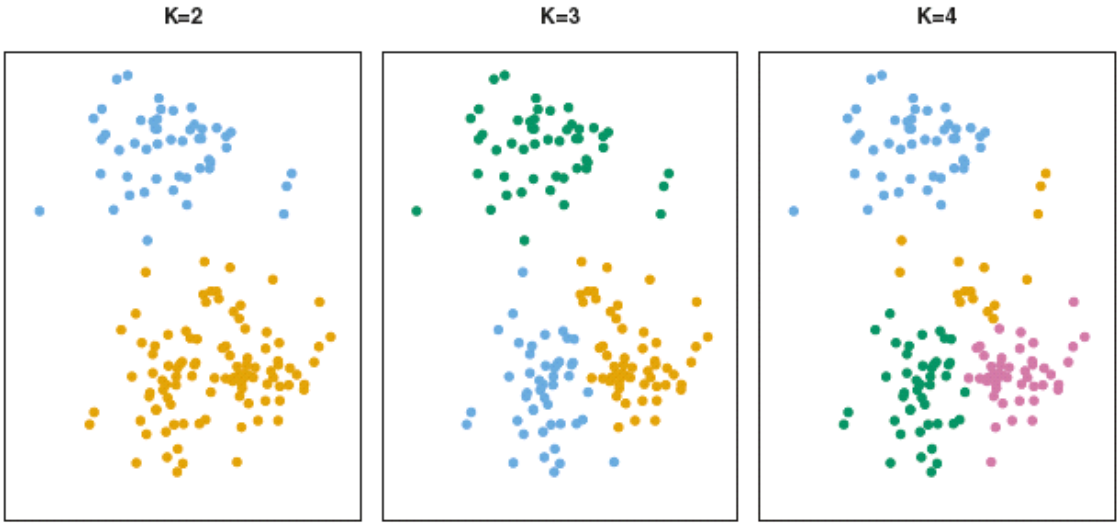
K-MEANS CLUSTERING

- K-means clustering, we seek to partition the observations into a pre-specified number of clusters.

- To perform K-means clustering, we must first specify the desired number of clusters K, then the K-means algorithm will assign each observation to exactly one of the K clusters.



- We want to partition the observations into K clusters such that the total within-cluster variation, summed over all K clusters, is as small as possible.

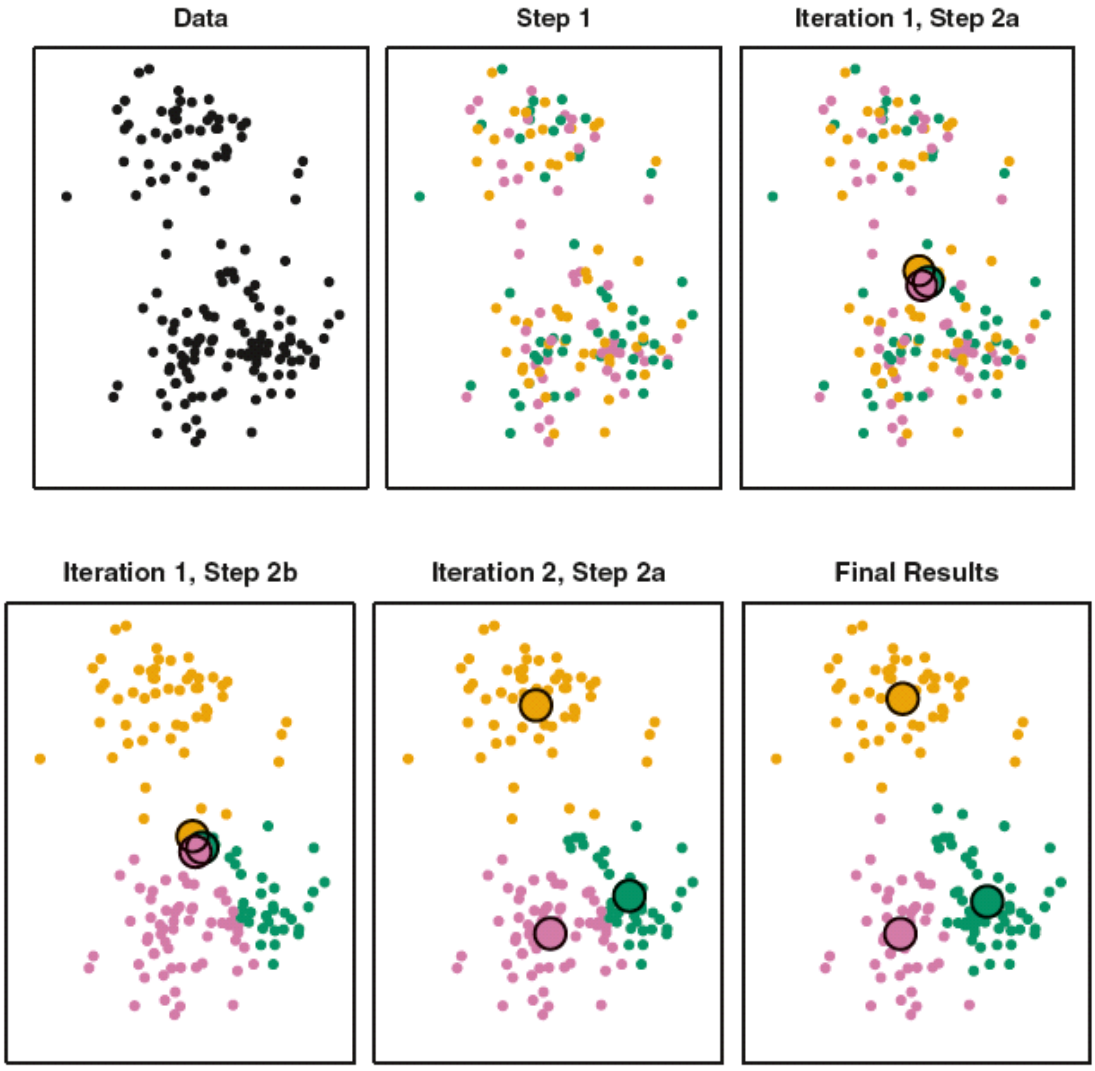
Algorithm

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:

(a) For each of the K clusters, compute the cluster centroid. The K^{th} cluster centroid is the vector of the p feature means for the observations in the K^{th} cluster.

(b) Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).



10.3.3 Practical Issues in Clustering

Clustering can be a very useful tool for data analysis in the unsupervised setting. However, there are a number of issues that arise in performing clustering. We describe some of these issues here.

- In the case of hierarchical clustering,
  - What dissimilarity measure should be used?
  - What type of linkage should be used?
  - Where should we cut the dendrogram in order to obtain clusters?
- In the case of K-means clustering, how many clusters should we look for in the data?

Each of these decisions can have a strong impact on the results obtained. In practice, we try several different choices, and look for the one with the most useful or interpretable solution. With these methods, there is no single right answer—any solution that exposes some interesting aspects of the data should be considered.

Reference taken from the "An Introduction to Statistical Learning".