Clustering methods

Mate Akos

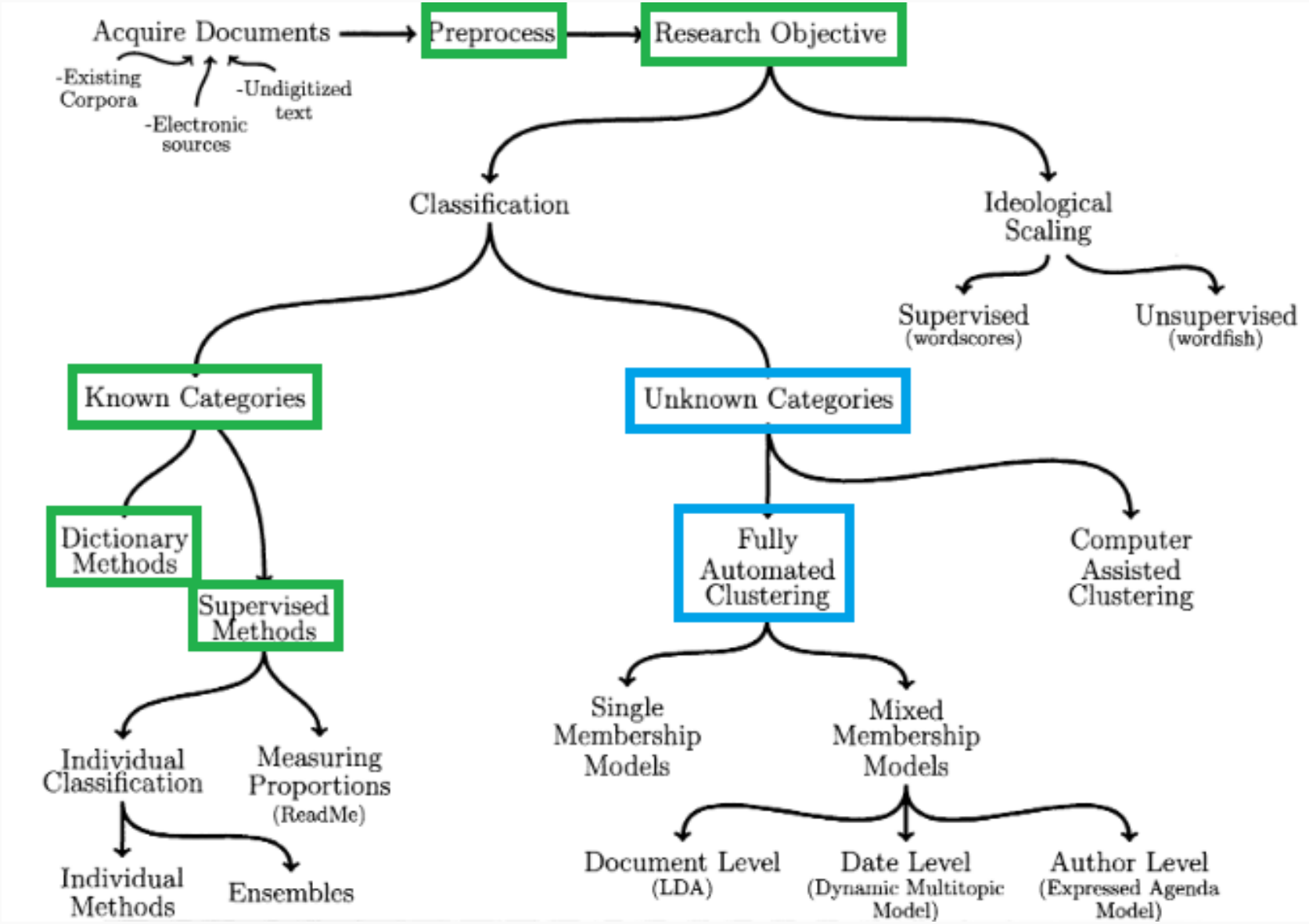
2019 October

TK PTI

For today

1. Unsupervised methods
2. Clustering: Key concepts
3. K-means
4. Hierarchical clustering
5. Challanges

Where we are


Acquire Documents
-Existing Corpora
-Electronic sources
-Undigitized text
Preprocess
Research Objective
Classification
Ideological Scaling
Supervised (wordscores)
Unsupervised (wordfish)
Known Categories
Unknown Categories
Dictionary Methods
Supervised Methods
Fully Automated Clustering
Computer Assisted Clustering
Individual Classification
Measuring Proportions (ReadMe)
Single Membership Models
Mixed Membership Models
Individual Methods
Ensembles
Document Level (LDA)
Date Level (Dynamic Multitopic Model)
Author Level (Expressed Agenda Model)

Refresher: unsupervised learning

Unsupervised learning:

- **Clustering problems**, we don't know the number of categories of our response variable
- The clustering algorithm finds clusters in the data without any "supervision"
- We don't have any labelled data
- Useful in exploratory analysis or discovering clusters in our data
- We might never know the "true" number of clusters

Clustering

What is clustering?

- finding **subgroups** or **clusters** in the data

Clustering

What is clustering?

- finding **subgroups** or **clusters** in the data
- Objective: partition the data into distinct groups (these subgroups are homogeneous)

Clustering

What is clustering?

- finding **subgroups** or **clusters** in the data
- Objective: partition the data into distinct groups (these subgroups are homogeneous)
- Problem: unknown number of such clusters

Clustering

What is clustering?

- finding **subgroups** or **clusters** in the data
- Objective: partition the data into distinct groups (these subgroups are homogeneous)
- Problem: unknown number of such clusters
- Essentials: similarity or dissimilarity measures between our data points

Clustering

What is clustering?

- finding **subgroups** or **clusters** in the data
- Objective: partition the data into distinct groups (these subgroups are homogeneous)
- Problem: unknown number of such clusters
- Essentials: similarity or dissimilarity measures between our data points
- Use cases:

Clustering

What is clustering?

- finding **subgroups** or **clusters** in the data
- Objective: partition the data into distinct groups (these subgroups are homogeneous)
- Problem: unknown number of such clusters
- Essentials: similarity or dissimilarity measures between our data points
- Use cases:
  - business intelligence (costumer categories)
  - biology
  - text analysis
  - web search
  - etc.

Clustering techniques

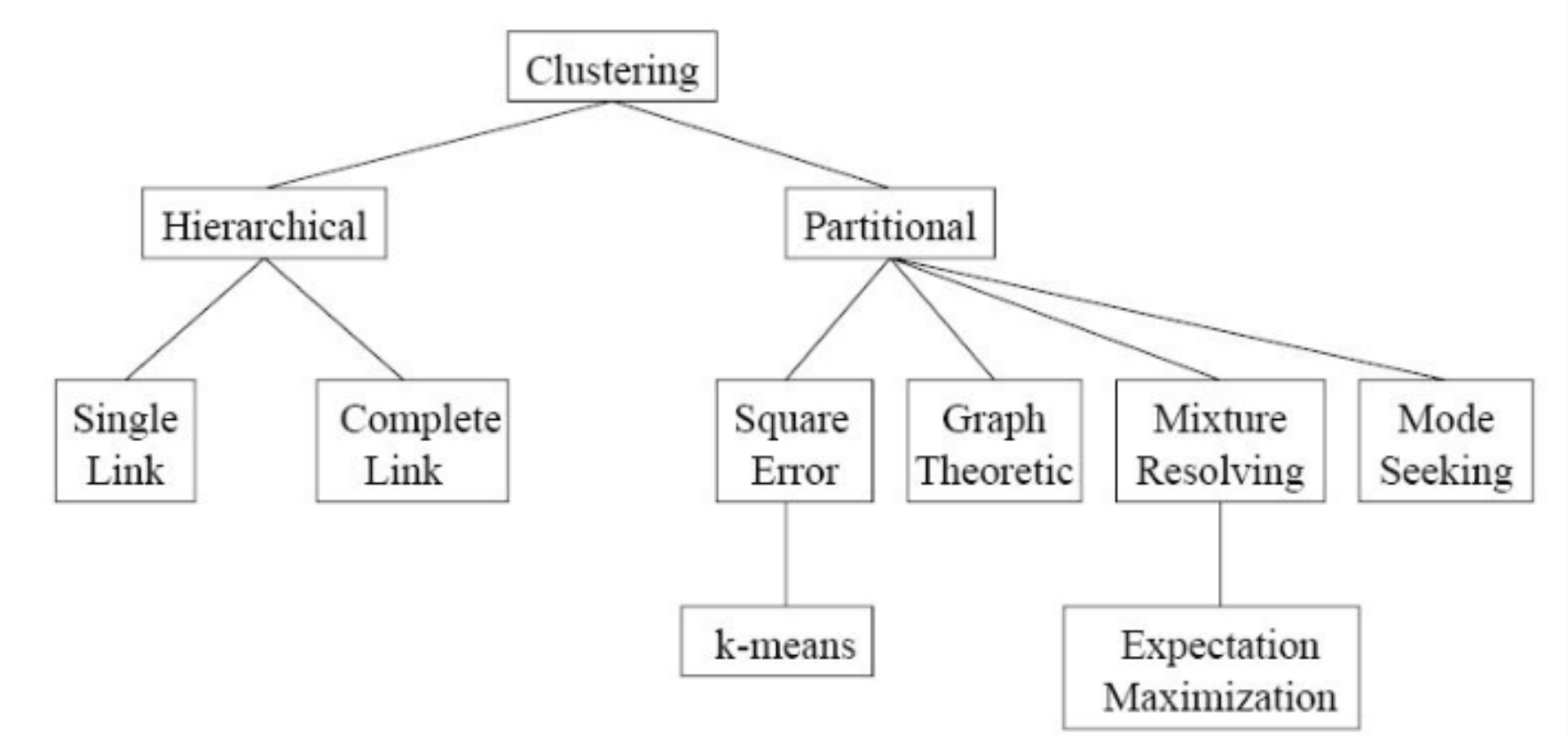


Figure 2.1: Classification of clustering methods.

Galpin, Ixent, et al. "Data requirements, data management and analysis issues, and query-based functionalities." Deliverable D2 1 (2009).

Clustering methods

Our focus in this session:

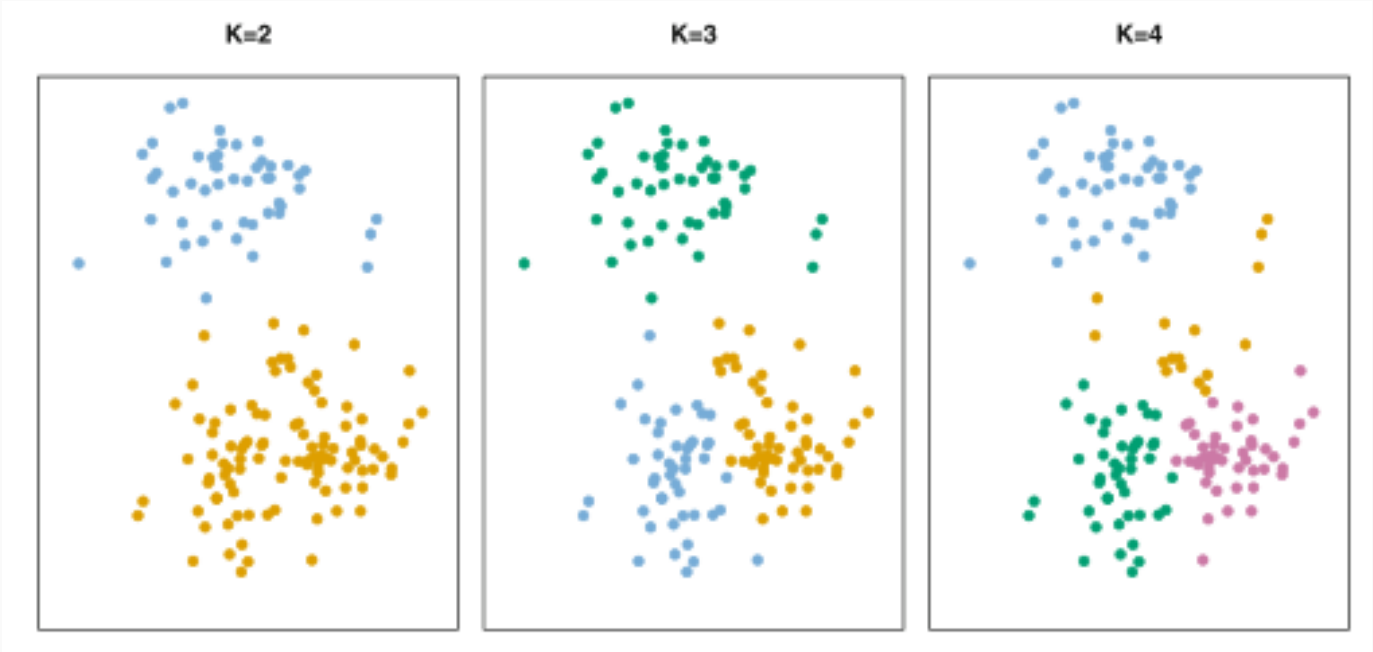
- K-means clustering
- Hierarchical clustering

K-means clustering

Simple method for finding K clusters in our data

- K is defined by the researcher
- Each observation gets classified into **one** category
- Goal: minimize within cluster variation and maximize between cluster differences
    - Closest point: minimizing the sum of squared errors: $SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2$
- Class labels are the result of the clustering
- Computationally cheap algorithm: O(n) time complexity

K-means clustering example



James, Witten, Hastie and Tibshirani (2017), p.387

K-means clustering algorithm

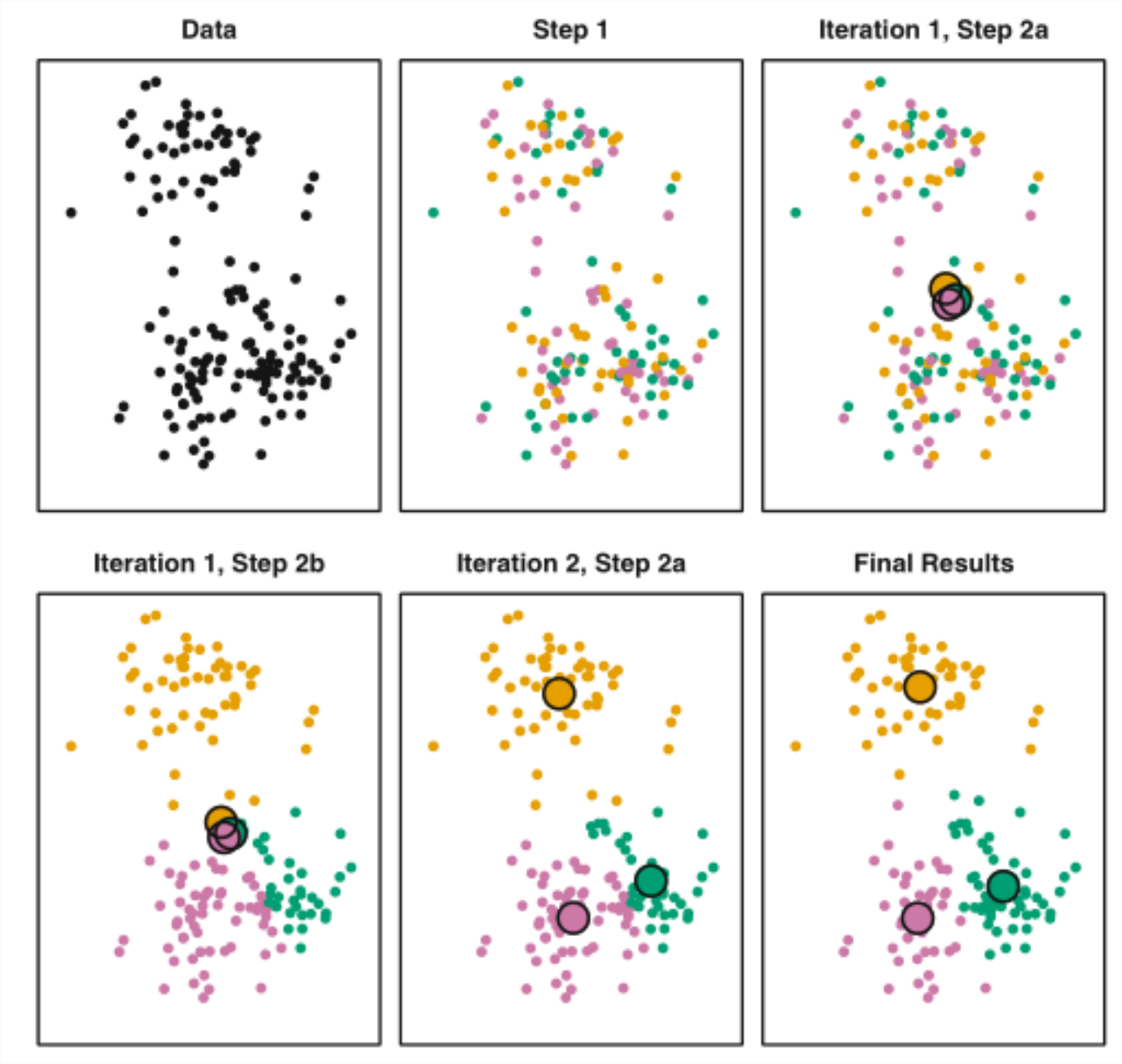
Algorithm 10.1 *K-Means Clustering*

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).

James, Witten, Hastie and Tibshirani (2017), p.388

The distance measure used is up to our discretion (e.g.: Manhattan, Minkowski, etc.)

K-means clustering algorithm

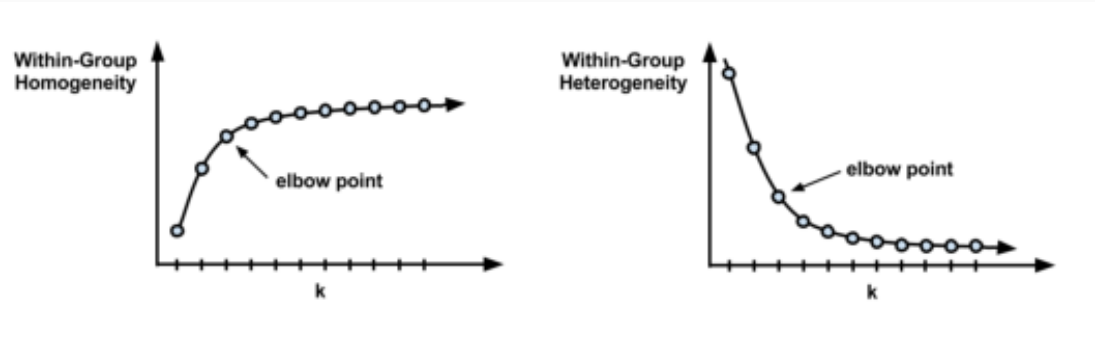


Problems with K-means

- We don’t know K
- Sensitivity to initial position of the centroids
- Can be trapped in a local minimum

How many K?

- **Prior information**: need to cluster into a specific amount of categories
- **Elbow method**: choose the one K value prior to diminishing total within-cluster sum of square (wss)
- $minimize(\sum_{k=1}^{k} W(C_k))$, where C_k is is the k^{th} cluster and the $W(C_k)$ is the within-cluster variation
- **Gap Statistic Method**: Compare within cluster dispersion on our actual data to an appropriate reference null distribution. (see more: Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 411-423.)



Hierarchical clustering

Agglomerative (bottom-up clustering)

- No need to specify the number of clusters before
- Visual representation: dendogram
- We can decide on the number of clusters by ”cutting the dendogram”

Dendogram interpretation

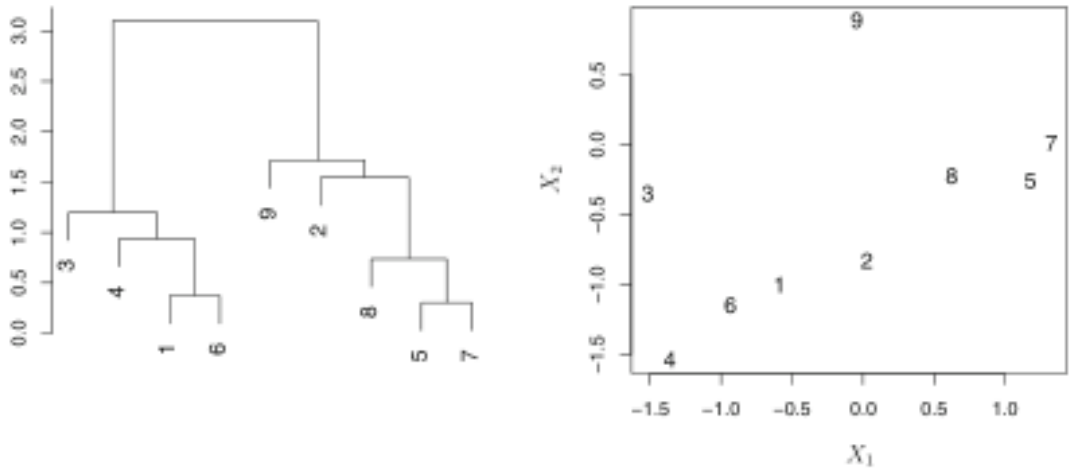


FIGURE 10.10. *An illustration of how to properly interpret a dendrogram with nine observations in two-dimensional space.* Left: *a dendrogram generated using Euclidean distance and complete linkage. Observations* 5 *and* 7 *are quite similar to each other, as are observations* 1 *and* 6. *However, observation* 9 *is* no more similar to *observation* 2 *than it is to observations* 8, 5, *and* 7, *even though observations* 9 *and* 2 *are close together in terms of horizontal distance. This is because observations* 2, 8, 5, *and* 7 *all fuse with observation* 9 *at the same height, approximately* 1.8. Right: *the raw data used to generate the dendrogram can be used to confirm that indeed, observation* 9 *is no more similar to observation* 2 *than it is to observations* 8, 5, *and* 7.

James, Witten, Hastie and Tibshirani (2017), p.393

The algorithm of agglomerative hierarchical clustering

1. Treat each item as its own cluster (with total of n clusters)
2. Create distance/similarity matrix (D_0) from the $N(N-1)/2$ pairwise distances of each cluster
3. Find the smallest distance in the matrix and merge the two items
4. Recalculate the distance matrix D_1 with the new cluster added.
5. Repeat step 3-4 until stop condition is reached (all items are merged in a single cluster)

Adressing similarity: Linkages

When to merge groups of observations?

- **Linkages** define distance between groups
- Four main type:
  1. **Complete**
  2. Single
  3. **Average**
  4. Centroid

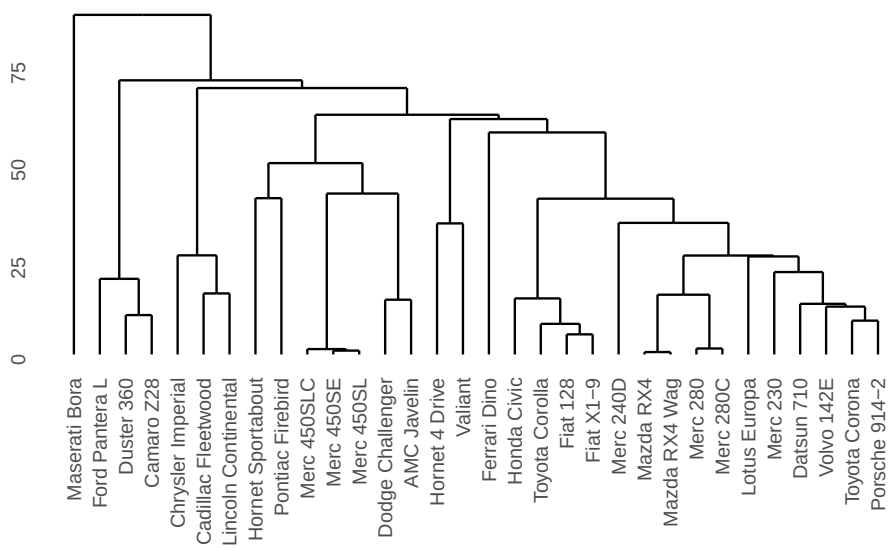
Linkages

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*.

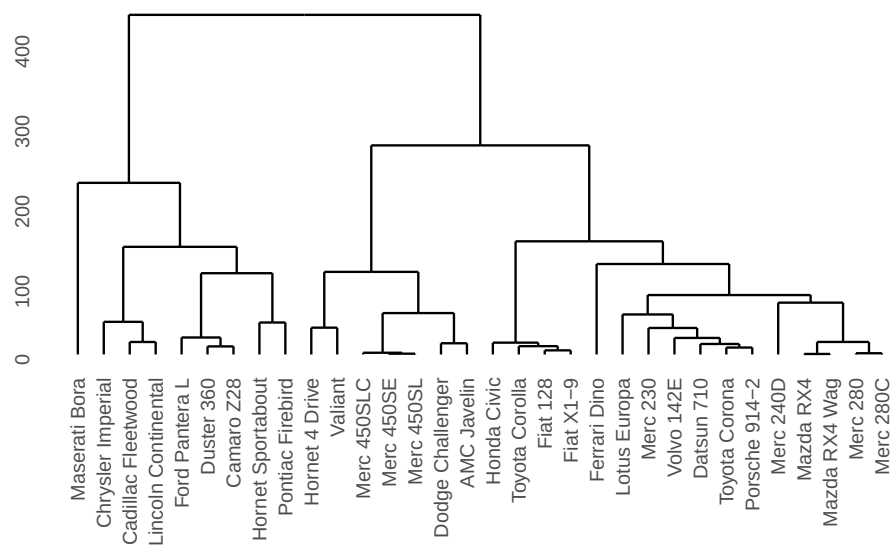
James, Witten, Hastie and Tibshirani (2017), p.395

Linkages comparison

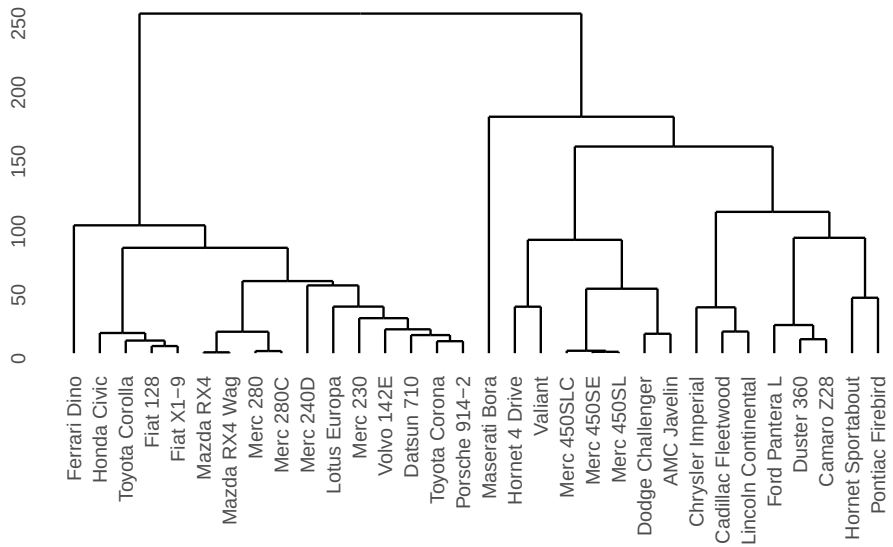
Single linkages



Complete linkages



Average linkages



Pros and Cons of Hierarchical clustering

Yay

- Deterministic
- No prior information needed on number of clusters
- Hierarchical relation between groups of data

Nay

Pros and Cons of Hierarchical clustering

Yay

- Deterministic
- No prior information needed on number of clusters
- Hierarchical relation between groups of data

Nay

- Computationally more expensive
- ”Cutting” the dendogram is arbitrary and subjective
- on the feature level we’ll likely get collocations on the base level
- No best solution between average or complete linkage, but choice alters outcome!