Cluster Analysis:

Cluster analysis is a data mining technique used to partition a dataset into groups or clusters of similar data points. The goal is to identify natural groupings within the data based on similarity or distance measures. Cluster analysis is an unsupervised learning technique, meaning it doesn't require labeled data for training.

Types of Clustering:

Exclusive Clustering: Each data point belongs to exactly one cluster.

Overlapping Clustering: Data points can belong to multiple clusters simultaneously.

Hierarchical Clustering: Clusters are organized in a hierarchical tree-like structure.

Probabilistic Clustering: Clusters are defined based on probability distributions.

Fuzzy Clustering: Data points have degrees of membership in multiple clusters.

Categorization of Major Clustering Methods:

Partitioning Methods: Divide the dataset into non-overlapping clusters.

Hierarchical Methods: Build a hierarchy of clusters.

Density-Based Methods: Group data points based on

density.
Grid-Based Methods: Partition the data space into a finite number of cells.
Model-Based Methods: Assume a probabilistic model for each cluster.
Partitioning Methods:
Partitioning methods aim to partition the dataset into a predetermined number of clusters. Popular partitioning methods include:

K-Means: Iteratively assigns data points to the nearest cluster centroid and updates centroids based on the mean of the assigned points.
K-Medoids (PAM): Similar to K-Means, but uses actual data points (medoids) as cluster representatives.
CLARA (Clustering Large Applications): An extension of K-Medoids designed to handle large datasets by sampling.
Hierarchical Methods:
Hierarchical clustering methods create a tree-like hierarchy of clusters. Two notable methods are:

CURE (Clustering Using Representatives): Divides the dataset into smaller subclusters, then selects representative points to merge similar clusters.

Chameleon: Adapts the similarity metric based on the local density of data points, allowing it to handle clusters of varying shapes and densities.

Density-Based Methods:

Density-based methods group together data points that are closely packed in dense regions. Examples include:

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Identifies clusters as dense regions separated by areas of lower density.

OPTICS (Ordering Points To Identify the Clustering Structure): Generates a reachability plot to identify clusters of varying densities and shapes.

Wave Cluster: Utilizes a wavelet transform to identify clusters in multi-scale data.

CLIQUE:

CLIQUE (CLustering In Quest) is a grid-based clustering algorithm that partitions the dataset into cells and identifies dense regions (cliques) within these cells. It is particularly suitable for high-dimensional datasets.

Current Trends: Text Mining & Web Mining:

Text Mining: Analyzing unstructured text data to extract meaningful insights and patterns. Current trends include:

Sentiment analysis for understanding opinions and emotions expressed in text.
Named entity recognition for identifying entities such as names, locations, and organizations.
Topic modeling for discovering themes or topics within large collections of documents.
Web Mining: Extracting knowledge and patterns from web data. Current trends include:
Web content mining to extract information from web pages.
Web structure mining to analyze the link structure of the web.
Web usage mining to understand user behavior and preferences on the web.