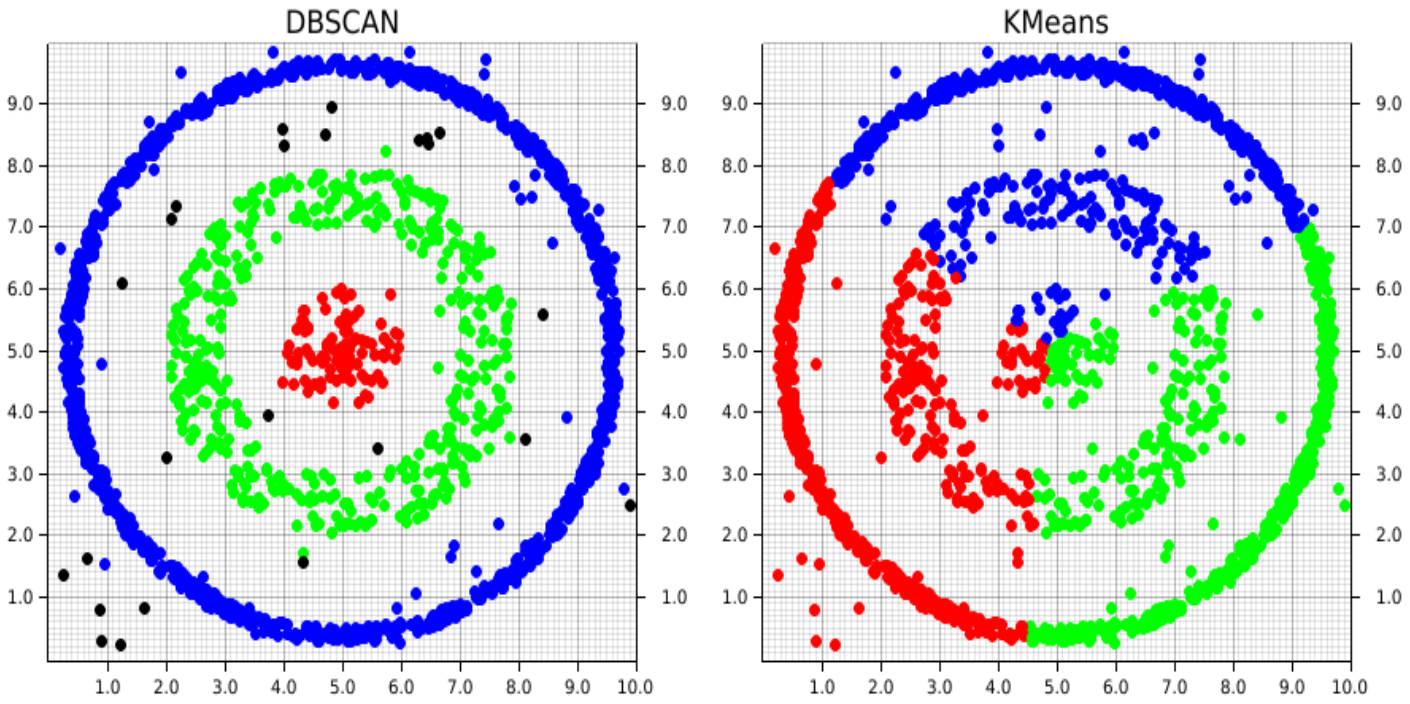
Hierarchical Clustering

- Working Principle:
- Builds a hierarchy of clusters by either merging or splitting them. Two types: Agglomerative (bottom-up) and Divisive (top-down).

- Advantages:
- No need to specify the number of clusters in advance. Can capture nested clusters.

- Disadvantages:
- Computationally expensive for large datasets. Sensitive to noise and outliers.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)



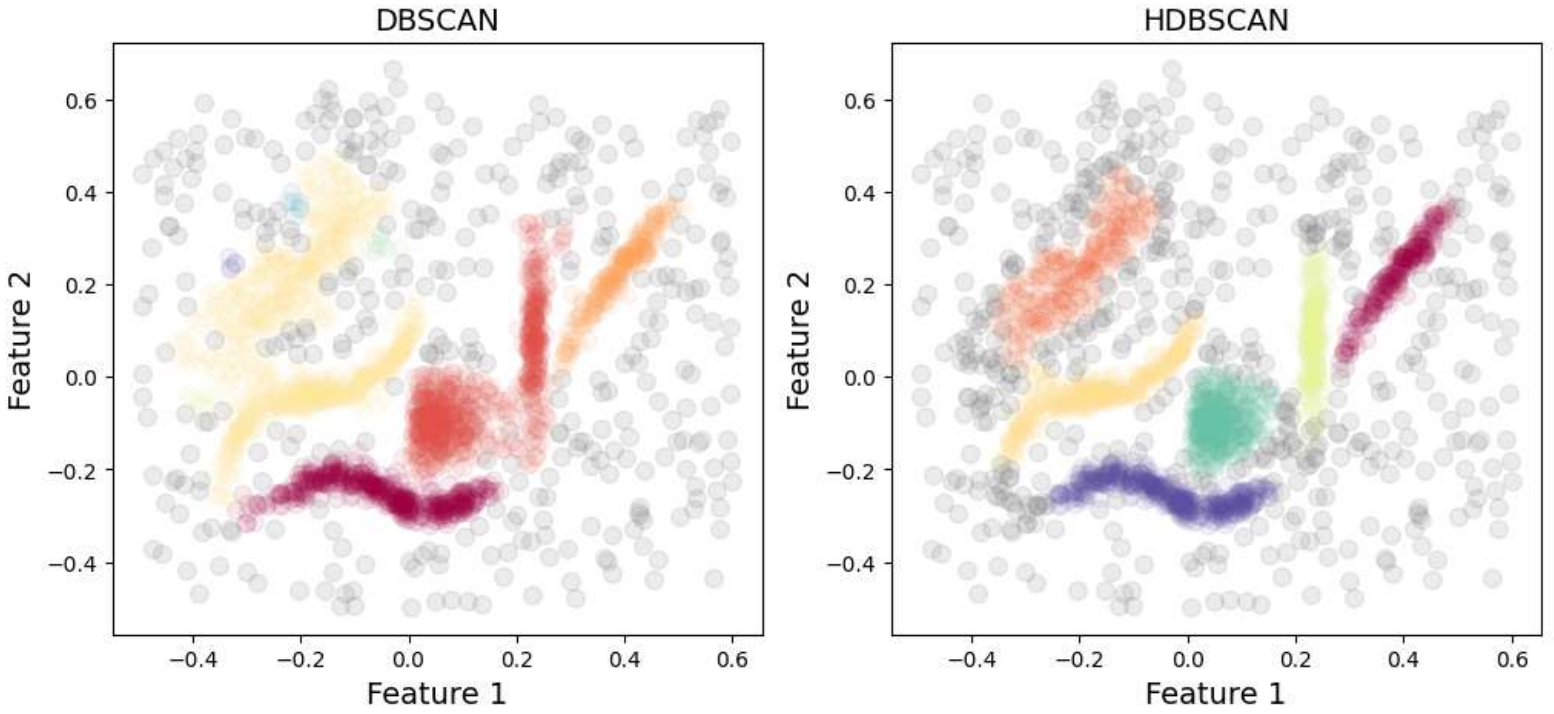
DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- Working Principle:
- Identifies clusters based on the density of points. Points in high-density areas form clusters; noise points are outliers.

- Advantages:
- Can find arbitrarily shaped clusters. Robust to outliers.

- Disadvantages:
- Does not work well with varying densities. Sensitive to parameter settings.

HDBSCAN(Hierarchical Density-Based Spatial Clustering of Applications with Noise)



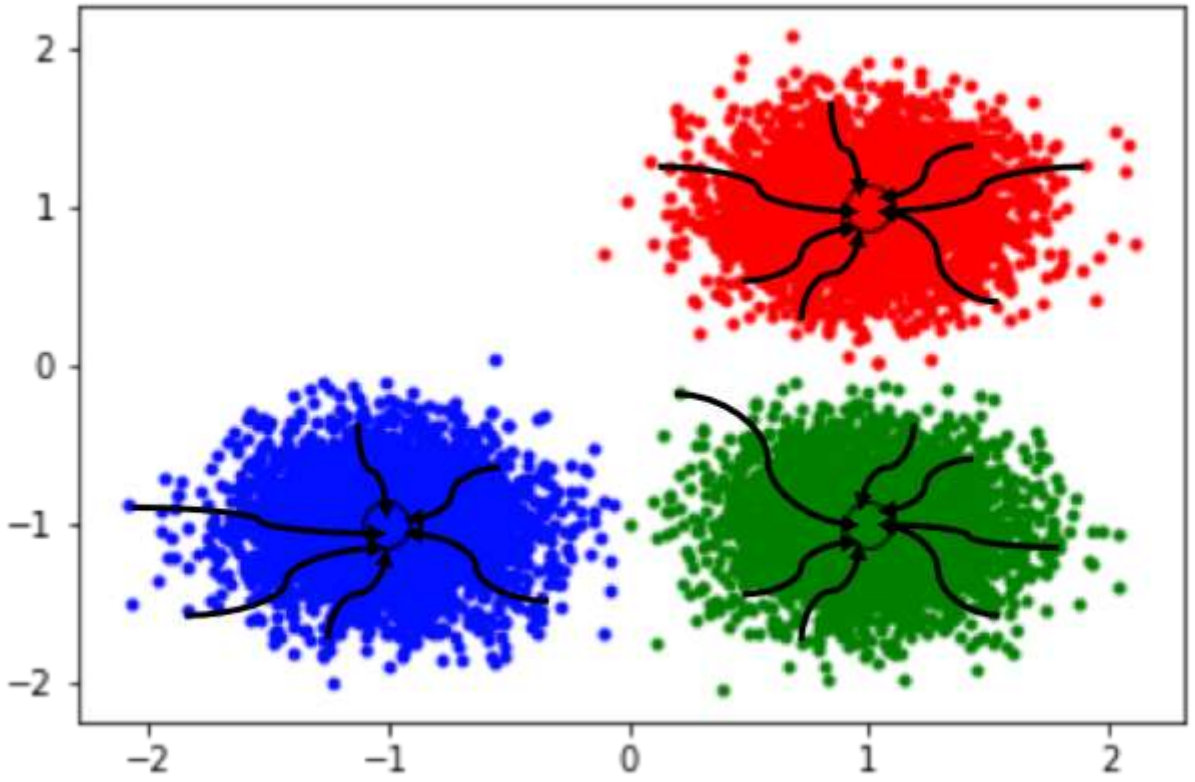
HDBSCAN(Hierarchical Density-Based Spatial Clustering of Applications with Noise)

- Working Principle:
- Converts the dataset into a minimum spanning tree. Identifies dense regions in the data and builds a hierarchy of clusters. Prunes the hierarchy to select the most significant clusters. Capable of identifying noise points (outliers).

- Advantages:
- Finds clusters of varying densities. Automatically determines the optimal number of clusters

- Disadvantages:
- May produce complex cluster hierarchies that are difficult to interpret.

Mean Shift Clustering



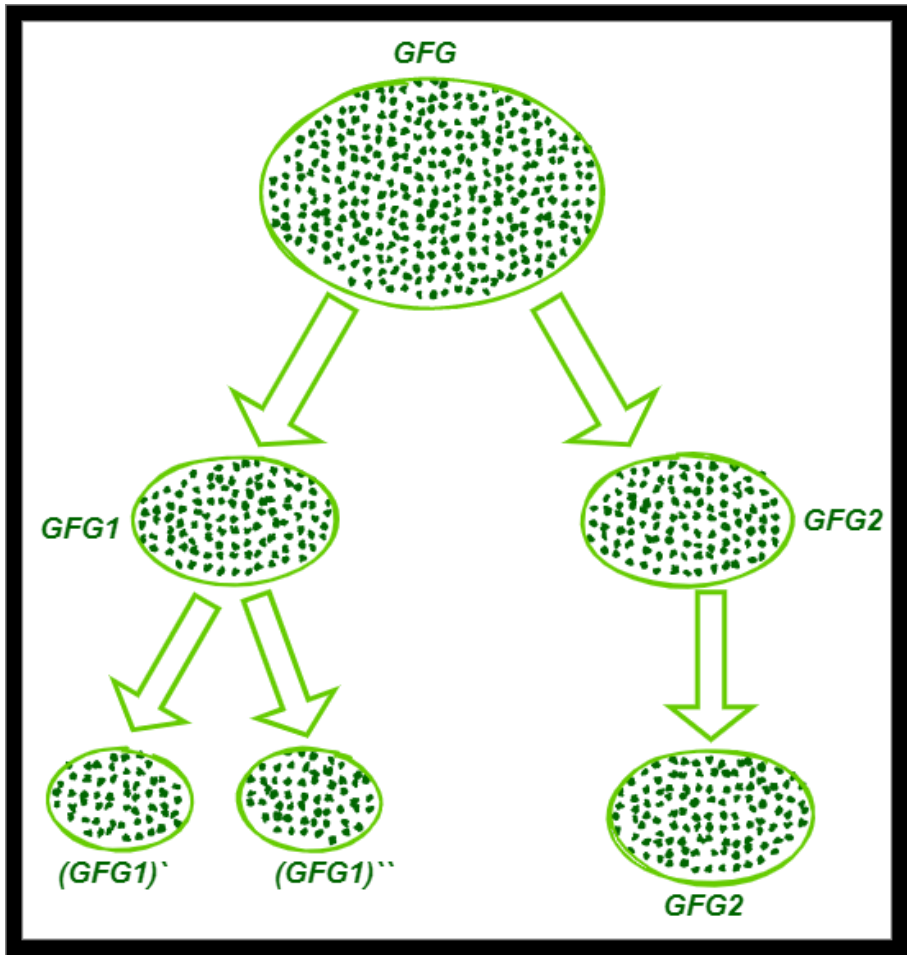
Mean Shift Clustering

- Working Principle:
- Iteratively shifts each data point to the mean of its neighbors. Clusters form at points where the data is dense.

- Advantages:
- Does not require specifying the number of clusters. Can identify clusters of arbitrary shape.

- Disadvantages:
- Computationally intensive. Choosing bandwidth (radius) is critical.

Bisecting K-Means

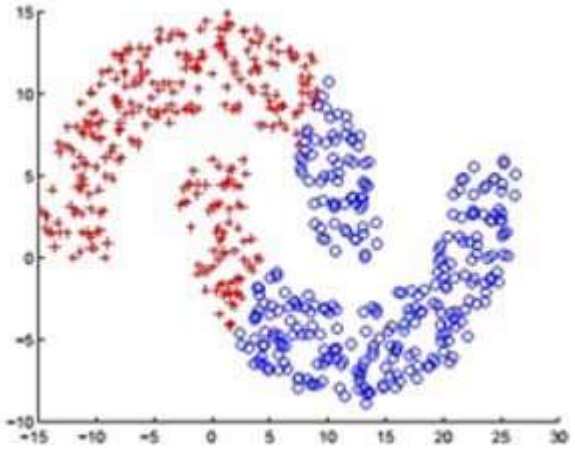


Bisecting K-Means

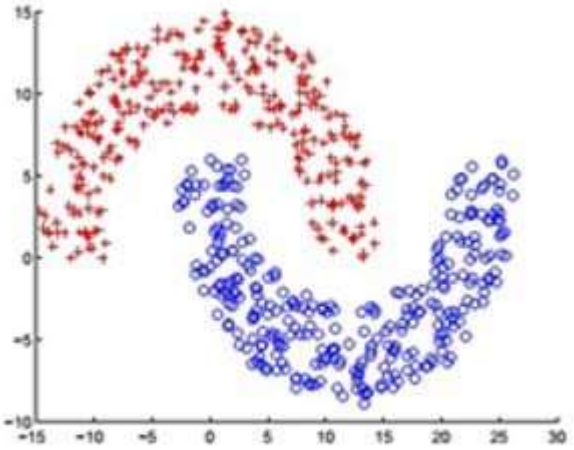
- Working Principle:
- Start with all points in one cluster. Repeat until the desired number of clusters is reached
- Advantages:
- Can model elliptical clusters. Provides a probabilistic cluster assignment.

- Disadvantages:
- Requires choosing the number of components (clusters). May not converge to the global optimum.

Spectral Clustering



(a) K-means



(b) Spectral Clustering

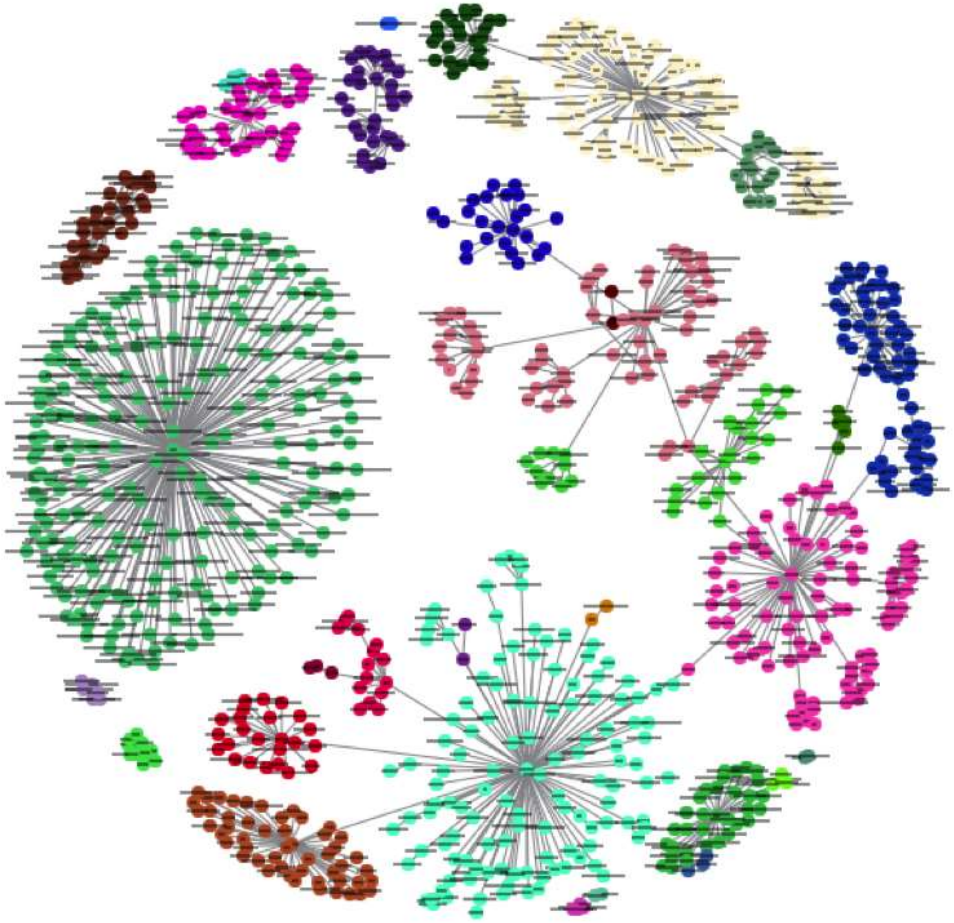
Spectral Clustering

- Working Principle:
- Uses graph theory and the spectrum of the similarity matrix to partition data. Clusters are identified from the eigenvectors of the graph Laplacian.

- Advantages:
- Can capture complex cluster structures. Effective with non-linear data.

- Disadvantages:
- Computationally expensive. Requires choosing the number of clusters.

Birch (Balanced Iterative Reducing and Clustering using Hierarchies)



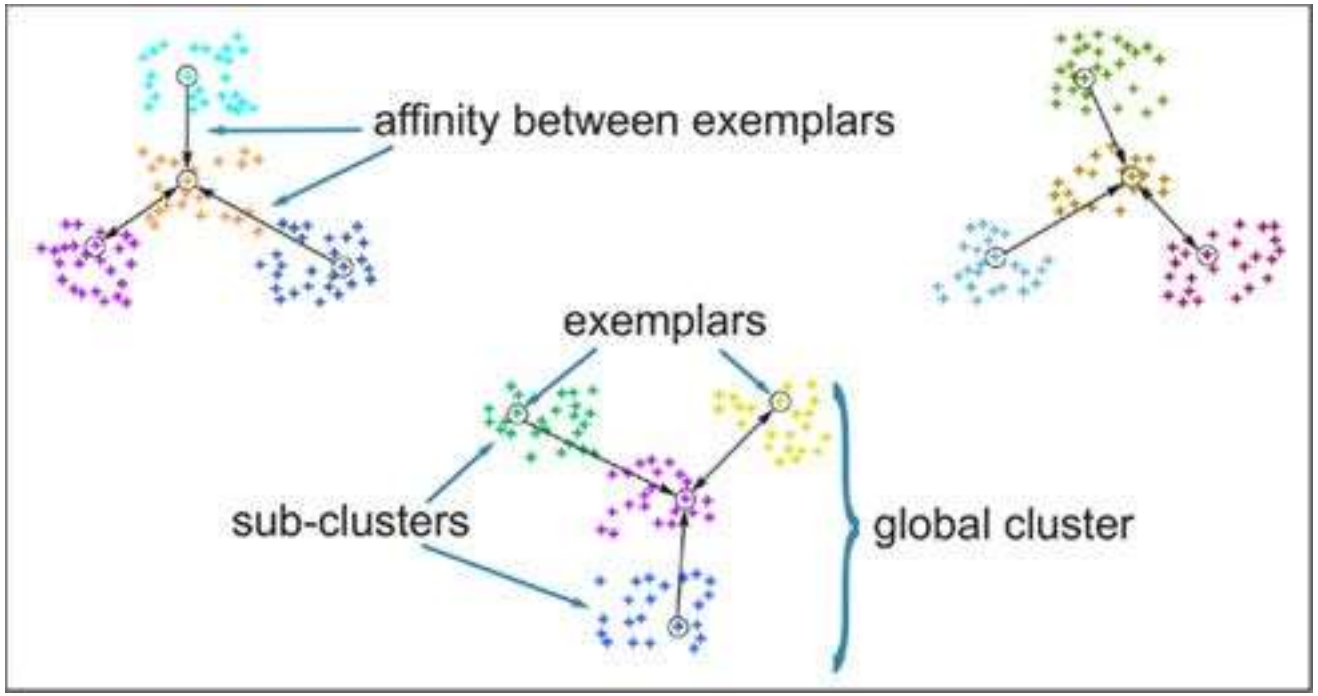
Birch (Balanced Iterative Reducing and Clustering using Hierarchies)

- Working Principle:
- Incrementally constructs a clustering feature tree (CF tree). Uses the CF tree to generate the final clusters.

- Advantages:
- Efficient with large datasets. Can handle noise effectively.

- Disadvantages:
- Sensitive to the threshold parameter. Less effective with very large datasets.

Affinity Propagation



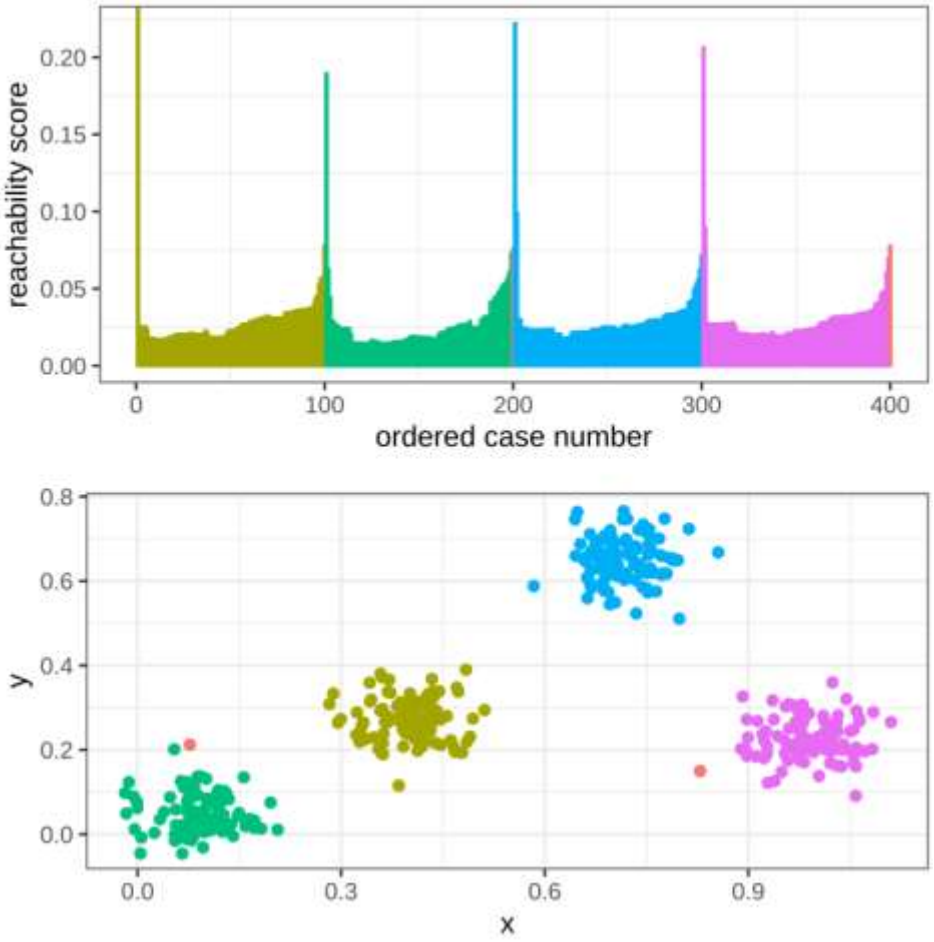
Affinity Propagation

- Working Principle:
- Exchanges messages between data points until convergence to identify exemplars. Clusters are formed around these exemplars.

- Advantages:
- Automatically determines the number of clusters. Handles clusters of various sizes.

- Disadvantages:
- Computationally expensive. Sensitive to the preference parameter.

OPTICS (Ordering Points To Identify the Clustering Structure)



OPTICS (Ordering Points To Identify the Clustering Structure)

- Working Principle:
- Similar to DBSCAN but generates an ordering of points to extract cluster structure. Can identify clusters of varying density.

- Advantages:
- Handles clusters of varying shapes and densities. No need to specify the number of clusters.

- Disadvantages:
- Sensitive to parameter selection. Interpretation of results can be complex.

Thank You