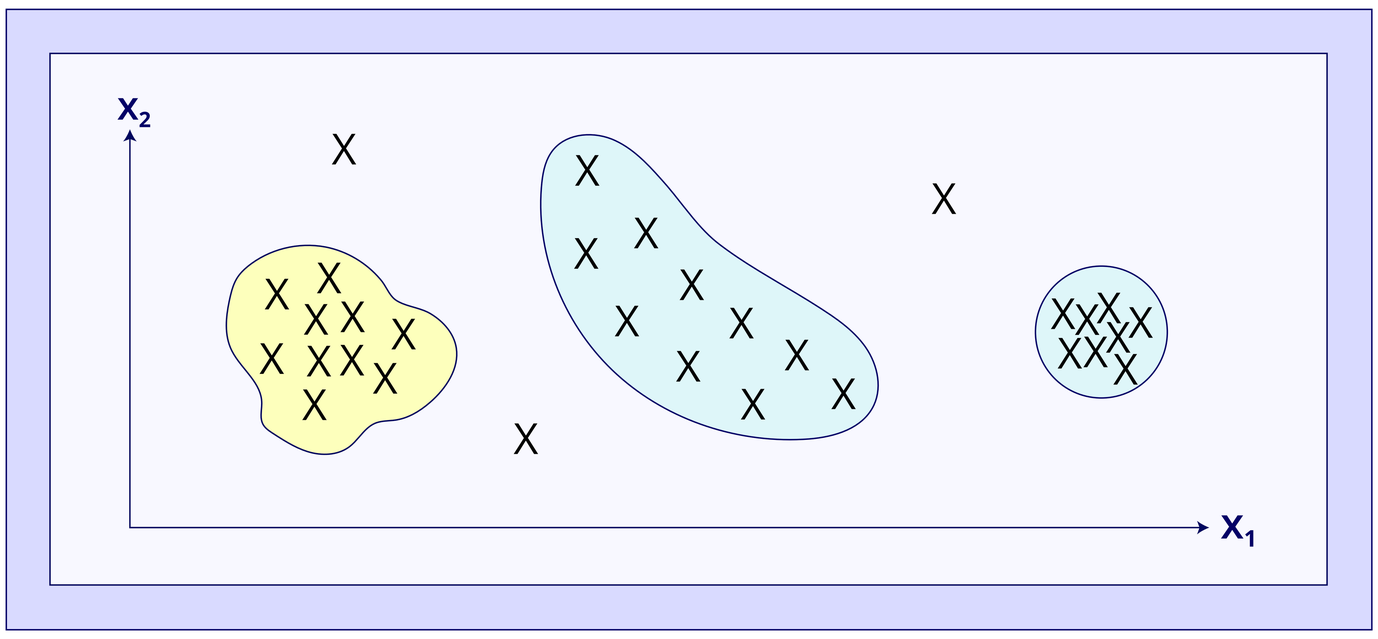
Clustering in Machine Learning

educative

WHAT IS CLUSTERING?

Clustering is used to identify patterns and group similar data points together, making it easier to analyze and understand large datasets.

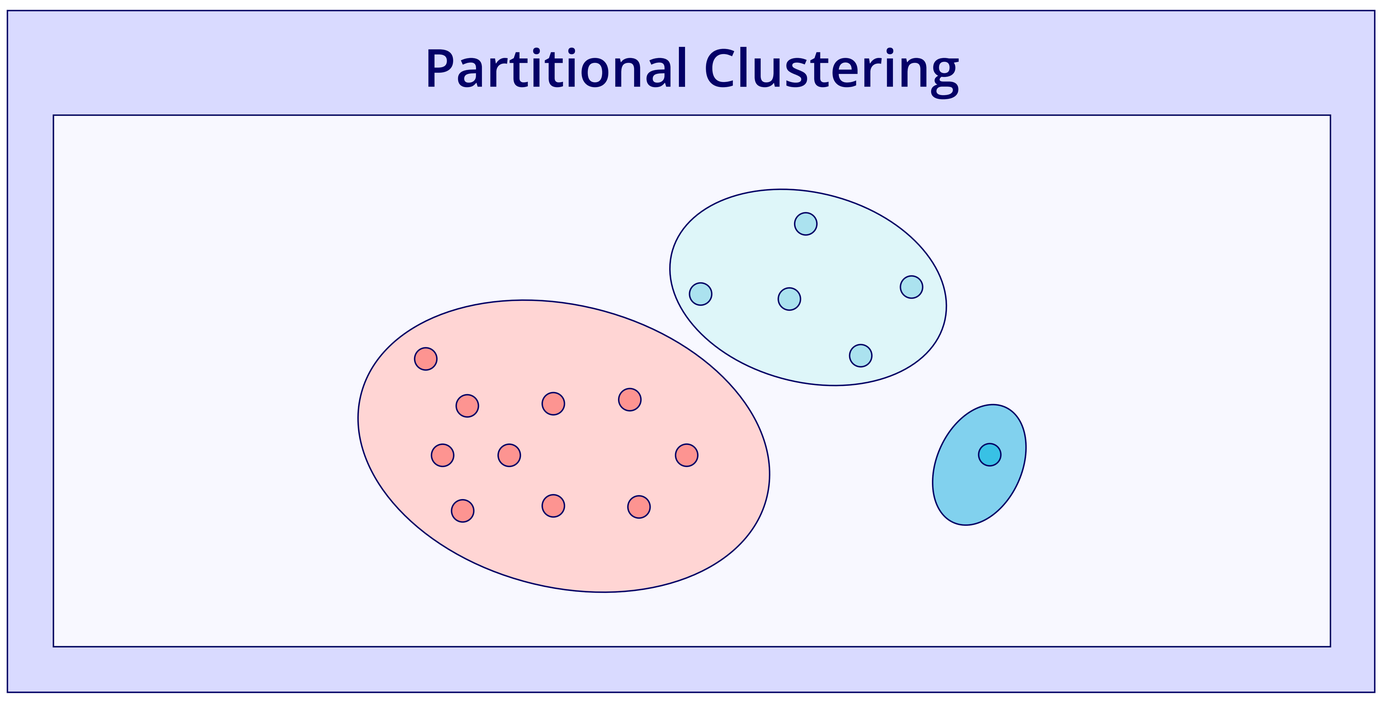
Example: Clustering groups of millions of social media posts by topic to identify trending subjects.



TYPES OF CLUSTERING

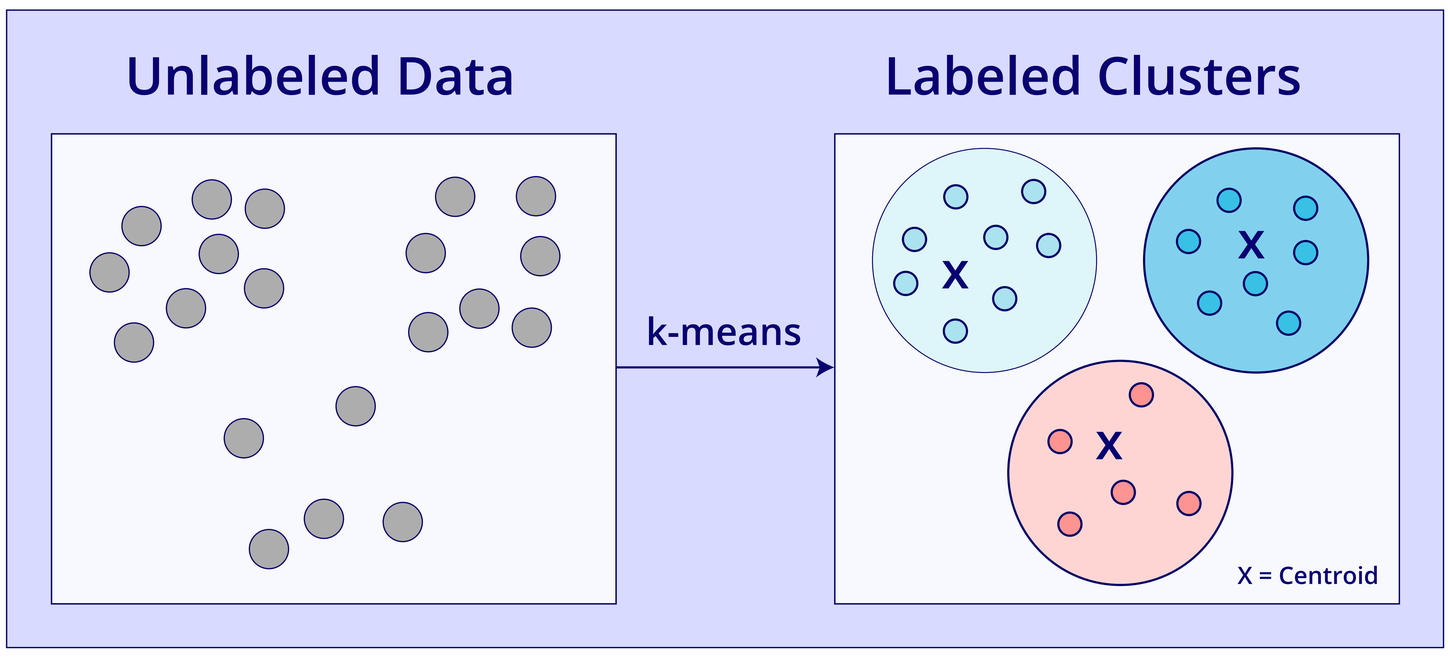
1. Partitional clustering

Partitional clustering is a method of grouping data into non-overlapping subsets based on their similarities, which helps identify patterns or structures in the data for better understanding and analysis.



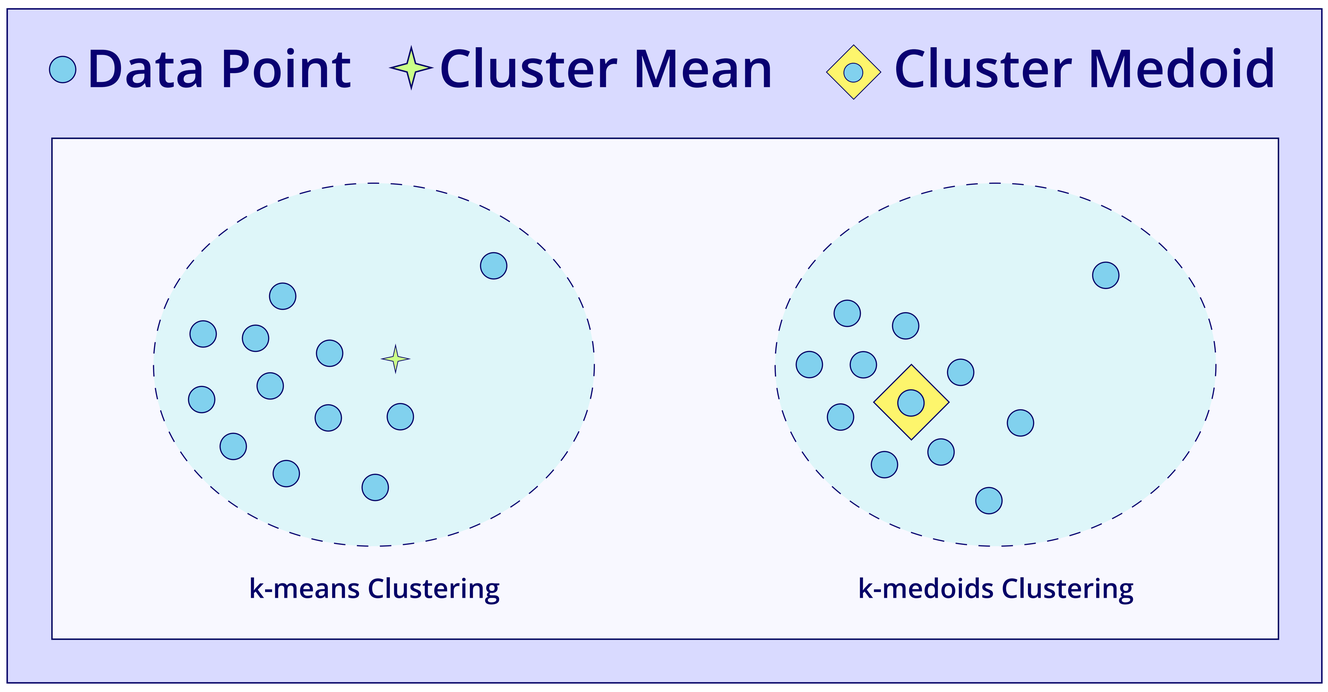
- **K-means clustering**

K-means clustering groups data into *k* clusters by iteratively assigning points to the nearest cluster center and recalculating the centers, helping to uncover patterns in the data.



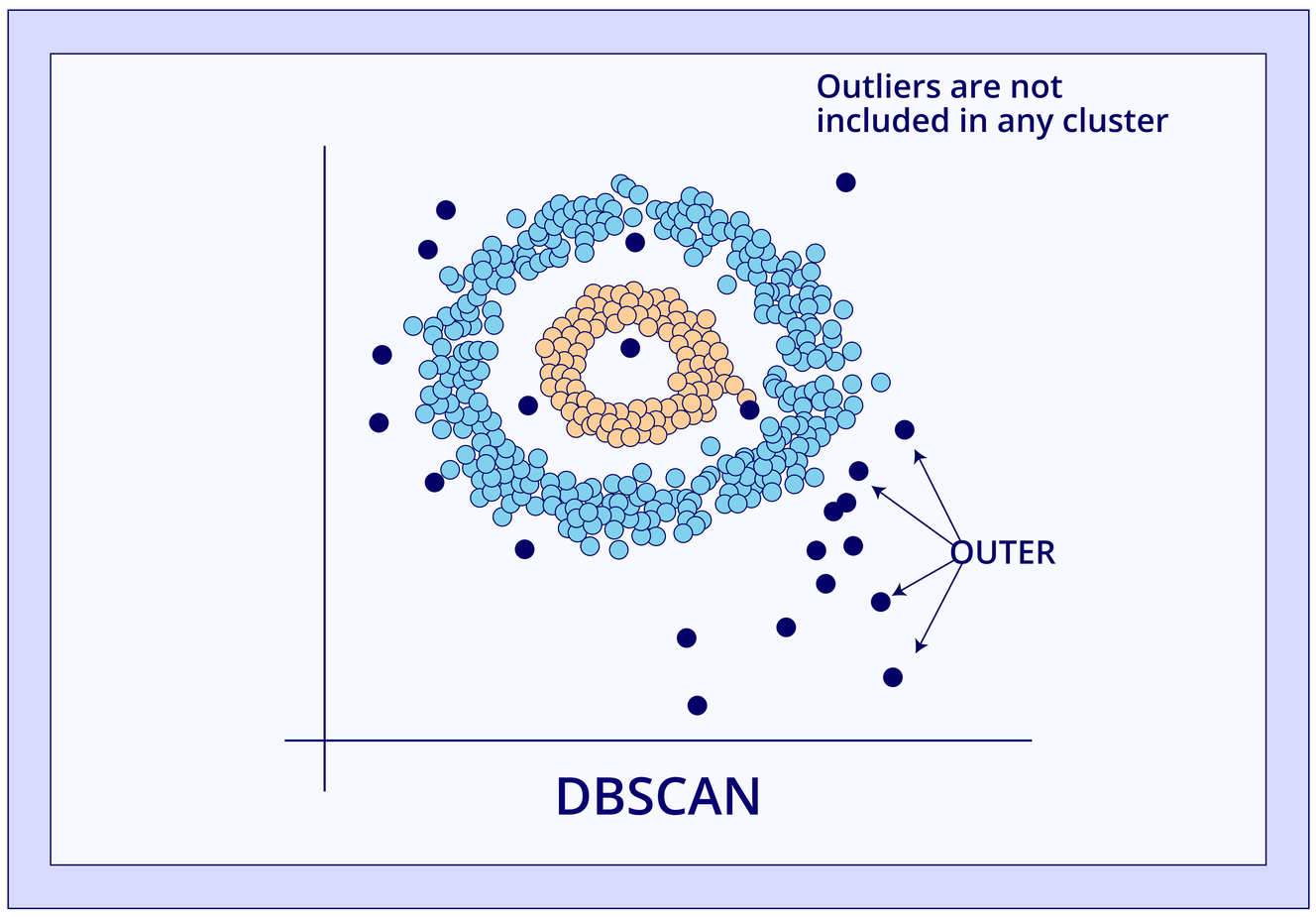
- **K-medoids clustering**

K-medoids clustering groups data into *k* clusters by assigning points to the nearest representative point (medoid) and updating the medoids to minimize total distance, providing a more robust method against outliers compared to K-means clustering.



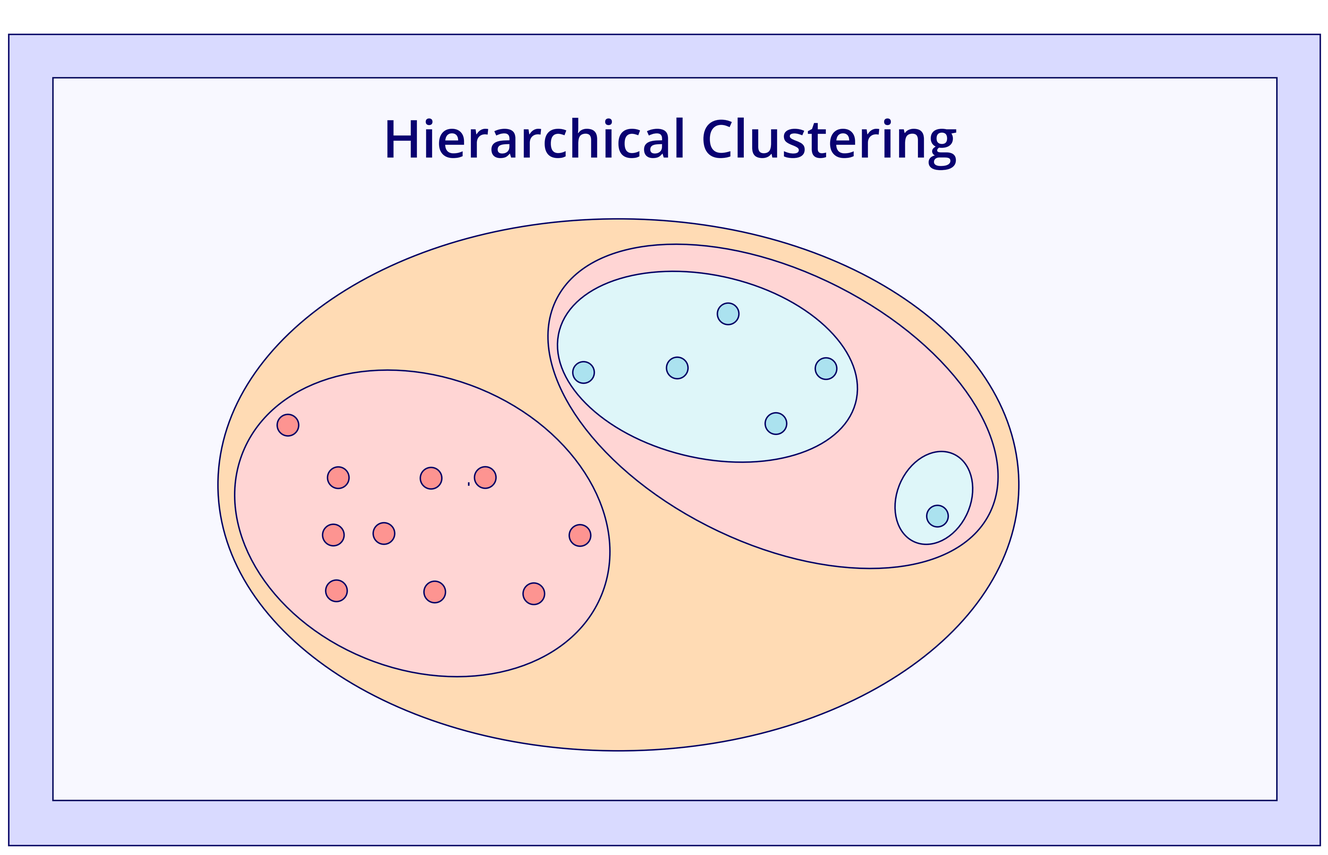
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**

DBSCAN clusters data by identifying dense regions based on a distance threshold, allowing it to effectively find clusters of arbitrary shapes and handle noise.



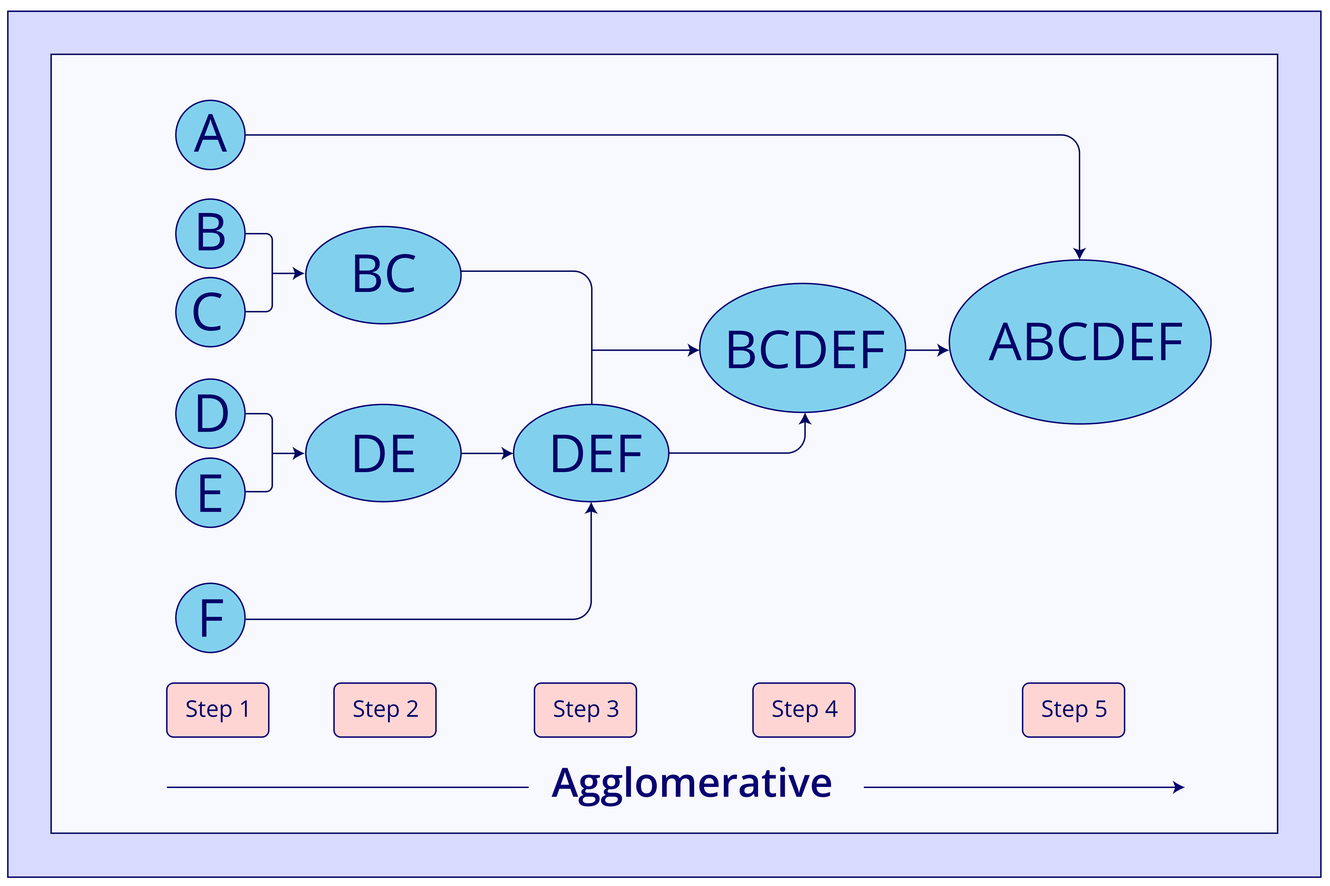
2. Hierarchical clustering

Hierarchical clustering groups data by progressively merging or splitting clusters based on similarity, creating a tree-like structure that reveals the nested relationships in the data.



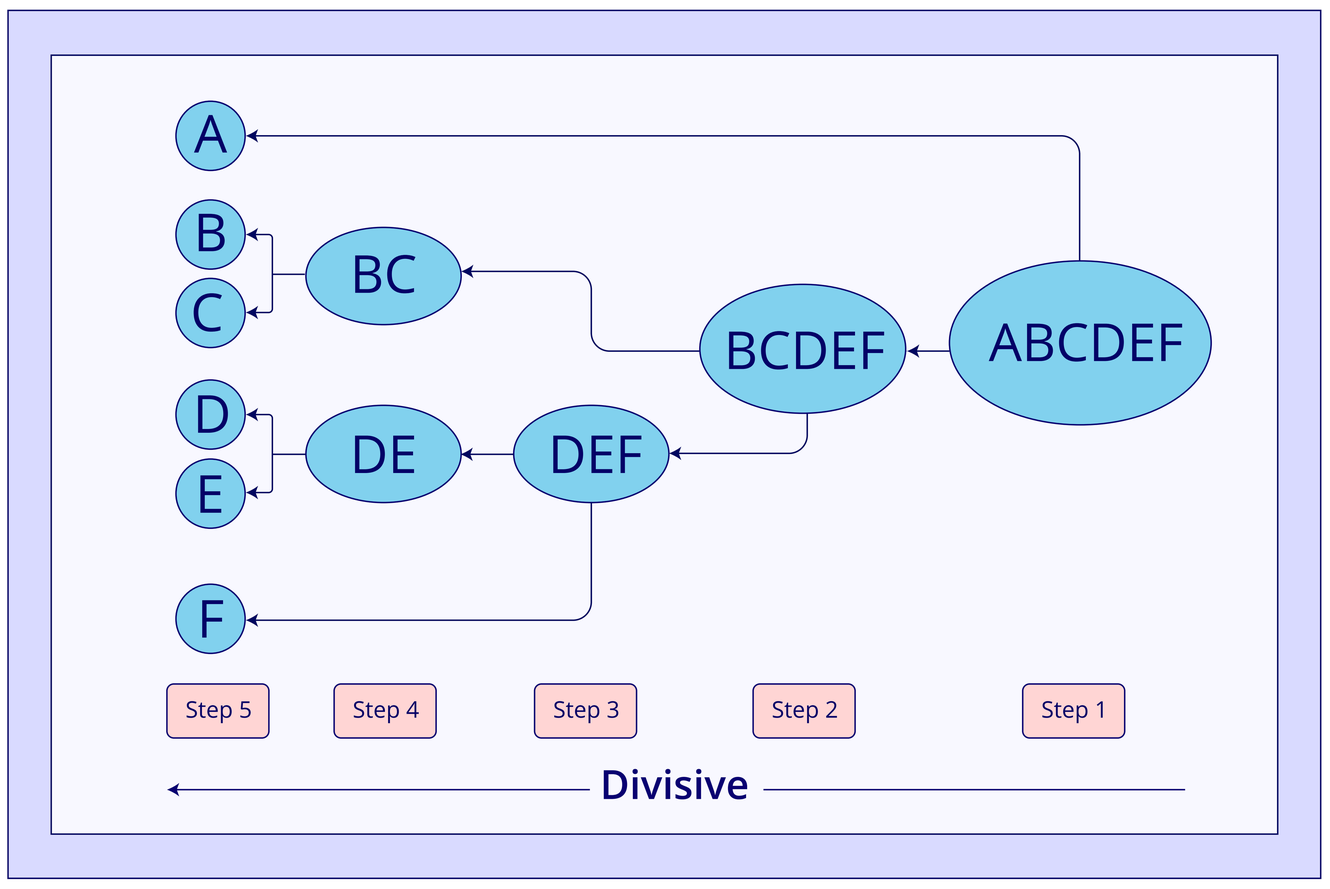
- **Agglomerative clustering algorithms**

Agglomerative clustering starts with each data point as its own cluster and repeatedly merges the closest pairs of clusters until all points are grouped, helping to uncover data hierarchies.



- **Divisive clustering algorithms**

Divisive clustering starts with all data points in a single cluster and repeatedly splits the clusters into smaller groups based on dissimilarities, helping to reveal data structure from a top-down approach.



- **CURE (Clustering Using Representatives)**

CURE clusters data by selecting representative points within each cluster and shrinking them toward the center, which helps in handling large datasets and capturing clusters of arbitrary shapes.