Clustering Techniques

Data Mining & Neural Networks

Dr. Wilmer Garzón

Director, Master's Program in Data Science
Department of Computer Engineering

Escuela Colombiana de Ingeniería
Universidade da Coruña

2025

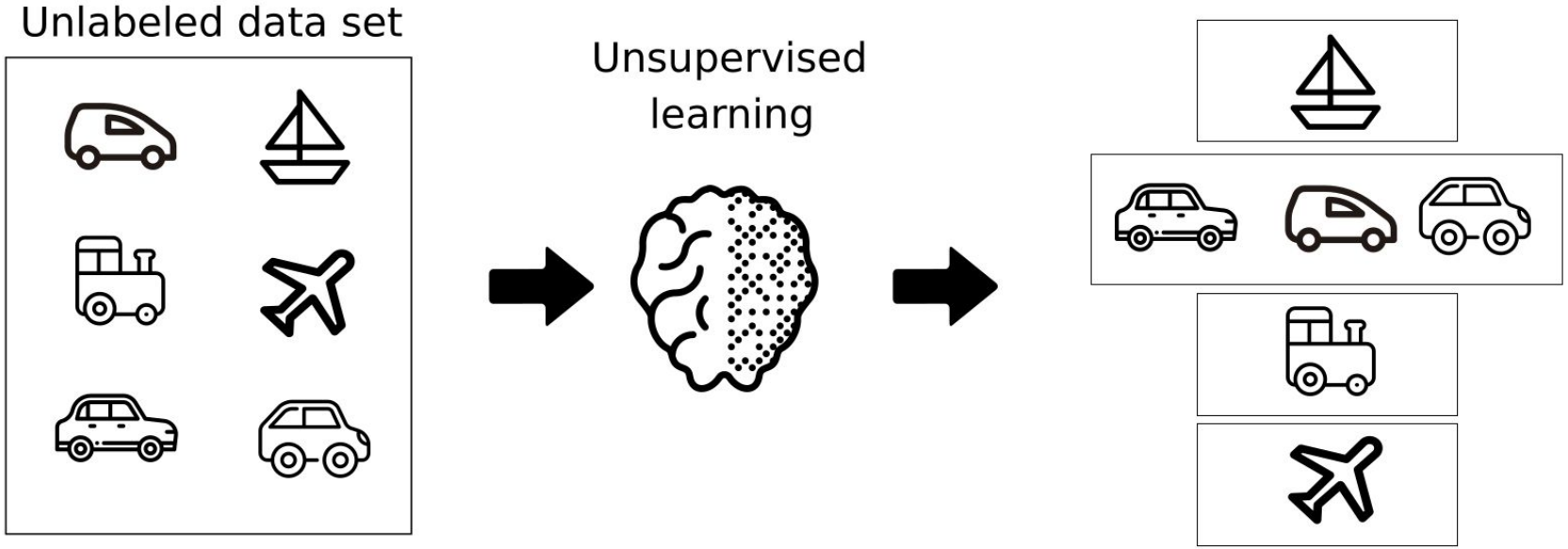
UNIVERSIDAD

Introduction to Clustering

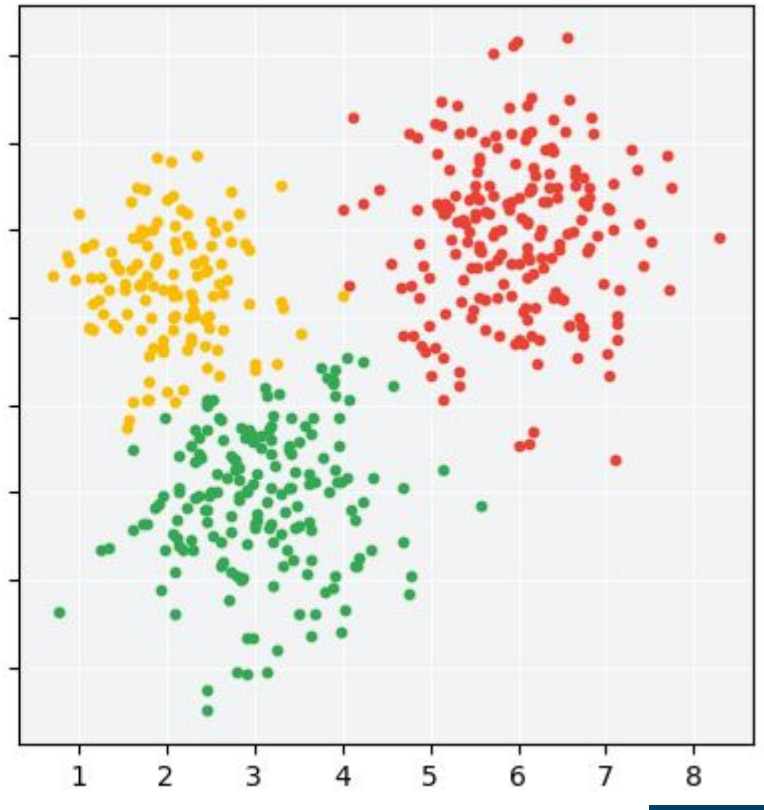
- Clustering is an **unsupervised learning** technique used to group similar data points based on inherent patterns.
- Clustering **does not rely on labeled data**.
- It is widely used in **customer segmentation, image analysis, anomaly detection, and bioinformatics**.
- The goal is to maximize **intra-cluster similarity and minimize inter-cluster similarity**.

Unsupervised Learning

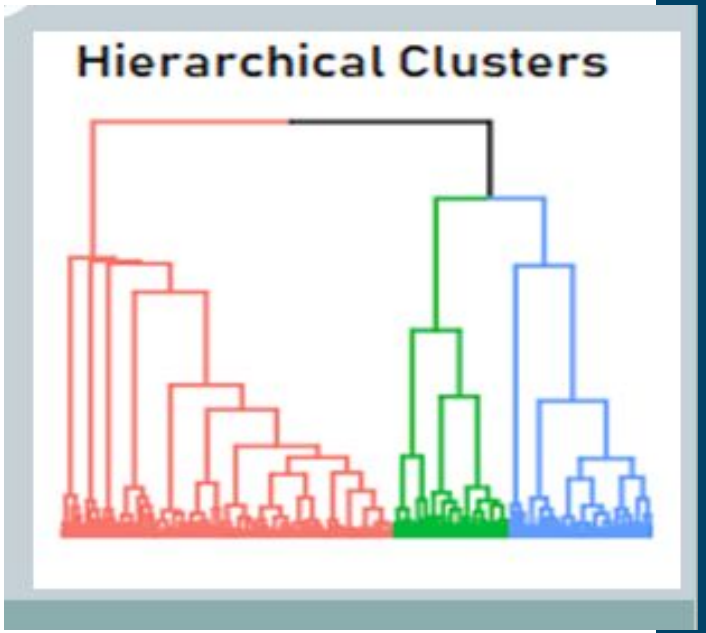
Clustering example



What is Clustering?



- Clustering is the process of **dividing a dataset into groups (clusters)** such that data points in the same group.
- It helps **uncover hidden patterns** in data without prior labels.
- For example: in marketing, clustering can identify customer segments based on purchasing behavior.
- Common similarity measures include **Euclidean distance**, **cosine similarity**, and **Manhattan distance**.



Types of Clustering Techniques

Clustering techniques can be broadly categorized into:

- **Partitioning Methods** (e.g., K-Means)
- **Hierarchical Methods** (e.g., Agglomerative Clustering)
- **Density-Based Methods** (e.g., DBSCAN)
- **Model-Based Methods** (e.g., GaussianMixture Models)

Choosing the right technique depends on the dataset's shape, size, and distribution.

K-Means Clustering

K-Means Clustering

- It is a partitioning method that divides data into **k** clusters by minimizing the within-cluster sum of squares.

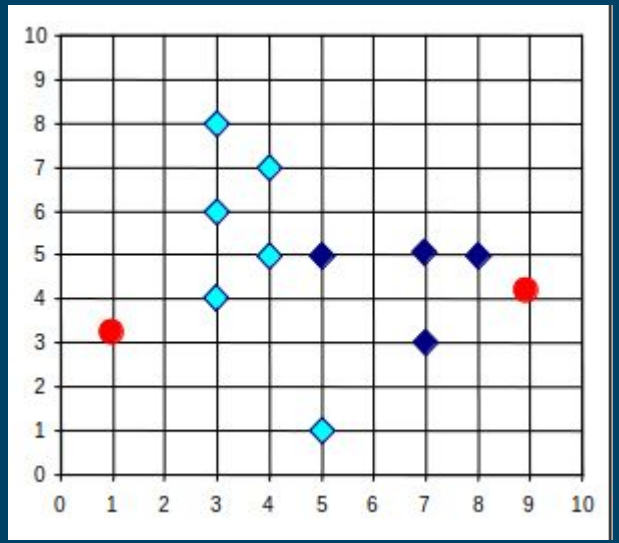
Step-by-Step:

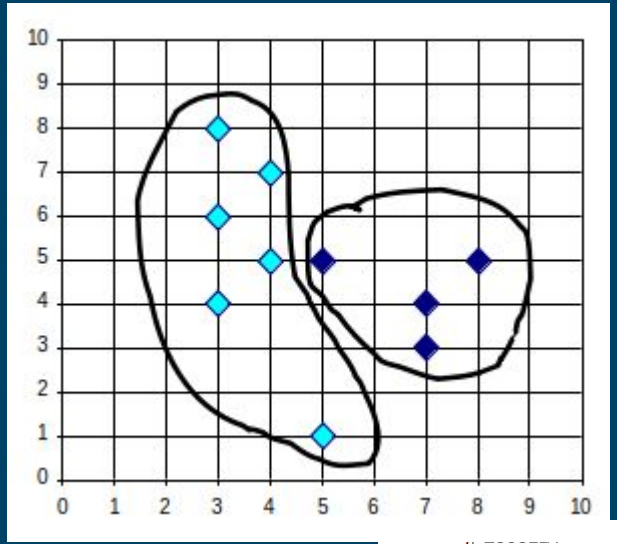
1. Choose the number of clusters (K)
How many clusters you want to form.

2. Initialize K centroids
Randomly select K data points as the initial centroids.

3. Assign points to nearest centroid
For each data point, assign it to the cluster with the closest centroid (based on distance, usually Euclidean).

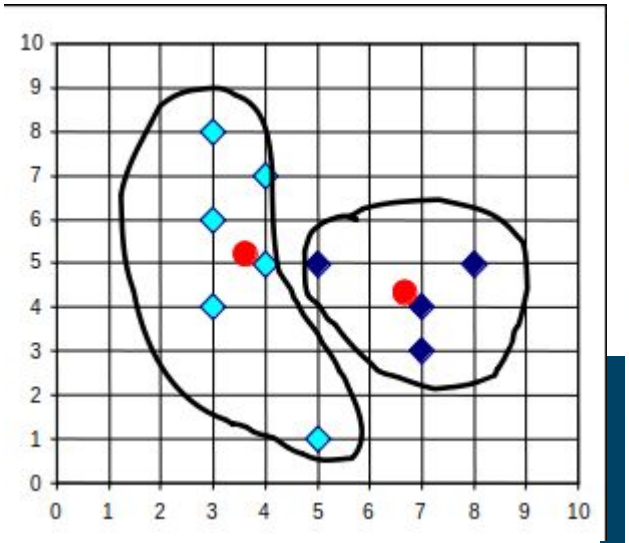
4. Update centroids
Recalculate the centroid of each cluster by taking the mean of all points assigned to it.

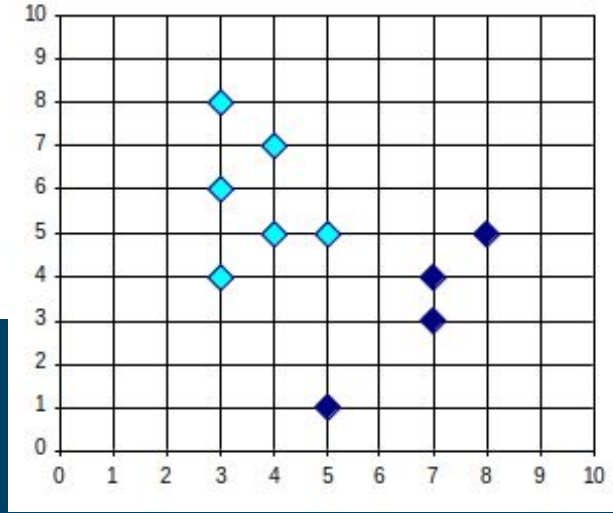


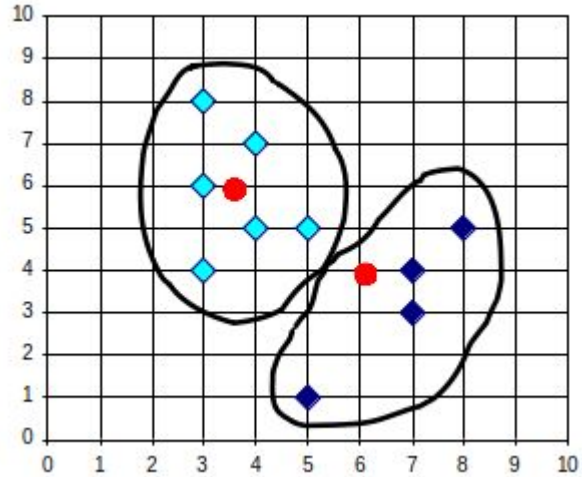


K-Means Clustering

- **Repeat steps 3 and 4**
- Continue reassigning points and updating centroids until the assignments no longer change (convergence) or a maximum number of iterations is reached.
- Output the final clusters: Return the K clusters and their centroids.

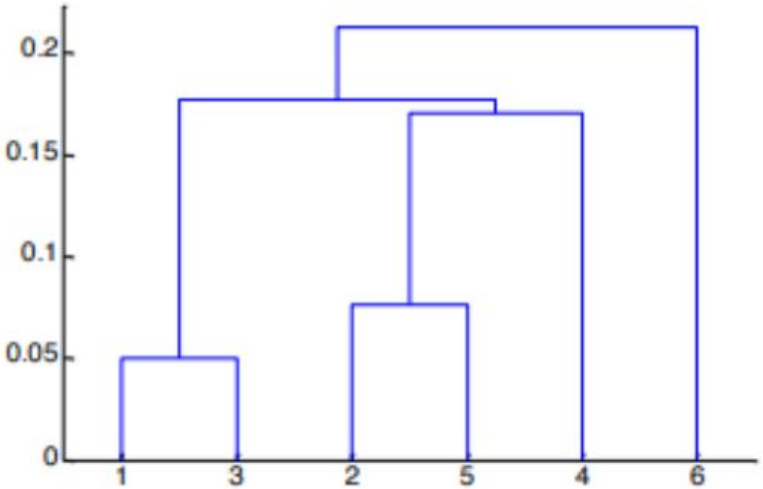


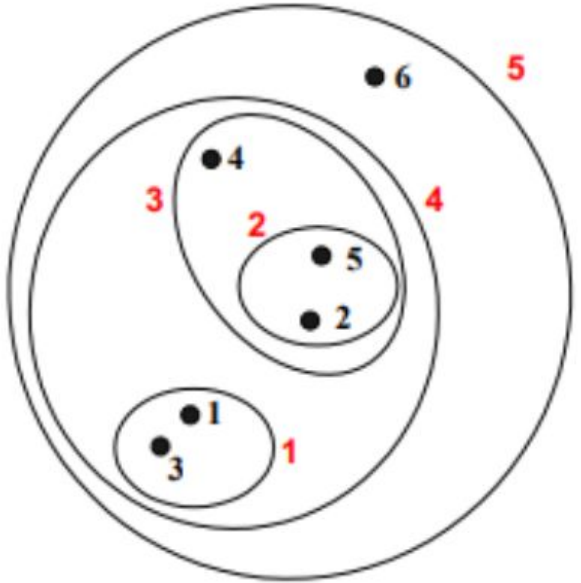




Hierarchical Clustering

Hierarchical clustering builds a tree-like structure (dendrogram) of nested clusters.





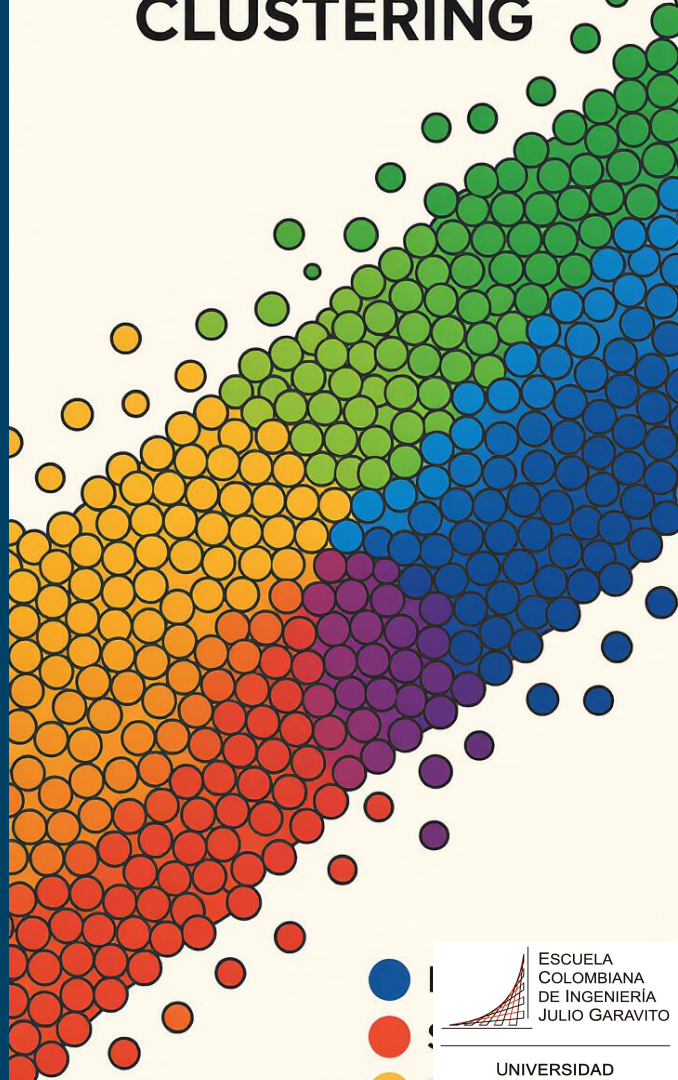
Other Clustering Techniques:

DBSCAN (Density-Based Spatial Clustering)

- Identifying clusters of GPS coordinates in urban planning.

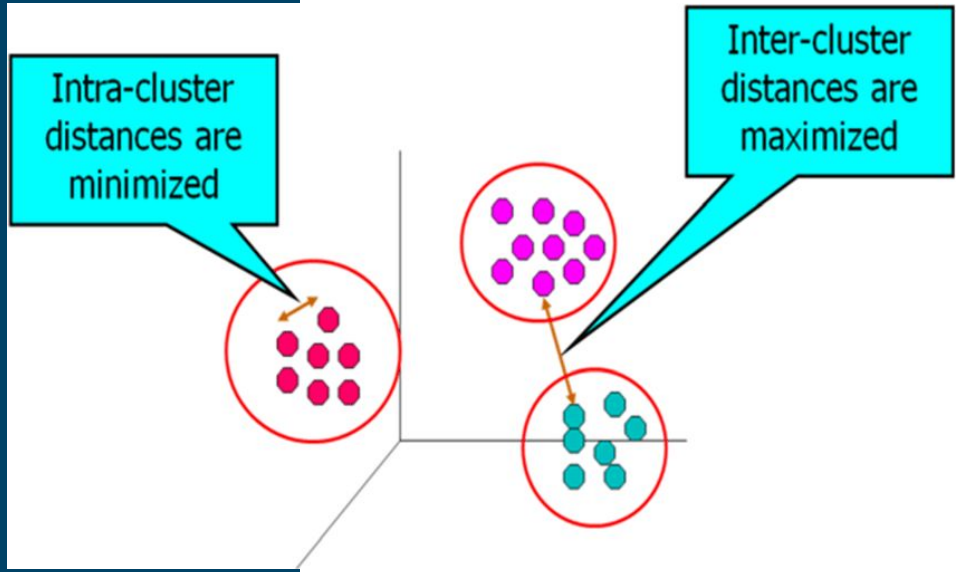
Gaussian Mixture Models (GMM)

- It is suitable for elliptical clusters and overlapping data.



Evaluation Metrics for Clustering

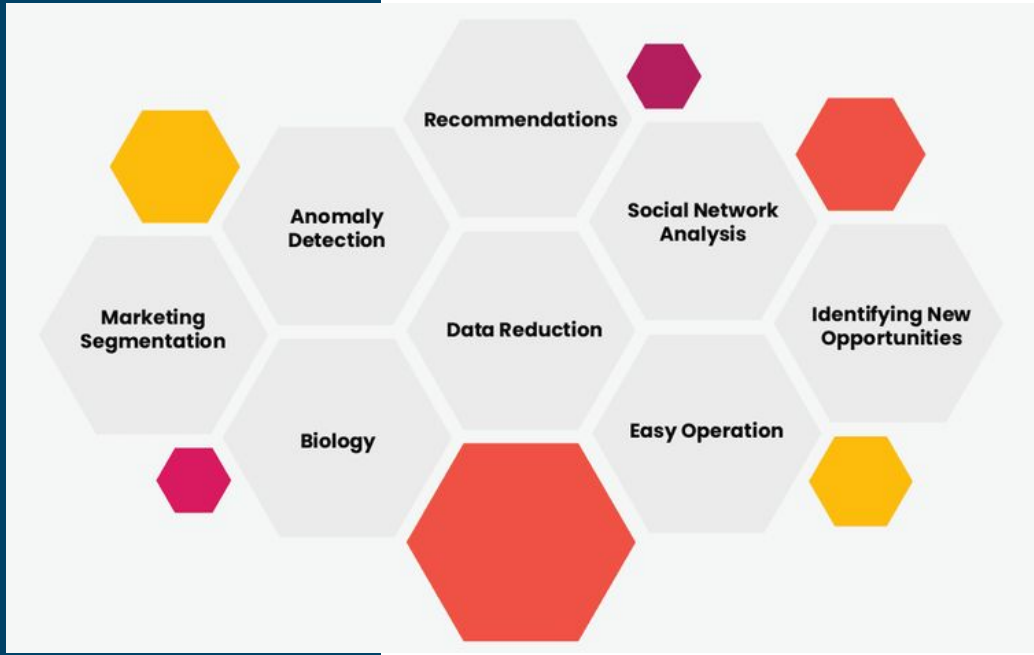
- **Silhouette Score**: Measures how similar a point is to its own cluster vs. others.
- **Davies-Bouldin Index**: Lower values indicate better clustering.
- **Adjusted Rand Index (ARI)**: Compares clustering with ground truth (if available).
- **Elbow Method**: Helps determine optimal k in K-Means.
- **Visual inspection** using scatter plots or dendrograms is also useful.



Applications of Clustering

Clustering is used in various domains:

- **Marketing**: Customer segmentation
- **Healthcare**: Disease subtype identification
- **Finance**: Fraud detection
- **Image Processing**: Object recognition
- **Social Networks**: Community detection



Challenges in Clustering

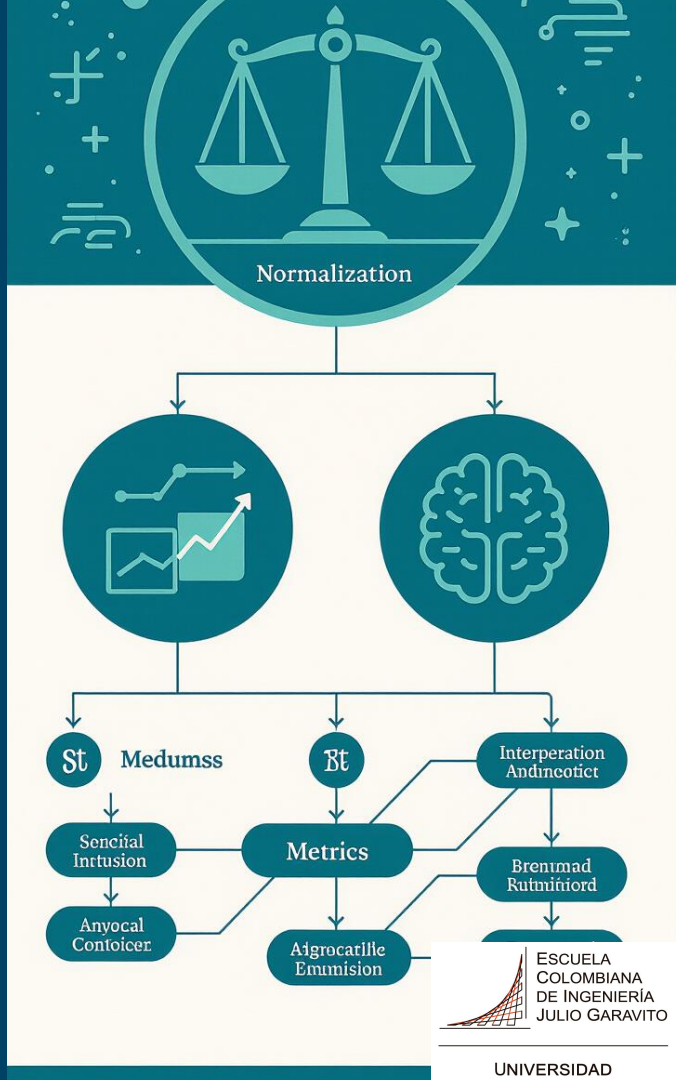
Key challenges include:

- Choosing the **right number of clusters**
- **Handling high-dimensional data**
- **Dealing with noise and outliers**
- **Scalability to large datasets**
- **Interpreting clusters meaningfully**

Summary and Best Practices

- Understand your data before choosing a clustering method.
- Normalize features to avoid bias due to scale.
- Use multiple metrics and visualizations to evaluate results.
- Combine clustering with domain knowledge for interpretation.
- Experiment with different algorithms and parameters.

Clustering is a powerful tool when used thoughtfully and iteratively.



INTER NATIONAL SUMMER SCHOOL
UNIVERSIDADE DA CORUÑA

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

UNIVERSIDAD