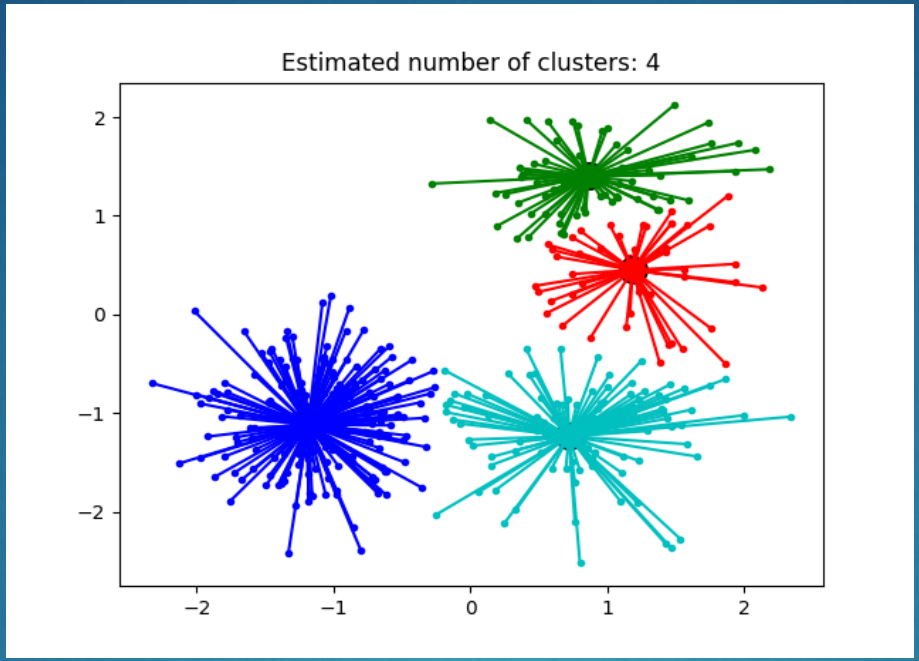
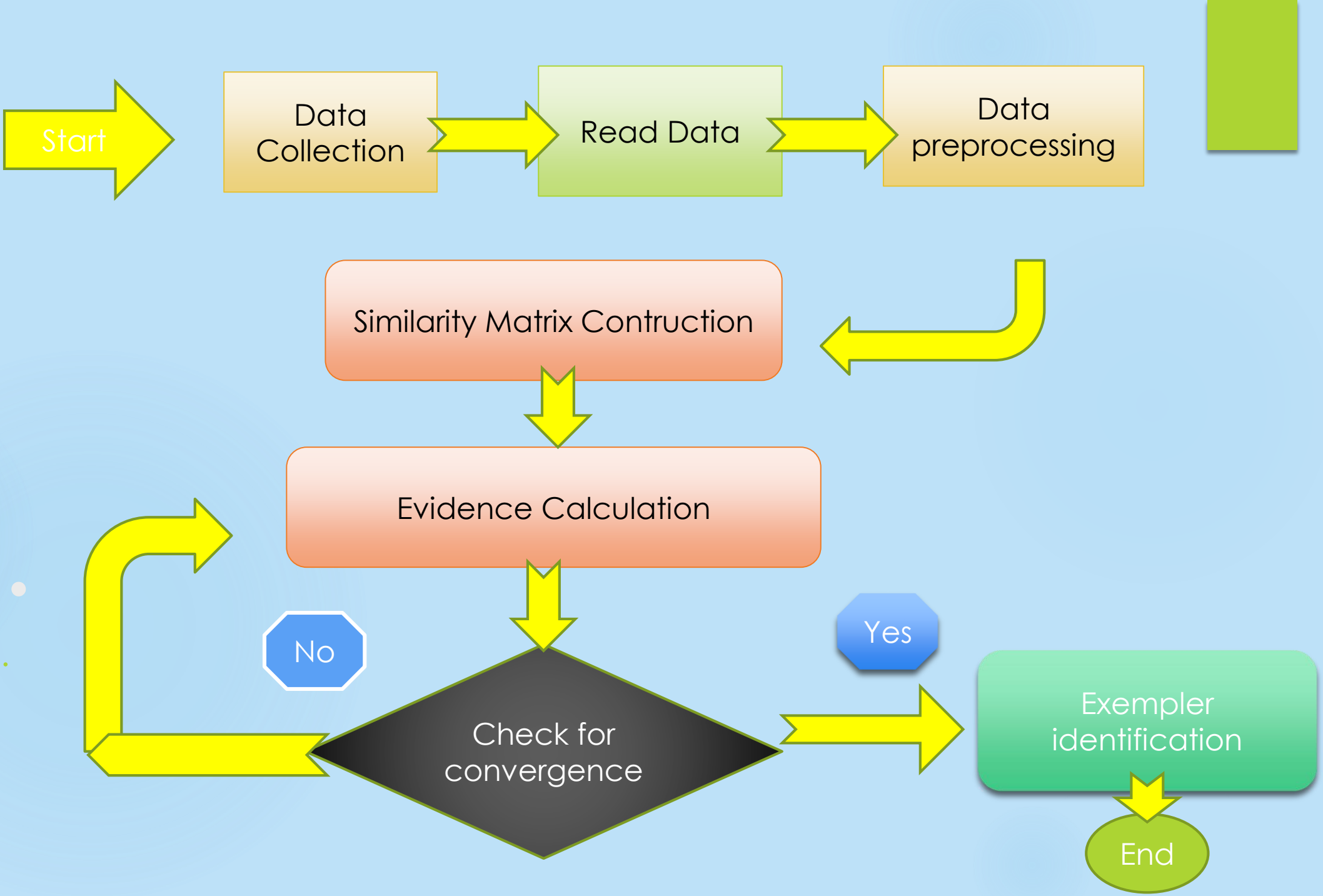
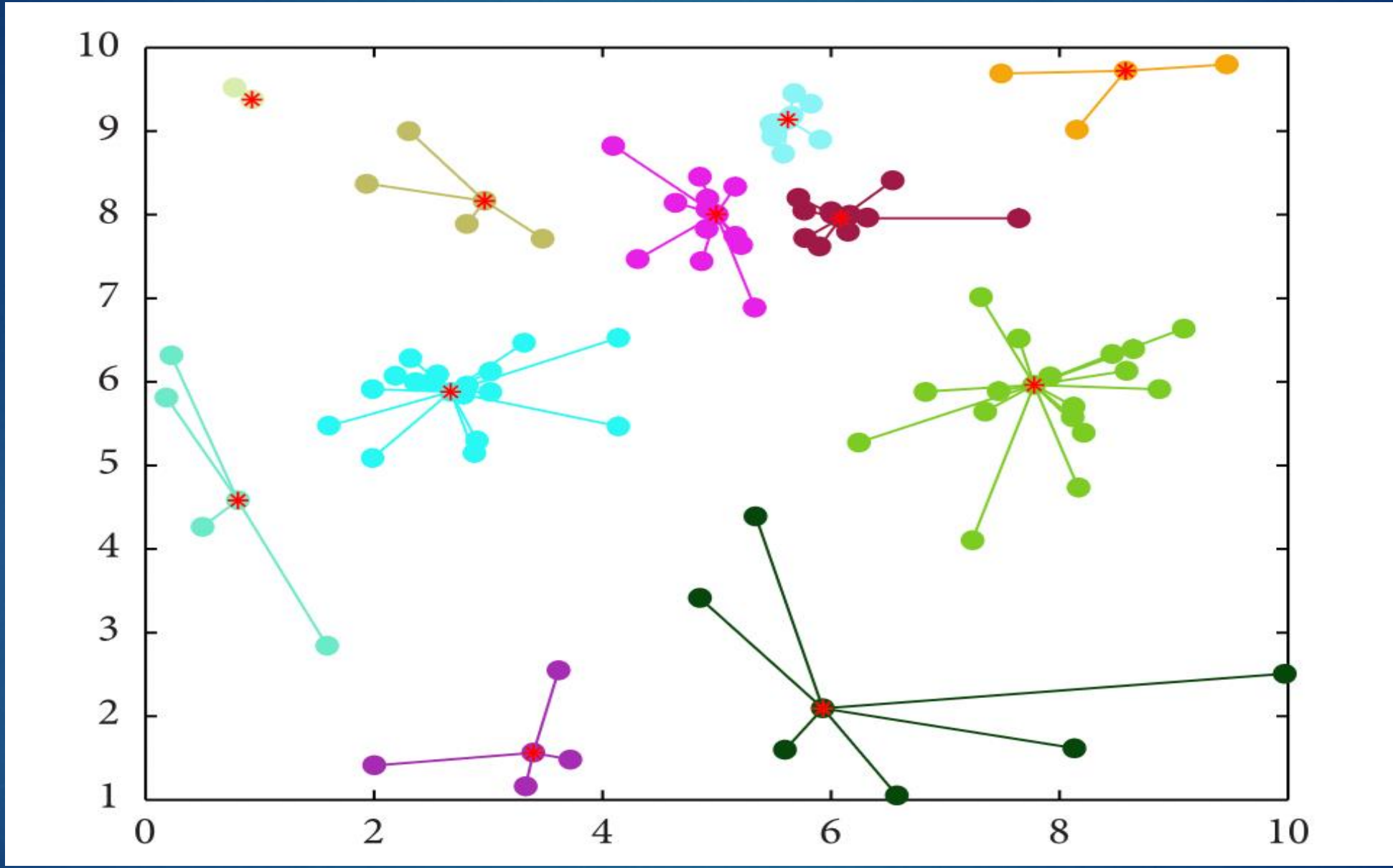
CLUSTERING ALGORITHMS

AFFINITY PROPAGATION CLUSTERING



Affinity propagation creates clustering by sending message between pair of samples until convergence


Start
Data Collection
Read Data
Data preprocessing
Similarity Matrix Contruction
Evidence Calculation
No
Yes
Check for convergence
Exempler identification
End


10
9
8
7
6
5
4
3
2
1
0
2
4
6
8
10

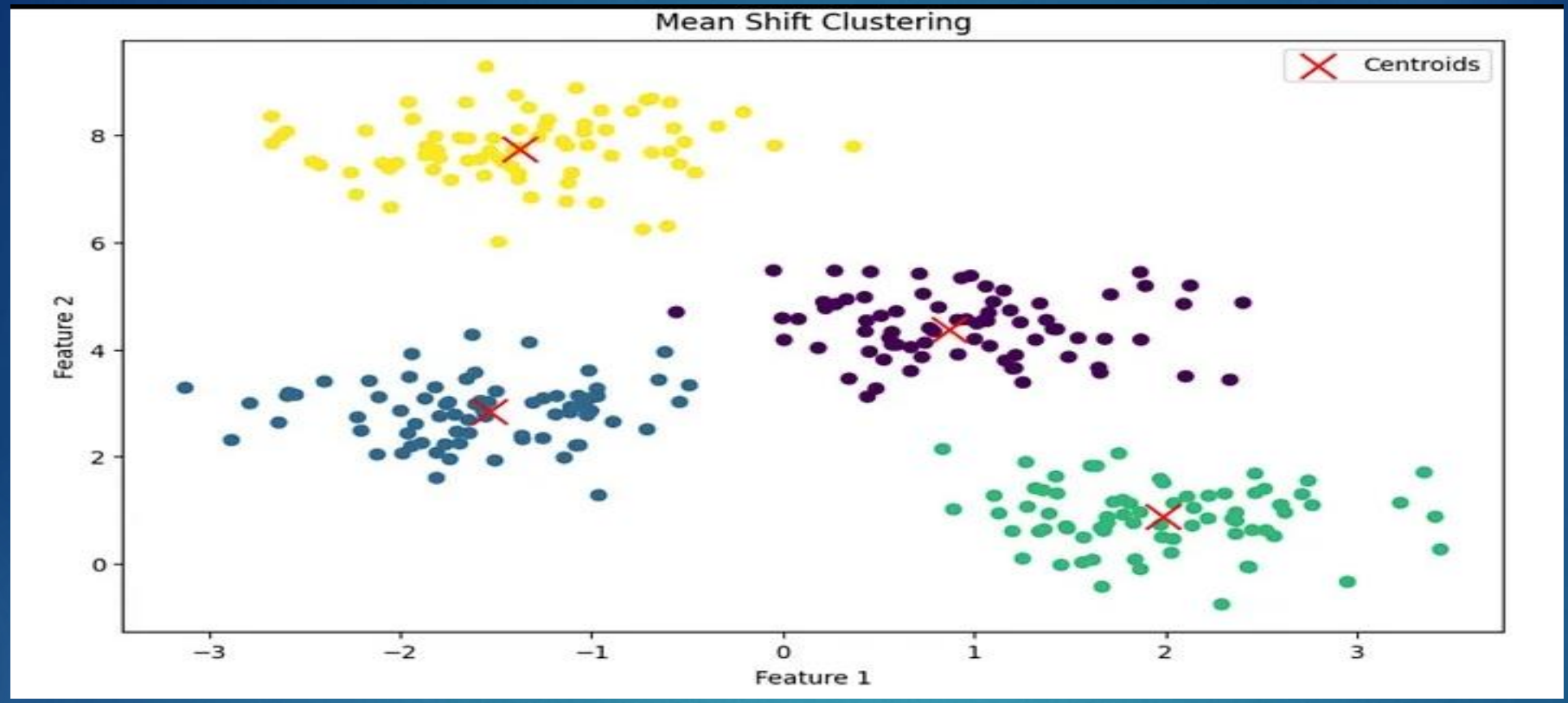
Merites:

It has better performance and lower clustering error

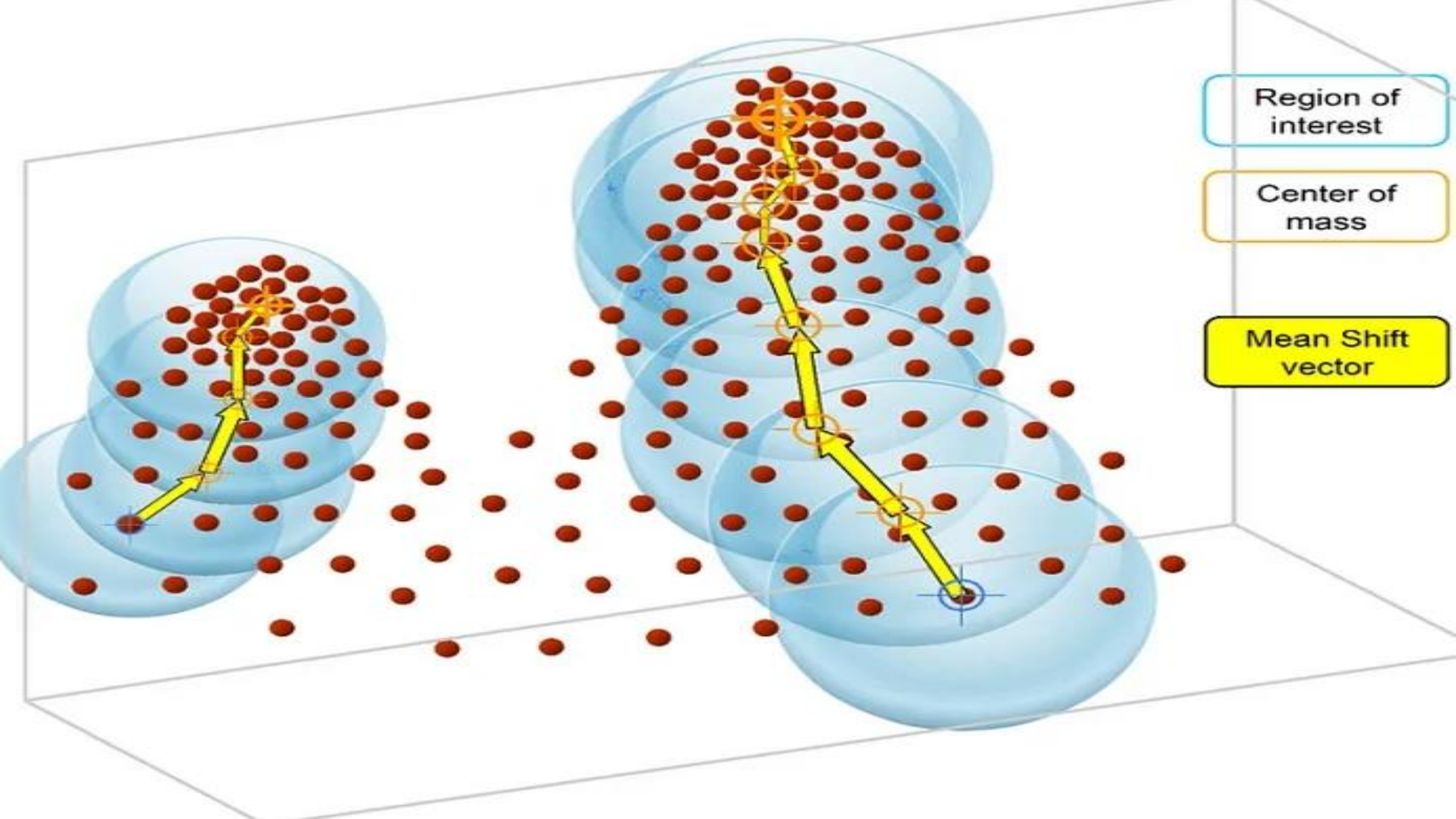
Demerites:

Making it difficult to scale large dataset.we do not have any control on the number of clusters but in some application we need no.of clusters

MEAN SHIFT CLUSTERING



Mean shift algorithm basically assigns the datapoints to the clusters iteratively by shifting points towards the highest density of datapoints


Region of interest
Center of mass
Mean Shift vector

Advantages:

*It can handle arbitrary data shapes and sizes

*It does not require prior knowledge of the number of cluster.

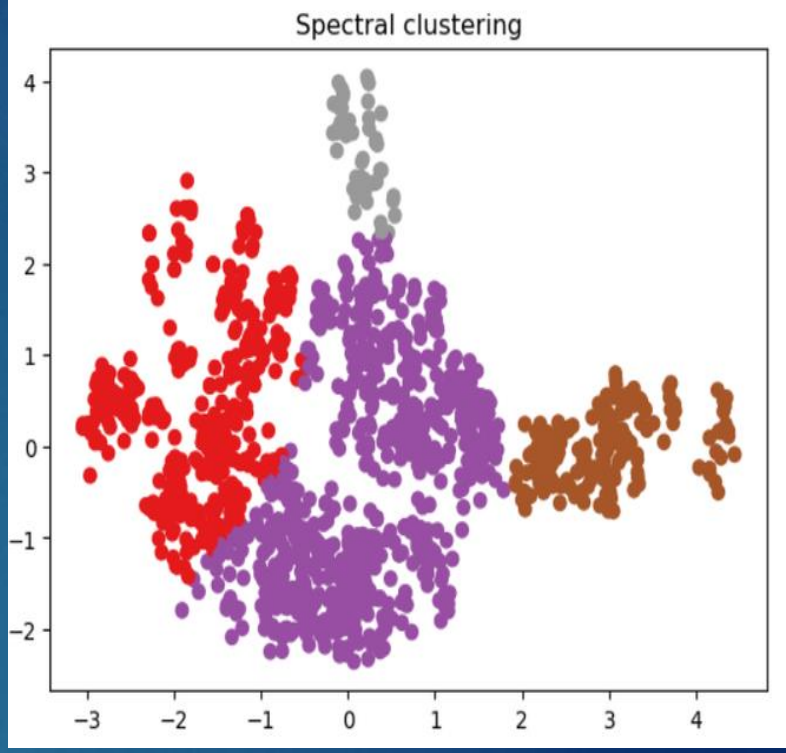
*It can work well with noisy data.

Disadvantage:

*It is computationally expensive especially for large datasets.

*It is sensitive to choice of bandwidth, which affects the cluster sizes and shapes

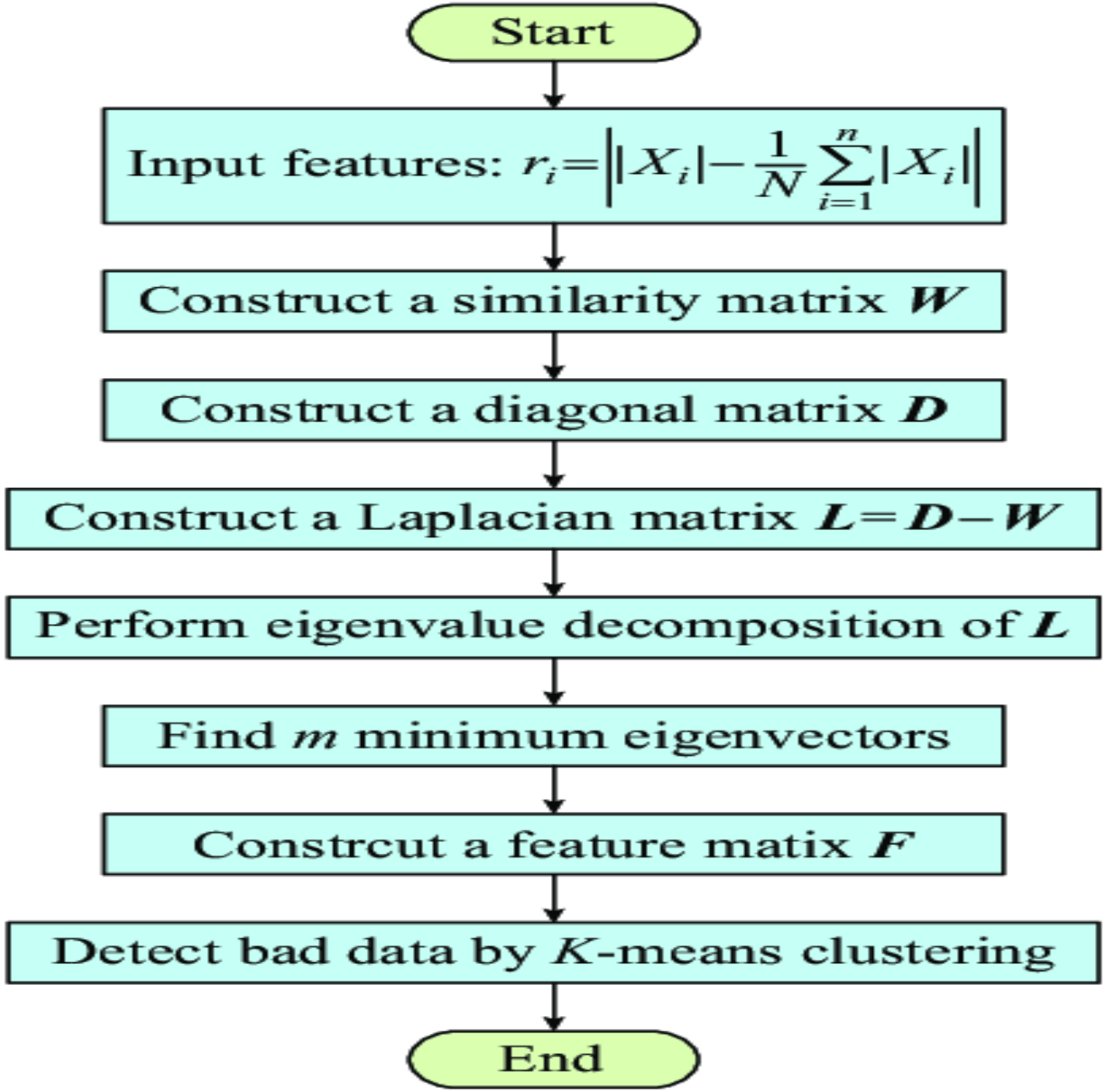
SPECTRAL CLUSTERING ALGORITHUM

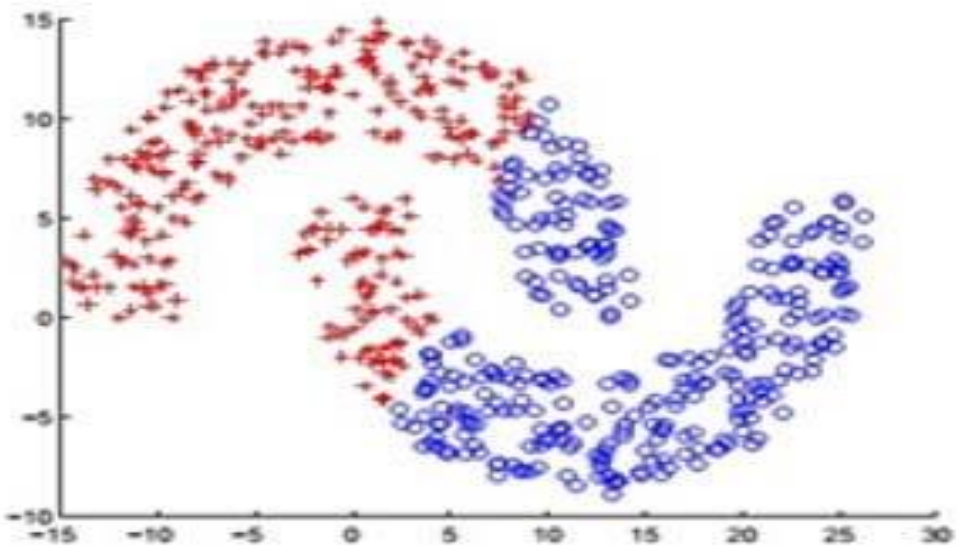
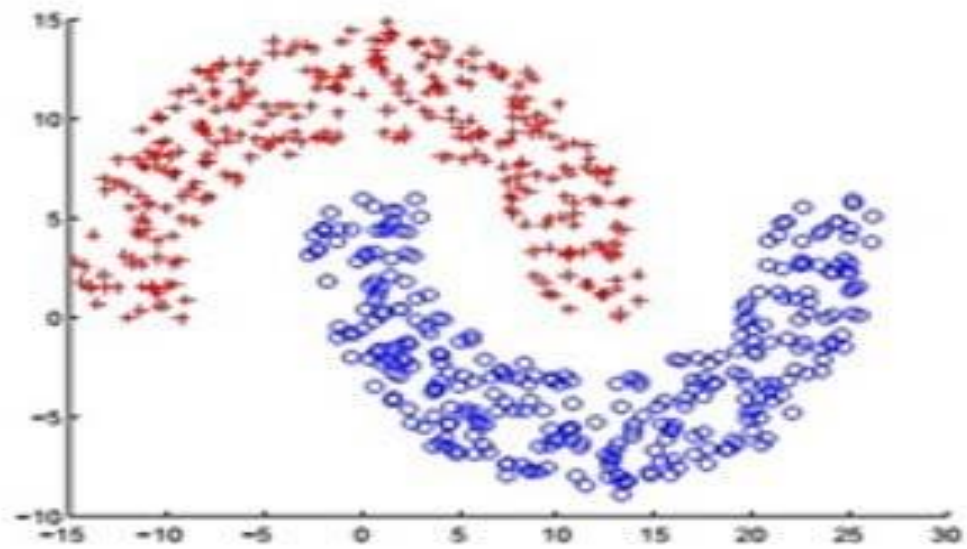


*In spectral clustering data points as nodes of a graph. Thus , spectral clustering is a graph portioning problem

*Low dimensional space and easily segregated to the cluster.

*Cluster within convex boundary.


Start
Input features: $r_i=\left\|\|X_i\|-\frac{1}{N}\sum_{i=1}^{n}\|X_i\|\right\|$
Construct a similarity matrix $\boldsymbol{W}$
Construct a diagonal matrix $\boldsymbol{D}$
Construct a Laplacian matrix $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$
Perform eigenvalue decomposition of $\boldsymbol{L}$
Find m minimum eigenvectors
Constrcut a feature matix $\boldsymbol{F}$
Detect bad data by K-means clustering
End

K- mean clustering	Spectral clustering
 (a) K-means	 (b) Spectral Clustering
It will assumes that the clusters are spherical or round or parabola ,within k-radius of cluster centroid.	**This well overcome the two problems**
Many iteration to find the cluster centroid	It does not follow any fixed shape

Advantages:

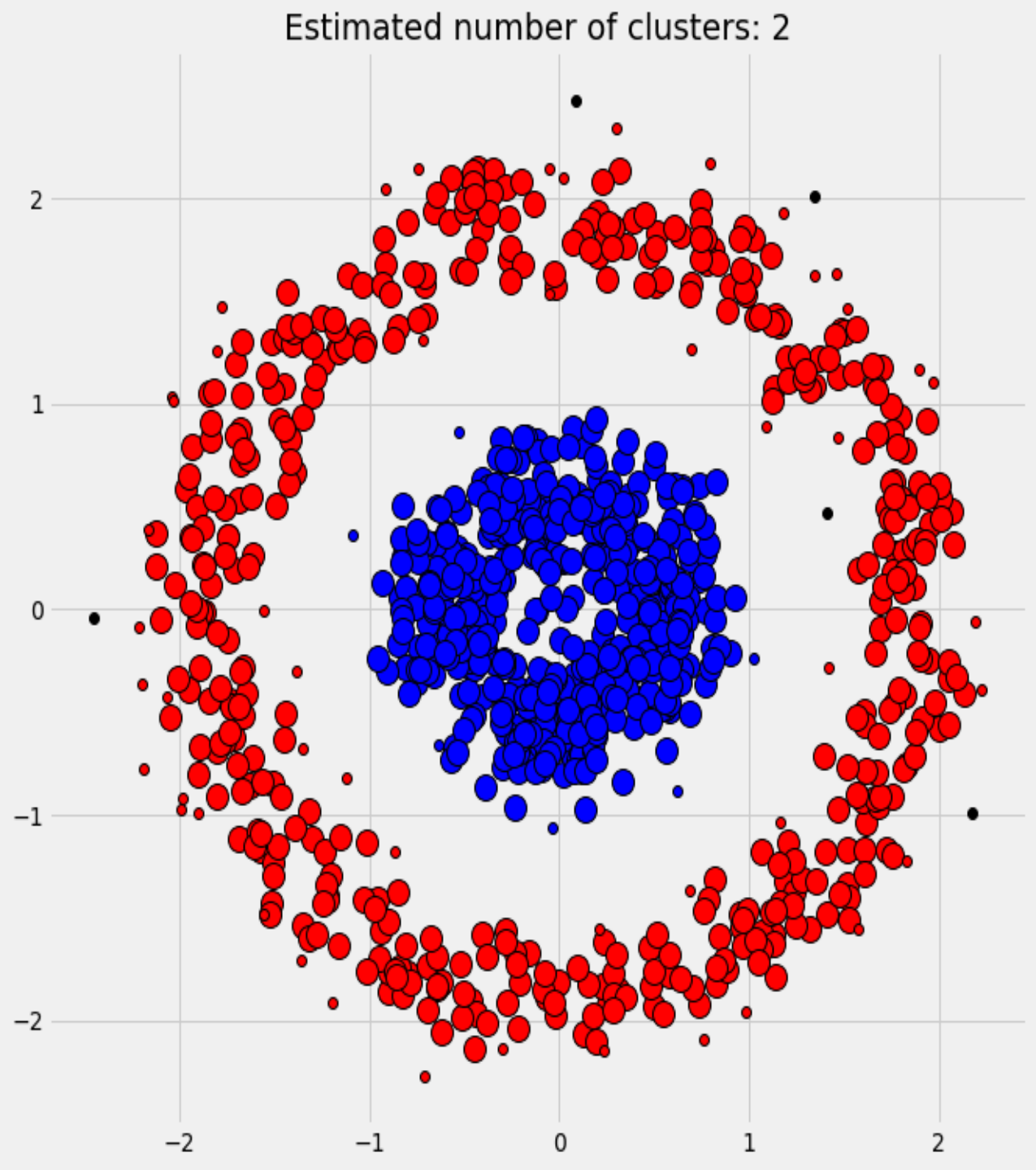
*Flexible of clustering shapes and size
*Applicable for high dimensional dataset.
*It will handle the non-linear data also.

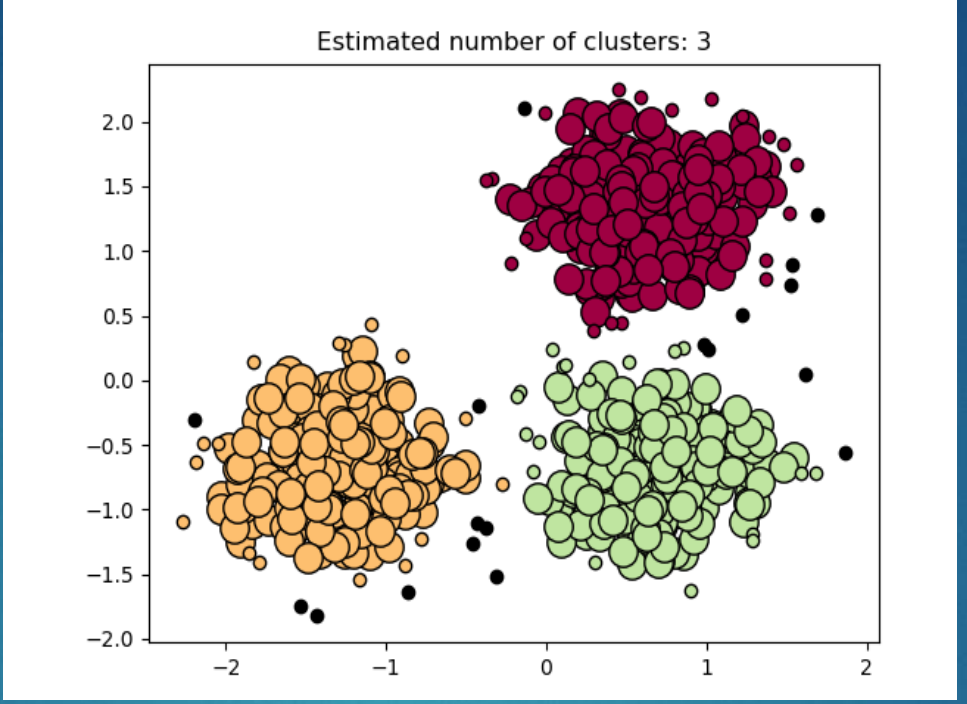
Disadvantages:

*Relatively slow.
*Need to select number of clusters.

DBSAN

Clustering





DBSCAN—Density Based Spatial Clustering of Application with Noise

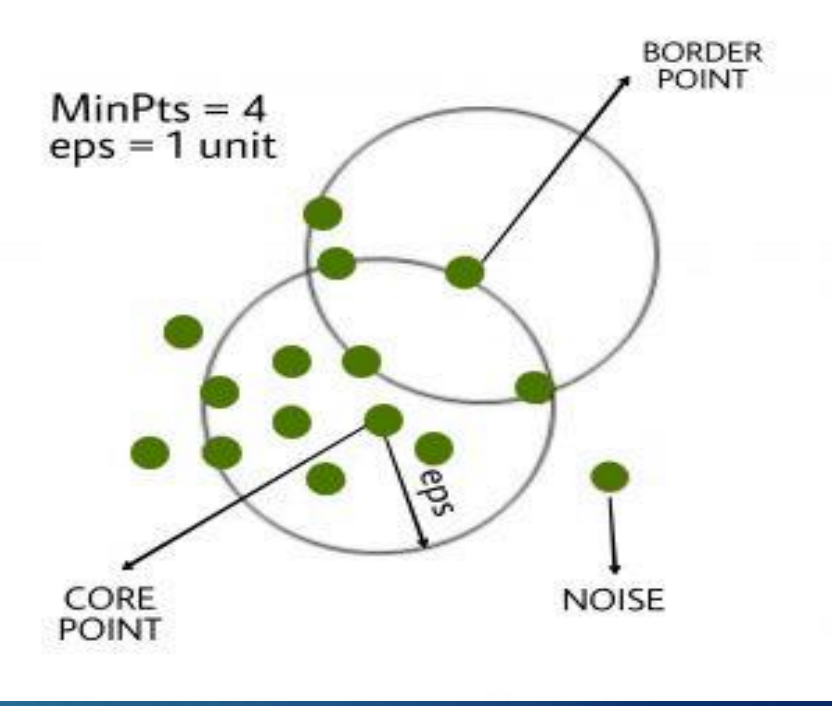
DBSCAN is a density based clustering algorithm that works on the assumption that clusters are dense regions in space separarted by region of lower density. It groups densely grouped data points into a single cluster.

In this algorithm, we have 3 types of data points.

Core Point: A point is a core point if it has more than MinPts points within eps.

Border Point: A point which has fewer than MinPts within eps but it is in the neighborhood of a core point.

Noise or outlier: A point which is not a core point or border point



Steps Used In DBSCAN Algorithm

1.Find all the neighbor points within eps and identify the core points or visited with more than MinPts neighbors.

2.For each core point if it is not already assigned to a cluster, create a new cluster.

3.Find recursively all its density-connected points and assign them to the same cluster as the core point.

3. iterate through the remaining unvisited points in the dataset.

DBSCAN	K-Means
In DBSCAN we need not specify the number of clusters.	K-Means is very sensitive to the number of clusters so it need to specified
Clusters formed in DBSCAN can be of any arbitrary shape.	Clusters formed in K-Means are spherical or convex in shape
DBSCAN can work well with datasets having noise and outliers	K-Means does not work well with outliers data. Outliers can skew the clusters in K-Means to a very large extent.
In DBSCAN two parameters are required for training the Model	In K-Means only one parameter is required is for training the model

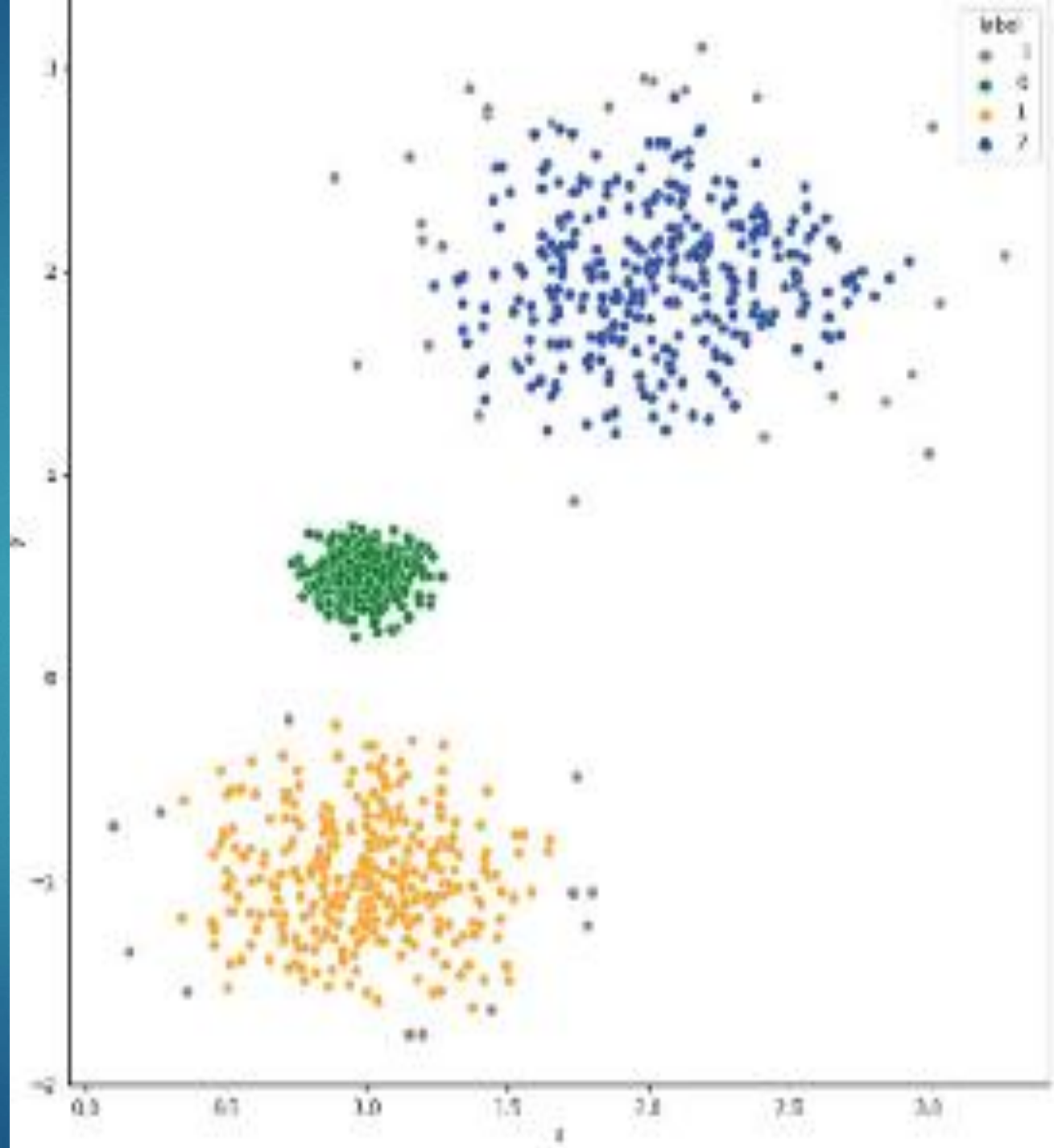
Advantages:

*DBSCAN is great at separating high-density clustering from low-density clustering.
*It will handle the noise and outlier.

Disadvantages:

it struggles with cluster of similar density.
*Struggle with high dimentionality data.

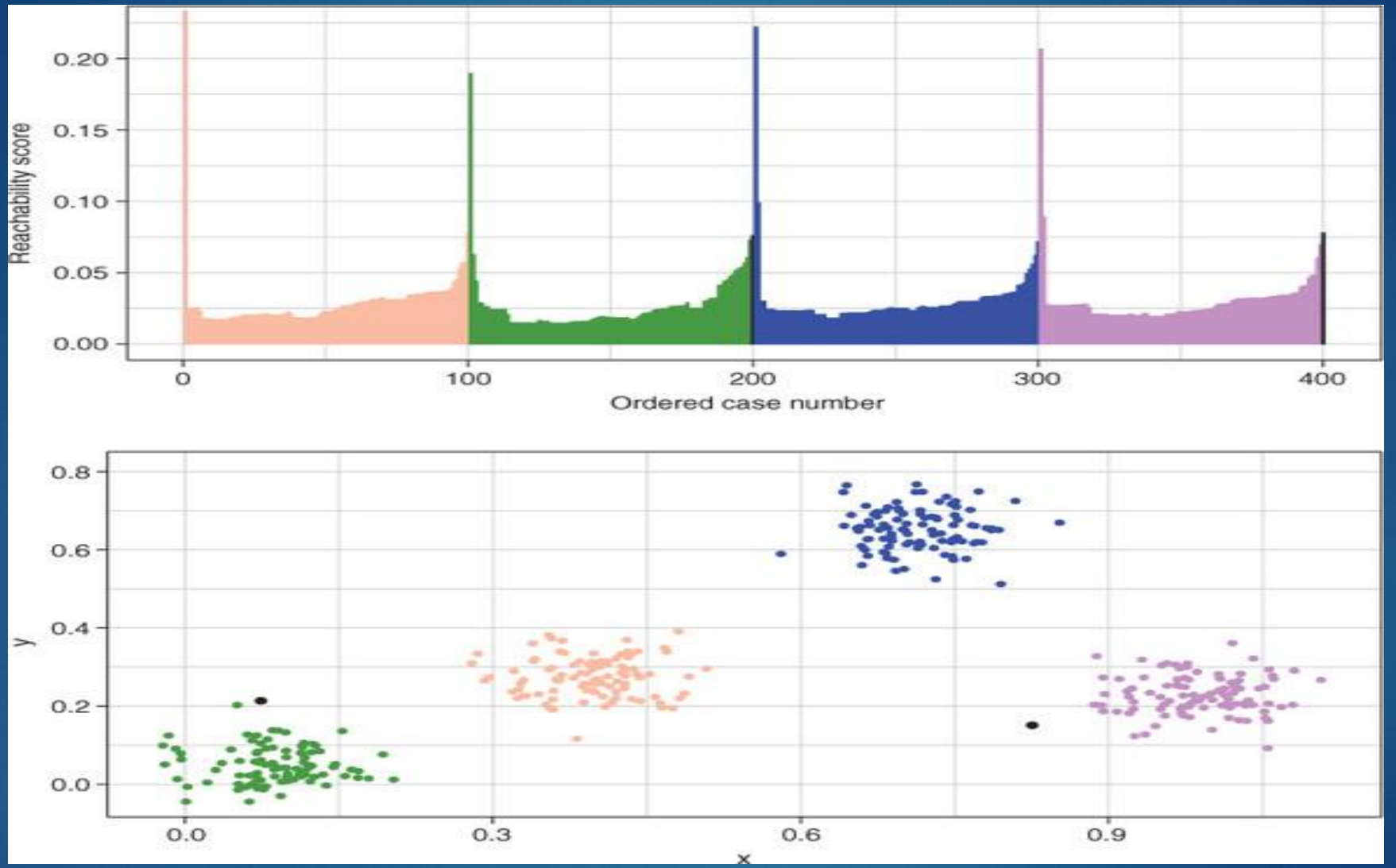
OPTICS Clustering



OPTICS– Ordering Points To Identify Structure

OPTICS is a Density based clustering algorithm, similar to DBSCAN but it can extract clusters of varying densities and shapes

The by representing of the data by creating ordered list of points call the reachability plot


Reachability score
0.20
0.15
0.10
0.05
0.00
0
100
200
300
400
Ordered case number
y
0.8
0.6
0.4
0.2
0.0
0.0
0.3
0.6
0.9
x

Advantages:

*OPTICS clustering does not require a predefined number of cluster in advance.

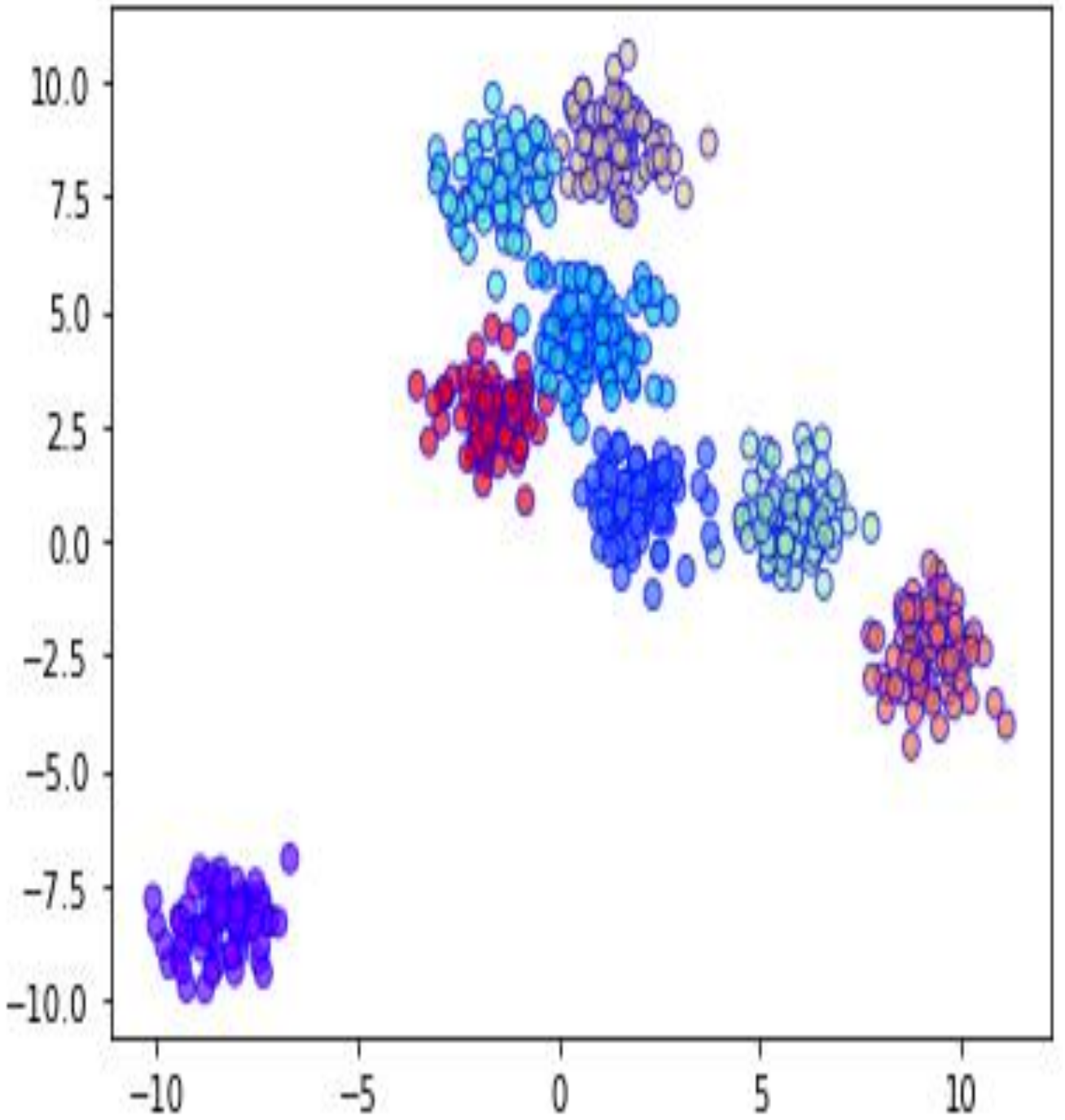
*Cluster can be any shape.

Disadvantages:

*It fails if there are no density drops between clusters

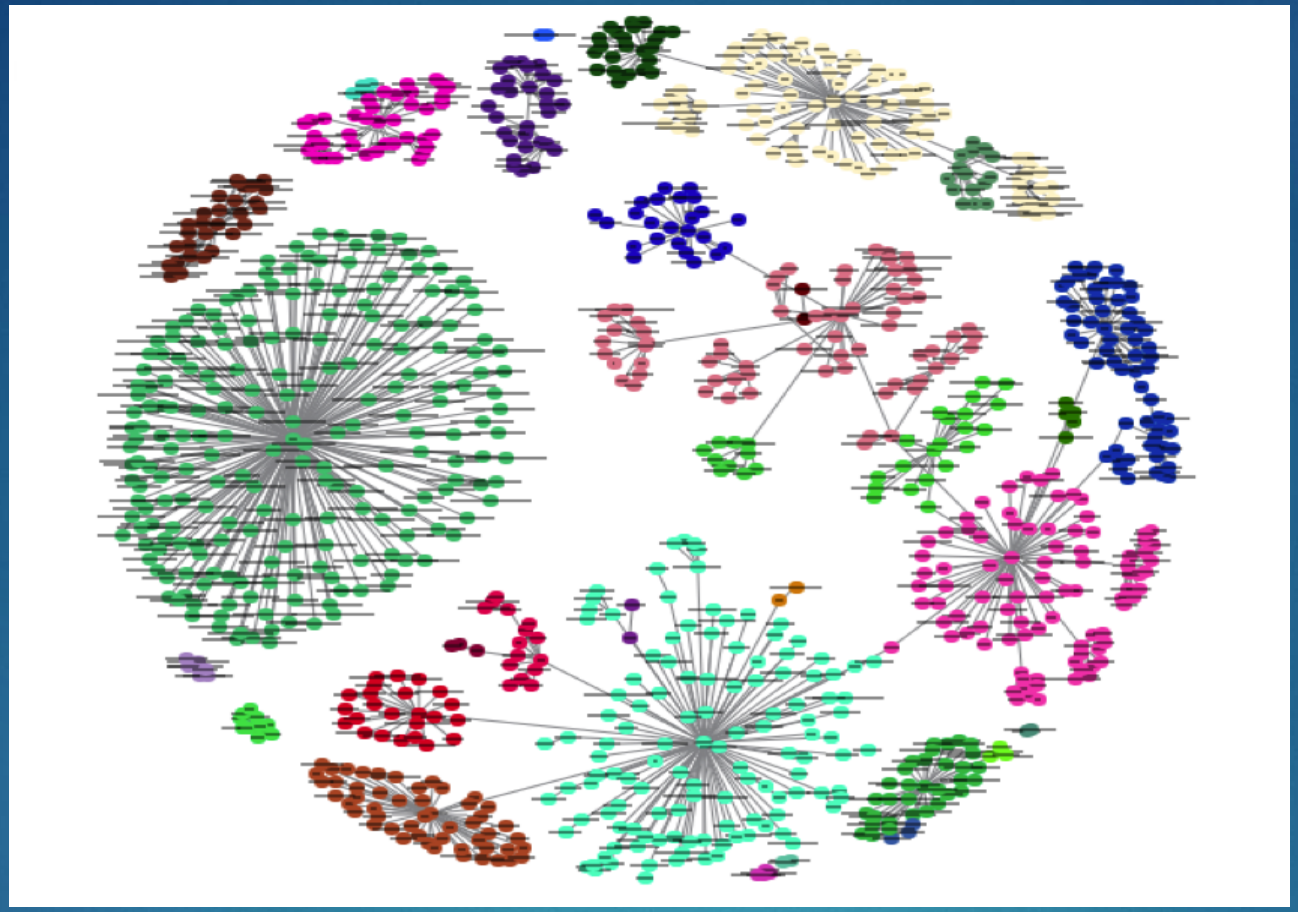
.

BIRCH CLUSTERING



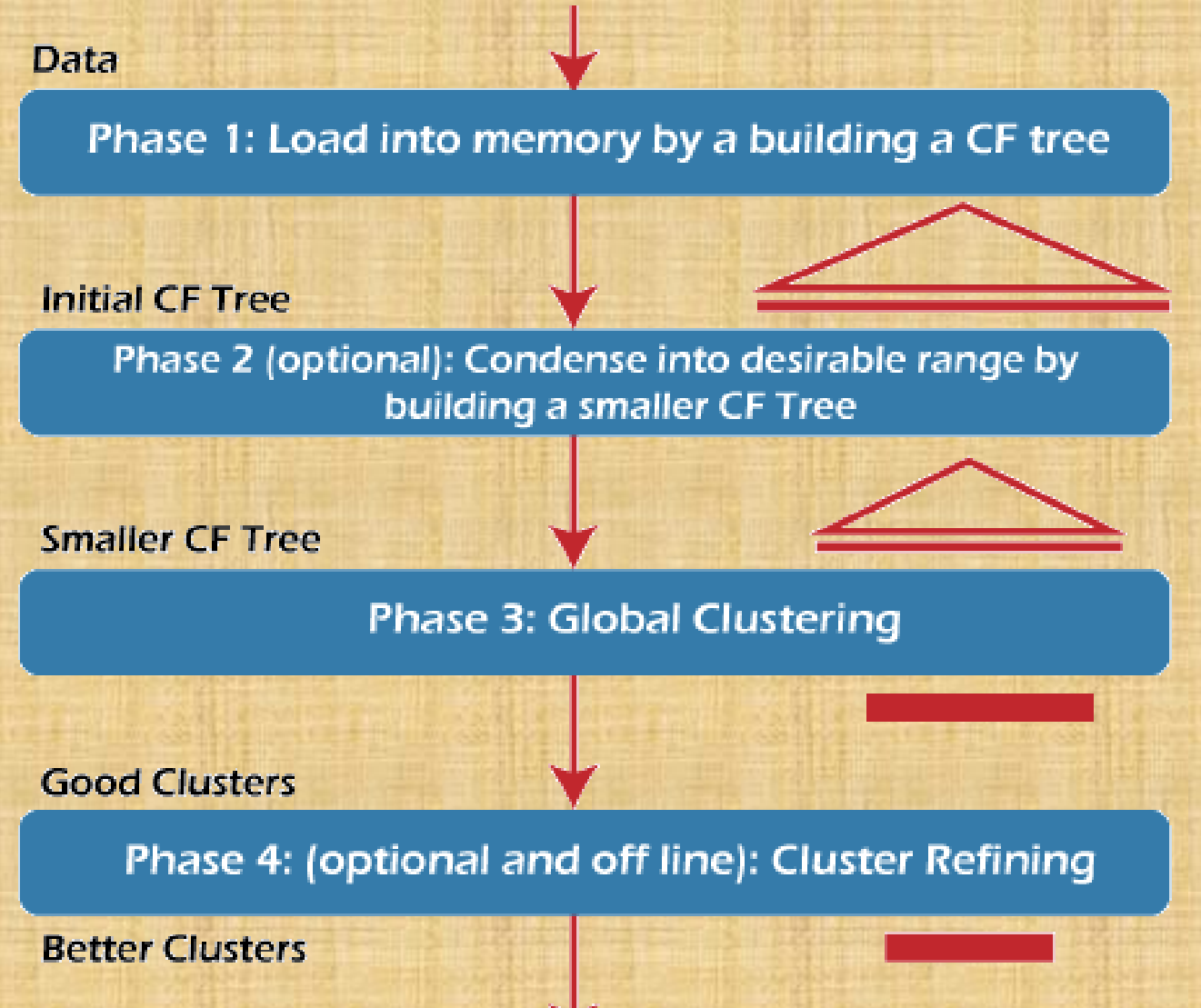
BIRCH– Balanced Iterative Reducing and Clustering using Hierarchies

Clustering algorithms like K-means clustering do not perform clustering very efficiently and it is difficult to process large datasets with a limited amount of resources (like memory or a slower CPU). So, regular clustering algorithms do not scale well in terms of running time and quality as the size of the dataset increases. This is where BIRCH clustering com



clustering algorithm that can cluster large datasets by first generating a small and compact summary of the large dataset that retains as much information as possible.

The BIRCH Clustering Algorithm



Advantages:

*It is useful for performing precise clustering on large dataset.

Disadvantages:

*A metric attribute is any attribute whose values can be represented in Euclidean space i.e., no categorical attributes should be present.

.