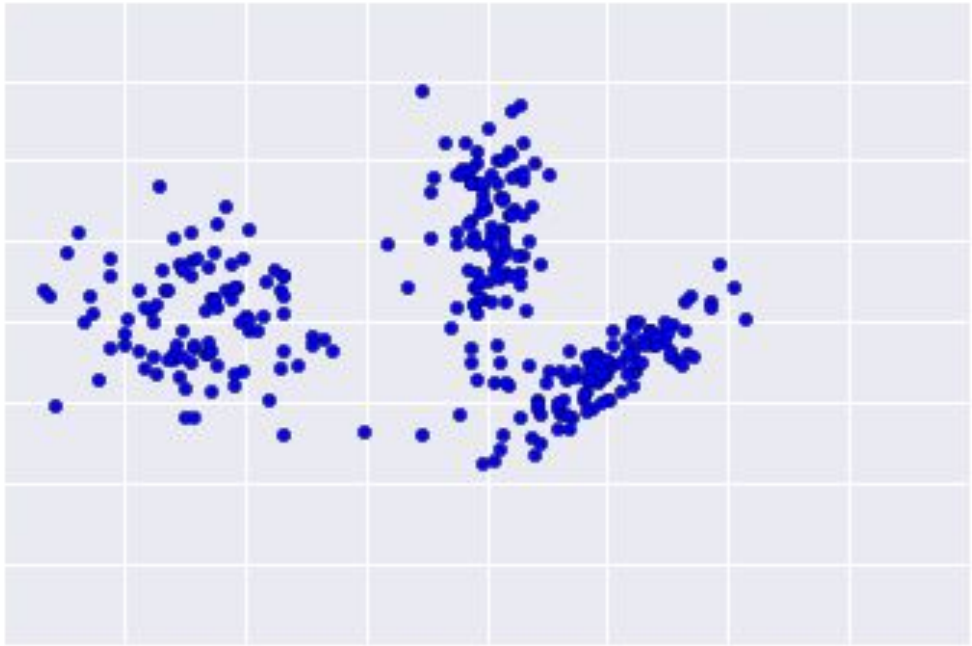
Clustering

Exploratory data analysis

In the last class, we have seen how PCA can be used to for exploratory data analysis.

Another possible way to explore the data is to see if it contains clusters.

Clustering



What is a cluster

Clusters are intuitively simple to understand, but harder to define mathematically.

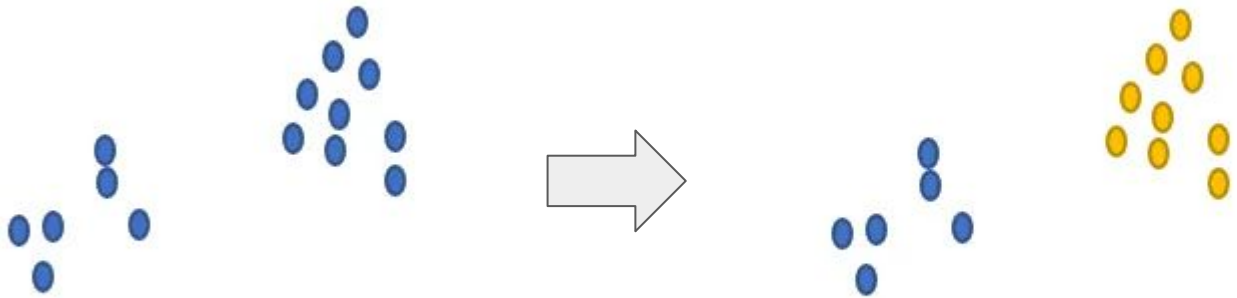
Clusters are groups of similar points.

Points in the same cluster have a higher similarity than points in different clusters.

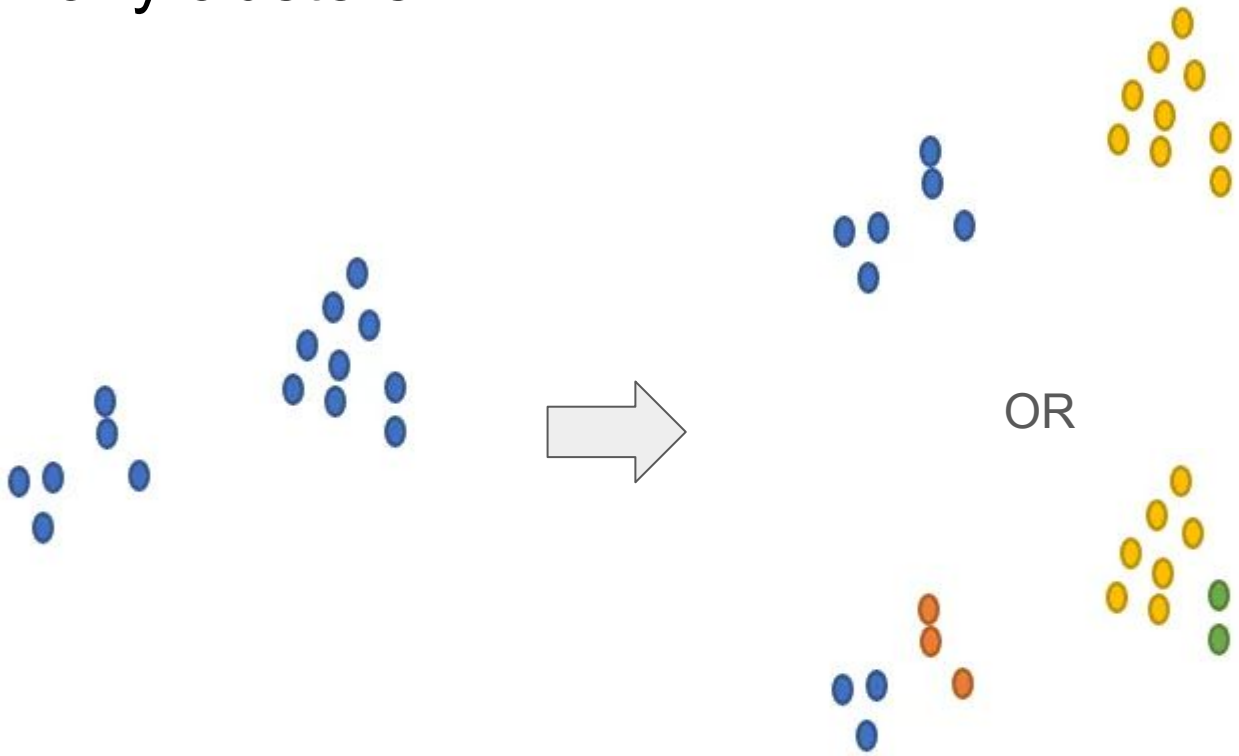
Here, we use distance as a measure of similarity.

We can cluster observations or variables

How many clusters?



How many clusters?

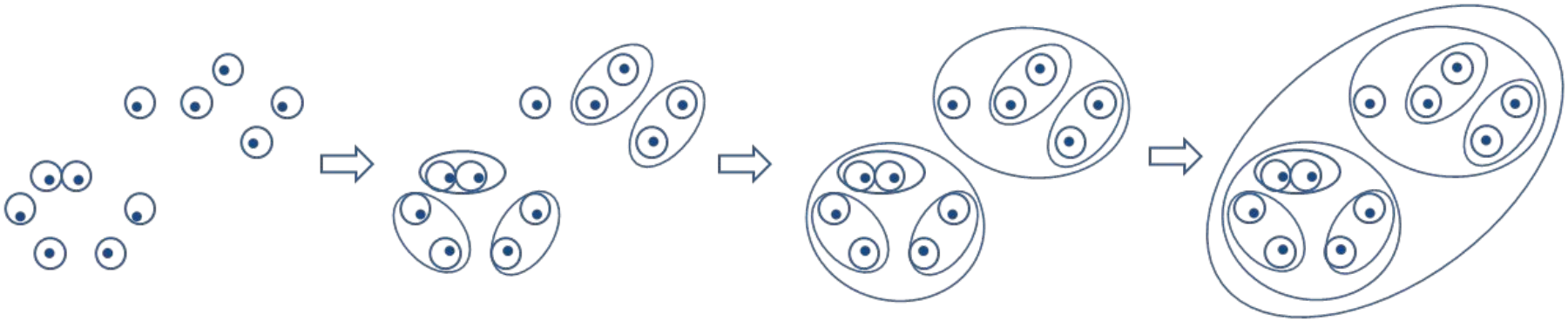


Hierarchical clustering

Hierarchical clustering avoids answering this question by giving you all possible answers at once!

The output of a hierarchical clustering algorithm is a hierarchy of cluster sizes (as the name suggests).

Hierarchical clustering



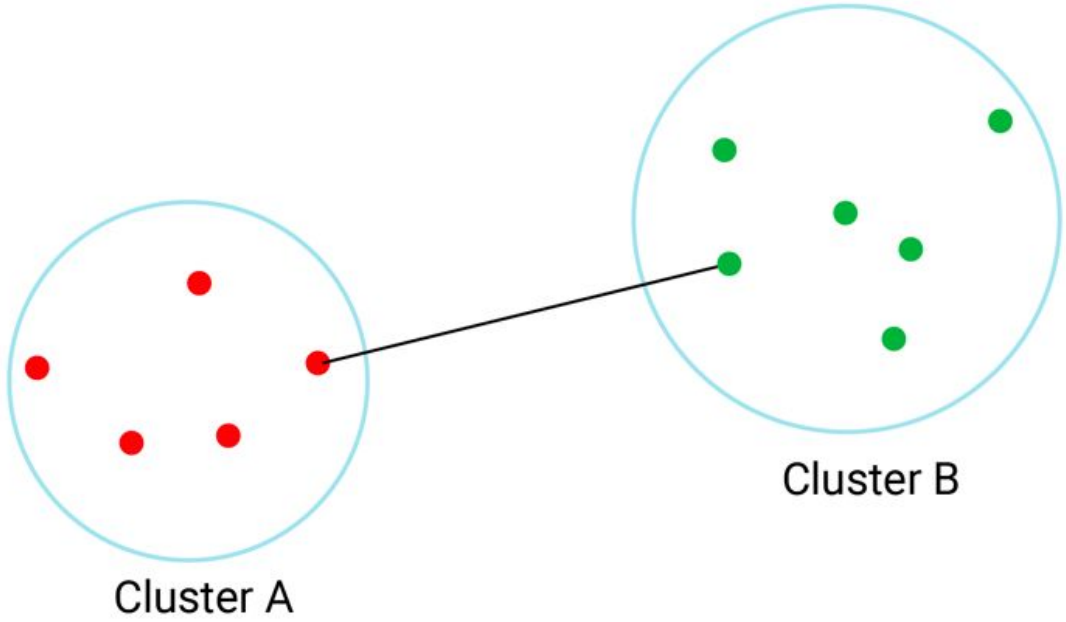
Linkage

Linkage is the measure of distance **between two clusters**.

It is based on the distance between points of the two clusters.

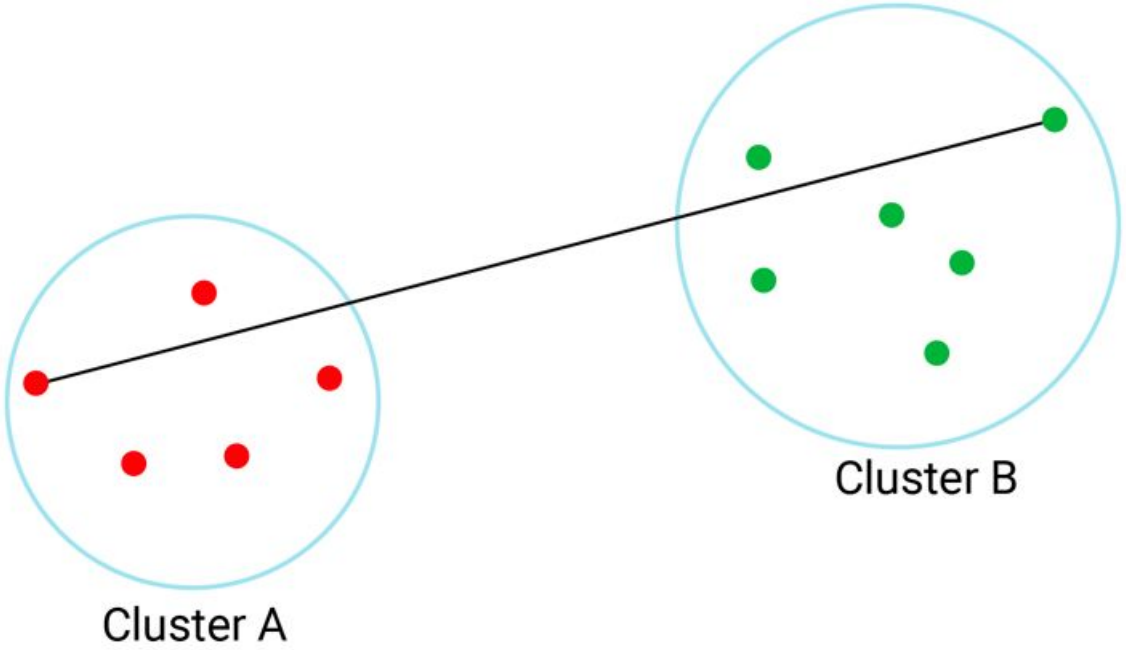
Single Linkage

The distance between two clusters is the **smallest** distance between any pair of points in different clusters



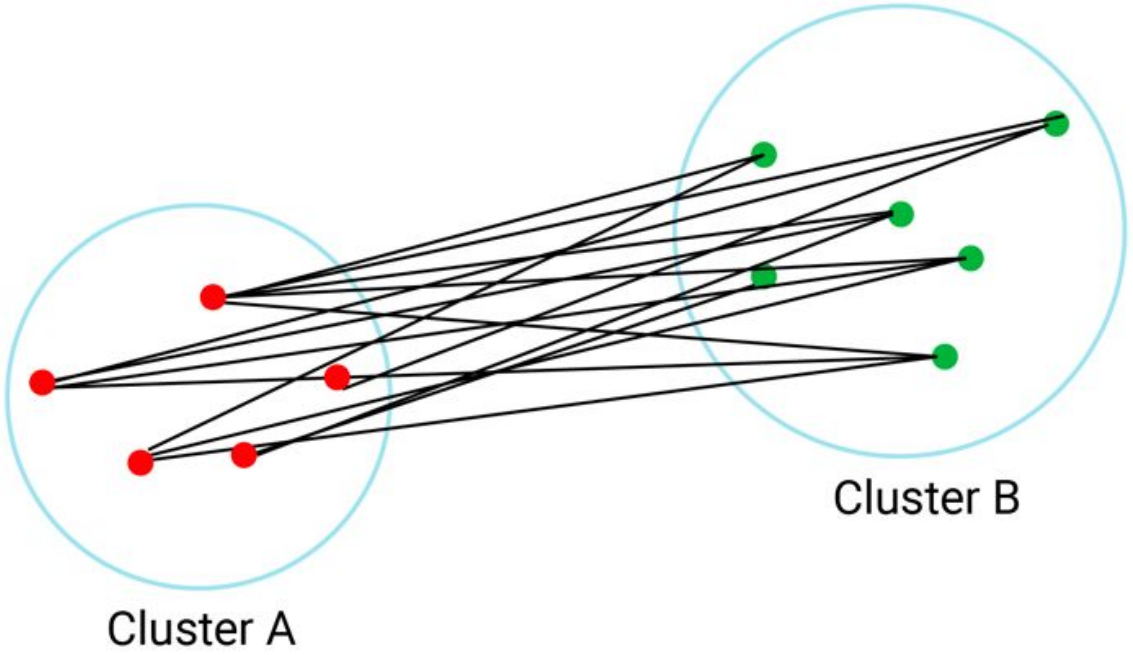
Complete linkage

The distance between two clusters is the **largest** distance between any pair of points in different clusters



Average linkage

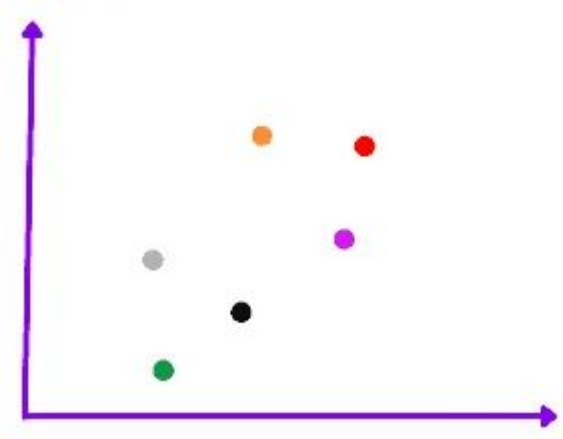
The distance between two clusters is the **average** distance between any pair of points in different clusters

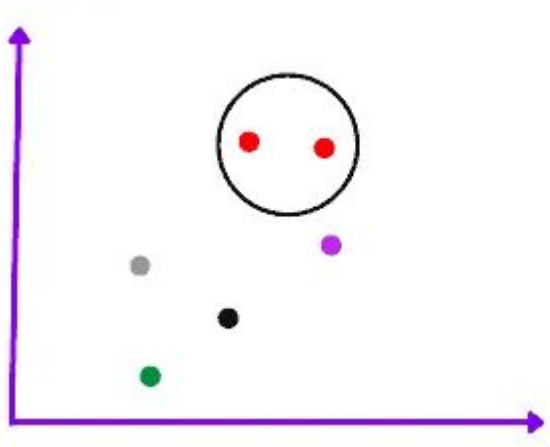


Hierarchical clustering steps

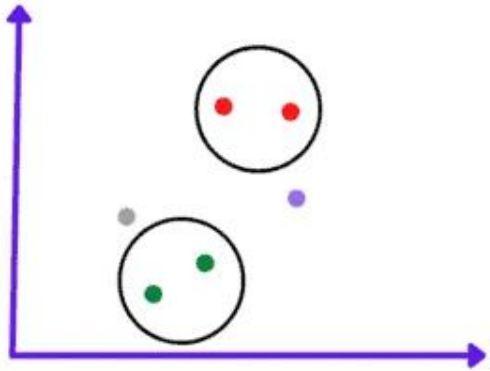
1. Start with each point in its own cluster
2. Calculate the linkage between any two clusters
3. Merge the two closest clusters
4. Recalculate the linkage
5. Repeat steps 3 and 4 until only one cluster is left

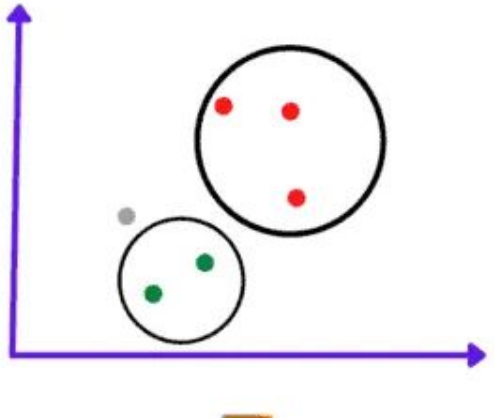
Hierarchical clustering steps



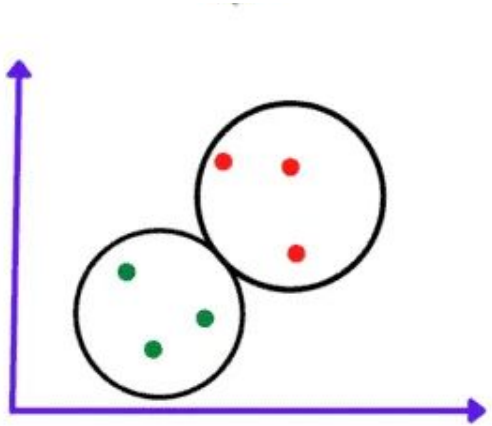


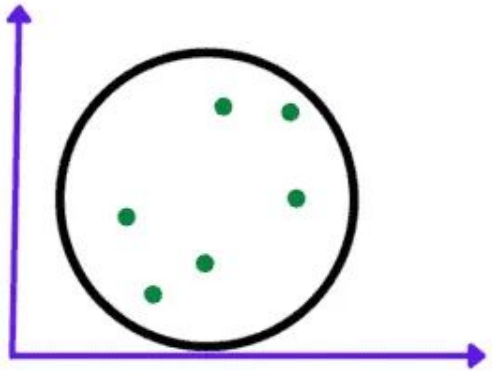
Hierarchical clustering steps



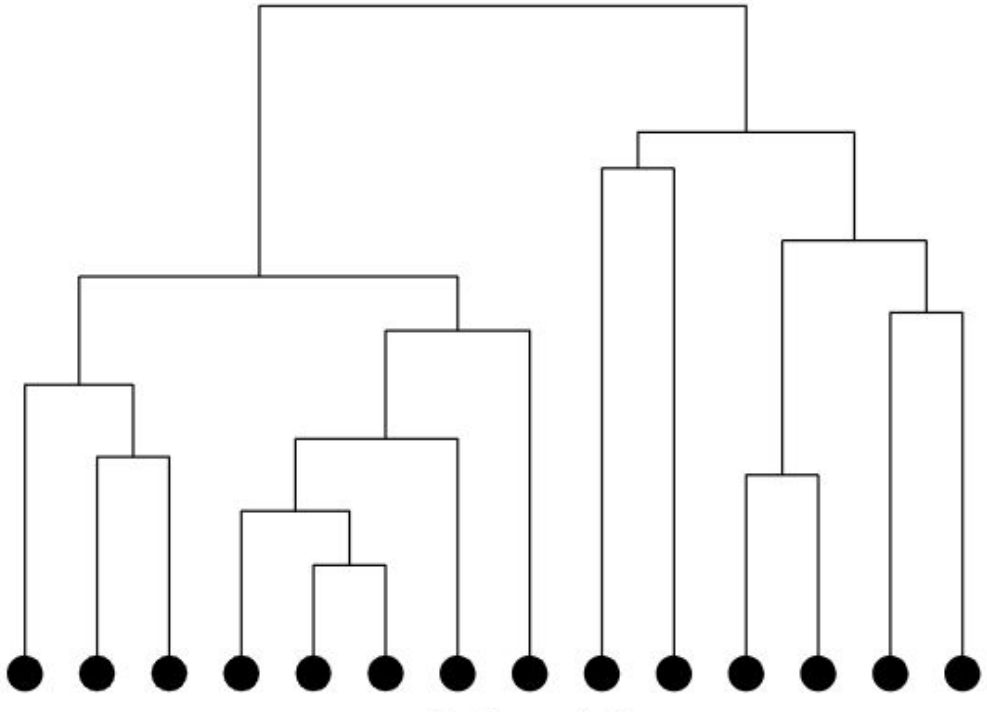


Hierarchical clustering steps

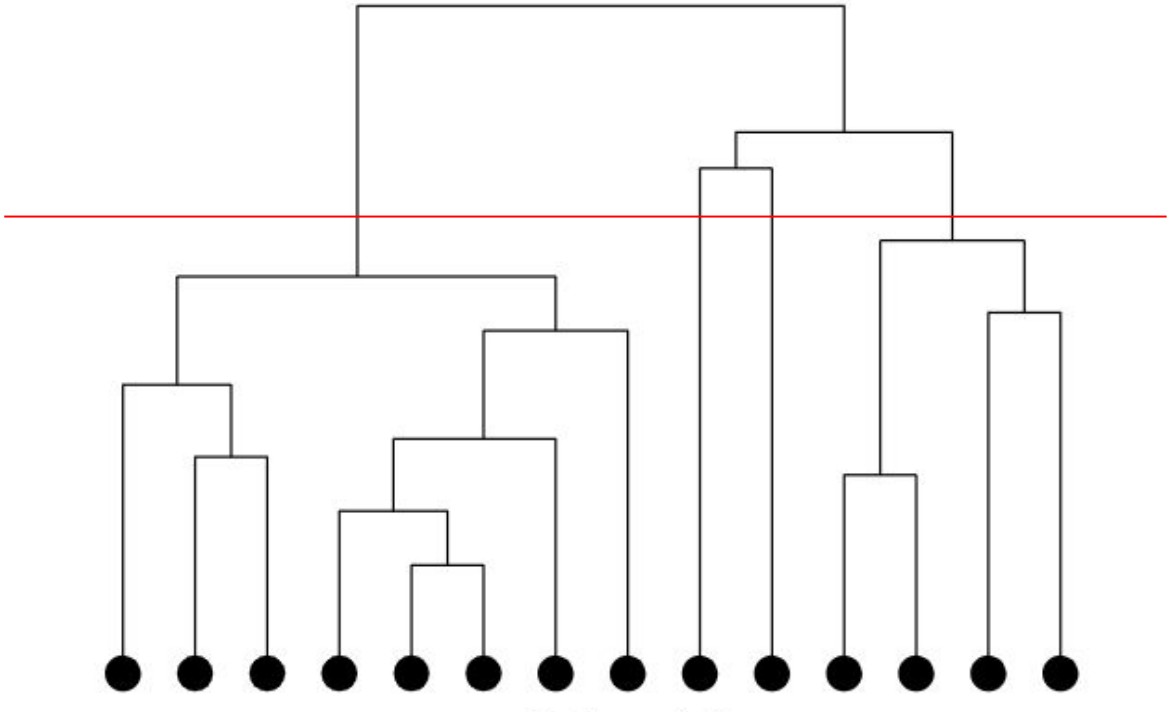




The dendogram

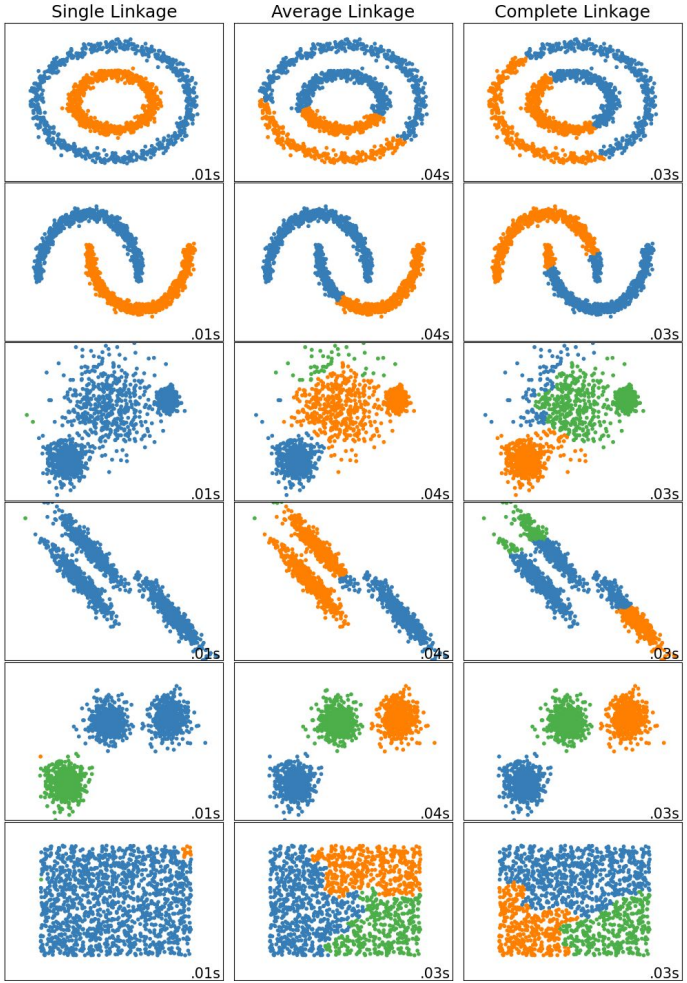


The dendogram



Disadvantages of Hierarchical clustering

1. Expensive to compute: all pairwise distances need to be computed.
    a. Not feasible in large data sets
2. Difficult to interpret.
    a. Sensitive to choice of linkage.
    b. Dendogram is sensitive to outliers.


Single Linkage
Average Linkage
Complete Linkage
.01s
.04s
.03s
.01s
.04s
.03s
.01s
.04s
.03s
.01s
.04s
.03s
.01s
.04s
.03s
.01s
.03s
.03s

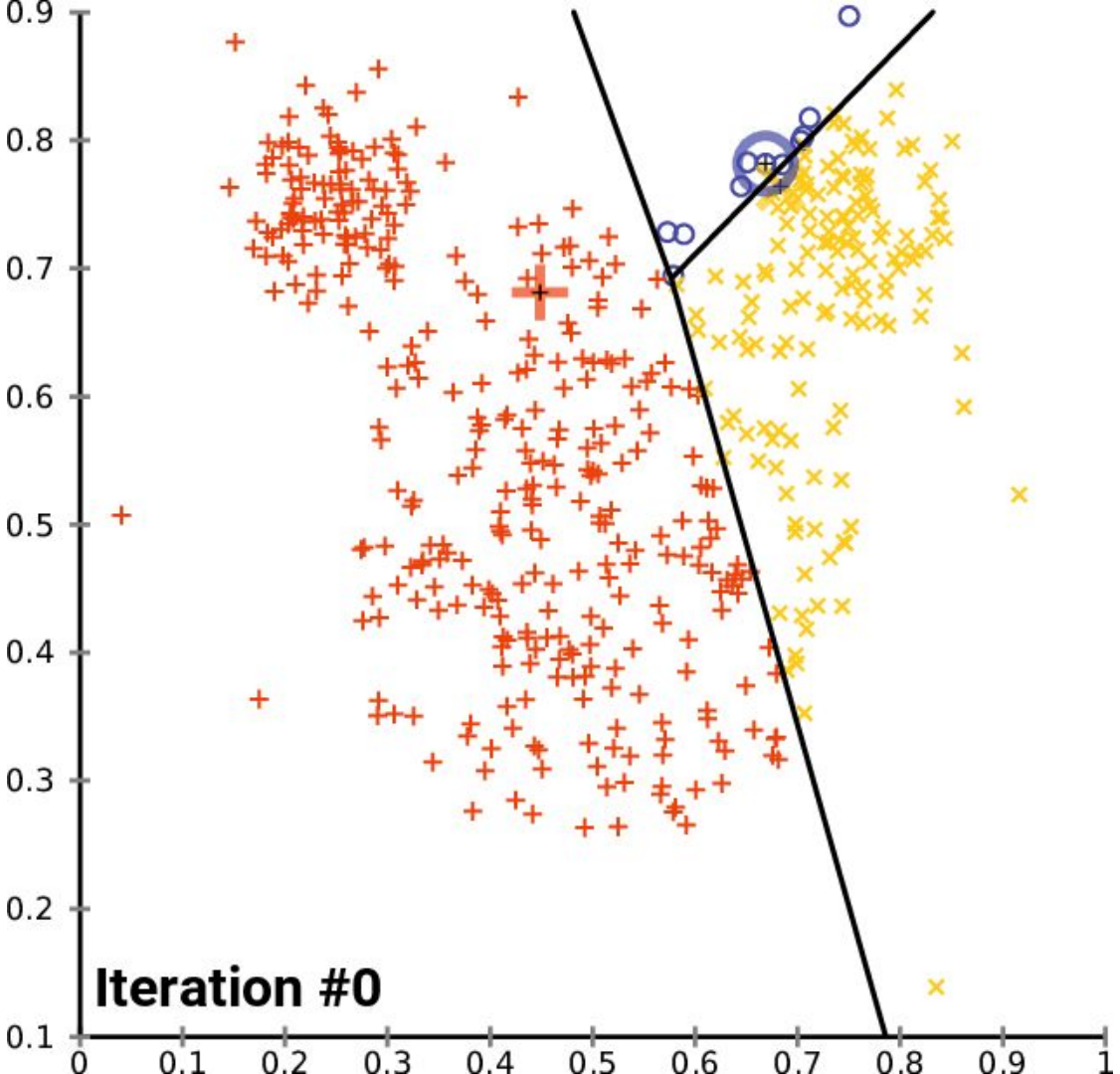
K-means clustering

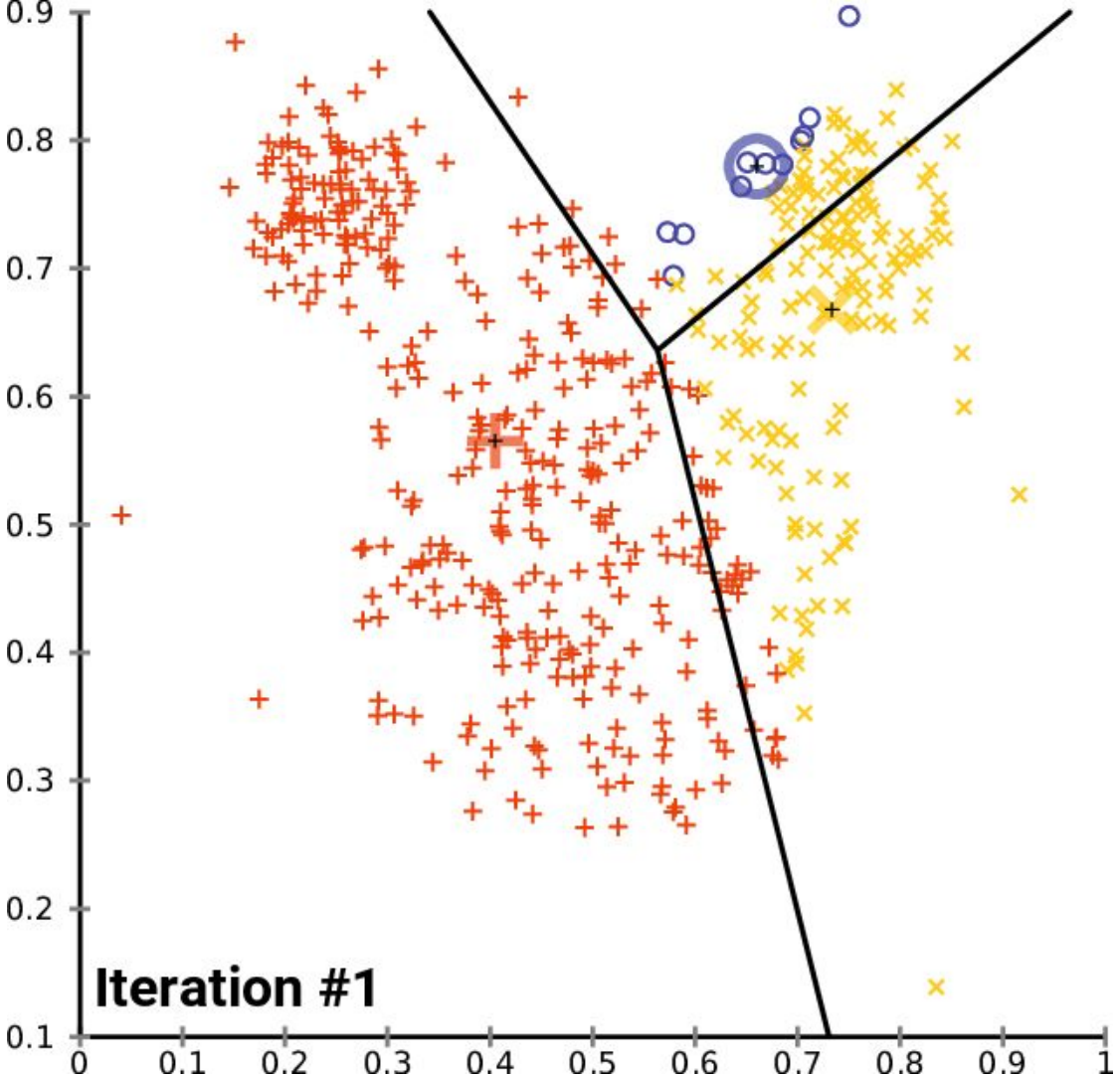
This clustering algorithm is a lot faster to run, but comes with some extra assumptions.

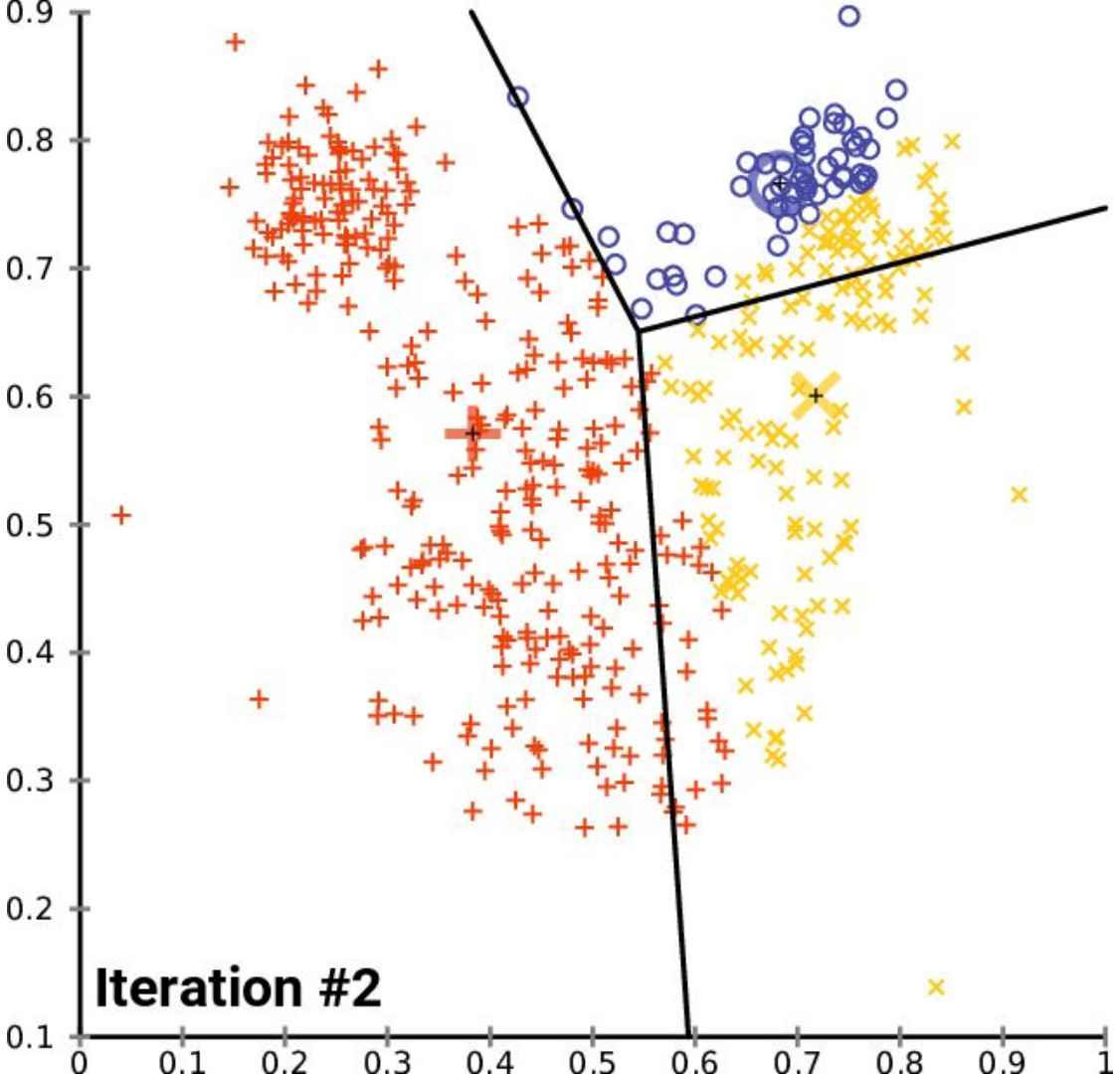
The first is that we need to know in advance how many cluster we expect in the dataset.

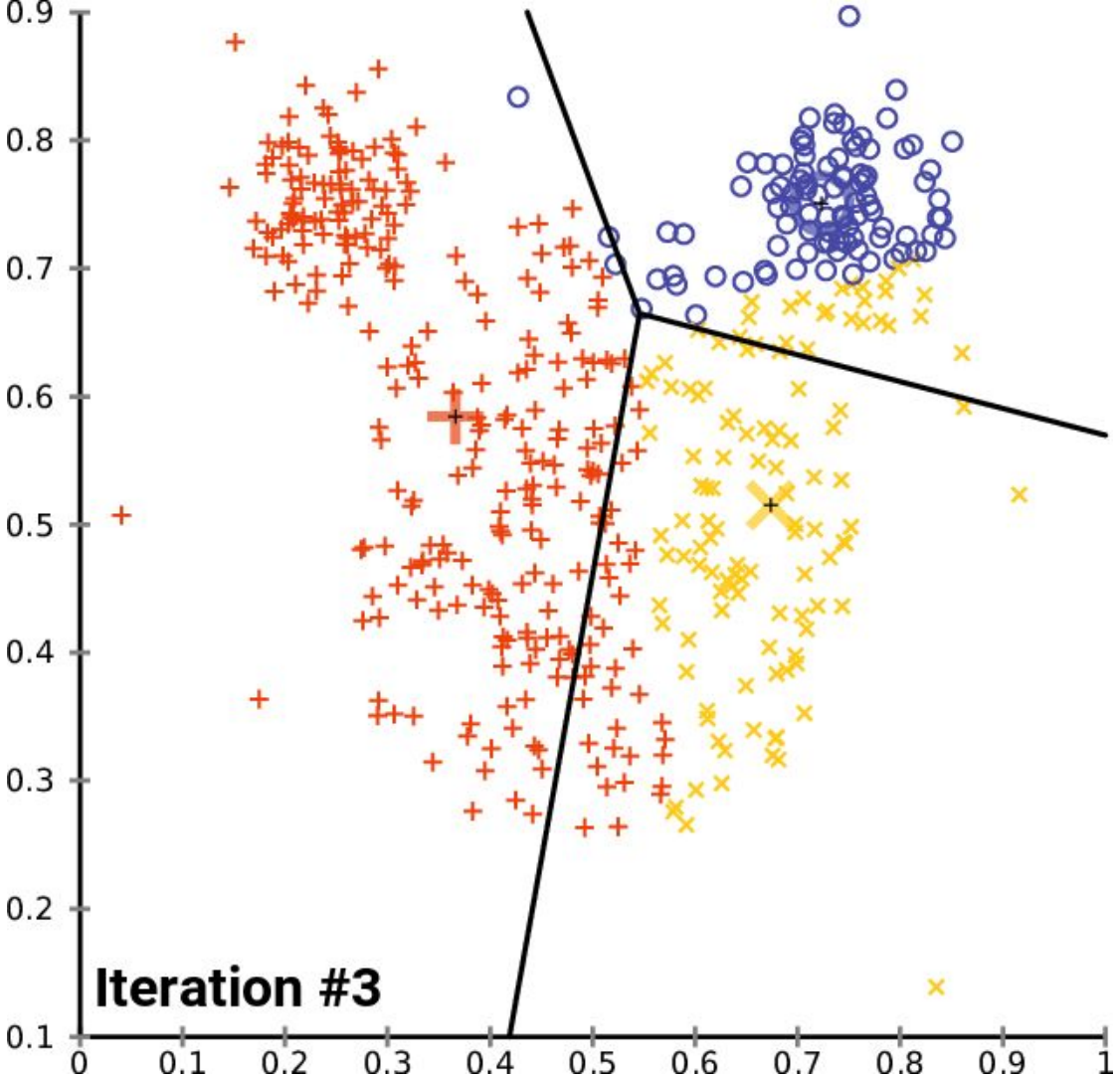
The algorithm

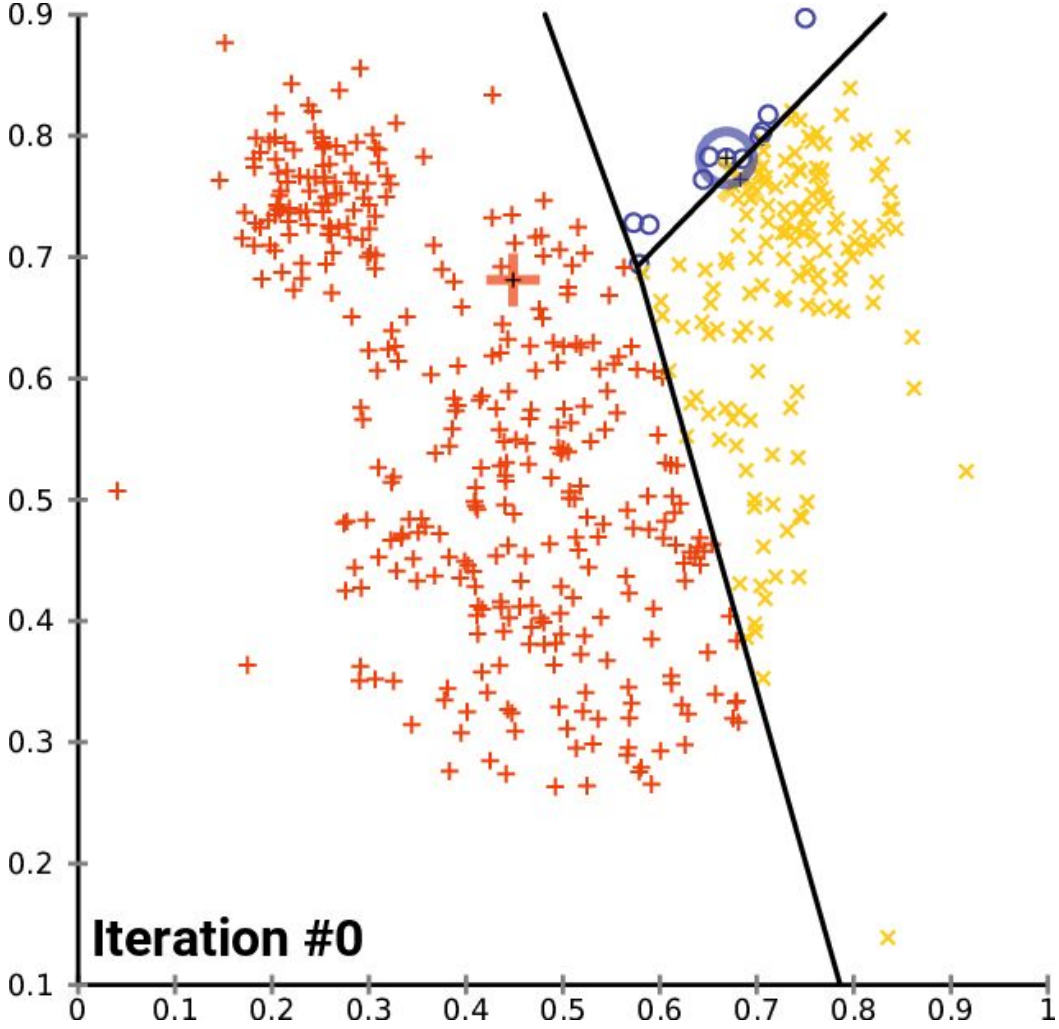
1. Select the number of clusters: **k**
2. Pick **k** random points, these will be your initial cluster means.
3. Assign each data point to a cluster, by calculating which cluster mean is closest to it.
4. Recalculate the mean of each cluster.
5. Repeat steps 3 and 4 until the means do not change.


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Iteration #0


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Iteration #1


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Iteration #2


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Iteration #3


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Iteration #0

Disdvantages

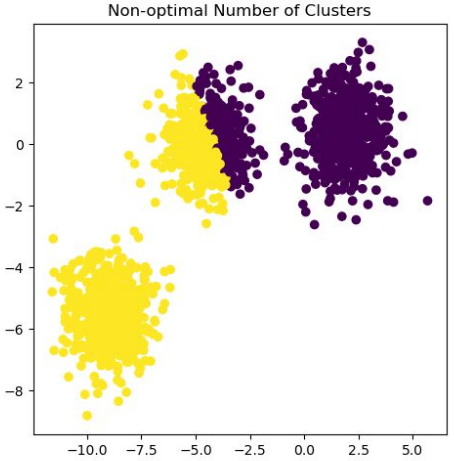
Contains extra assumptions

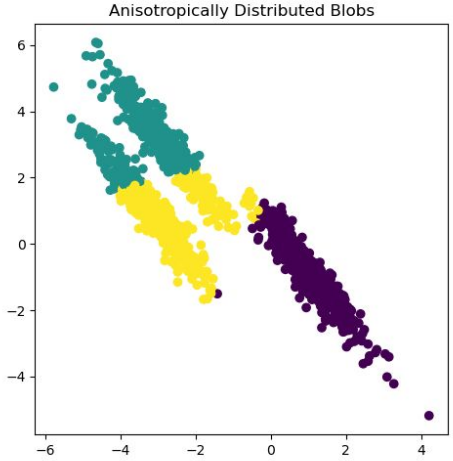
- Clusters are "spherical"
- Clusters have the same number of points
- Clusters have the same variance

It also relies on random initialization, which can make results hard to reproduce.

Unexpected KMeans clusters

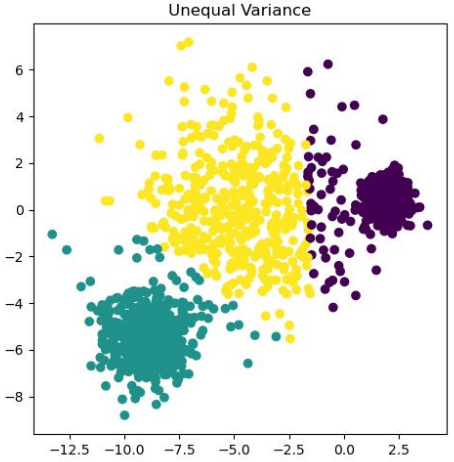
Bad number of clusters

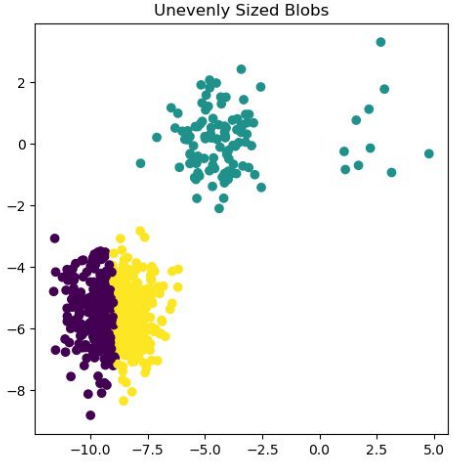




Non-spherical clusters

Unequal variance



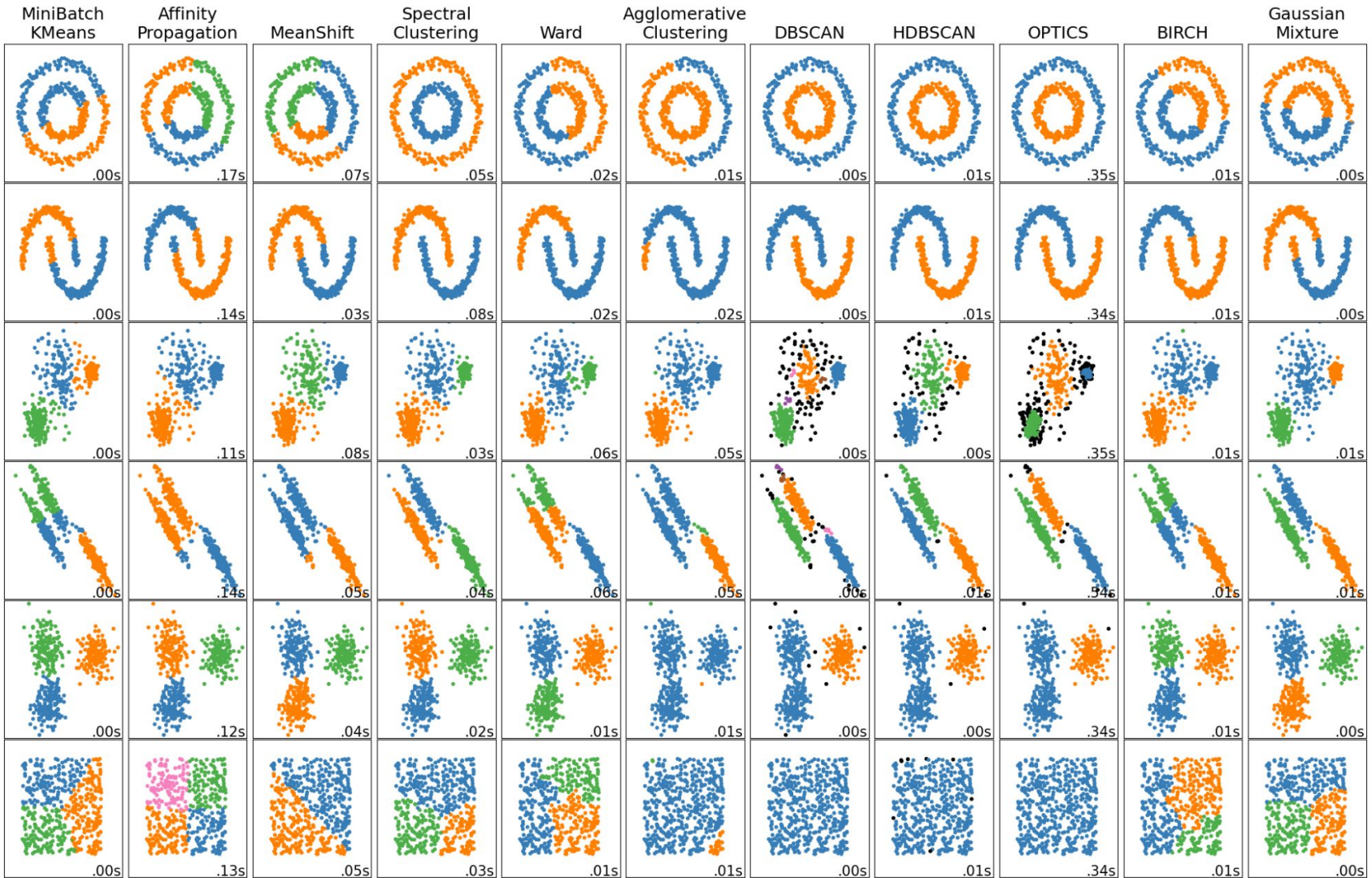


Unequal cluster size

Other clustering methods

Because of the difficulty to define clusters, many different clustering algorithms were developed.

They each have their strengths and weaknesses, and no algorithm performs better overall.


MiniBatch KMeans
Affinity Propagation
MeanShift
Spectral Clustering
Ward
Agglomerative Clustering
DBSCAN
HDBSCAN
OPTICS
BIRCH
Gaussian Mixture
.00s .17s .07s .05s .02s .01s .00s .01s .35s .01s .00s
.00s .14s .03s .08s .02s .02s .00s .01s .34s .01s .00s
.00s .11s .08s .03s .06s .05s .00s .00s .35s .01s .01s
.00s .14s .05s .04s .06s .05s .00s .01s .34s .01s .01s
.00s .12s .04s .02s .01s .01s .00s .00s .34s .01s .00s
.00s .13s .05s .03s .01s .01s .00s .01s .34s .01s .00s

Data normalization

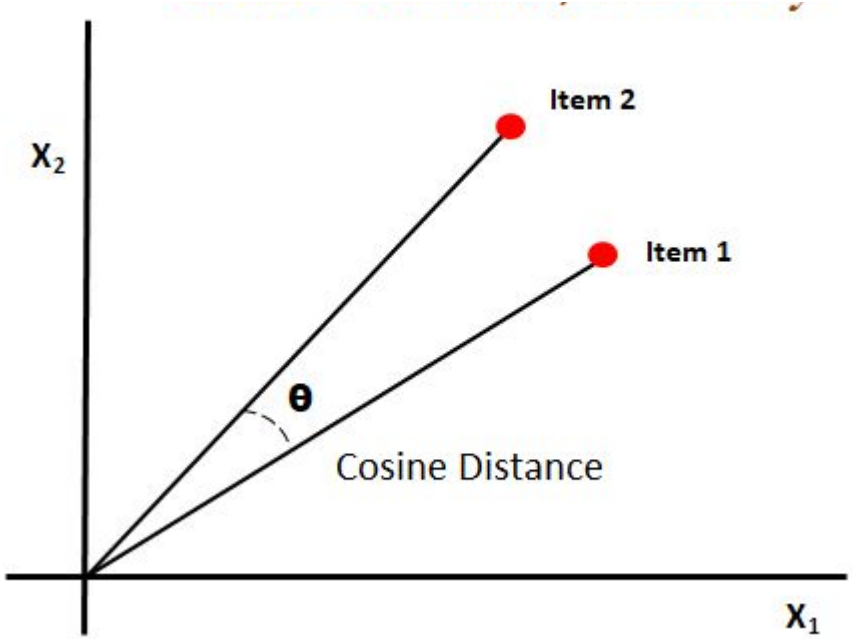
You should also standardize your data before clustering, such that all variables have the same variance.

Question: Why?

Cosine similarity

Useful when you do not want the magnitude of the vectors to play a role.

E.g. single cell transciptomics

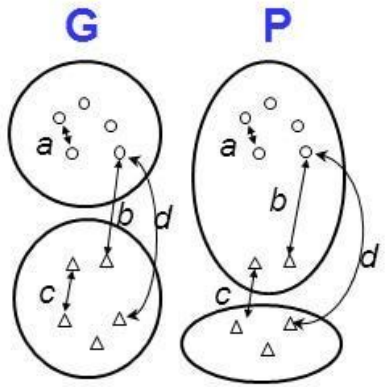


Rand index

The Rand index compares how similar two clustering results are.

It also can be used to benchmark the performance of an algorithm by comparing it to the true labels.

Rand index



Agreement: a, d

Disagreement: b, c

$$RI(P,G) = \frac{a+d}{a+b+c+d}$$

The Adjusted Rand Index takes into account how large these can be by random chance

$$ARI = \frac{RI - E(RI)}{1 - E(RI)}$$

Visualization

Motivation

We want to visualize the structure of multidimensional data though a plot.

We already have one way of doing so: PCA

PCA is very good for capturing linear relationships, but we will now look at 2 **non-linear** visualization methods.

t-SNE

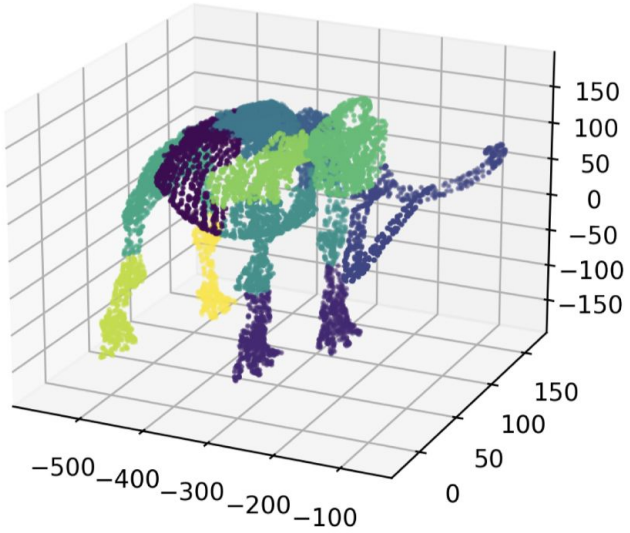
t-Distributed Stochastic Neighbor Embedding.

As a dimensionality reduction technique, it tries to preserve the relative distance between data points.

That is, close point remain close, and **distant points remain distant**.

PCA vs t-SNE

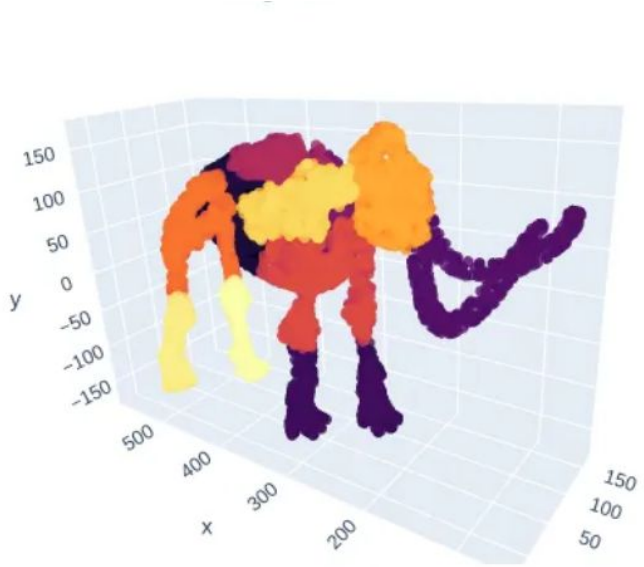
Original data



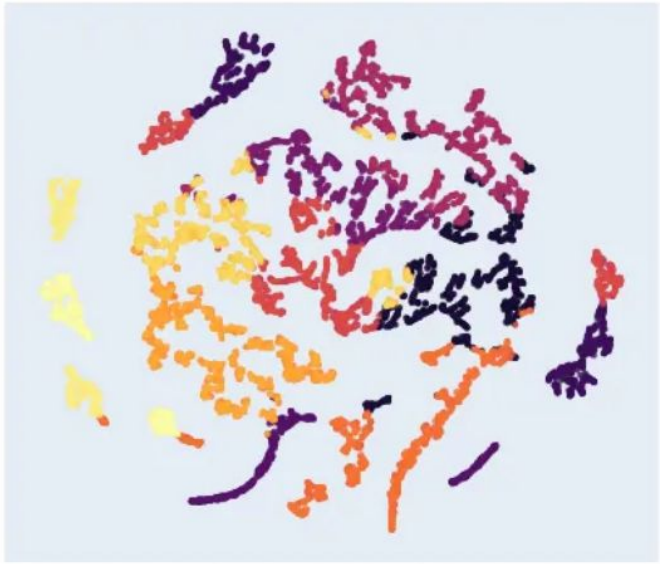
PCA

PCA vs t-SNE

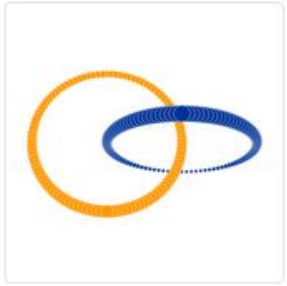
Original data



t-SNE



Perplexity controls the balance between local and global structure



Original

Perplexity: 2
Step: 5,000

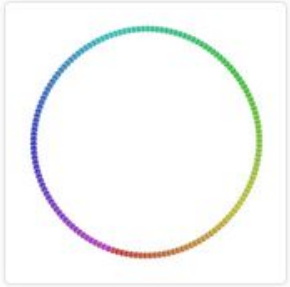
Perplexity: 5
Step: 5,000

Perplexity: 30
Step: 5,000

Perplexity: 50
Step: 5,000

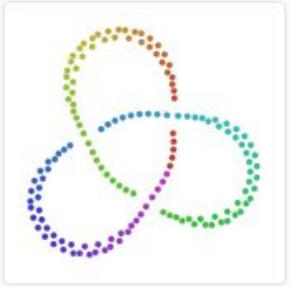
Perplexity: 100
Step: 5,000

Original

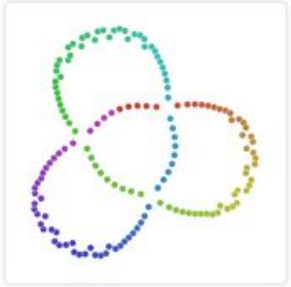


Perplexity: 2
Step: 5,000

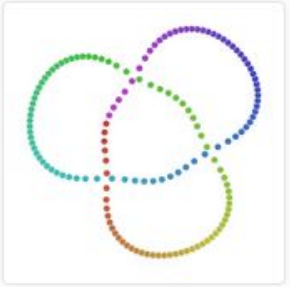
Perplexity: 5
Step: 5,000



Perplexity: 30
Step: 5,000

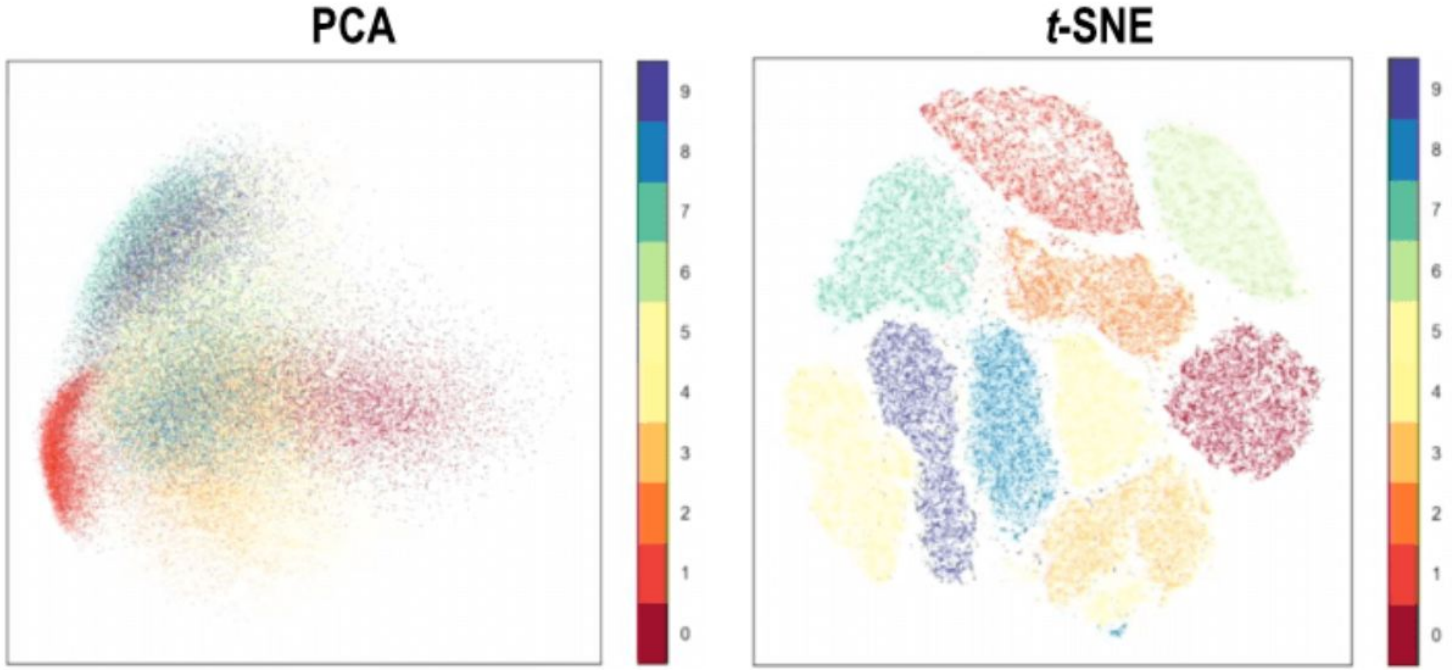


Perplexity: 50
Step: 5,000



Perplexity: 100
Step: 5,000

t-SNE is good for observing clusters

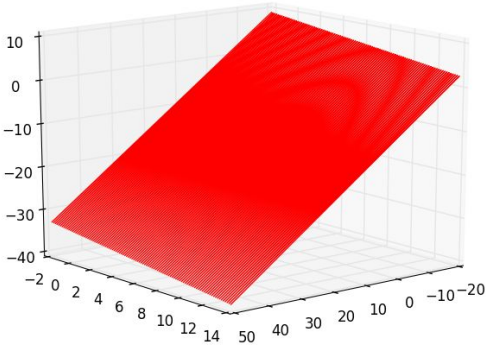


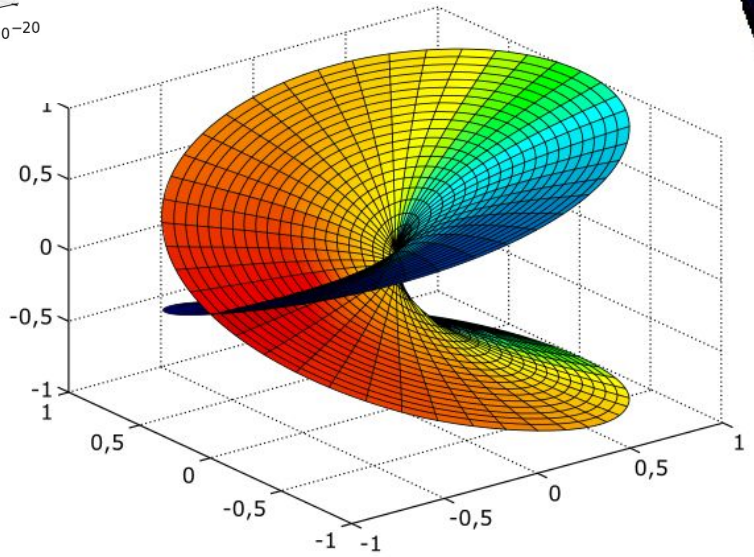
UMAP

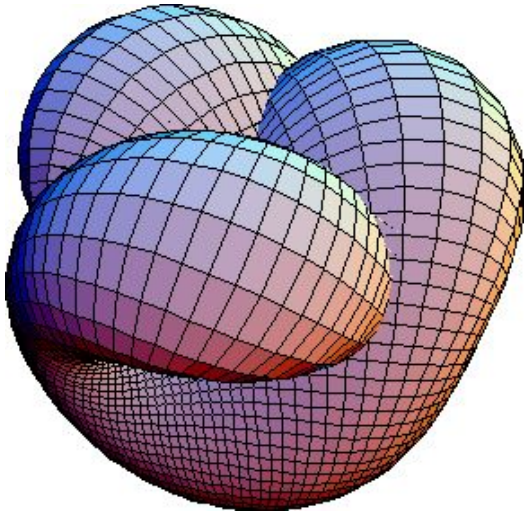
Uniform Manifold Approximation and Projection

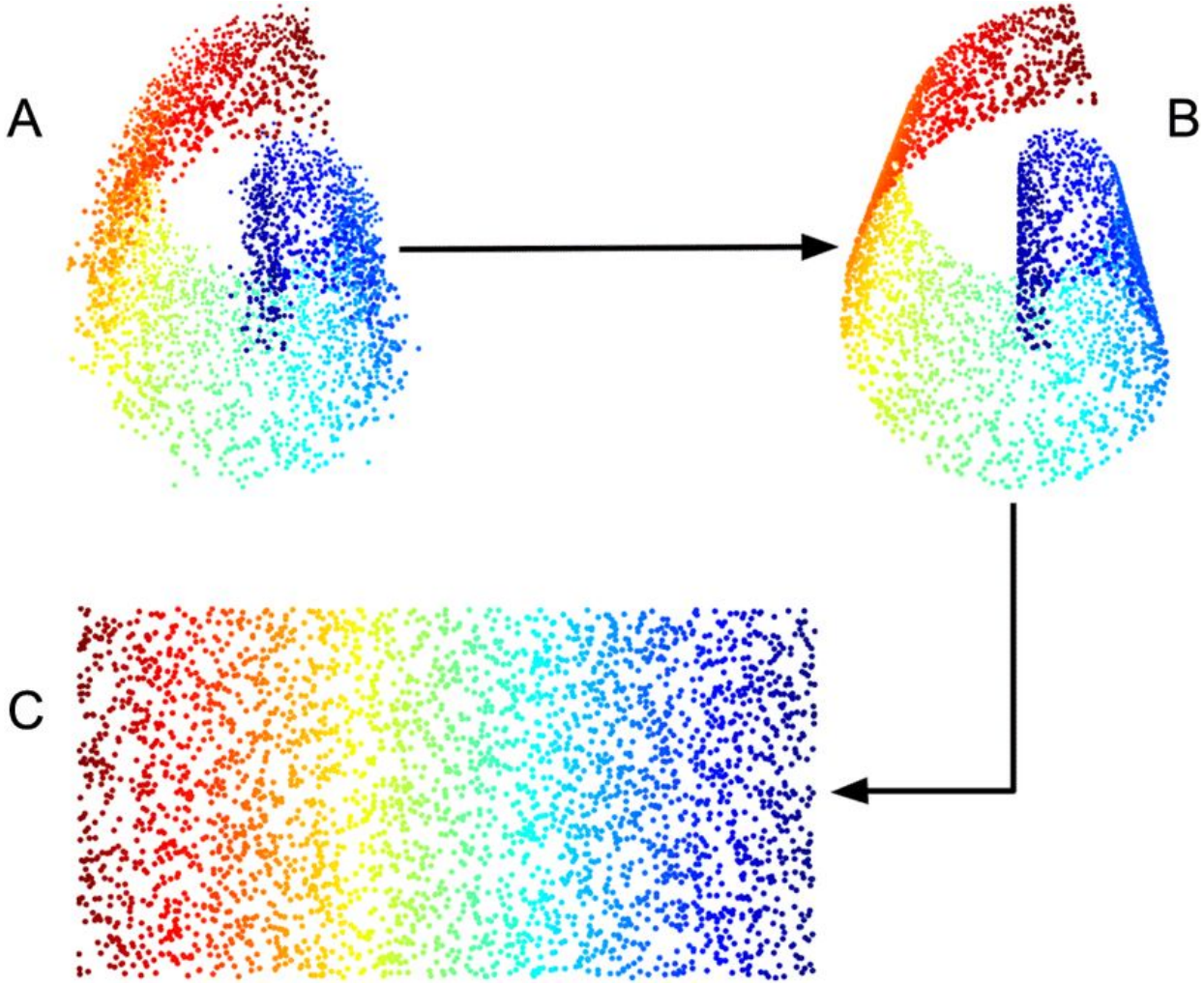
UMAP tries to find a low dimensional (2D) manifold in which the data is uniformly distributed.

What is a manifold?

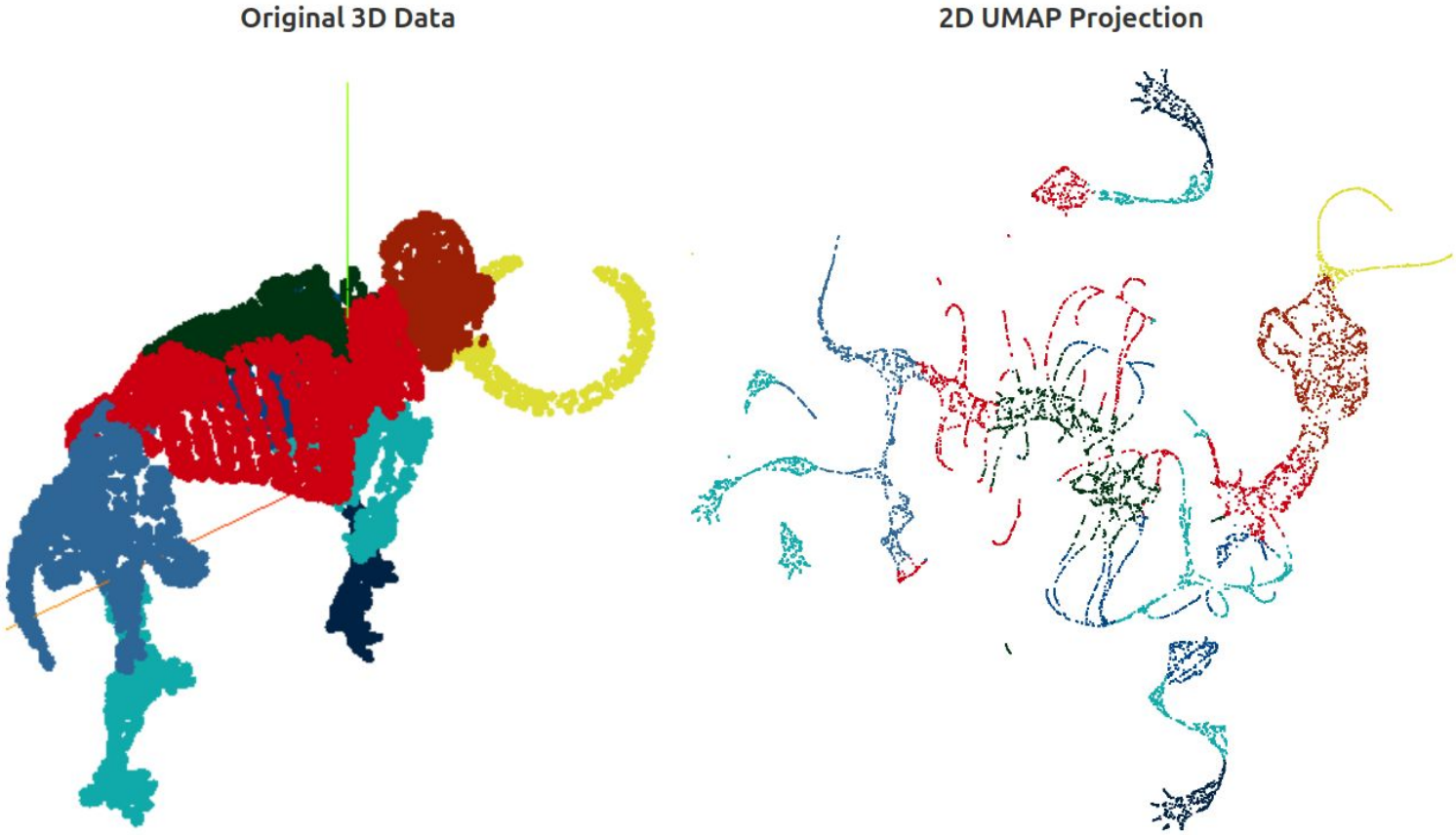




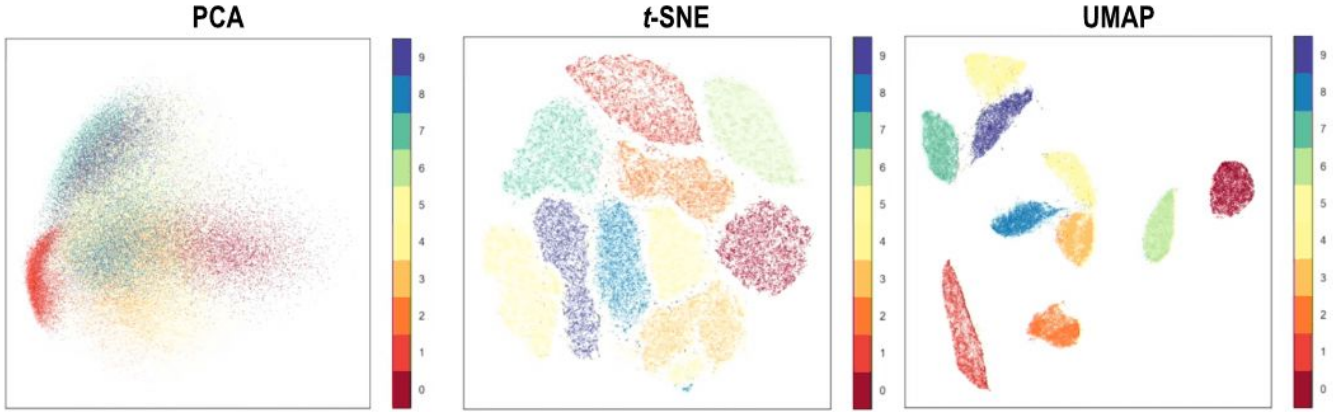



A
B
C

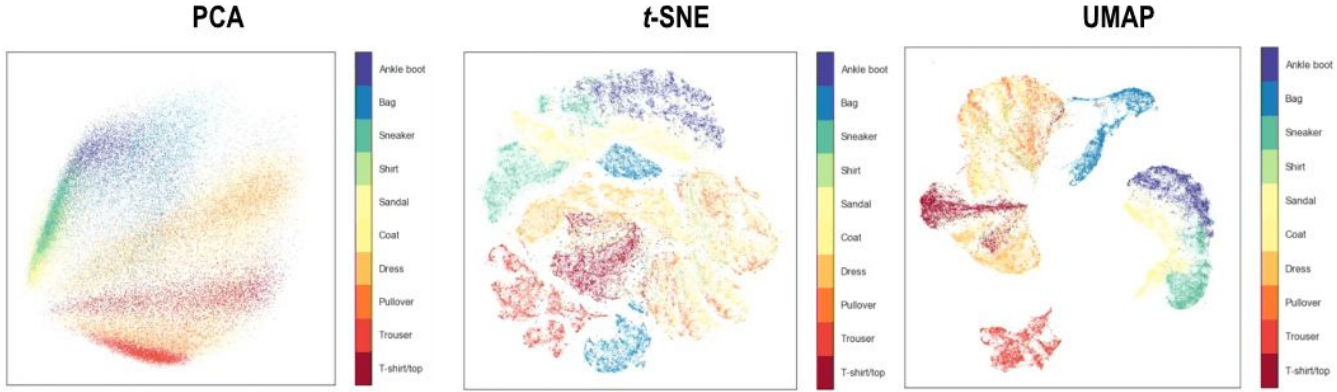
UMAP is good for observing higher order structures



MNIST Digits



Fashion MNIST



Caution

t-SNE and UMAP should be used only for visualization purposes.

Because of the non-linear transformation, it is not possible to use the results of t-SNE and UMAP for further statistical analysis.