# 23) K-Means Clustering and Hierarchical Clustering

Vitor Kamada

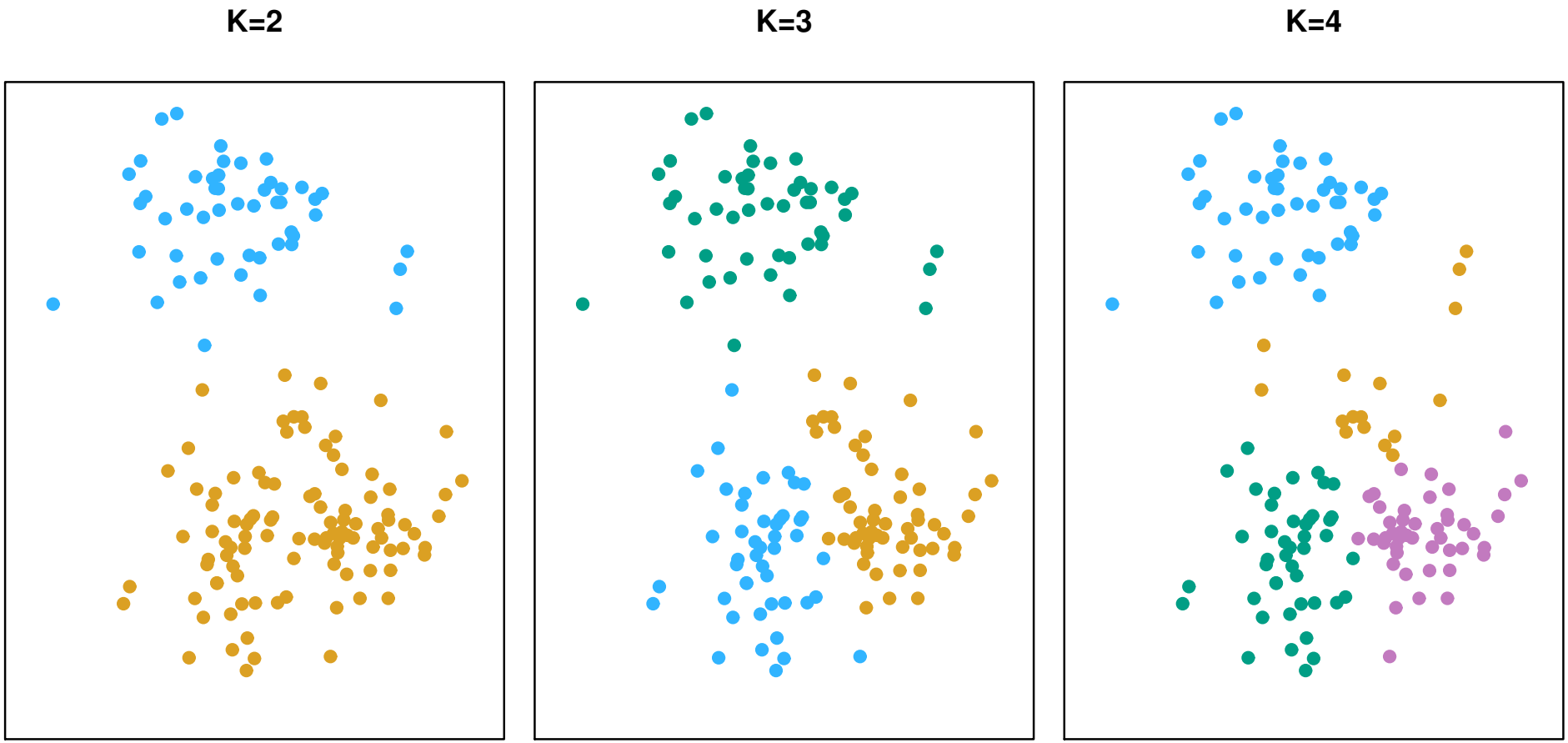
January 2020

# Reference

Tables, Graphics, and Figures from:

James et al. (2017): Ch 10.3

# Simulated Data Set with 150 Observations



# K-Means Clustering

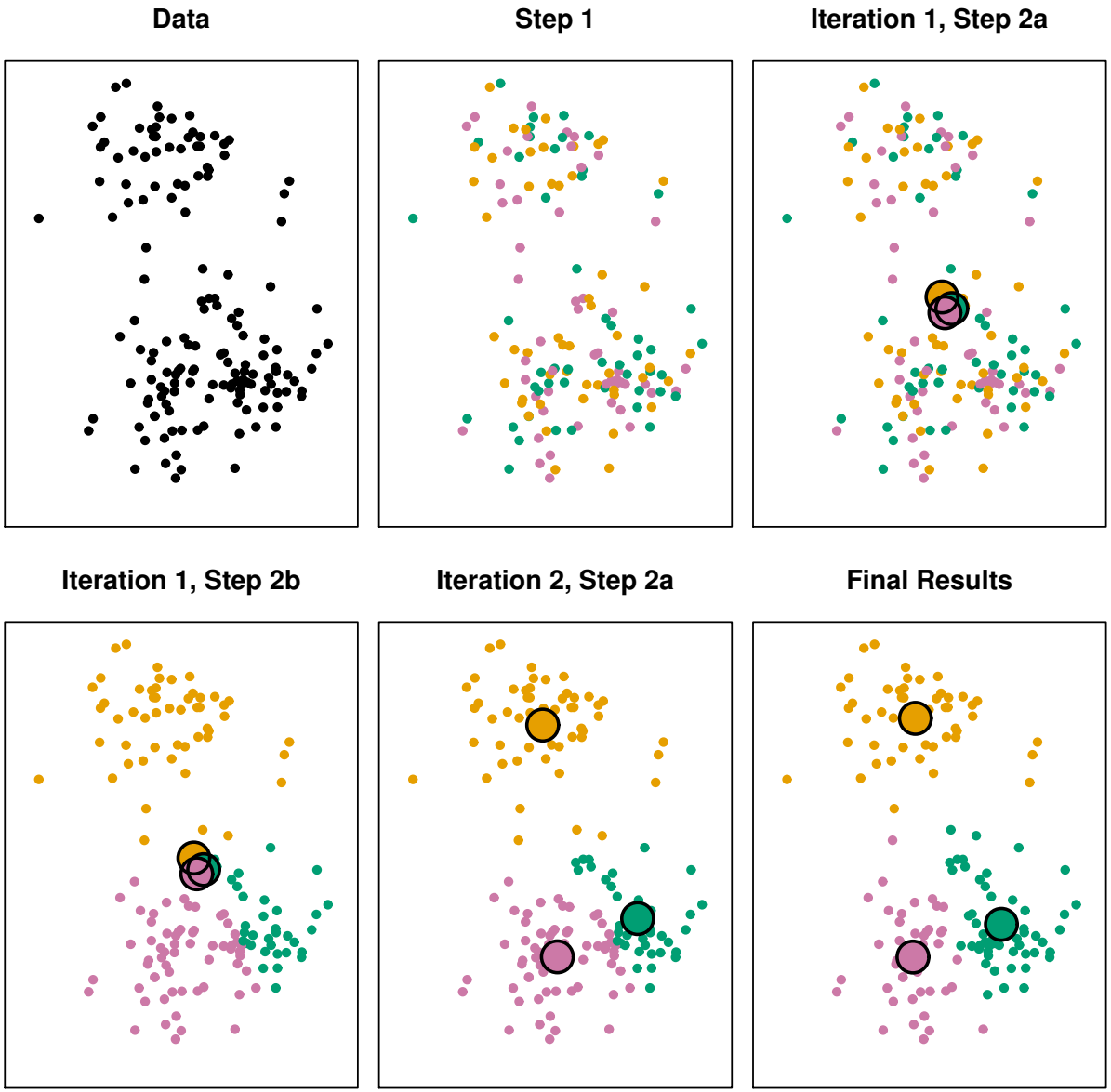
$$C_1 \cup C_2 \cup ... \cup C_K = \{1, ..., n\}$$

$$C_k \cap C_{k'} = \emptyset \text{ for all } k \neq k'$$

$$\underset{C_1,...,C_k}{Minimize}\{\sum_{k=1}^{K} W(C_k)\}$$

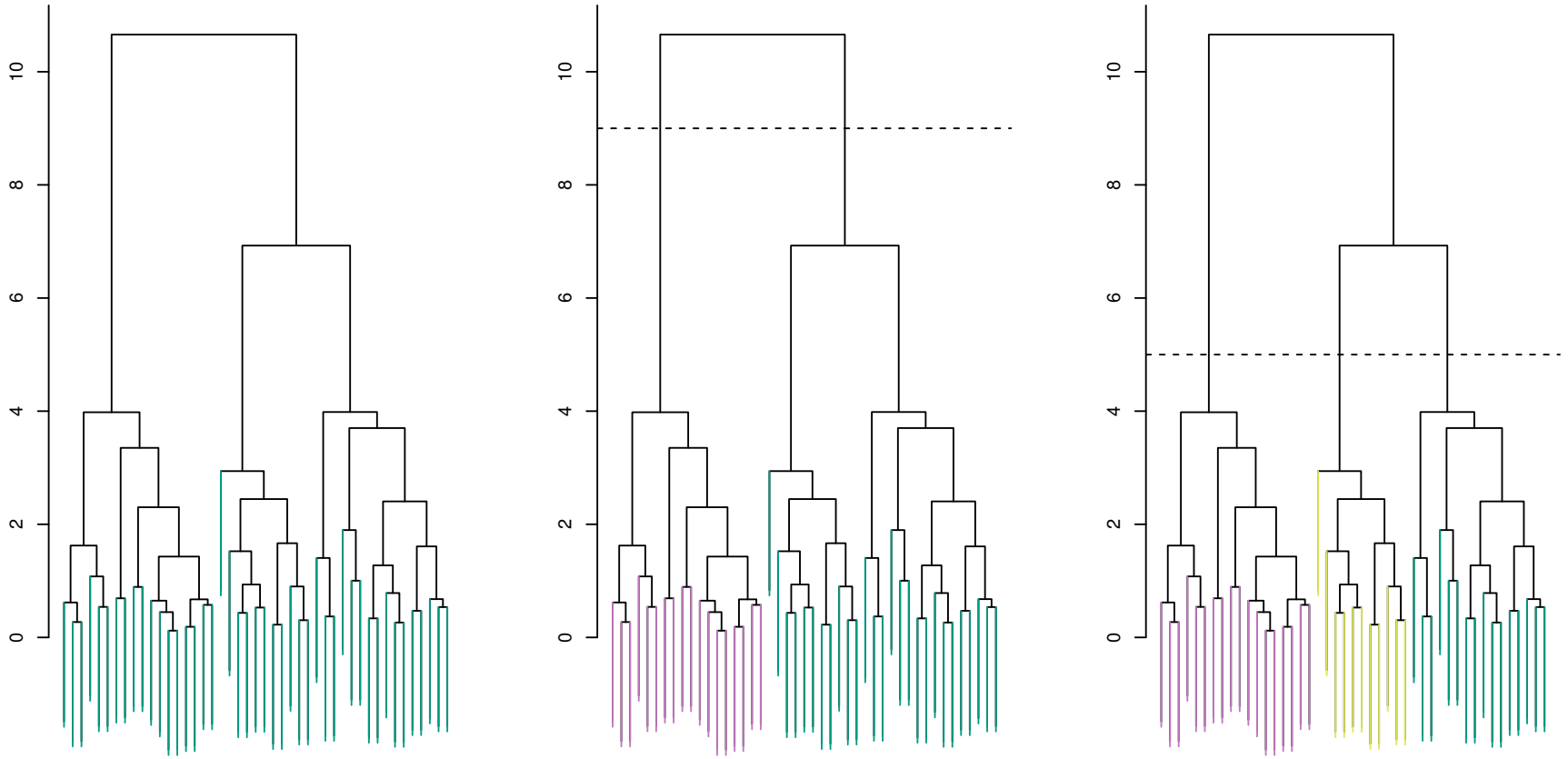
$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

# K-Means Clustering Algorithm

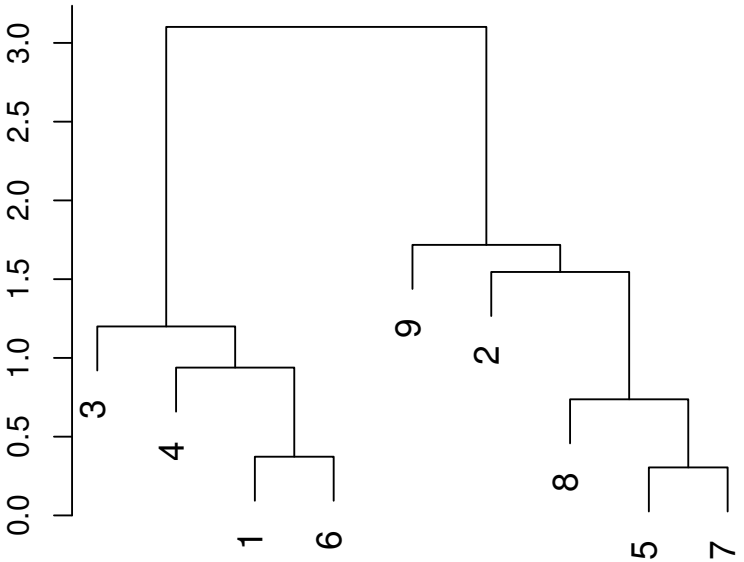


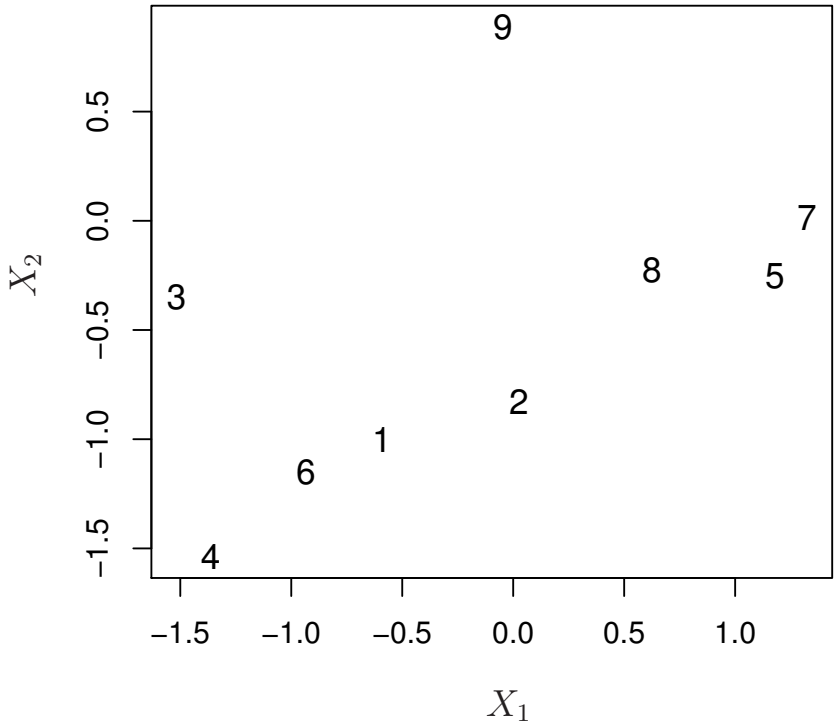
# Different Random Assignment

# Hierarchical Clustering - Dendrogram



# 9 Observations: Euclidean Distance and Complete Linkage



# Measure of Dissimilarity ($d(G, H)$)

## Single Linkage or Nearest-Neighbor

$$\min_{i \in G, i' \in H} d_{ii'}$$

## Complete Linkage or Furthest-Neighbor

$$\max_{i \in G, i' \in H} d_{ii'}$$

## Group Average

$$\frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'}$$

# Average, Complete, and Single Linkage