# CSCI 491: Data Visualization

## 16- Clustering (Hierarchical Clustering)

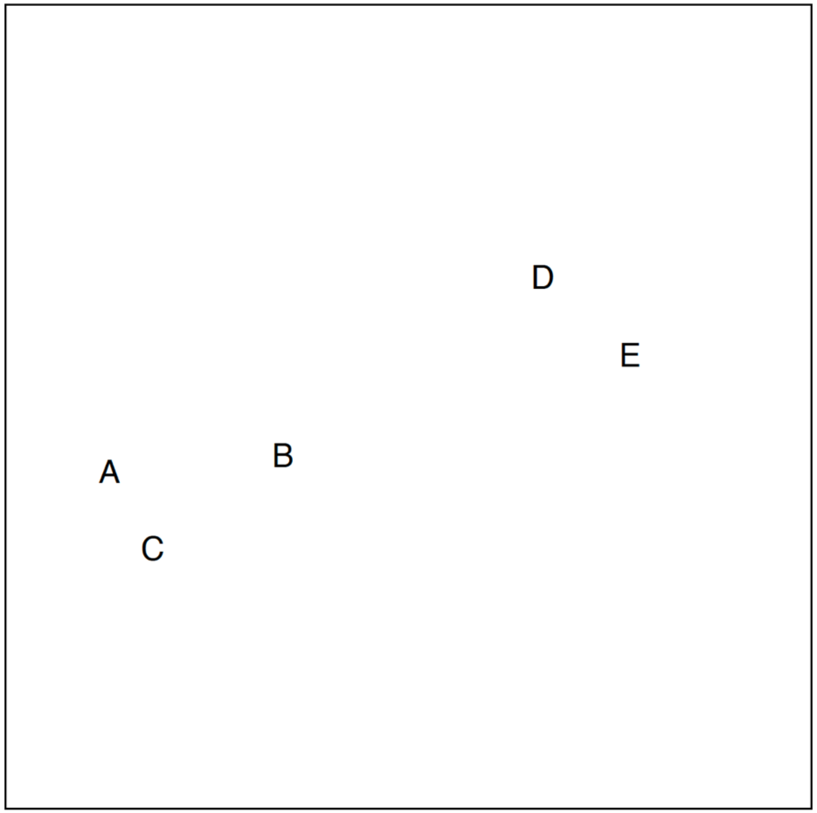
MONTANA STATE UNIVERSITY

# Hierarchical Clustering

- $K$-means clustering requires us to pre-specify the number of clusters $K$. This can be a disadvantage
- *Hierarchical clustering* is an alternative approach which does not require that we commit to a particular choice of $K$.
- In this lecture, we discuss *bottom-up* or *agglomerative* clustering. This is the most common type of hierarchical clustering, and refers to the fact that a dendrogram is built starting from the leaves and combining clusters up to the trunk.

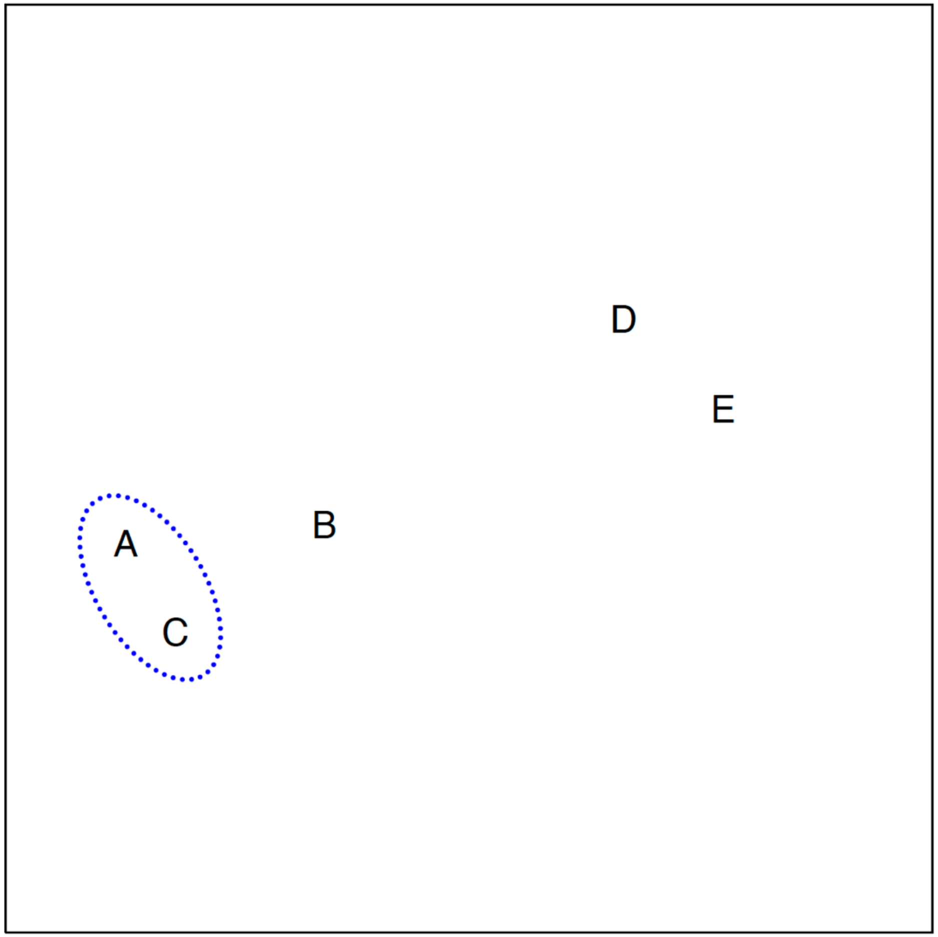
# Hierarchical Clustering: the idea

- Builds a hierarchy in a “bottom-up” fashion...



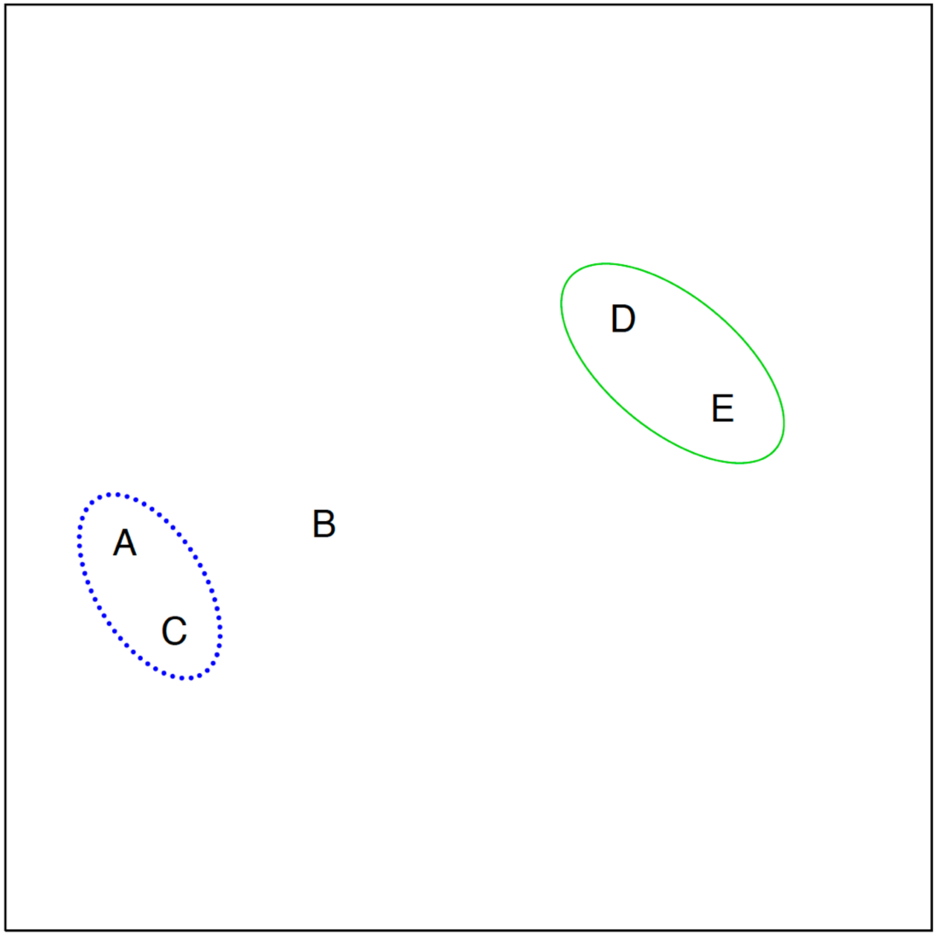
# Hierarchical Clustering: the idea

- Builds a hierarchy in a “bottom-up” fashion...



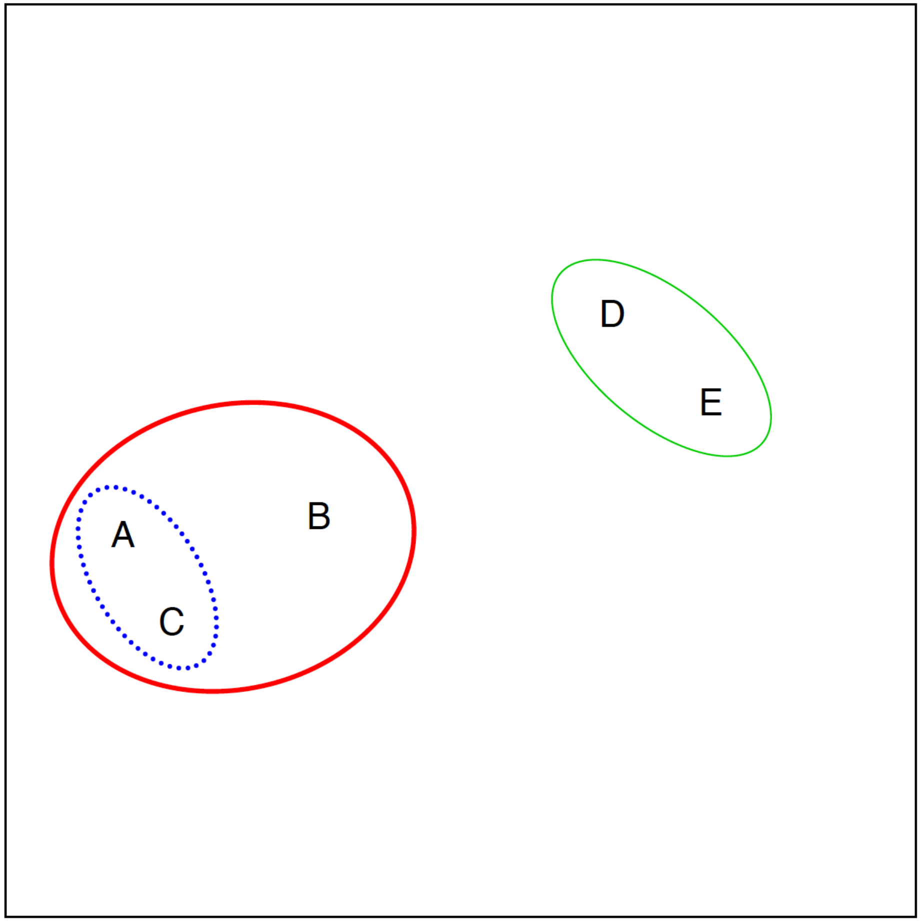
# Hierarchical Clustering: the idea

- Builds a hierarchy in a “bottom-up” fashion...



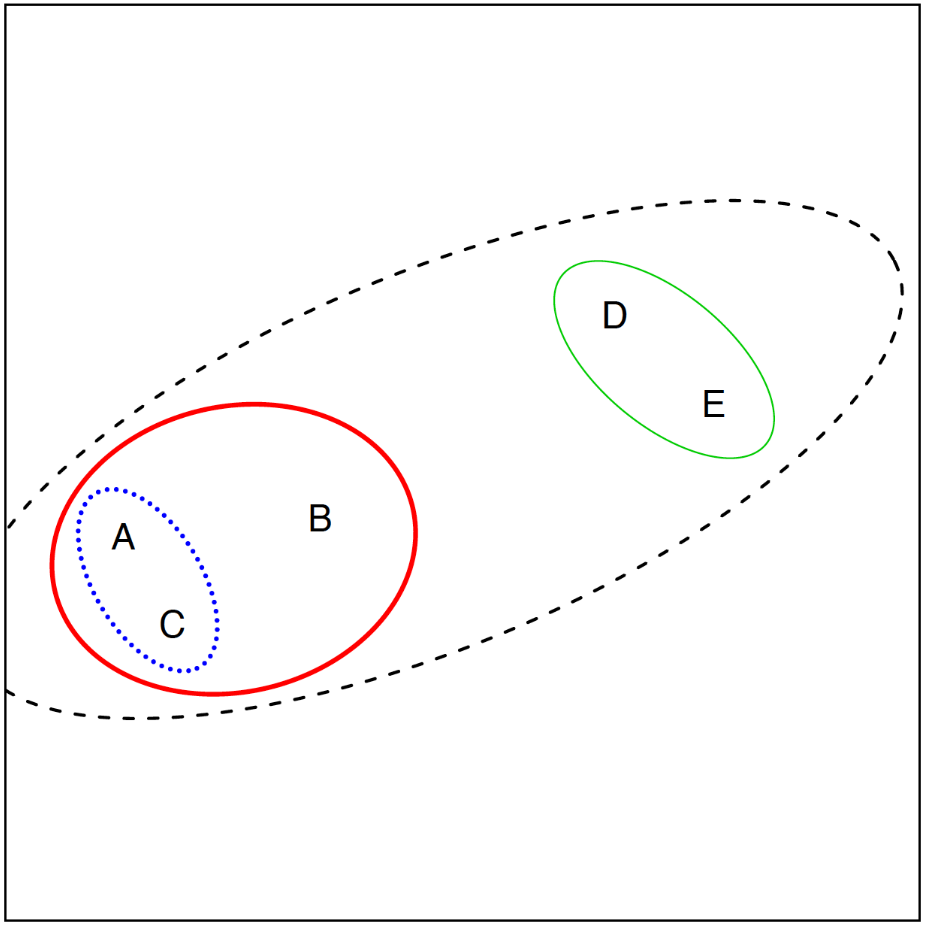
# Hierarchical Clustering: the idea

- Builds a hierarchy in a “bottom-up” fashion...



# Hierarchical Clustering: the idea

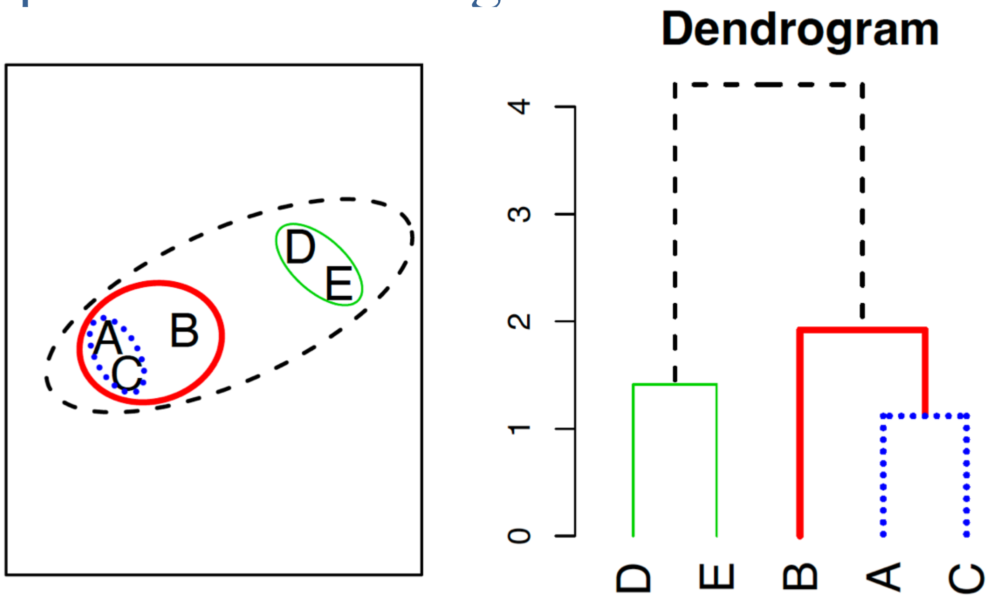
- Builds a hierarchy in a "bottom-up" fashion...



# Hierarchical Clustering Algorithm

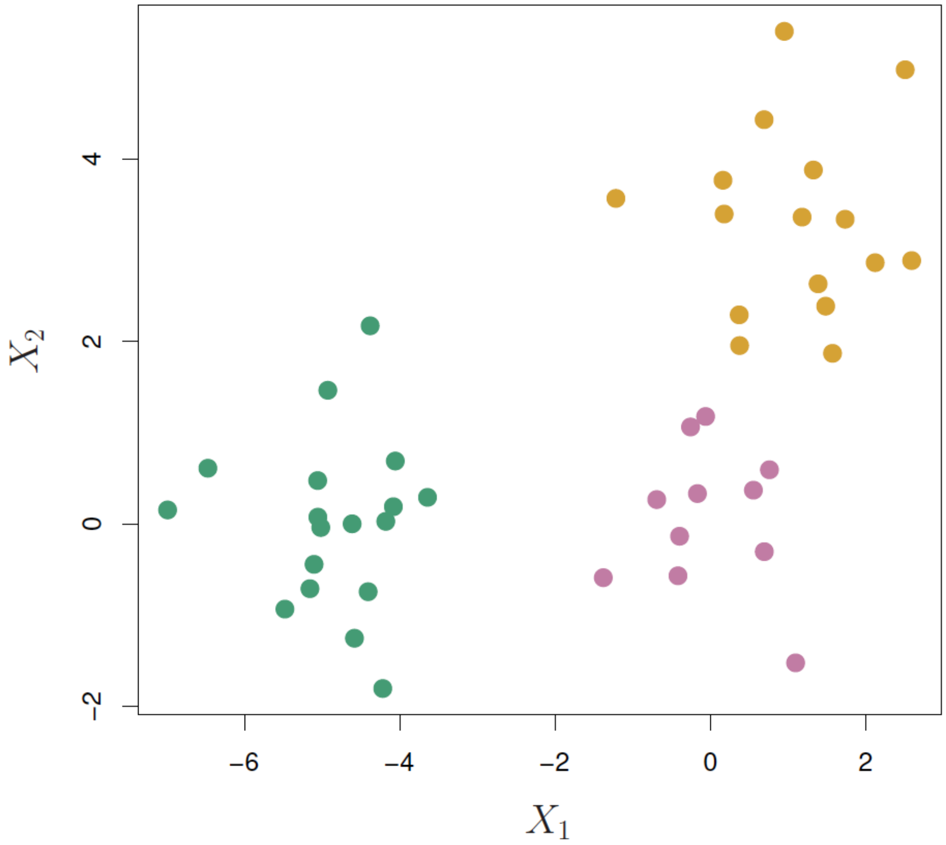
The approach in words:

- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat.
- Ends when all points are in a single cluster.

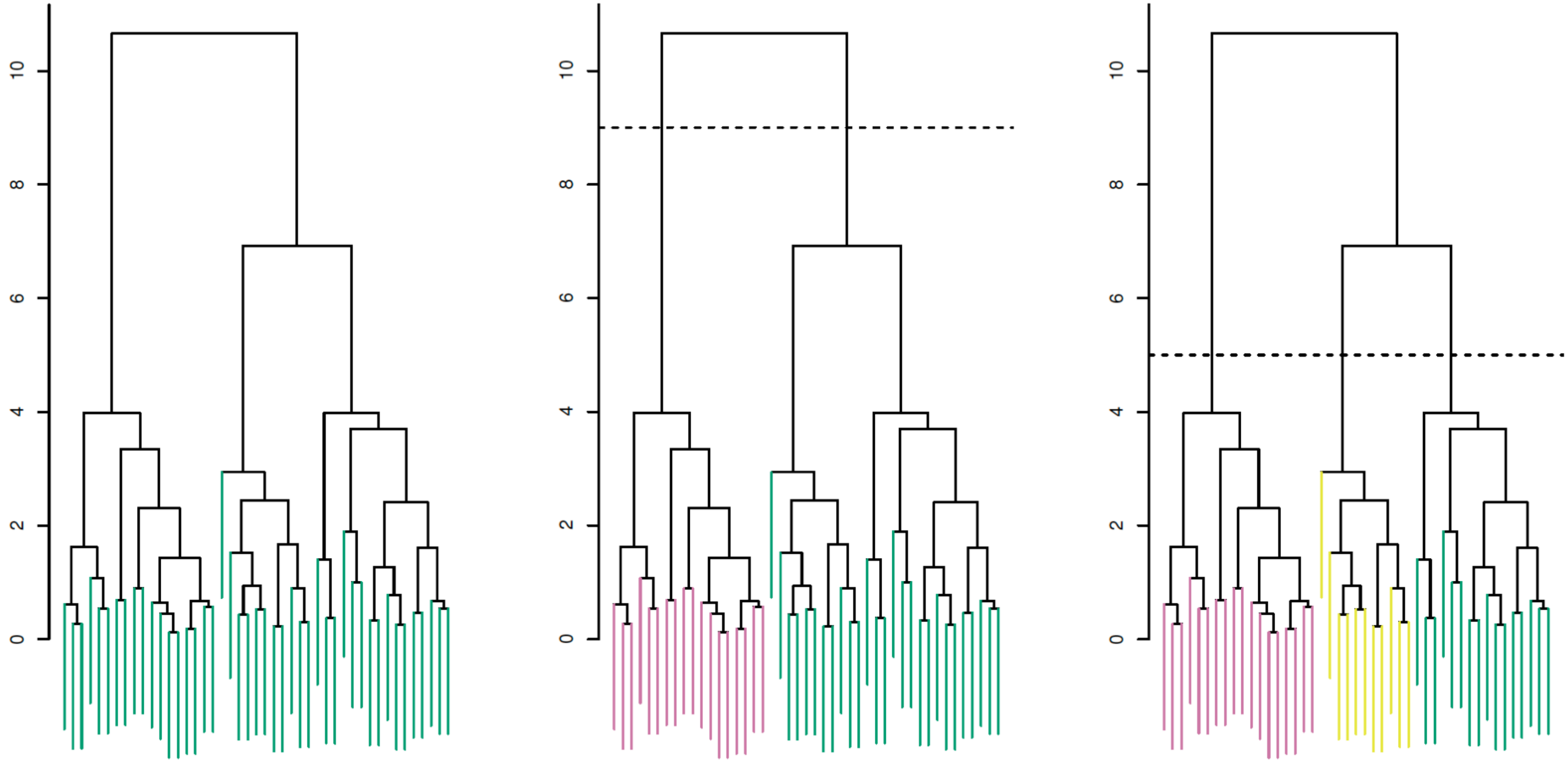


# An Example

45 observations generated in 2-dimensional space. In reality there are three distinct classes, shown in separate colors. However, we will treat these class labels as unknown and will seek to cluster the observations in order to discover the classes from the data.



# Application of hierarchical clustering



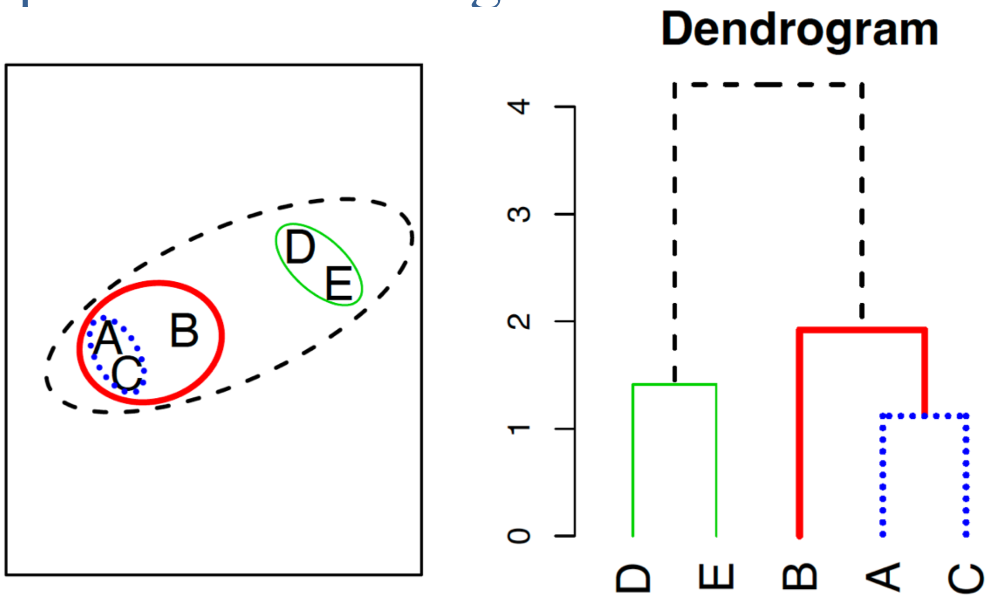
# Details of previous figure

- *Left*: Dendrogram obtained from hierarchically clustering the data from previous slide, with complete linkage and Euclidean distance.
- *Center*: The dendrogram from the left-hand panel, cut at a height of 9 (indicated by the dashed line). This cut results in two distinct clusters, shown in different colors.
- *Right*: The dendrogram from the left-hand panel, now cut at a height of 5. This cut results in three distinct clusters, shown in different colors. Note that the colors were not used in clustering, but are simply used for display purposes in this figure

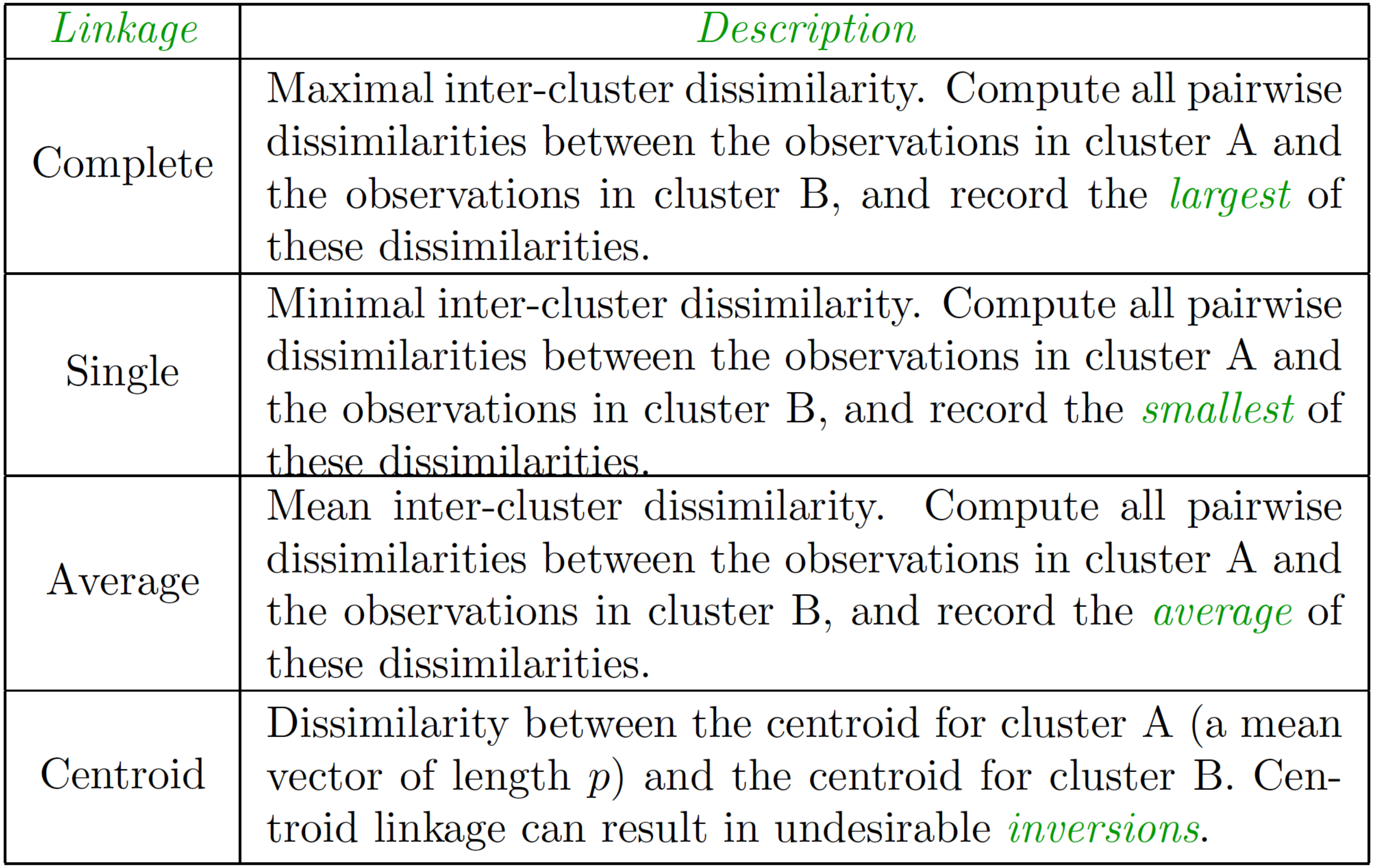
# Hierarchical Clustering Algorithm

The approach in words:

- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat.
- Ends when all points are in a single cluster.



# Types of Linkage



| *Linkage* | *Description* |
|---|---|
| Complete | Maximal inter-cluster 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 inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. |
| Average | Mean inter-cluster 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*. |

# Practical Question

- Do you think scaling of the variables matters?
- Is always the Euclidian distance suitable as dissimilarity measure?
- What type of linkage should be used?
- How many clusters to choose?
- What features should we use to drive the clustering?
- How to validate the clusters?

# Metrics for clustering

- Inertia (Sum of within-cluster-variations):
- The simplest metric for clustering is inertia which is based on the distance from each data point to the center of the cluster

$$i_k = \frac{1}{|C_k|} \sum_{X_j \in C_k} \|X_j - \mu_k\|^2$$

$$I = \sum_k i_k.$$

# Metrics for clustering

- Silhouette: The silhouette coefficient is a clustering validation metric that measures the similarity of an object to its own cluster (cohesion) compared to other clusters (separation).
- For a data point $Xj$, the silhouette coefficient can be computed as follows. Let $aj$ be the mean distance between data point $Xj$ and all the other points in the cluster that $Xj$ belongs to. And let $bj$ be the mean distance between data point $Xj$ and all the points in the next nearest cluster. Then the silhouette coefficient for point $Xj$ is given by:

$$s_j = \frac{b_j - a_j}{\max(a_j, b_j)} .$$

- When this is averaged over all the points in a given cluster, it measures how tightly grouped the points in that cluster are. When it is averaged over all points in the data set, it measures how well the data have been clustered.

# Discussion

- What is the range of possible values for inertia? Does it depend on the scale of the features in the data?
- What is the range of values for the silhouette coefficient? Does this imply advantages or disadvantages between the two metrics?
- When will the silhouette coefficient be equal to 1? When will it be equal to 0?