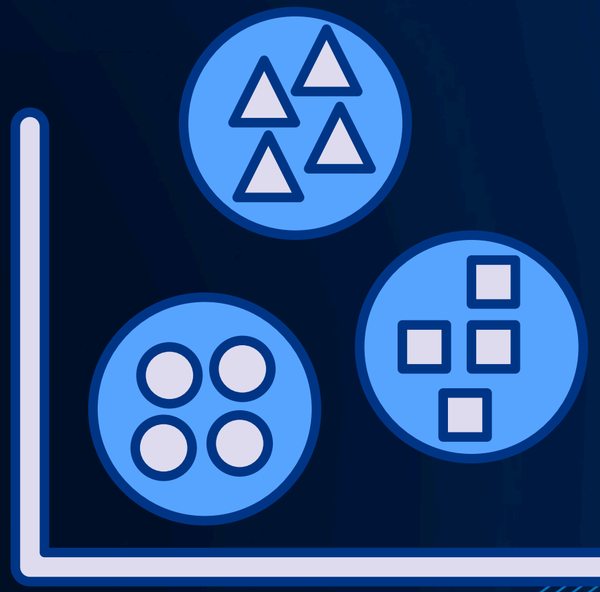
Understanding K-Means Clustering

By Luiza Stepanyan



ABI

What is Clustering?

- Clustering is an unsupervised learning technique.
- Goal: Group similar data points together based on some similarity metric.
- No predefined labels — the algorithm learns patterns and forms natural groupings.
- Real-world uses: customer segmentation, image compression, anomaly detection, etc.

Types of Clustering Algorithms

- K-Means Clustering (Focus of this presentation)
- Hierarchical Clustering: Builds nested clusters in a tree-like structure (dendrogram).
- DBSCAN: Density-based clustering that can find arbitrarily shaped clusters.
- Gaussian Mixture Models (GMM): Probabilistic model assuming data points are from multiple Gaussian distributions.
- Note: Each method has strengths/limitations based on data shape, scale, and noise.

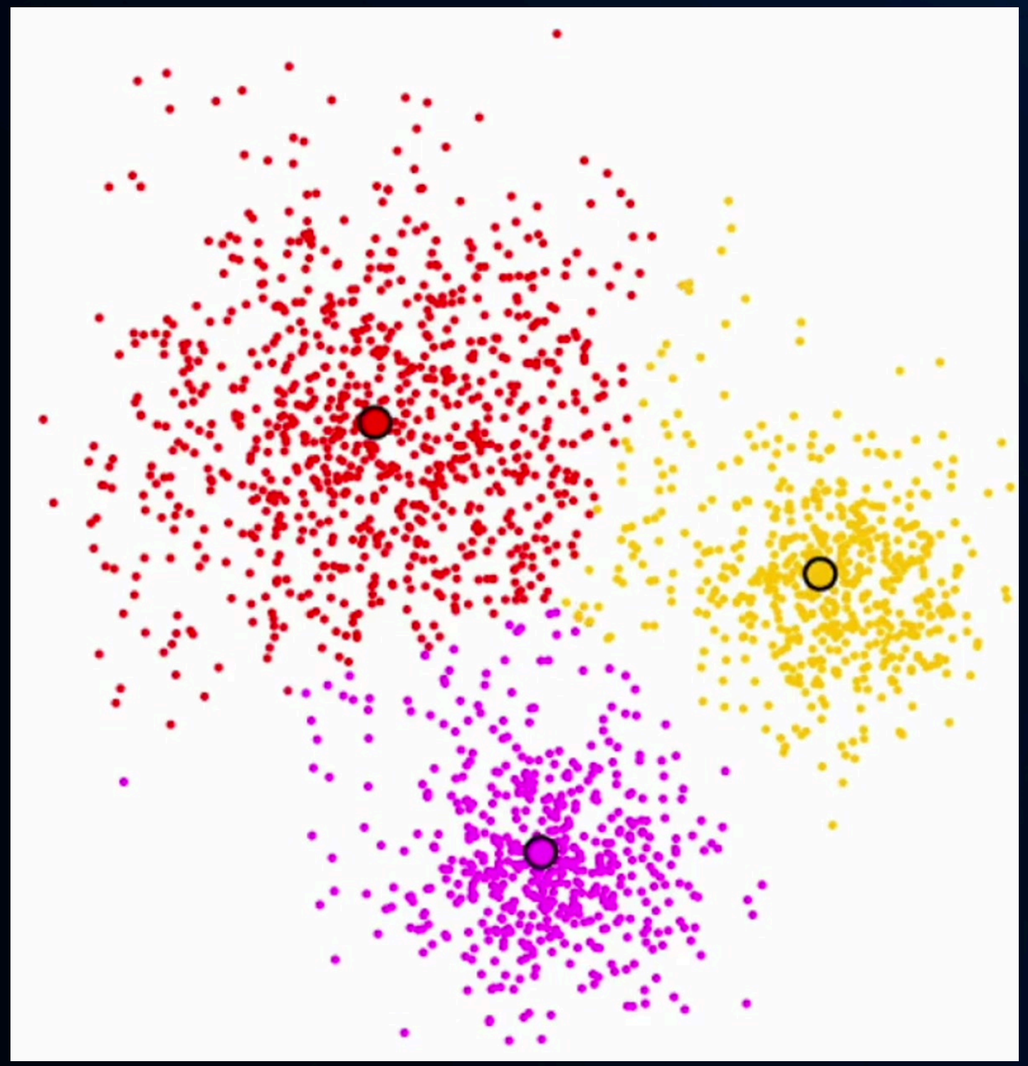
What is K-Means Clustering?

Objective: Partition n data points into k clusters such that each point belongs to the cluster with the nearest mean (centroid).

- **It minimizes intra-cluster variance (i.e., within-cluster sum of squares).**
- **Assumes clusters are spherical and similar in size.**

Step-by-Step: K-Means Algorithm

1. Choose k (number of clusters).
2. Randomly initialize k centroids.
3. Assign each point to the nearest centroid → Form clusters.
4. Recalculate centroids by taking the mean of all points in each cluster.
5. Repeat steps 3–4 until:
    - Centroids do not change (convergence).
    - Or a maximum number of iterations is reached.



Visual Representation

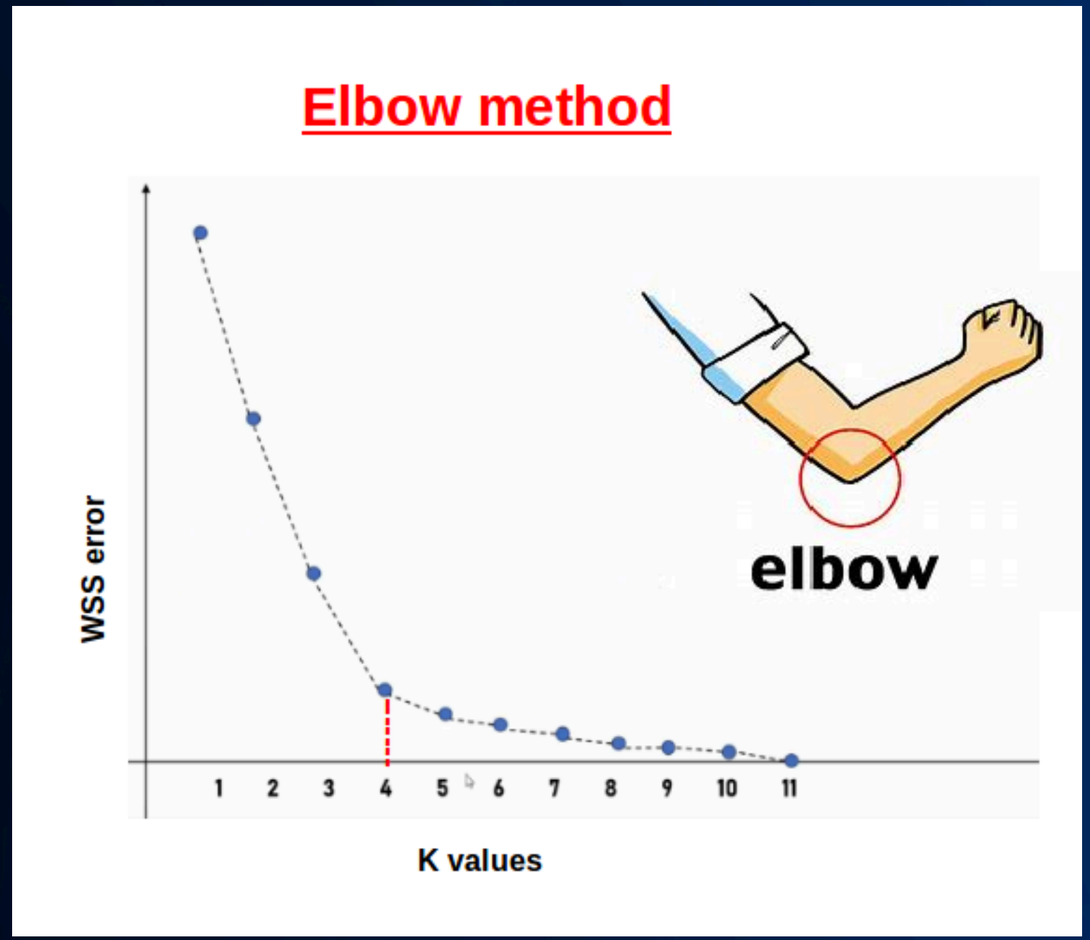
How to Choose the Number of Clusters (k)?

Elbow Method:

- Plot Sum of Squared Errors (SSE) vs. k.
- "Elbow" point = ideal number of clusters.

Silhouette Score: Measures cohesion & separation of clusters.

Trial and error, domain knowledge, or more advanced techniques (e.g., Gap Statistic).



The K-Means Optimization Goal

Objective: Minimize the within-cluster sum of squares:

$$\sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

Where:

- Ci: cluster i
- μi: centroid of cluster i
- x: data point

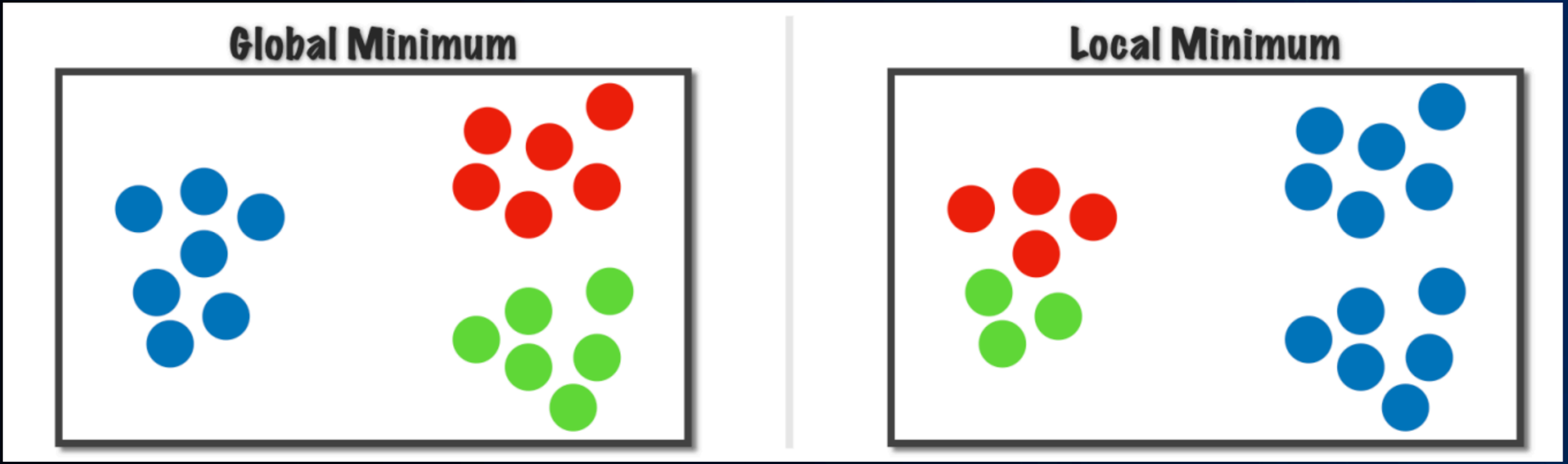
Why Use K-Means?

- Simple and fast.
- Works well on large datasets.
- Easy to implement and interpret.
- Efficient on spherical, equally-sized clusters.

K-Means be like:

When K-Means Falls Short

- Needs pre-specification of k.
- Not suitable for non-spherical or overlapping clusters.
- Sensitive to outliers and initial centroid positions.
- May converge to local minima.
-



Variants & Improvements

- K-Means++: Better centroid initialization for faster convergence.
- Mini-Batch K-Means: For large datasets – faster & scalable.
- Multiple Runs: Use different initializations and pick the best.
- Combine with PCA for dimensionality reduction before clustering.

K-Means in the Real World

- Healthcare:
  - Grouping patients based on symptoms/genomics.
- Marketing:
  - Customer segmentation.
- Computer Vision:
  - Image segmentation, color quantization.
- Finance:
  - Fraud detection, market segmentation.

Takeaways

- K-Means is a powerful, simple clustering tool for unsupervised learning.
- Works best when clusters are well-separated and spherical.
- Careful initialization and evaluation are key to performance.
- Always explore your data and try multiple clustering approaches when necessary.

Thank You So Much!