Cluster analysis: basic concepts and methods

What is cluster analysis?

Cluster analysis or simply clustering is the process of partitioning a set of data objects (or observations) into subsets.

- **Scenario:**
  - As the Director of Customer Relationships at a retail company, managing millions of customers individually is inefficient.
  - Need to organize customers into smaller groups for effective management.
- **Goal:**
  - Group customers based on similarities, ensuring each group has common business patterns.
  - Avoid mixing customers with significantly different behaviors.
- **Objective:**
  - Develop targeted customer relationship campaigns for each group, based on shared features.
- **Challenge:**
  - Unlike classification, group labels are unknown; need to **discover** groupings.
  - Manually analyzing large customer datasets is impractical.
- **Solution:**
  - Use **clustering techniques** to automatically find meaningful customer groups.

What is cluster analysis?

- What is a cluster?
  - A cluster is a collection of data objects which are
    - Similar (or related) to one another within the same group (i.e., cluster)
    - Dissimilar (or unrelated) to the objects in other groups (i.e., clusters)
- **Cluster analysis** (or clustering, data segmentation, ...)
  - Given a set of data points, partition them into a set of groups (i.e., clusters), such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups
- Typical ways to use/apply cluster analysis
  - As a stand-alone tool to get insight into data distribution, or
  - As a preprocessing (or intermediate) step for other algorithms

Applications of Clustering

- **Business Intelligence:**
  - Grouping customers based on purchase patterns for targeted marketing.
  - Organizing projects by similarities (e.g., type, duration, complexity) to improve management.
- **Image Recognition:**
  - Grouping photos with similar features (e.g., faces, backgrounds).
  - Automatically categorizing images without manual labeling.
- **Web Search:**
  - Organizing search results into clusters for easier access.
  - Grouping web pages by topic for better information retrieval.
- **Data Segmentation:**
  - Dividing large datasets into smaller groups based on similarities.
- **Outlier Detection:**
  - Identifying unusual data points that don’t fit into any cluster.
  - Examples: Detecting credit card fraud based on unusual spending patterns.
- **Exploratory Data Analysis:**
  - Helps in understanding data characteristics and relationships.

Clustering vs. Classification

Clustering

- **Unsupervised learning**: Groups are unknown and need to be discovered.
- Learns by finding patterns in the data without predefined labels.

Classification

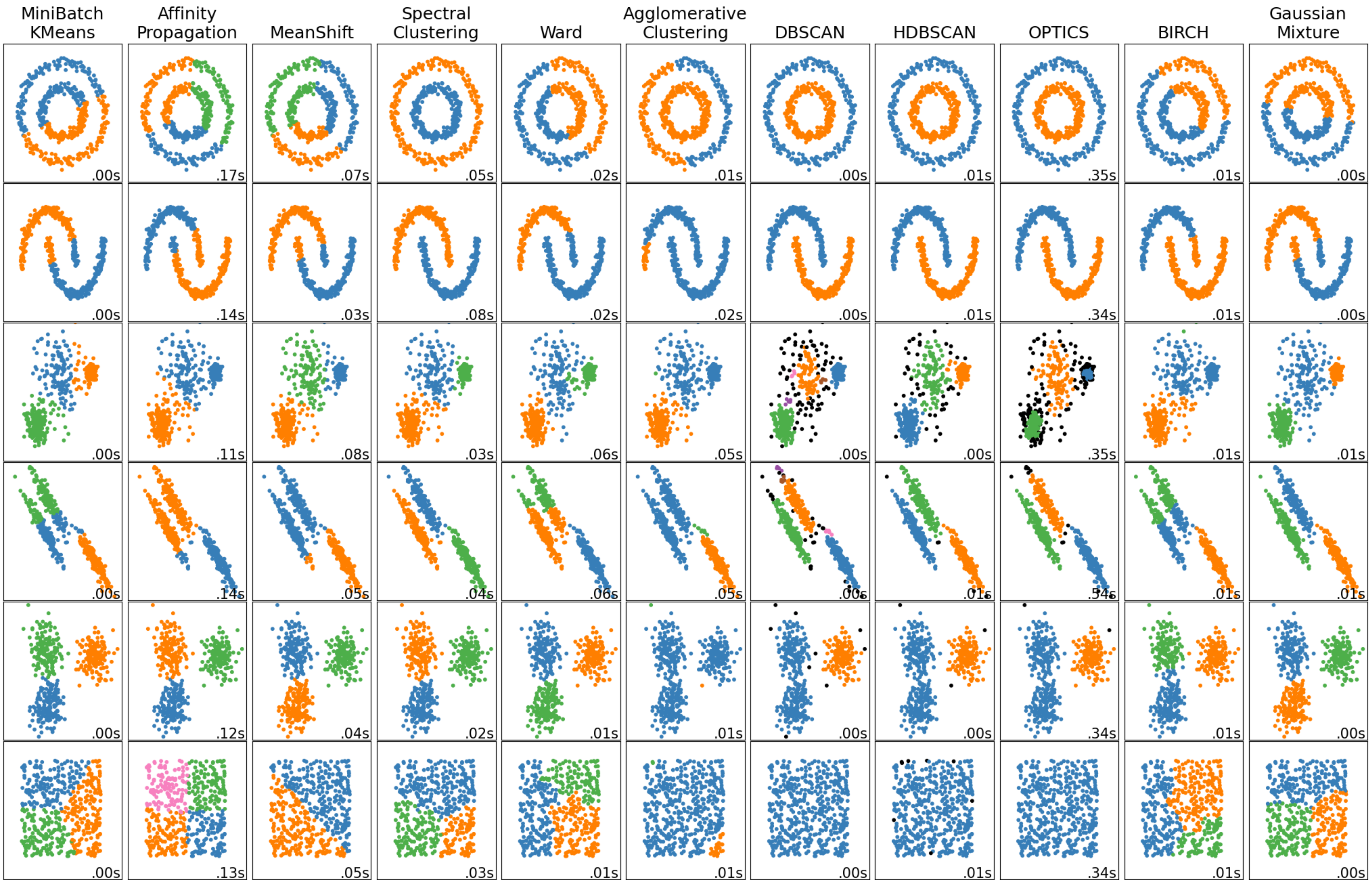
- **Supervised learning**: Uses known labels to train a model.
- Learns by examples where class membership is given.

Types of Data and Clustering Challenges

- Complex Data Types:
  - Can work with text, images, graphs, and mixed data types.
- High-Dimensional Data:
  - Clustering data with many features or attributes can be challenging.
- Scalability:
  - Techniques need to handle very large datasets effectively.

Methods and Tools for Clustering

- Common Algorithms:
  - K-means, K-medoids for partitioning data based on distance.
  - Agglomerative clustering
  - Density based clustering
- Tools:
  - Built into statistical software (e.g., SPSS, SAS).


MiniBatch KMeans
Affinity Propagation
MeanShift
Spectral Clustering
Ward
Agglomerative Clustering
DBSCAN
HDBSCAN
OPTICS
BIRCH
Gaussian Mixture
.00s .17s .07s .05s .02s .01s .00s .01s .35s .01s .00s
.00s .14s .03s .08s .02s .02s .00s .01s .34s .01s .00s
.00s .11s .08s .03s .06s .05s .00s .00s .35s .01s .01s
.00s .14s .05s .04s .06s .05s .00s .01s .54s .01s .01s
.00s .12s .04s .02s .01s .01s .00s .00s .34s .01s .00s
.00s .13s .05s .03s .01s .01s .00s .01s .34s .01s .00s

Requirements for cluster analysis

Clustering is a challenging research field. In this section, we will discuss about the requirements for clustering as a data mining tool, as well as aspects that can be used for comparing clustering methods.

Requirements for Cluster Analysis algorithms

- Clustering is complex and has several requirements to be effective in data mining.
- Different methods exist, and each has its strengths and weaknesses.
- **Ability to Handle Different Types of Data**
  - Algorithms should cluster various data types: numerical, categorical (e.g., colors or brands), text, images, etc.
  - Example: Clustering customer profiles that include age (numerical), preferences (categorical), and past purchases (text).

Requirements for Cluster Analysis algorithms (cont.)

- **Scalability** (Handling Large Datasets):
  - Must work well with large databases, sometimes containing millions of data points.
  - Algorithms need to process big data without sampling too much, which could cause biased results.
  - Example: Clustering search engine data with billions of queries.
- Discovering **Clusters with Arbitrary Shapes** (Beyond Spherical Clusters):
  - Some methods find clusters of similar size and shape, but clusters can be irregularly shaped.
  - Example: Identifying the boundary of a wildfire in satellite images, which is not a perfect circle.

Requirements for Cluster Analysis algorithms (cont.)

- **Reducing the Need for Domain Knowledge**:
  - Many methods require setting parameters (e.g., number of clusters), which is difficult for complex data.
  - Algorithms should help users explore the data without needing detailed knowledge upfront.
- **Robustness to Noisy Data:**
  - Real datasets often have errors, missing values, or outliers.
  - Clustering should be able to deal with these issues and still produce meaningful results.
  - Example: Sensor data with occasional faulty readings.

Requirements for Cluster Analysis algorithms (cont.)

- **Incremental Clustering and Input Order Sensitivity:**
  - Algorithms should handle updates without needing to start from scratch.
  - Results should be consistent regardless of the data's input order.
  - Example: Adding new customer data over time without reprocessing the entire dataset.
- **Clustering High-Dimensional Data**: Challenges with Many Features:
  - High-dimensional data (many attributes) can be sparse and difficult to cluster.
  - Example: Document clustering, where each keyword is a dimension, resulting in thousands of dimensions.
- **Constraint-Based Clustering** Clustering with Conditions:
  - Sometimes clustering needs to follow specific rules or constraints.
  - Example: Deciding the locations for electric vehicle charging stations while considering space availability and power networks.
- **Interpretability and Usability** Making Results Understandable:
  - Clustering outcomes should be easy to interpret and use.
  - Example: Grouping customers in ways that are meaningful for marketing strategies.

Different Ways to Compare Clustering Methods

- When you use different clustering techniques or settings on the same data, you might get different results. To figure out which clustering is better, we can compare them using a few key factors:
- **Single vs. Multilevel Clustering**
  - Single-Level Clustering:
    - Groups all the data at one level, without any hierarchy.
    - Example: Dividing customers into groups where each group is managed separately.
- **Multilevel Clustering:**
  - Creates a hierarchy of clusters, with bigger groups that can be divided into smaller subgroups.
  - Example: Organizing documents into general categories like "sports" and "politics," and then further splitting "sports" into "football," "basketball," etc.

Different Ways to Compare Clustering Methods (cont.)

- **Separation of Clusters**
  - Mutually Exclusive Clusters: Each data point belongs to only one cluster.
  - **Overlapping Clusters**
    - Data points can belong to multiple clusters.
    - Example: A document could relate to both “science” and “technology” topics.
- **Similarity Measure:**
  - How Similarity is Calculated:
    - Methods often use distance between points (like straight-line distance in space) to determine similarity.
    - Different applications may use other measures, like how connected points are in a network.
  - Impact on Clustering:
    - Distance-based methods work well for clusters that are round in shape.
    - Other methods (like density-based) can find clusters of any shape.

Different Ways to Compare Clustering Methods (cont.)

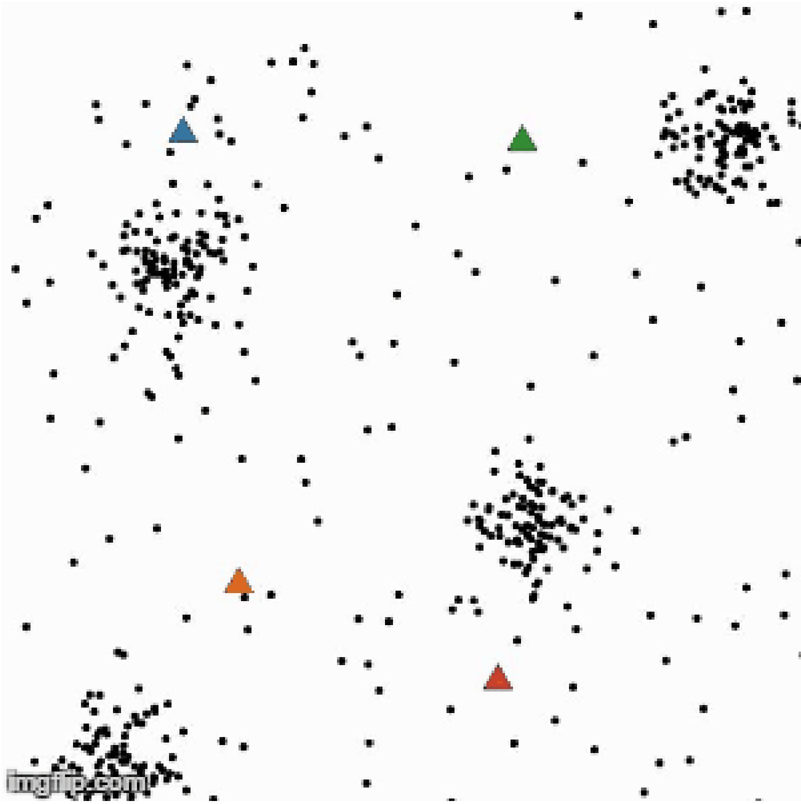
- Clustering in Full Space vs. Subspace
    - Full Space Clustering:
    - Considers all features (dimensions) in the data.
    - Works well for data with only a few features.
- Subspace Clustering:
    - Focuses on certain relevant features or dimensions, ignoring irrelevant ones.
    - Useful for high-dimensional data (lots of features), where some features might not help in finding meaningful clusters.

Overview of Basic Clustering Methods

There are many clustering techniques, and they often share features across different categories. We will explore the main categories.

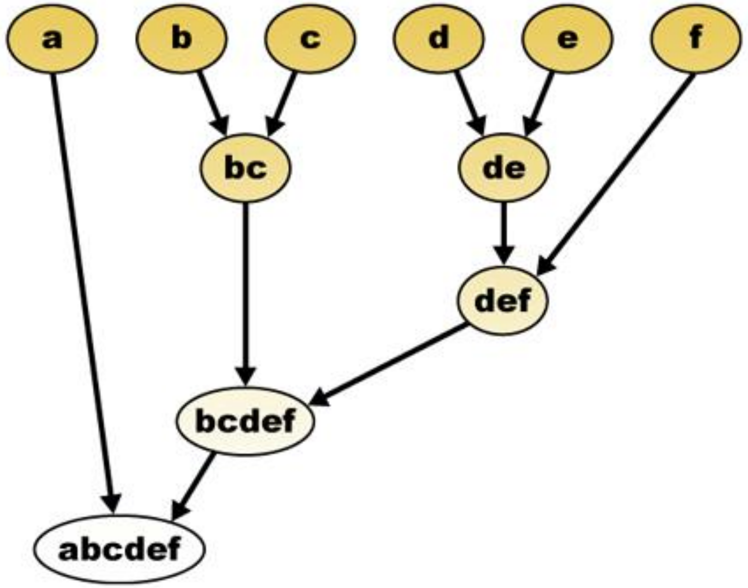
Partitioning Methods

- Partitioning methods divide a dataset of “n” objects into “k” clusters, where “k” is specified by the user and is much smaller than “n”.
  - Each object is assigned to one cluster
- **How They Work:**
  - The method starts by creating an initial partitioning of the data.
  - It then uses an iterative process to improve the partitioning by moving objects between clusters to make the groups more distinct. Visualization
  - The goal is to have objects in the same cluster be similar (close together) and objects in different clusters be different (far apart).
- **Challenges and Limitations:**
  - Achieving the best possible partitioning can be very difficult because it may require evaluating all possible ways to divide the data.
  - Instead, simpler approaches like **k-means** and **k-medoids** are used to get a solution that is “good enough” by improving the partitioning step by step.
  - These methods work well for data with clusters that are roughly spherical in shape, but struggle with clusters that have complex or irregular shapes.



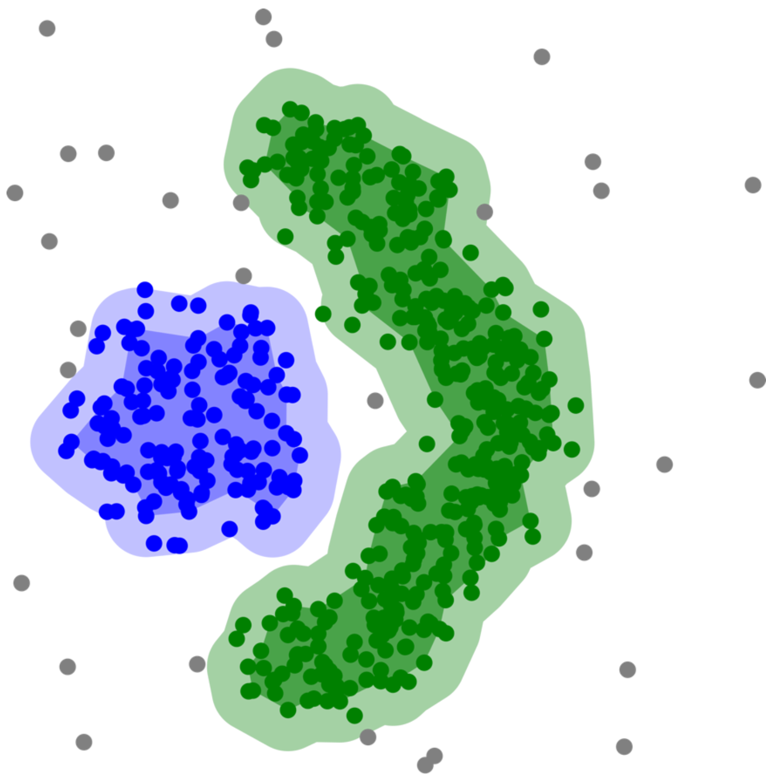
Hierarchical Methods

- Hierarchical methods create a tree-like structure of clusters, either by:
- **Agglomerative (Bottom-Up)** Approach: Start with each data point as its own cluster and merge them step by step until there is one big cluster.
- **Divisive (Top-Down)** Approach: Start with all data points in one big cluster and split them into smaller clusters step by step.
- How They Work:
  - The merging or splitting is based on the similarity (or distance) between clusters.
  - This process can be extended to find clusters in subspaces (specific parts of the data).
- Challenges and Limitations:
  - Once a cluster is merged or split, it cannot be undone, which can lead to errors that cannot be corrected.
  - Despite this, the fixed structure helps reduce computational effort.



Density-Based and Grid-Based Methods

- These methods look for areas in the data where points are densely packed together.
- The idea is to grow a cluster as long as the surrounding density exceeds a certain threshold.
- **Density-Based Clustering:**
    - Groups data points into clusters based on the number of points in a neighborhood.
    - Can find clusters with arbitrary shapes, unlike partitioning methods that work best with spherical shapes.
    - Useful for identifying outliers or noise (points that don't belong to any cluster).
- **Grid-Based Clustering:** Divides the data space into a grid of cells.
    - Clusters are formed based on the density of points in these cells.
    - The advantage is fast processing, as the time depends more on the number of cells rather than the number of data points.
- Integration with Other Methods: Sometimes combined with hierarchical methods to improve performance.
- Blended Approaches: Some clustering algorithms use ideas from multiple methods.



[DBSCAN Visualization]