Distance Measures and Hierarchical Clustering

Unsupervised ML 1

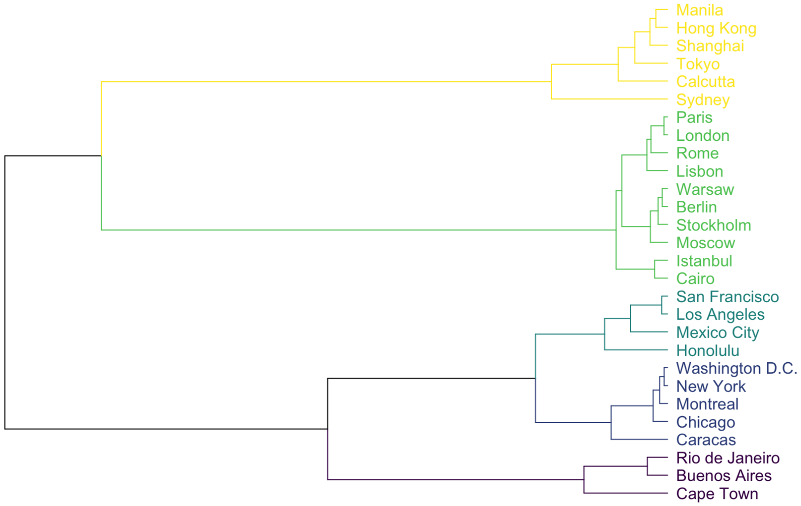
Brock Tibert

October 23, 2021

Boston University Questrom School of Business

Outline for Today

- Unsupervised Machine Learning Overview
- The Usage of Distance Metrics
- Cluster Analysis via Hierarchical Clustering (Hclust)



All of our classes will be recorded and posted to Resources > Recorded Meetings

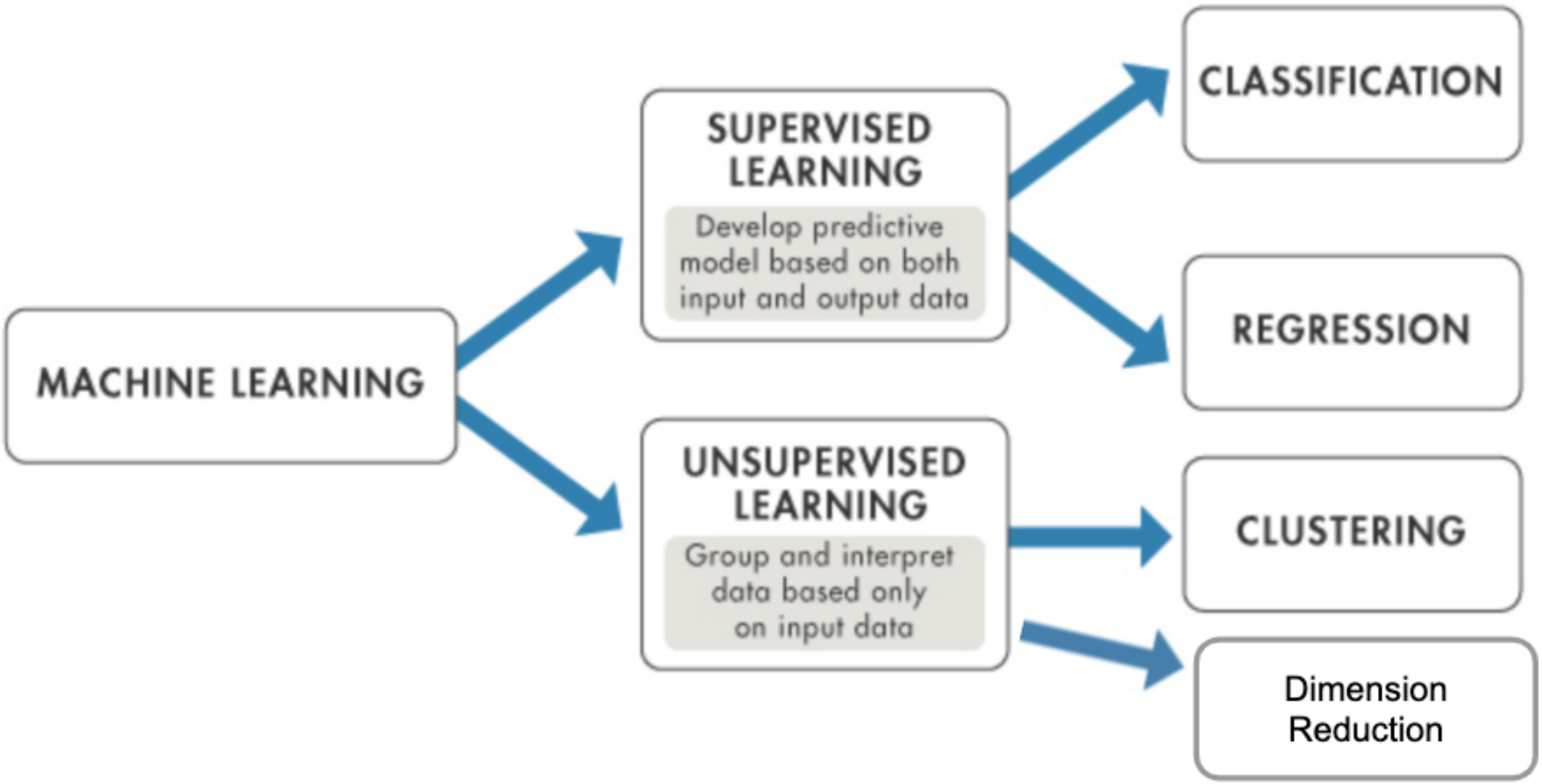
CourseKey

Check-in is done via the speakers playing a sound in class only

4 / 27

Boston University Questrom School of Business

ML Landscape - Big Picture



Boston University Questrom School of Business

Pattern Discovery

"...the discovery of interesting, unexpected, or valuable structures in large data sets."

David Hand

Unsupervised Learning - Applications

Clustering:

- Marketing Contexts for Customer Segmentation and Persona Development
- Market Segmentation for Retail Site Planning or Urban Development
- Information Retrieval on the web
- Biology (similar genes or organisms)
- Sports Analytics (Player similarities)

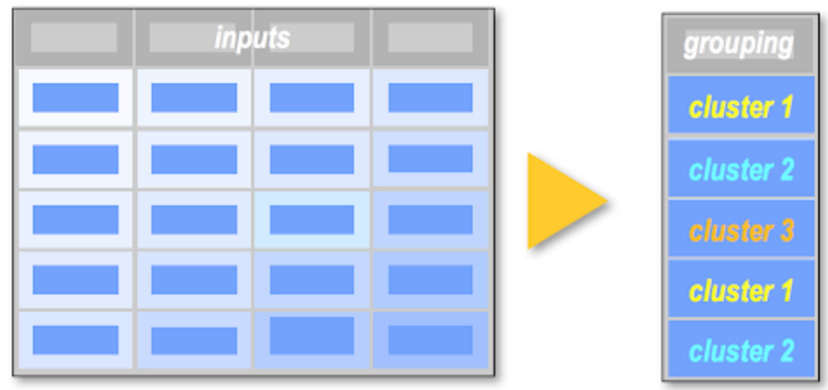
Dimension Reduction:

- Smaller search space with little loss in *information*
- Latent construct identification

UML can be used downstream in SML tasks and even help with data annotation tasks!

Cluster Analysis - Bigger Picture

- Group of cases (observations/rows) based on similarities in input values
- Unlike supervised learning, no label exists, so cluster labels are generated
  - Sometimes we elect to remove variables that could act as targets in SML tasks.
  - We do this avoid remove impact on cluster determination and to profile later.



When we have clusters, we can:

- Use the input as a categorical value for SML
- Profile the segments to tell a story and take action

BU
Boston University Questrom School of Business

9 / 27

The Two Methods We Will Explore

Hierarchical Clustering (Hclust)

- Also referred to as agglomerative clustering
- A "bottom-up" approach
- Intuitive approach and let's us as analysts determine our cluster solution

K-Means Clustering

- We set the number of clusters up-front, this is K
- The algorithm uses K to identify clusters from our search space
- This is usually done by minimizing the distance from each point to it's cluster center
- This is the topic for next week

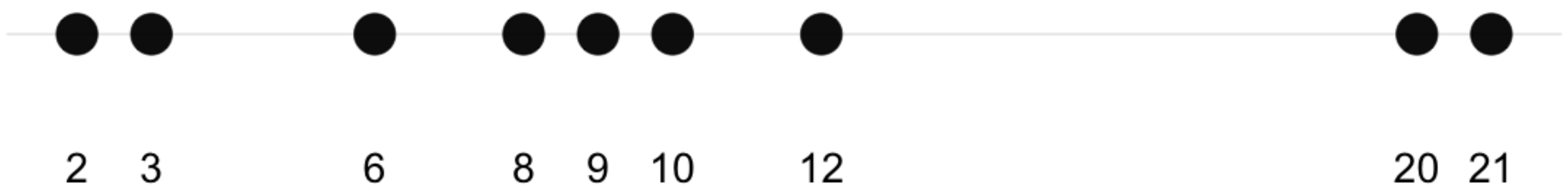
- In either case, we are using a concept of distance to join, or cluster, our records

BU
Boston University Questrom School of Business

Distance

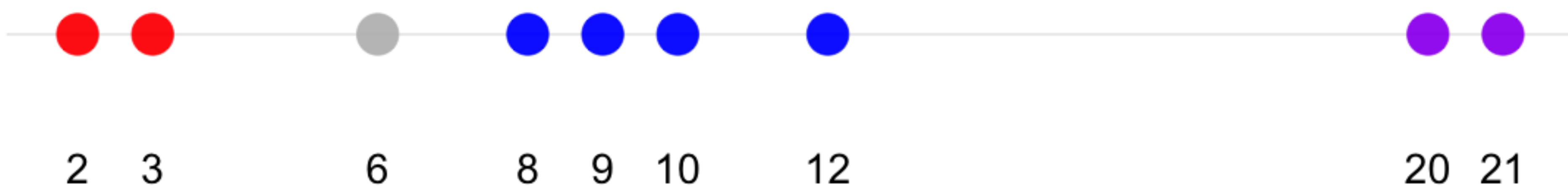
Distance Intro

Let's start with a simple, 1-dimensional example. How would you group the observations to *minimize* distance?



Distance Intro

Let's start with a simple, 1-dimensional example. How would you group the observations to *minimize* distance?



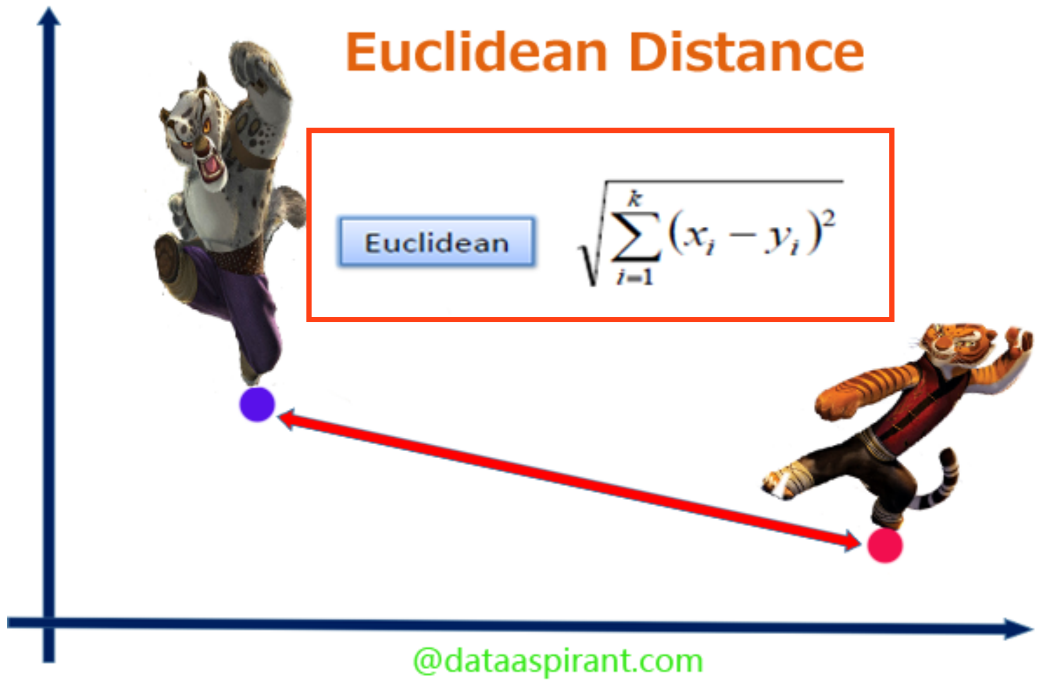
What do you think we should do with the point in grey?

Boston University Questrom School of Business

Distance Measures

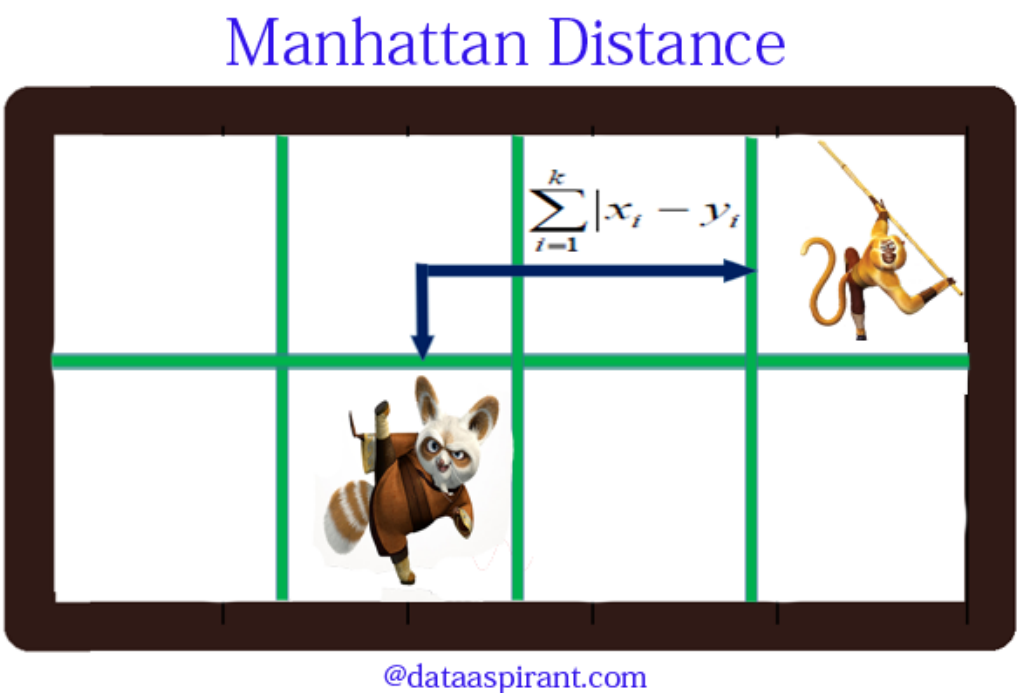
Euclidean Distance

- Straight-line Distance
- One of the most used approaches across a number of techniques
- Works for numerical inputs only



Manhattan Distance

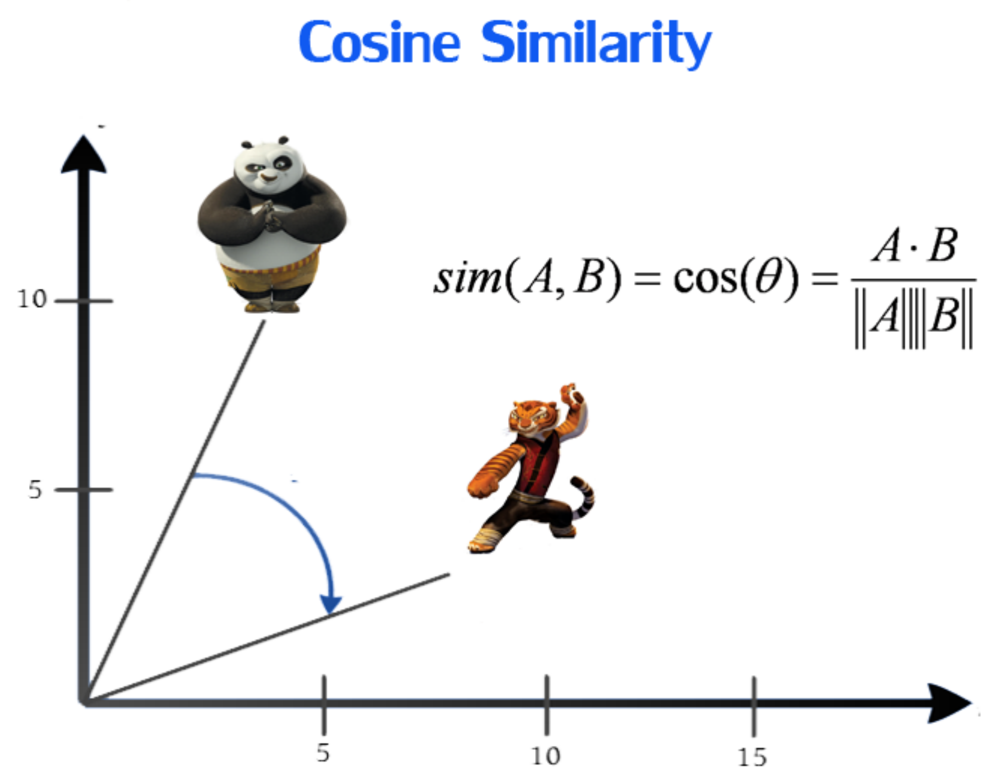
- Total line distance
- Just like navigating or walking a grid-based city (e.g. Manhattan)
- Works for numerical inputs only
- Could be better if the domain of the problem maps to grid-like issues
    - Optimizing route planning
- Also, could be a better choice if you have a **large** number of columns or want to place less emphasis on outliers



Boston University Questrom School of Business

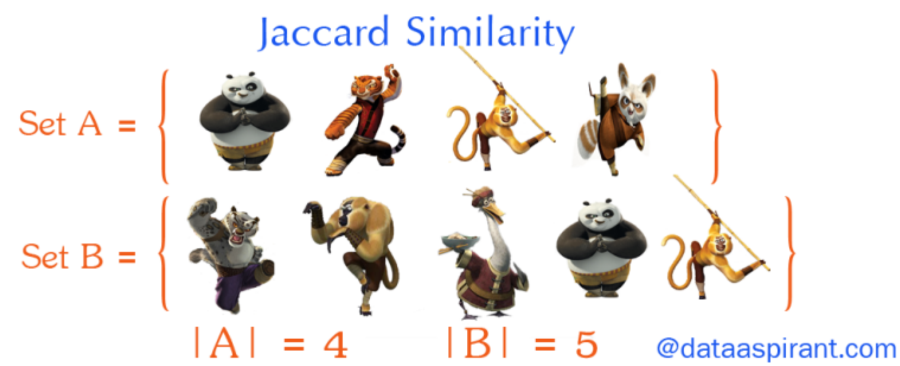
Cosine

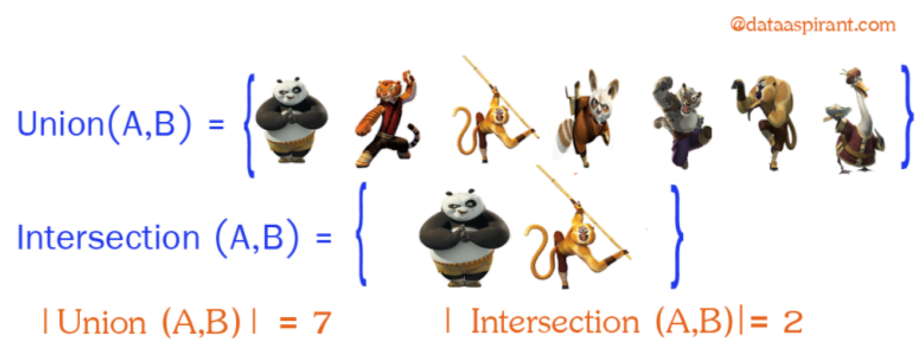
- The magnitude is not measured, but the cosine of the angle between the two vectors
- Used fairly often in recommender systems when we are calculating distance based on product ratings or when comparing word/document embeddings
  - The numeric columns are ratings of a movie or a product (e.g. Netflix or Amazon)
- For pairwise comparisons, all items/columns (embedding) are considered



Jaccard

- More appropriate for categorical data, not numeric observations on a number line
- If we have categorical data, we can make it *numeric* by dummy-encoding, also called one-hot encoding, of the data
- Think of the items as off or 0/1, where 1 is True, or "On" or "Present"
- For the pairwise comparison of records, the total items across both are considered, and use the overlap to determine how similar they are as a ratio of the total items in common





↳ become to Categorical relationships

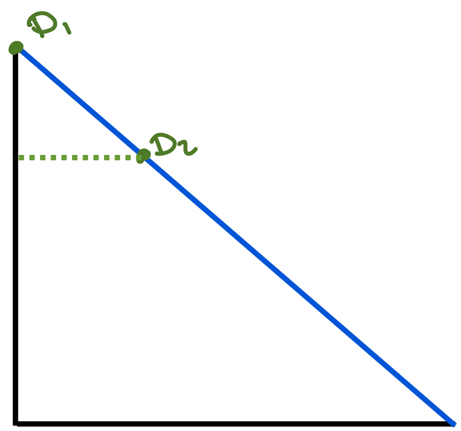
Boston University Questrom School of Business

Distance Summary

- There are many other distance metrics available, but these are the 4 that I see appear in real-world solutions (or are at least, considered in an analysis)
- As mentioned earlier, the distance measure is applied pairwise, that is, all records are compared against each other

> Question: What is the distance when a record is compared with itself? → 0

→ there are no differences in attributes

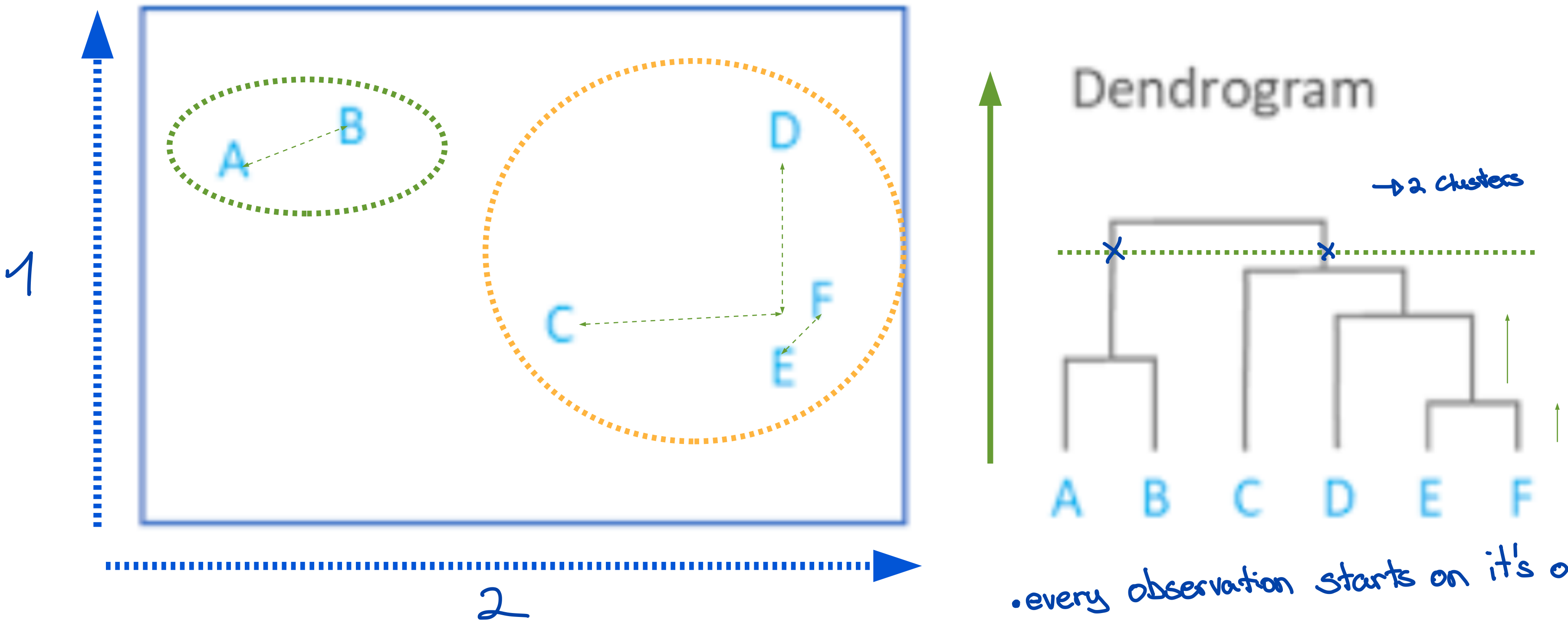


Boston University Questrom School of Business

Hierarchical Clustering

A Bottom-up Approach

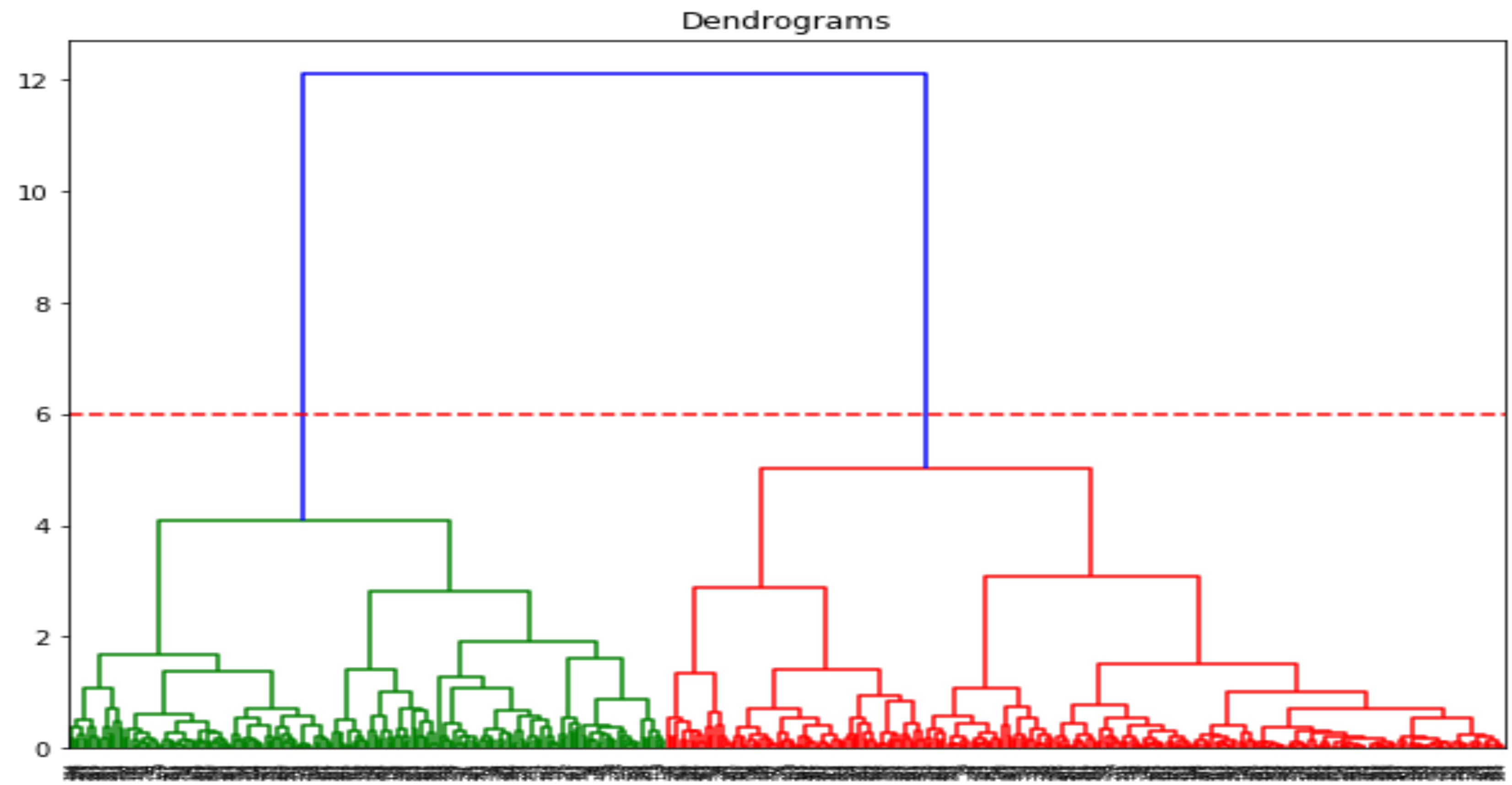
2-dimensions



• every observation starts on it's own

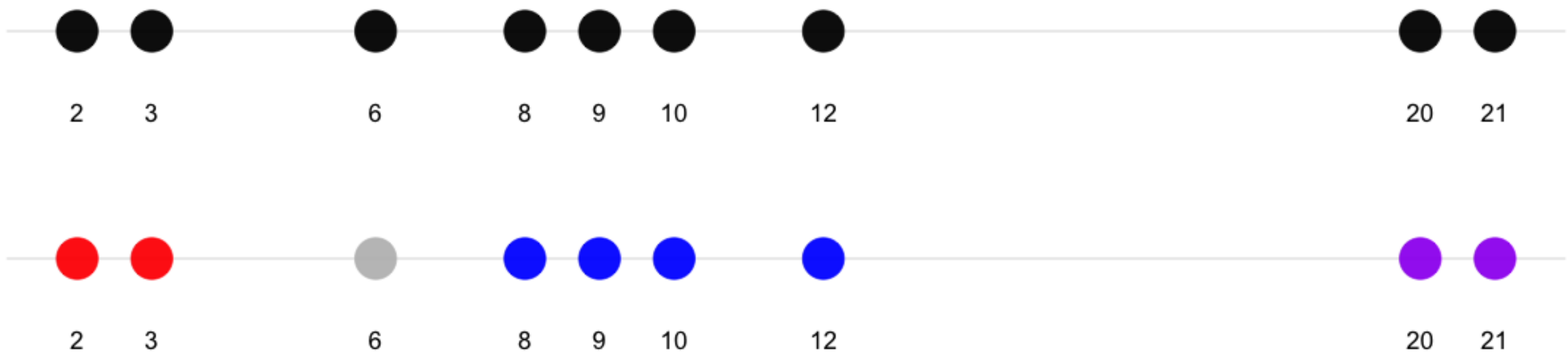
Boston University Questrom School of Business

Another Example



Boston University Questrom School of Business

Remember our 1-D Example?



How do we group/cluster point 6?

BU
Boston University Questrom School of Business

Linkage Methods

Boston University Questrom School of Business

Linkage Methods

Single Linkage

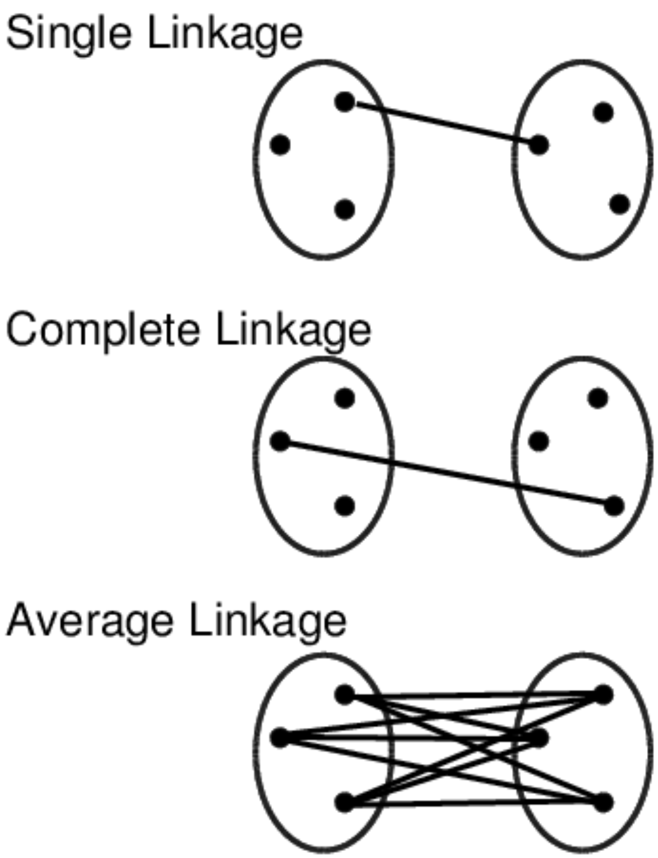
- The shortest connection between items/clusters

Complete Linkage

- The farthest connection between items/clusters

Average Linkage

- The average connection distance between all items/clusters in consideration



Boston University Questrom School of Business

Different Linkage Methods

Your turn = Classify the points

1. Single Linkage?
2. Complete Linkage?
3. Average Linkage?

Boston University Questrom School of Business

Let's write some code