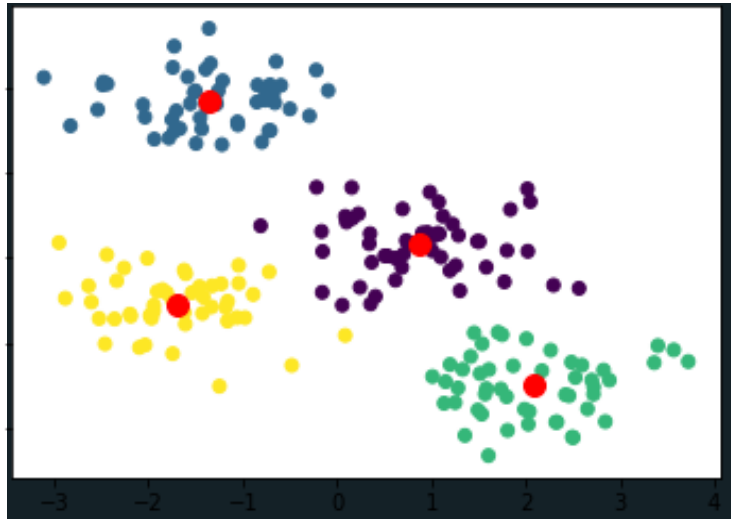
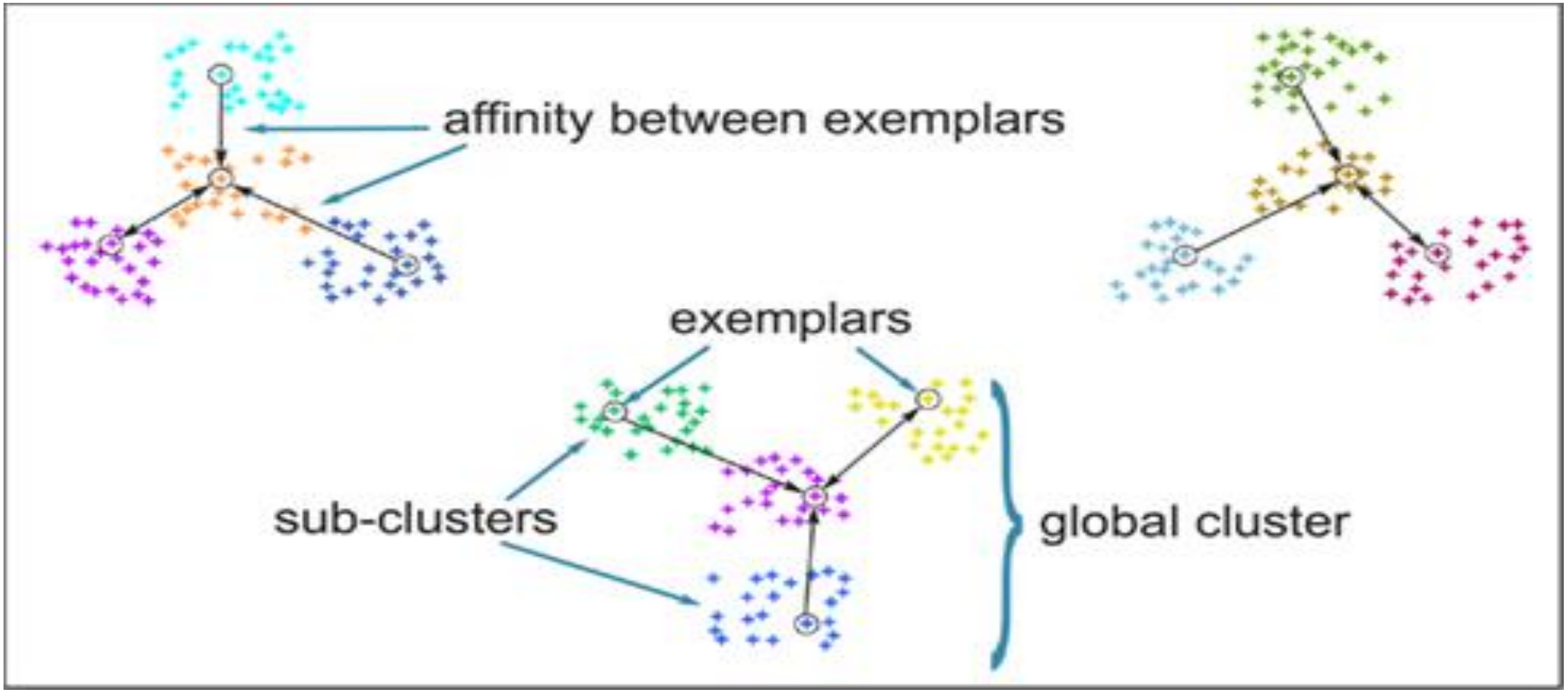


Clustering Methods & Applications

AFFINITY PROPAGATION

- Affinity propagation does not require the number of clusters to be determined before running the algorithm.
- Automatically determine the number of clusters.
- Datapoints in a dataset communicate with each other.
- Exemplar is a single datapoint represents all datapoints in each cluster.
- Exemplar then define clusters by finding which datapoint belongs to it.
- Datapoints are assigned to their cluster automatically by communicating with exemplar which is nearest to them.

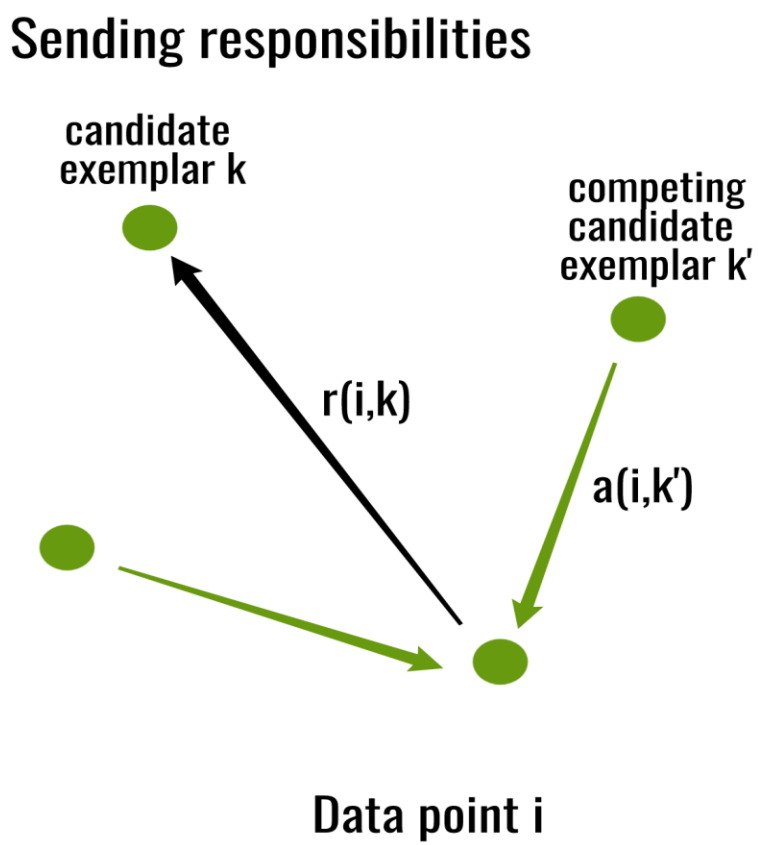
CLUSTER FORMATION BY EXEMPLAR

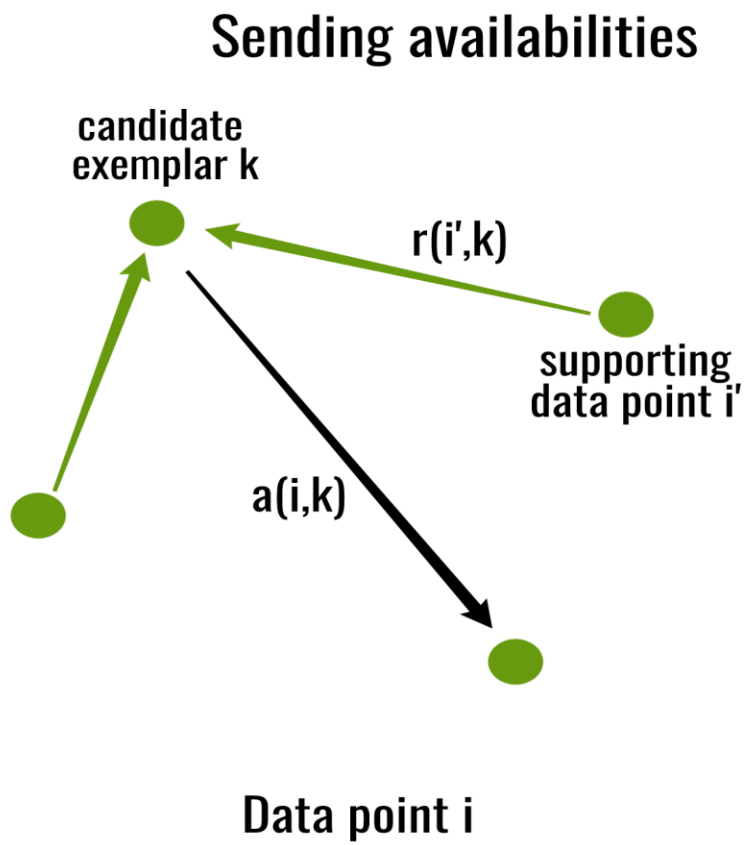


AFFINITY PROPAGATION - DATAPOINT SEGREGATION IN CLUSTERS

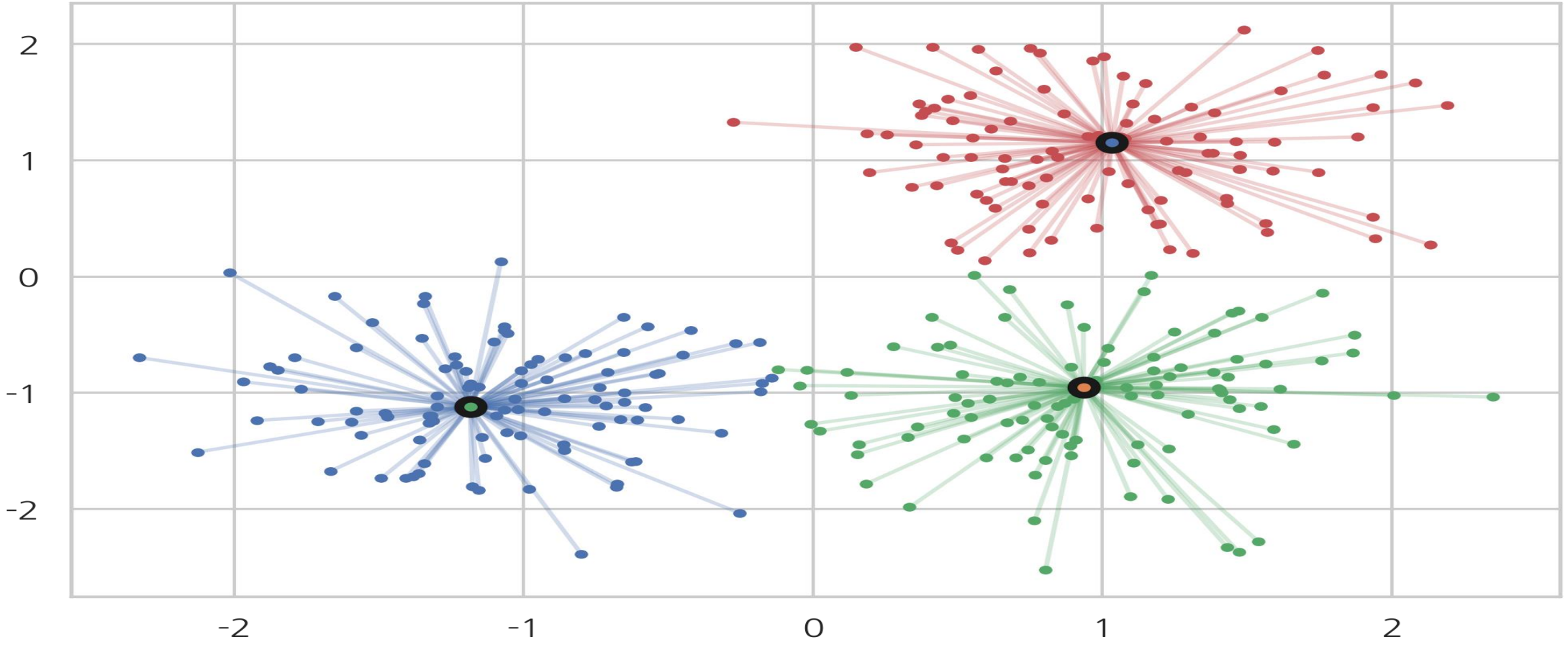
- Affinity propagation segregate the datapoints into appropriate clusters by the following two methods.
- Responsibility(r) is the measure of how well suited the datapoint(Convergence).
- Availability (a) reflects the accumulated evidence that a datapoint choose another(by calculating the shortest distance) as its exemplar (divergence).
- Responsibility and Availability flow will be shown in the below picture.

CLUSTER FORMATION PROCESS

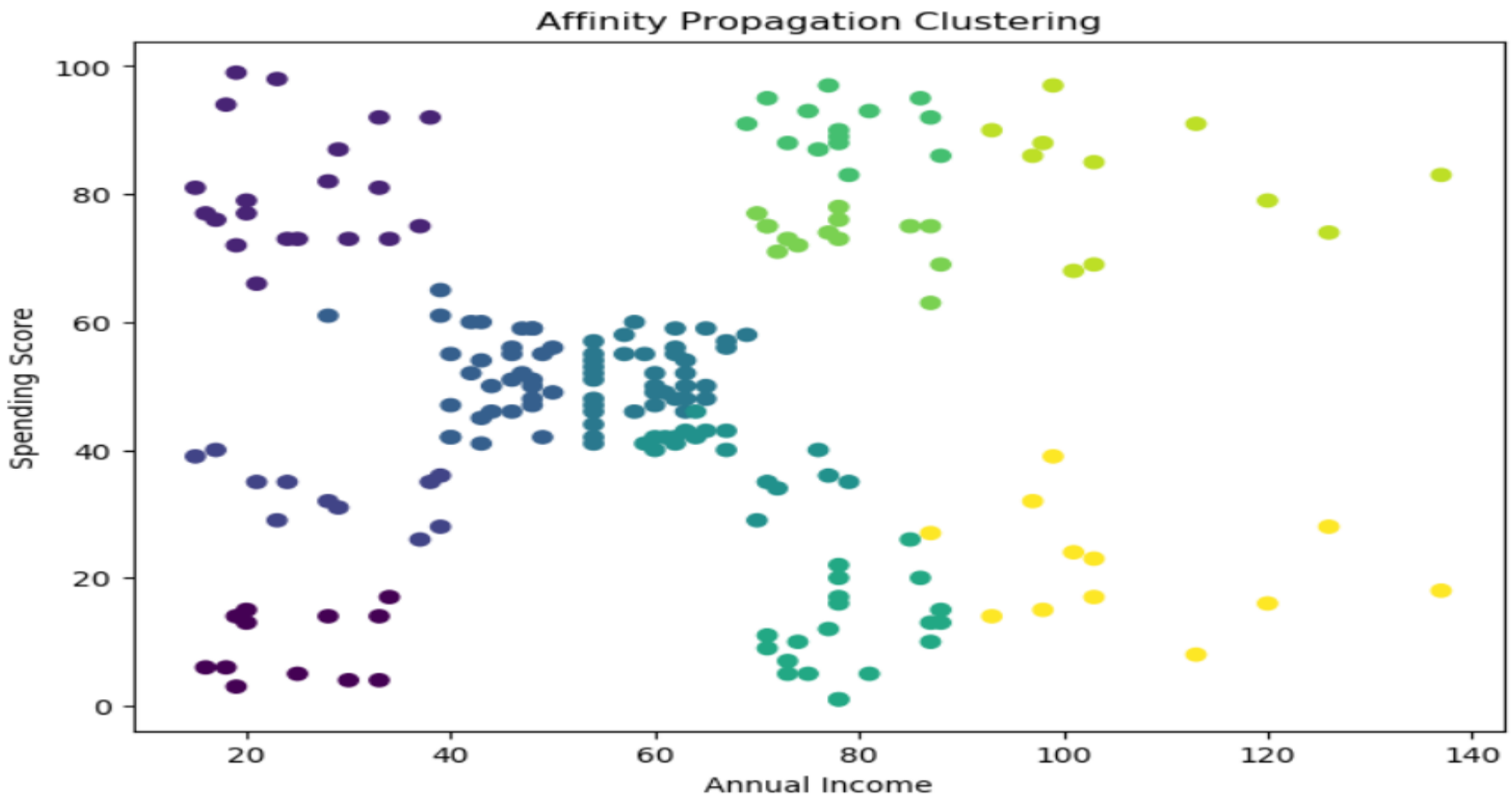




CLUSTER DIAGRAM



OUTPUT OF AFFINITY PROPAGATION FOR SAMPLE DATASET



APPLICATIONS OF AFFINITY PROPAGATION

Image Segmentation:

- Affinity Propagation can be used for segmenting images into meaningful regions or objects based on similarity measures between pixels or image patches. It's particularly useful in computer vision tasks where identifying distinct objects or regions in an image is essential.

Gene Expression Analysis:

- In bioinformatics, Affinity Propagation can be applied to analyze gene expression data for clustering genes or samples based on similarity in expression patterns. This can aid in identifying genes with similar functions or samples with similar characteristics.

Natural Language Processing (NLP):

- Affinity Propagation can be used in NLP tasks such as document clustering or topic modeling. It can cluster documents based on similarity in their content or identify representative documents (exemplars) for each cluster, which can facilitate tasks like document summarization or information retrieval.

Social Network Analysis:

- Affinity Propagation can be applied to analyze social networks by clustering individuals based on similarity in their interactions or attributes. This can help identify communities or influential nodes within the network, providing insights into social dynamics or information diffusion processes.

Market Segmentation:

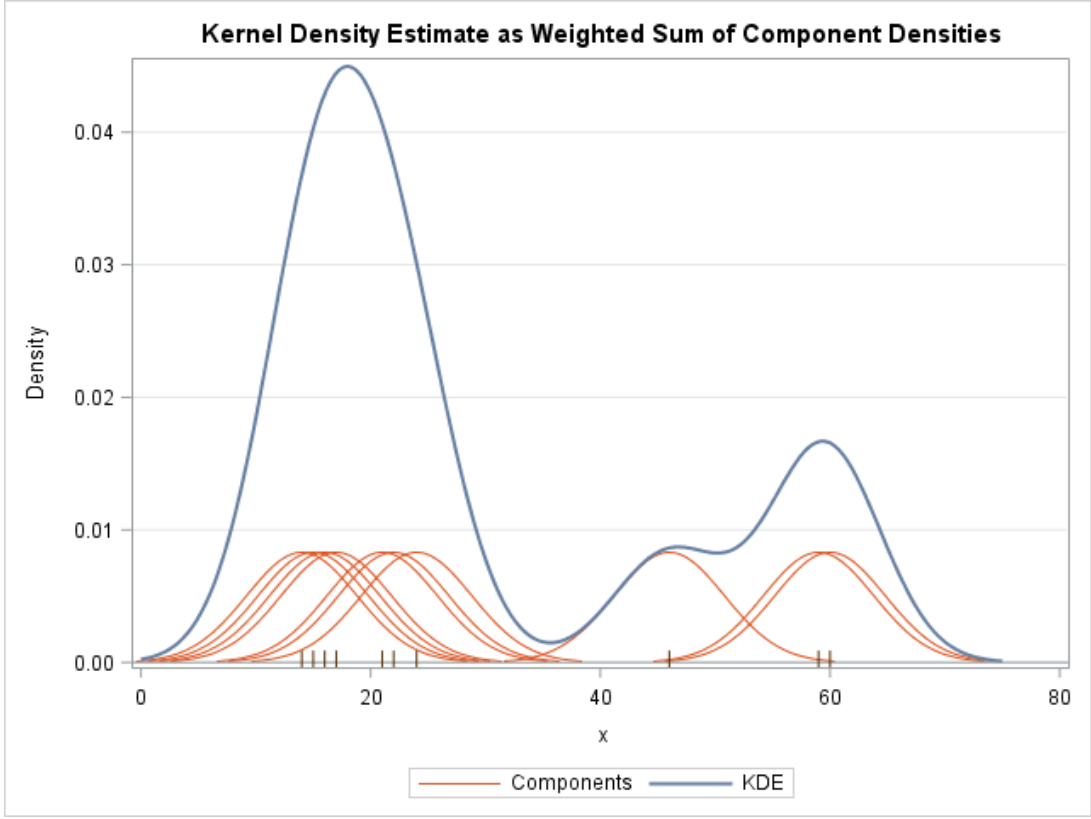
- In marketing analysis, Affinity Propagation can be used to segment customers based on their purchasing behavior, demographic characteristics, or preferences. This can aid businesses in targeted marketing strategies, product recommendations, and customer relationship management.

MEAN SHIFT CLUSTERING

- Mean shift clustering is a non-parametric clustering algorithm that aims to discover dense regions in a dataset.
- It does not require you to specify the number of clusters beforehand, making it an adaptive approach to data exploration.
- Mean shift algorithm is based on the concept of Kernel Density Estimation(KDE).

KERNAL DENSITY ESTIMATION (KDE)

- KDE is a method for estimating the underlying distribution of a set of datapoints.
- It would help you draw a continuous line that represents the density of those dots, showing where they are most likely to appear.
- It works by providing weights to each datapoint.



How Mean shift clustering works?

1.Initialization:

Start with datapoints assigned to cluster of their own.

2.Mode Seeking:

For each datapoint, calculate the mean shift vector, which points towards the densest region of points in its vicinity.

3.Update centroids:

This step effectively shifts datapoints towards denser regions of the dataspace.

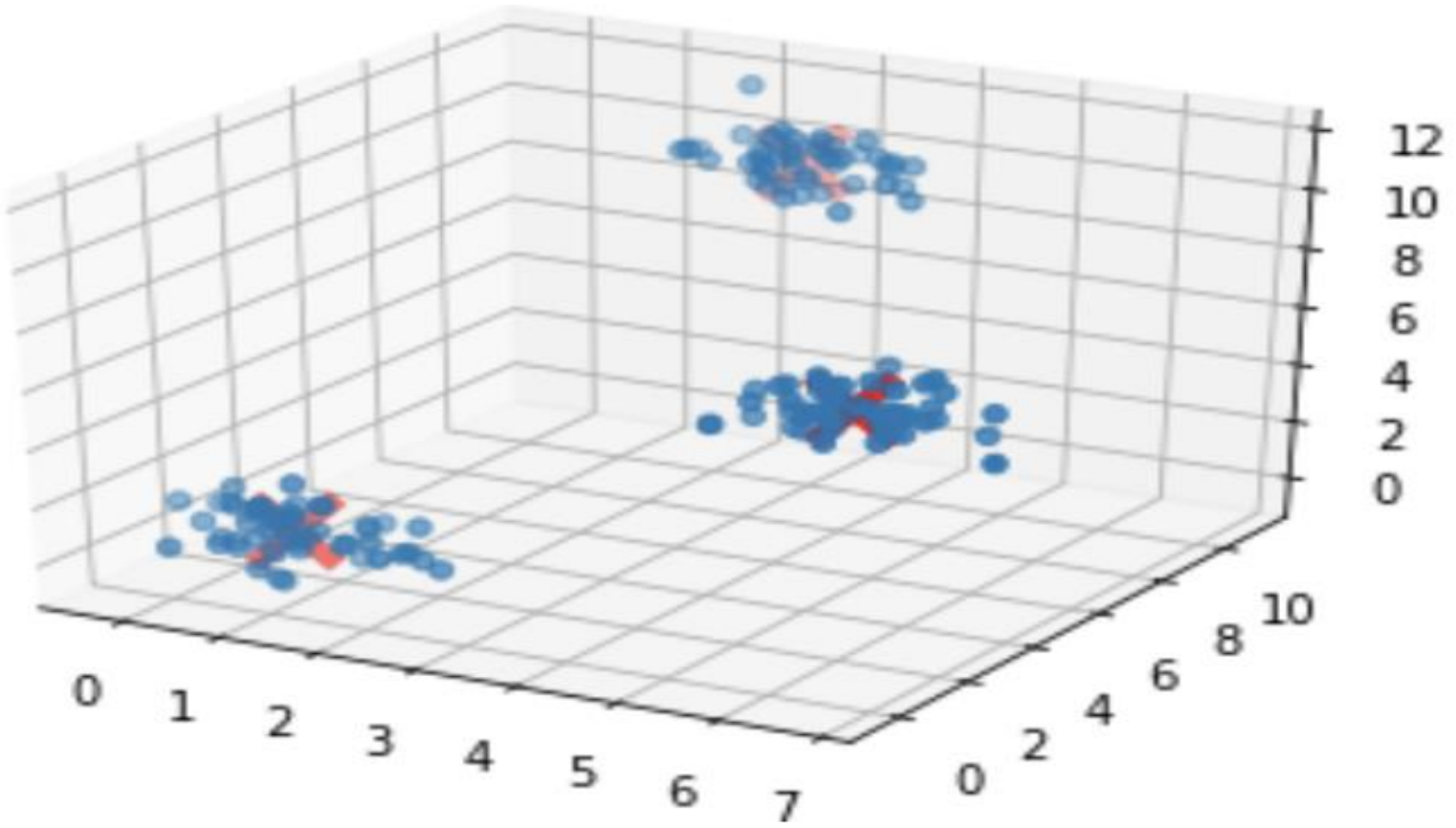
4.Iteration:

Repeat step 2 and 3 until convergence, which occurs when no datapoint moves significantly along its mean shift vector.

5.Cluster formation:

Group datapoints into clusters based on their final locations.

CLUSTER IMAGE



MEAN SHIFT CLUSTERING

ADVANTAGES

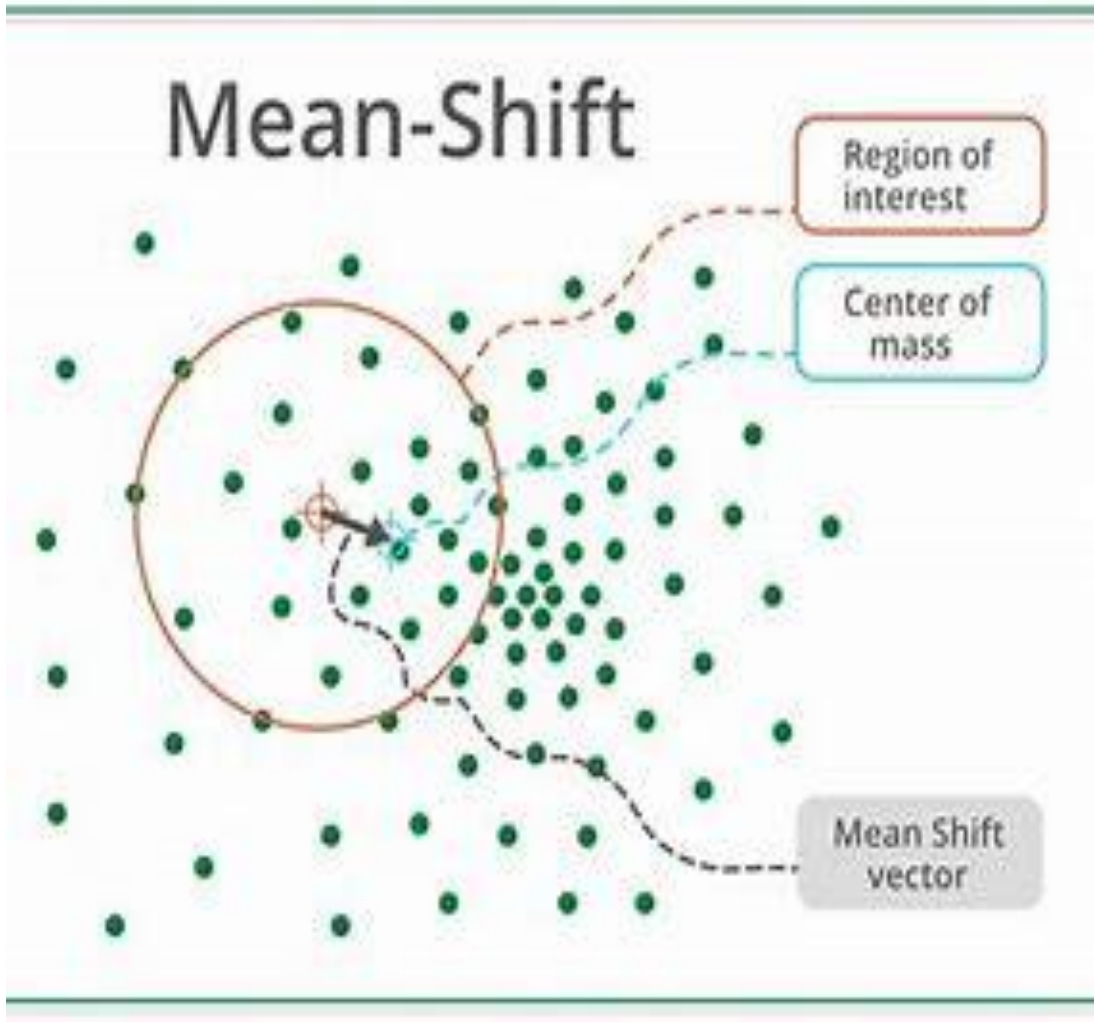
- It has only one parameter, which determines the number of clusters.
- There is no issue of local maxima and assumptions of total clusters.
- It also models the complex-shaped clusters.

DISADVANTAGES

- It doesn’t work in the case of high dimension, where clusters change unexpectedly.
- We cannot have direct control over the number of clusters.

APPLICATIONS

- IMAGE SEGMENTATION
- USER SEGMENTATION
- ANOMALY DETECTION



SPECTRAL CLUSTERING

Definition:

Spectral clustering is a powerful technique for clustering data points based on their similarity.

Overview:

It uses the eigenvalues of a similarity matrix to reduce the dimensionality of the data and then performs clustering in this reduced space.

- Spectral clustering is a versatile and powerful technique for clustering high-dimensional data.
- It offers advantages in terms of flexibility, scalability, and robustness compared to traditional clustering algorithms.
- Understanding spectral clustering can open doors to a wide range of applications in various fields.

How Spectral Clustering Works

Step 1: Construct a Similarity Matrix

- Measure pairwise similarities between data points (e.g., using Gaussian kernel or k-nearest neighbors).
- Represent similarities in a matrix form.

Step 2: Compute the Graph Laplacian

- Define a graph Laplacian matrix from the similarity matrix.
- Laplacian matrix captures the graph structure and data connectivity.

Step 3: Compute Eigenvalues and Eigenvectors

- Compute the eigenvalues and eigenvectors of the Laplacian matrix.
- Eigenvalues represent the "spectral" information of the data.

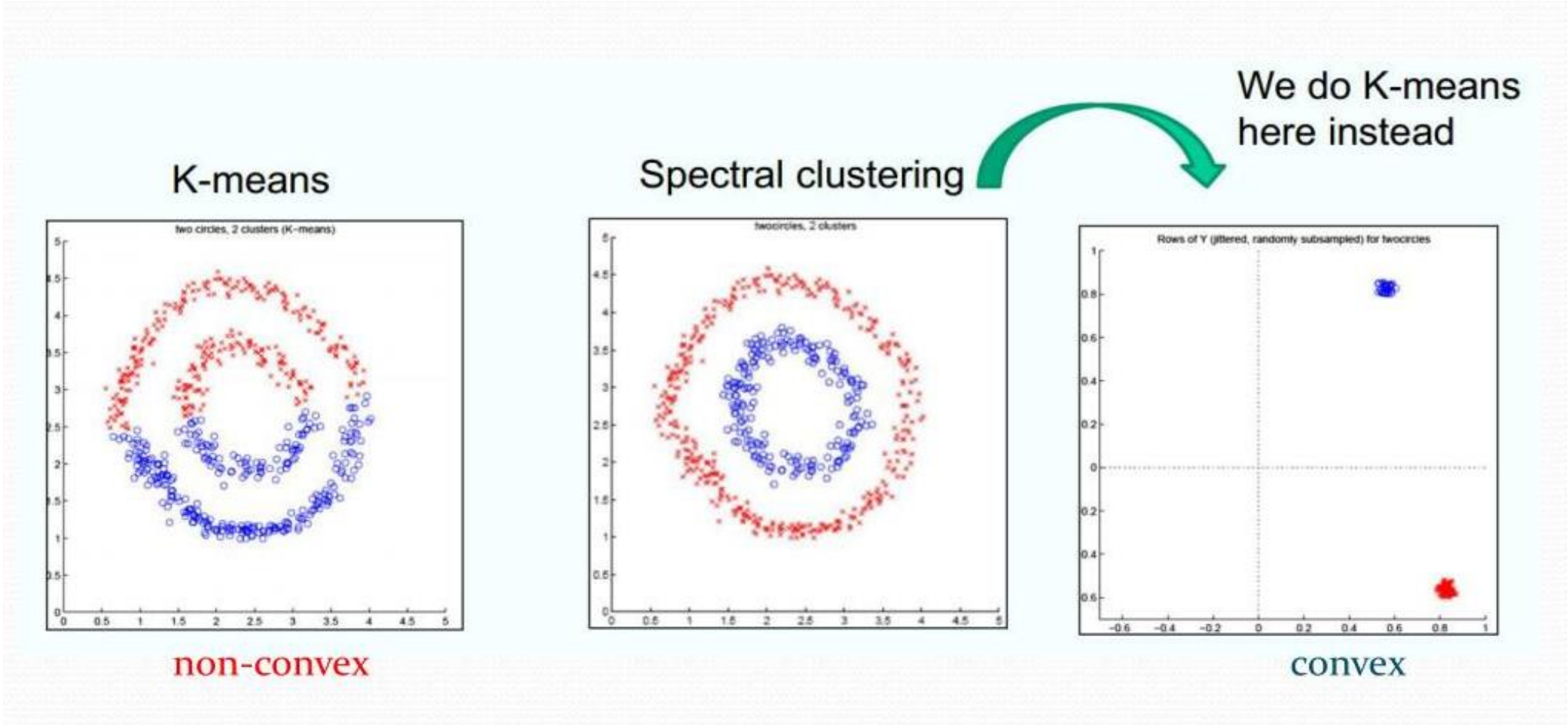
Step 4: Dimensionality Reduction

- Select a subset of eigenvectors corresponding to the smallest eigenvalues (low-frequency components).
- Form a new feature space using these eigenvectors.

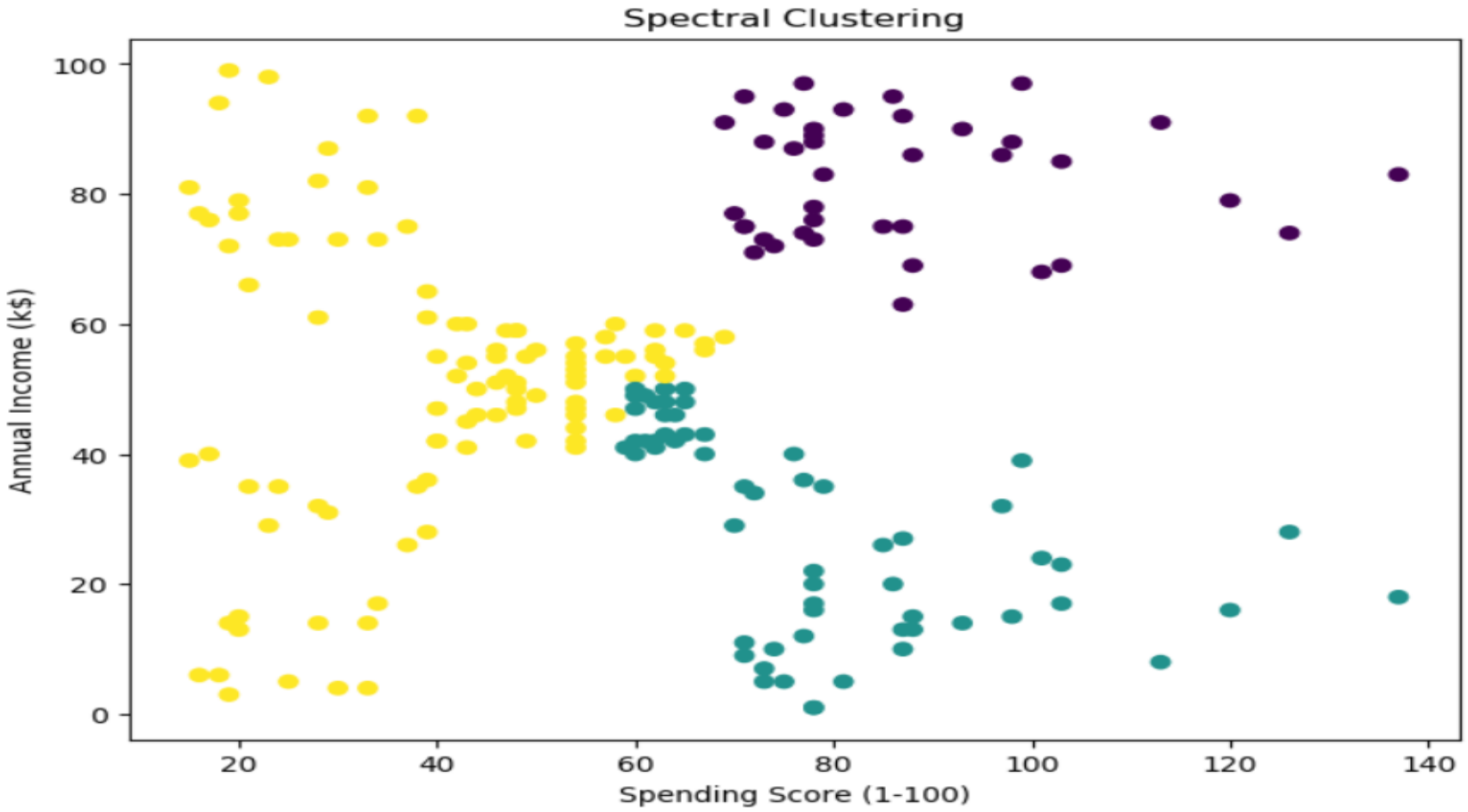
Step 5: Clustering

- Perform clustering (e.g., K-means) in the reduced feature space.
- Each data point is assigned to the cluster with the nearest centroid.

CLUSTER IMAGE



SPECTRAL OUTPUT



Advantages of Spectral Clustering

- **Flexibility:** Can handle non-linearly separable data and complex structures.
- **Scalability:** Suitable for large datasets due to dimensionality reduction.
- **Robustness:** Less sensitive to noise and outliers compared to traditional methods like K-means.

Applications of Spectral Clustering

- **Image Segmentation:** Partitioning images into meaningful regions based on pixel similarities.
- **Community Detection:** Identifying communities or clusters in social networks or biological networks.
- **Document Clustering:** Grouping similar documents based on content or topic similarity.

DBSCAN CLUSTERING

- **Definition:** DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular clustering algorithm in machine learning.
- **Overview:** It groups together closely packed points based on a notion of density, and it can discover clusters of arbitrary shape.
- DBSCAN is a powerful clustering algorithm suitable for various machine learning tasks.
- It offers advantages in terms of flexibility, robustness, and scalability compared to traditional clustering algorithms.
- Understanding DBSCAN can enhance the capability of machine learning practitioners in tackling clustering problems effectively.

HOW DBSCAN WORKS?

Core Concept: DBSCAN defines clusters as dense regions separated by sparser areas.

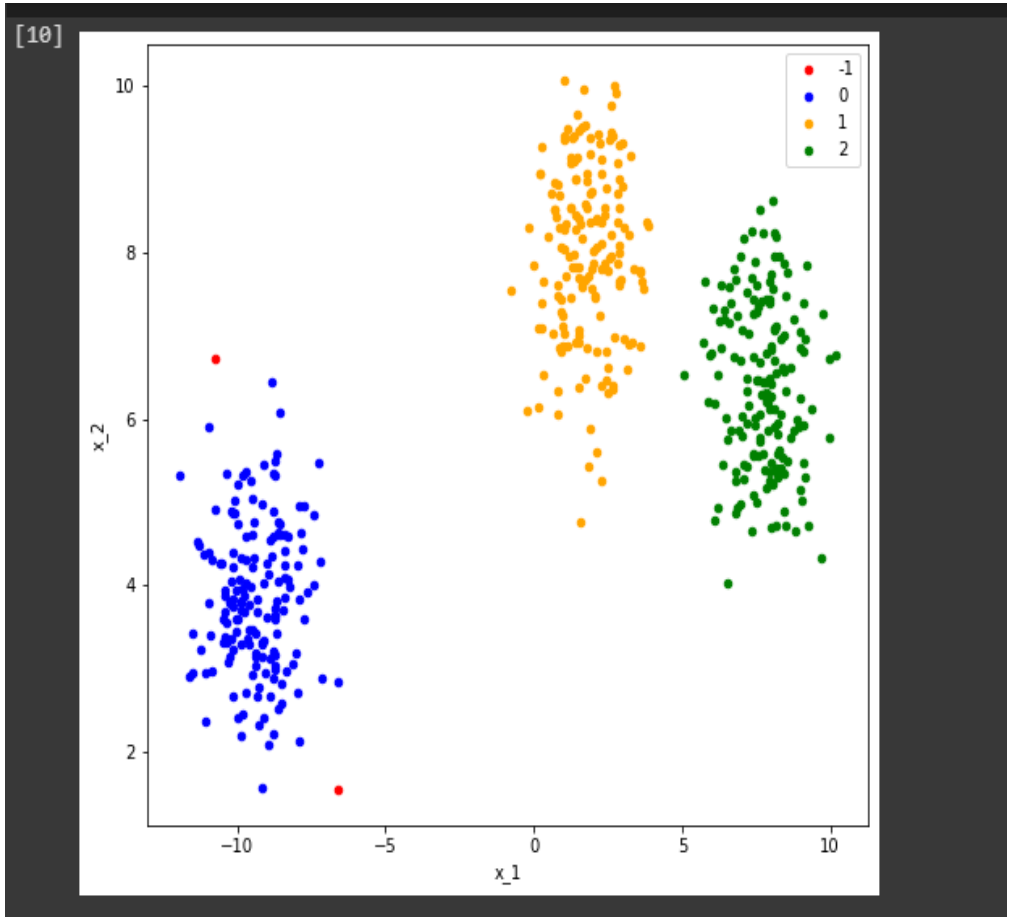
Parameters:

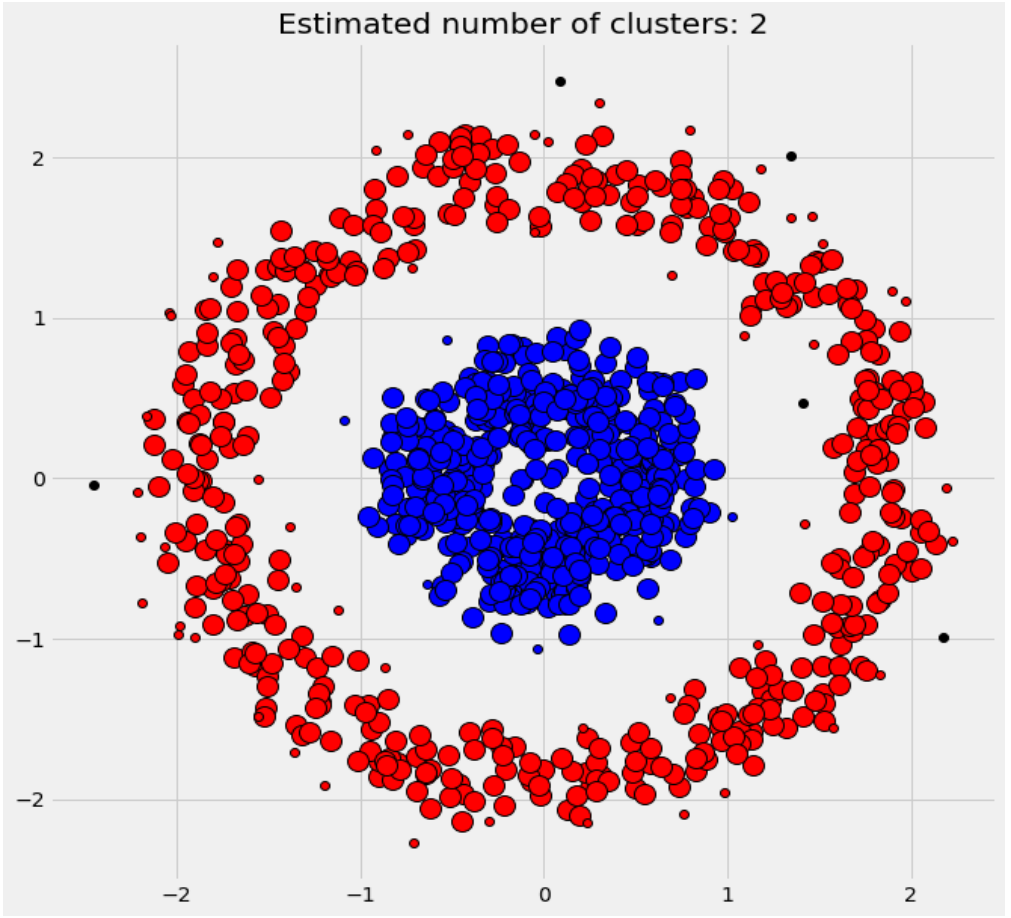
- **Epsilon (eps):** The maximum distance between two points for them to be considered as in the same neighborhood.
- **MinPts:** The minimum number of points required to form a dense region (core point).
- **Core Points:** Points with at least MinPts neighbors within a distance of eps.
- **Border Points:** Points within eps distance of a core point but with fewer than MinPts neighbors.
- **Noise Points:** Points that are neither core nor border points.

Algorithm:

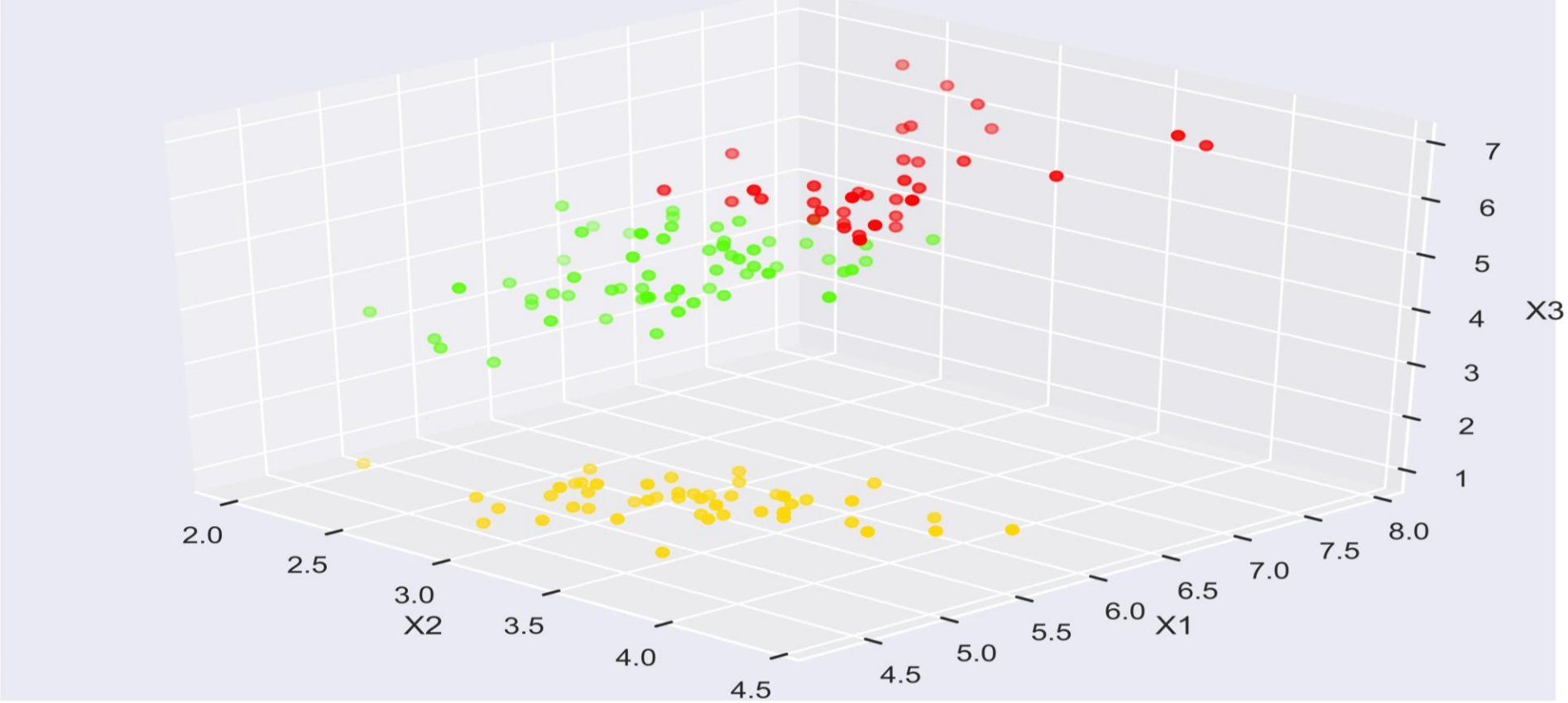
1. Randomly select a point not already assigned to a cluster.
2. If the point has at least MinPts neighbors within eps, start a new cluster with this point as a core point and expand it by adding all reachable points.
3. Repeat steps 1 and 2 until all points are assigned to a cluster or labeled as noise.

CLUSTER IMAGE





OUTPUT IMAGE



ADVANTAGES OF DBSCAN

- **Flexibility:** Can find clusters of arbitrary shape and handle noise well.
- **No Need for Pre-specifying Number of Clusters:** Automatically determines the number of clusters.
- **Robustness:** Less sensitive to outliers compared to K-means.
- **Suitable for Large Datasets:** Efficient algorithm for clustering large datasets.

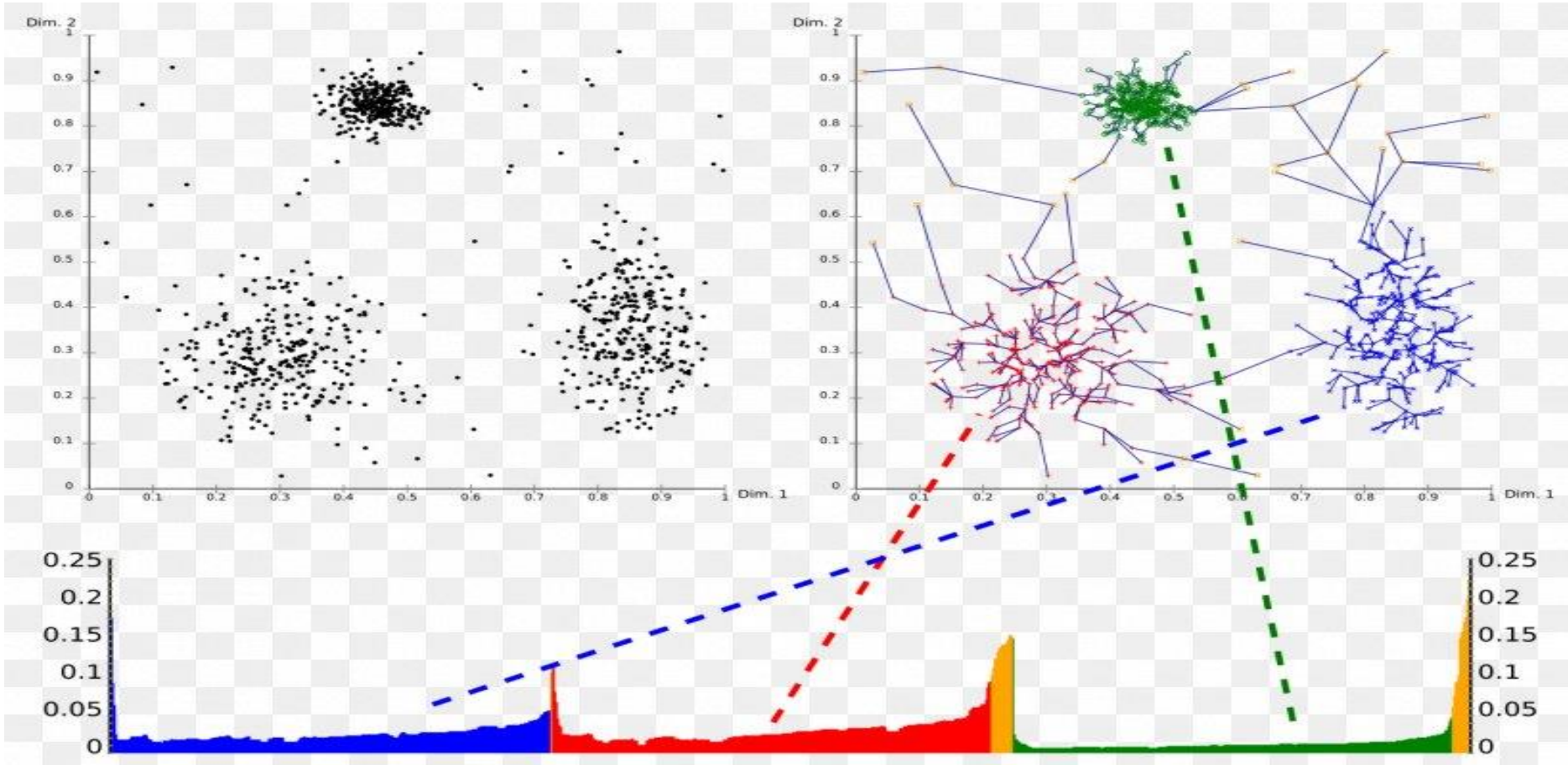
APPLICATIONS OF DBSCAN

- **Anomaly Detection:** Identifying outliers or anomalies in data.
- **Image Segmentation:** Partitioning images into regions based on pixel similarities.
- **Geographic Data Analysis:** Clustering geographical data to identify hotspots or regions of interest.

OPTICS CLUSTERING

- **Definition:** OPTICS (Ordering Points To Identify the Clustering Structure) is a density-based clustering algorithm in machine learning.
- **Overview:** It identifies clusters of arbitrary shape in a dataset by ordering points based on their density and connectivity.
- OPTICS is a versatile clustering algorithm suitable for various machine learning tasks.
- It offers advantages in terms of flexibility, robustness, and scalability compared to traditional clustering algorithms.
- Understanding OPTICS can enhance the capability of machine learning practitioners in exploring complex datasets and identifying meaningful patterns.

CLUSTER IMAGE



HOW OPTICS WORKS?

Core Concept: OPTICS extends DBSCAN by computing a reachability distance for each point, providing a more flexible approach to cluster identification.

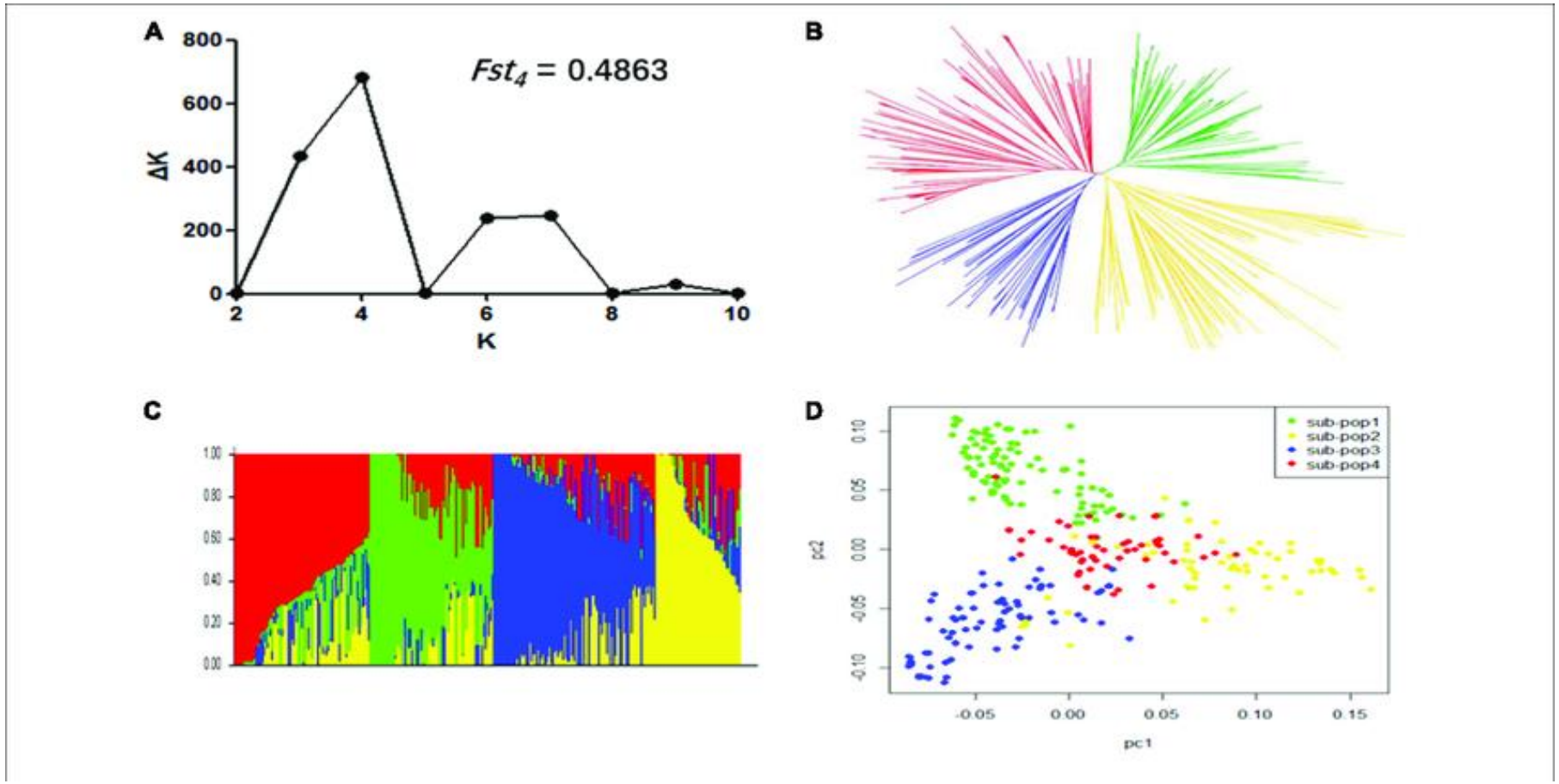
Parameters:

- MinPts: The minimum number of points required to form a cluster.
- Epsilon (eps): The maximum distance within which points are considered reachable.

Algorithm:

- Compute the reachability distance for each point in the dataset.
- Sort the points based on their reachability distances.
- Extract clusters by analyzing the reachability plot.
- Each valley in the reachability plot corresponds to a cluster, and points within the same valley belong to the same cluster.

OPTICS IMAGE



ADVANTAGES OF OPTICS

- **Flexibility:** Can handle clusters of arbitrary shape and varying densities.
- **Robustness:** Less sensitive to noise and outliers compared to K-means.
- No Need for Pre-specifying Number of Clusters: Automatically identifies clusters based on reachability distances.
- Suitable for Large Datasets: Scalable algorithm for clustering large datasets.

APPLICATIONS OF OPTICS

- **Spatial Data Analysis:** Clustering geographical data to identify spatial patterns or hotspots.
- **Anomaly Detection:** Identifying outliers or anomalies in high-dimensional data.
- **Network Analysis:** Detecting clusters or communities in social networks or biological networks.

BIRCH CLUSTERING

- Definition: BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is a hierarchical clustering algorithm in machine learning.
- Overview: It is designed to handle large datasets by incrementally clustering data points into a hierarchical structure.
- BIRCH is a powerful hierarchical clustering algorithm suitable for large-scale machine learning tasks.
- It offers advantages in terms of scalability, speed, and robustness compared to traditional clustering algorithms.
- Understanding BIRCH can enhance the capability of machine learning practitioners in handling big data and extracting meaningful insights.

HOW BIRCH WORKS?

Core Concept: BIRCH constructs a tree-like structure called the Clustering Feature Tree (CF Tree) to represent the dataset in a compact form.

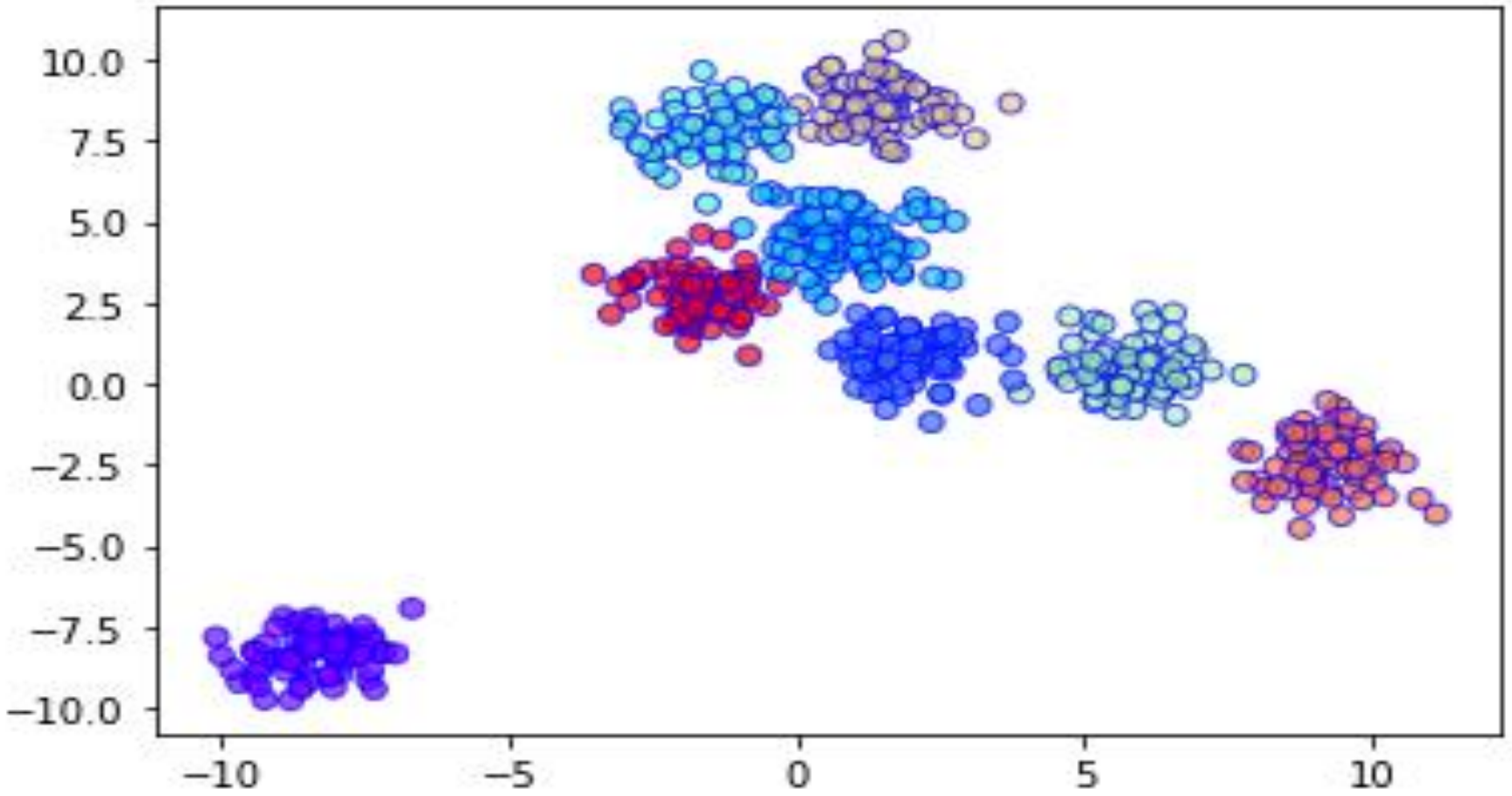
Parameters:

- Branching Factor: Maximum number of children allowed in each node of the CF Tree.
- Threshold: Maximum distance between points in a cluster.

Algorithm:

- Initialize an empty CF Tree.
- Incrementally insert data points into the CF Tree.
- Merge CF Tree nodes that satisfy the threshold condition.
- Extract clusters from the CF Tree.

CLUSTER IMAGE



ADVANTAGES OF BIRCH

- **Scalability:** Suitable for large datasets due to its efficient memory usage and incremental processing.
- **Speed:** Fast clustering algorithm, particularly for high-dimensional data.
- **Robustness:** Less sensitive to noise and outliers compared to K-means.
- **No Need for Pre-specifying Number of Clusters:** Automatically generates a hierarchical structure.

APPLICATIONS OF BIRCH

- **Customer Segmentation:** Clustering customers based on purchasing behaviour or demographic attributes.
- **Web Document Clustering:** Grouping similar documents for information retrieval or recommendation systems.
- **Network Traffic Analysis:** Identifying patterns or anomalies in network traffic data.