K-Means Clustering: Objective and Core Concepts

* **Objective:** The fundamental goal of K-Means is to group similar data points together based on their features and discover underlying patterns within the unlabeled data. To achieve this, K-Means partitions the dataset into a **fixed number (K)** of clusters.
* **Cluster:** A cluster is simply a collection of data points that have been aggregated together because they exhibit certain **similarities** (usually based on proximity in the feature space).
* **Centroid:** For K-Means to work, we first define a target number **'K'**, which specifies the number of clusters we want to find. Each cluster is represented by its center point, called the **centroid**. The centroid can be thought of as the location representing the "average" data point within that cluster. Initially, centroids can be actual data points or random points in the feature space.
* **Core Mechanism:** The K-Means algorithm essentially works by:
  1. Identifying **K** number of centroids (initially chosen randomly or strategically).
  2. Allocating **every data point** in the dataset to its **nearest centroid/cluster**.
  3. Updating the position of each centroid based on the points assigned to it.
  4. Repeating steps 2 and 3 until the cluster assignments stabilize.
* **The "Means" in K-Means:** The name "K-Means" comes from how the centroids are updated. The new position of each centroid is calculated as the **mean** (average) of all the data points currently assigned to that cluster.

K-Means Clustering: Basis and Steps (The Algorithm)

K-Means attempts to separate data samples into 'K' distinct groups while trying to ensure that the variation *within* each group (cluster) is as small as possible.

Optimization Goal: Minimizing Inertia

The algorithm aims to minimize a criterion known as **Inertia** or **Within-Cluster Sum-of-Squares (WCSS)**.

* **WCSS Definition:** This is the sum of the squared distances between each data point (xᵢ) and the centroid (μⱼ) of the cluster (Cⱼ) it belongs to, summed over all data points.
* Inertia (WCSS) = Σ [ Σ [ ||xᵢ - μⱼ||² ] ]
* for j=1 to K for xᵢ in Cⱼ

Essentially, we want to find K centroid positions (μⱼ) that make the data points within each cluster as close as possible to their respective centroid.

K-Means Algorithm Steps:

The algorithm proceeds iteratively as follows:

1. **Determine K:** Choose the desired number of clusters, 'K'. This is a crucial parameter that you need to specify beforehand.
2. **Initialize Centroids:** Randomly select 'K' distinct data points from the dataset to serve as the initial centroids, or place 'K' random points within the data space.
3. **Measure Distance:** For *each* data point in the dataset, calculate the distance between that point and *each* of the K centroids. The most common distance metric used is the **Euclidean Distance**.
   * **Euclidean Distance:** For two points p = (p₁, p₂, ..., p<0xE2><0x82><0x99>) and q = (q₁, q₂, ..., q<0xE2><0x82><0x99>) in an n-dimensional space:
   * d(p, q) = sqrt[ Σ [ (qᵢ - pᵢ)² ] ] for i = 1 to n

*Example:* Distance between p = (3, 4) and q = (7, 8) in 2D: d(p,q) = sqrt[ (7-3)² + (8-4)² ] = sqrt[ 4² + 4² ] = sqrt[ 16 + 16 ] = sqrt[32] ≈ 5.66

1. **Assign Points to Clusters:** Assign each data point to the cluster whose centroid is **nearest** (i.e., has the minimum distance calculated in Step 3).
2. **Update Centroids:** Recalculate the position of each of the K centroids. The new centroid position is the **mean** (average) of all the data points assigned to that cluster in Step 4.
3. **Repeat:** Repeat **Steps 3-5** (Measure Distance, Assign Points, Update Centroids) until a stopping criterion is met. Common criteria include:
   * The centroids no longer move significantly between iterations.
   * The data points no longer change cluster assignments.
   * A maximum number of iterations is reached.

Once the algorithm converges (stops repeating), the final positions of the centroids and the final cluster assignments for each data point represent the result of the K-Means clustering.