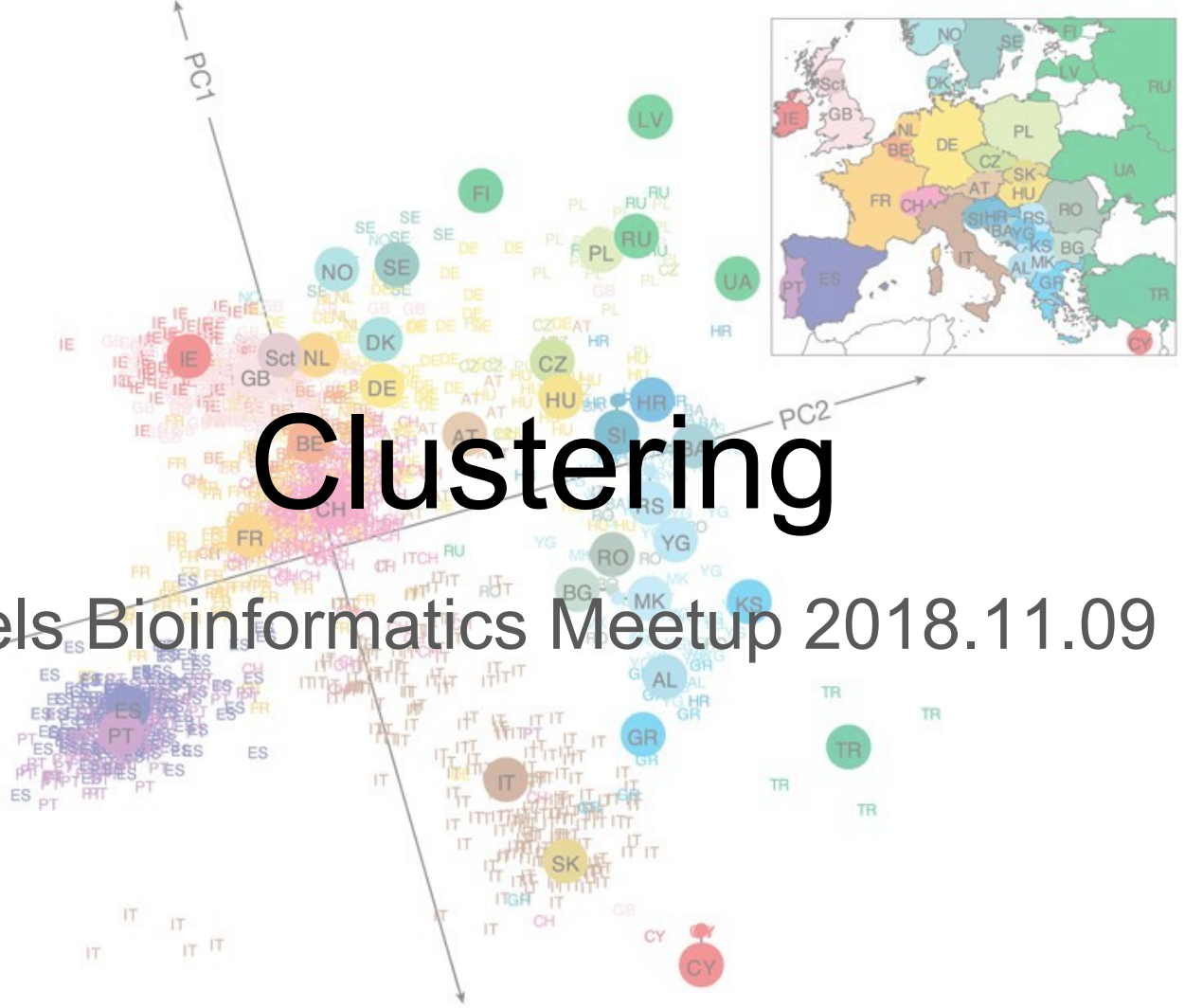
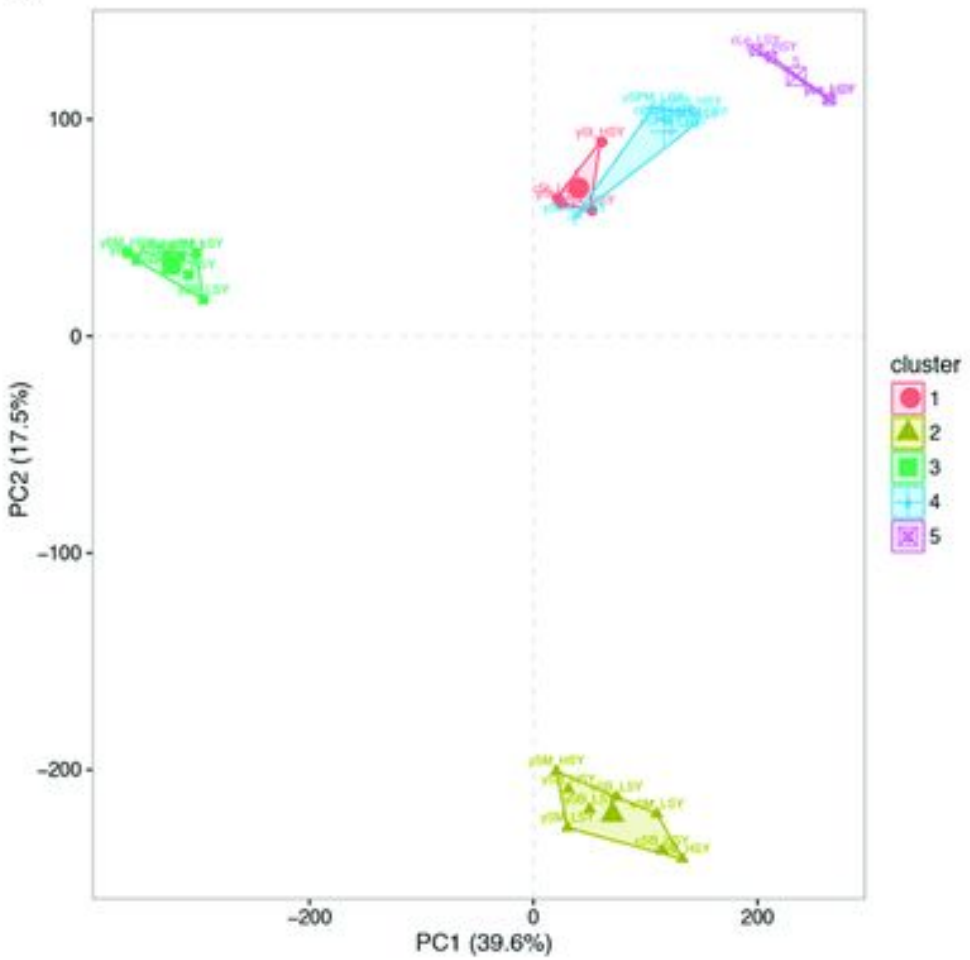


Clustering

Fels Bioinformatics Meetup 2018.11.09

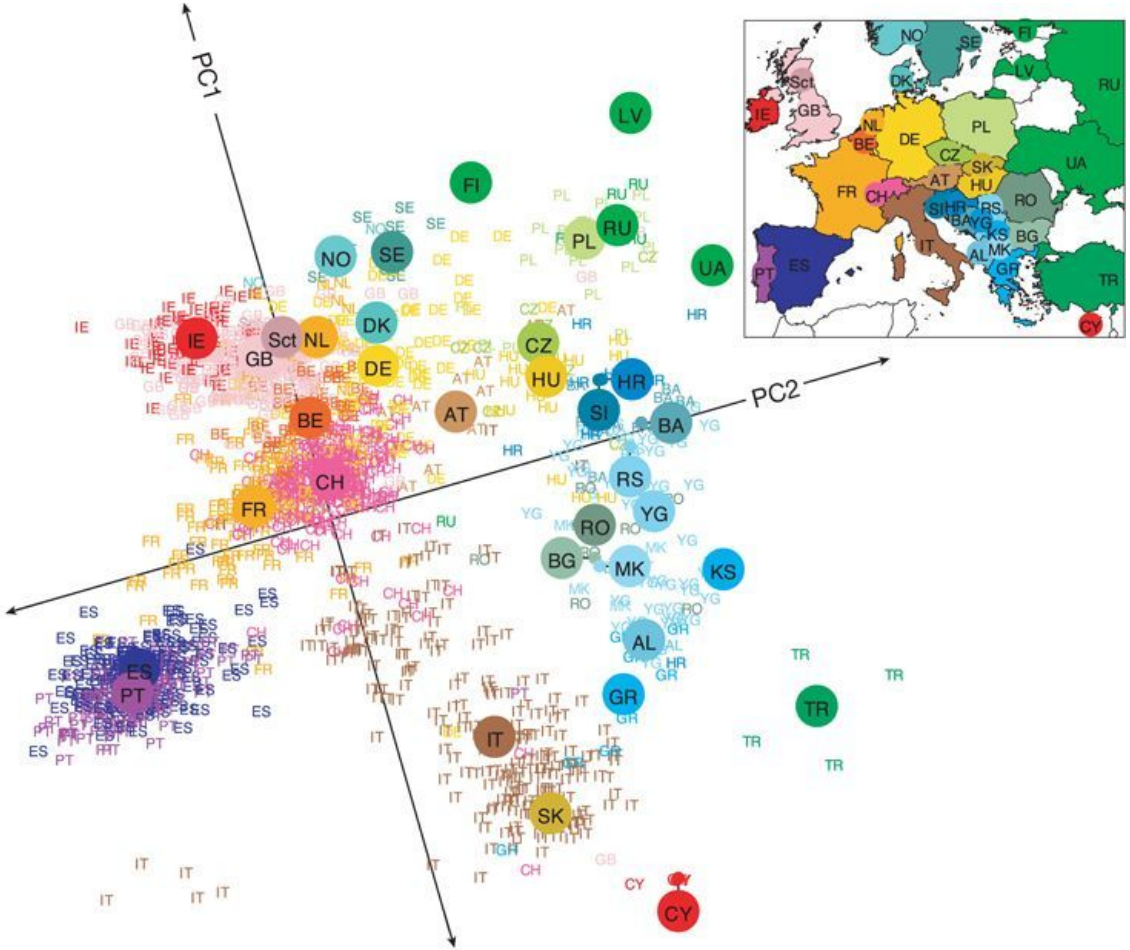
Why do we care about clustering?

- **see if samples cluster together**



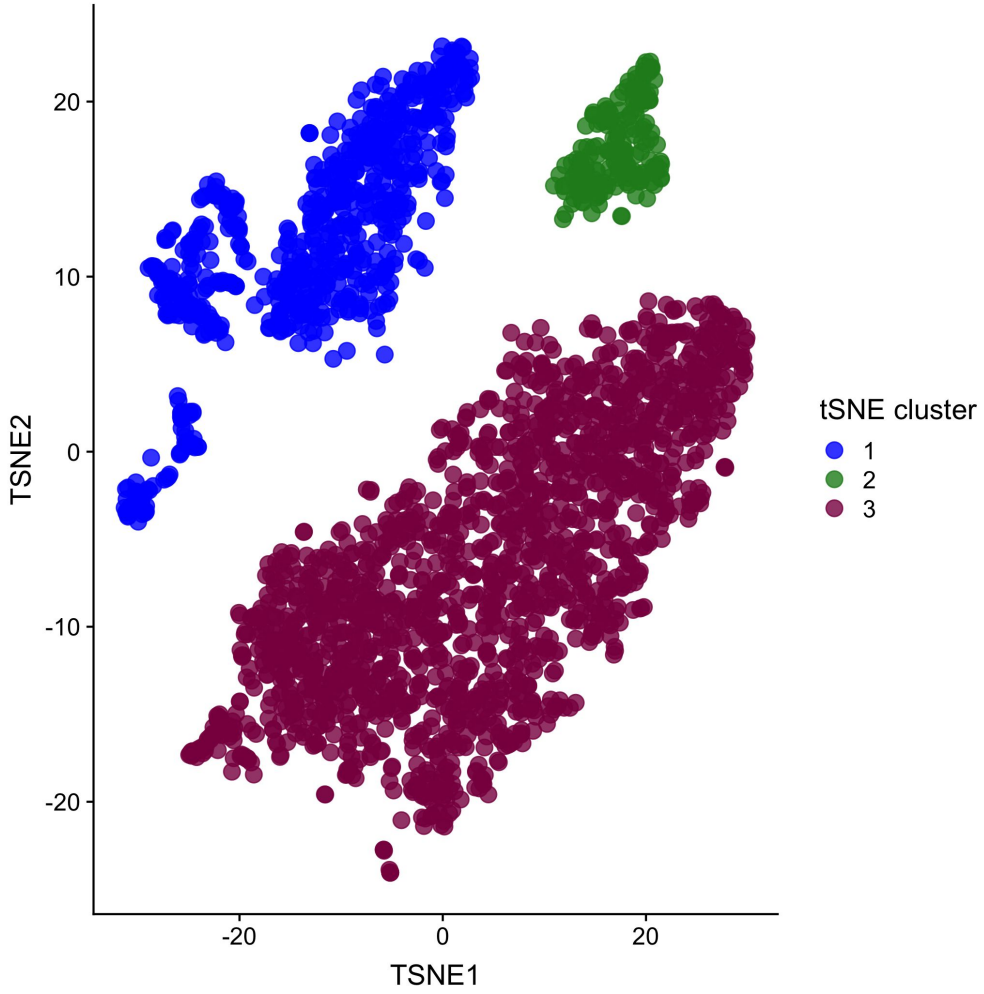
Why do we care about clustering?

- see if samples cluster together
- **see if individuals cluster together by some trait**



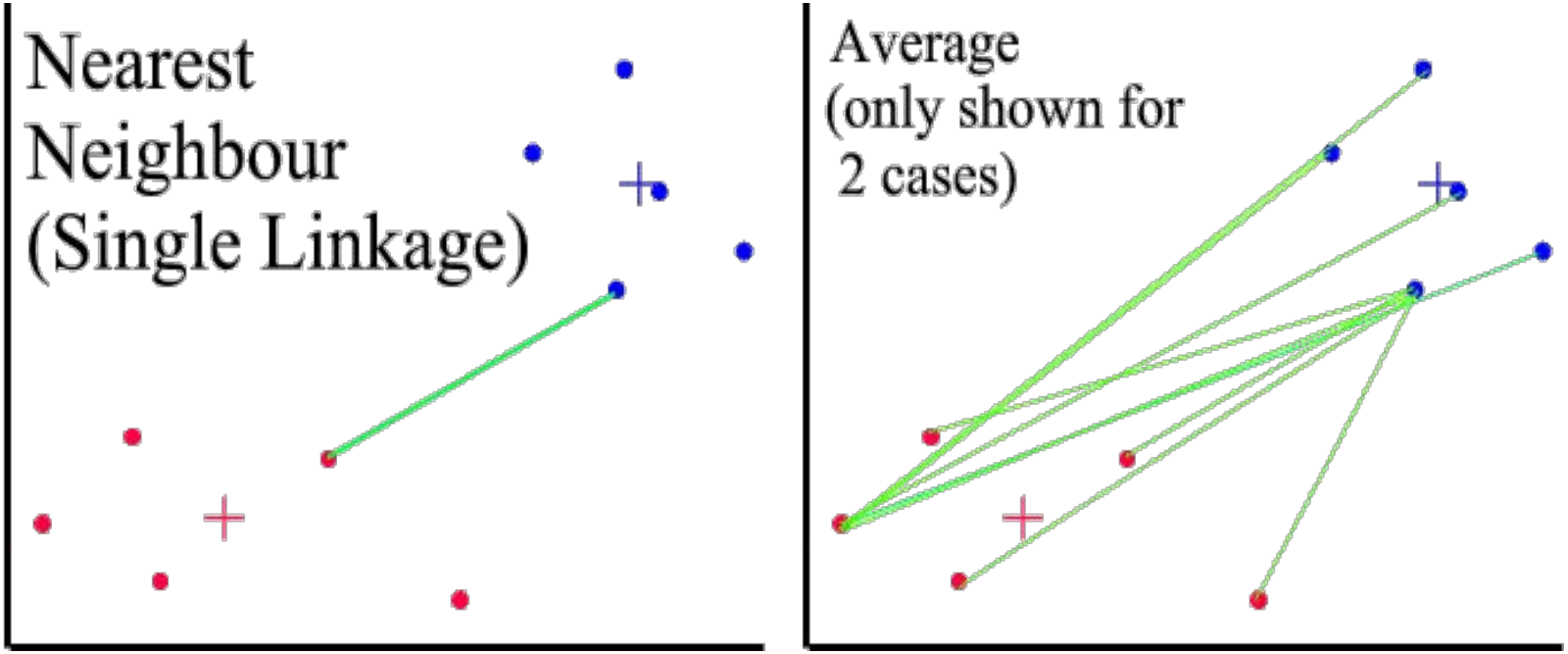
Why do we care about clustering?

- see if samples cluster together
- see if individuals cluster together by some trait
- **single cell**



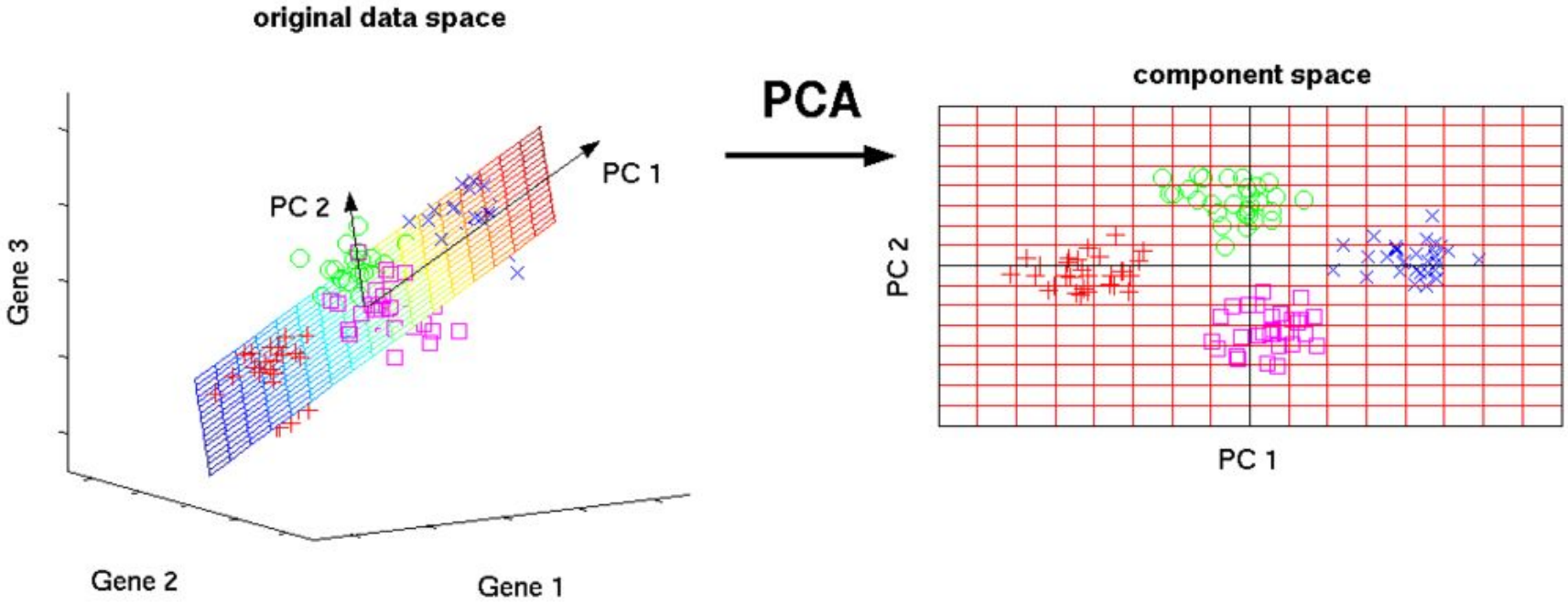
How does clustering work (on a high level)?

association



How does clustering work (on a high level)?

dimensionality reduction

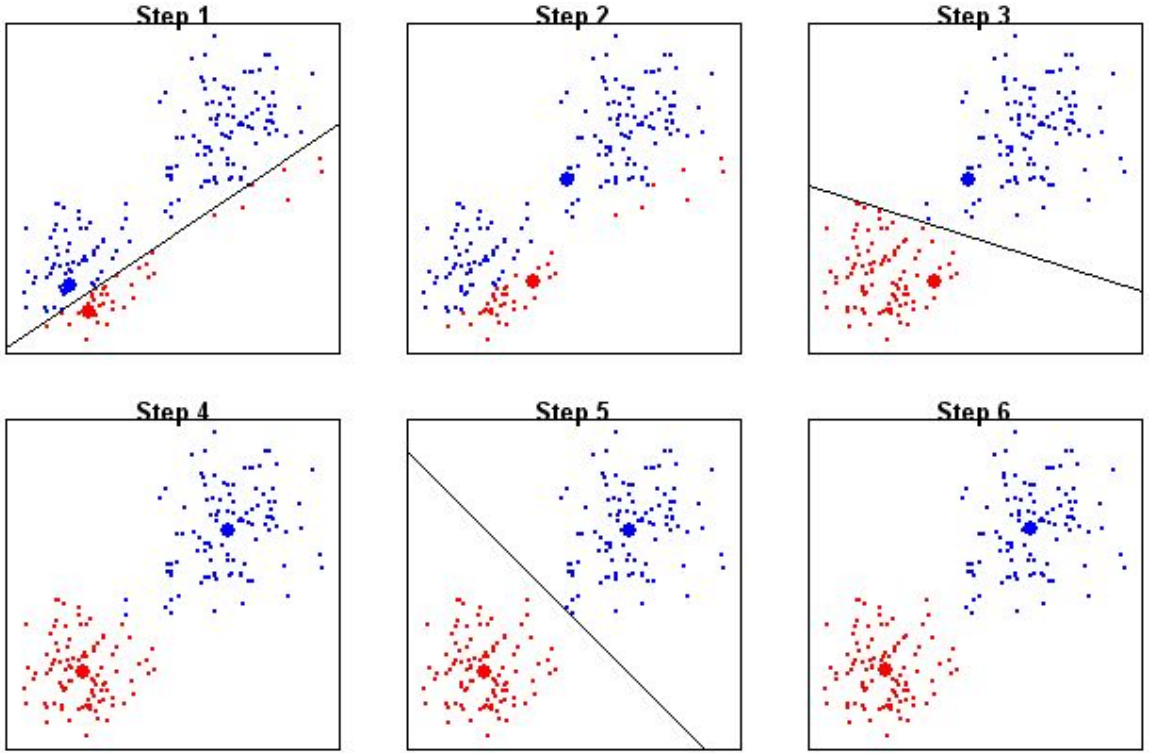


Clustering Methods

kmeans

How it Works

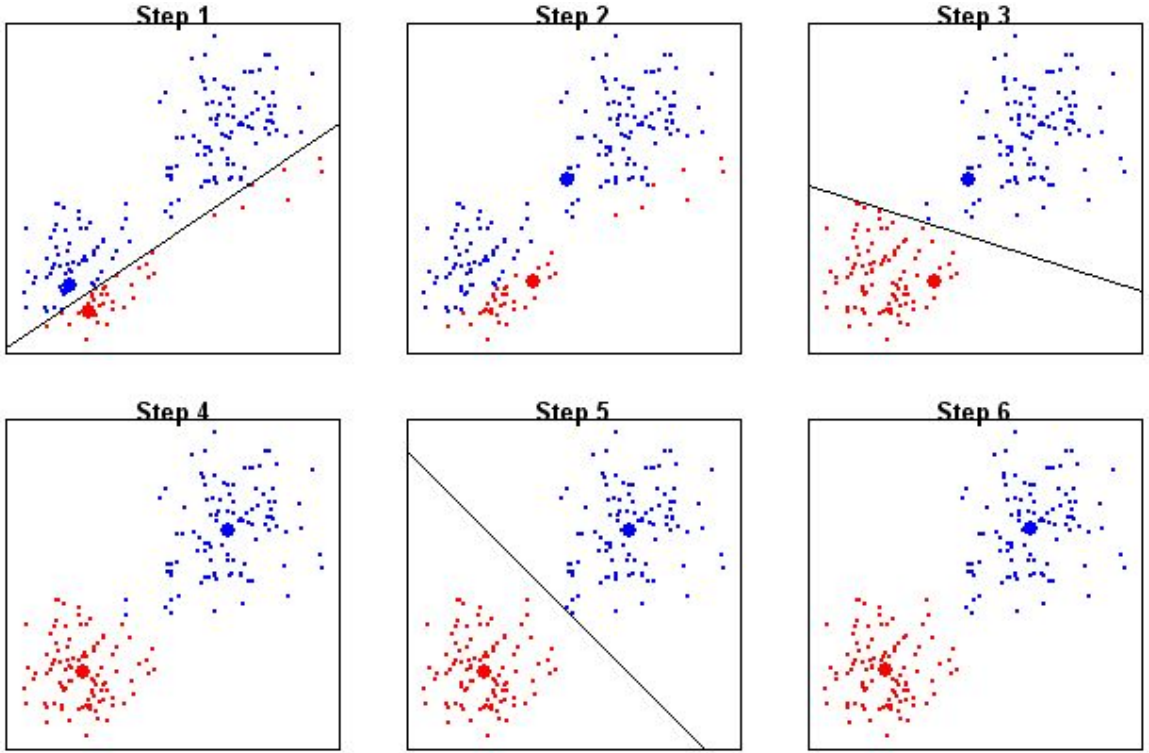
1. Pick number of clusters
2. Randomly assign center of cluster
3. Calculate the average of all points in the cluster
4. Move centroid to the average
5. Repeat steps 3 and 4 until nothing changes



kmeans

Pros/Cons

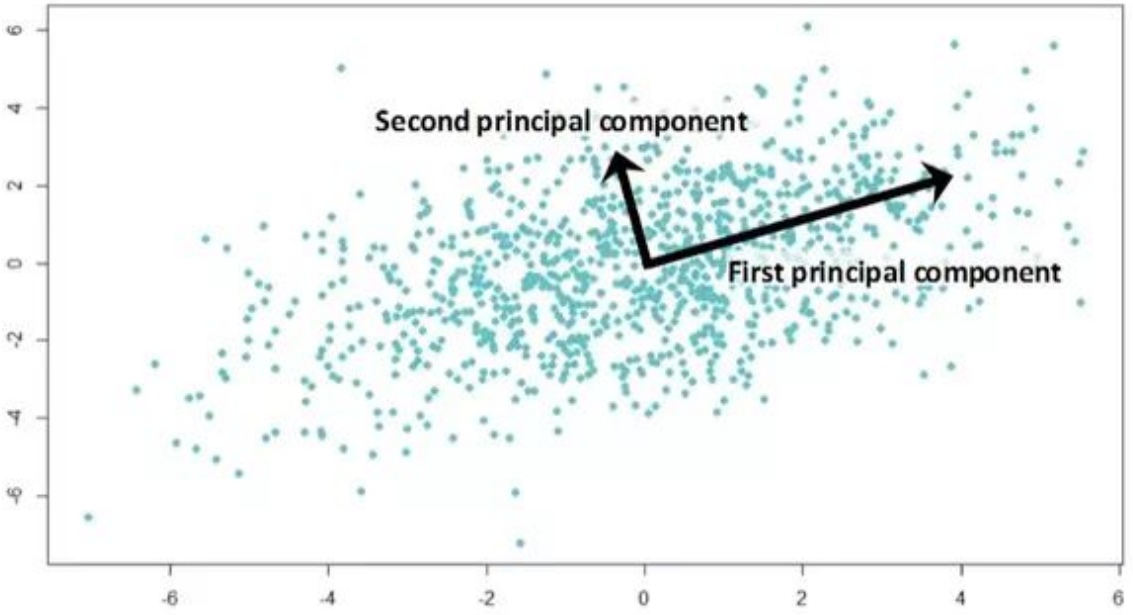
- very efficient
- works best when the data is in distinct, spherical clusters
- have to pick number of clusters, which can be tricky
- generally not the best for biological applications



Principal Components Analysis (PCA)

How it Works

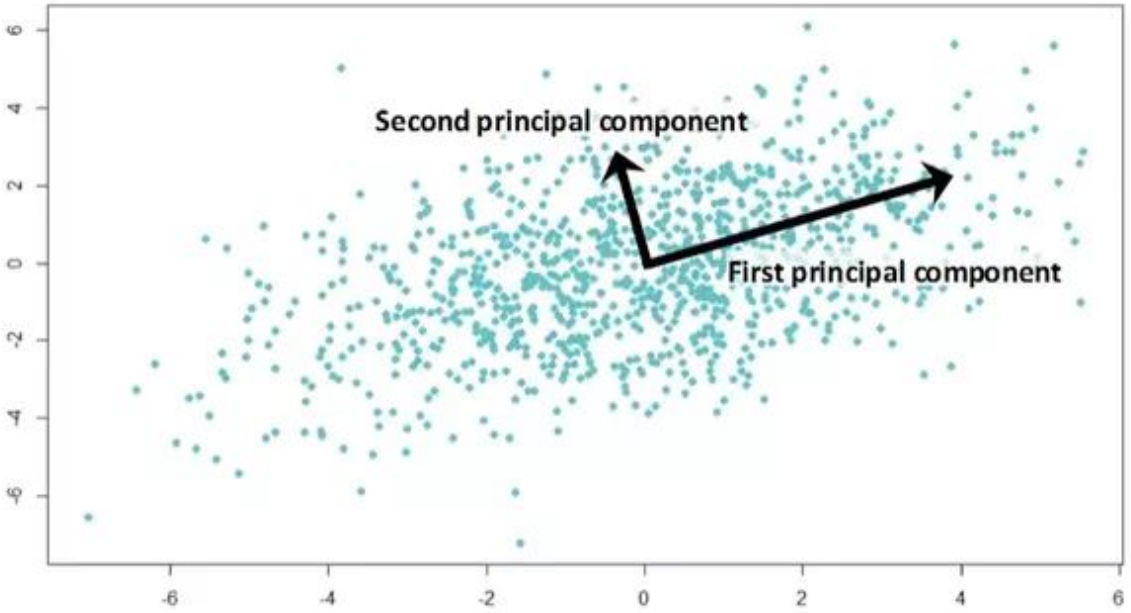
1. Project into higher dimensional space
2. Find axis with most variation and assign one-dimensional coordinates from it to PC1
3. Repeat step 2 for all variables



Principal Components Analysis (PCA)

Pros/Cons

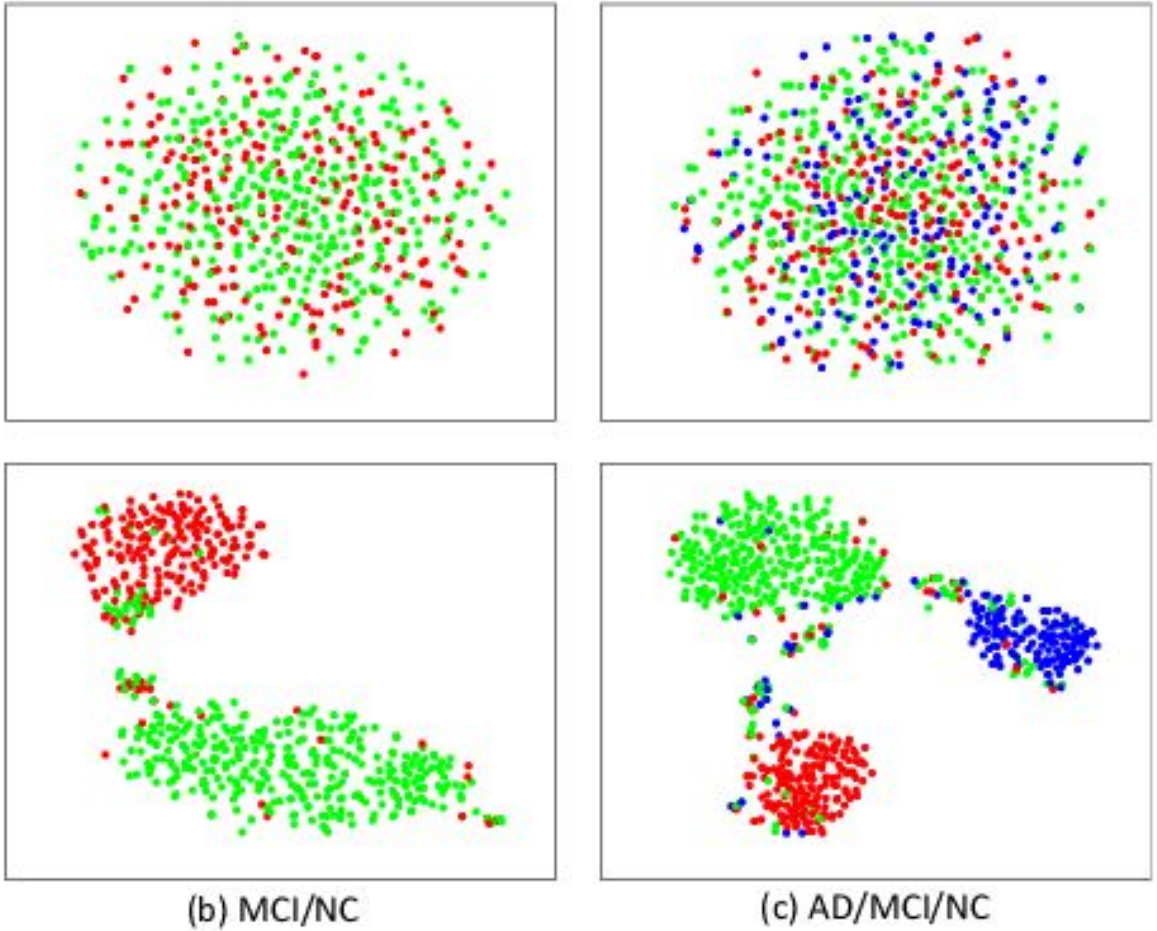
- good for linear data
- relatively efficient
- results are in order of relevance (i.e. PC1 is the most important)
- Visualization is easy to interpret
- Numeric results are hard to interpret
- can’t handle missing data



t-Distributed Stochastic Neighbor Embedding (tSNE)

How it Works

1. Calculate similarity between all points in higher dimensional space
2. Randomly project into lower (probably 2) dimensions
3. Recalculate similarities
4. Minimize difference between high dimension and low dimension similarities



t-Distributed Stochastic Neighbor Embedding (tSNE)

Pros/Cons

- works best with non-linear data (which is a lot of biological data)
- good for visualization
- very inefficient
- hard to interpret; size and distance between clusters have essential no meaning
- hard to interpret numbers returned as well
- not visually reproducible