Clustering-Machine Learning

What is mean by Clutering?

Clustering is an unsupervised machine learning technique designed to group unlabeled examples based on their similarity to each other.

Application:

Clustering is useful in a variety of industries. Some common applications for clustering:

1. Market segmentation,
2. Social network analysis,
3. Search result grouping,
4. Medical imaging,
5. Image segmentation,
6. Anomaly detection.

Types of clustering

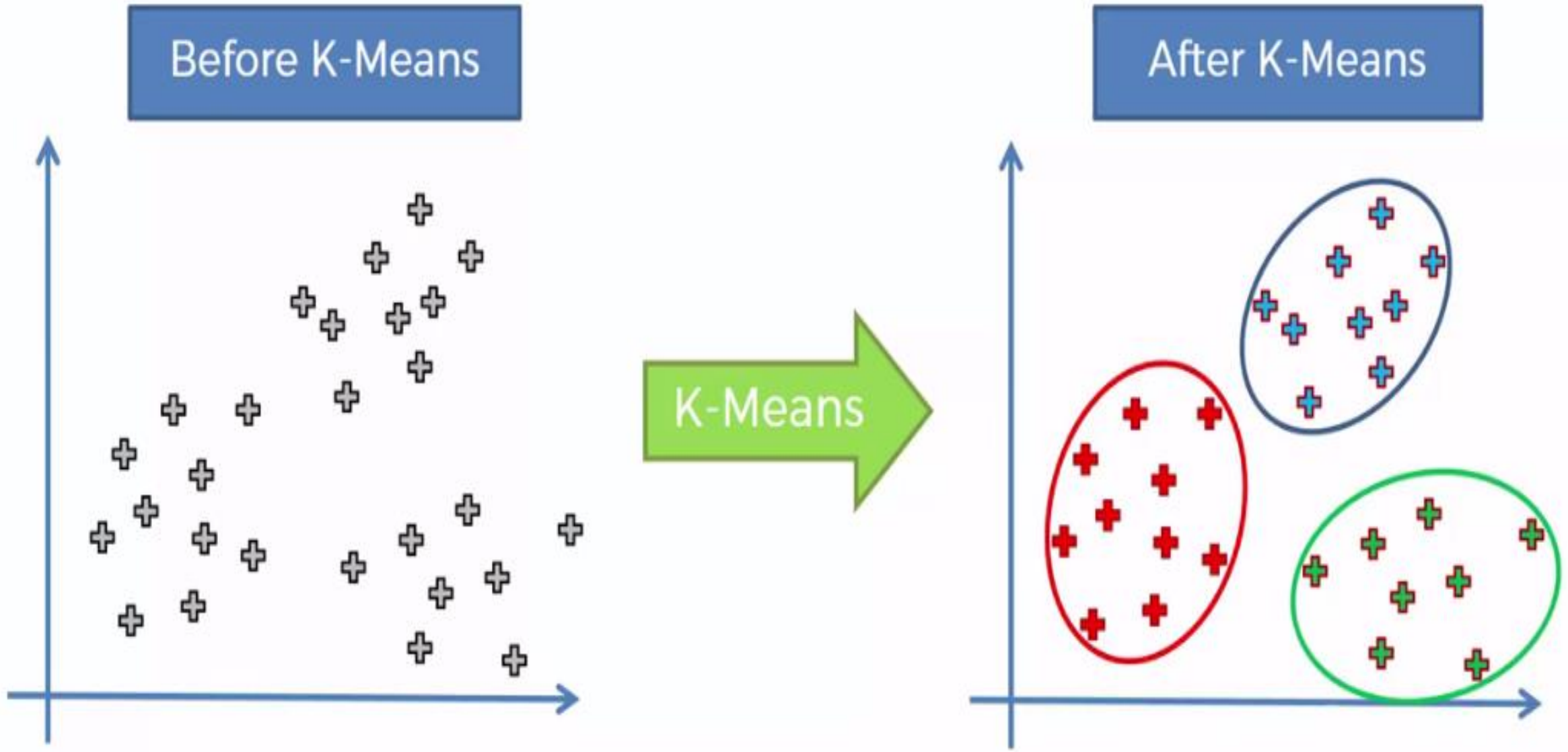
- Centroid-based clustering.
- Density-based clustering.
- Distribution-based clustering.
- Hierarchical clustering.

Clustering Algorithms

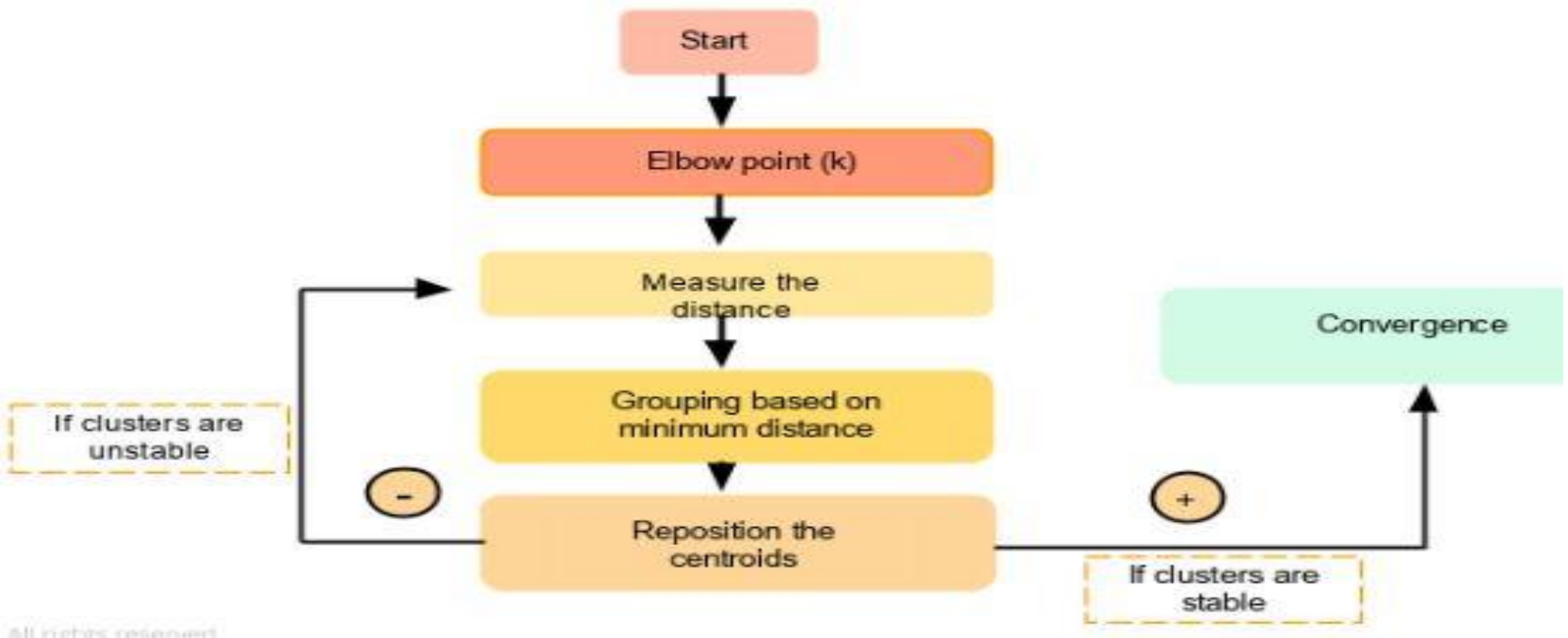
- 1.k-means Clustering
- 2.Agglomerative Clustering
- 3.Affinity propagation clustering
- 4.Mean shift clustering
- 5.Spectral
- 6.DBSCAN Clustering
- 7.HDBSCAN
- 8.OPTICS Clustering
- 9.BIRCH Clustering
- 10.Bisecting K-means Clustering

1.K-Means Algorithm

- K-means clustering is an unsupervised learning algorithm used for data clustering, which groups unlabeled data points into groups or clusters.
- It is one of the most popular clustering methods used in machine learning.
- K-means clustering is an iterative process to minimize the sum of distances between the data points and their cluster centroids


Before K-Means
K-Means
After K-Means

Working principle of k-Means algorithm



K-Means Application:

- customer segmentation,
- fraud detection,
- predicting account attrition,
- targeting client incentives,
- cybercrime identification, and
- delivery route optimization.

Advantage and Disadvantage

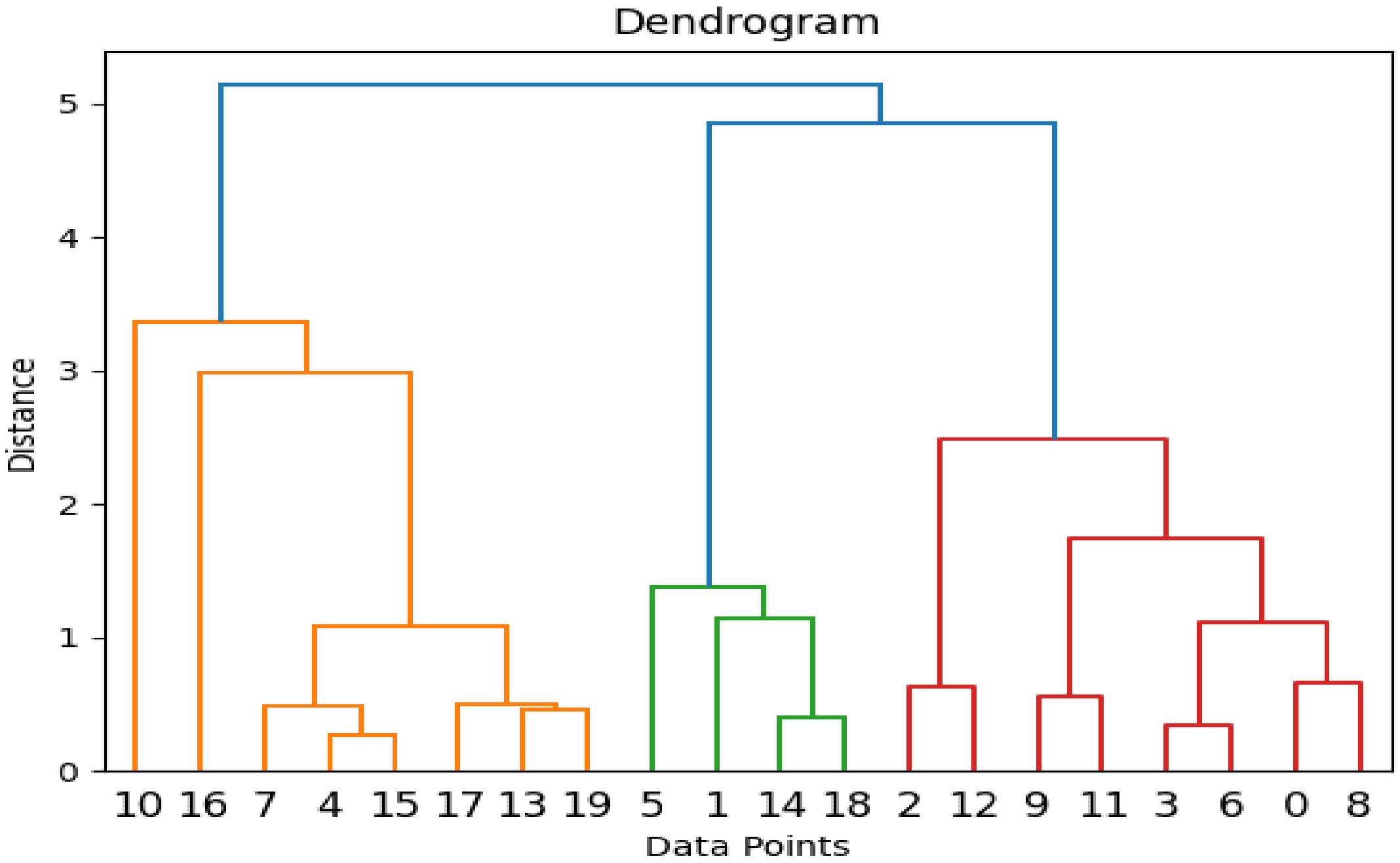
- **Advantage:**
- **Simple and easy to implement**: The k-means algorithm is easy to understand and implement, making it a popular choice for clustering tasks.
- **Fast and efficient**: K-means is computationally efficient and can handle large datasets with high dimensionality.

Disadvantage:

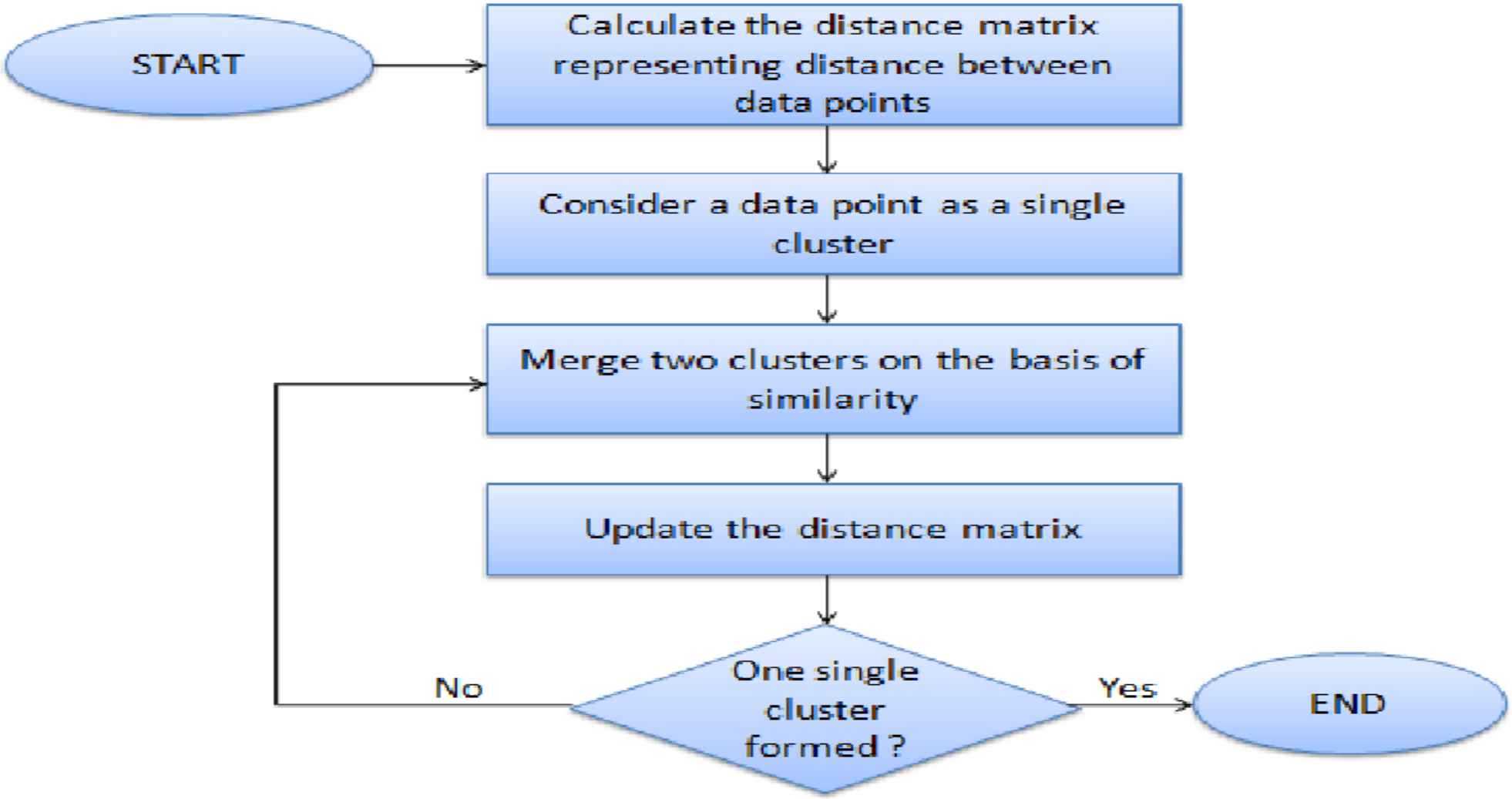
- Difficult to predict K-value
- With global cluster, it didn't work well.
- Different initial partitions can result in different final clusters

2.Agglomerative Clustering

- Agglomerative clustering is a hierarchical algorithm that uses a **bottom-up approach**.
- Each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- Each data point is initially considered a “cluster.”
- The algorithm proceeds by successively merging clusters using a selected linkage criterion.


Dendrogram
Distance
0
1
2
3
4
5
10 16 7 4 15 17 13 19 5 1 14 18 2 12 9 11 3 6 0 8
Data Points

Working principle of Agglomerative Clustering



Agglomerative Application

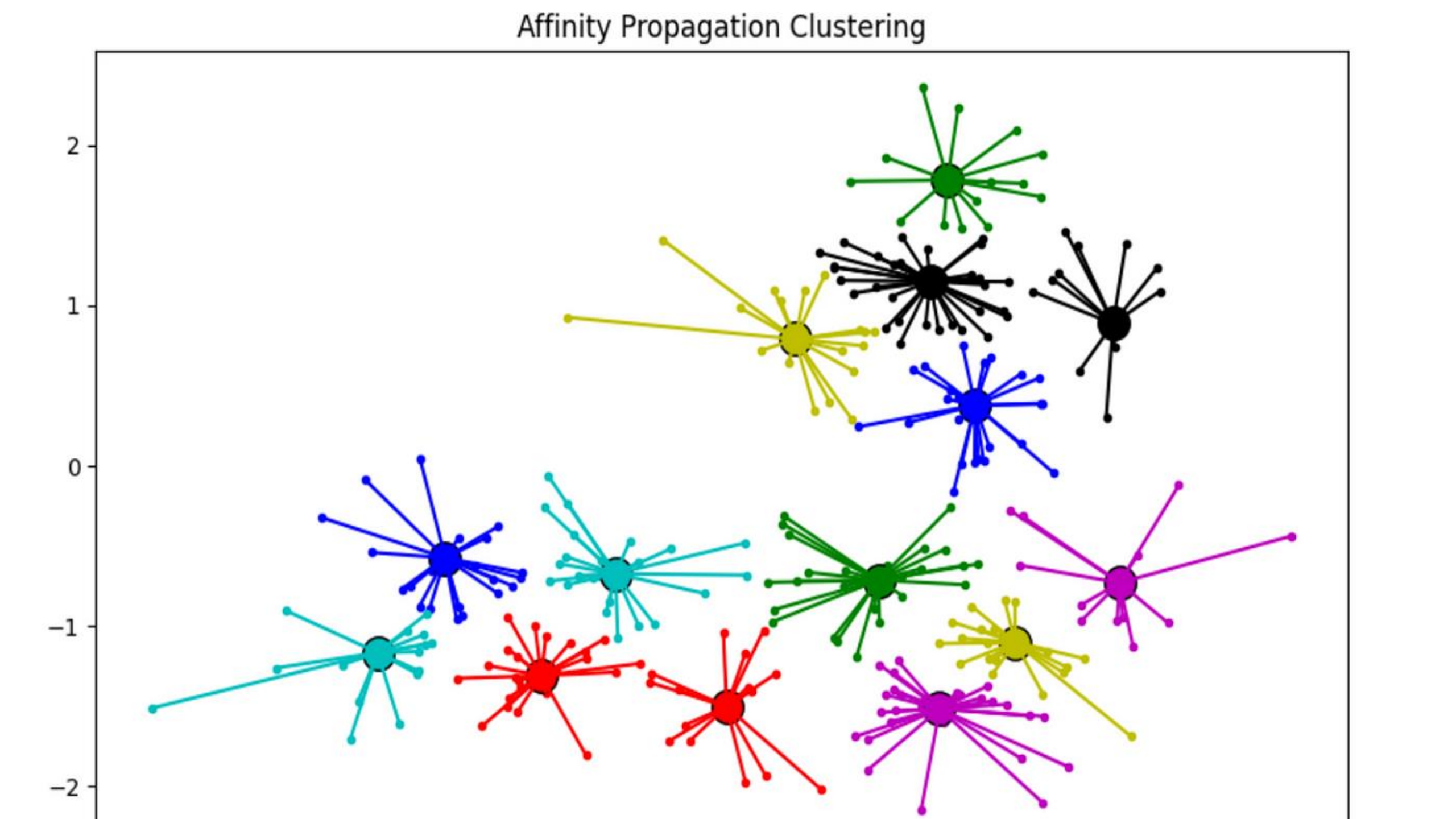
- the agglomerative clustering method is helpful in **phylogenetic analysis** which the process of grouping of n small groups into a single large group, where n is the number of data

Advantage and Disadvantage

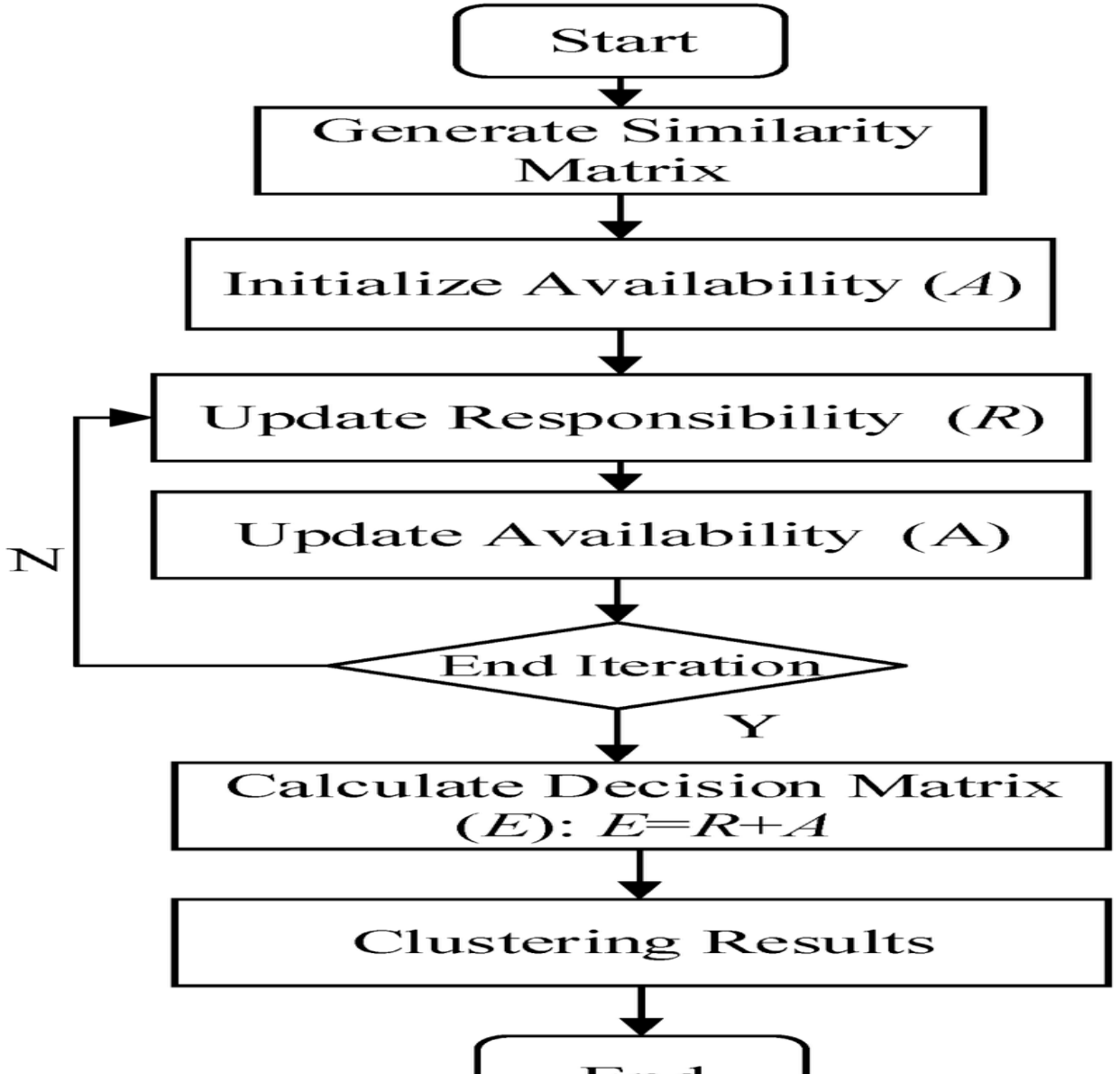
- **Advantage**:
- It works from the dissimilarities between the objects to be grouped together.
- **Disadvantage:**
- It is difficult to determine the right number of clusters in agglomerative clustering. ...
- It can be difficult to interpret the results of clustering. ...
- Agglomerative clustering may not produce the same results each time.

3.Affinity Propagation

- affinity propagation (AP) is a clustering algorithm based on the concept of "message passing" between data points.
- In the Affinity Propagation Method, all points are considered as potential centers of the cluster and the negative value of the Euclidean distance between two points determines their affinity.
- Thus, the greater the sum of the affinity, the greater the probability that the point is the center of the cluster.


Affinity Propagation Clustering
2
1
0
−1
−2

Working principle of Affinity propagation



Affinity propagation Application

- affinity propagation showed it is better for:
- certain computer vision
- computational biology tasks, (example)
- clustering of pictures of human faces and identifying regulated transcripts

Advantage:

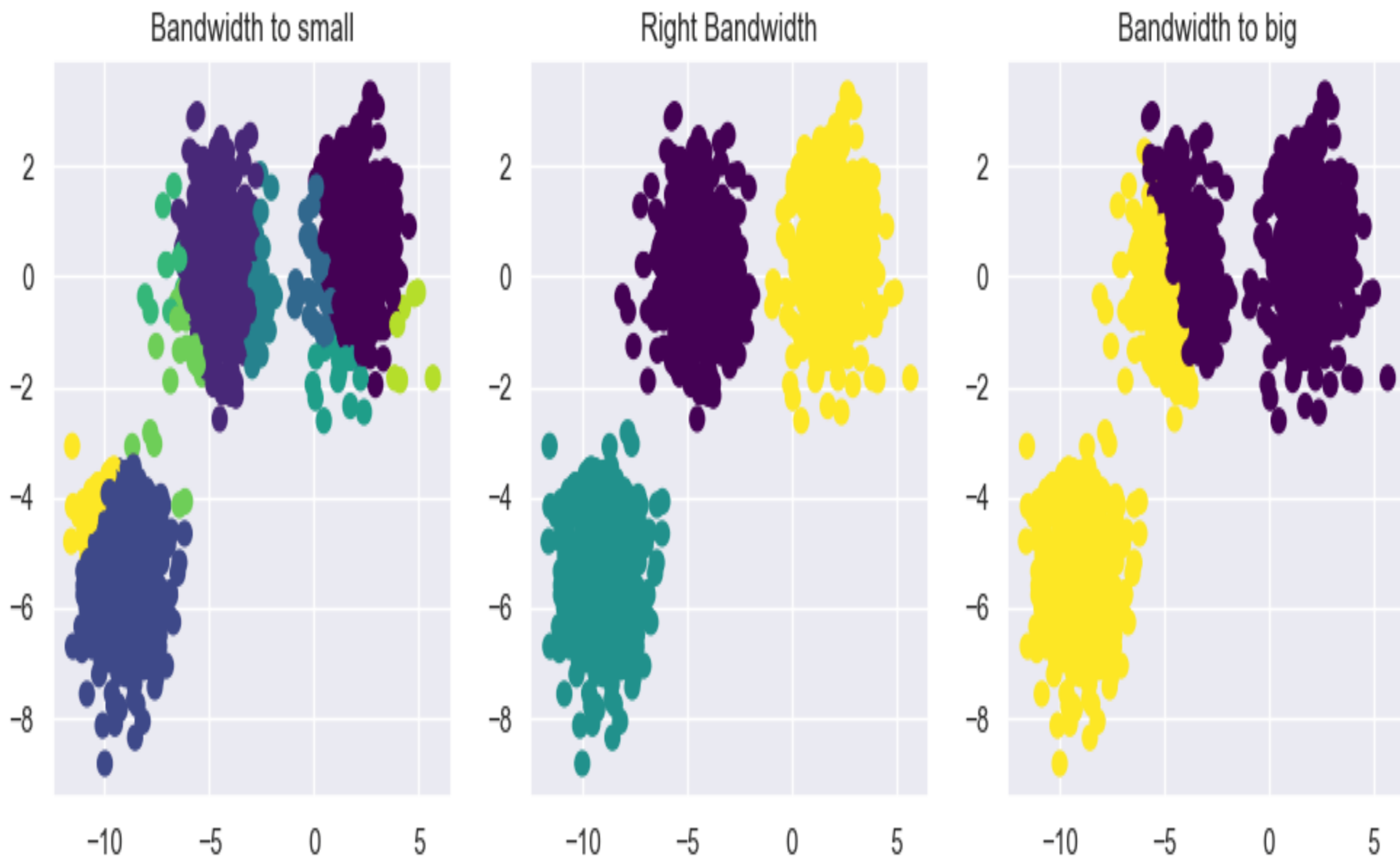
- Has better performance and lower clustering error
- Its ability to automatically determine the number of clusters
- Can be used to cluster data with complex relationships and non-linear structures.
- Can be used in a wide range of applications, including image segmentation, customer segmentation, and gene expression analysis

Disadvantage:

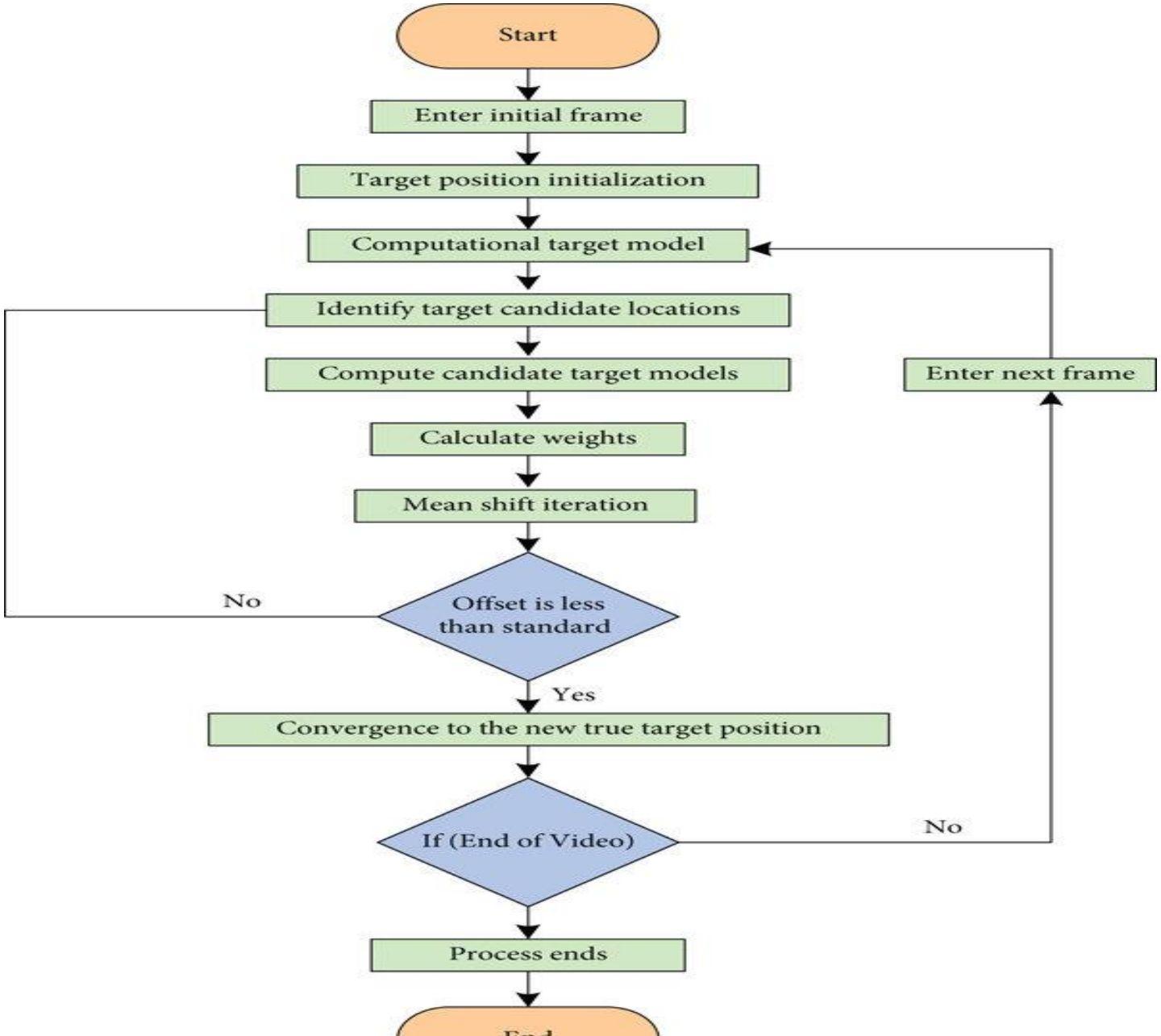
- Can be computationally expensive, especially for large datasets, making it unsuitable for large-scale clustering problems.
- May not always produce the best results compared to other clustering algorithms, such as K-Means or Gaussian Mixture Models.
- Can be sensitive to the choice of similarity metric used to measure the similarities between data points.
- Can produce multiple exemplars for a single cluster, making it difficult to interpret the results of the clustering process.

4.Mean Shift Clustering

- Mean Shift is an unsupervised clustering algorithm, that aims to discover blobs in a smooth density of samples.
- It is a **centroid-based algorithm** that works by updating candidates for centroids to be the mean of the points, within a given region (also called bandwidth)
- The mean-shift algorithm is an efficient approach to tracking objects whose appearance is defined by histograms


Bandwidth to small
Right Bandwidth
Bandwidth to big
2
0
−2
−4
−6
−8
−10
−5
0
5

Working principle of Mean Shift Clustering



Application:

- object recognition,
- image processing and analysis
- computer vision techniques such as image denoising, image segmentation, motion tracking, etc.

Advantage and Disadvantage

Advantage:

- Mean Shift is a non-parametric technique, which means you don't need to specify the number of clusters beforehand.
- ability to automatically determine the number of clusters and adapt to the shape and size of the data distribution

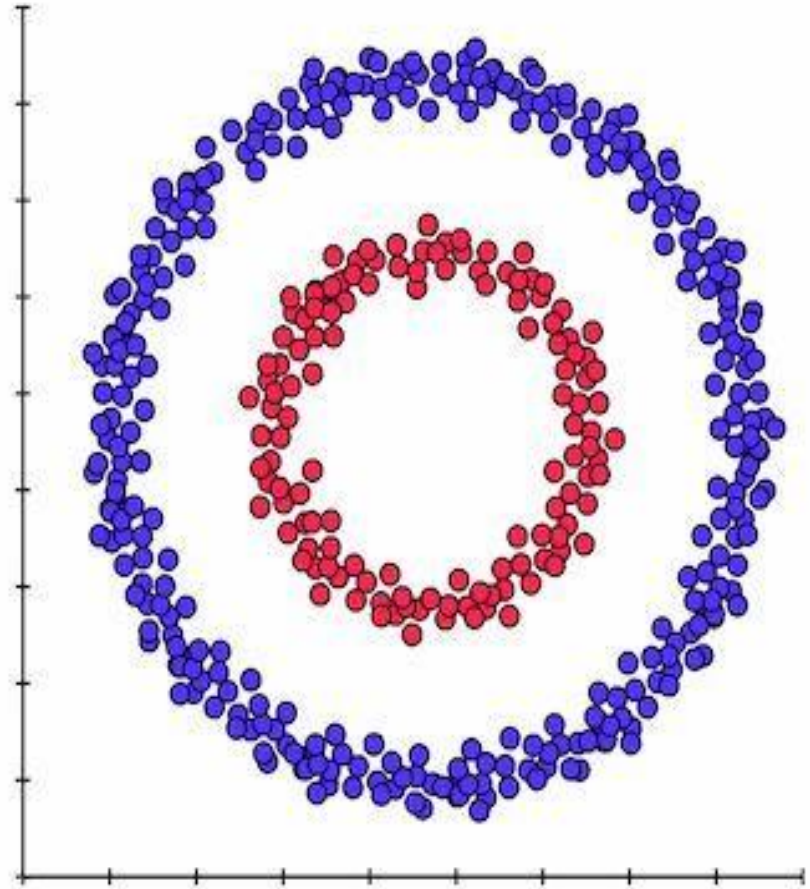
Disadvantage:

- The output depends on bandwidth size, and its selection may not be trivial.
- The mean-shift algorithm is relatively slow, especially for large datasets, and has poor scalability with high-dimensional data.
- The computation time for this algorithm is generally O(n log n) to $O(n^2)$.

5.Spectral clustering

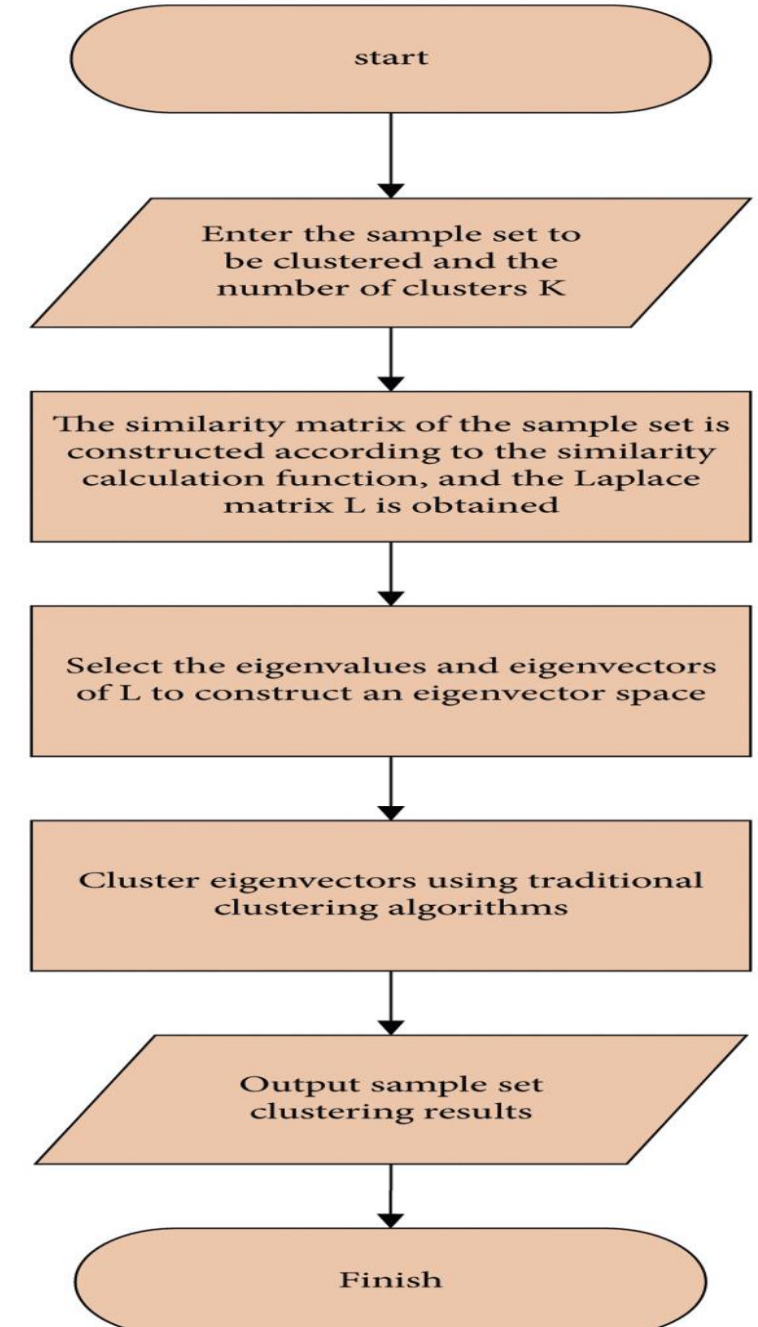
- Spectral clustering is one of the components of machine learning and artificial intelligence.
- It is a data partitioning algorithm based on **spectral graph theory** and **linear algebra**.
- The idea is to segment a graph into several small groups with similar or nearby values.
- The spectral methods for clustering usually involve taking the **top eigen vectors** of some matrix based on the **distance between points** (or other properties) and then using them to cluster the various points.

SPECTRAL CLUSTERING



YoungWonks

Working principle of spectral clustering



Application:

- Spectral Clustering is a technique used to group together data points of similar behavior in order to analyze the over all data
- Spectral Analysis. The modern methods of **time series analysis** are often used to **simplify complicated waveforms such as EEG**. Many industrial applications involve such methods, as **electric-Circuits**
- **signal processing** (television, radar, astronomy, etc.), and
- **voice recognition.**

Advantage and Disadvantage

Advantage:

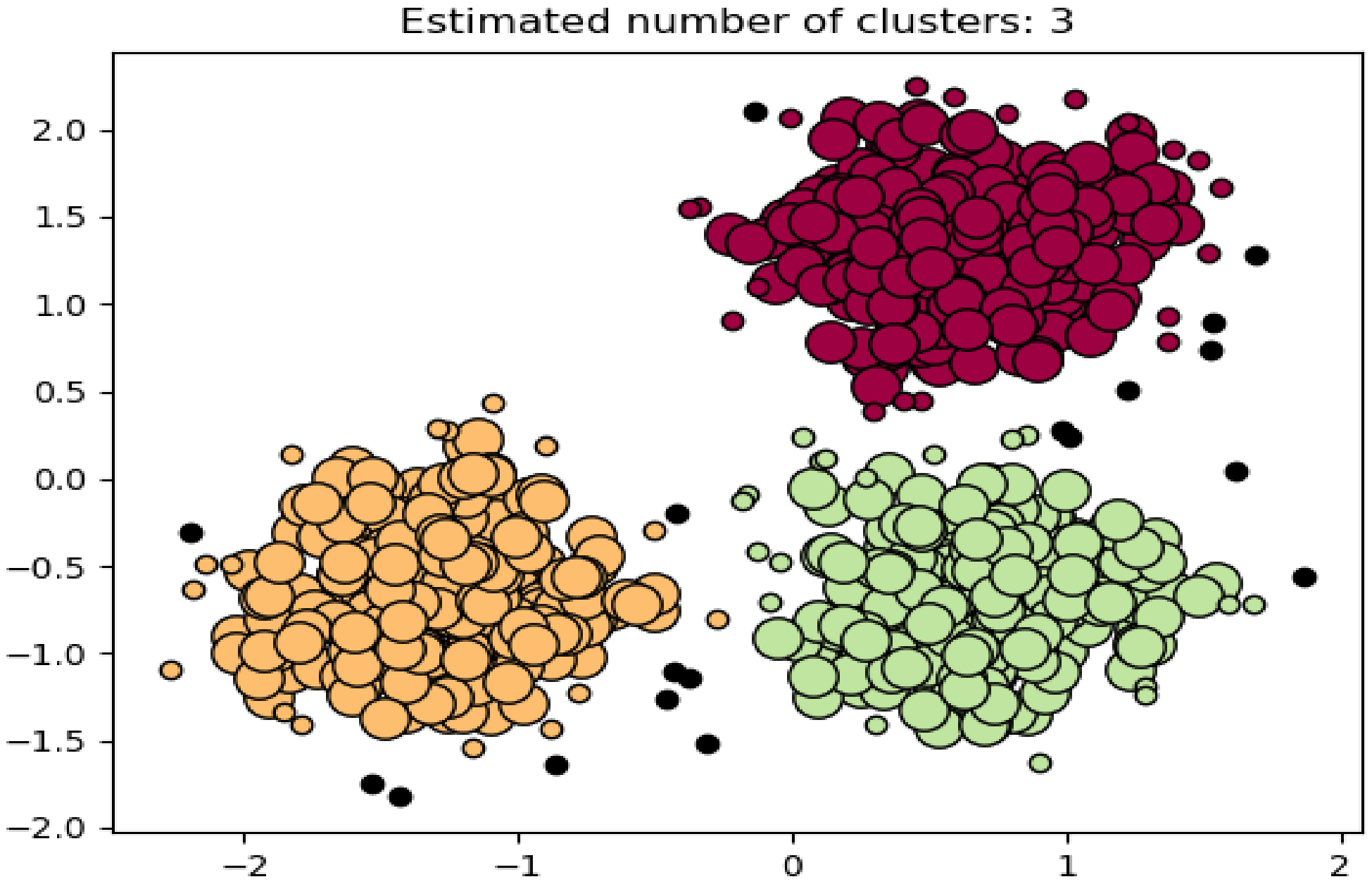
- It offers a powerful solution for complex, non-convex cluster structures.
- Spectral clustering is useful when the clusters have a non-linear shape, and it can handle noisy data better than k-means.

Disadvantage:

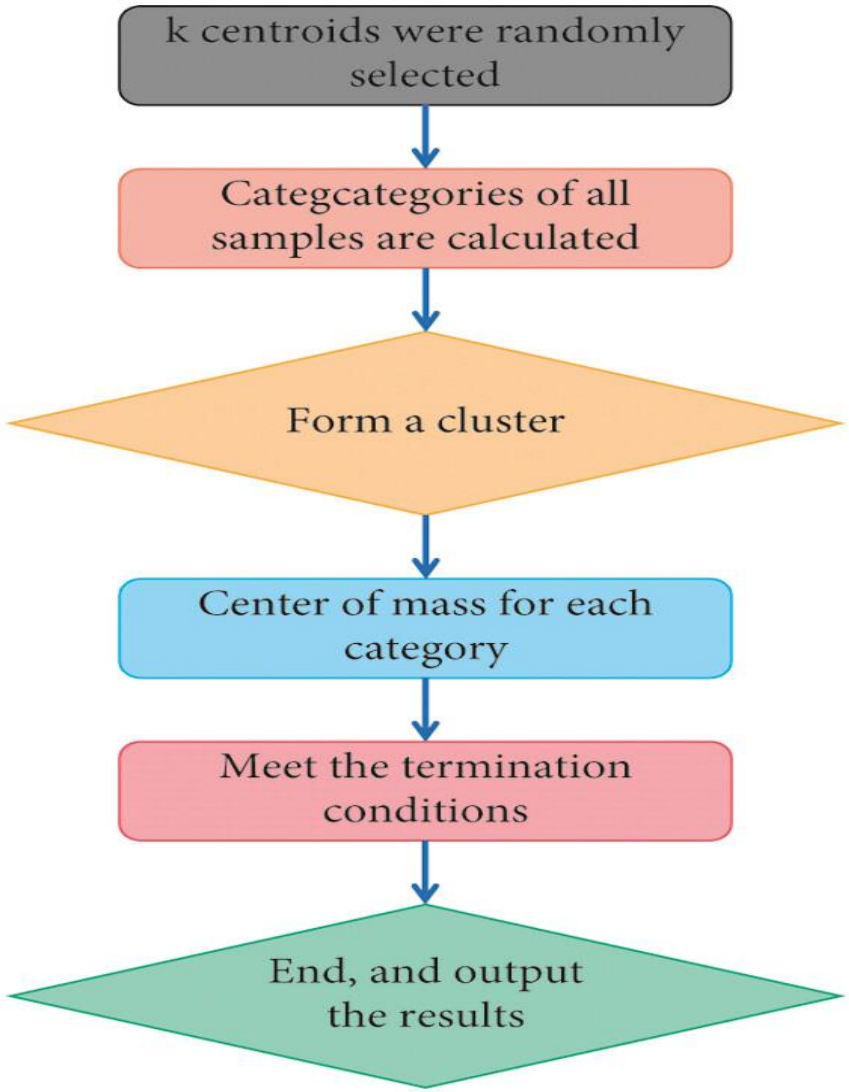
- In comparison to other clustering techniques like k-means clustering, Spectral Clustering has the drawback of being rather slow.
- If your dataset has a large number of data points, a faster algorithm will be better.

6.DBSCAN-Clustering

- **Density-Based Spatial Clustering of Applications with Noise -** (DBSCAN) clustering.
- Is a popular clustering algorithm used in machine learning and data mining to **group points in a data set that are closely packed together based on their distance to other points**.


Estimated number of clusters: 3
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2
−1
0
1
2

Working principle of DBSCAN Clustering



Application:

- DBSCAN is broadly used in many applications such as
- market research,
- pattern recognition,
- data analysis, and
- image processing

Advantage and Disadvantage

Advantage:

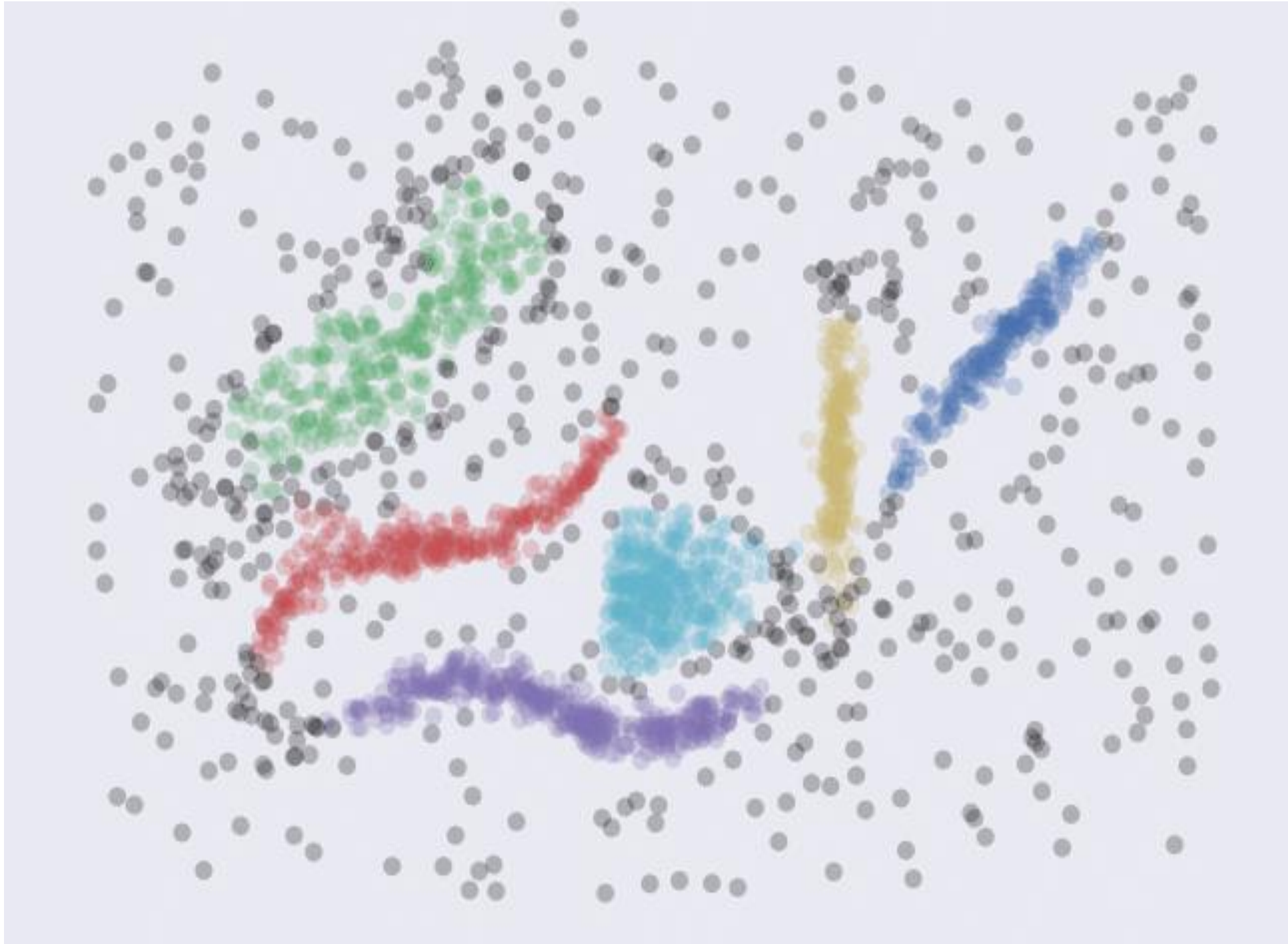
- DBSCAN is great at separating high-density clusters from low-density clusters,
- DBSCAN can be used to detect clusters that are oddly or irregularly shaped, such as clusters that are ringshaped.
- DBSCAN is used to handle clusters of multiple sizes and structures and is not powerfully influenced by noise or outliers.

Disadvantages:

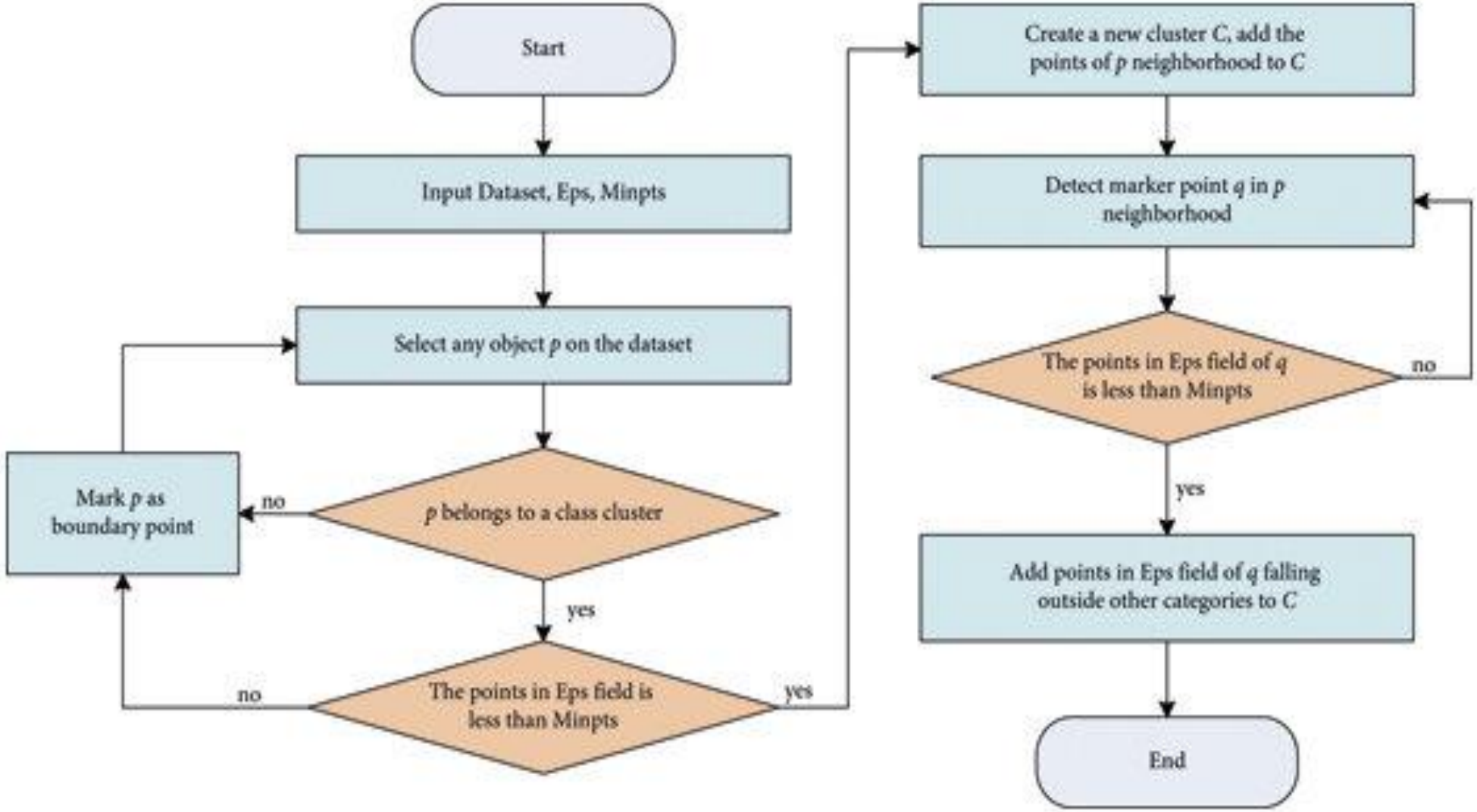
- DBSCAN struggles with clusters of similar density.
- Struggles with high dimensionality data. If given data with too many dimensions, DBSCAN suffers.

7.HDBSCAN-Clustering

- **Hierarchical Density-Based Spatial Clustering of Applications with Noise.**
- The algorithm essentially seeks areas in the dataset where there are lots of data points (high density), and separates these regions from areas with few data points (low density).
- While HDBSCAN can perform well on low to medium dimensional data the performance tends to decrease significantly as dimension increases.
- In general HDBSCAN can do well on up to **around 50 or 100 dimensional data,**



Working principle of HDBSCAN



Application of HDBSCAN

- It uses high-density regions to identify clusters and views isolated or low-density points as noise.
- The hierarchical structure is used primarily today for **storing geographic information** and file systems. Currently, hierarchical databases are still widely used especially in applications that require very high performance and availability such as **banking, health care, and telecommunications.**

Advantage and Disadvantage

Advantage :

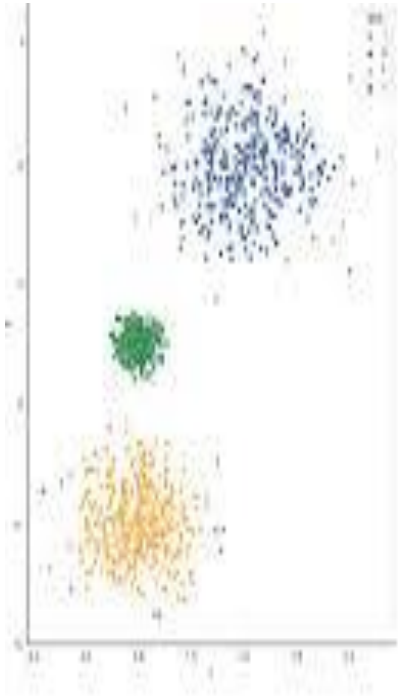
- Using HDBSCAN for big data clustering offers several advantages, including its ability to handle varying cluster densities and its robustness to noise
- HDBSCAN's hierarchical structure enables it to uncover clusters of different shapes and sizes.

Disadvantage:

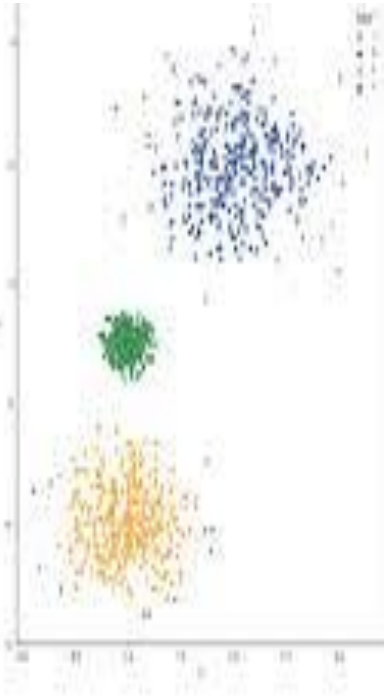
Computationally Intensive: HDBSCAN can be computationally expensive, particularly for large datasets, due to the construction of the minimum spanning tree and the calculation of mutual reachability distances

8.OPTICS-Clustering

- OPTICS clustering refers to "**Ordering Points To Identify the Clustering Structure",** an algorithm used in the field of data mining and machine learning for cluster analysis.

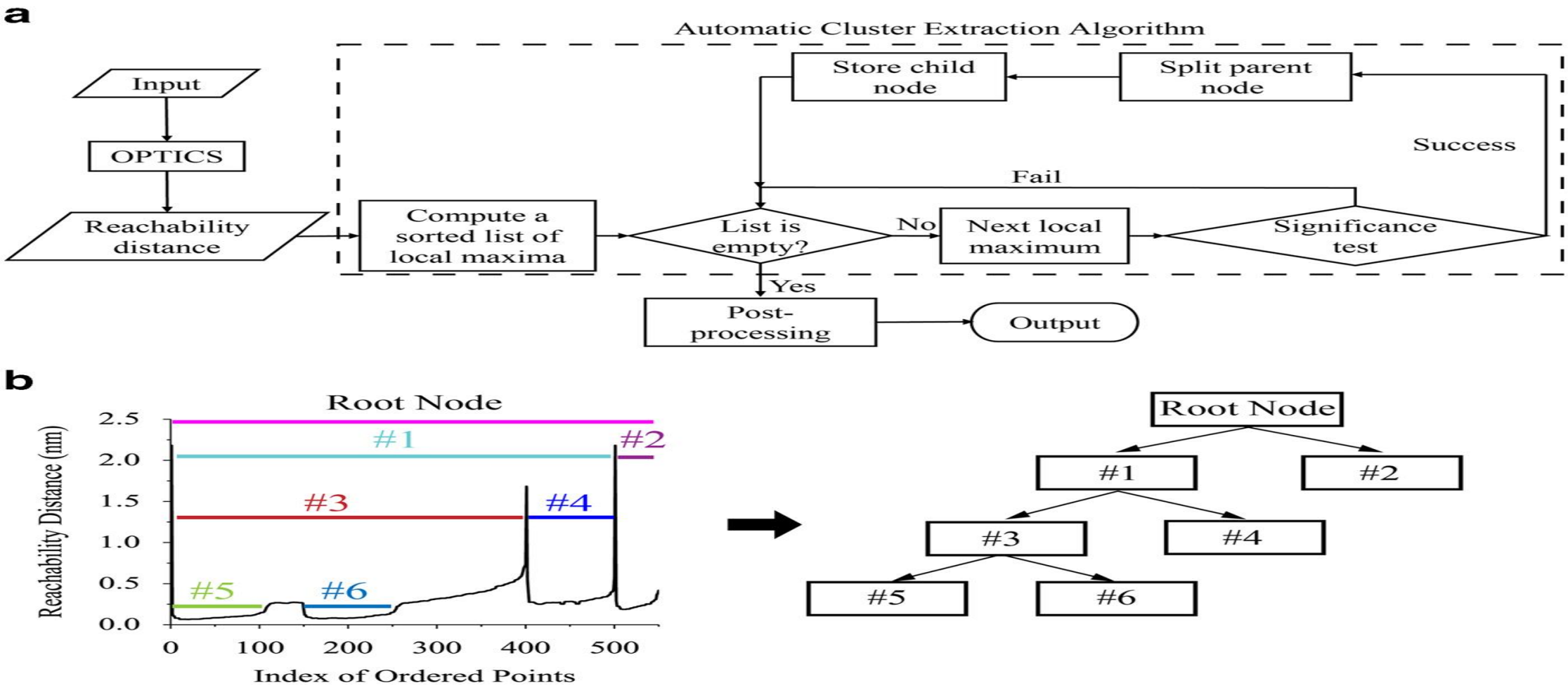


(a) OPTICS



(b) BLOCK-OPTICS

Working principle of OPTICS Clustering



Application

- Clustering using OPTICS by a **MAQ Software analyzes and identifies data clusters**. The algorithm relies on density-based clustering, allowing users to **identify outlier points** and closely-knit groups within larger groups.
- This visual includes adjustable clustering parameters to control hierarchy depth and cluster sizes.

Advantage and Disadvantage

Advantage:

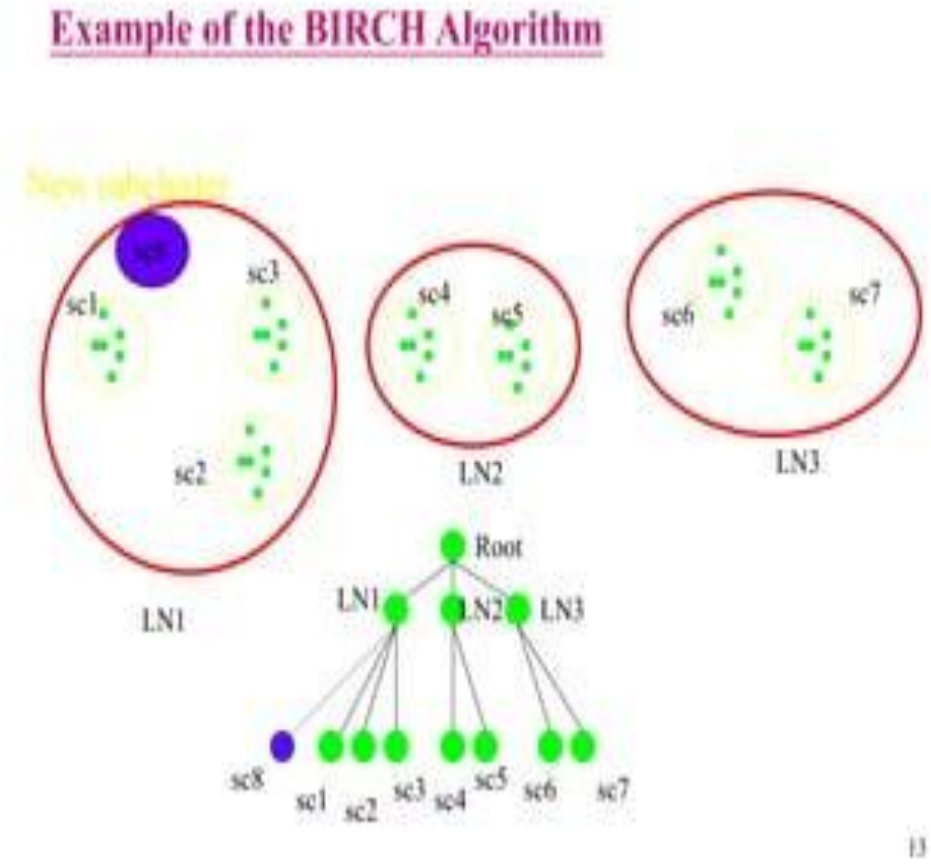
- **Flexible Clustering**: OPTICS allows for the identification of clusters with different shapes, sizes, and densities, making it suitable for diverse datasets.
- One of the main advantage of OPTICS over DBSCAN, is that it does not require to set the number of **clusters in advance.**

Disadvantage:

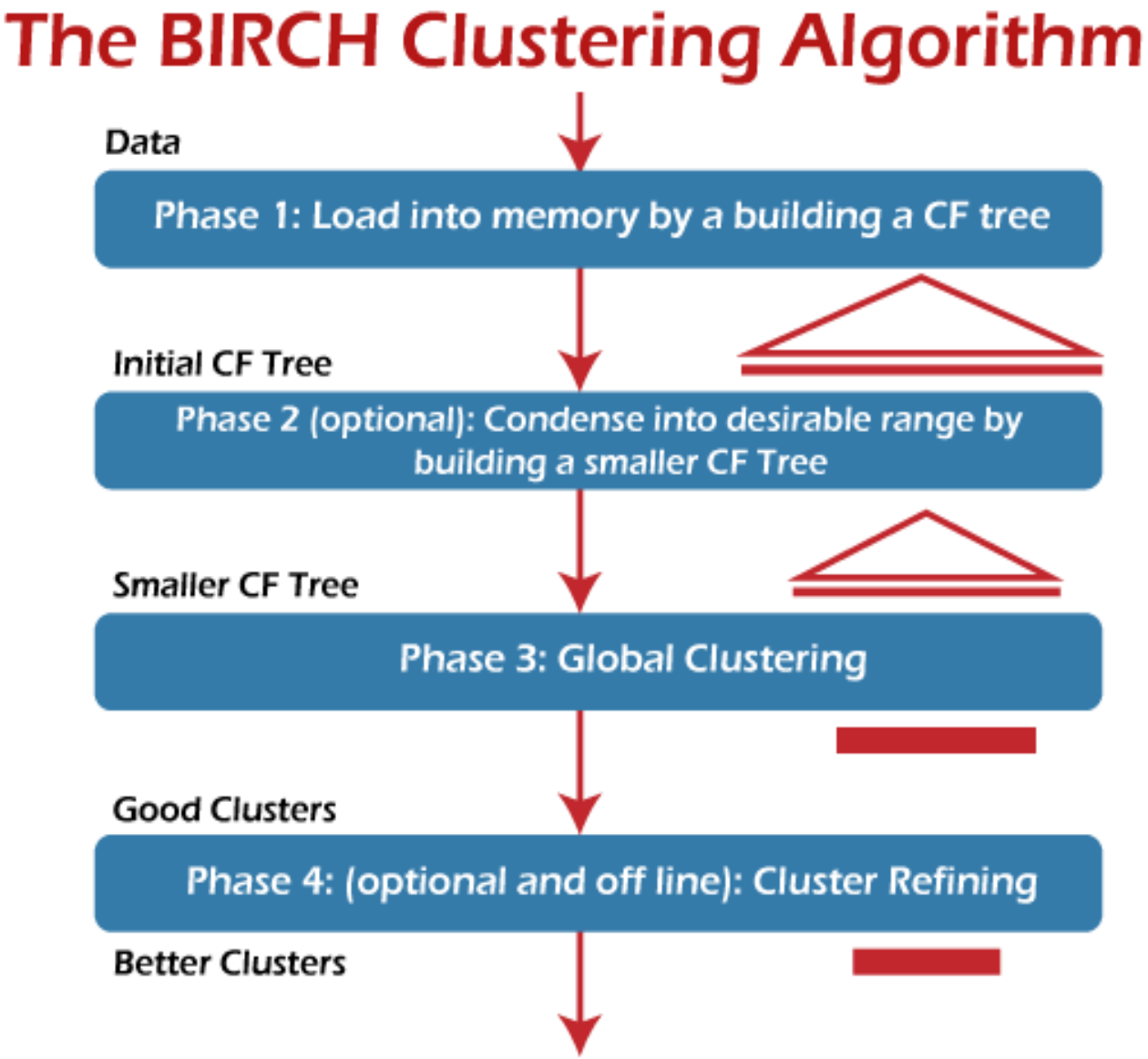
- **Sensitivity to parameters** – OPTICS requires careful tuning of its parameters, such as the min_samples and xi parameters, which can be challenging.
- **Computational complexity** – OPTICS can be computationally expensive for large datasets, especially when using a high min_samples value

9.BIRCH-Clustering

- **BIRCH (Balanced iterative reducing and clustering using hierarchies)** is an unsupervised data mining algorithm used to perform **hierarchical clustering over particularly large data-sets**.
- BIRCH incrementally and dynamically clusters incoming multi-dimensional metric data points to try to produce the best quality clustering with the available resources. (i.e., available memory and time constraints).



Working principle of BIRCH



Application:

- BIRCH algorithm With modifications, it can also be **used to accelerate k-means clustering and Gaussian mixture modeling with the expectation-maximization algorithm**.
- **BIRCH** is a **clustering algorithm** using which we can **cluster large datasets** by first producing a brief summary about the dataset that preserves information

Advantages and Disadvantages

Advantages:

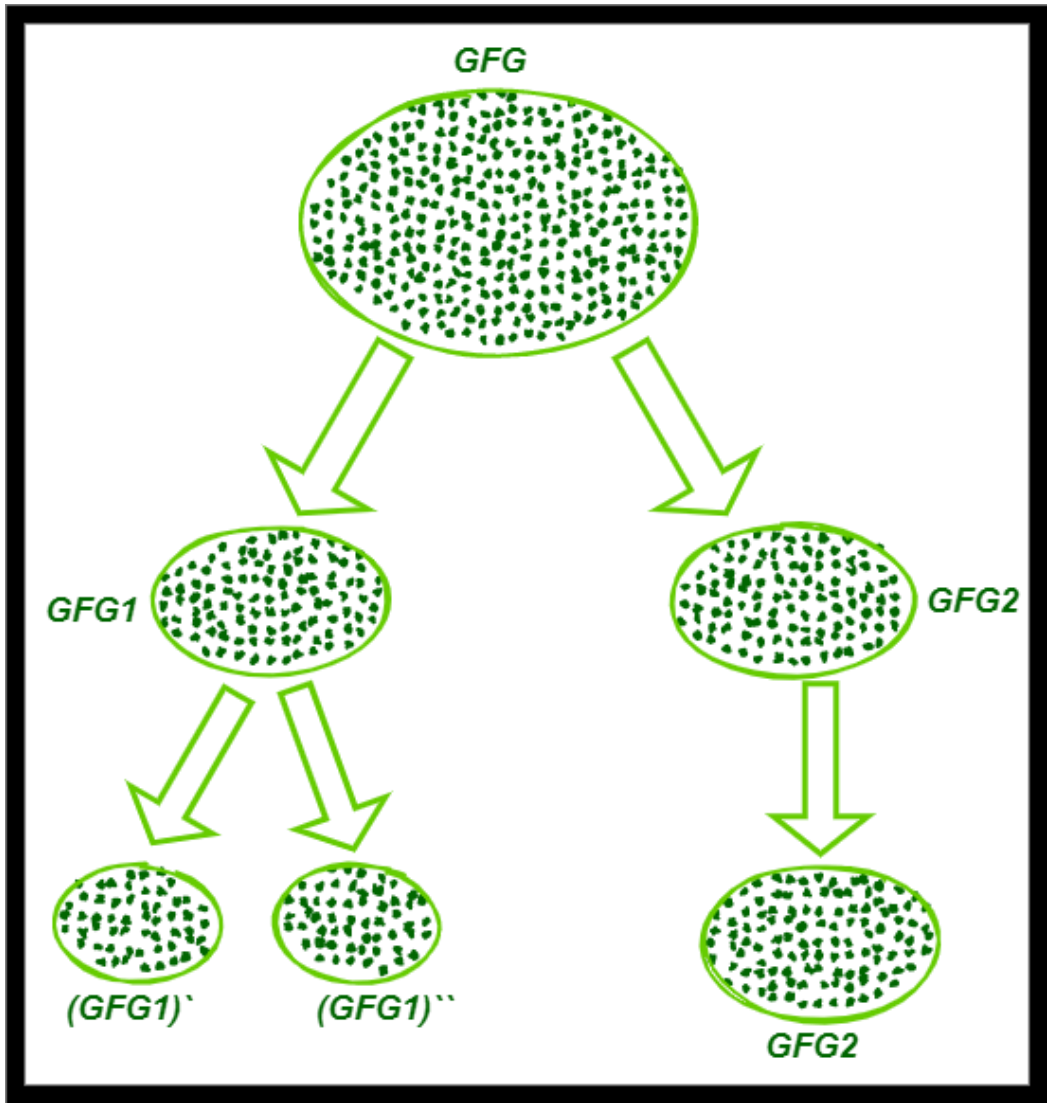
- BIRCH is useful for performing precise Clustering on large datasets
- An main advantage of BIRCH is its ability to incrementally and dynamically cluster incoming, multi-dimensional metric data points to produce the best quality clustering for a given set of resources (memory and time constraints). In most cases, BIRCH only requires a single scan of the database.

Disadvantages:

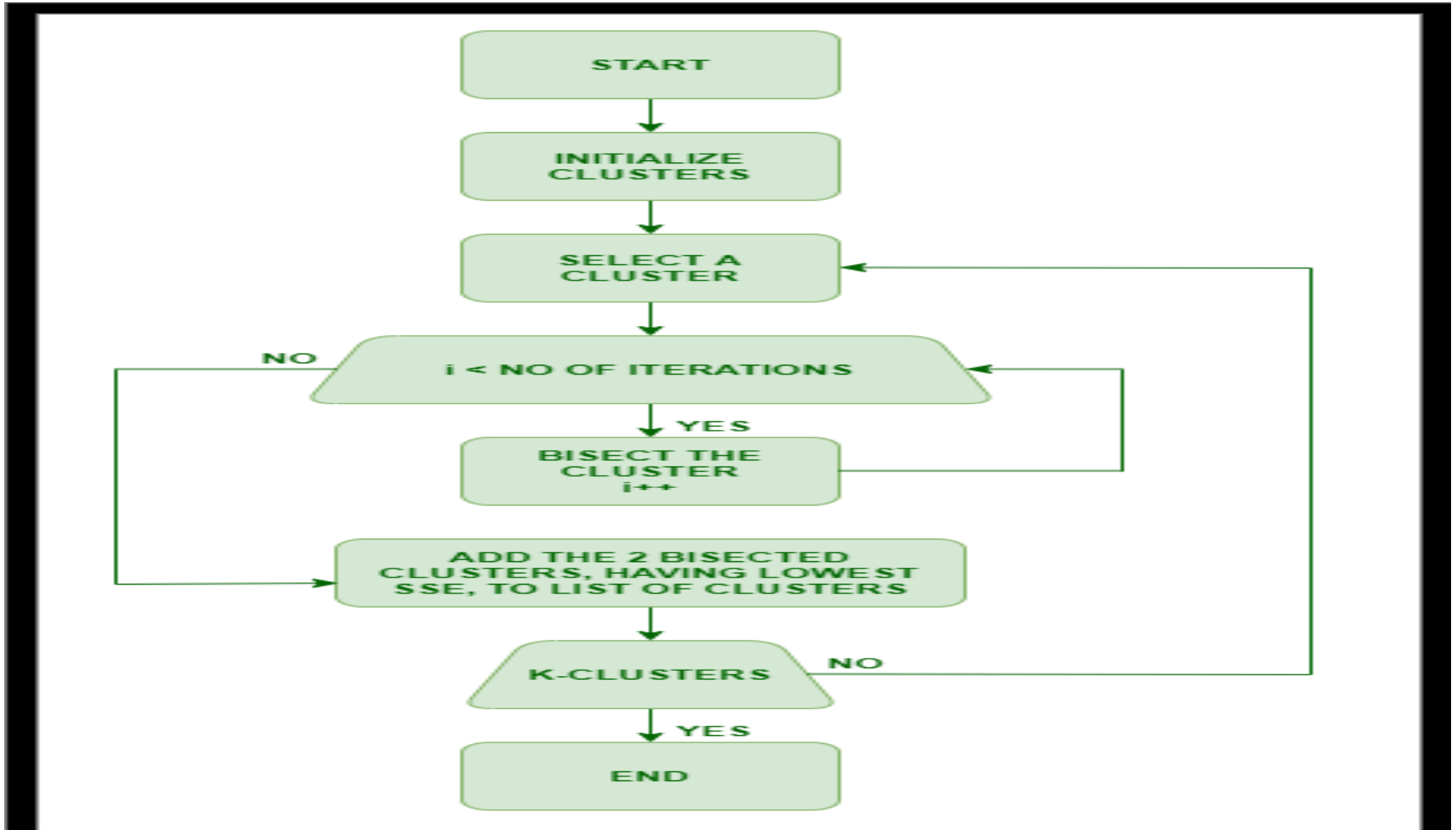
- BIRCH has one major drawback, it can **only process metric attributes**. A metric attribute is an attribute whose values can be represented in Euclidean space, i.e., no categorical attributes should be present.

10.Bisecting K-means Clustering

- The bisecting k-means clustering algorithm combines k-means clustering with **divisive hierarchy clustering.** With bisecting k-means, **you get not only the clusters but also the hierarchical structure of the clusters of data points.**
- bisecting k-means algorithm splits **one cluster into two sub clusters at each bisecting step** (by using k-means) until k clusters are obtained.



Working principle of Bisecting K-means Clustering



Application:

- It is used to **separate a set of instances** (vectors of double values) **into groups of instances** (clusters) according to their similarity.

Advantages:

- Bisecting K- means is even more efficient than the regular K-means algorithm.
- The advantages of the bisecting technique in **dental radiography** are **increased accuracy**,
- **simplicity of use, and shorter exposure time.**

Disadvantages:

it has some major drawbacks like quality of the resulting clusters **heavily** depends on **the selection of initial centroids**, clusters produced are of **varying sizes**, hence unbalanced and may also lead to empty clusters.

the bisecting technique include **image distortion**, and **excess radiation due to increased angulations exposing the eyes and thyroid.**