Module 5B: Unsupervised Learning

A smattering of options: PCA, permutations, bootstrap

K-means clustering algorithm

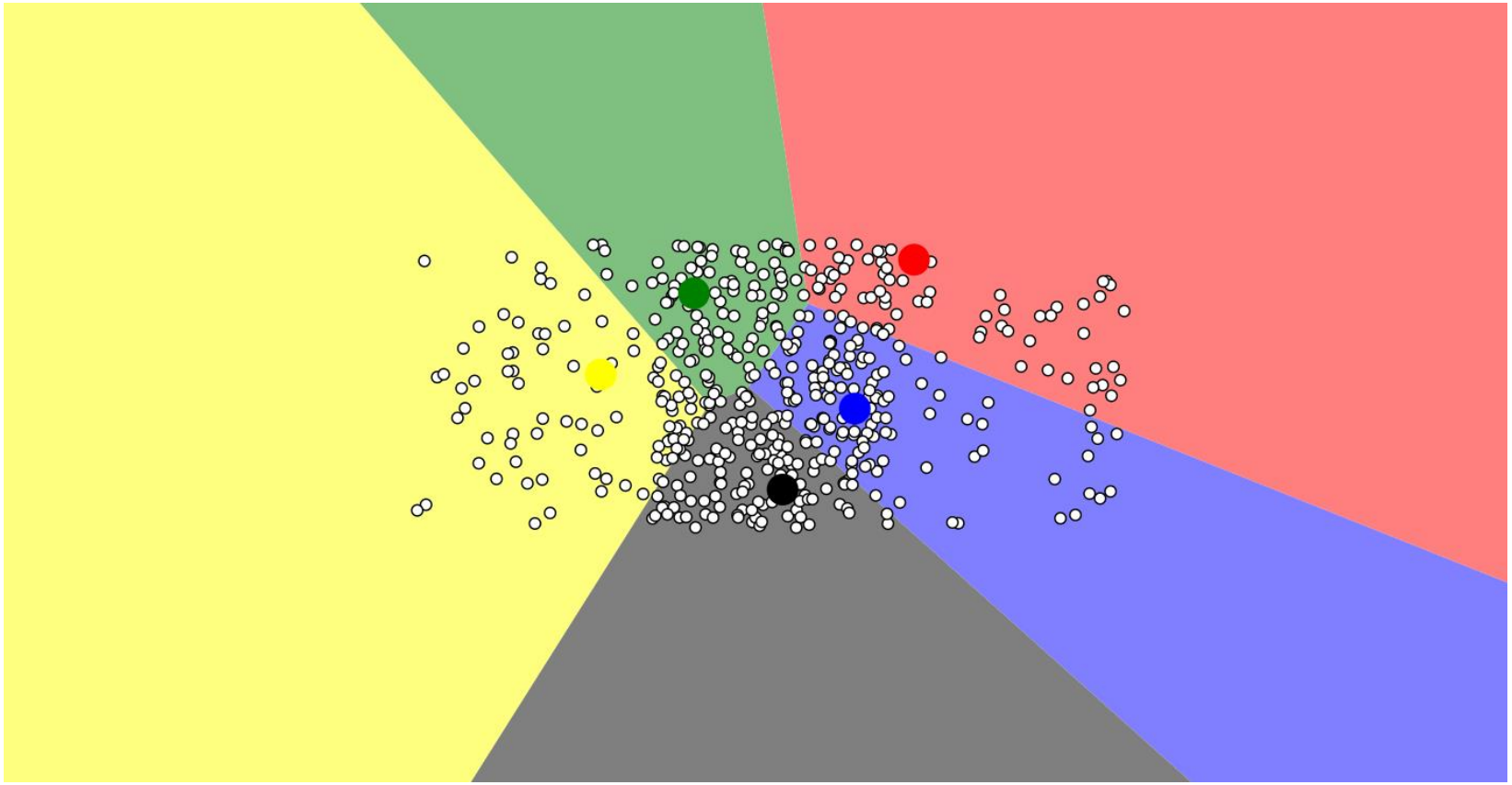
Clusters group data points together that share similarities

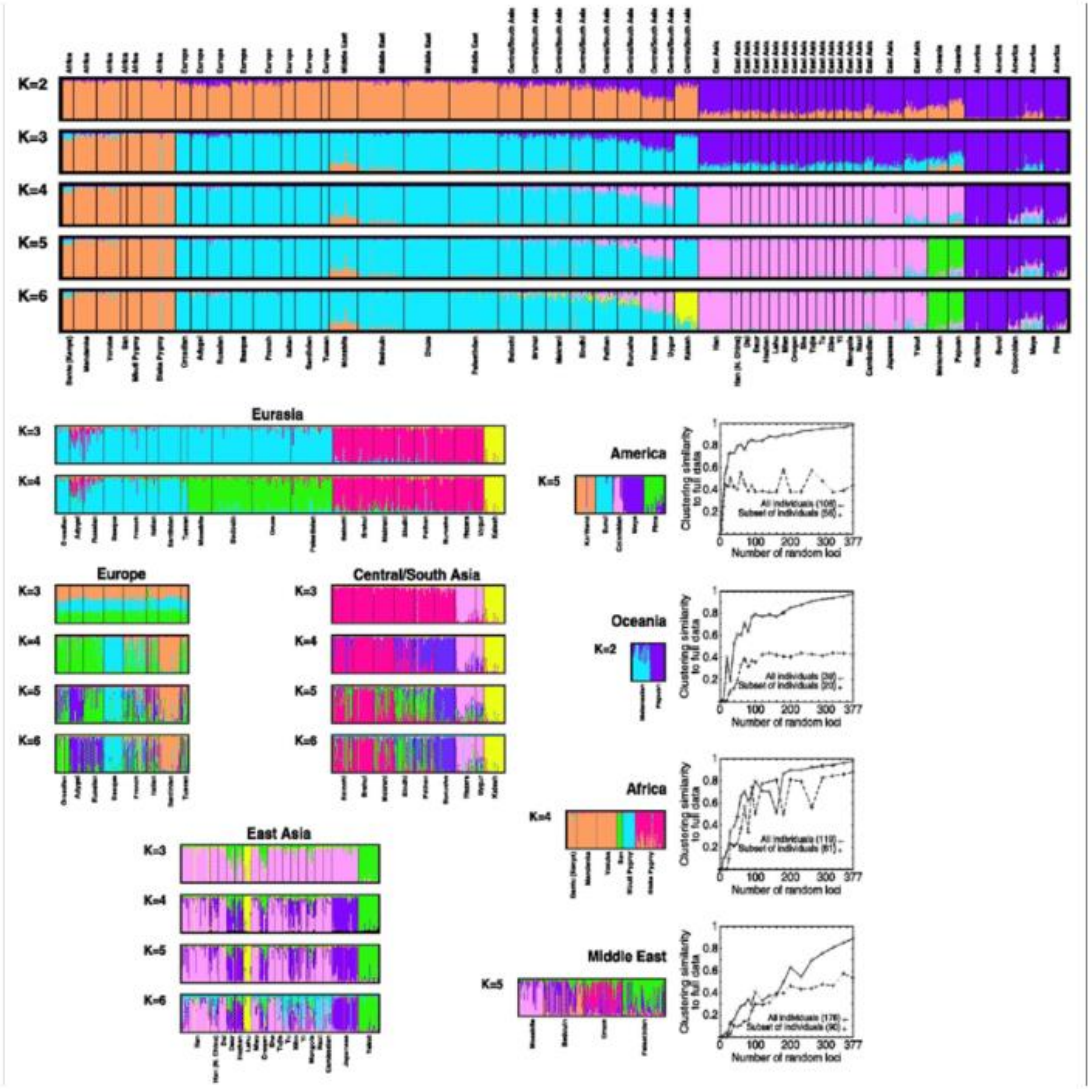
- N data points; k clusters → for each data point, assign it to one of the k groups
- Criteria: distance of data point from the center value (mean) of the kth group
- <u>Method:</u>

  1. **Random creation of k clusters** with centroids
  2. **Assignment** of each data point to a cluster based on shortest Euclidean distance to centroid
  3. **Updated centroids** (updating mean to include newly assigned point values), and the process repeats until a 'good' cluster is found (the value of the centroids stop changing between iterations)
- Sensitive to noise (data needs to be highly separated); highly dependent on initial assignment of centroids and k; can get stuck in local minima
- Pros: fast!

Code free way of visualizing k means!

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





From Rosenberg et al. (2002) estimated population structure for the 52 sampled populations of the HGDP-CEPH panel for pre-chosen values of K = 2 through K = 6. Each cluster (K) is represented by a different color. Each individual is a vertical line, which depicts an estimate of that individual's membership in each cluster (multiple colors indicate membership in more than one cluster). Thin black lines denote individual populations. Population labels are shown at the bottom of the figure, while broad regional labels are listed at the top of the figure. While broad geographic clustering occurs, note that many individuals share genetic similarities with more than one cluster. This is particularly true within continents and for individuals from populations at the borders of continents