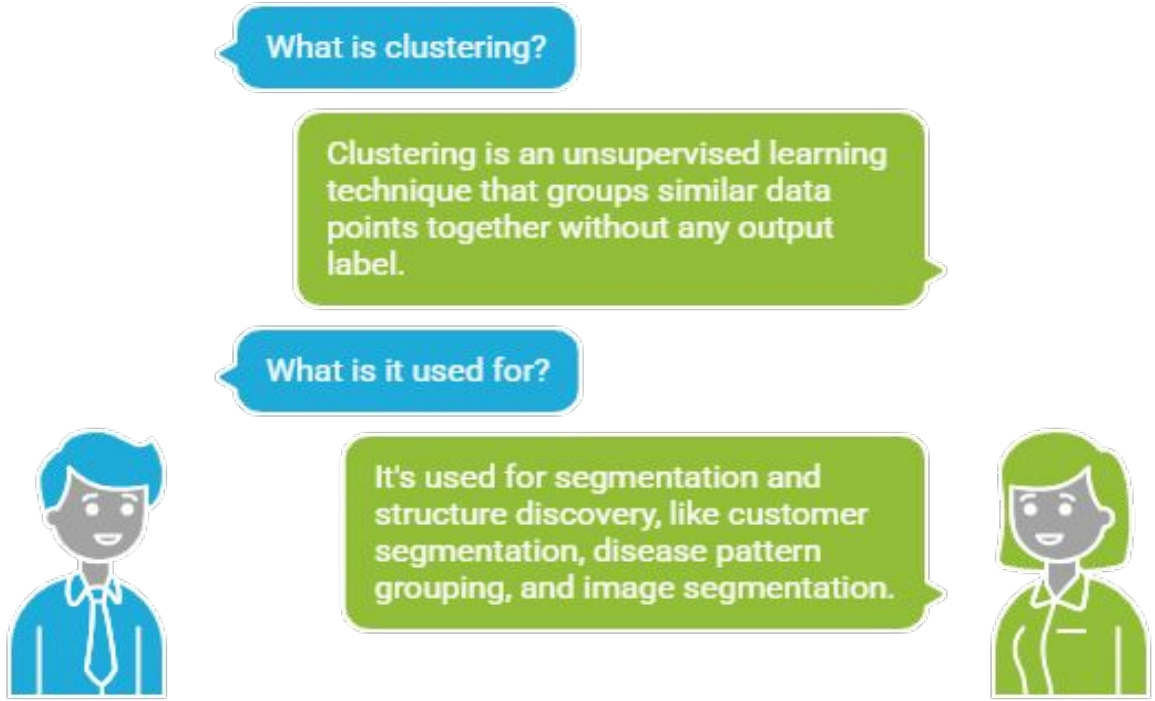
Machine Learning

Clustering Algorthims

A Conceptual and Practical Guide

Prepared by: Keerthana R

What is Clustering?



Which clustering algorithm should be used?

Graph-based

Ideal for identifying clusters in complex networks (e.g., Spectral Clustering).

Hierarchical-based

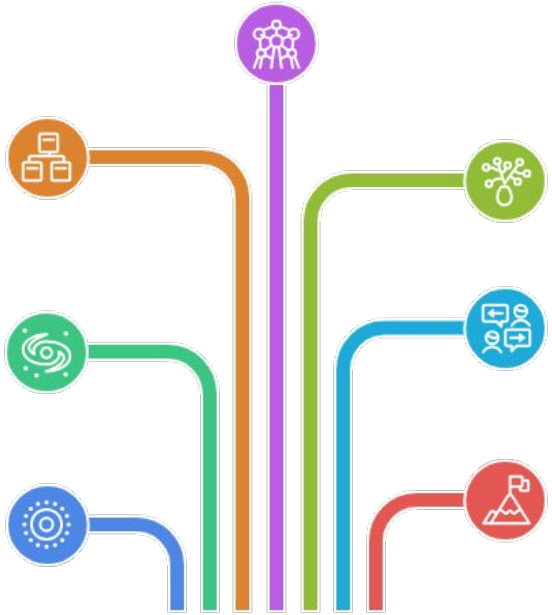
Useful for understanding cluster relationships and creating a dendrogram (e.g., Agglomerative Clustering).

Density-based

Effective for clusters with irregular shapes and varying densities (e.g., DBSCAN).

Centroid-based

Suitable for spherical clusters with similar sizes (e.g., K-Means).



Tree-based

Efficient for large datasets with a hierarchical structure (e.g., BIRCH).

Message-passing

Good for finding clusters without specifying the number of clusters (e.g., Affinity Propagation).

Mode-seeking

Effective for finding clusters with high-density regions (e.g., Mean Shift).

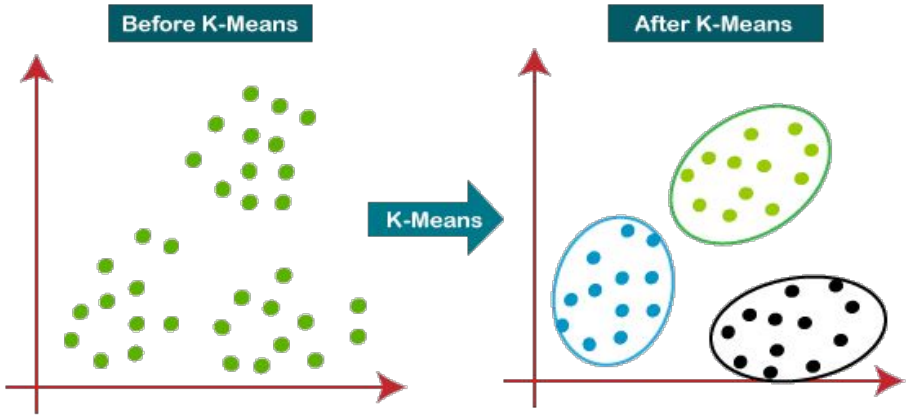
K-Means Clustering

Divides data into K clusters using centroids.

Objective:

Minimize Within Cluster Sum of Squares (WCSS)

$$WCSS = \Sigma\Sigma(x_i - \mu_k)^2$$



K-Means Clustering Optimization Process

Choose K

Select the number of clusters

Assign Points

Assign data points to nearest centroids

Repeat Until Convergence

Continue until WCSS is minimized

Initialize Centroids

Set initial cluster centers

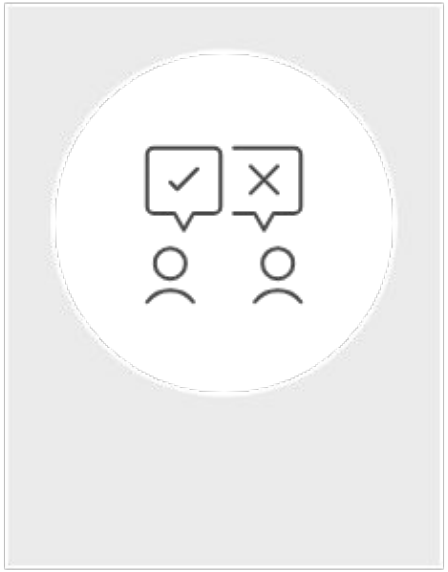
Update Centroids

Recalculate centroids based on assignments

K-Means Clustering

Pros

- Simple and fast
- Spherical clusters



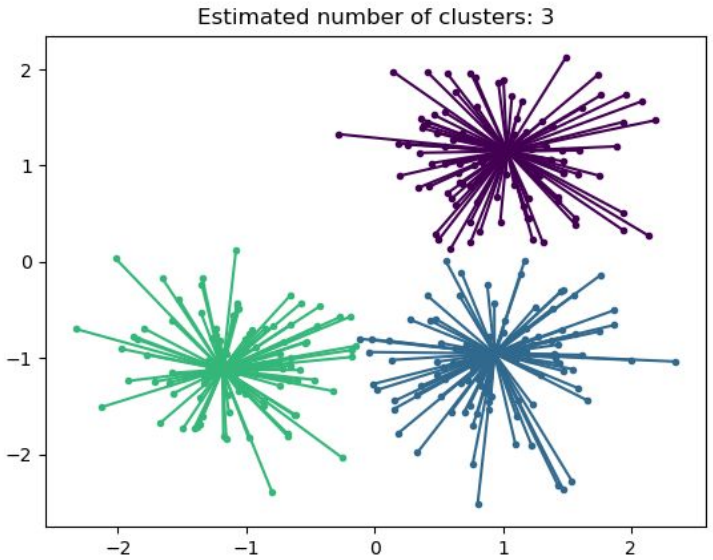
Cons

- Specify K
- Sensitive to outliers
- Irregular shapes

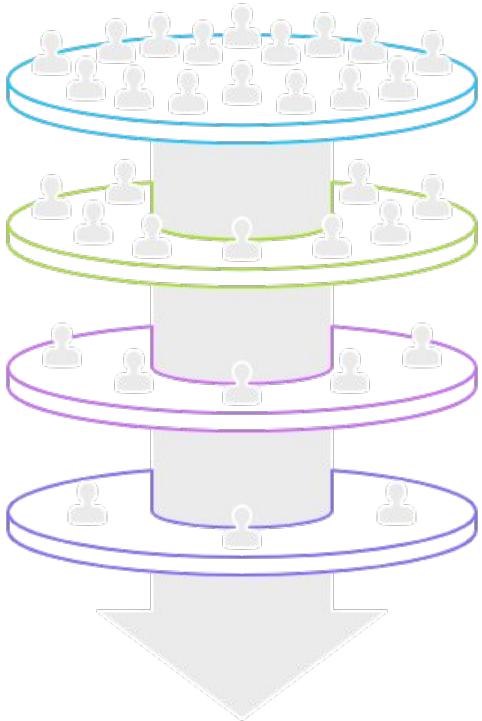
Affinity Propagation

Clusters are formed by exchanging messages between data points.

No need to predefine number of clusters.



Affinity Propagation Clustering Process



Message Exchange

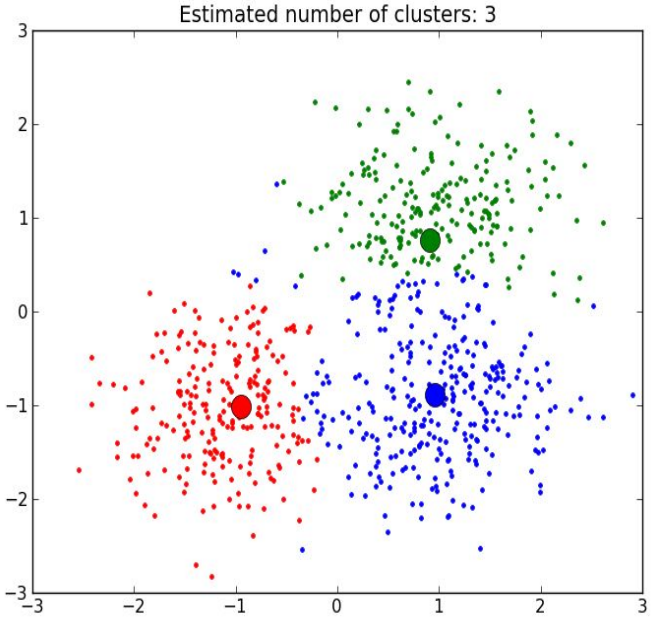
Responsibility Calculation

Availability Calculation

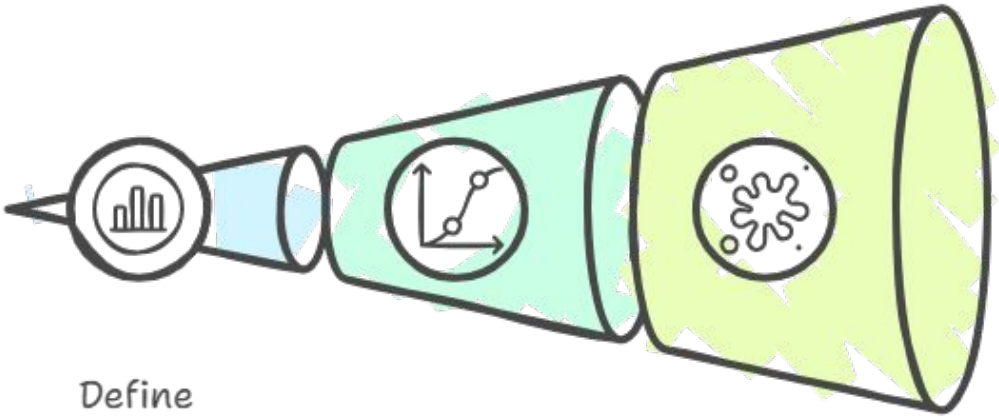
Exemplar Selection

Mean Shift Clustering

Shifts cluster center towards high-density regions. No need to specify K.



Mean Shift Clustering Process



Define Bandwidth

Set the radius for density estimation

Shift Points

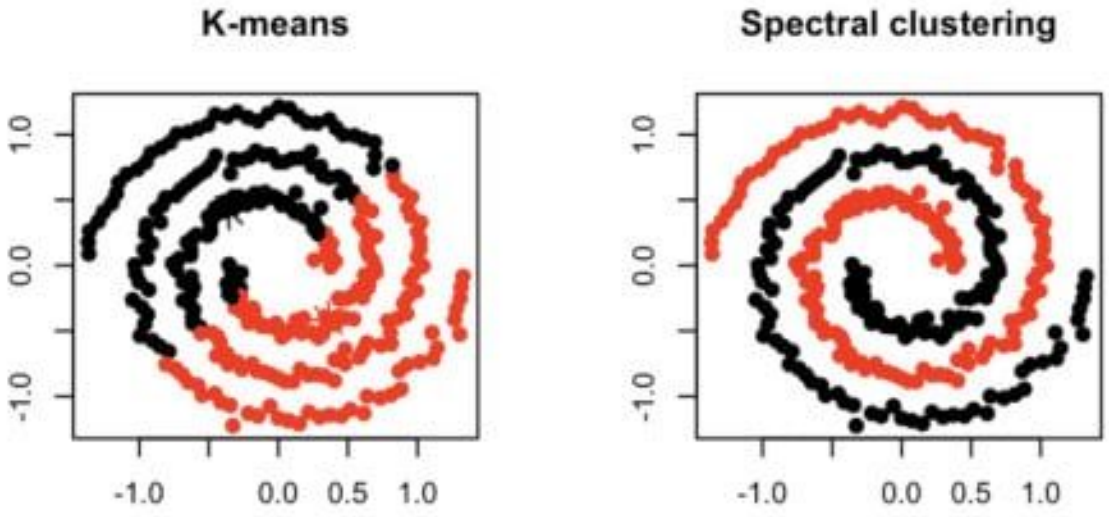
Move points towards dense areas

Converge to Mode

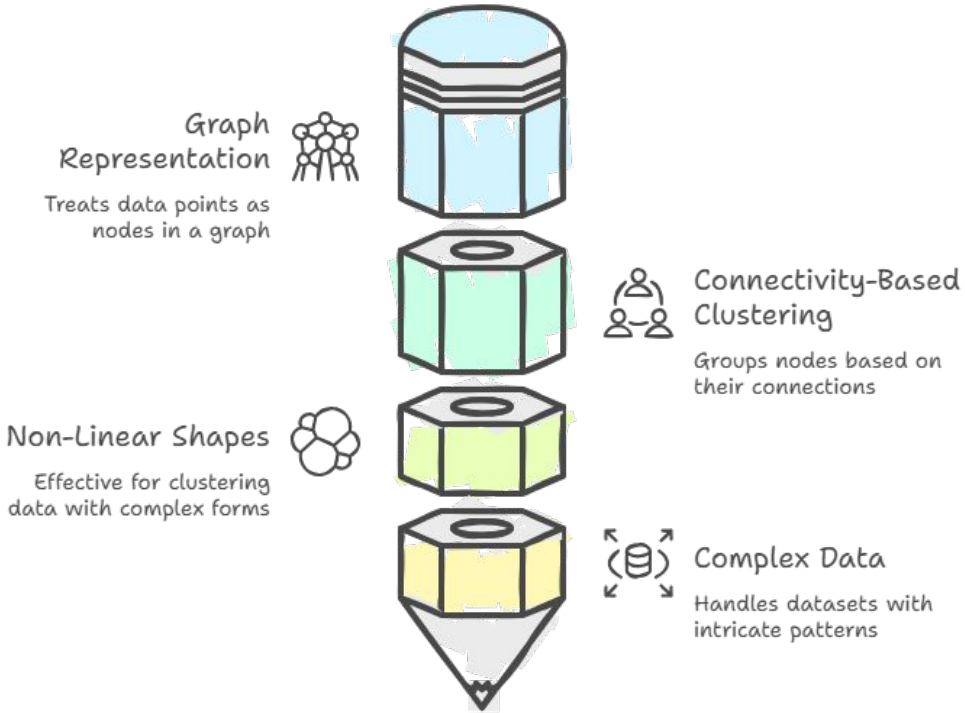
Points settle at density peaks

Spectral Clustering

Treats data points as graph nodes.

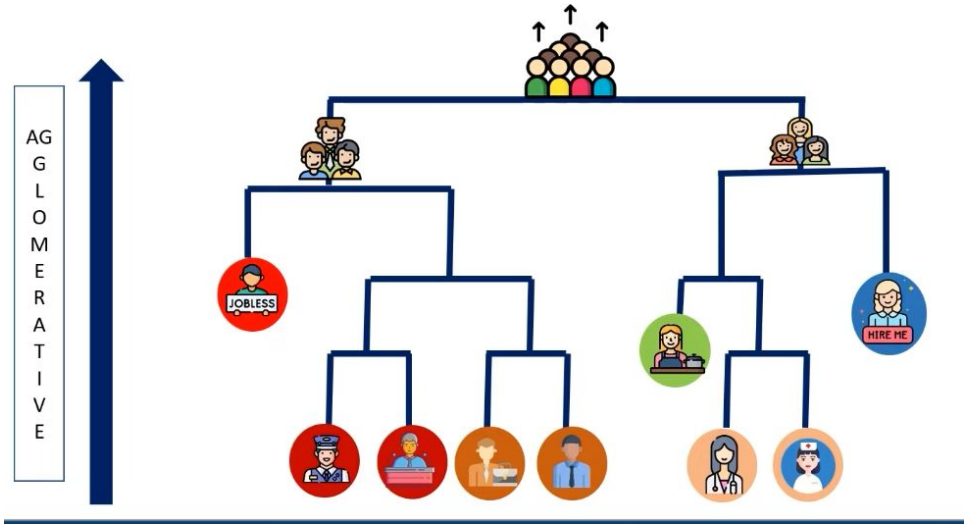


Understanding Spectral Clustering

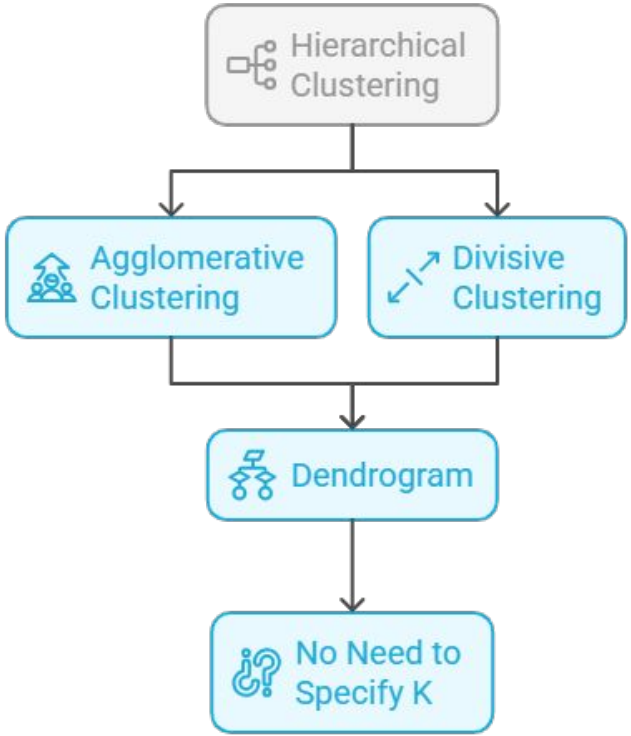


Hierarchical Clustering

Hierarchical clustering groups data based on similarity by building a tree-like structure of nested clusters

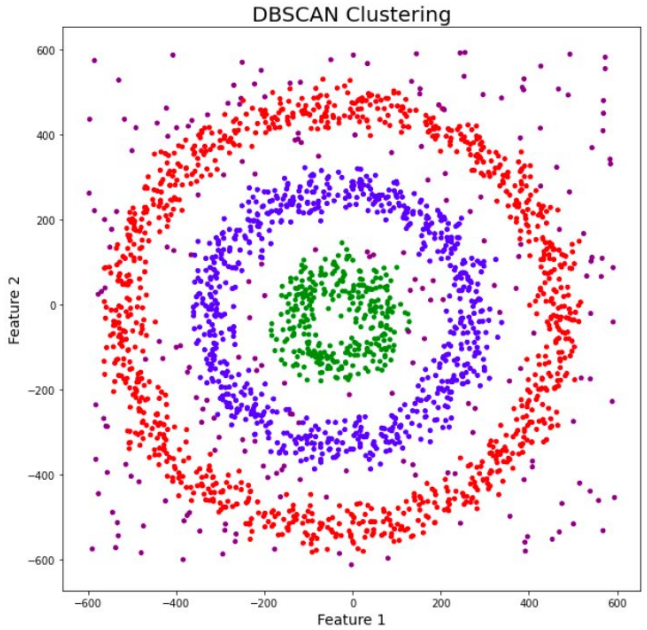


Hierarchical Clustering Process

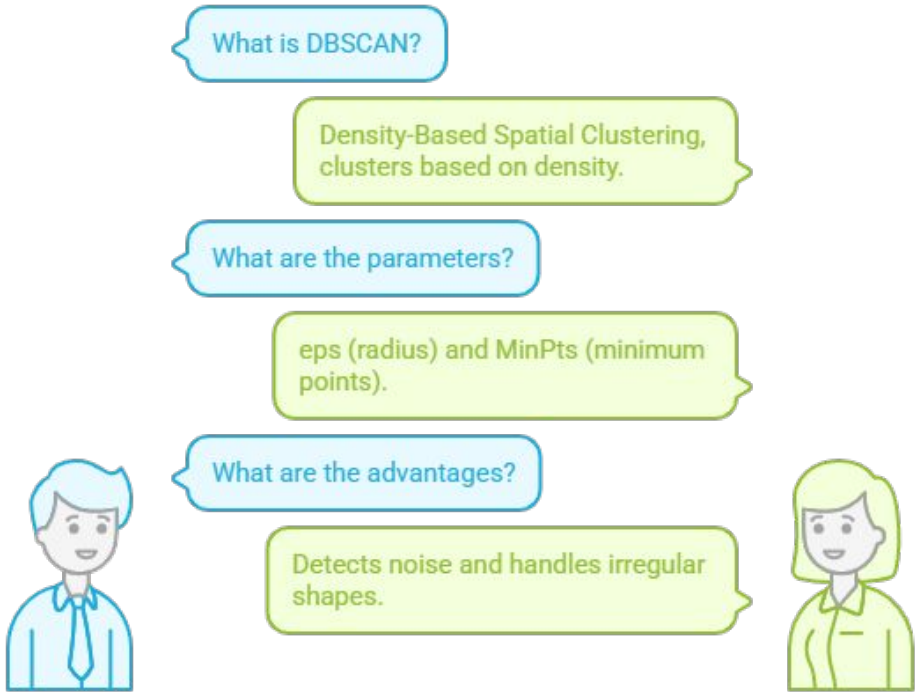


DBSCAN

Density-Based Spatial Clustering

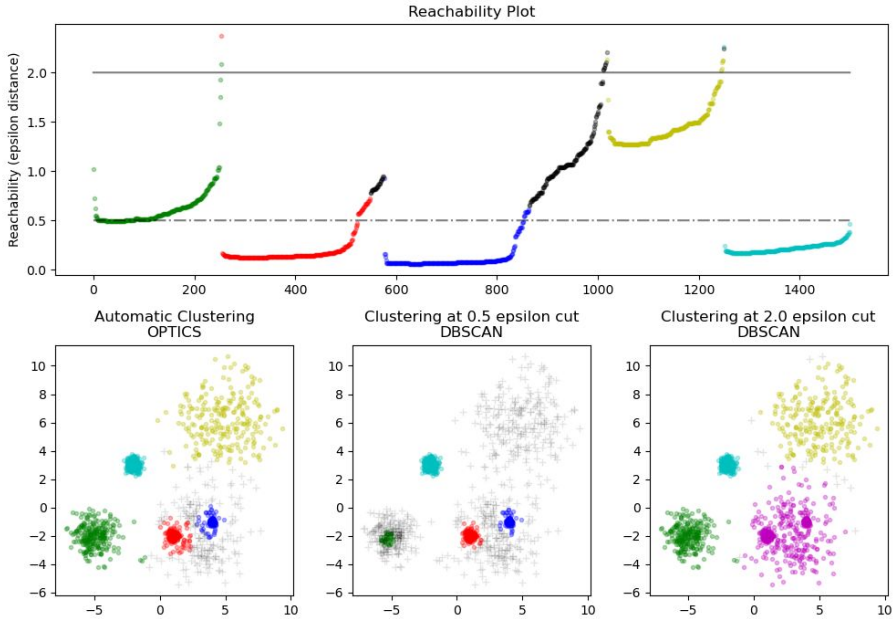


DBSCAN Clustering

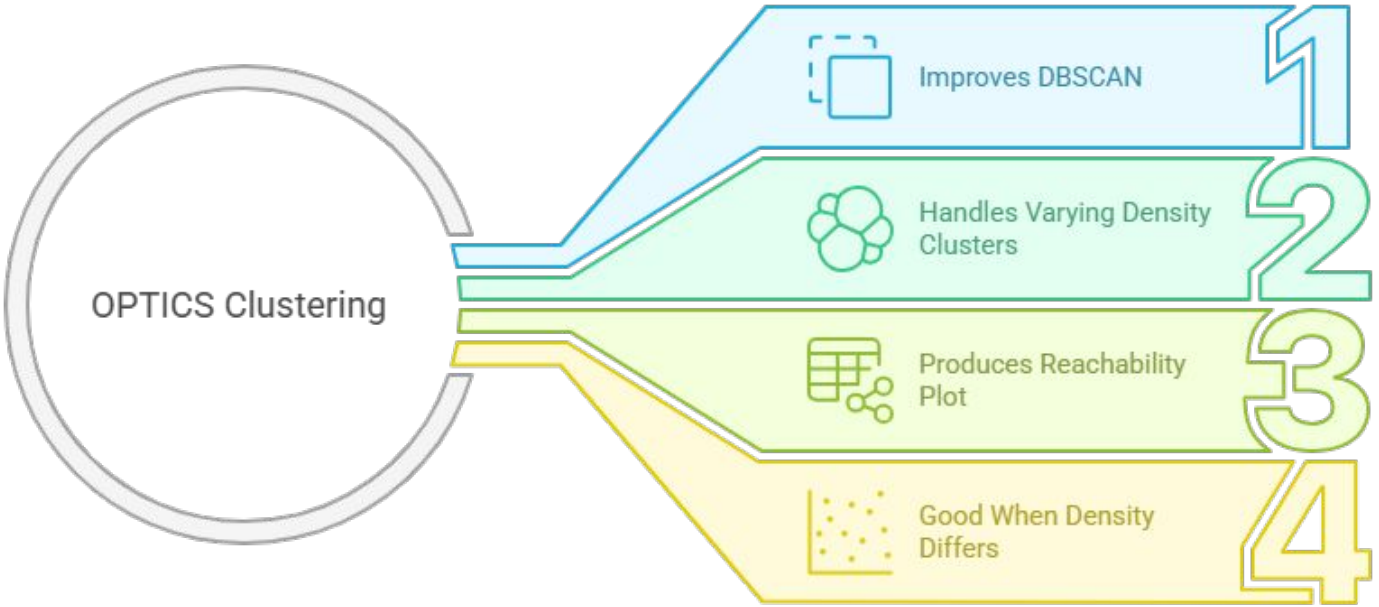


OPTICS

A reachability plot is a graph that helps visualize clustering structures. It shows the reachability distance of each point in the dataset.



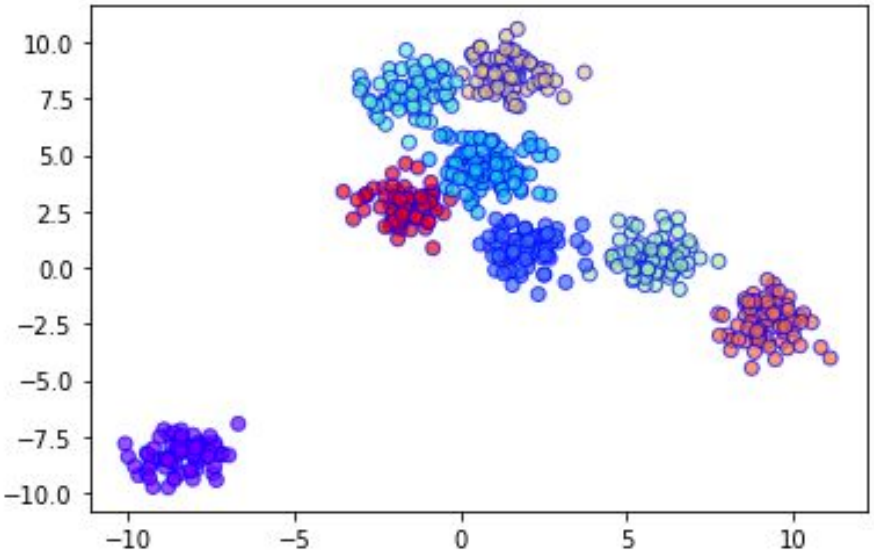
Unveiling OPTICS Clustering



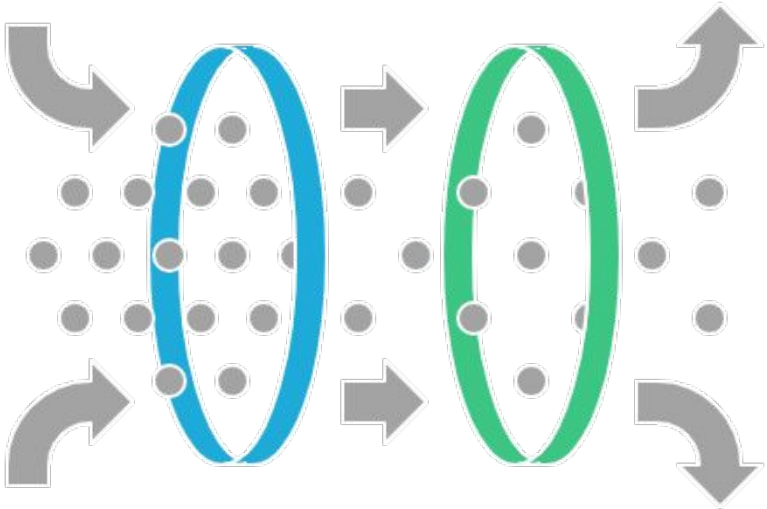
BIRCH

Balanced Iterative Reducing and Clustering using Hierarchies.

Designed for large datasets.



BIRCH Clustering Process



Build CF Tree

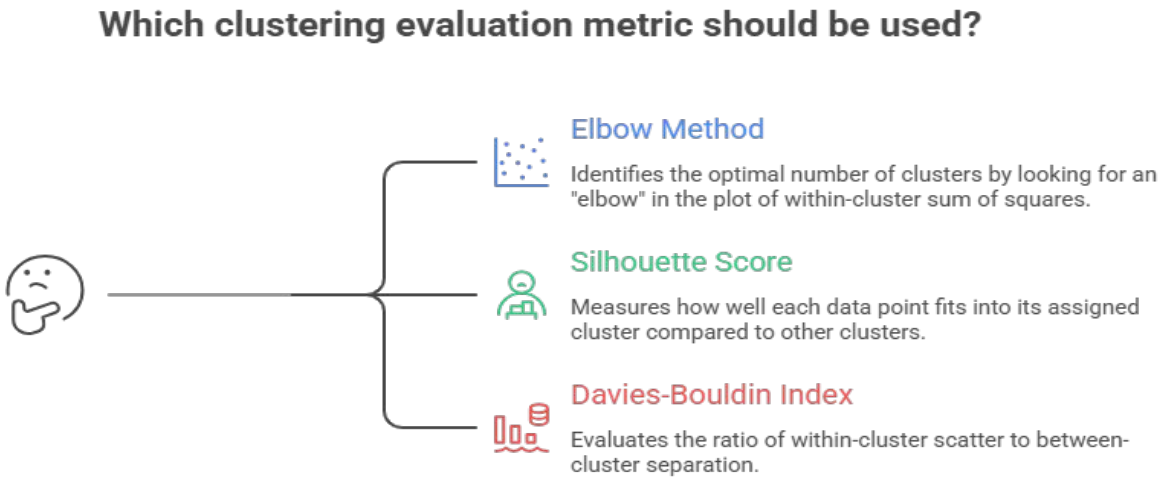
Create a hierarchical structure for data

Global Clustering

Apply clustering to the CF Tree

Clustering Evaluation

Since clustering has no labels, evaluation uses internal metrics:



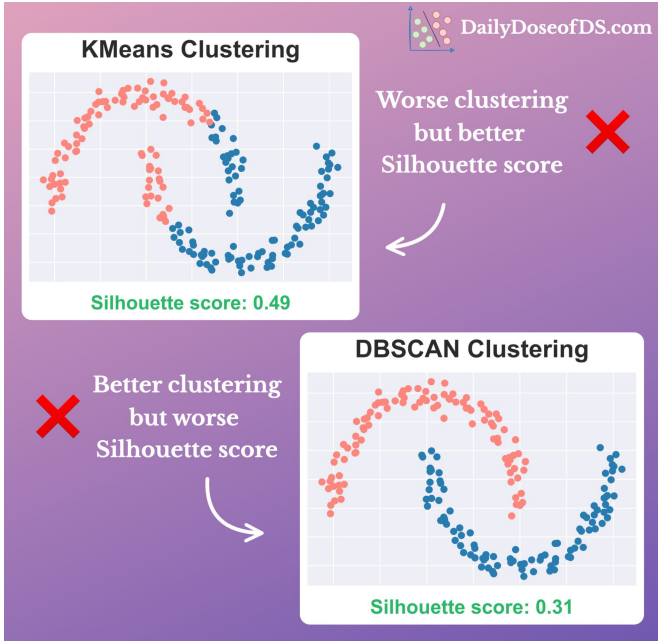
Silhouette Formula:

$$S = (b - a)/max(a, b)$$

Where:

a = intra-cluster distance

b = nearest cluster distance



Algorithm Comparison — Practical Checklist

Use this quick reference to match algorithms to dataset needs: whether a method requires K, handles noise, finds irregular shapes, and how it scales.

Algorithm	Needs K	Handles Noise	Irregular Shapes	Large Dataset
K-Means	Yes	No	No	Yes
DBSCAN	No	Yes	Yes	Medium
OPTICS	No	Yes	Yes	Medium
BIRCH	Yes (leaf summaries)	No	Medium	Yes
Spectral	Yes	No	Yes	No

Note: "Medium" indicates suitability depends on implementation and data dimensionality. Always pre-process (normalize, reduce dimensionality) before choosing an algorithm.

Conclusion and Key Takeaways

1

Clustering Importance

Essential technique for uncovering insightful patterns in large datasets.

2

Algorithm Selection

Choice must consider data structure, distribution, and cluster shapes.

3

Evaluation Metrics

Utilize internal methods like Silhouette Score for clustering effectiveness.

4

No One-size-fits-all

Different clustering algorithms best for varying dataset characteristics.

THANK YOU!

These algorithms play a critical role in discovering hidden patterns, grouping similar data points, and enabling intelligent decision-making in real-world machine learning applications.

I hope this presentation provided a clear conceptual understanding of clustering techniques and their practical significance.

Keerthana R
keerthanauxd@gmail.com