Clustering

Clustering
Presented by David John Baker
December 30th, 2019

// FLATIRON SCHOOL

Outline

What is Clustering?

- Definitions of Clustering
- When To Use
- Case Study: Marketing Segmentation

K-Means Clustering

- Selecting K
- Advantages/Disadvantages

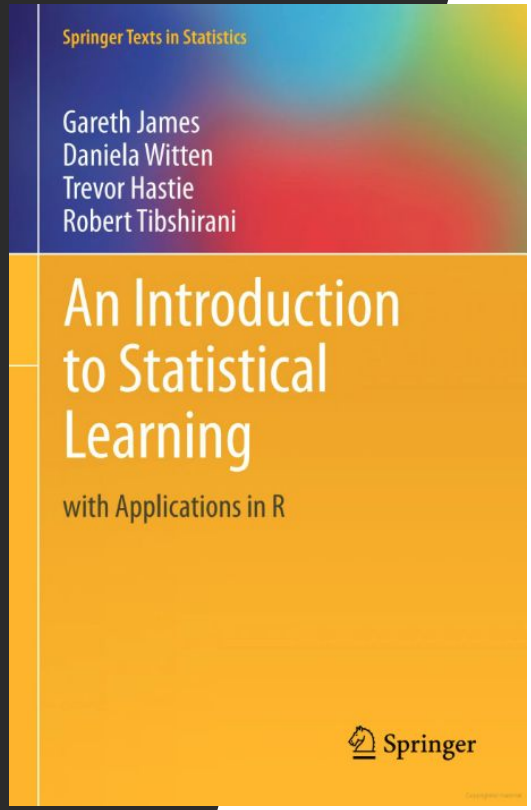
Hierarchical Clustering

- Dendrograms
- Deciding Where to Cut

Dangers of Unsupervised Learning

- When ML Goes Wrong!!

Outline



What is Clustering?

- *Definitions of Clustering*
- *When To Use*
- *Case Study: Marketing Segmentation*

K-Means Clustering

- *Selecting K*
- *Advantages/Disadvantages*

Hierarchical Clustering

- *Dendrograms*
- *Deciding Where to Cut*

Dangers of Unsupervised Learning

- *When ML Goes Wrong!!*

In what type of situation would you want to find subgroups within a larger population???

(don’t say market segmentation…)

What is Clustering???

Set of techniques for finding subgroups or clusters in a larger population

In order to find out what subgroups ARE you need to know how similar something is…?

Discussion Question:

Which genres of music are closer to one another?

Rock **Jazz** **Bluegrass** **Country**

Discussion Question: Answer?

In order to find similarity between groups, you need to define what you even mean by similar. Your definition of similar depends on what variables you have and how you measure them.

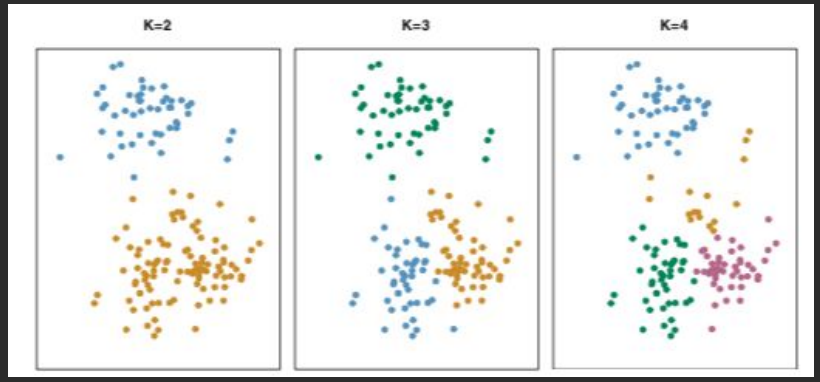
PCA vs Clustering

- PCA: Low Dimensional Representation of Data that explains portion of the data
- Clustering: Find homogeneous subgroups

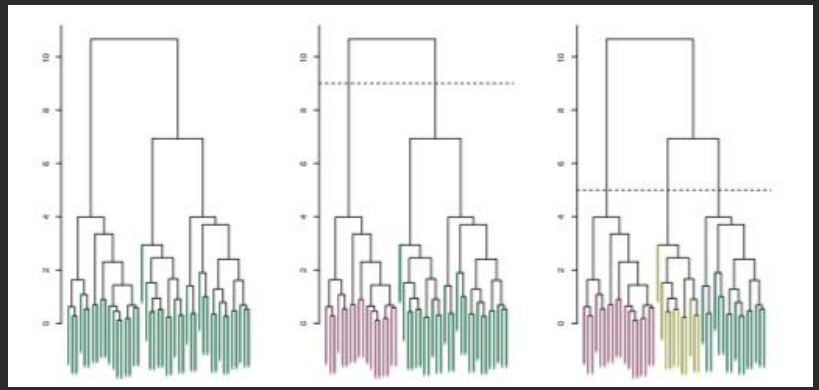
How are these different?

Types of Clustering

K-Means Clustering



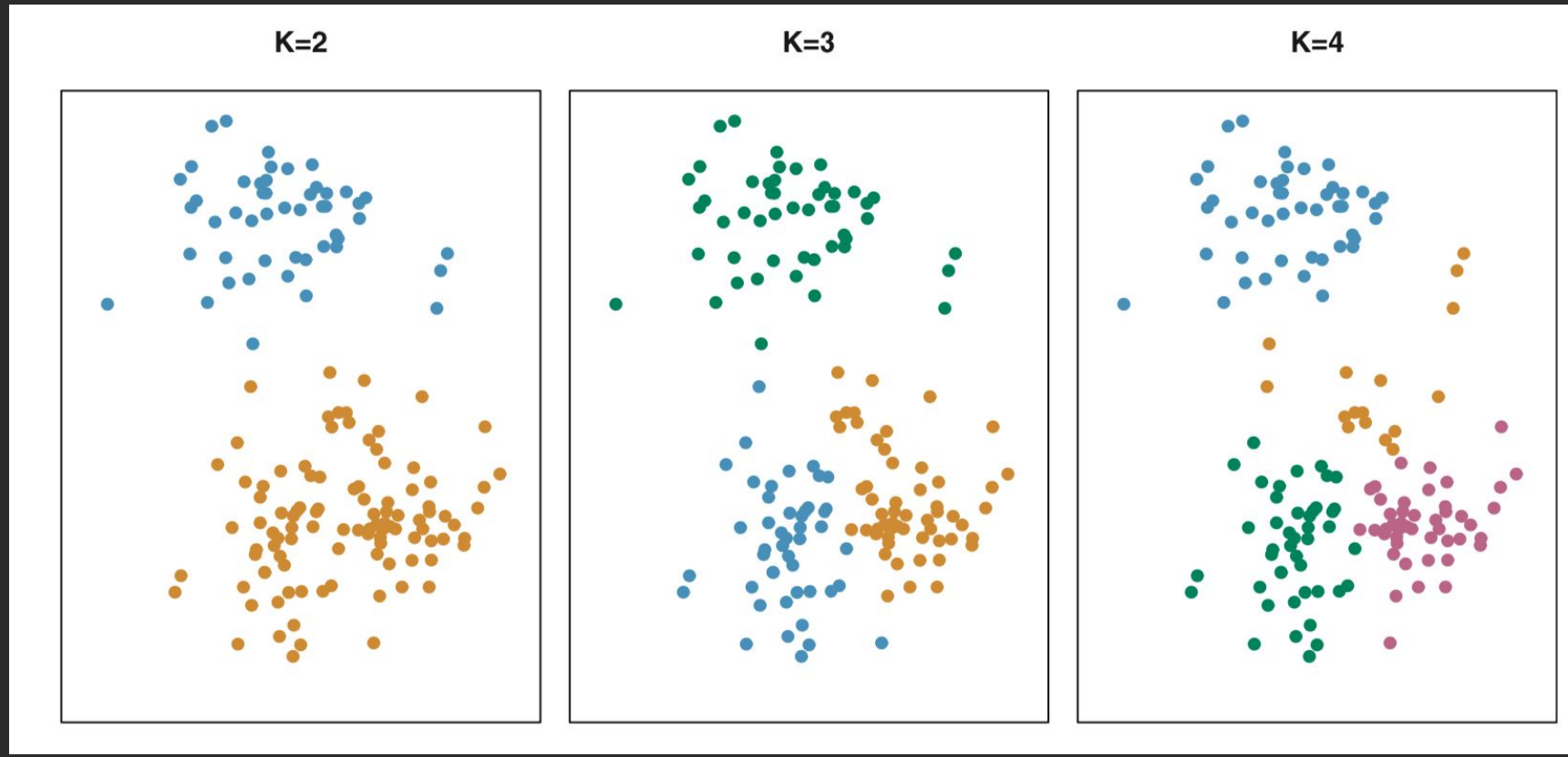
Hierarchical Clustering



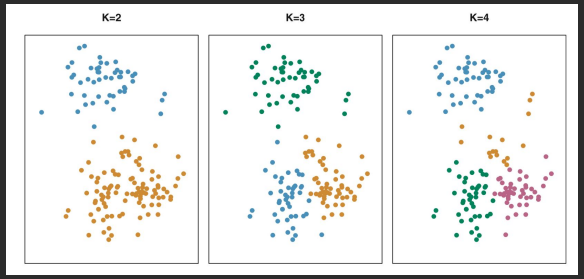
Types of Clustering

K-Means Clustering

Hierarchical Clustering

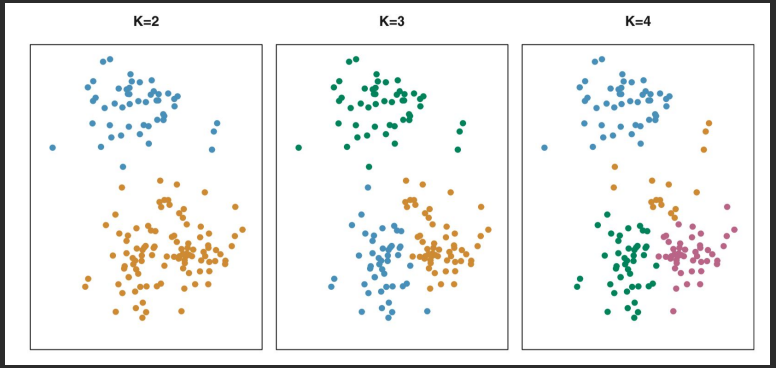

K=2
K=3
K=4

K Means Basics



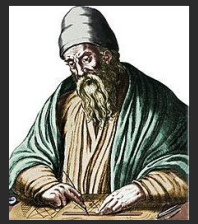
- Each observation belongs to a cluster
- No overlapping clusters!
- We want to MINIMIZE within cluster variation

Within Cluster Variation

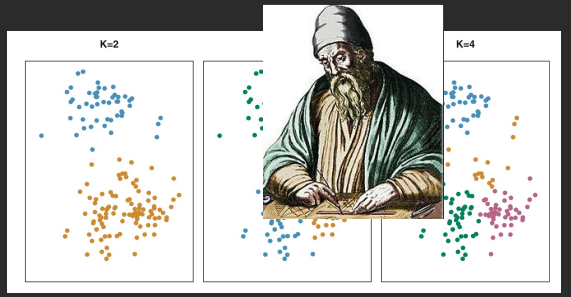


- Most common is *Squared Euclidean Distance*

- Sum all squared pairwise distances
- Divide by number of observations
- But how to do it?!

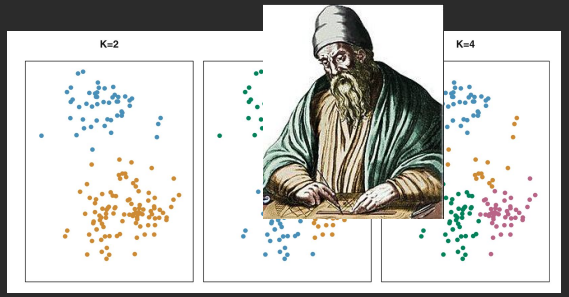


Computation Problem



- There are K^n ways to partition n observations into K clusters!!
- Where do you even begin !?

Computation Problem



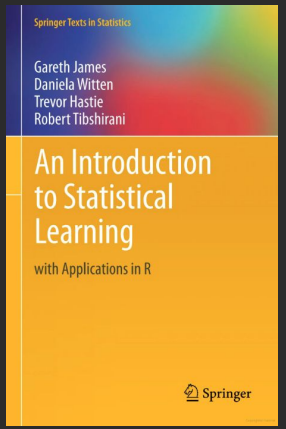
Algorithm 10.1 *K-Means Clustering*

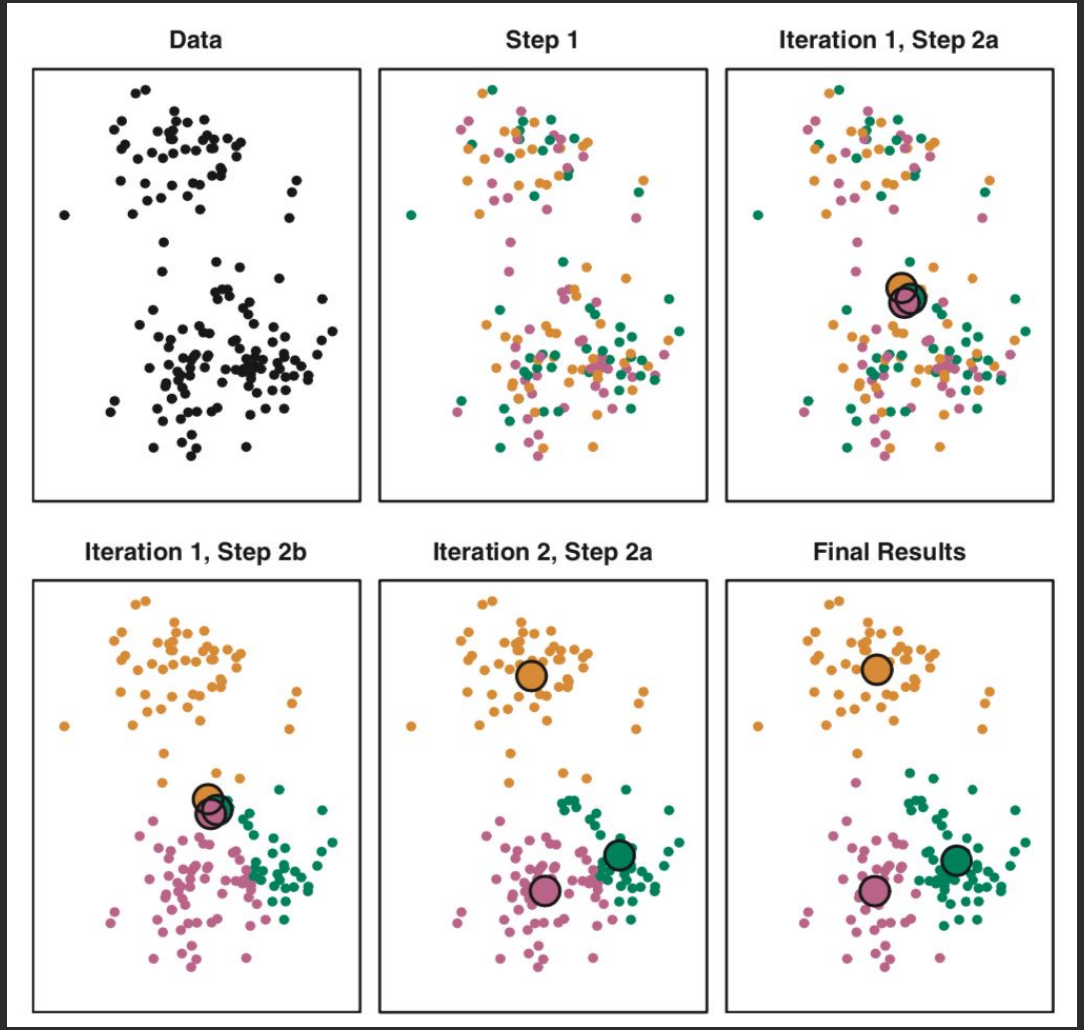
1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:

   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

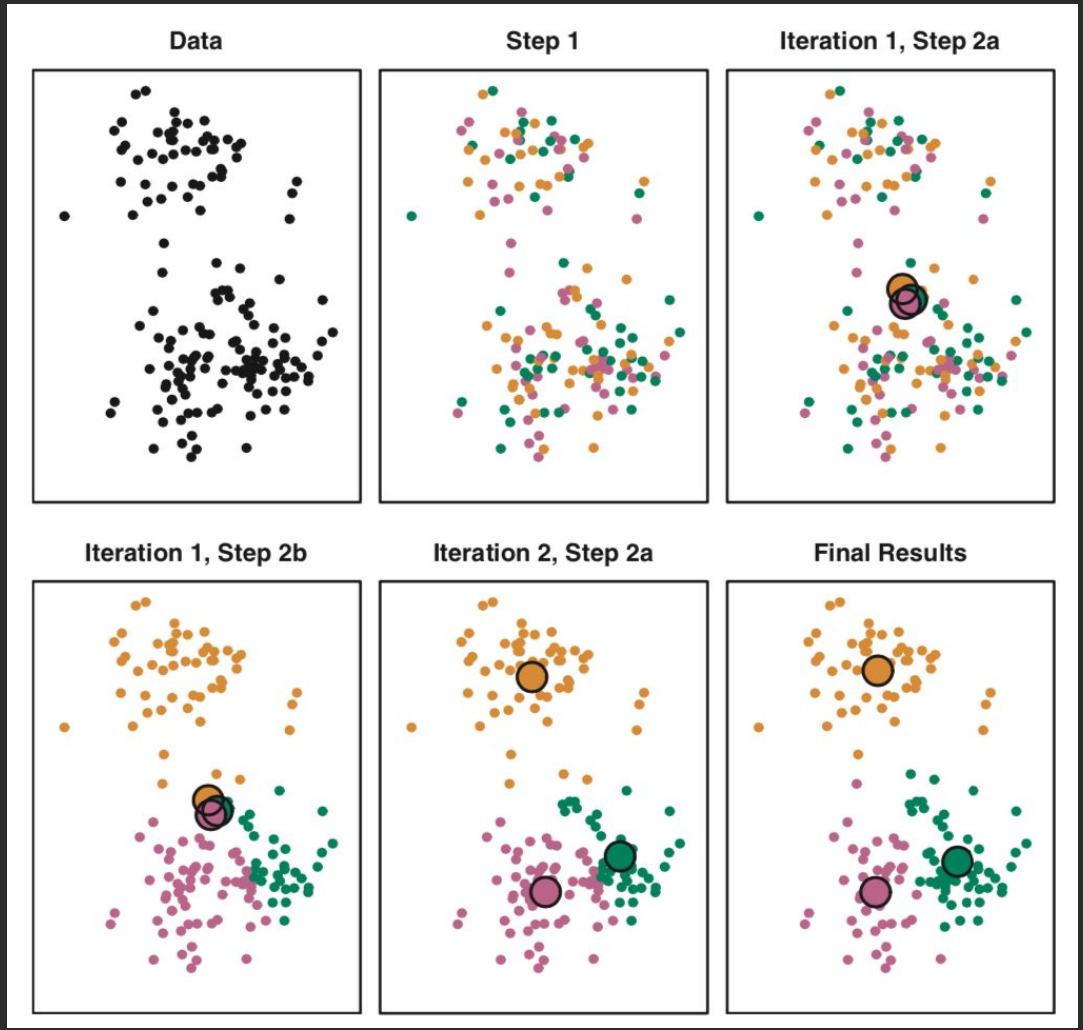
   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).




Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

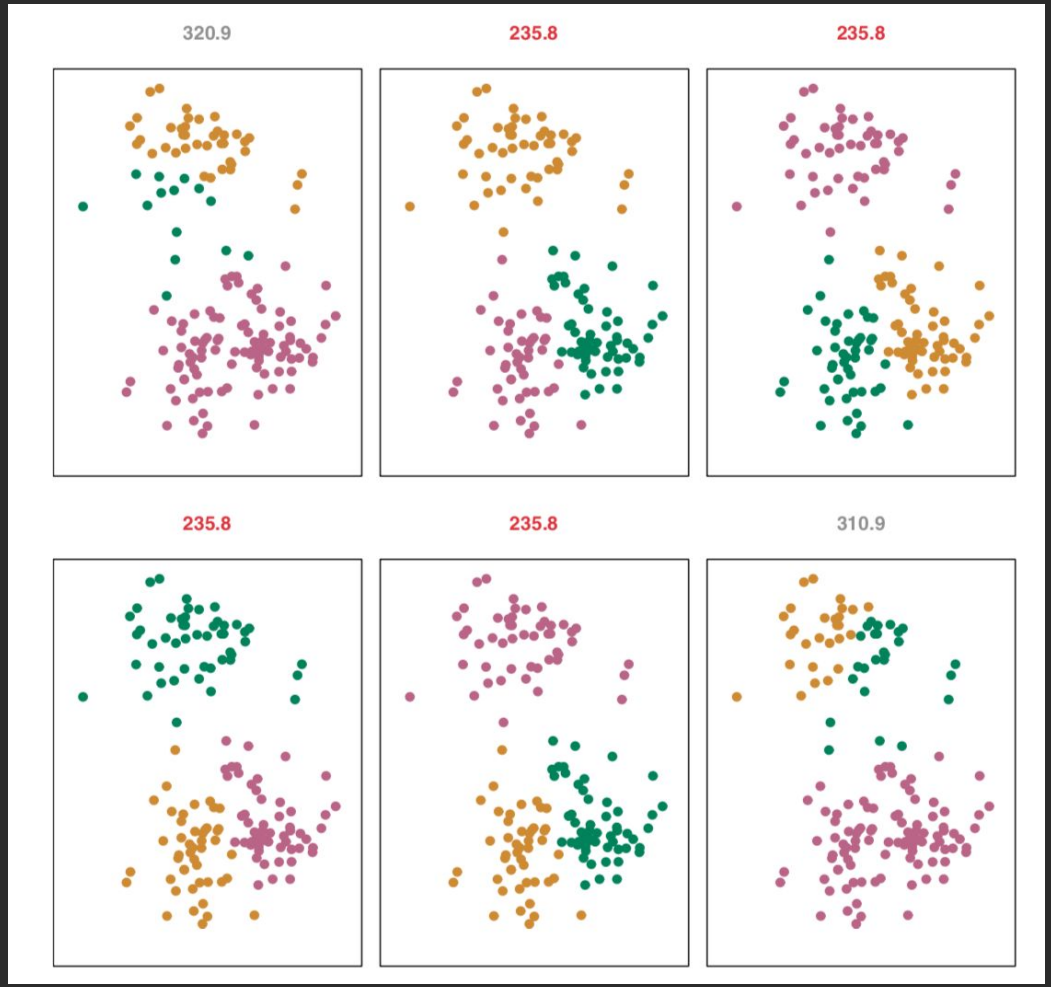
Algorithm 10.1 *K-Means Clustering*

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 kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).



Starting State Problems…

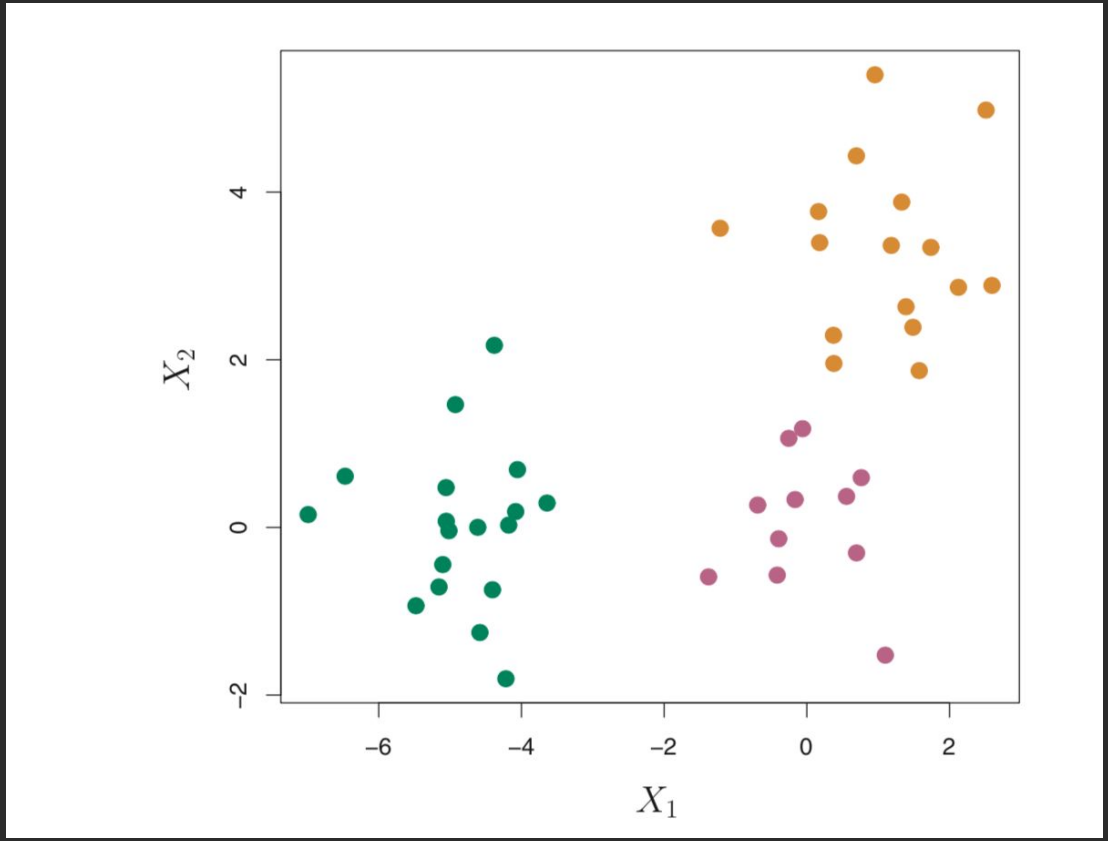
Need lower Objective!! (red)



Hierarchical Clustering

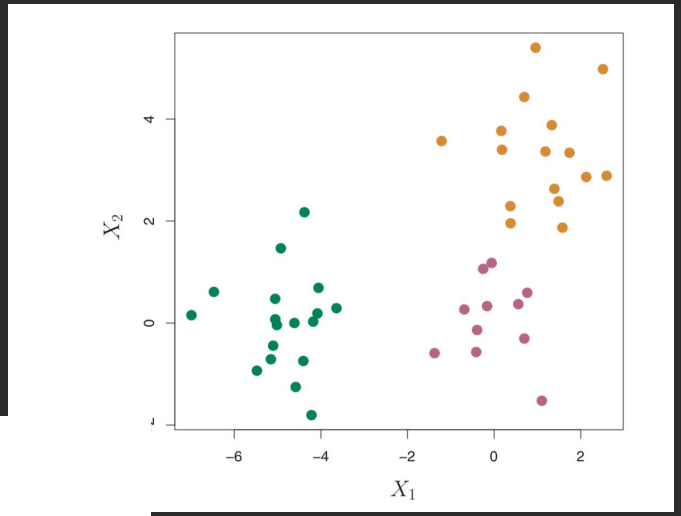
- What if you don’t know K? (Top down)
- What if there are clusters within clusters worth writing home about?!
- Gotta go bottom (agglomerative) up!
- More common form
- Build it bottom up!

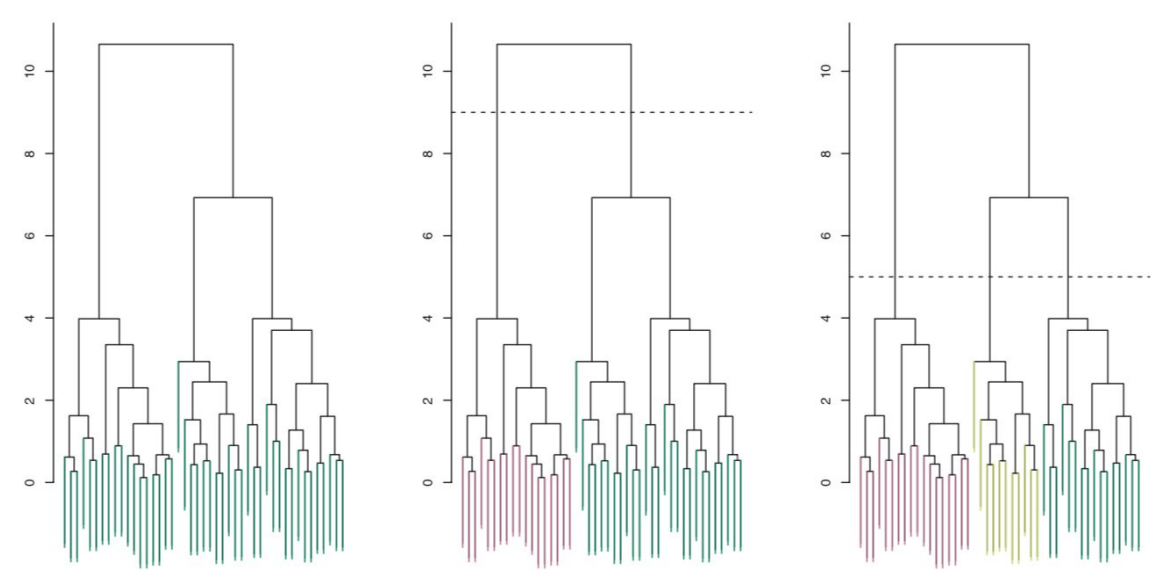
How do we recover these clusters?



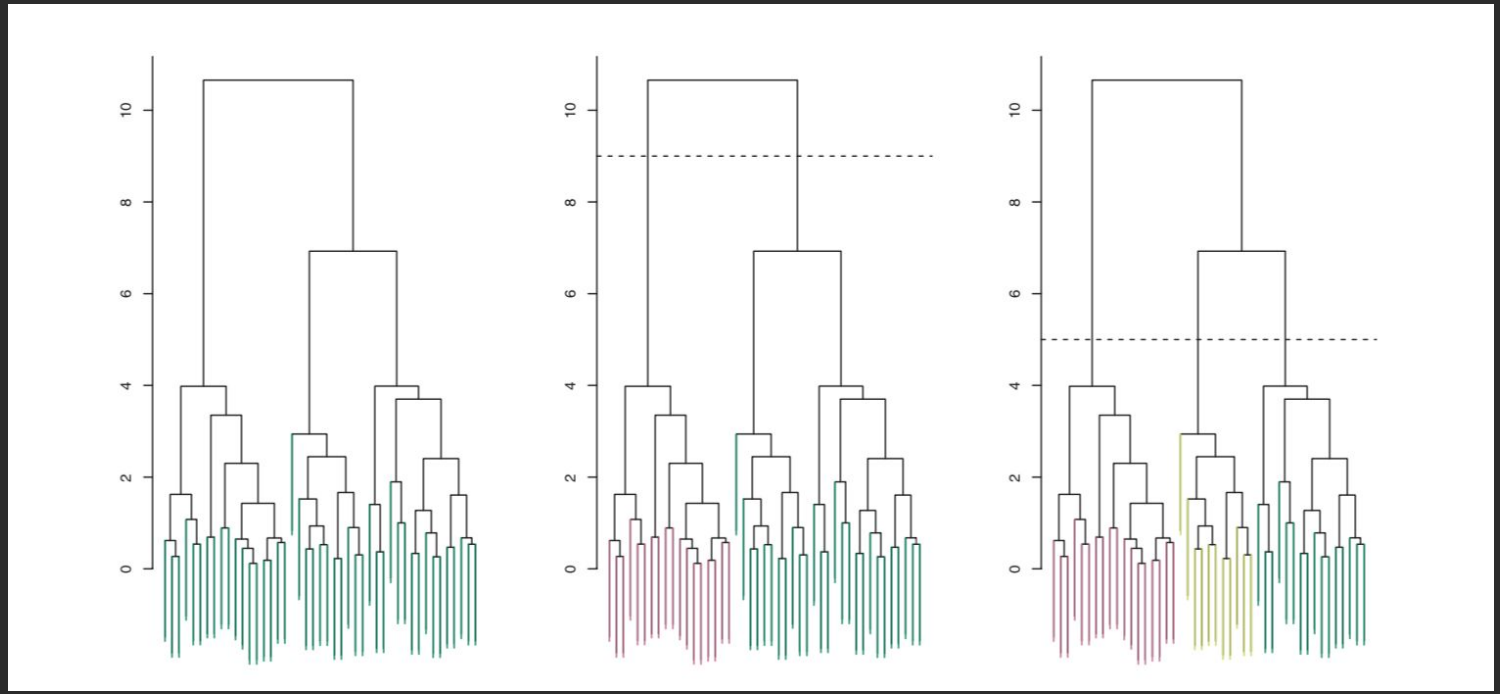
How do we recover these clusters?

Build tree from leaves to branches!



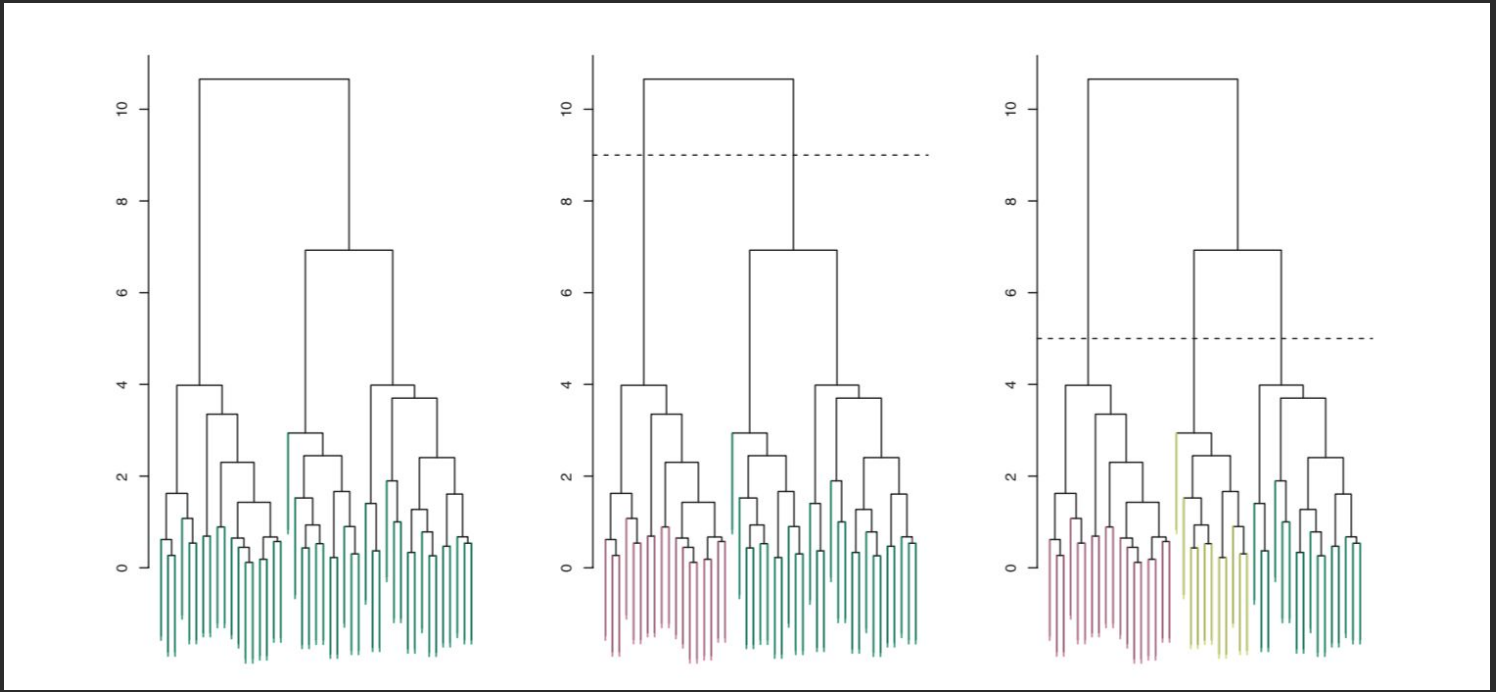


Dendrogram

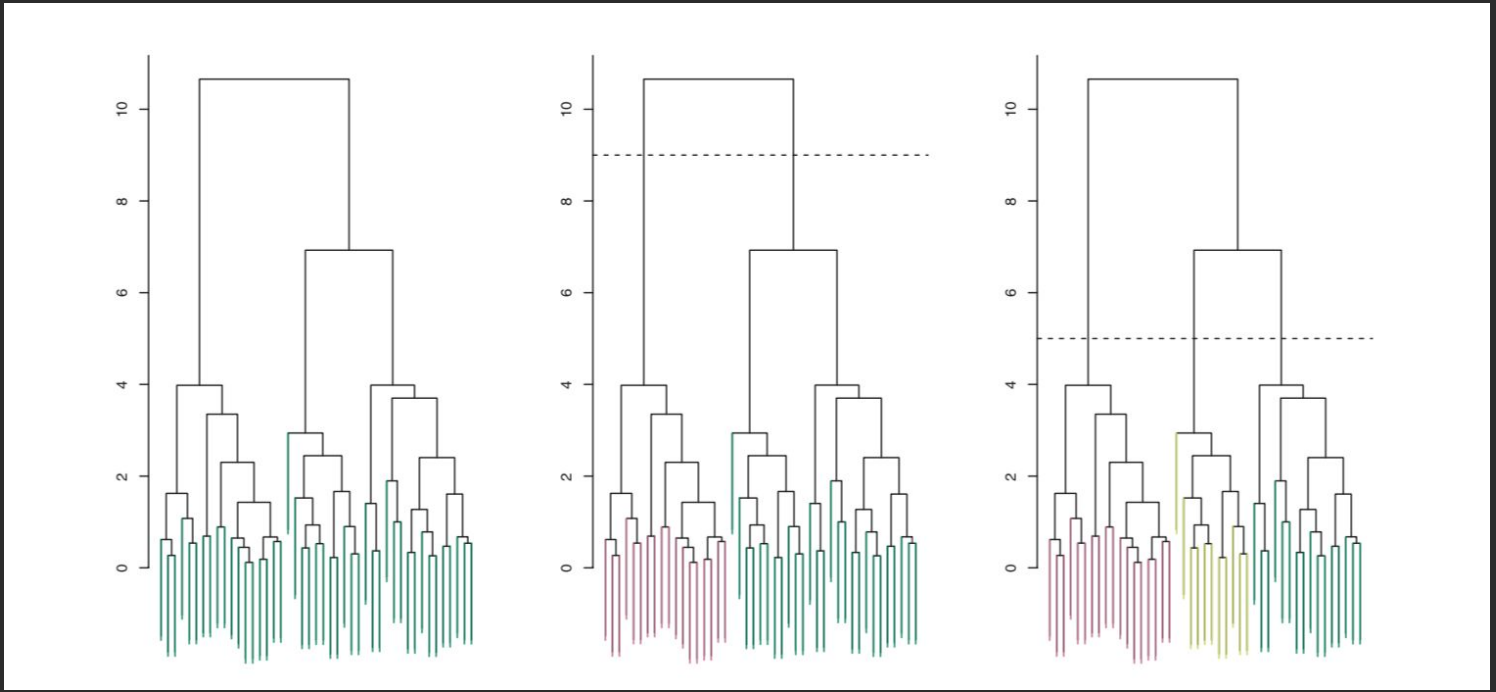


As we go up tree, observations become more different from one another

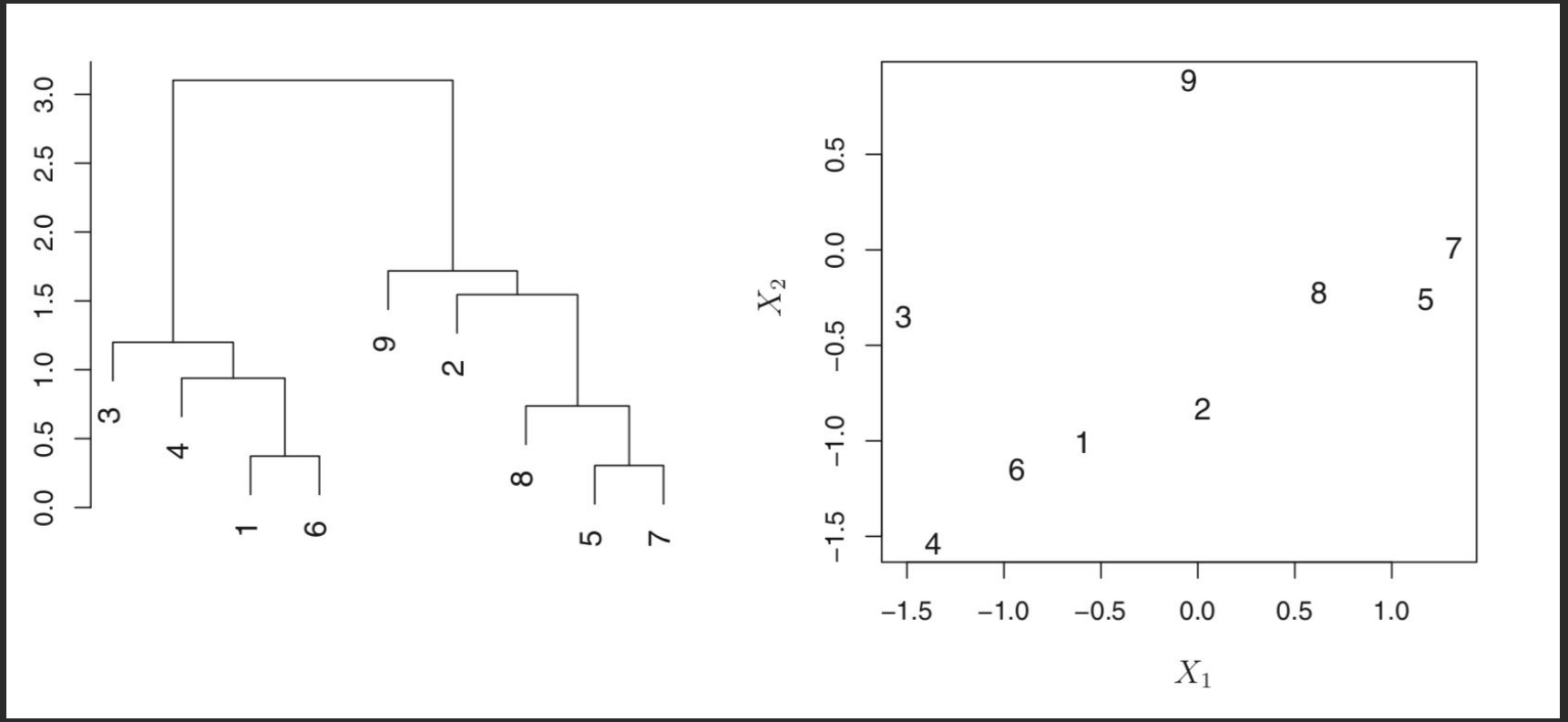
Lower in tree == More Similar to each other



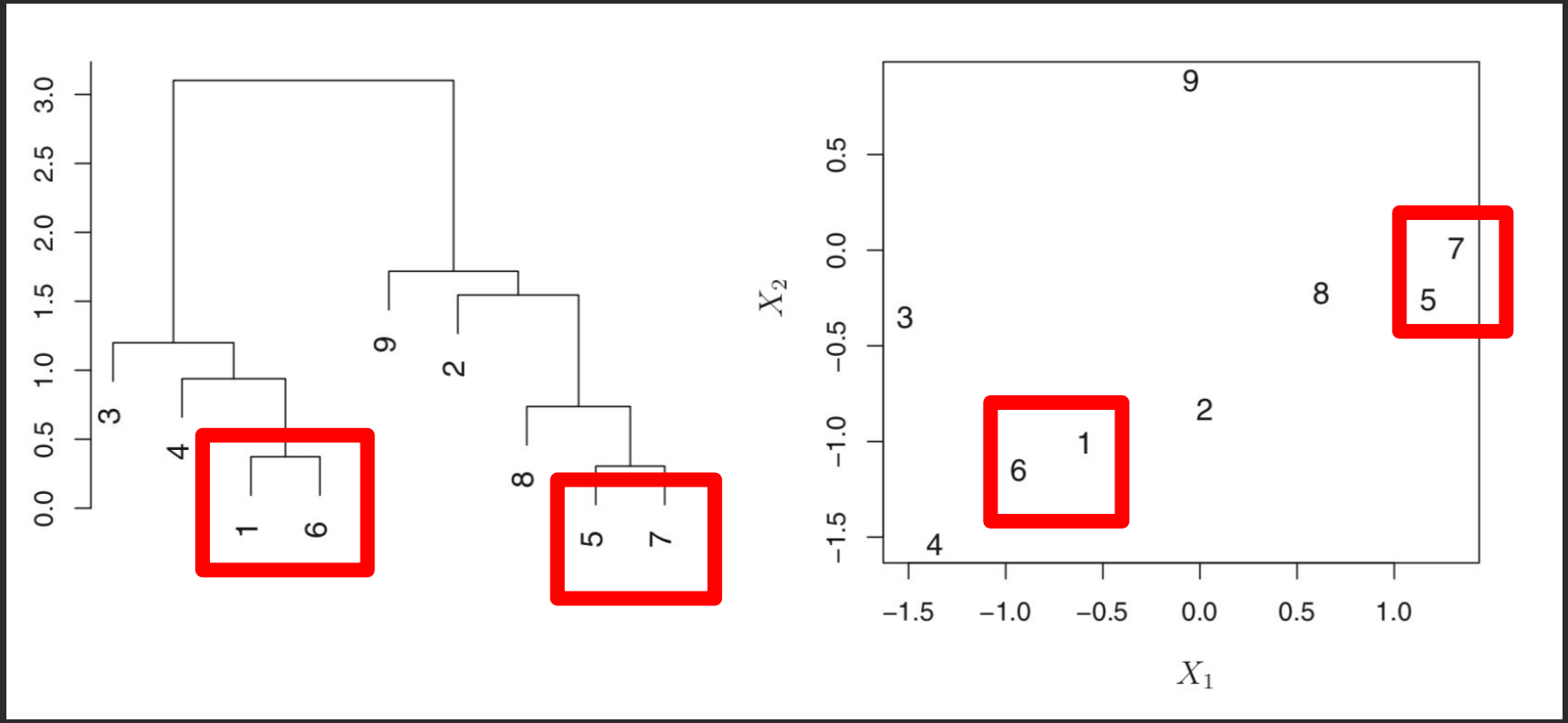
Height of fusion per vertical indicates similarity



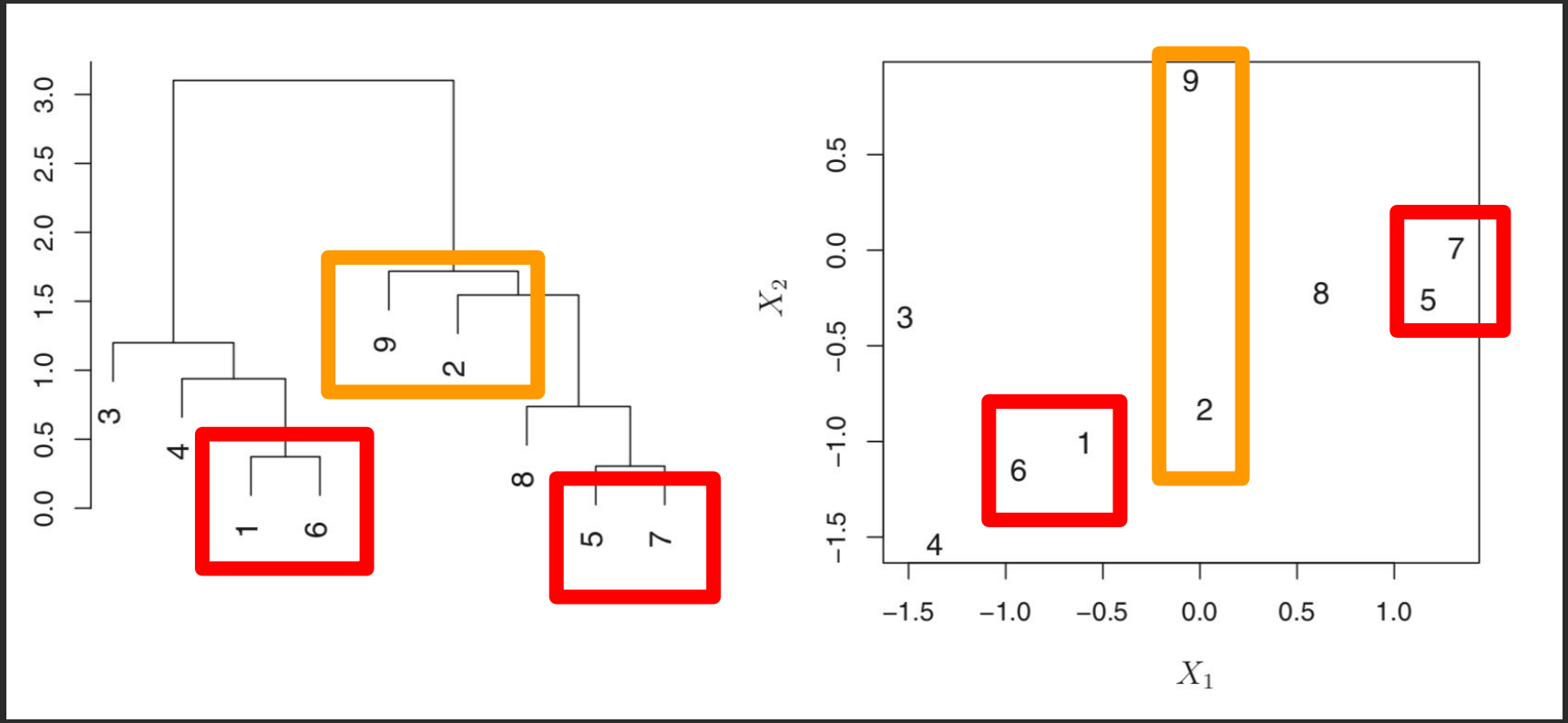
Height of fusion per vertical indicates similarity



Clustering Danger Zone!

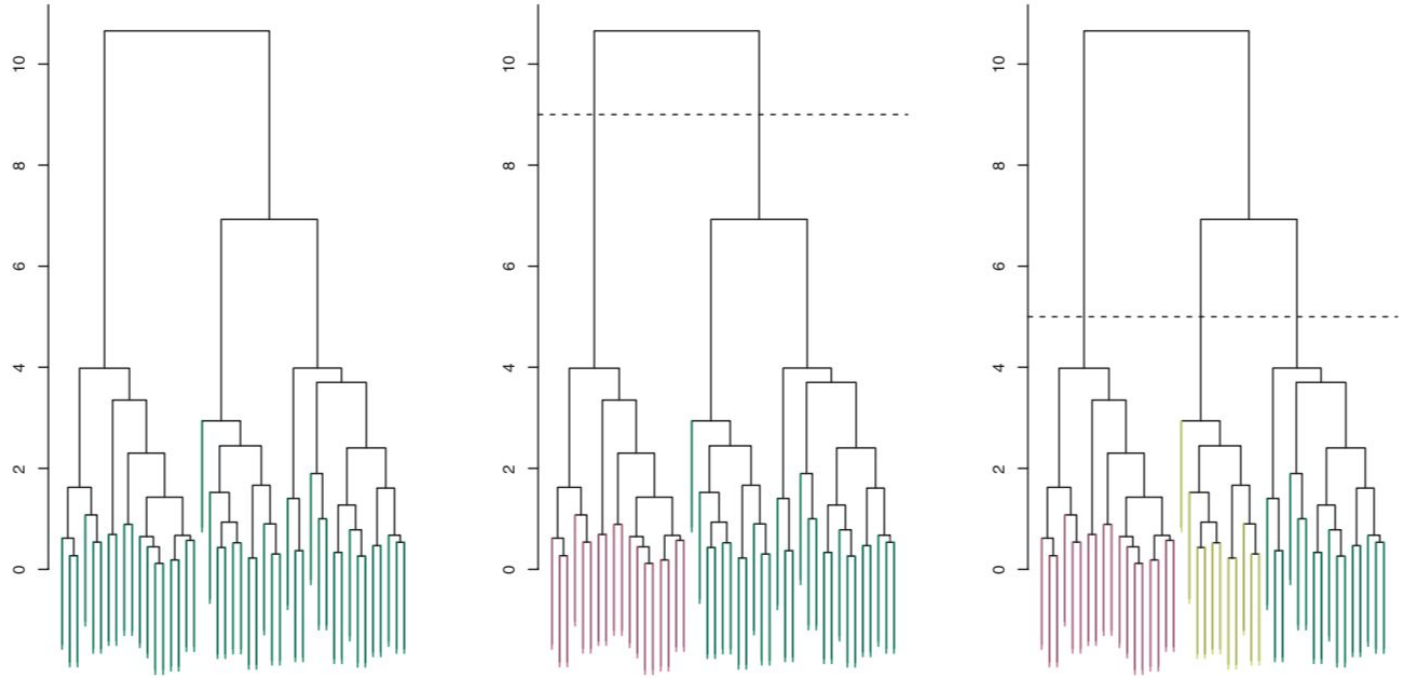


Clustering Danger Zone!



Clustering Danger Zone!

DO NOT MAKE CONCLUSIONS ON HORIZONTAL AXIS!!



Make horizontal cut, get distinct clusters

One dendrogram, many clusters!!

Hierarchical Clustering Algorithm

Algorithm 10.2 *Hierarchical Clustering*

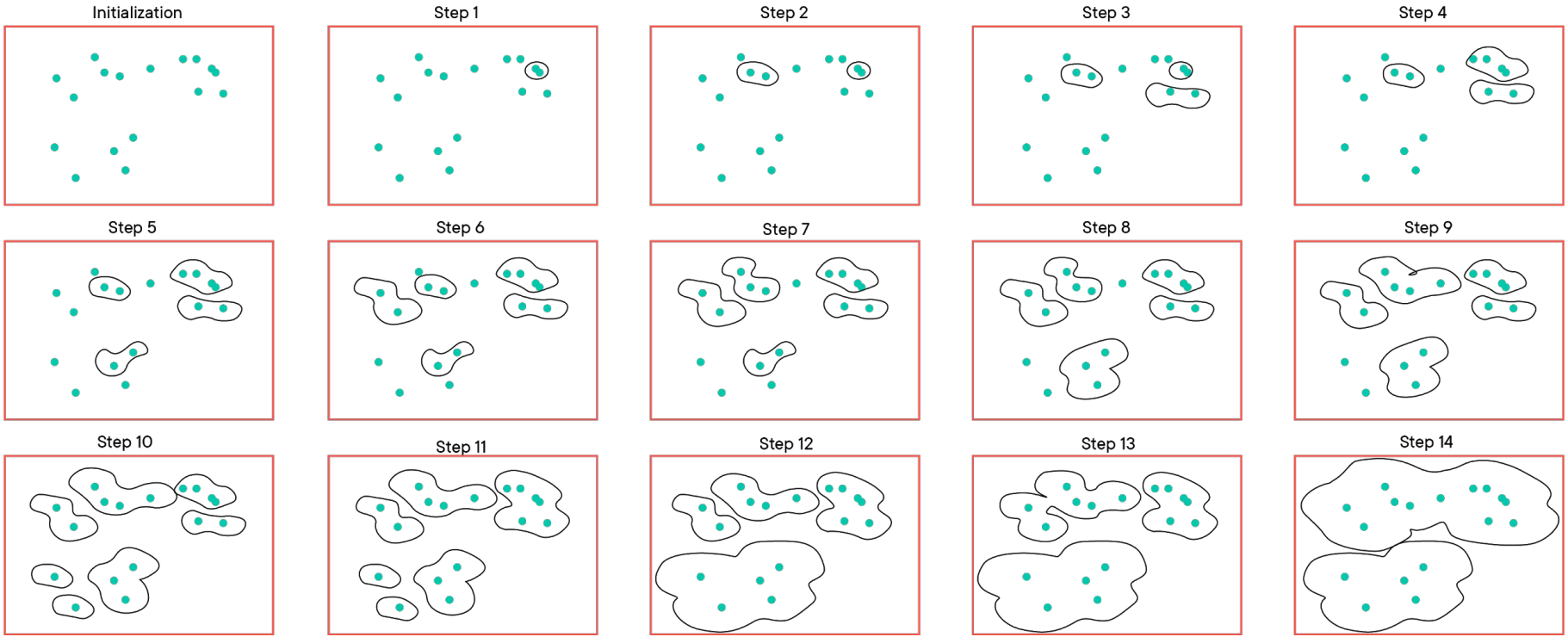
1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

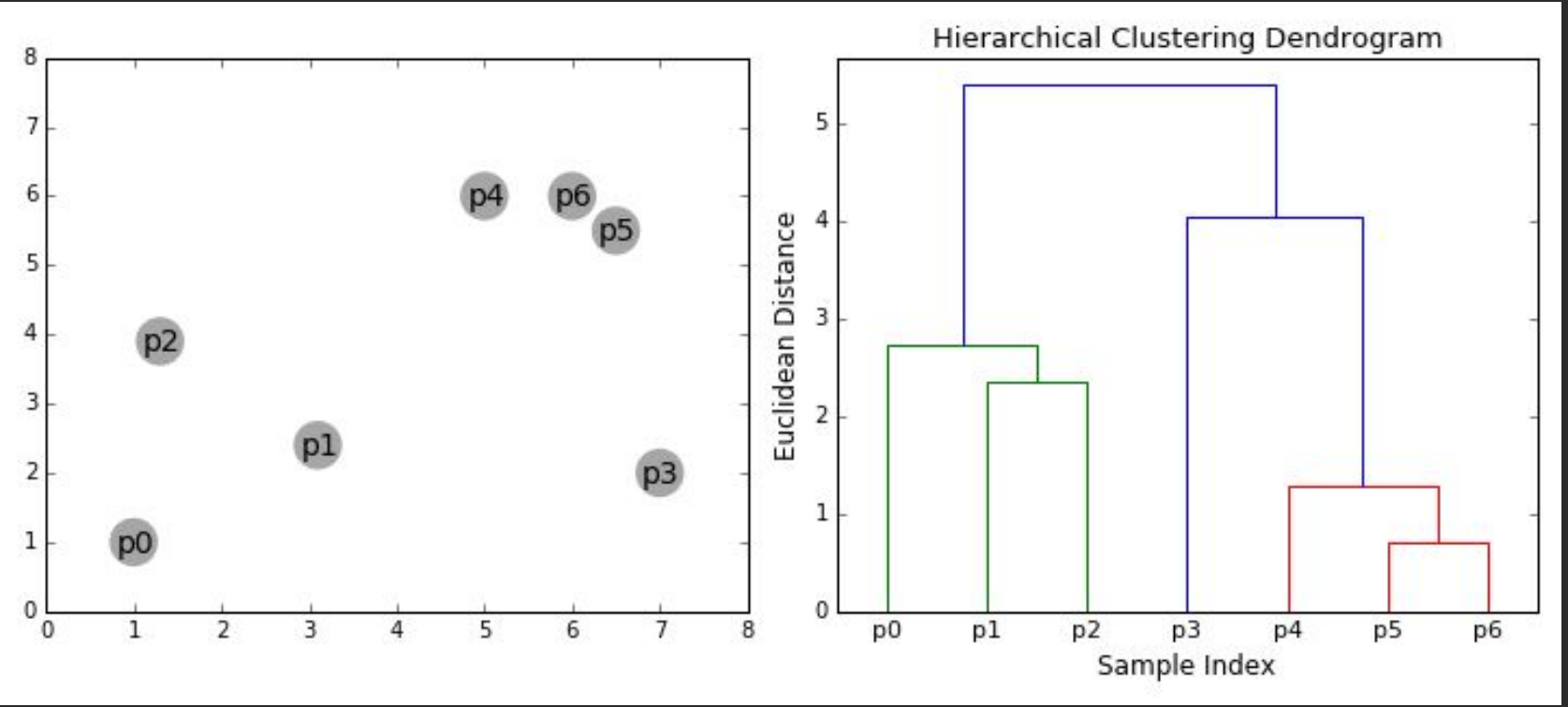
2. For $i = n, n-1, \ldots, 2$:

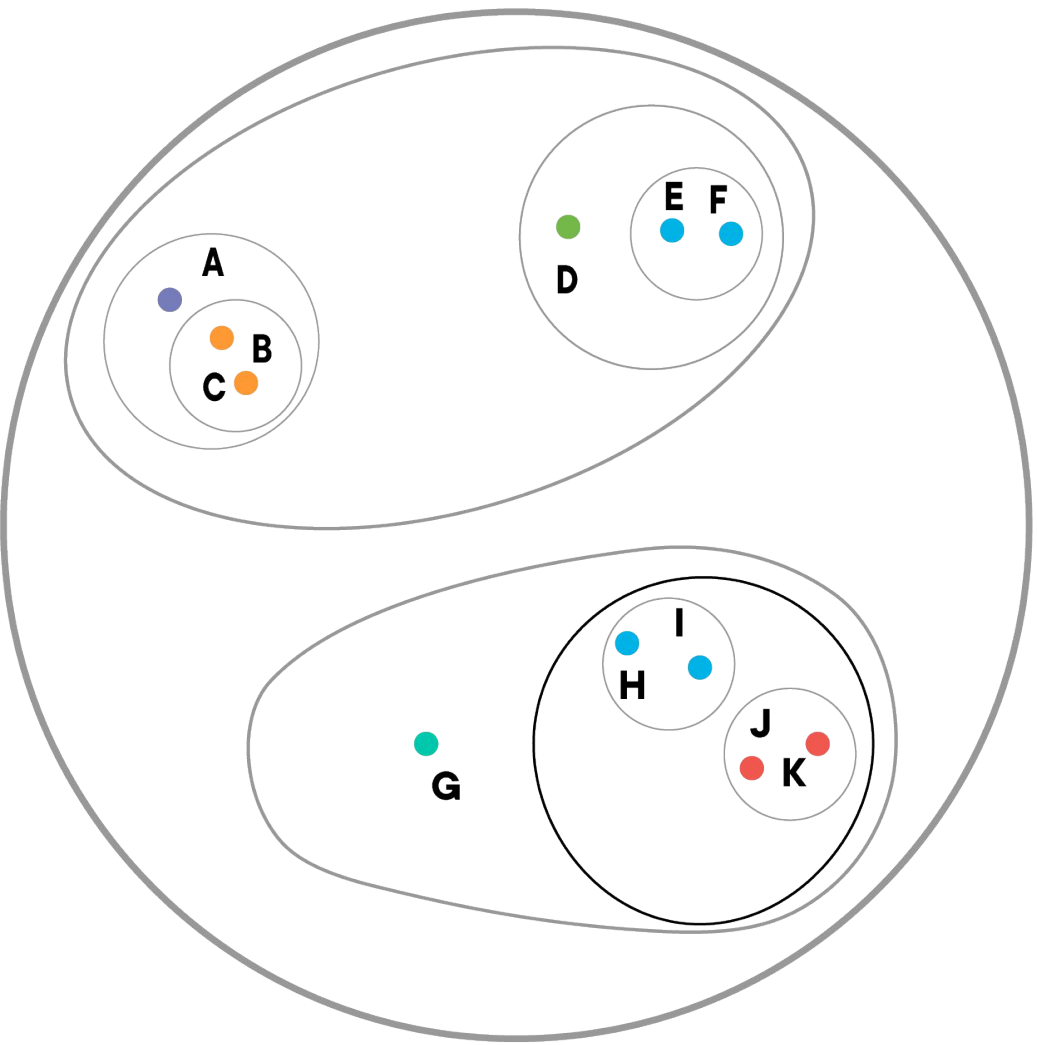
   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

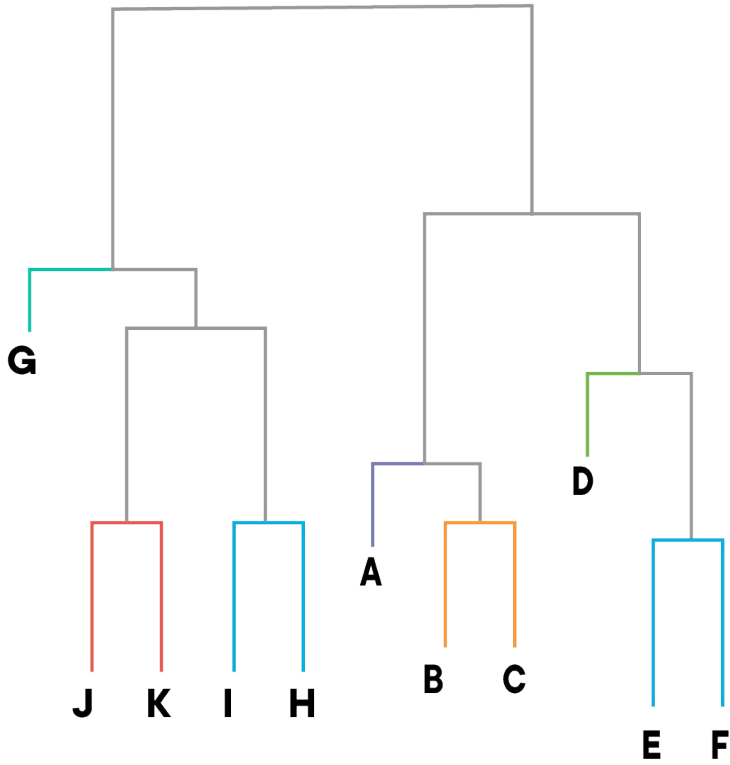
   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

Hierarchical Clustering Algorithm




p0
p1
p2
p3
p4
p5
p6
Hierarchical Clustering Dendrogram
Euclidean Distance
Sample Index


A
B
C
D
E
F
G
H
I
J
K

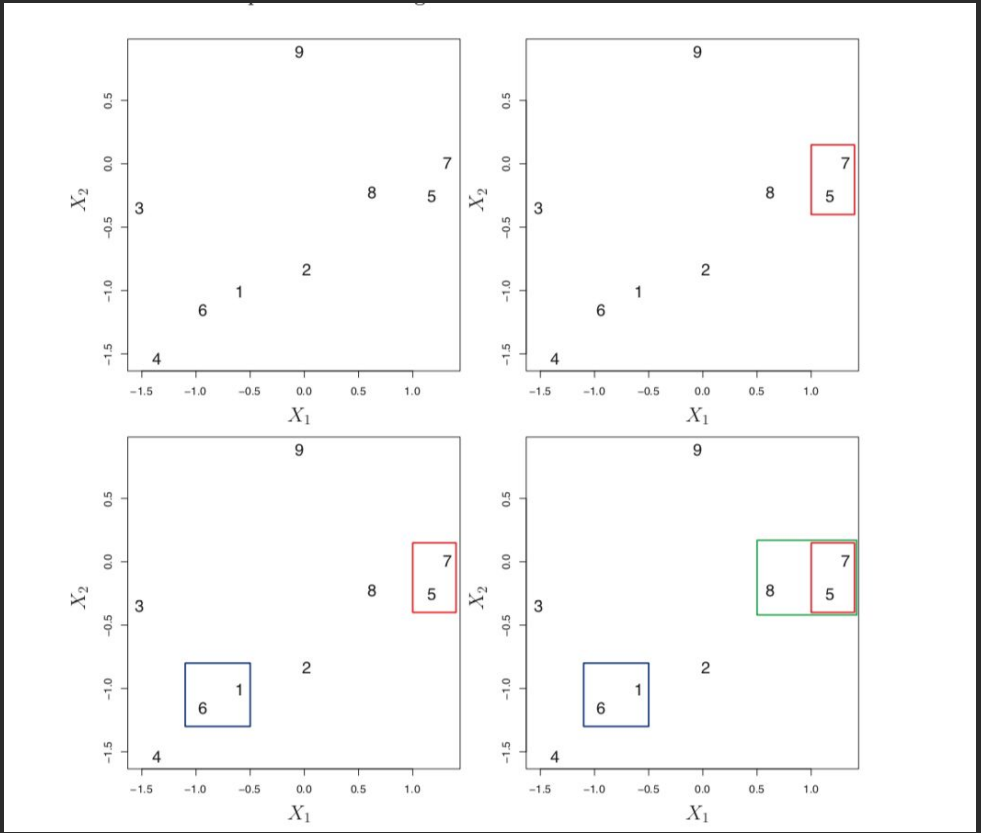

G
J
K
I
H
A
B
C
D
E
F

Sneaky Problems….

Why {5,7} with {8}?

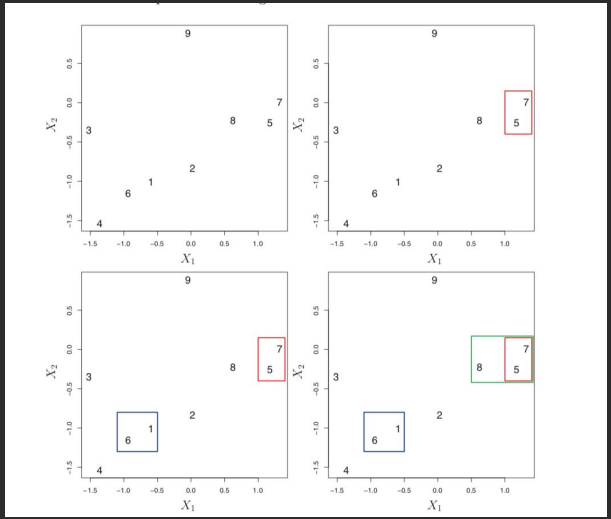
Unequal # obs?

Obs = groups



Linkage:

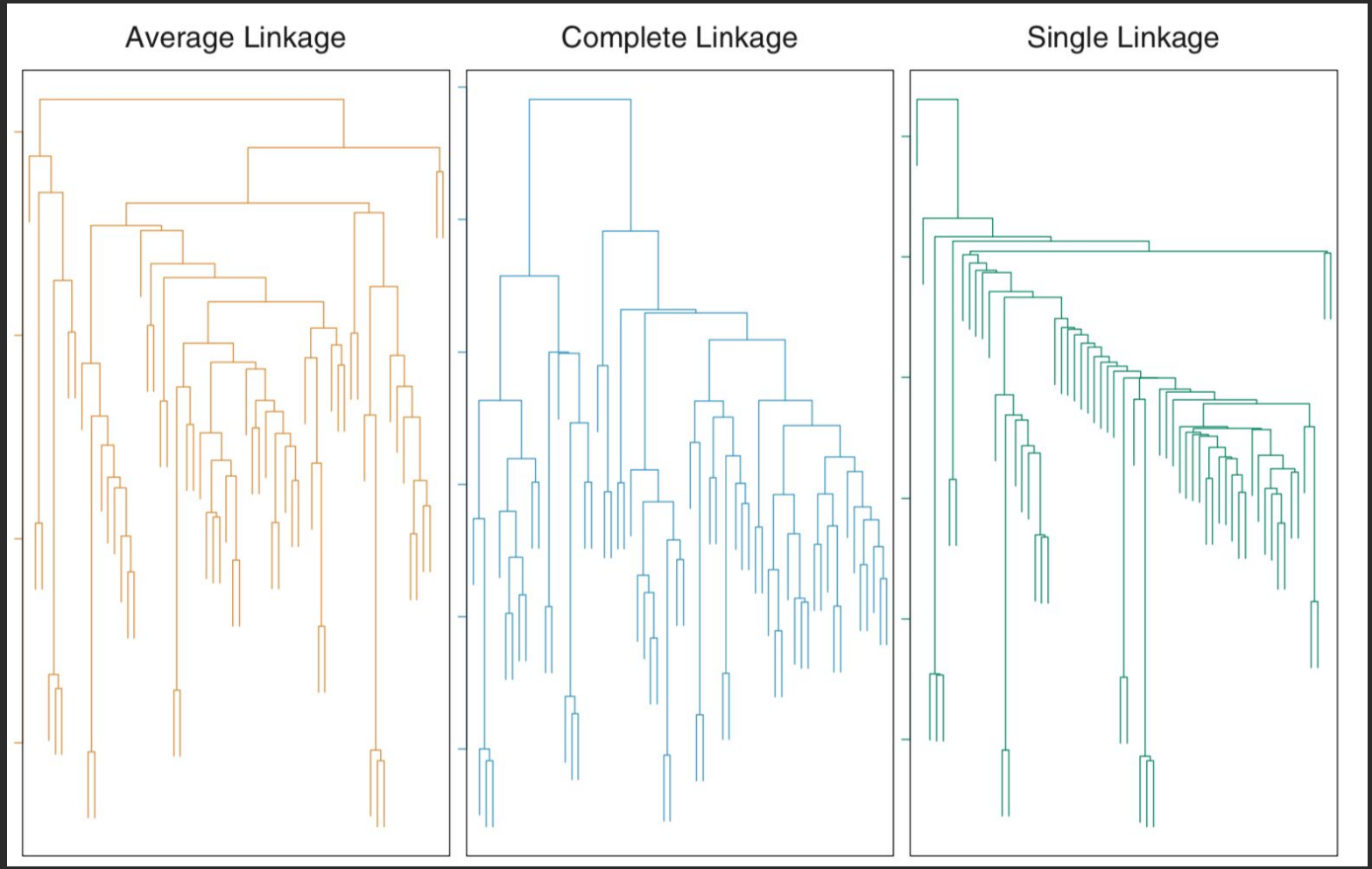
Defines dissimilarity between groups of observations!



Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

More Balanced Dendrograms

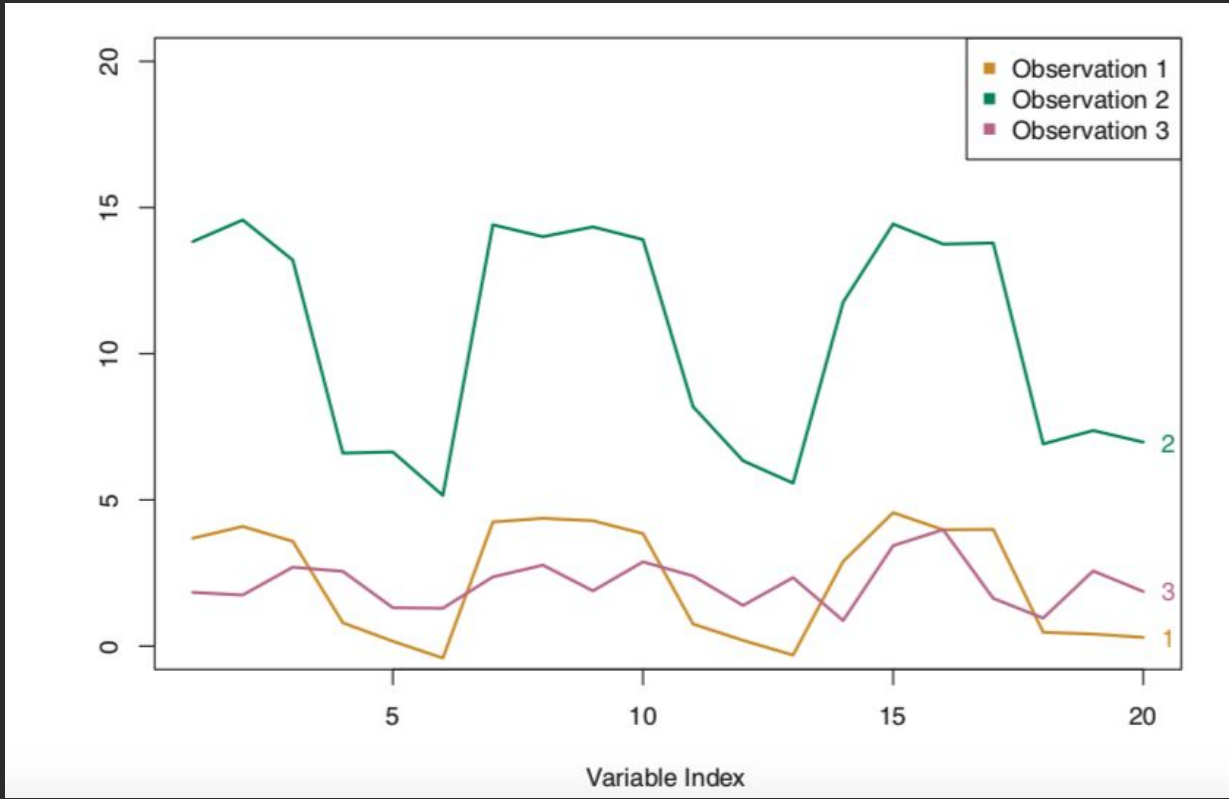

Average Linkage
Complete Linkage
Single Linkage

How dissimilar?

- Euclidean Distance
- Correlation
  - Shapes of profiles rather than magnitude

1 + 3 have small Euclidean Distance, weak correlation

1 + 2 have a similar shape so we can capture this with correlation!



Practical Considerations

- Standardize Variables?
- What dissimilarity metrics?
- What type of linkage?
- Where to cut dendrogram?
- How many clusters (for K means)?

Come up with a practical application of clustering analysis and decide how you would answer the the following questions and say why..

- **Standardize Variables?**
- **What dissimilarity metrics?**
- **What type of linkage?**
- **Where to cut dendrogram?**
- **How many clusters (for K means)?**

Discussion Questions:

Come up with a practical application of clustering analysis and decide how you would answer the the following questions and say why..

How would you go about validating your clusters?

Future Reading

https://scikit-learn.org/stable/modules/clustering.html#k-means

TO THE NOTEBOOK!!