COMM 605

Week 10: Unsupervised machine learning: Topic modeling

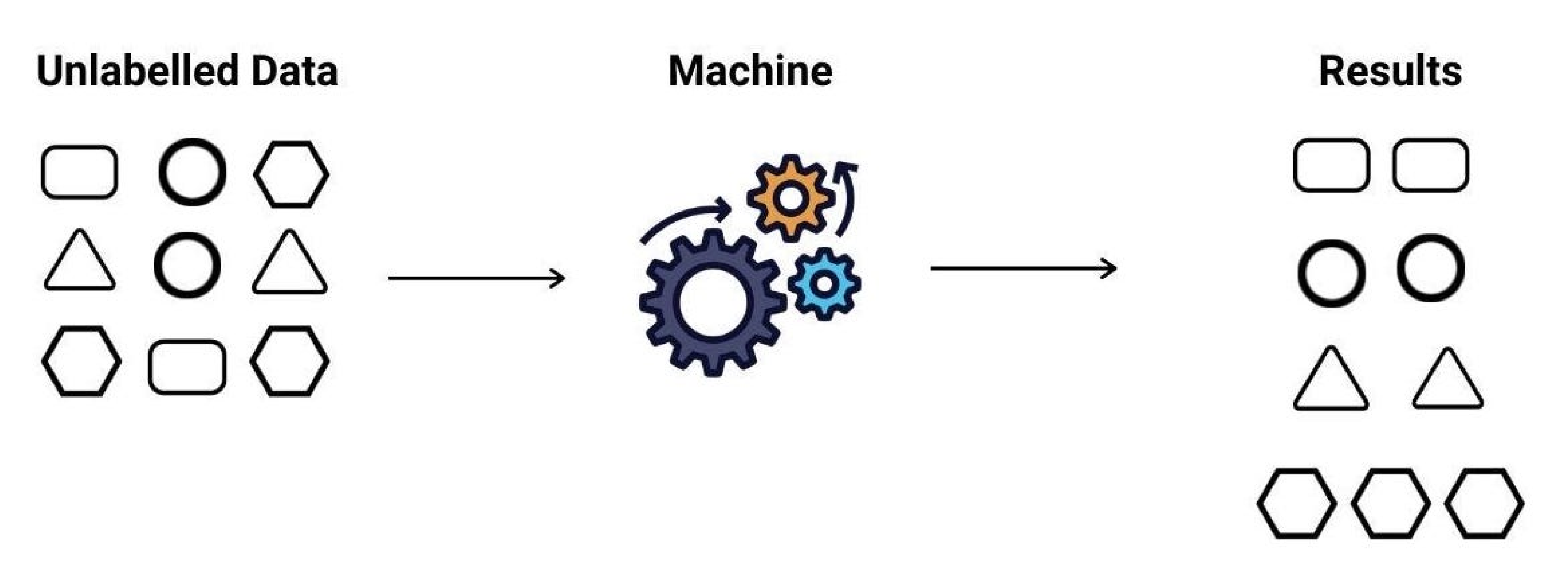
Ryan Y. Wang, Ph.D.

ryanwang@rit.mail.edu

School of Communication

October 30, 2023

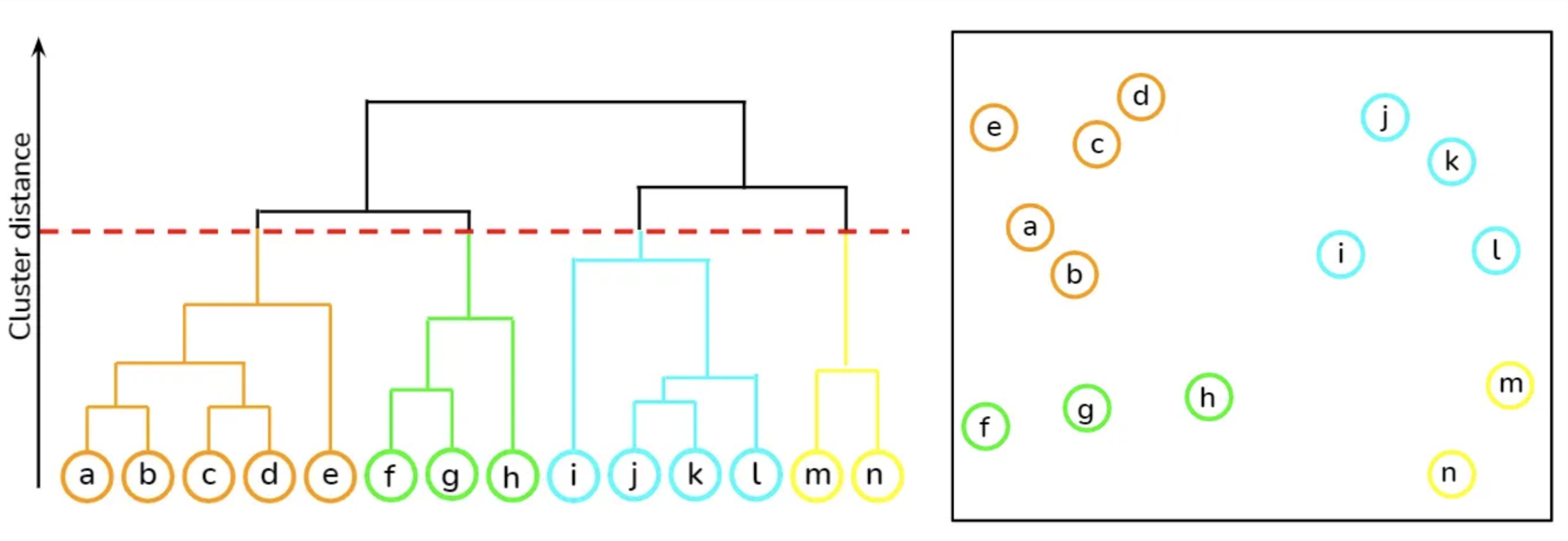
Unsupervised machine learning



Unsupervised machine learning

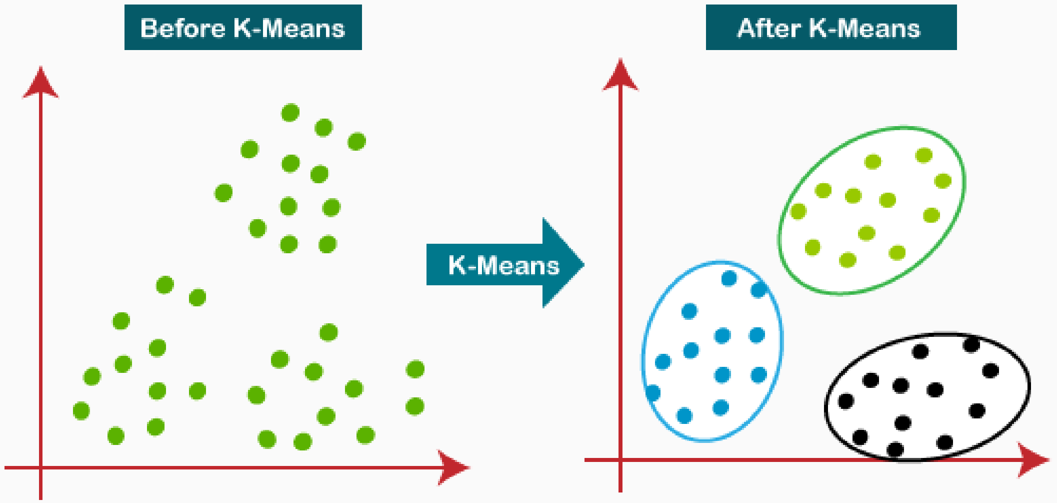
In the case of input data of images of different shapes:

- Clustering (Unsupervised classification): The goal is to find homogeneous subgroups within the data. The grouping is based on similiarities (or distance) between observations. The result of a clustering algorithm is to group the observations (features) into distinct (generally non-overlapping) groups.
  - Hierarchical clustering



Unsupervised machine learning

- Clustering (Unsupervised classification):
    - K-means clustering



Unsupervised machine learning

In the case of input data of images of different shapes:

- Dimensionality Reduction: Dimensionality could be understand the number of variables, characteristics or features present in the dataset (e.g., color pixel values, size, and shapes). The goal is to summarize the data in a reduced number of dimensions, i.e. using a reduced number of variables.
    - PCA (Principal Component Analysis)
    - t-SNE (t-distributed Stochastic Neighbor Embedding)

Unsupervised machine learning in textual data

Since a document-term matrix (DTM) is a matrix, you can also apply these unsupervised machine learning techniques to the DTM to find groups of words or documents.

- Topic modeling: We group words and documents into *topics*, consisting of words and documents that co-vary
- Goal: Given a corpus, find a set of topics, consisting of specific words and/or documents, that minimize the mistakes we would make if we try to reconstruct the corpus from the topics

Empirical examples

Murashka, Liu & Peng (2021)

- Topics: Fitspiration on Instagram
- RQ: topics (in comments) to posts (objectification features)
  - Posts: human coding (N = 2000)
  - Comments: topic modeling (N = 35263) -> K = 3 ((p. 1543)
  - Multilevel analysis

Empirical examples

Yang, Sun & Taylor (2022)

- Topic: CSR on Facebook
- RQ: public response to Fortune 500 companies's discussion on their COVID-19 pandemic CSR actions
    - Post: topic modeling (N = 9977 posts from 469 companies)
        - K = 20 (p.6)
        - Classifying into three themes:community information update, organizational crisis response and organizational contribution (network)
    - Public response:
        - Behavioral engagement outcome: comment and share
        - Emotional engagement outcome: like, love, sad, angry

Thank you!