WEEK 10: CLUSTERING

Dr. Kai Li

School of Information Sciences

University of Tennessee, Knoxville

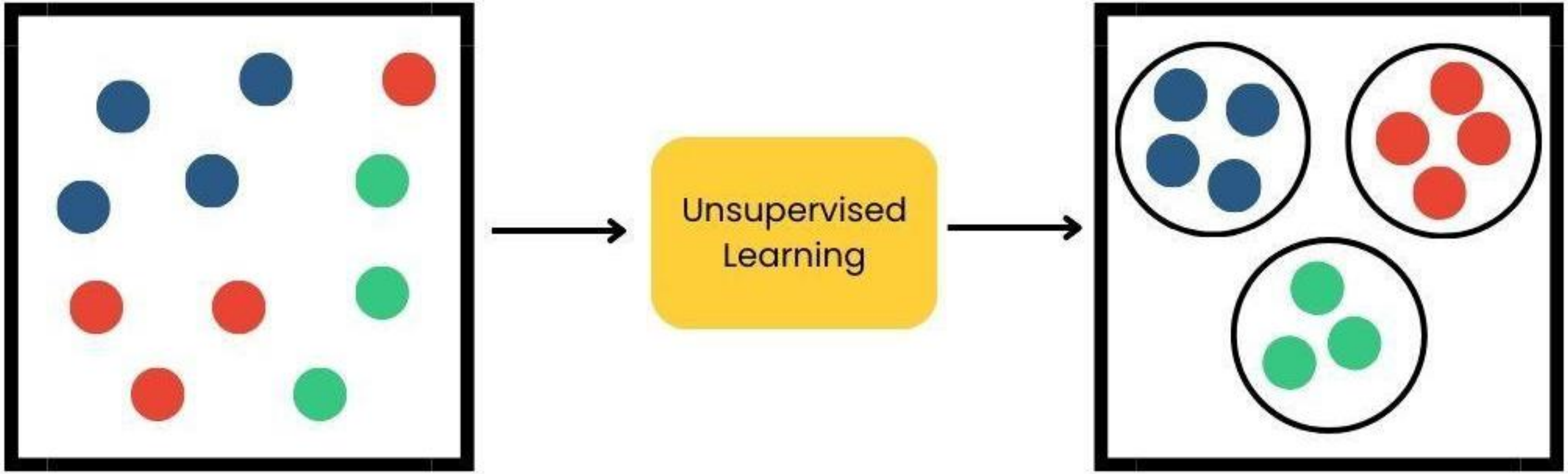
Spring 2025

Review / Overview

- Last week:
  - *Classification methods*
- Your thought on group activity?
- Final project idea sharing!
- This week:
  - *Final project idea*
  - *Unsupervised learning:*
    - Clustering

Unsupervised learning

- Unlike supervised learning, in unsupervised learning, we don’t give algorithm data to learn, instead we allow the model to discover patterns without any guidance or instruction.



CLUSTERING

Clustering is a central technique in unsupervised learning. It strives to group all observations into natural, similar groups (*clusters*).

Why clustering?

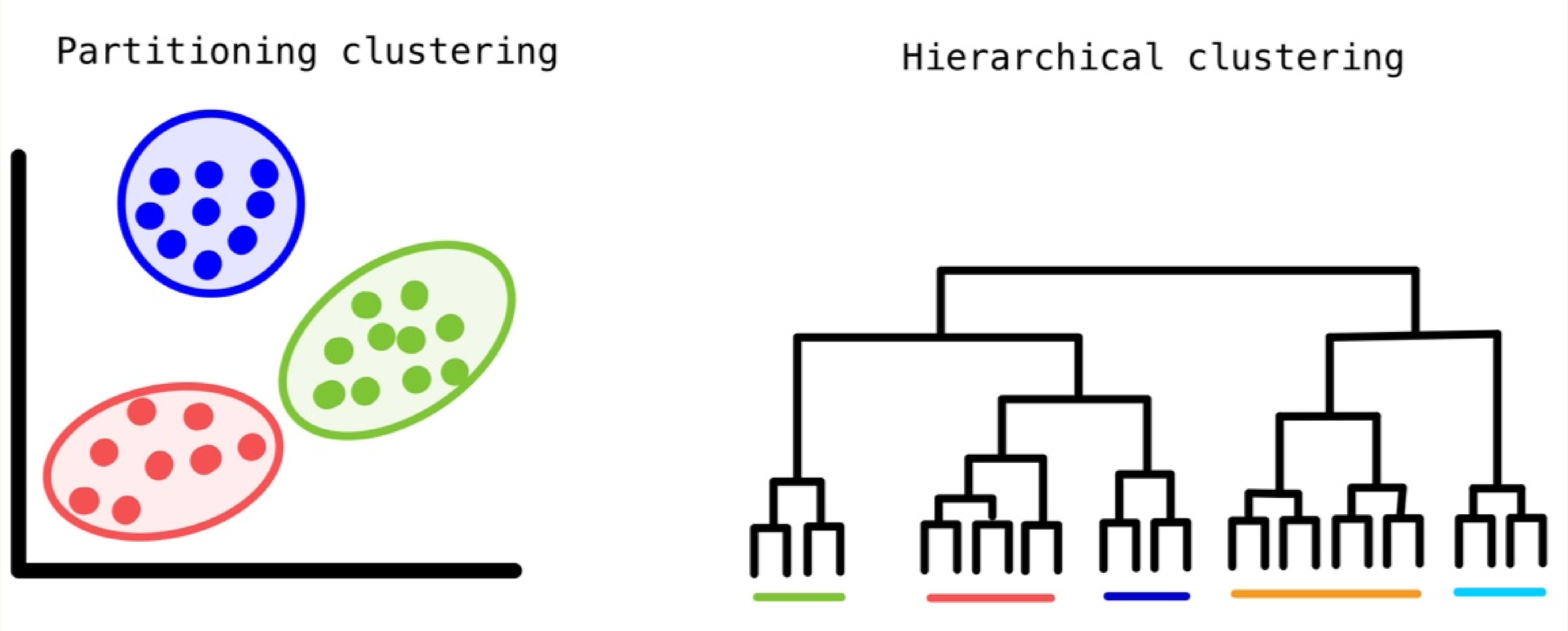
- There could be a few reasons for doing clustering:
  - *Understand the internal structure of the data*
  - *Classifying data into clusters*
- For example:
  - *We can cluster users (user groups) based on their behavior.*
  - *Understand how individual instances are similar with each other.*

Two additional comments

- Given that there is no train and test data, clustering is inevitably subjective.
  - *But of course, it is based on the algorithm’s best guess.*
  - *We can still evaluate the model’s performance by looking at how well the model works (by not comparing with baseline data).*
- The algorithm can use whatever variables we feed into the model to define similarity.
  - *Some techniques, such as K-means, do not support categorical variable at all.*

Two types of clustering methods

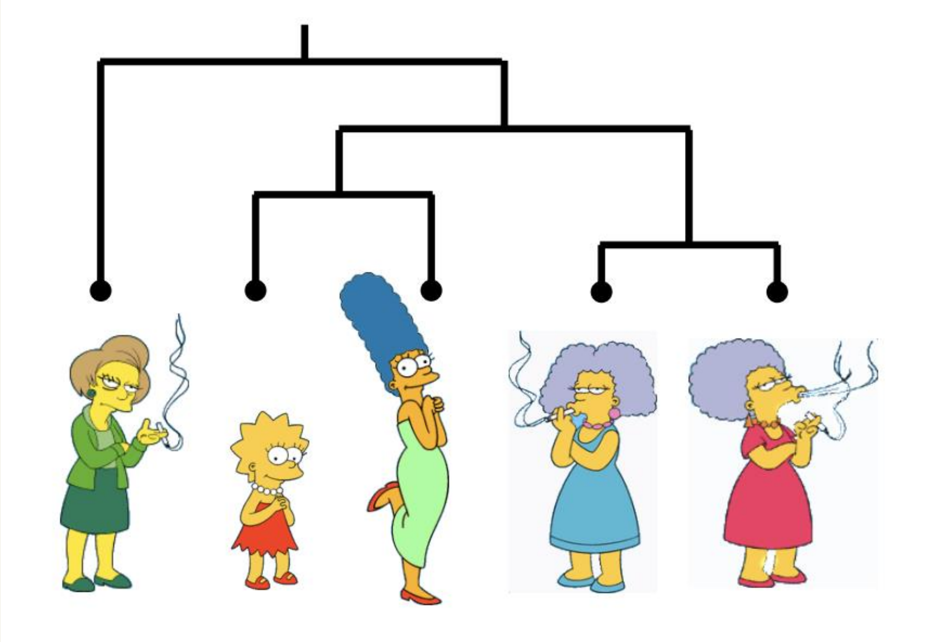
- Hierarchical clustering:
  - *It is a bottom-up approach: for each instance, the algorithm will look for other similar items and grow clusters.*
- Partitional clustering
  - *It is a top-down approach: the algorithm will calculate the clusters and then assign every instance to each cluster.*



https://medium.com/@chyun55555/how-to-find-the-optimal-number-of-clusters-with-r-dbf84988388b

Hierarchical clustering

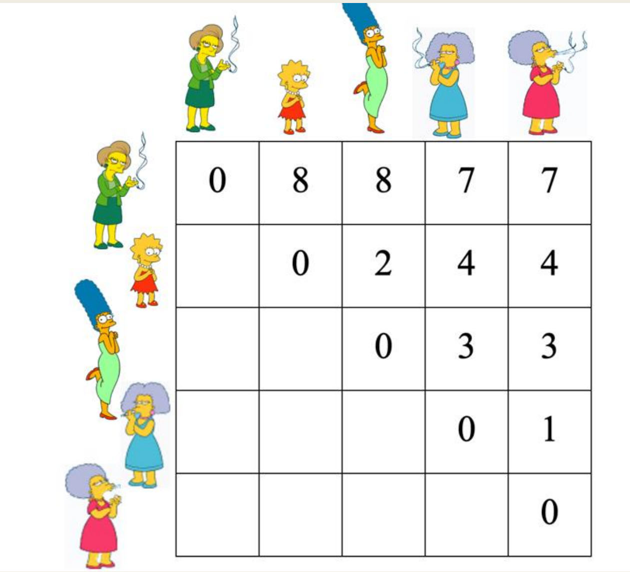
- The basic idea of hierarchical clustering is that for each item, we should try to find other items that can be merged with it based on the distance between items.
- While we are not trying to define distance in this class, it is the general similarity between things, which can be calculated in many ways.



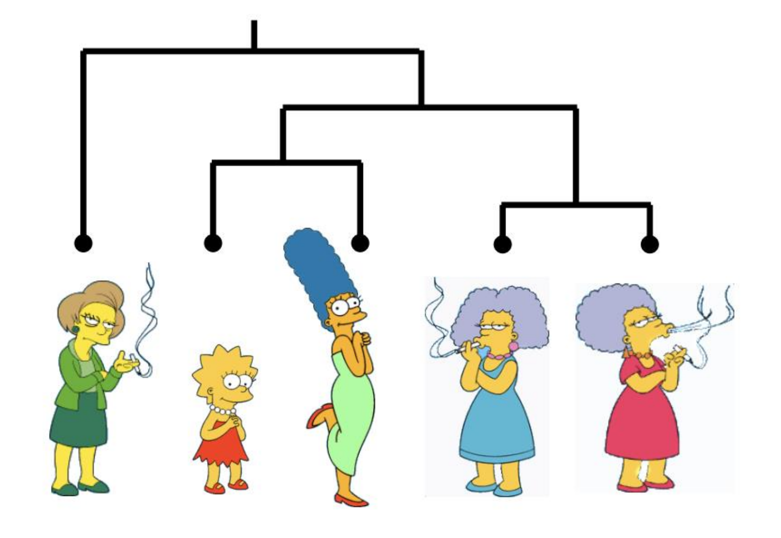
This example from: https://www.cs.cmu.edu/~epxing/Class/10701/slides/clustering.pdf

Visualization of hierarchical clustering

- A distance matrix (top) between all items is the basis of creating the dendrogram (below).
- The dendrogram will be computed automatically based on the process describe above.

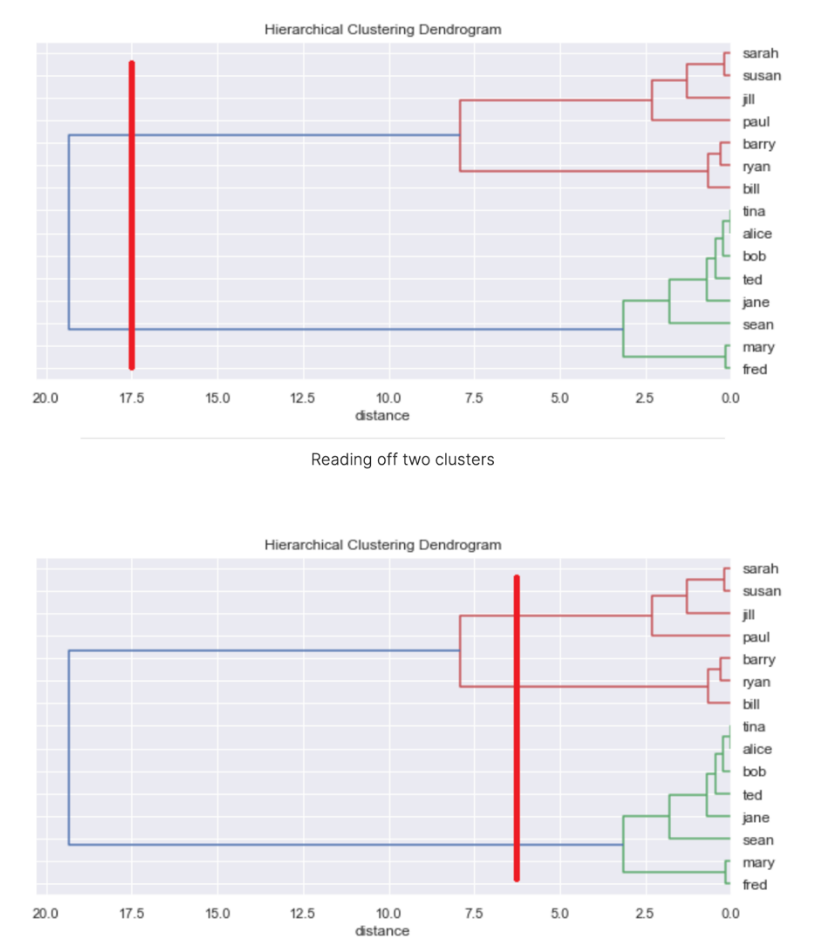


0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0



Determining the number of clusters in hierarchical clustering

- As the results of hierarchical clustering are just the tree structure (dendrogram), we will have to decide how many clusters we want to have, and this decision will have direct impacts on our results.
- Generally, we can make this decision by choosing any meaningful number of clusters.
  - *For example, we already know that there should be k groups in the data.*



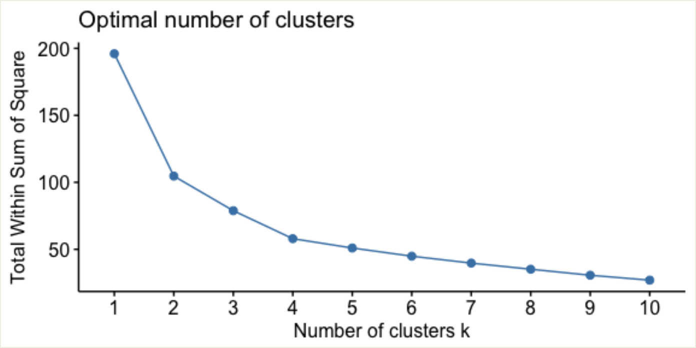
Reading off two clusters

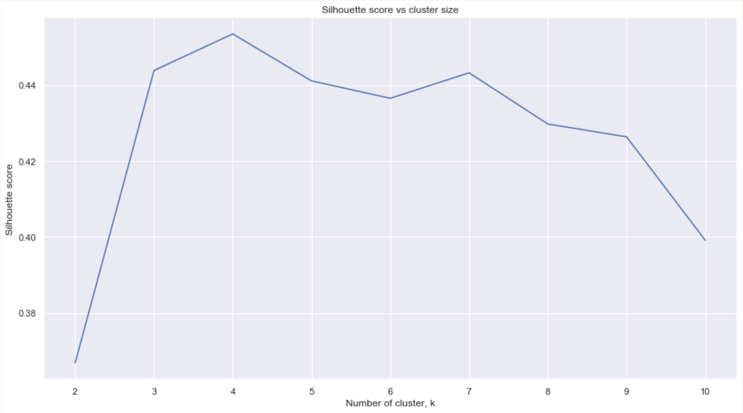
Partitional clustering

- Partitional clustering just focuses on putting all items in different clusters.
- In one popular algorithm, K-means clustering, we first define the number of clusters (number K) and calculate the center of each cluster. And then, we assign all items into the nearest clusters.
  - *The process is normally iterative: we will need to do a few rounds of adjustment to have the best clusters.*
  - *There are a few different ways to determine what the value of K should be (to be discussed in the next slide).*

Determining the K number

- Again, we can use our domain knowledge to determine the number of K.
- But a more popular method is the elbow method.
  - *In the graph, we need to choose the "elbow point" (the point after which there is a linear relationship) → so in this case, 4 is the best choice.*
- Another method is silhouette method.
  - *In this case, we need to choose the number with the highest silhouette score (also 4).*
- We can also use these methods in the h-cluster method.





Hierarchical vs. K-means

- Generally, clustering method does not support categorical variable very well and we can:
  - *Remove the variables; or*
  - *For hierarchical clustering, use some special distance-calculation methods.*
- K-means clustering requires pre-existing knowledge, i.e., the number of K.
- It is generally a good practice to scale the data before applying the clustering method, especially for hierarchical clustering.

Clustering vs. Classification

- Both the approaches are both focusing on assigning individual members to different groups.
  - *So we can still classify items using any of the method.*
- However, clustering is different from classification in that:
  - *Supervised vs. unsupervised*
    - So in clustering method, we are not training nor evaluating the results.
    - We also do not have to have predefined knowledge (i.e., the model).