Clustering

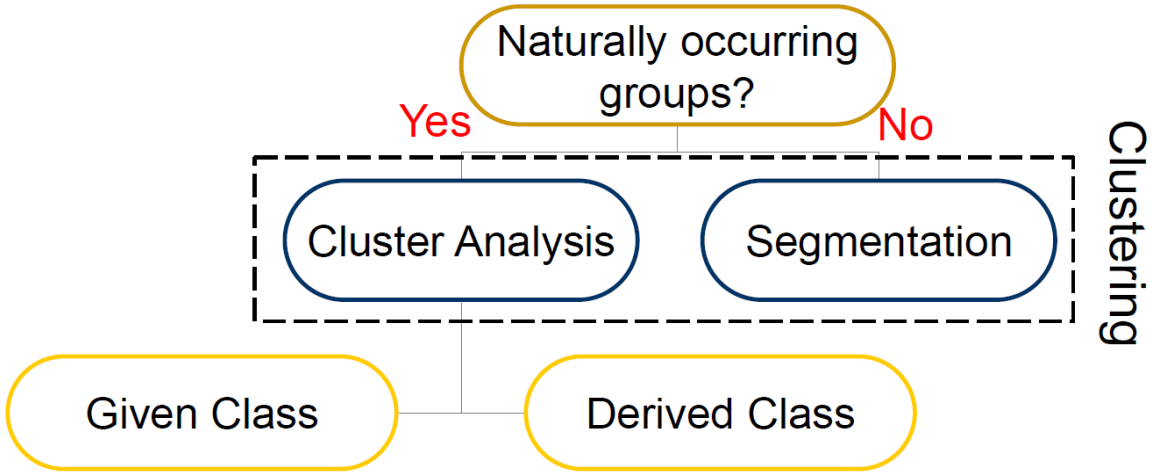
DSA 6000: Data Science and Analytics, Fall 2019

Wayne State University

Cluster Analysis

- **Cluster analysis** is a set of methods for constructing a (hopefully) sensible and informative classification of an initially unclassified set of data, using the variable values observed on each individual.

  Everitt (1998), *The Cambridge Dictionary of Statistics*



The goal of clustering is to partition data into groups so that the observations within a group are as similar as possible to each other, and as dissimilar as possible to the observations in other groups.

Pattern Discovery

- Clustering is searching for patterns in complex data.
- Patterns can lead to business decisions.
- Are there demographic characteristics to identify people who are more likely to preorder books at a premium price point?
- What kinds of people are most likely to be at the food court on a Saturday afternoon?
- What sorts of complaints are most common for different call centers?

Example: Customer Types

- While you have thousands of customers, there are really only a handful of major types into which most of your customers can be grouped.
  - ❑Bargain hunter
  - ❑Man/woman on a mission
  - ❑Impulse shopper
  - ❑Weary parent
  - ❑DINK (dual income, no kids)

Example: Fraud Detection

- Most fraudulent customer activity is difficult to identify by a single variable.
- Are there unusual combinations of behaviors that can help identify criminal activity or fraud?
  - Spending $250 on shoes is not unusual.
  - An online purchase by Johnny is not unusual.
  - Purchases in New York by Johnny are not unusual, although Johnny lives in Detroit.
  - Johnny buying $250 shoes online while he is in New York, that is unusual. Fraud alert!

Example: Store Location

- You want to open new grocery stores in the U.S. based on demographics. Where should you locate the following types of new stores?
  - ❑low-end budget grocery stores
  - ❑small boutique grocery stores
  - ❑large full-service supermarkets

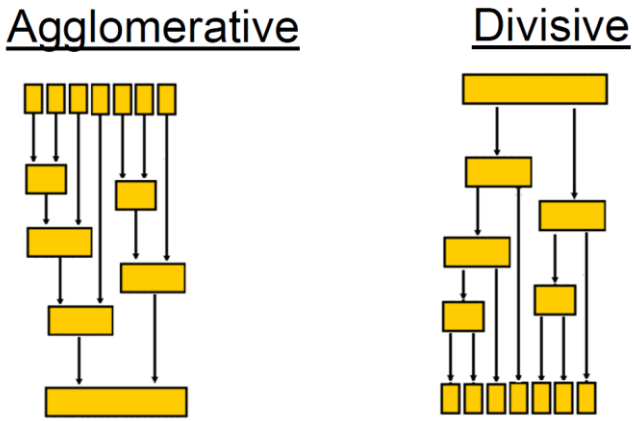
Example: Fashion Trends

- Based on the four styles of pants that your customers can purchase, can you identify stores as serving similar fashion types?
  - country-club dresser
  - fashion trendsetter
  - comfort kick-back dresser
- **Cluster profiling** is the derivation of a class label from a proposed cluster solution.
- The objective is to identify the combination of features that uniquely describe each cluster.

Types of clustering

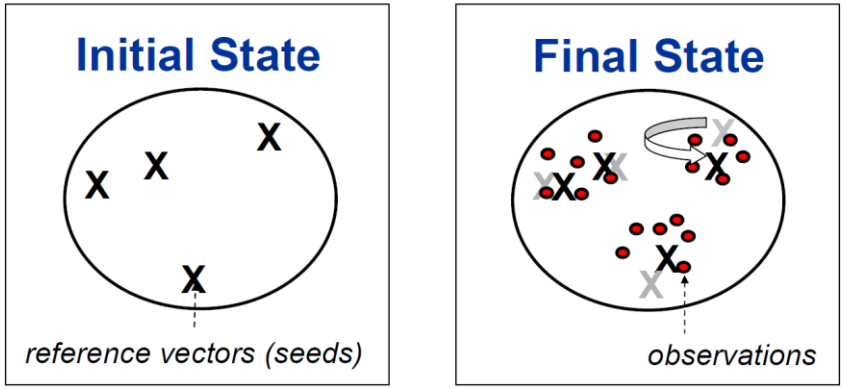
- Two major classes of clustering methods:
  - **Hierarchical clustering** (the backbone of cluster analysis)
  - Partitive (i.e., optimization) clustering (less used)

Hierarchical Clustering



- Do not scale up well.
- Previous steps are irrevocable: errors will prorogate.

Partitive Clustering



- Scale up well.
- Need user to guess the # of clusters
- Influenced by seed, outliers, order of the obs.