# Unsupervised Learning, Part 2

K-Means Clustering

Introduction to Statistical Learning, Section 10.3.1

# Clustering

A broad set of techniques for finding subgroups, or clusters, in a data set.

**Goal:** observations within each group are similar to each other (high intra-cluster similarity), while observations in different groups are different from each other (low inter-cluster similarity).

# K-Means Clustering

**Goal:** partition the dataset into *K* distinct, non-overlapping clusters

Want to minimize the total within-cluster variation.

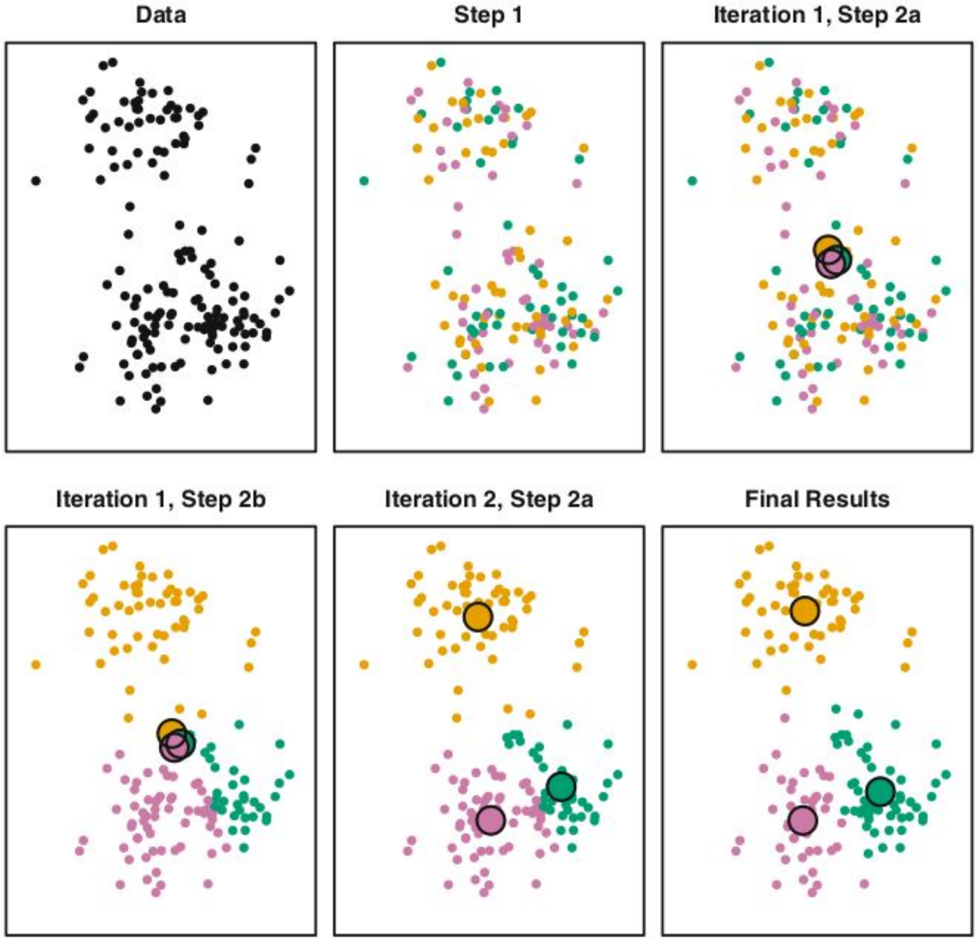
This algorithm depends on distance, so it’s best practice to standardize (center + scale) prior to performing clustering.

# K-Means Clustering

Stolen from *Introduction to Statistical Learning*

1) Randomly assign point to clusters
2) Find the centroid of each cluster
3) Assign each point to the nearest centroid
4) GOTO 2

This algorithm depends on the initial random distribution of clusters, so you can end up with different final results over different runs.



# K-Means Clustering

When doing *K*-Means clustering, we have to choose *K*, the number of clusters!

The “Elbow Method”:

- Run the algorithm over a range of number of clusters
- Plot the total within-cluster sum of squares
- Find an “elbow” in the plot, where the rate of improvement decreases.

# Example Notebook

## KMeans.Rmd