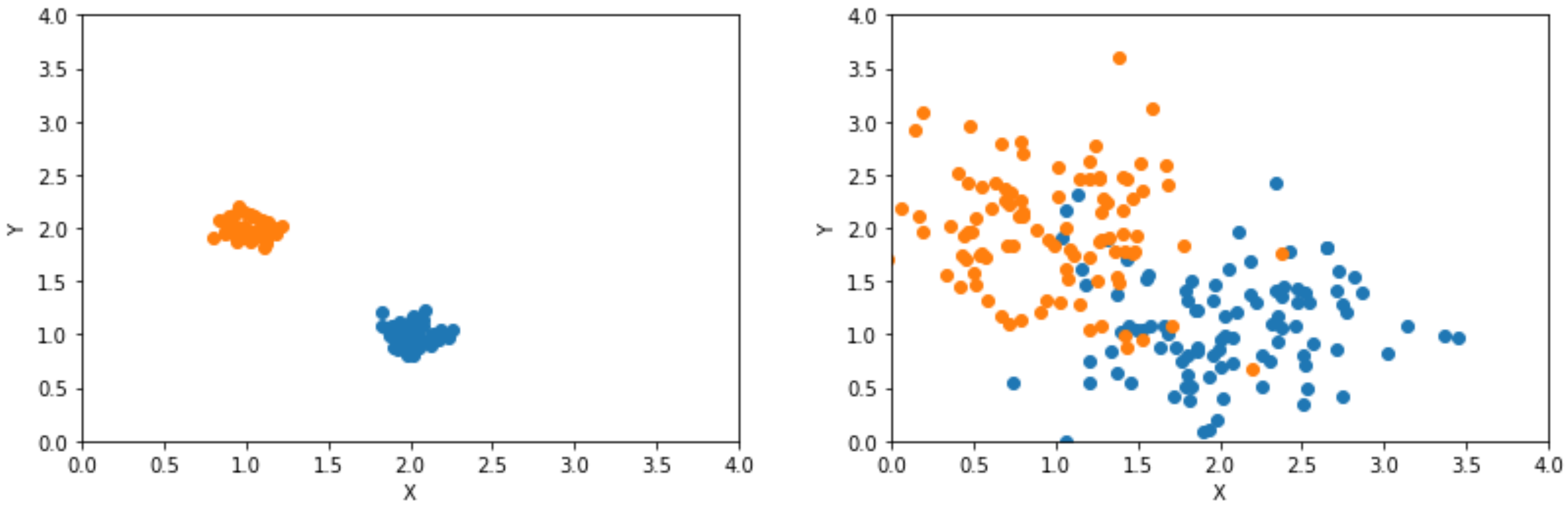
Clustering

Clustering

- Clustering is an unsupervised problem.
  - There are no targets
  - Find relationships and structure within data
- There are multiple methods used for clustering.
  - They all share the concept of distance
    - Clustering is about how “close” data is.
    - Close means a distance is measured.
    - Distance can be a real distance or a mathematical distance.

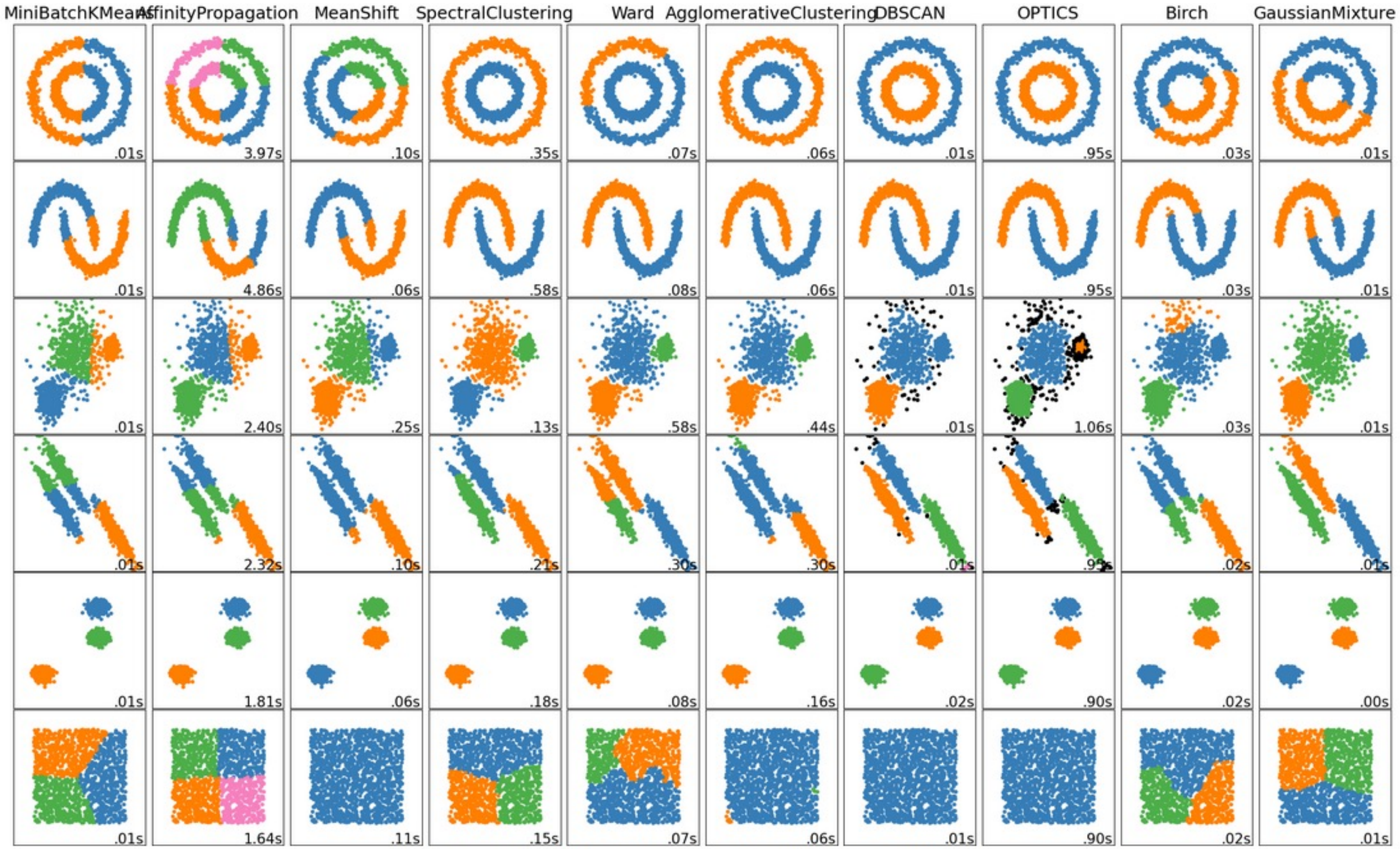
Clusters of Data



- Both clusters are centered at (2,1) (blue) and (1,2) (orange) but the distribution of the data is much tighter on the left. For the graph on the right, are there 2 distinct clusters? There are 2 distinct distributions!
- How many clusters are there? How do we decide?

Clustering Algorithms

- Different algorithms produce different results; some will classify outliers (DBSCAN, OPTICS) or try and guess the number of clusters
- Others require careful selection of initial conditions
- All the algorithms do well with isolated tightly grouped data; when the cluster shapes become complex is when the results diverge



A comparison of the clustering algorithms in scikit-learn

DataScience@SMU

Distance

Distance

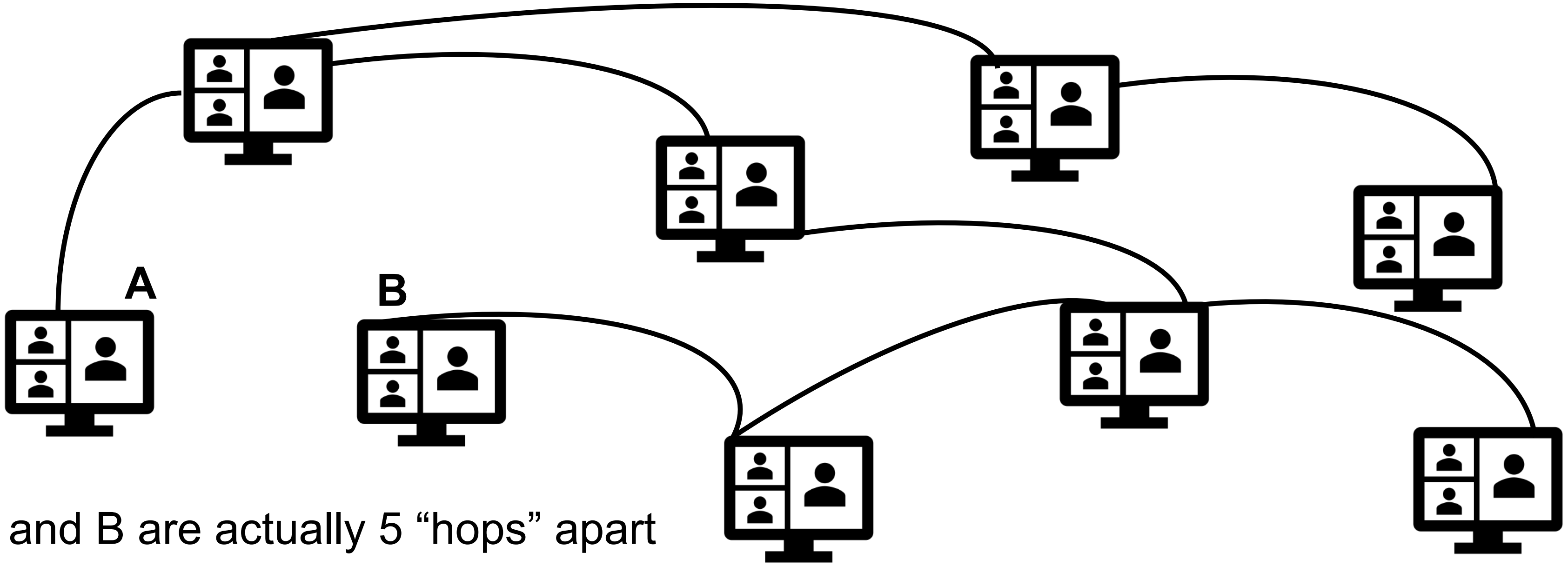
- While the basic concept of distance is familiar from the most basic courses in math, it does get an update.
- Standard distance (Euclidian distance): $(\sum_{i=1}^{n}(X_{1i} - X_{2i})^2)^{\frac{1}{2}}$
  - Aka the shortest distance between two points is a straight line.
- Think of the data as vectors in an abstract space.
- Data that is “nearby” will have a small distance:

$$X_{1i} = (x_1, y_1, z_1, etc....)$$

Nonstandard Distance

Manhattan

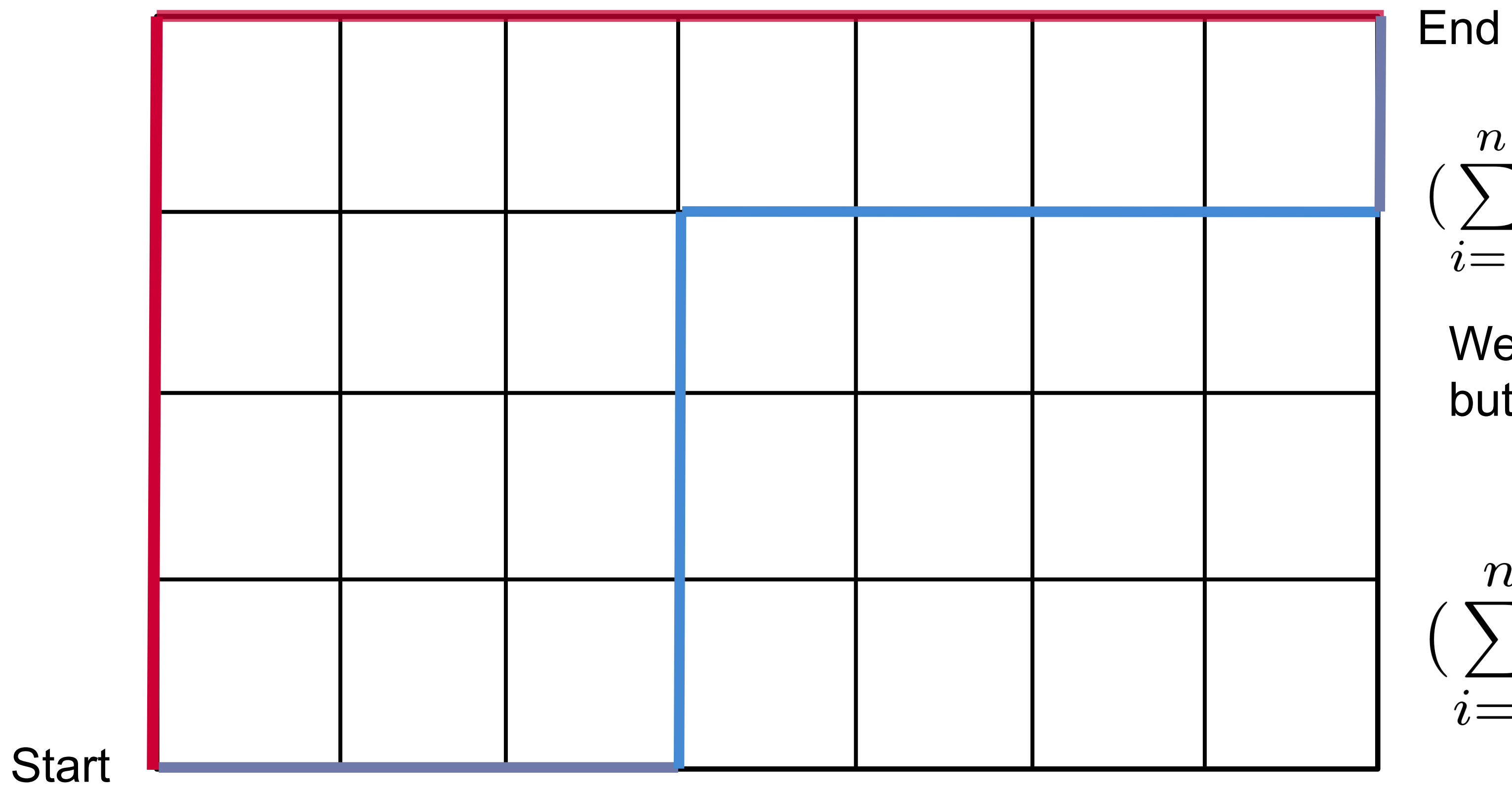
- How many “hops” between points
- Useful in graphs and networks
- Named after the grid of blocks in New York City



Computers A and B are actually 5 “hops” apart

Manhattan Distance Diagram in 2D

You may only "travel" along the grid, thus the red and blue paths have equal length (and the shortest path may not be unique!)



$$\left(\sum_{i=1}^{n} |X_{1i} - X_{2i}|^1\right)^{\frac{1}{1}}$$

Weird way to write it—but look at the pattern!

$$\left(\sum_{i=1}^{n} |X_{1i} - X_{2i}|\right)$$

A General Distance Formula

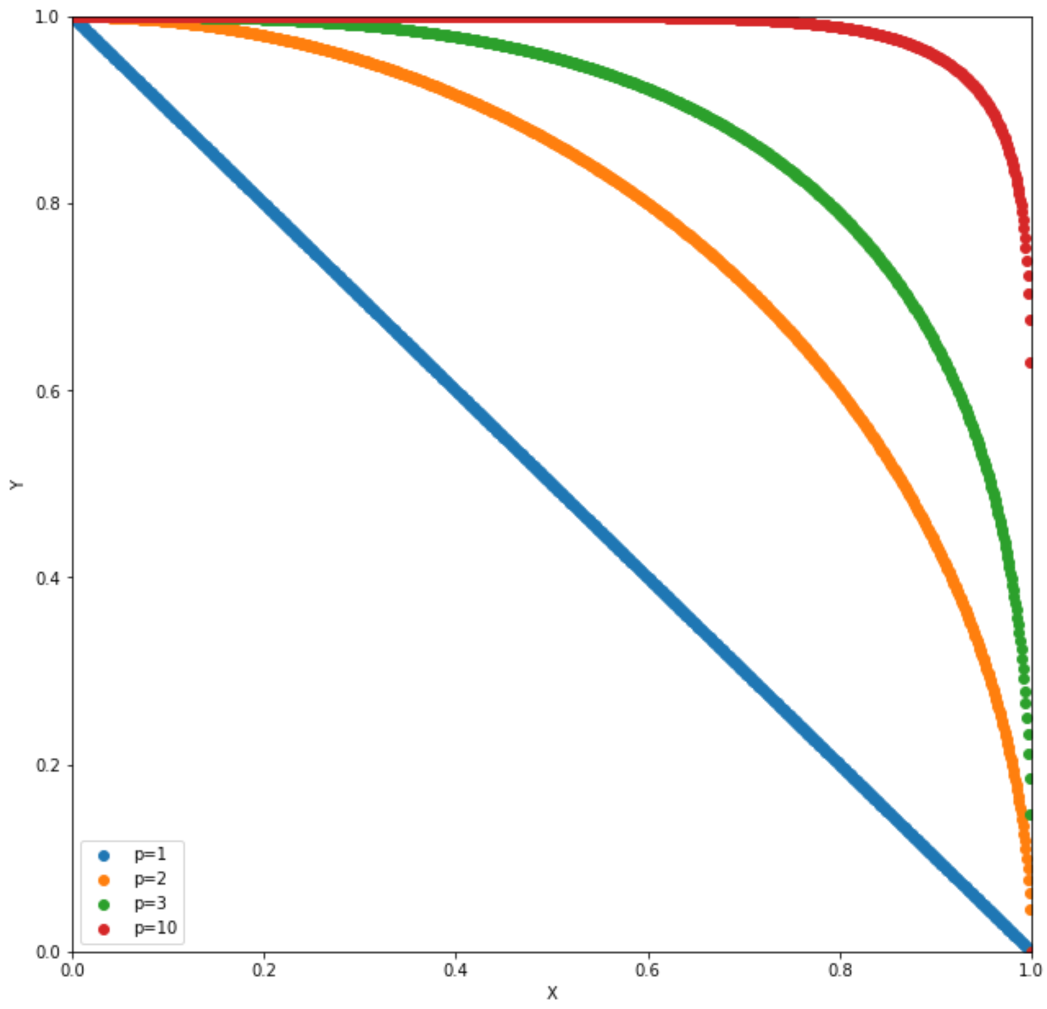
- Minkowski distance

$$\left(\sum_{i=1}^{n} |X_{1i} - X_{2i}|^p\right)^{\frac{1}{p}}$$

- p = 1, Manhattan distance
- p = 2, Euclidean distance
- p = ∞, Chebyshev distance
  - Not a curiosity—used in warehouses where cranes can move in x, y, z independently

Shapes of Equal Distance for Various P

- We can see that as p increases, the shape of positions that are the same distance from the center becomes a square (only 1st quadrant shown here).
- The p = 1 would be a diamond, while p = 2 is the traditional circle.
- The Chebyshev distance would be when something like a crane independent motors (move in X, Y at the same time). Thus, to get to any point on X = 1, from Y = 0 to Y = 1, take the same time (and the time in this case would be the distance metric).

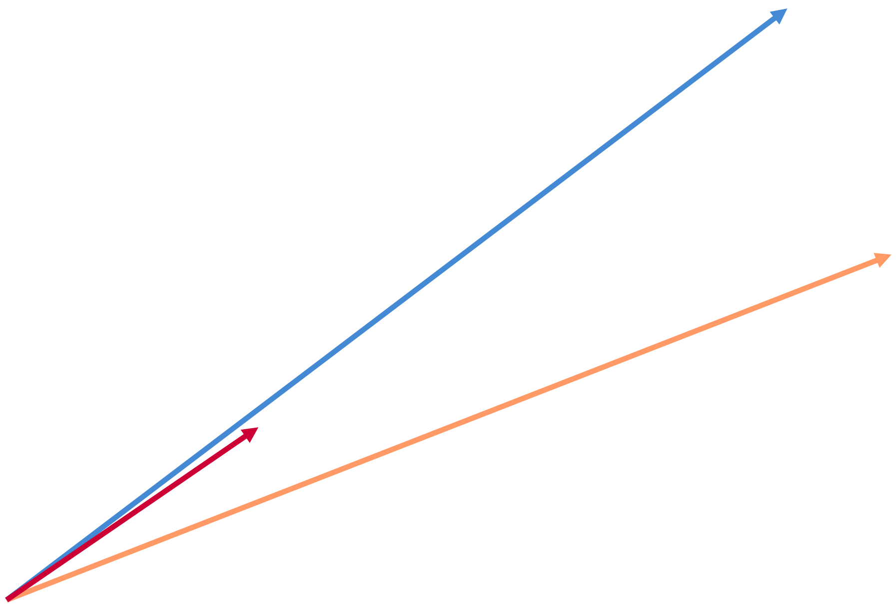


Cosine Distance

- Many times we want to look not at the vector position but the vector relationship, or if two vectors “point” in the same direction.
- If they point in the same direction, the angle between them is small.
  - If we take the cosine of that angle, we say similar vectors have a cosine similarity close to 1 ($\cos 0 = 1$).
  - If we want a **distance**, distance needs to be positive, and nearby things need to have a small distance. Thus, cosine **distance** is 1—cosine similarity.

Cosine Similarity and Distance

- Let’s say the angle between red and blue vectors is 2 degrees, while blue and orange are separated by 20 degrees.
- The cosine similarity of red and blue will be .999 and the cosine distance will be 0.001. Blue and orange have a cosine similarity of 0.939 and cosine distance of 0.061.
- However, if we give a rough estimate of Euclidean distance, blue and orange is about 3x closer than blue to red.
- Subjects like NLP use cosine distance.



Don’t forget about non-Euclidean geometry either!

DataScience@SMU