Text Analytics & Business Application

Text Clustering

Qinglai He
Department of Operations and Information Management
Wisconsin School of Business

WISCONSIN SCHOOL OF BUSINESS

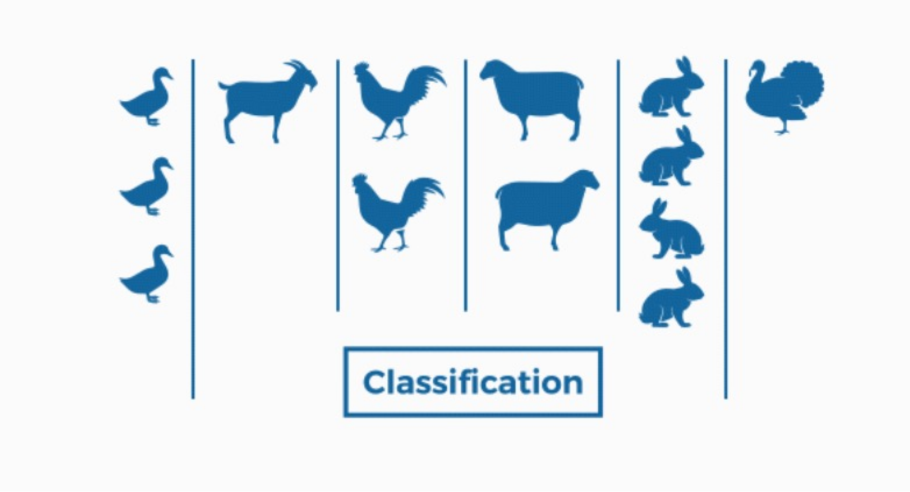
Outline of Text Clustering

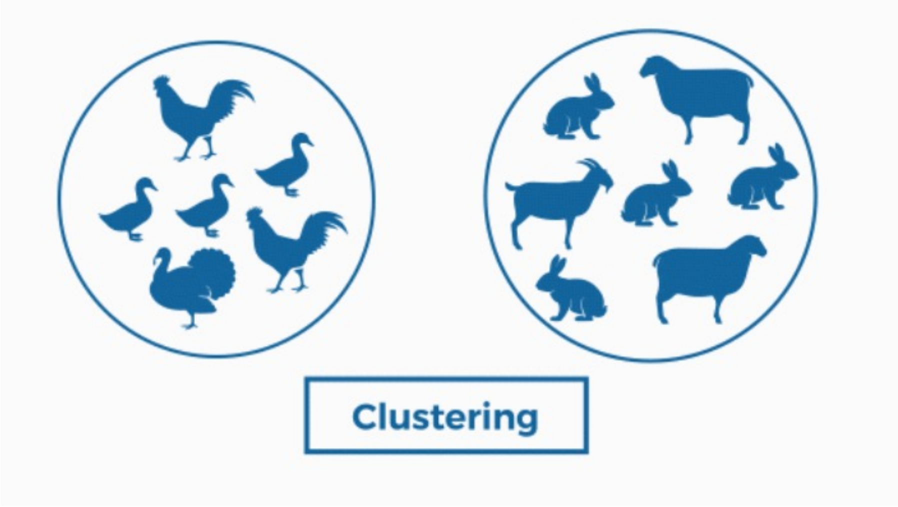
- Intro to text clustering
- Clustering method
    - K-means
    - Agglomerative clustering

Intro to Text Clustering

Classification vs. Clustering

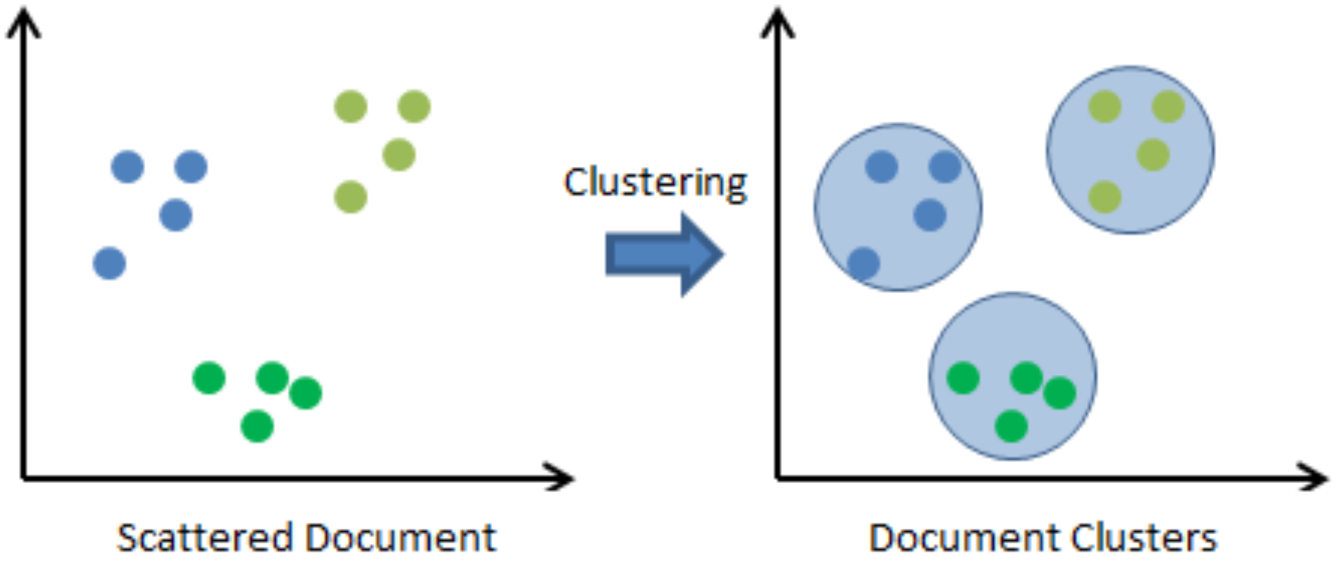
- **Classification** is supervised learning
  - It has labeled target variable
  - Example algorithms:
    - Logistic regression
    - Naive Bayes classifier
    - Support vector machines
- **Clustering** is unsupervised Learning
  - It does not have labeled target variable
  - Grouping the instances based on their similarity
  - Example algorithms:
    - K-means
    - Fuzzy algorithm
    - Gaussian (EM) clustering algorithm





Text Clustering

- Clustering can be incredibly useful for exploratory text analysis.
- With text data, each instance is a single document or utterance, and the features are its tokens, vocabulary, structure, metadata, etc.



Clustering by Document Similarity

- Many features of a document can inform similarity, from words and phrases to grammar and structure.
- For example:
  - We might group medical records by reported symptoms, saying two patients are similar if both have “nausea and exhaustion.”

Clustering by Document Similarity

- There are a number of different measures that can be used to determine document similarity:

String Matching	Distance Metrics	Relational Matching	Other Matching
Edit Distance - Levenstein - Smith-Waterman - Affine **Alignment** - Jaro-Winkler - Soft-TFIDF - Monge-Elkan **Phonetic** - Soundex - Translation	- Euclidean - Manhattan - Minkowski **Text Analytics** - Jaccard - TFIDF - Cosine similarity	**Set Based** - Dice - Tanimoto (Jaccard) - Common Neighbors - Adar Weighted **Aggregates** - Average values - Max/Min values - Medians - Frequency (Mode)	- Numeric distance - Boolean equality - Fuzzy matching - Domain specific **Gazettes** - Lexical matching - Named Entities (NER)

Applications of Text Clustering

Marketing Segmentation

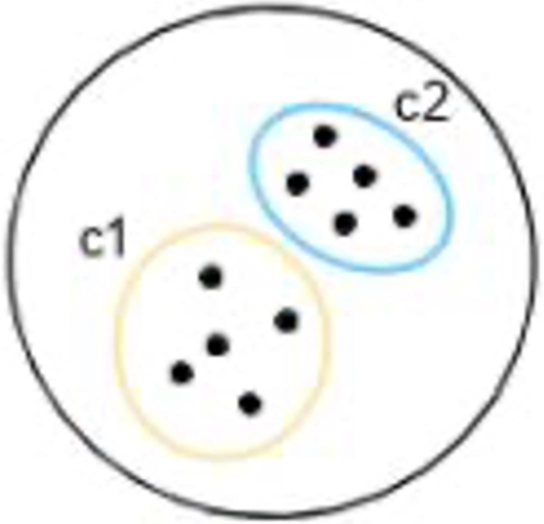
Search Engines

Clustering Method (1)

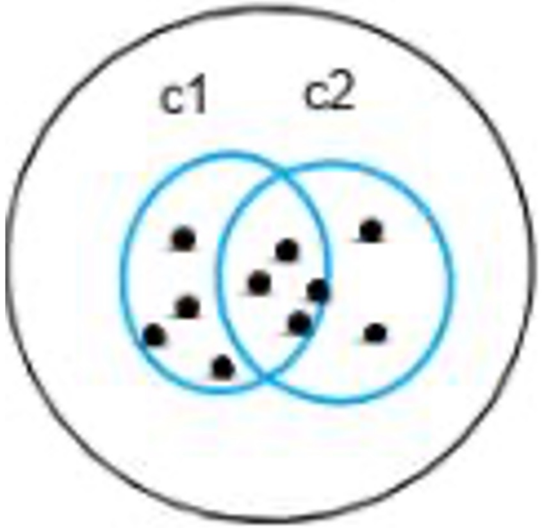
K-means

Clustering Methods

- The various types of clustering are:
  - Partitioning clustering
  - Hierarchical clustering
- Partitioning clustering
  - K-Means clustering
  - Fuzzy C-Means clustering
- Hierarchical clustering
  - Agglomerative clustering
  - Divisive clustering



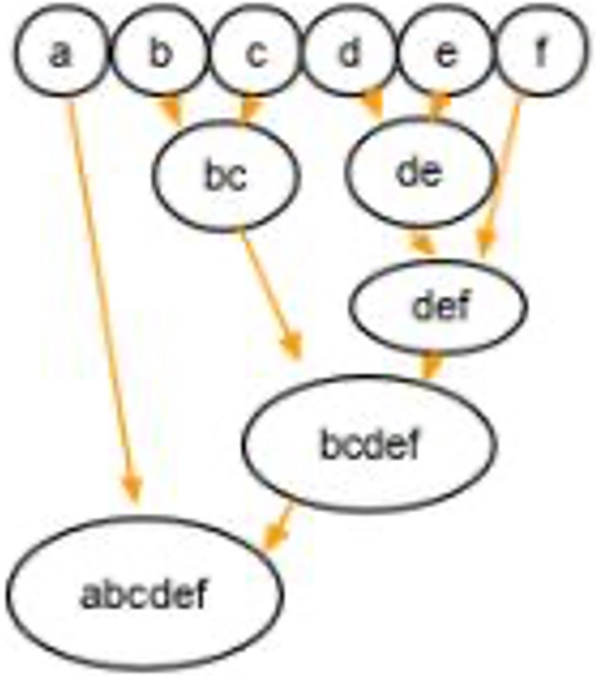
K-Means



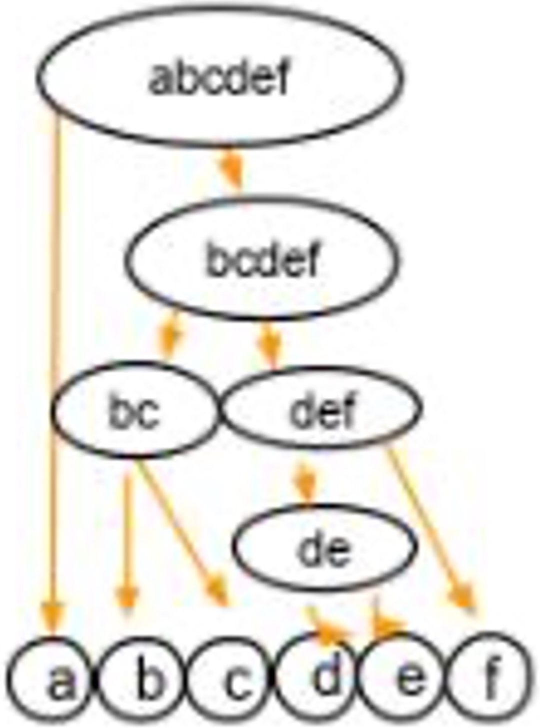
Fuzzy C-Means

Clustering Methods

- The various types of clustering are:
  - Partitioning clustering
  - Hierarchical clustering
- Partitioning clustering
  - K-Means clustering
  - Fuzzy C-Means clustering
- Hierarchical clustering
  - Agglomerative clustering
  - Divisive clustering



Agglomerative clustering



Divisive clustering

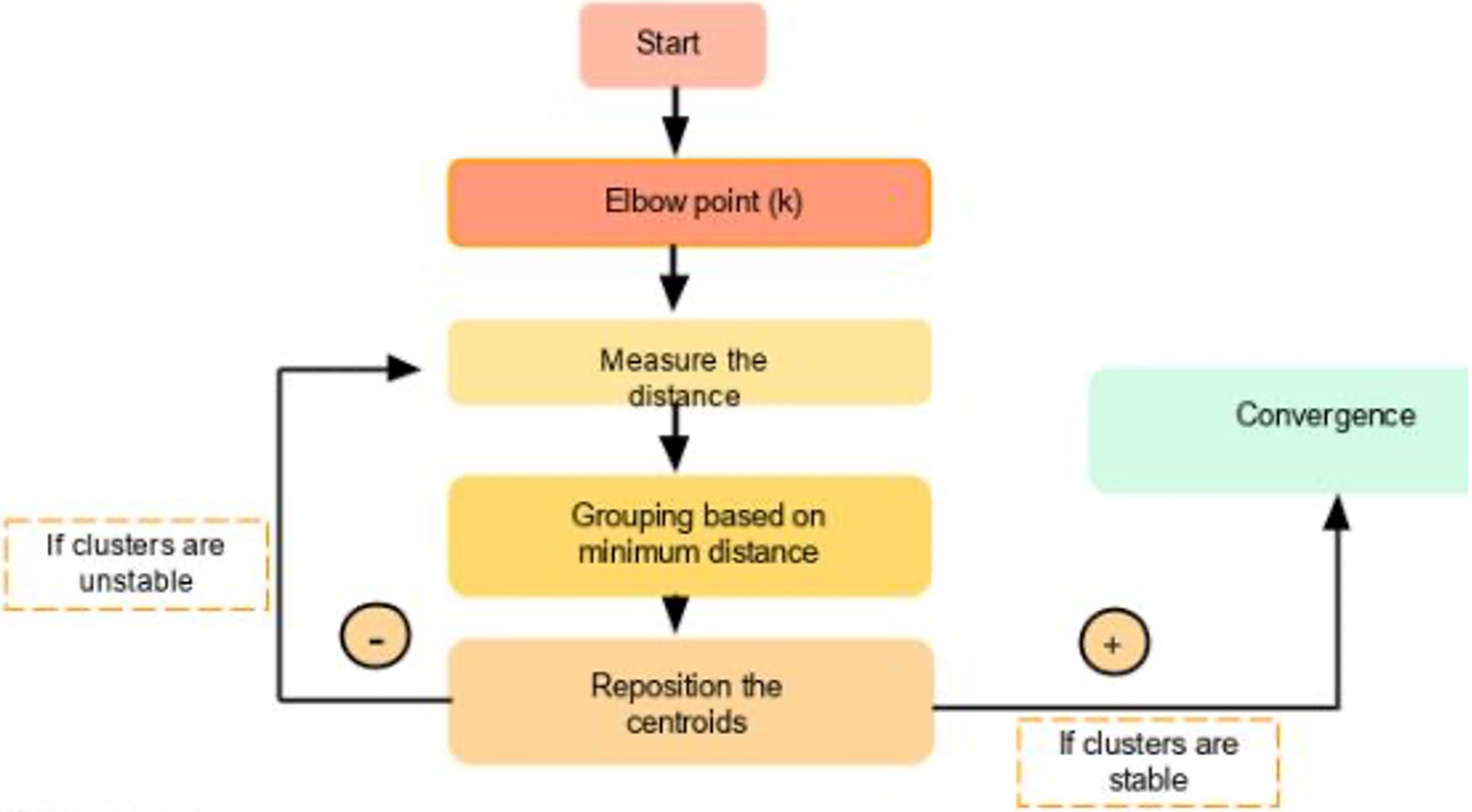
Partitioning Clustering

- **Partitioning clustering** separates documents into groups whose members share maximum similarity as defined by some distance metric.
- It partitions instances into groups that are represented by a central vector (the centroid) or described by a density of documents per cluster.
  - Centroids represent an aggregated value (e.g., mean or median) of all member documents and are a convenient way to describe documents in that cluster.

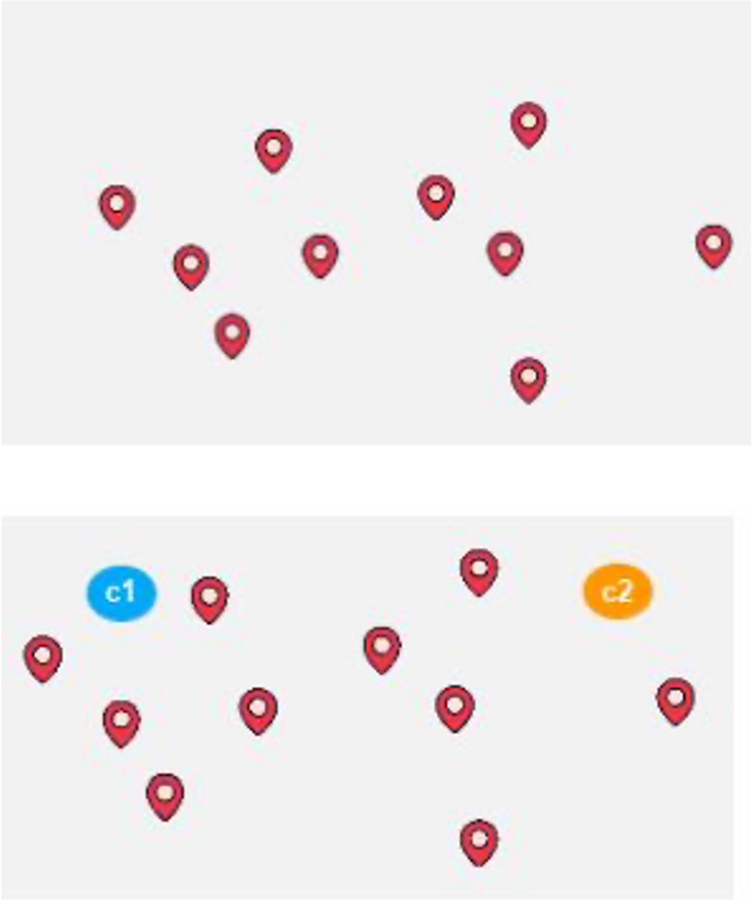
K-means Clustering

- The *k*-means algorithm is based on the choice of the initial cluster centers. The general process is below.
    1. Specify the *k* value
    2. Randomly assign *k* observations as cluster centers
    3. Assign each observation to its nearest cluster center
    4. Calculate cluster centroids
    5. Reassign each observation to a cluster with the nearest centroid
    6. Recalculate the cluster centroids, and repeat step 5
    7. Stop when reassigning observations can no longer improve within-cluster dispersion.
- Dispersion is defined as the sum of Euclidean distances of observations for their respective cluster centers.
- Results from *k*-means clustering are highly sensitive to the random process for finding the initial cluster centers as well as implementing specific algorithms.

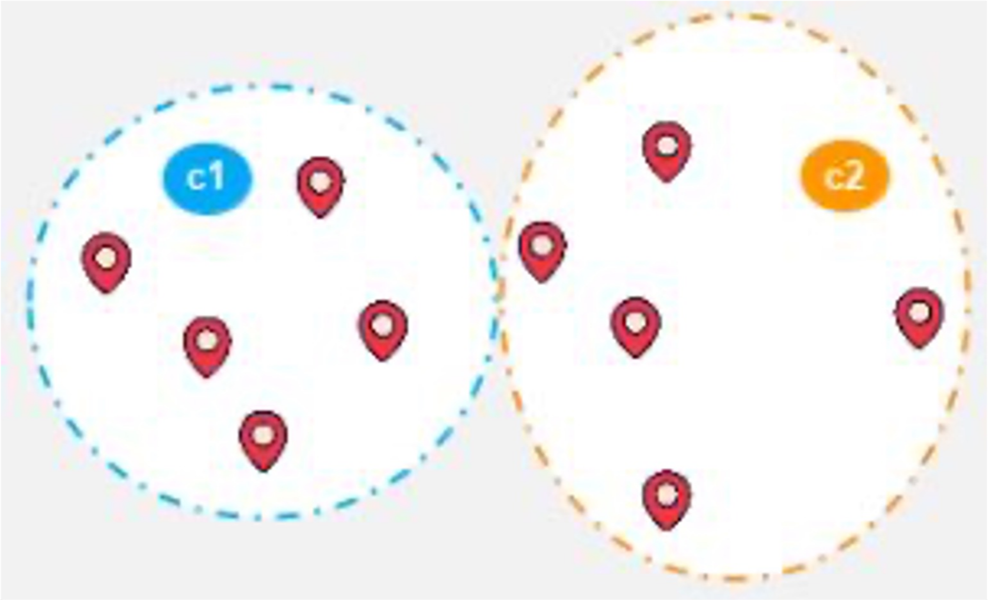
K-means Clustering



K-means Clustering Steps

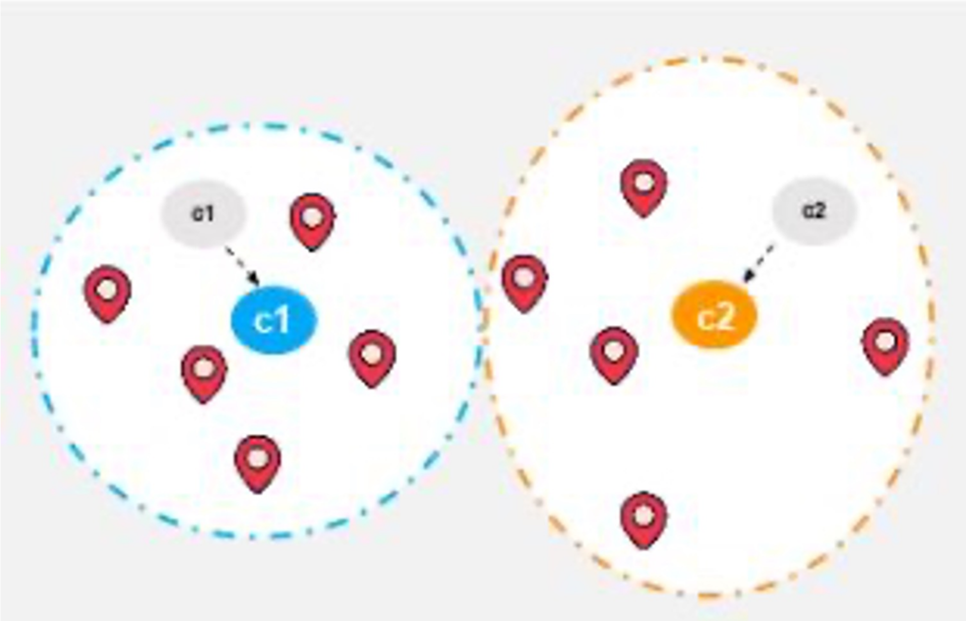


Randomly initiate two cluster centroids

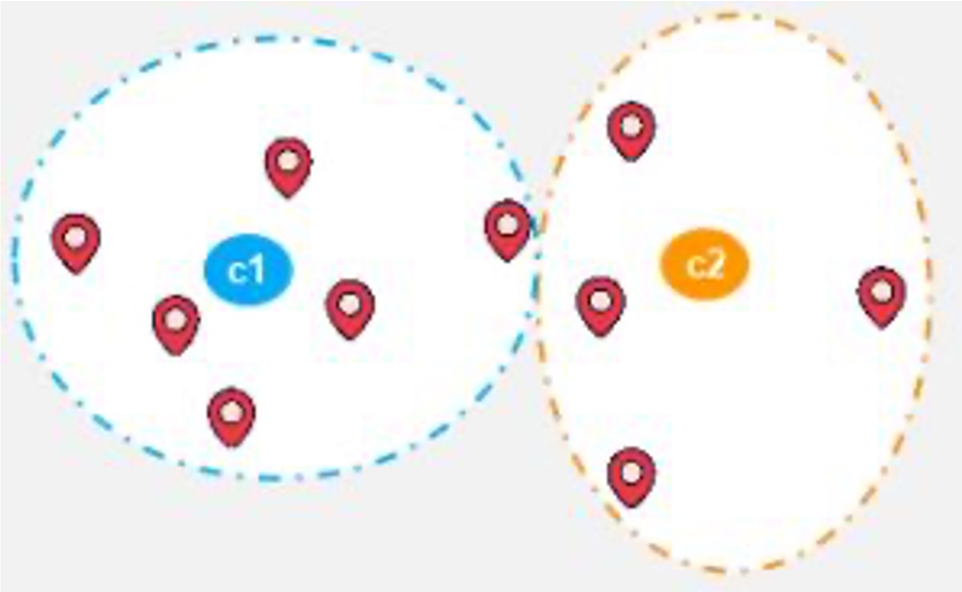


Based on the distance between each data point and centroids, assign each point to a nearest centroid. Then, we form two groups.

K-means Clustering Steps



Compute the actual centroids for each group. Reposition the initial random centroids to the actual centroids.



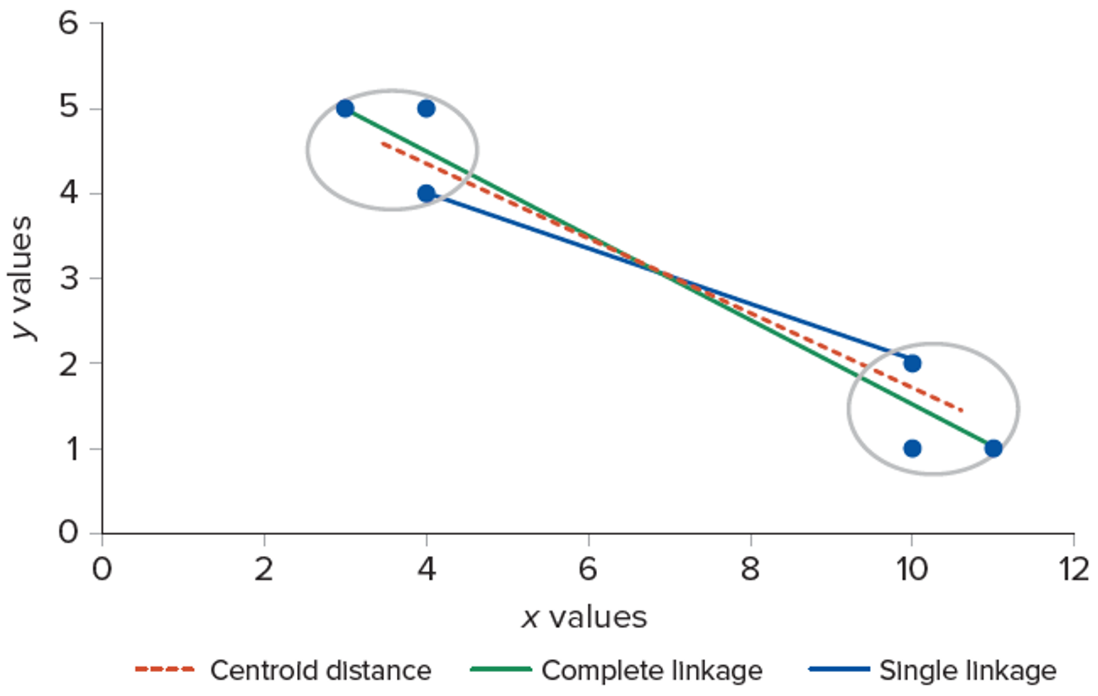
Iterate the centroid update many times until the cluster becomes static.

K-means Clustering

- The objective is to divide the sample into a prespecified number *k* of **non-overlapping** clusters so that each of these *k* clusters is as homogenous as possible.
- <u>The number of clusters *k* needs to be specified</u> prior to performing the analysis.
- We may experiment with different values of *k* until we obtain a desired result.
- In addition, we may have prior knowledge or theories about the subjects under study and can determine the appropriate number of clusters based on domain knowledge.
- The *k*-means clustering method can only be applied to data with numerical variables. For categorical variables, we need to convert them into numerical.

Multiple Linkage Methods to Evaluation (dis)Similarity Between Clusters

- **Single**: nearest distance between a pair of observations not in the same cluster
- **Complete**: farthest distance between a pair of observations not in the same cluster
- **Centroid**: distance between the center/centroid or mean values of the clusters
- **Average**: average distance between all pairs of observations not in the same cluster
- **Ward's**: uses error sum of squares (ESS/WCSS), which is the squared difference between individual observations and the cluster mean; measures the loss of information that occurs when observations are clustered.



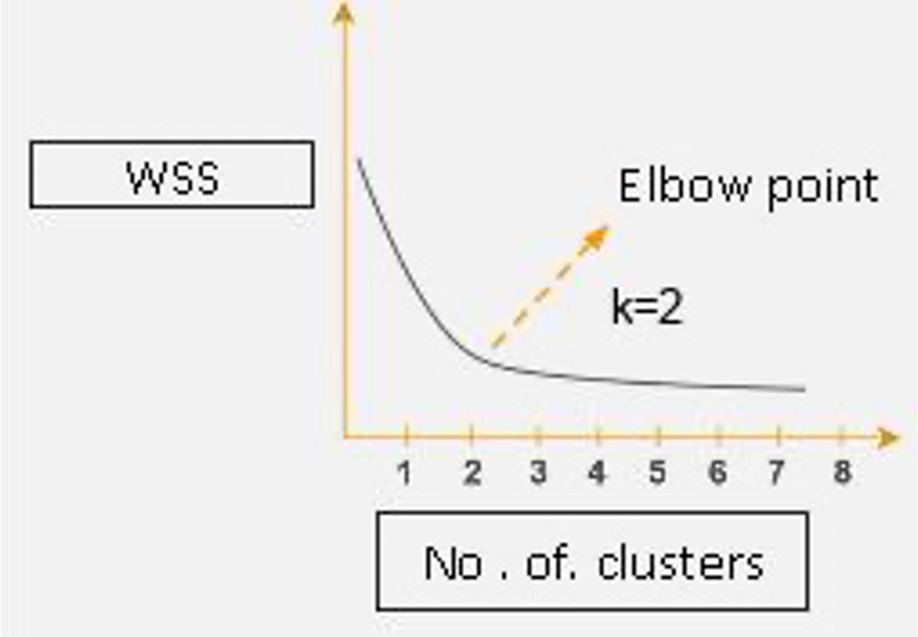
How Should We Choose the Optimal K?

Elbow technique

We need to calculate within-sum-of-squares (WSS or WCSS). WSS is defined as the sum of the squared distance between each member of the cluster and its centroid.

$$WSS = \sum_{i=1}^{m} (x_i - c_i)^2$$

Where x_i = data point and c_i = closest point to centroid



K=2 is the optimal value. There is a gradual change in the value of WSS as the K value increase from 2. Beyond that, increasing the K will not dramatically change the value of WSS.


QUIZ

QUIZ

Q1. K-means is an iterative algorithm, some steps need to be done repeatedly. Which are the repeated steps?

A. Assign each point to its nearest cluster
B. Update the cluster centroids based on the current assignment
C. Using the elbow method to choose K
D. Test on the test dataset

Answer: A,B

Q2. What is the minimum number of variables/features required to perform clustering?

A. 0
B. 1
C. 2
D. 3

Answer: B

Q3. If we run K-Means clustering twice, is it expected to get the same clustering results?

A. Yes
B. No
C. Yes, as long as we use the same data
D. Yes, as long as we use the same distance measure

Answer: B

QUIZ

Q4. The ideal clustering results should have _

A. High intra-cluster similarity (within a cluster)
B. Low intra-cluster similarity
C. High inter-cluster similarity (between clusters)
D. Low inter-cluster similarity

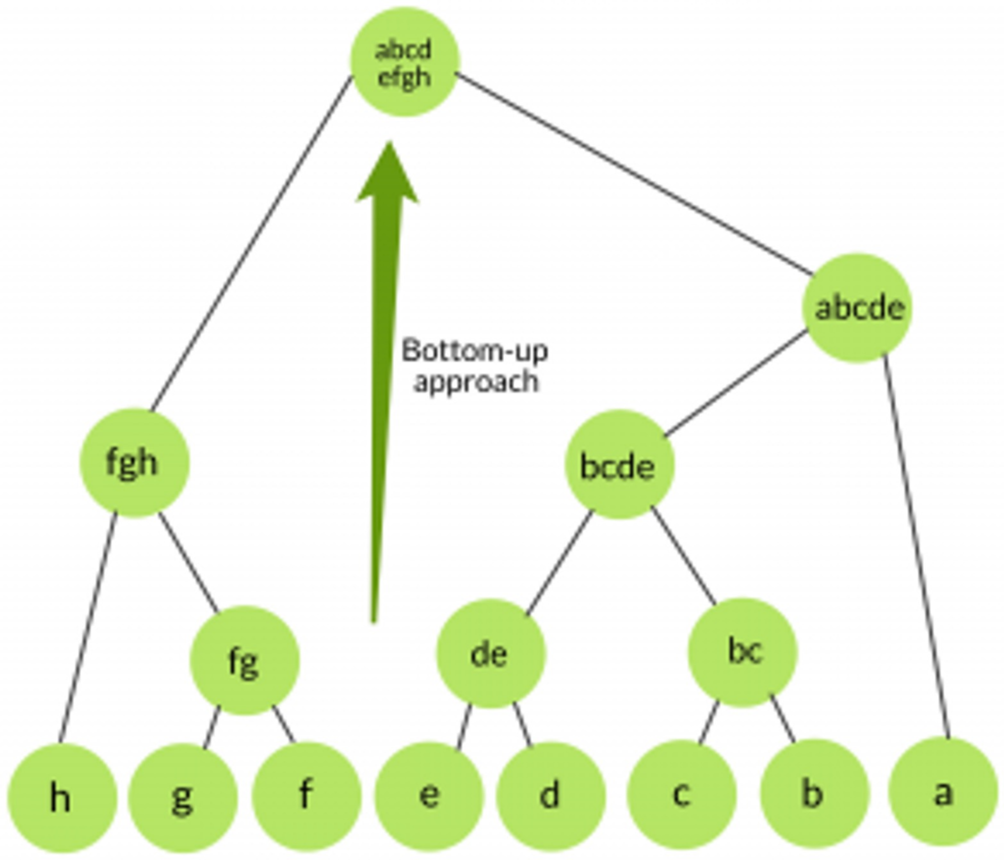
Answer: A, D

Clustering Method (2)

Agglomerative Clustering

Agglomerative clustering

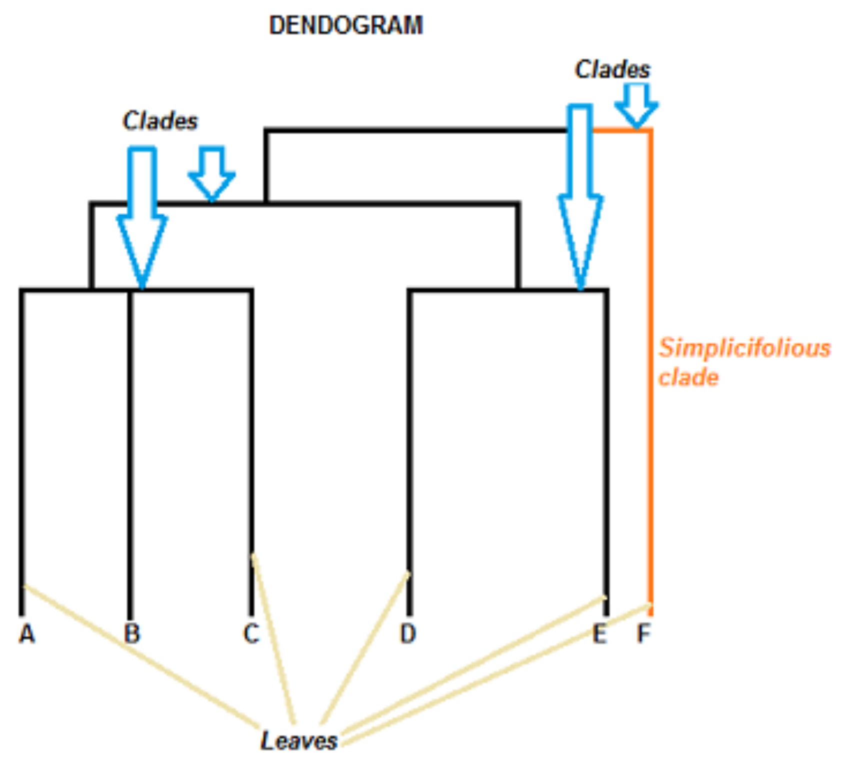
- With AGNES, each observation in the data initially forms its own cluster.
- The algorithm then successively merges these clusters into larger clusters based on their similarity until all observations are merged into one final cluster, referred to as a root.
- Uses (dis)similarity measures.
    - Numeric variable: Euclidean distance or Manhattan distance
    - Categorical variable: matching, Jaccard's coefficient



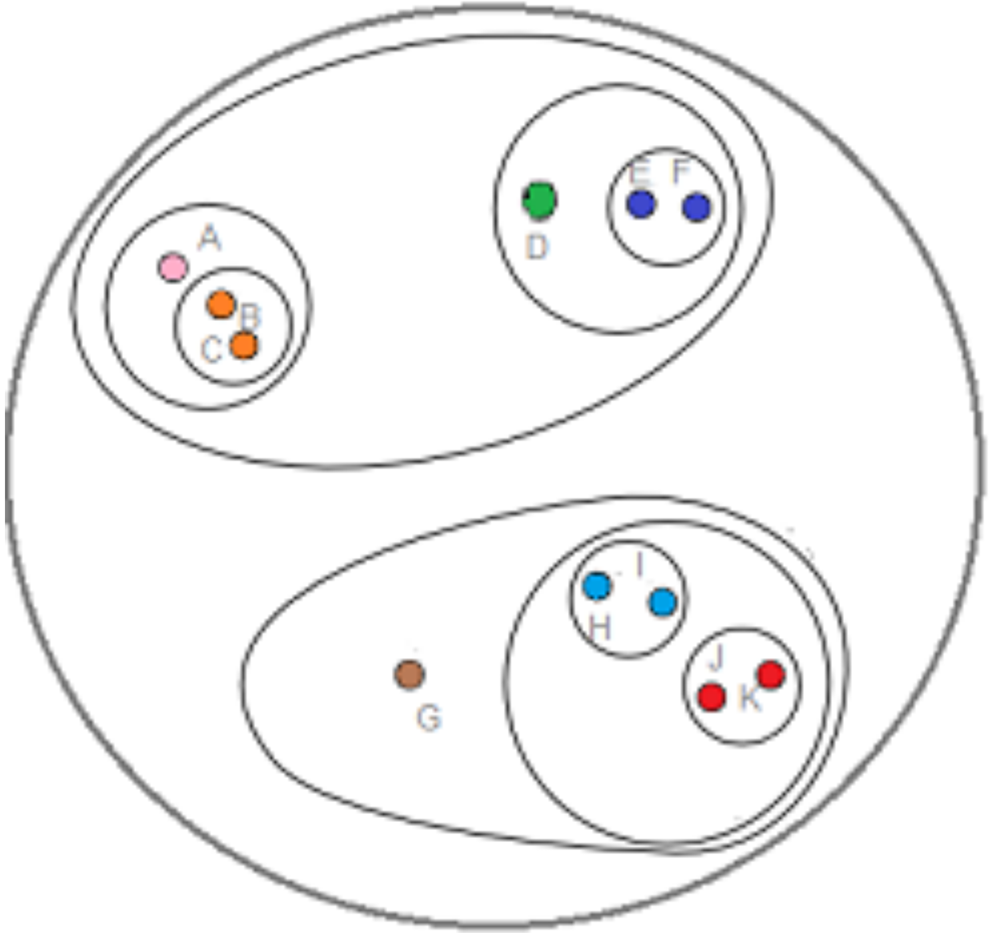
Hierarchical agglomerative clustering

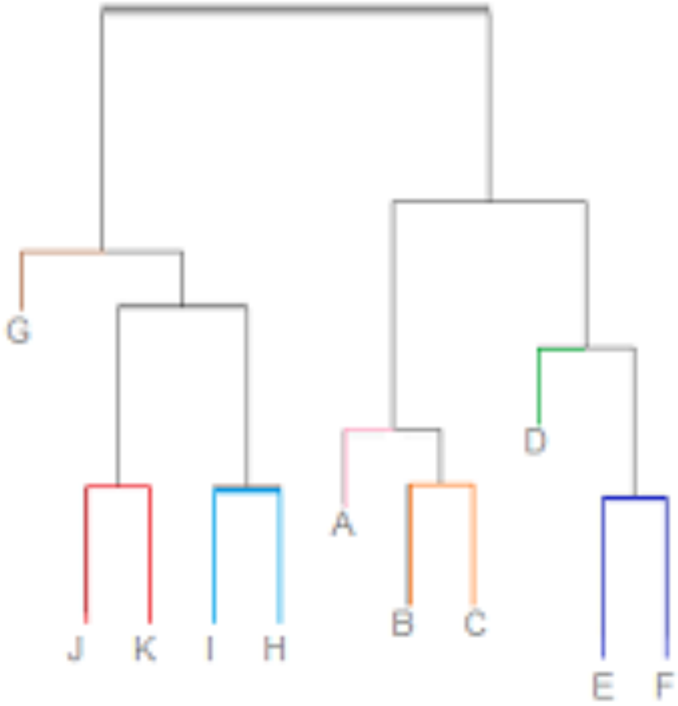
Agglomerative clustering

- Once AGNES completes the clustering process, data are usually represented in a tree-like structure.
  - Called a dendrogram
  - Branches are clusters
  - An observation is a “leaf”
  - Visually inspect the clustering result and determine the appropriate number of clusters
- The height of each branch (cluster) or sub-branch (sub-cluster) indicates how dissimilar it is from the other branches or sub-branches with which it is merged.
- The greater the height, the more distinctive the cluster is from the other clusters.

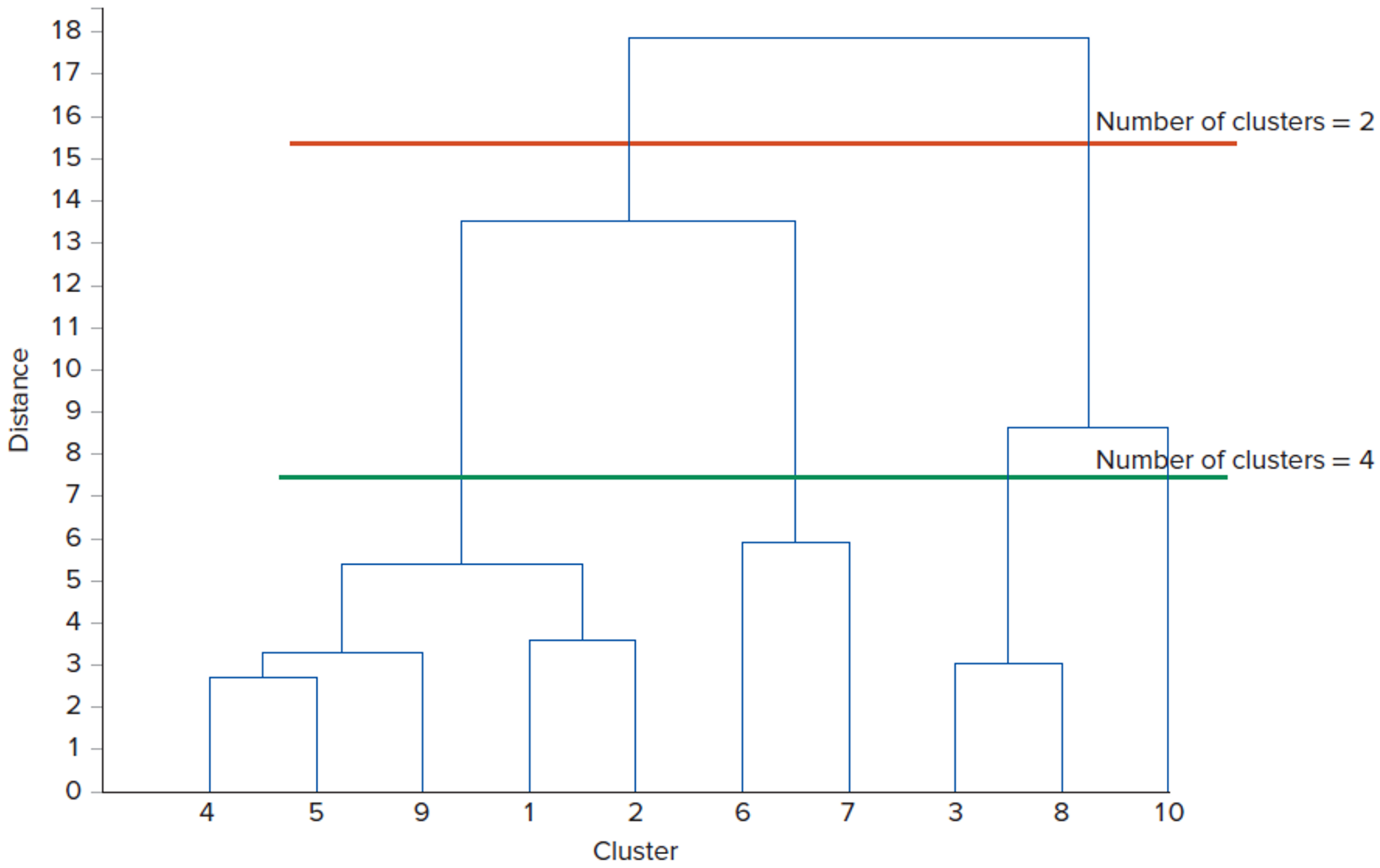


AGNES Dendrogram



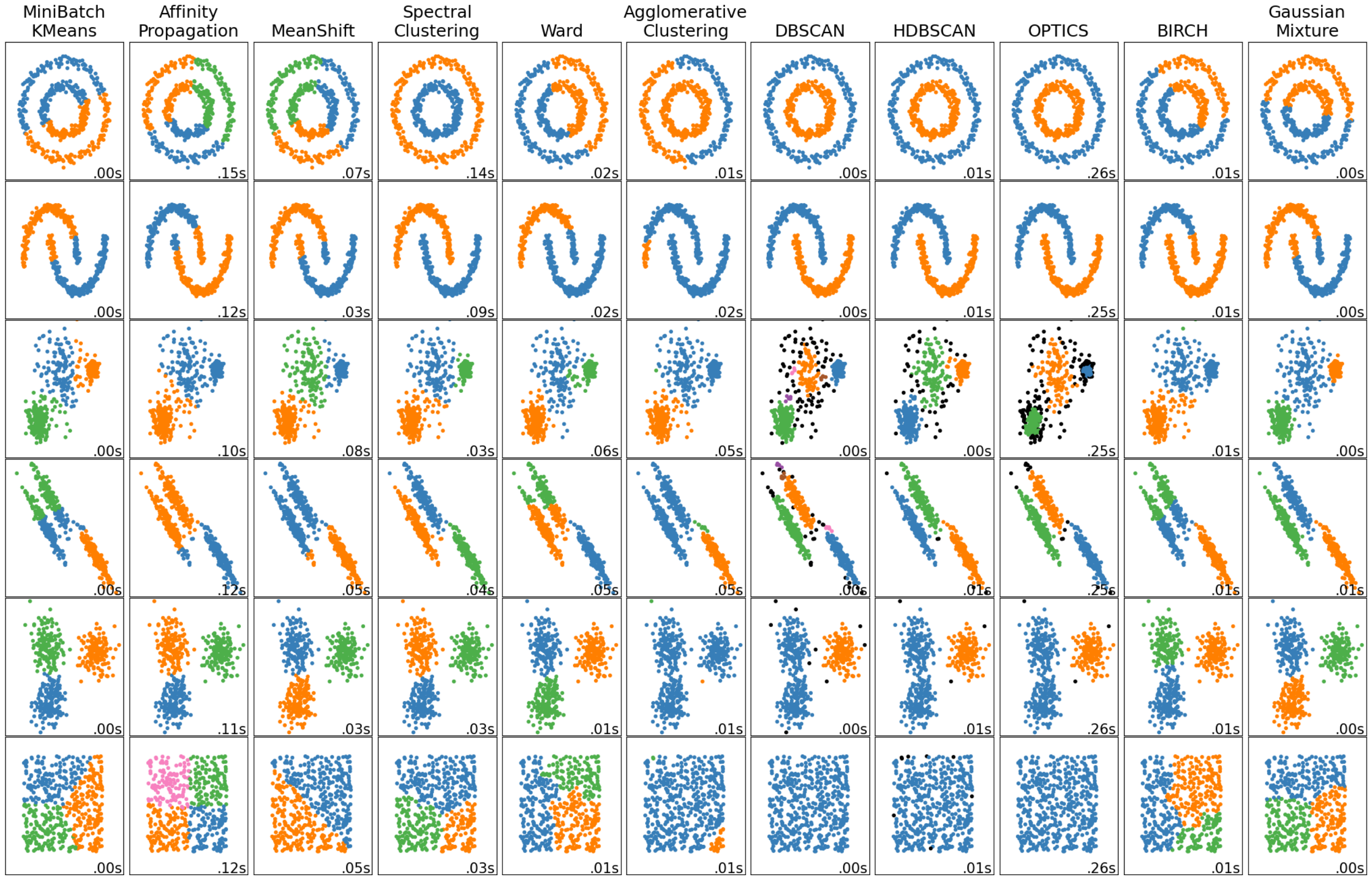


Agglomerative Clustering (Dendrogram)



Other popular clustering methods

- DBSCAN
- BIRCH
- GMM
- Fuzzy clustering
- …



https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html

Exercises using Google Colab