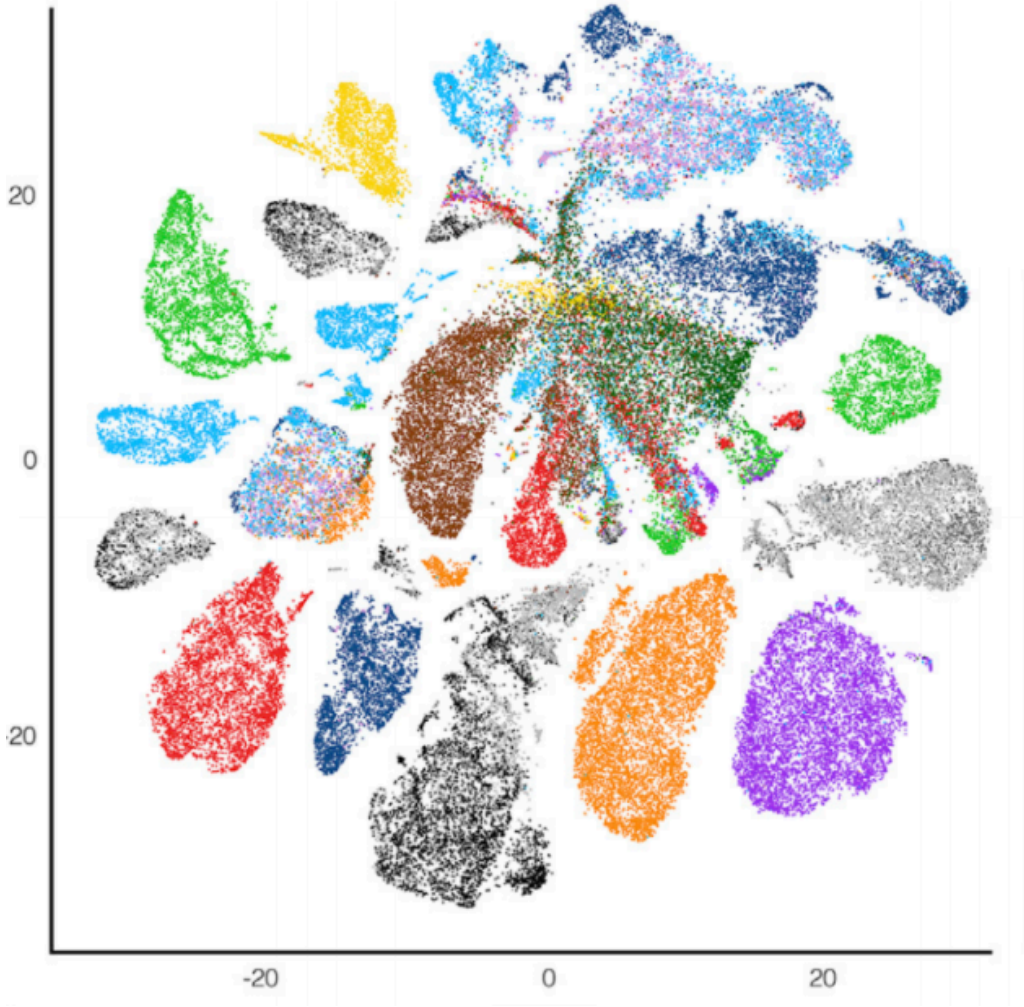
Clustering Methods in R

2020-07-15

What is clustering? And why do it?

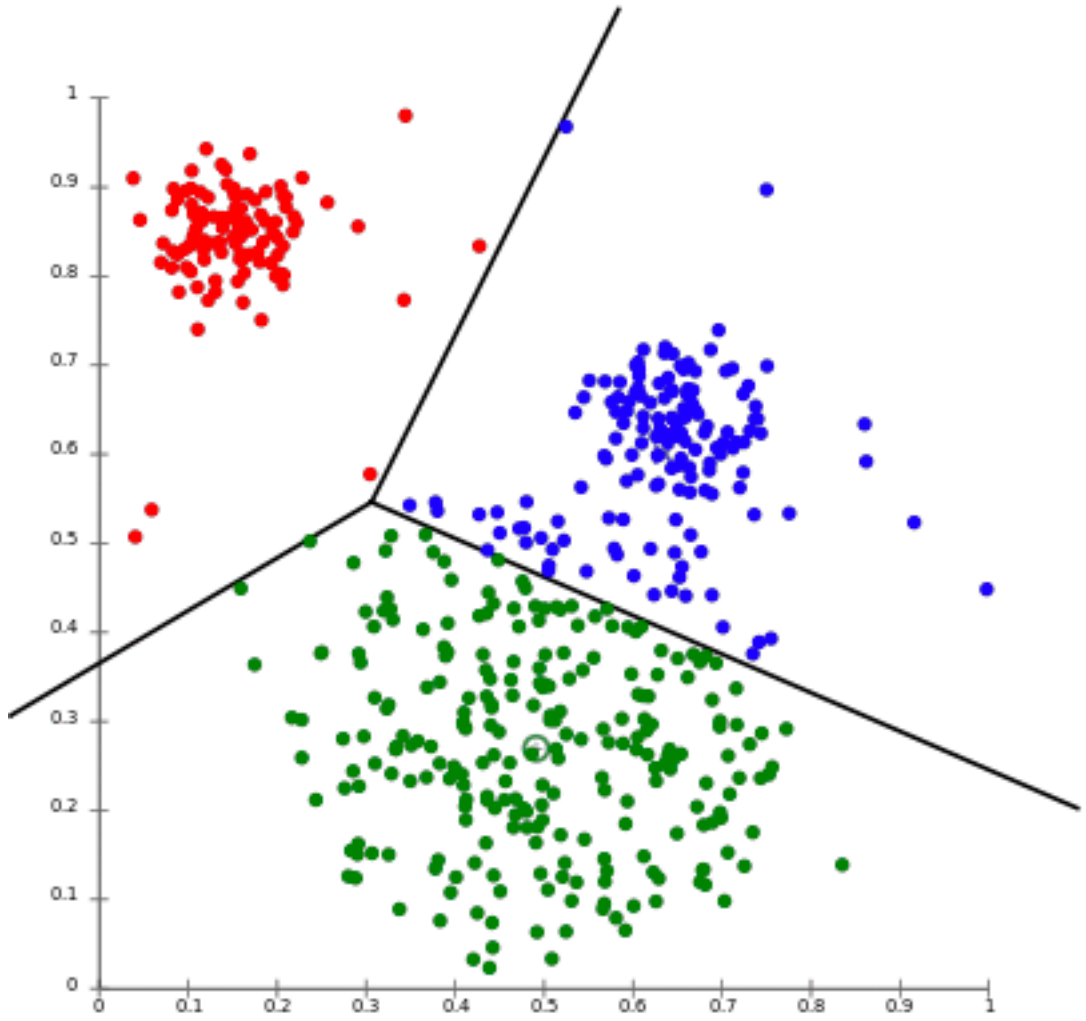
- clustering – grouping objects together into clusters by similarity
- Many, many different algorithms that try to solve this problem
- Clustering can identify patterns of variation in the data
  - Unwanted clustering like batch effect
  - Wanted clustering like by condition or to identify cell type in single cell sequencing



K-means Clustering

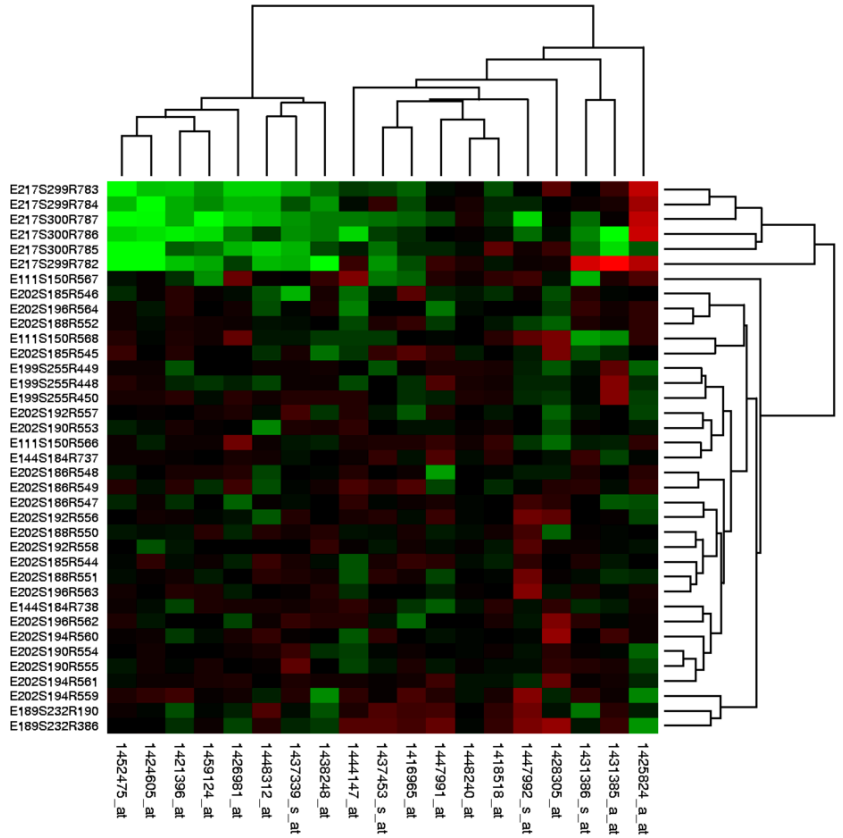
1. Pick number of clusters
2. Randomly pick a point to be the center of each cluster (centroid)
3. Assign all the datapoints to a cluster based on the nearest centroid.
4. Move the centroids
5. Repeat steps 3 and 4 until no points change clusters

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/



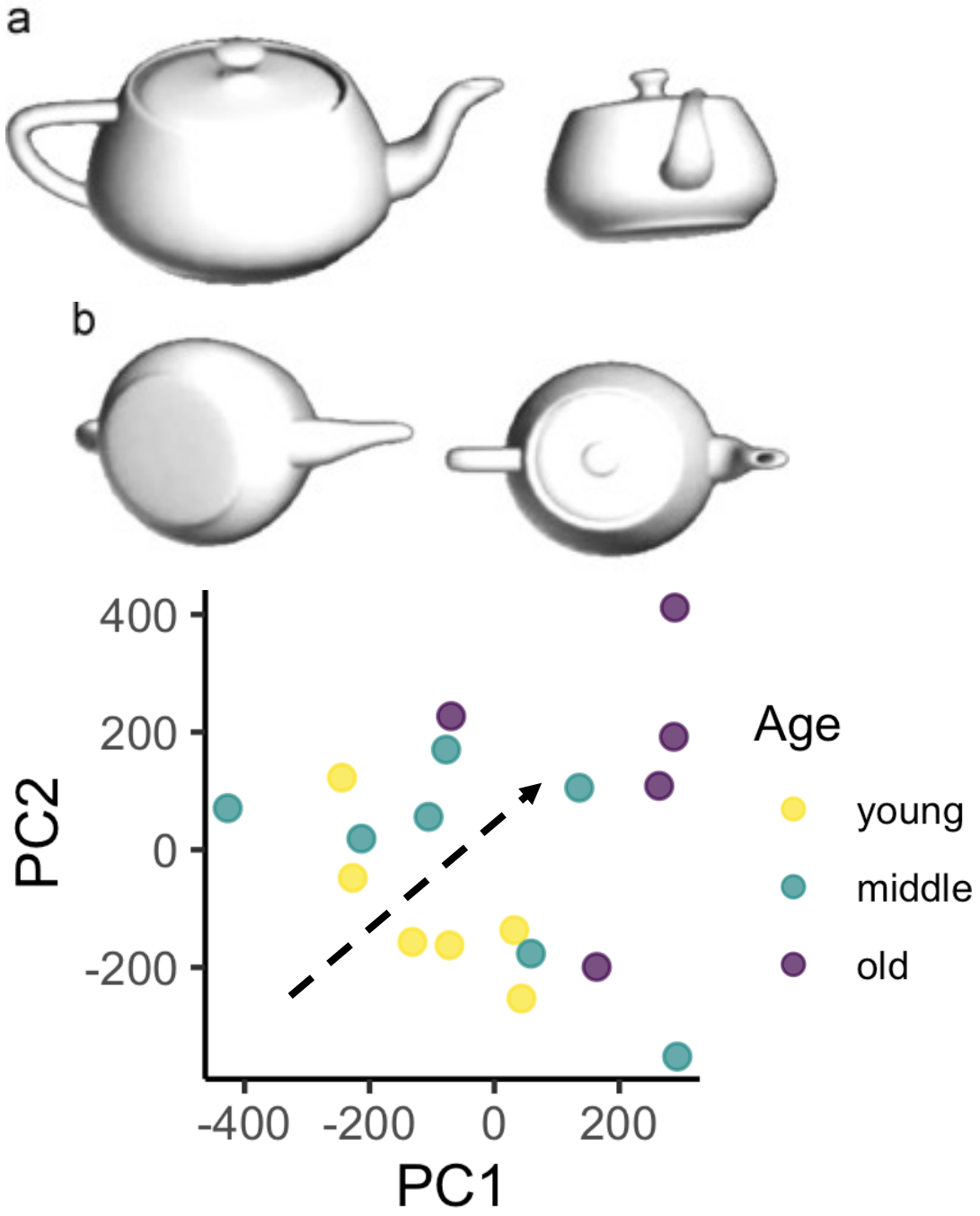
Heatmap

- A **heatmap** shows magnitude of some data as color in two dimensions (one color for lowest values transitioning to the other color for the highest values); cells are clustered by some algorithm and the dendrogram on the top and/or sides shows the relationships
- To calculate the clustering, first a measure of similarity is calculated, then a clustering algorithm is applied to the similarity scores.
- Many different algorithms can be used to calculate the clustering



PCA

- Like calculating a line of best fit, but with more than 2 dimensions
- To calculate a PCA
  - Data can be normalized by scaling and centering it (z scoring). Whether you do this affects the final outcome
  - Do some linear algebra to calculate principal components
- PC1 always explains the most variation, then PC2, then PC3 etc.
- Dimensionality reduction technique; goal is to retain most of the important information while simplifying the data



DEMO