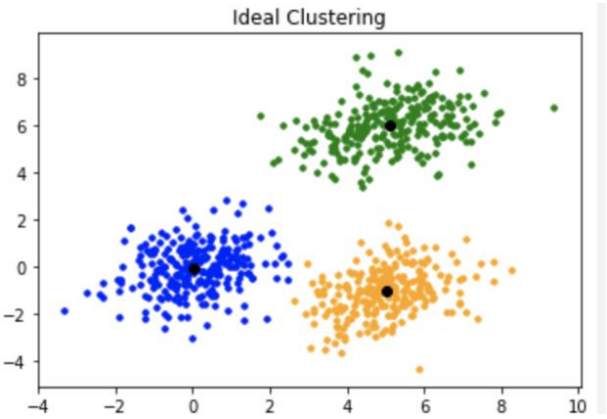
# K-Means Clustering

K-Means is an unsupervised machine learning algorithm used to group similar data points into K distinct clusters based on feature similarity.

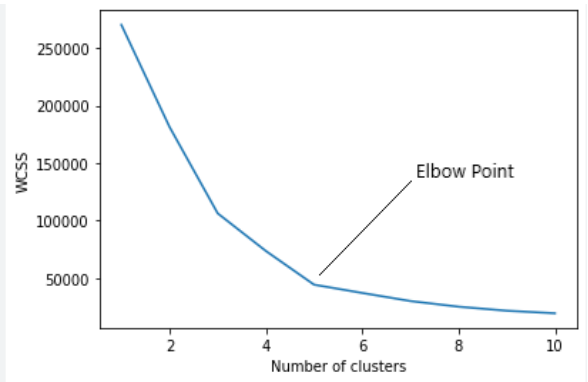
| **How K-Means Works:** |
|---|

| |
|---|
| 1. **Choose K** (number of clusters). |
| 2. Randomly **initialize K centroids**. |
| 3. Assign each data point to the **nearest centroid** (forming K clusters). |
| 4. **Recompute centroids** as the **mean** of all points in each cluster. |
| 5. **Repeat steps 3 and 4** until centroids no longer move significantly (convergence). |




## Determining the Optimal K (Number of Clusters):

**Elbow Method (Using WCSS):** The Elbow Method is a technique used to determine the optimal number of clusters (K) in K-Means clustering.
It does this by measuring how the Within-Cluster Sum of Squares (WCSS) changes with varying values of K.




- **WCSS (Within-Cluster Sum of Squares):**

**It** measures the **total squared distance between each point and the centroid of its cluster**.

- For each K, compute WCSS.
- Plot **K vs. WCSS**.
- The point where the **rate of decrease sharply slows down** forms an **"elbow"** – this is the optimal K.

**Goal:** Minimize WCSS but avoid overfitting with too many clusters.

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

Where:

- $K$: Number of clusters
- $C_i$: Cluster i
- $\mu_i$: Centroid of cluster i
- $x$: Data points in cluster $C_i$

# K-Means++:

K-Means++ is an improved initialization method for K-Means clustering that selects initial centroids in a smarter way to reduce the chances of poor clustering and speed up convergence.

**Steps:**

1. Randomly choose the **first centroid** from the data points.
2. For each remaining point, compute the **distance to the nearest centroid already chosen**.
3. Choose the next centroid with a **probability proportional to the square of this distance**.
4. Repeat until **K centroids** are chosen.

**Benefit:** Avoids poor clustering that can result from random initialization in standard K-Means.

# Silhouette Score

**Silhouette Score** is a metric used to evaluate the **quality of clustering**. It quantifies how well a data point fits within its own cluster compared to other clusters.

$$\text{Silhouette Score} = \frac{b - a}{\max(a, b)}$$

Where:

- aa = average distance of the point to all other points in the **same cluster** (intra-cluster distance)
- bb = average distance of the point to all points in the **nearest neighbouring cluster**

**Interpretation:**

- **Close to 1** → good clustering
- **~ 0** → point is on the cluster boundary
- **Negative** → point is assigned to the wrong cluster

**Use:** Helps choose the optimal number of clusters **K**.