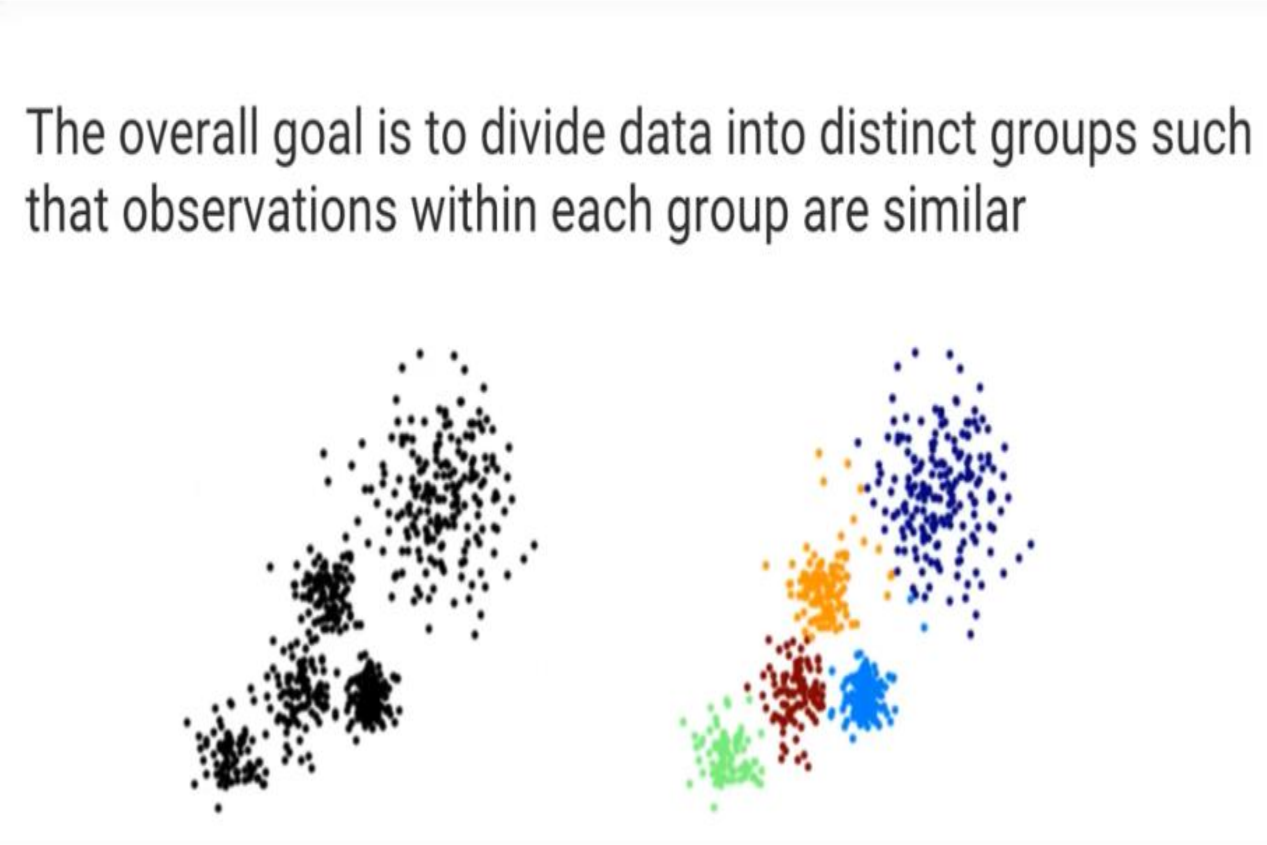
K MEANS Clustering

- Unsupervised Learning Algorithm
- Attempts to group similar clusters in your data
- A typical Clustering Problem will
- Cluster similar documents
- Cluster customers based on features/behaviour
- Market segmentation
- Product segmentation
- Identify similar physical groups

When to Go for K-Means Clustering

- When your data is not labelled (no headers)
- When you know what your target classes are going to be. For example if have some data, without labels but you know you want only 2or specific classes as outcome



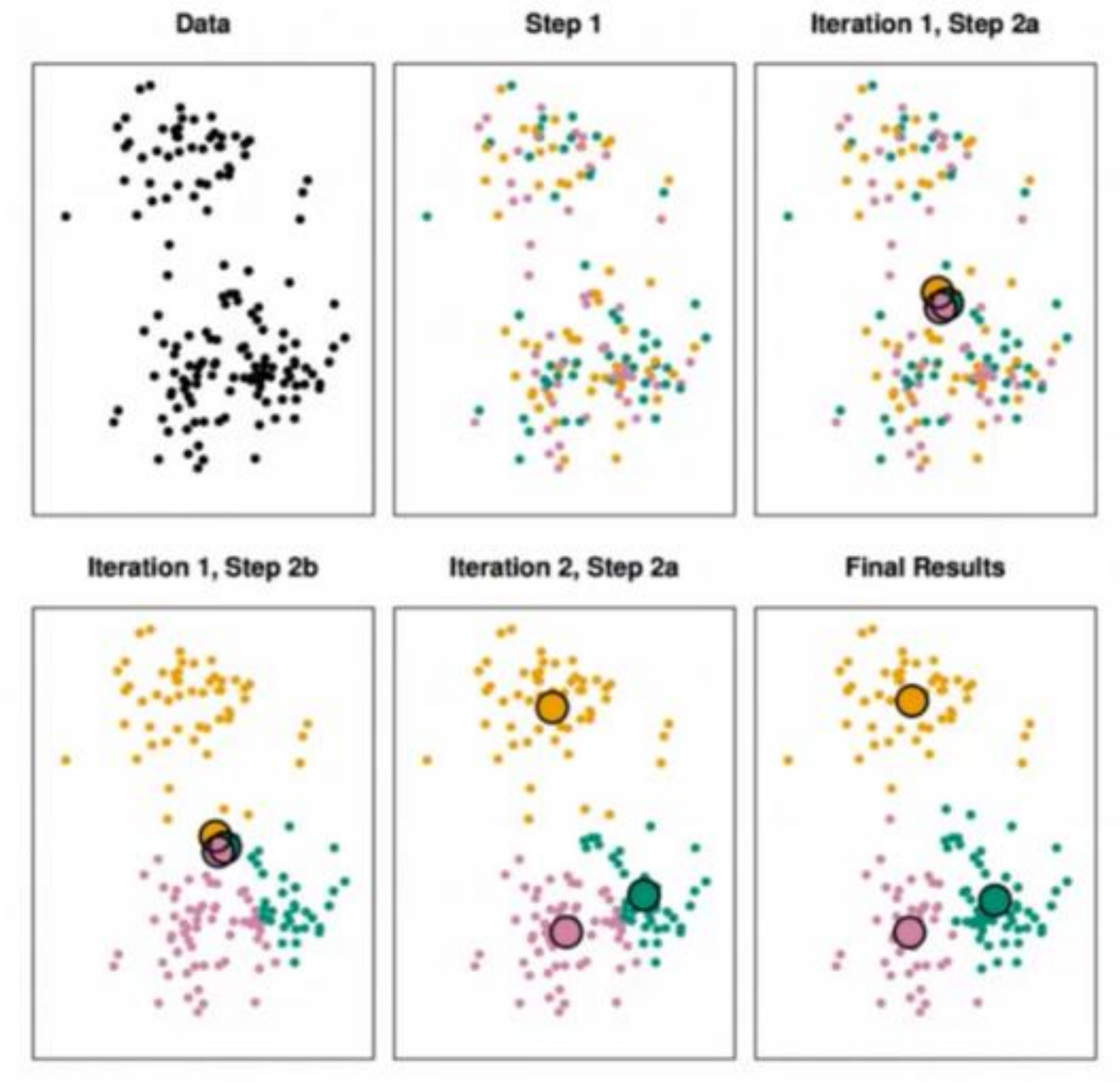
The overall goal is to divide data into distinct groups such that observations within each group are similar

Unlabelled training data
K-means clusters the data into 5 different cluster

K-Means – How Does It Work

- Choose a number of Clusters “K”
- Randomly assign each data point to a Cluster
- Until the clusters stop changing, do the below
  - For each cluster, compute the cluster centroid (Center of cluster) by taking the mean vector of points in the cluster
  - Assign each data point to the cluster for which the centroid is the closest

K-Means Process



❑ **Data**

All observations are plotted. No groups yet

❑ **Step1**

Each observation is randomly assigned to a cluster. Observations are shown by different Colors

❑ **Iteration1-Step2a**

Cluster centroids are shown (large discs orange). Initially these centroids overlap as the initial cluster assignments are chosen at random

❑ **Iteration1-Step2b**

Each data point is assigned to nearest centroid.

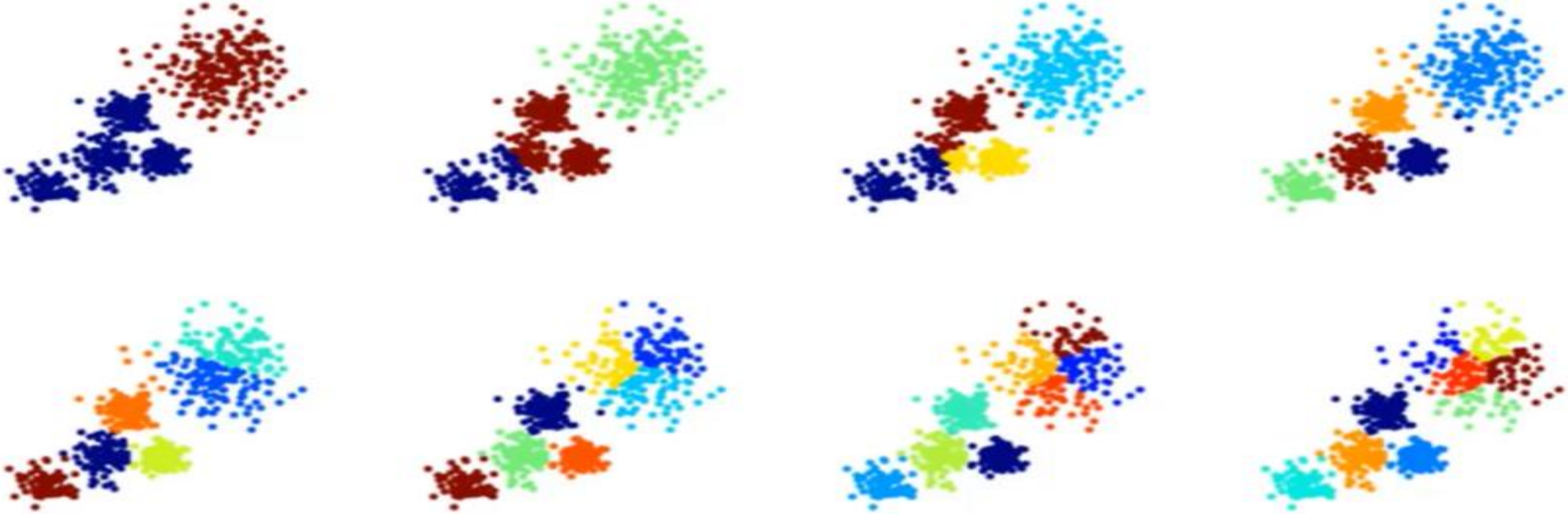
❑ **Iteration 2-Step2a**

Iterate until no new clusters are assigned. See clusters are already forming

❑ **Final results**

after 10 or more iterations

Choosing K Value



❑ **Choosing the best K Value -** Elbow Method can be used to choose the value of K (clusters)

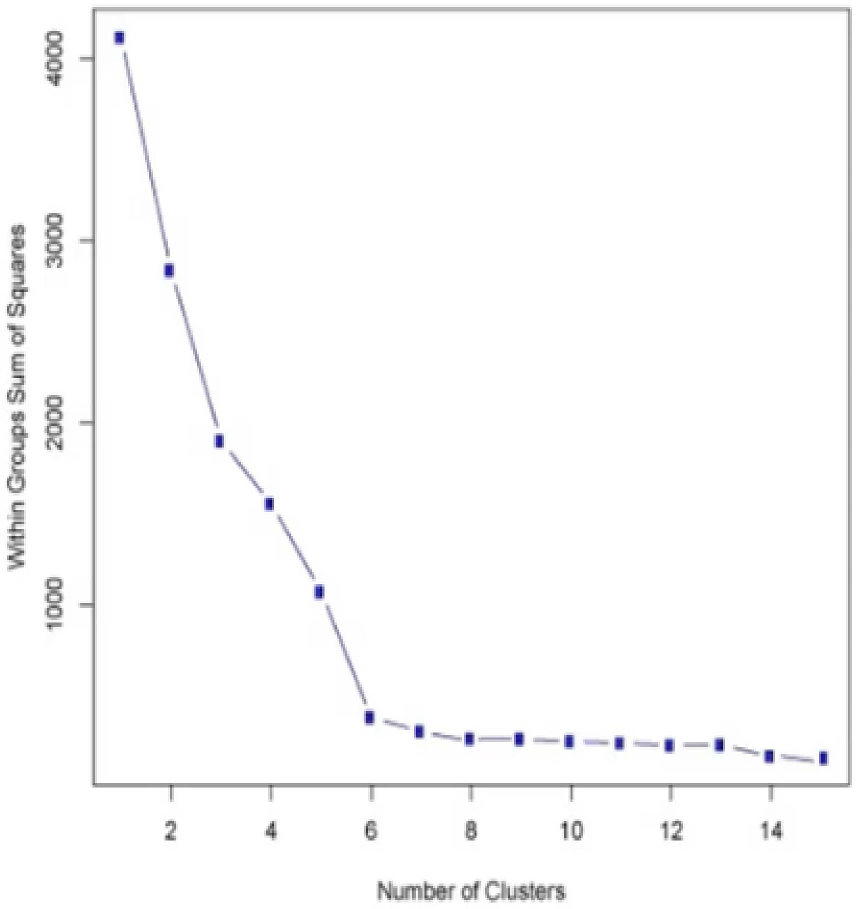
- ▪ Compute the Sum of Squared Errors(SSE) for some values of K
- ▪ SSE = Sum of [Square of (Distance between each data point in a cluster and its centroid)]

❑ **Plot K against SSE - Note as error decreases K gets larger.**

- When Number of clusters increases the size gets smaller, so the error (distortion) will also be smaller

❑ **Elbow Method**

- ❑ Helps in choosing a value of K where the error decreases suddenly
- ❑ A “elbow effect” (as shown in graph) will be produced



- X Axis indicates the K value or number of clusters
- Y Axis – Within group Sum of Squares (SSE)

- We need a K where the SSE does not decrease any more as K increases

- That point is at a K value of 6 Where we are able to see the error does not significantly decrease anymore.

- That would be the ideal K value for our clustering process

- Always remember that we will not be able to find out the perfect k value through this process.
- But this along with your domain knowledge can prove vital in the clustering