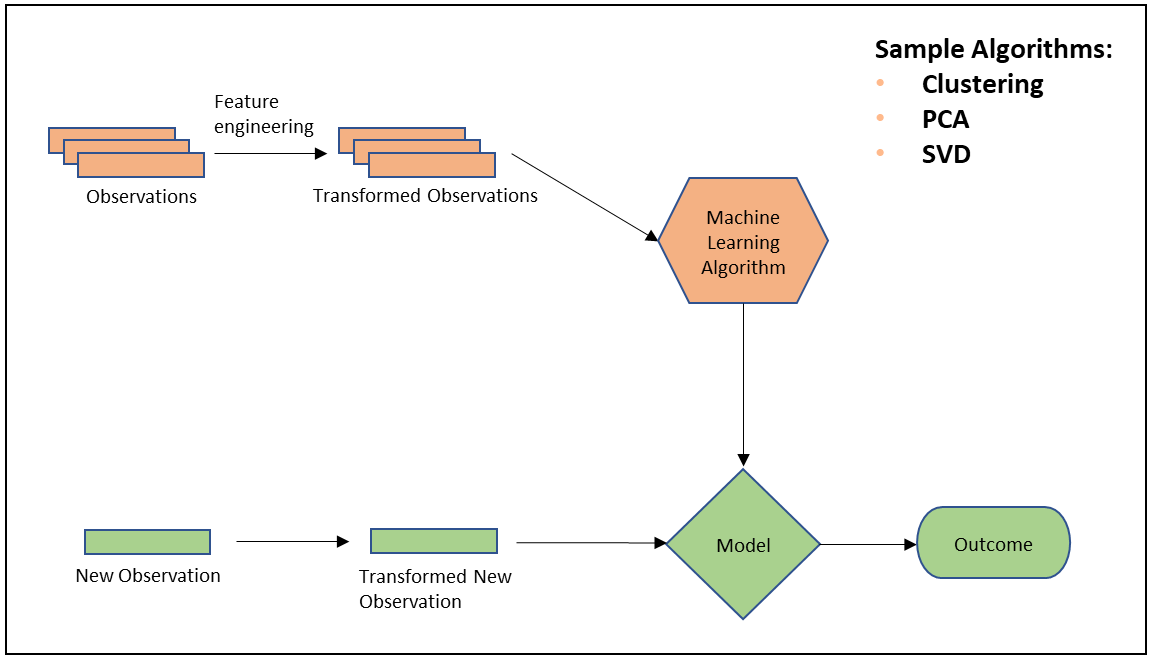
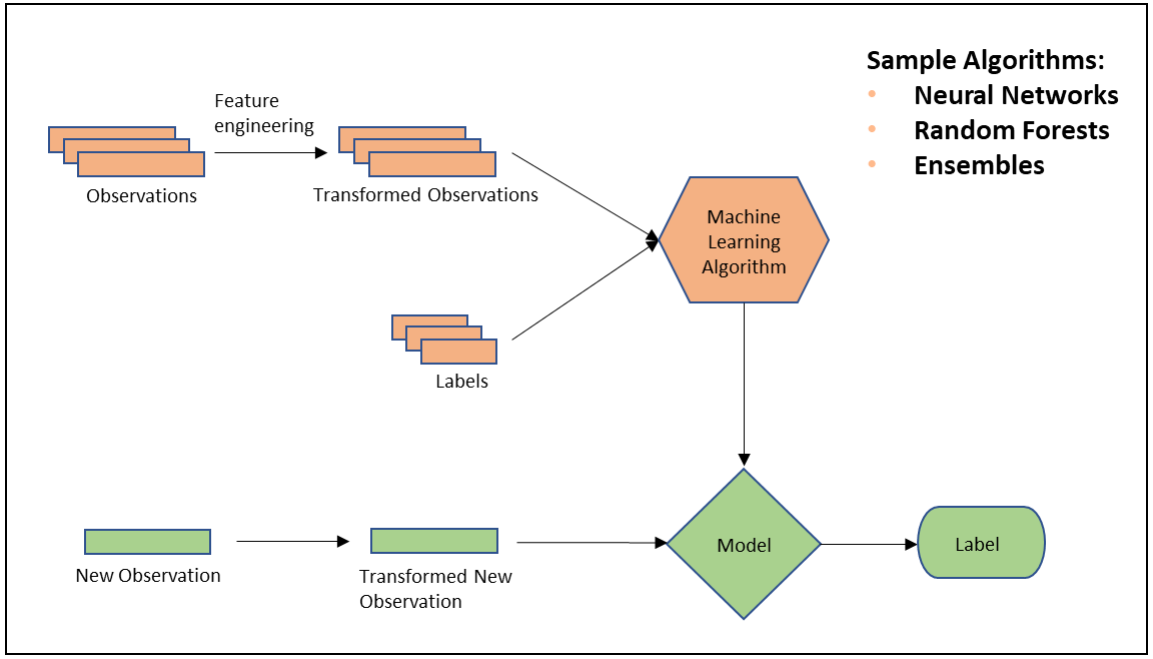
Machine Learning Live
Session #8

Supervised vs. Unsupervised Learning

- Unsupervised learning: no labels associated with observations
  - Try to infer relationships between the observations or between the features
  - Useful for data visualization and dimension reduction
- Supervised learning: each observation is associated with a label
  - Try to infer a relationship between the features and labels
    - The label acts as a teacher that supervises the learning process
  - Use the relationship to predict label for a new observation

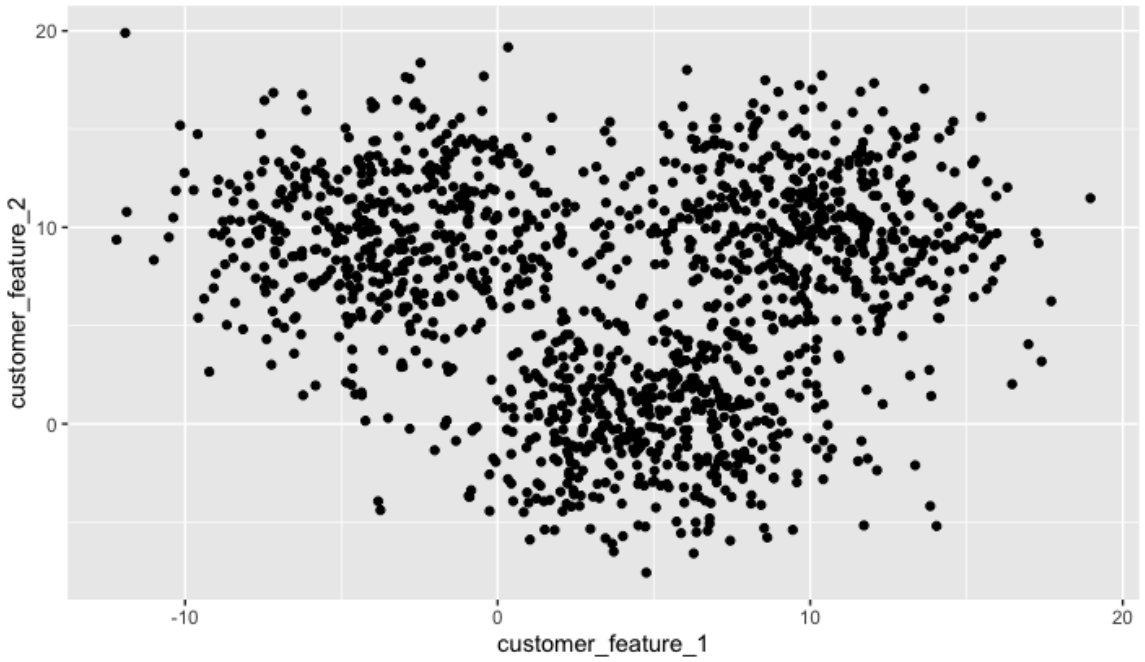


Unsupervised Learning



Supervised Learning

Supervised vs. Unsupervised Learning



Unsupervised Learning

- Observations (here, customers) have no labels
- Use unsupervised learning to explore and learn about customers
  - E.g., do sub-groups of customers exist, with each sub-group exhibiting similar characteristics?
  - This is called customer segmentation

Supervised Learning

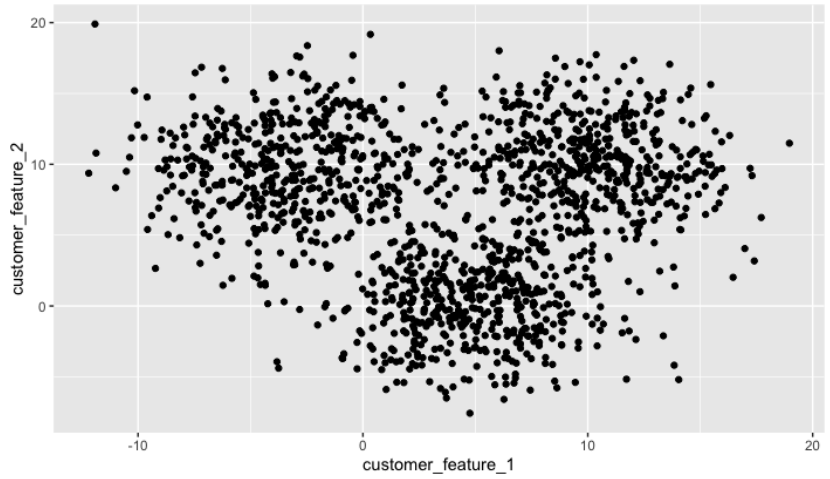
- Each observation (here, customer) is associated with a label
  - E.g., whether the customer left or stayed
- Use supervised learning to predict the label for new customers

Unsupervised Learning

- Unsupervised learning methods try to infer relationships between features or between observations
  - More subjective than supervised learning, since there is no clear objective
- Unsupervised learning is an important step in the machine learning process
  - Exploring and visualizing the data
  - Dimension reduction
- Unsupervised learning methods discussed in this course are:
  - Principal component analysis (PCA)
  - k-means clustering
  - Hierarchical clustering
- **Important: these methods are intended for numerical features only**

Clustering

- Clustering is the process of grouping similar observations in the data
  - Commonly used for customer segmentation
  - Group (i.e., cluster) similar customers together based on their demographics and purchasing behaviors
  - Target each group (i.e., cluster) with a specific marketing campaign

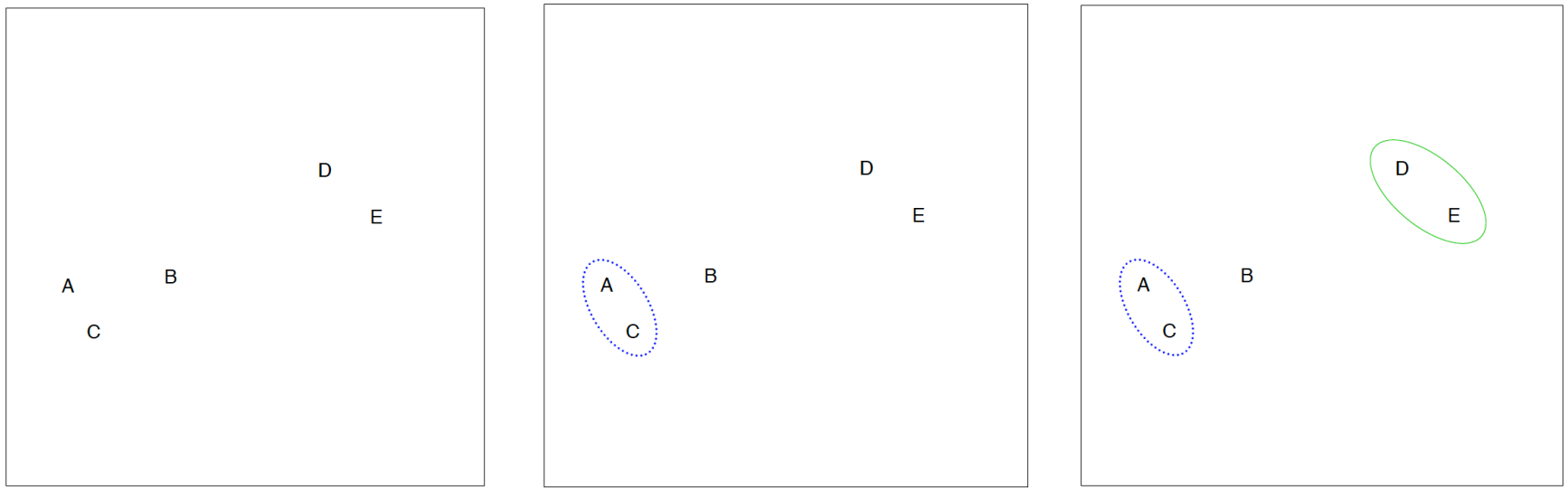


Clustering

- There are two main types of clustering:
  - Hierarchical agglomerative clustering
    - Build the clusters in a bottom-up manner
    - Start with each observation as its own cluster, end with all observations in one cluster
    - Result is a tree-like visualization of the observations, called the dendrogram, that displays all the possible clusters obtained from the hierarchical clustering
  - Partitioning clustering (k-means)
    - Specify the number of clusters (k) in advance

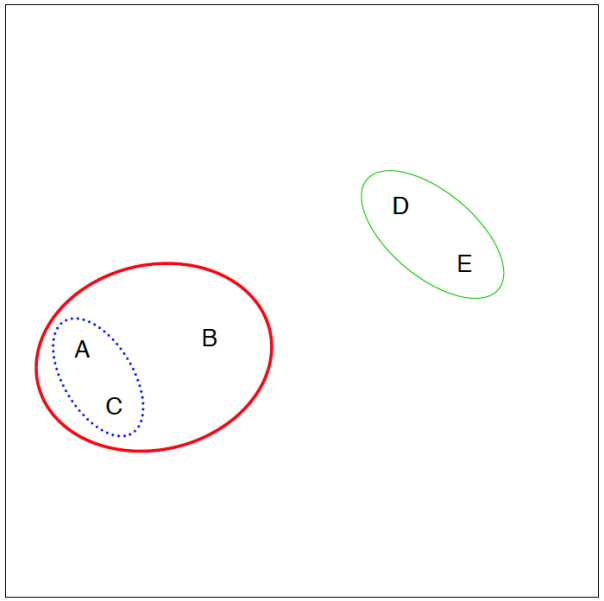
Hierarchical Agglomerative Clustering

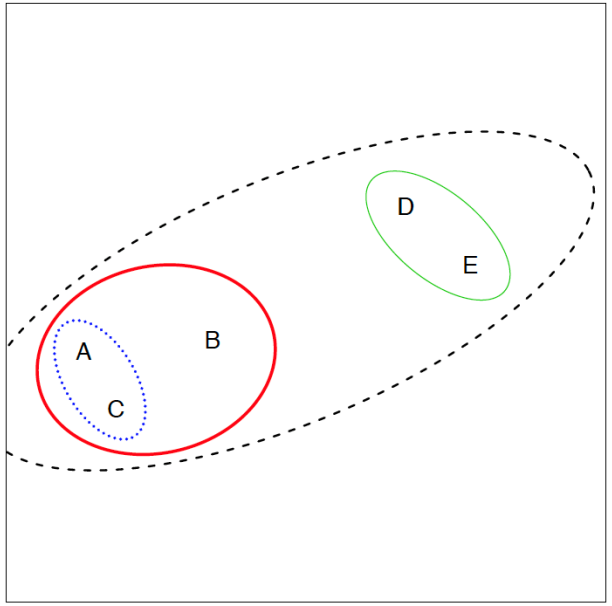
- In hierarchical agglomerative clustering, we build the clusters in a bottom-up approach



Hierarchical Agglomerative Clustering

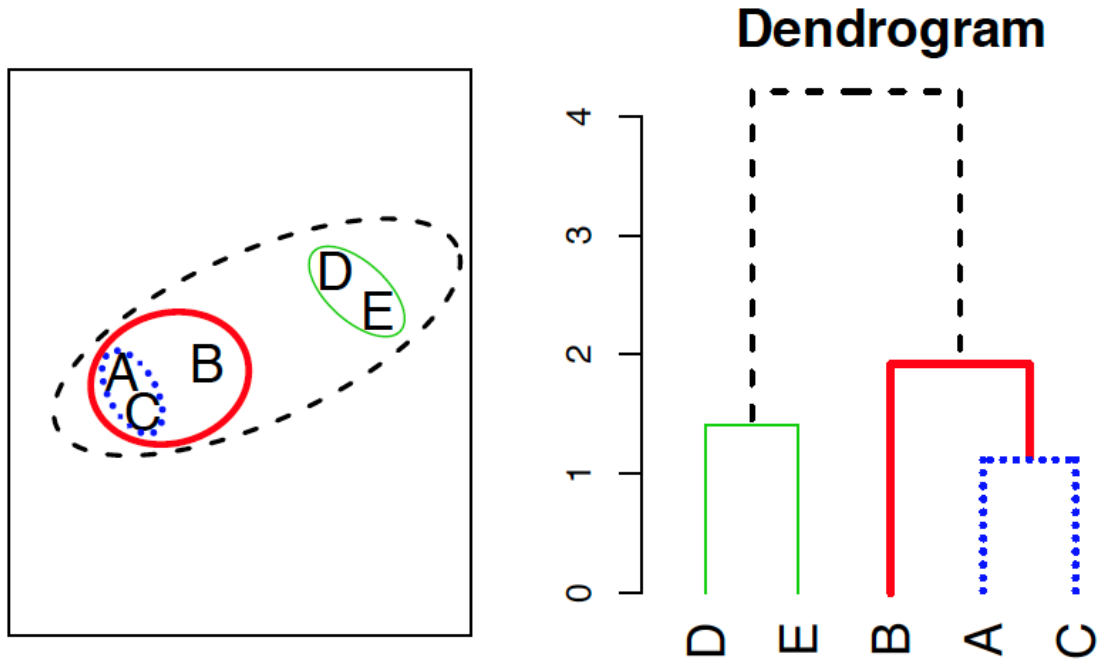
- In hierarchical agglomerative clustering, we build the clusters in a bottom-up approach





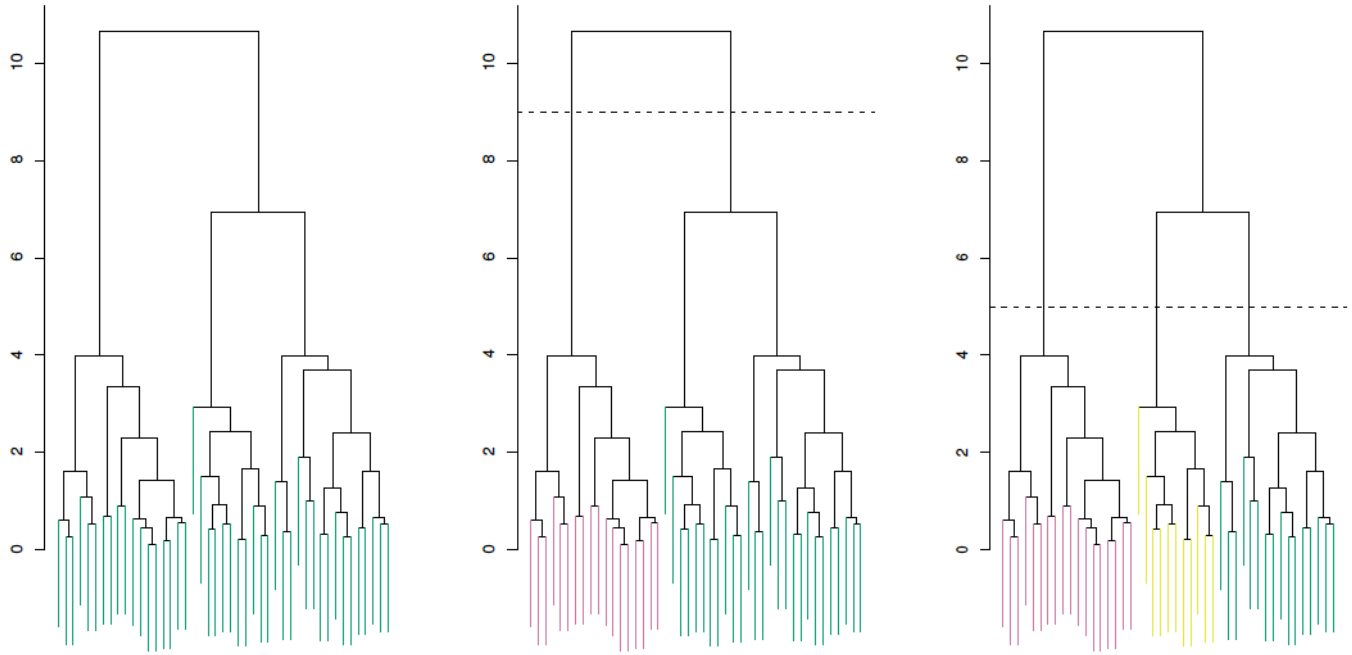
Hierarchical Agglomerative Clustering

- The dendrogram summarizes how the hierarchical agglomerative clustering proceeded



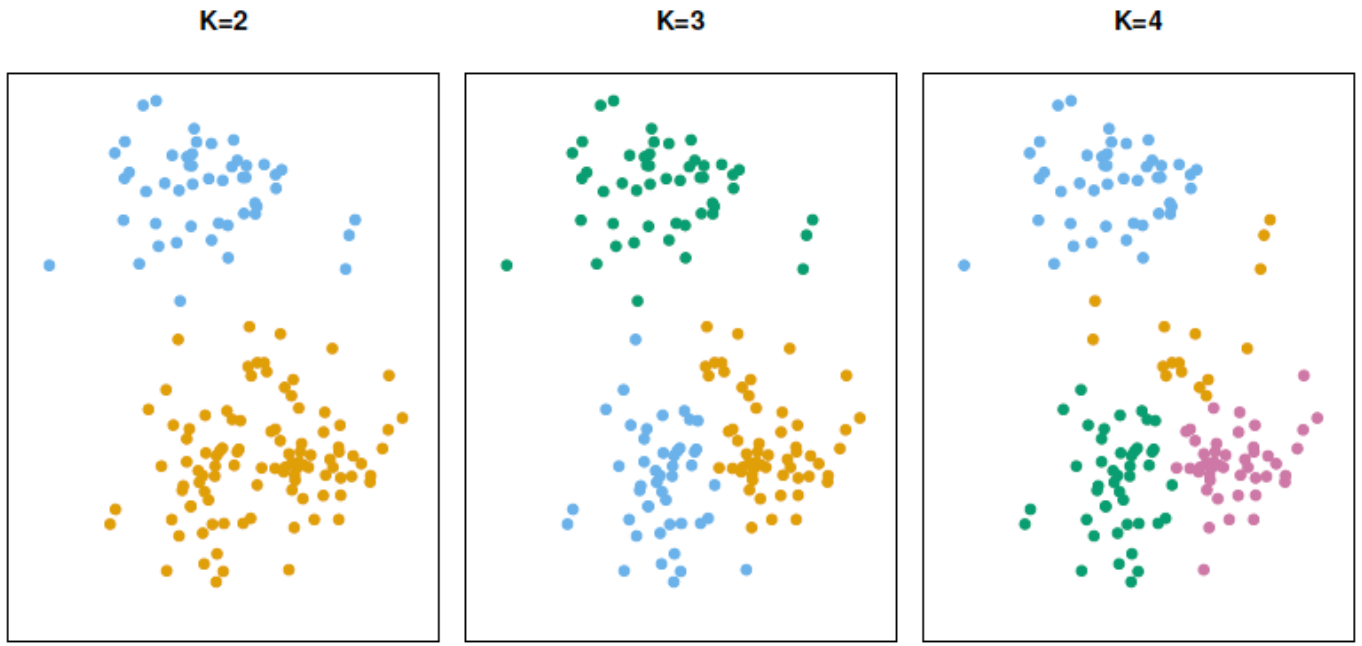
Hierarchical Agglomerative Clustering

- The dendrogram can help us select the number of clusters by specifying the cut-point



Partitioning Clustering

- In partitioning clustering (k-means), we specify the number, k, of clusters we want to identify



Steps of a Cluster Analysis

1. Choose the variables to include in the cluster analysis
2. Identify and remove outliers
3. Scale the data
4. Select a clustering algorithm
5. Evaluate one or more cluster solutions
6. Select and run the final cluster solution
7. Interpret the clusters

Steps of a Cluster Analysis

1. Choose the variables to include in the cluster analysis
   - Requires knowledge of the data
   - Select the variables that may be important for identifying and understanding differences between groups in the data
   - Example: customer segmentation analysis
     - Demographic variables
     - Purchasing behavior

Steps of a Cluster Analysis

2. Identify and remove outliers
   - Univariate outlier: an observation that has an extreme value in one variable
     - Use the "outliers" R package
   - Multivariate outlier: an observation that has extreme values in at least two variables
     - Use the "mvoutlier" R package

Steps of a Cluster Analysis

3. Scale the data
   - Scale each variable to have a mean of 0 and a standard deviation of 1
     - Otherwise, variables with larger magnitudes may have more influence on the clusters, since they will dominate the distance calculation
   - In R, we use “scale(data)”

income	age
50000	28
75000	46
47000	34
100000	42
65000	56
90000	39
56000	25

income_scaled	age_scaled
-0.9368211	-0.9874276
0.29583823	0.69386807
-1.0847402	-0.4269957
1.52849753	0.3202468
-0.1972255	1.62792125
1.03543381	0.04003085
-0.6409828	-1.2676436

Steps of a Cluster Analysis

4. Select a clustering algorithm
   - Hierarchical agglomerative clustering is better suited for smaller problems
     - It is difficult to visualize the dendrogram for larger problems
   - Partitioning clustering can handle larger problems
     - Need to choose number of clusters (k) beforehand
   - Can try both (with several variations for each) and compare results
     - This is often done to see gauge the robustness of the results

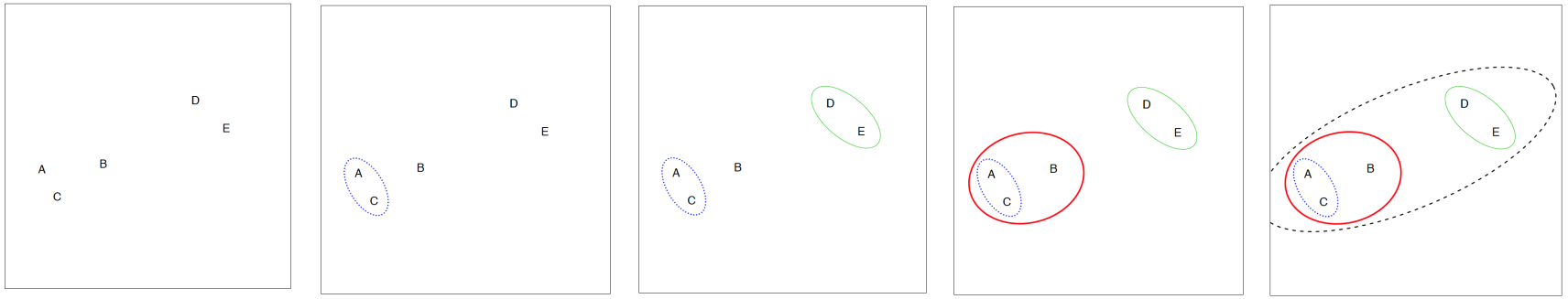
Steps of a Cluster Analysis

4. Select a clustering algorithm

- Need to specify parameters of the chosen clustering algorithm
  - E.g., need to specify definition of distance between two clusters for hierarchical clustering (see below)
  - E.g., need to specify k for agglomerative (k-means) clustering

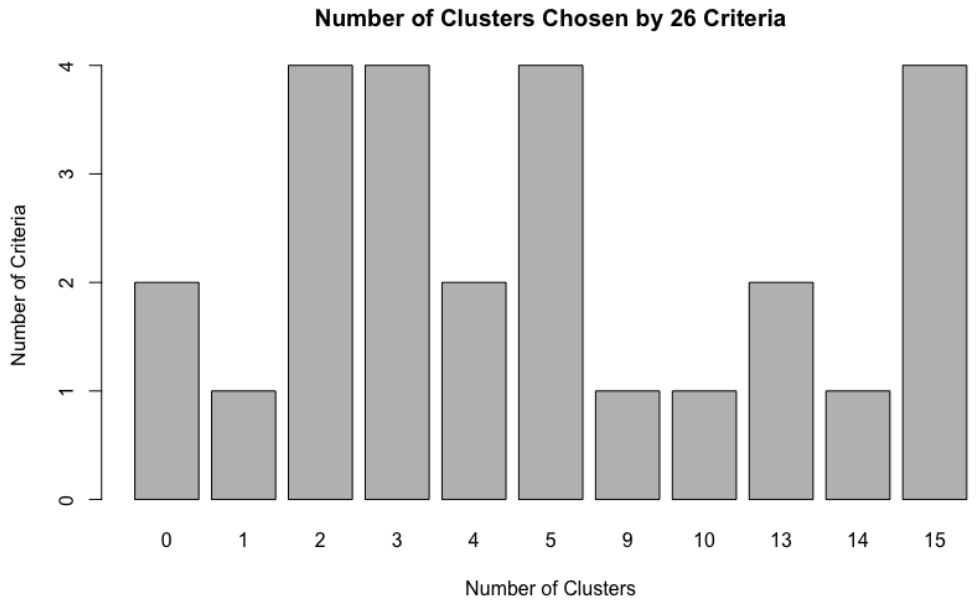
Table 16.1: Hierarchical clustering methods

Cluster method	Definition of the distance between two clusters
Single linkage	Shortest distance between a point in one cluster and a point in the other cluster.
Complete linkage	Longest distance between a point in one cluster and a point in the other cluster.
Average linkage	Average distance between each point in one cluster and each point in the other cluster (also called UPGMA [unweighted pair group mean averaging]).
Centroid	Distance between the centroids (vector of variable means) of the two clusters. For a single observation, the centroid is the variable's values.
Ward	The ANOVA sum of squares between the two clusters added up over all the variables.



Steps of a Cluster Analysis

5. Evaluate one or more cluster solutions
   - Use NbClust() to evaluate cluster solutions specified using a range (minimum and maximum number of clusters to evaluate)



Steps of a Cluster Analysis

6. Select and run the final cluster solution
   - Use hclust() for hierarchical clustering
     - Use cutree() to obtain the number of clusters selected from NbClust()
   - Use kmeans() for k-means clustering
     - Specify the “centers” argument to be the number of clusters selected from NbClust()

Steps of a Cluster Analysis

7. Interpret the clusters
   - Use the aggregate() function to aggregate the rows in the data by their assigned clusters