Introduction to Data Science

BRIAN D'ALESSANDRO

ADJUNCT PROFESSOR, NYU

FALL 2016

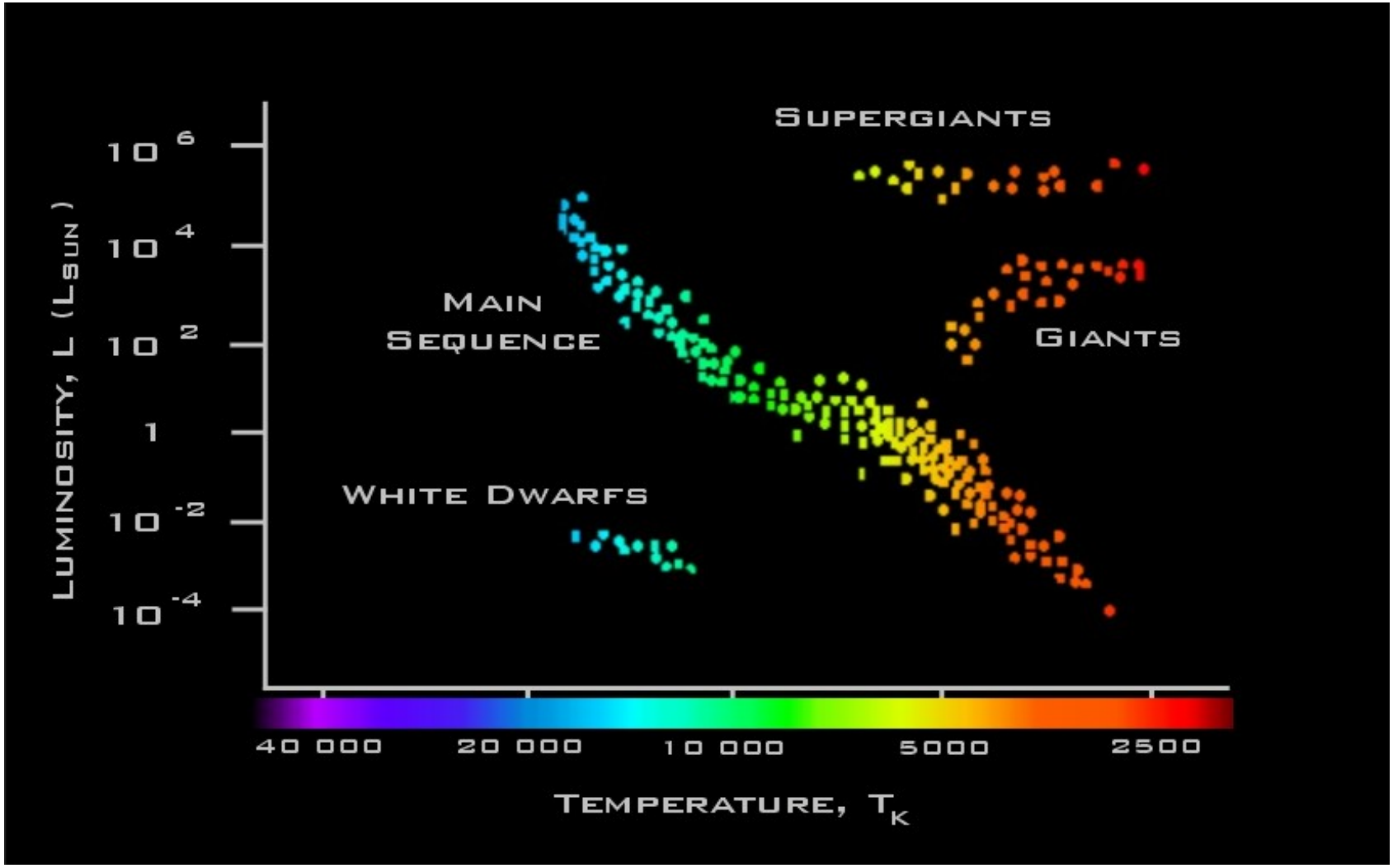
Fine Print: these slides are, and always will be a work in progress. The material presented herein is original, inspired, or borrowed from others' work. Where possible, attribution and acknowledgement will be made to content's original source. Do not distribute, except for as needed as a pedagogical tool in the subject of Data Science.

CLUSTERING

BASIC CLUSTERING GOAL

How do I find distinct groupings of similar objects?

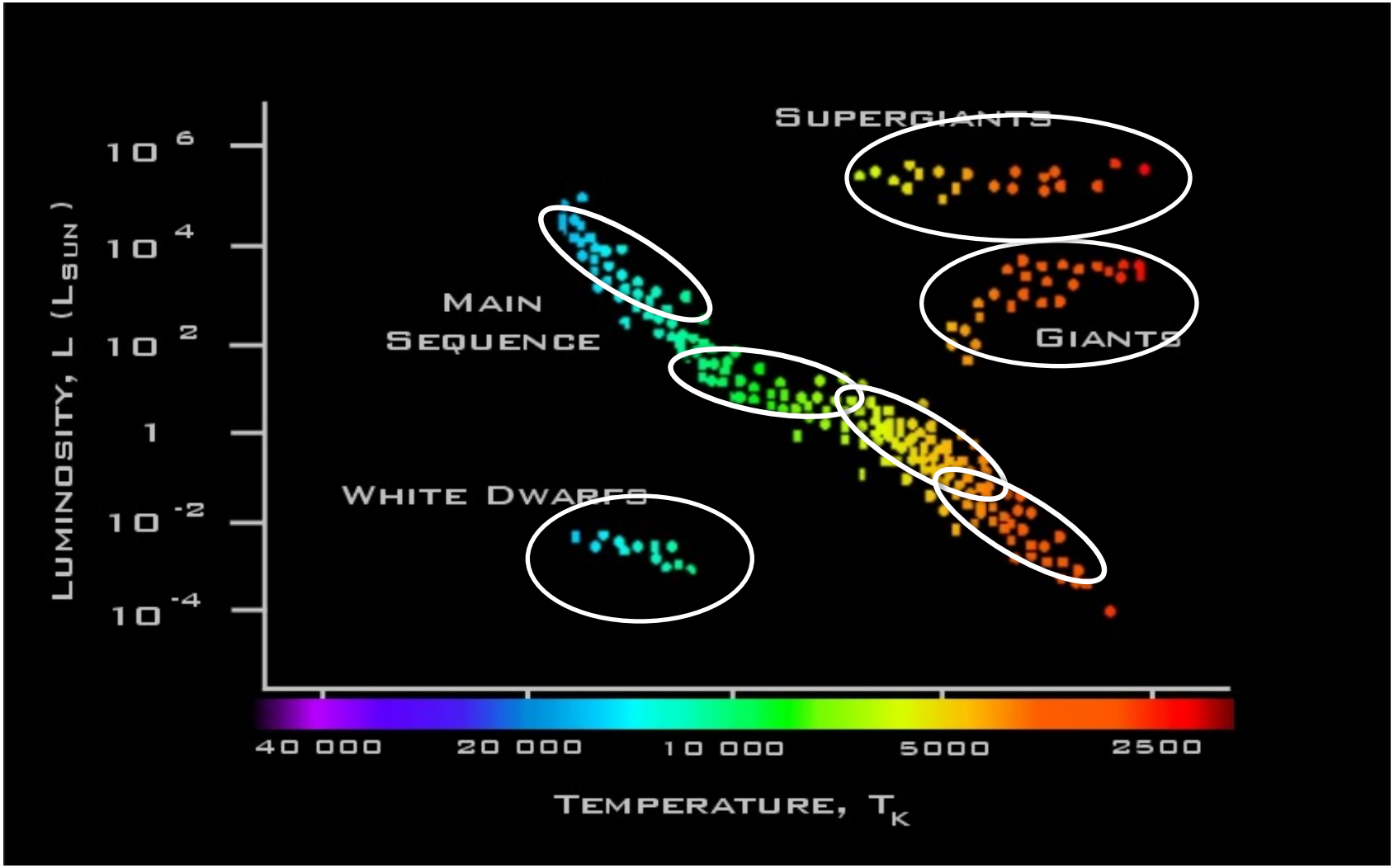
An object is an row in a data matrix X with a feature vector X_i



BASIC CLUSTERING GOAL

Compute similarity/distance between objects

Create distinct groups that minimize intra-group distance and maximize inter-group distances



2 TECHNIQUES

K-Means (partitioning)

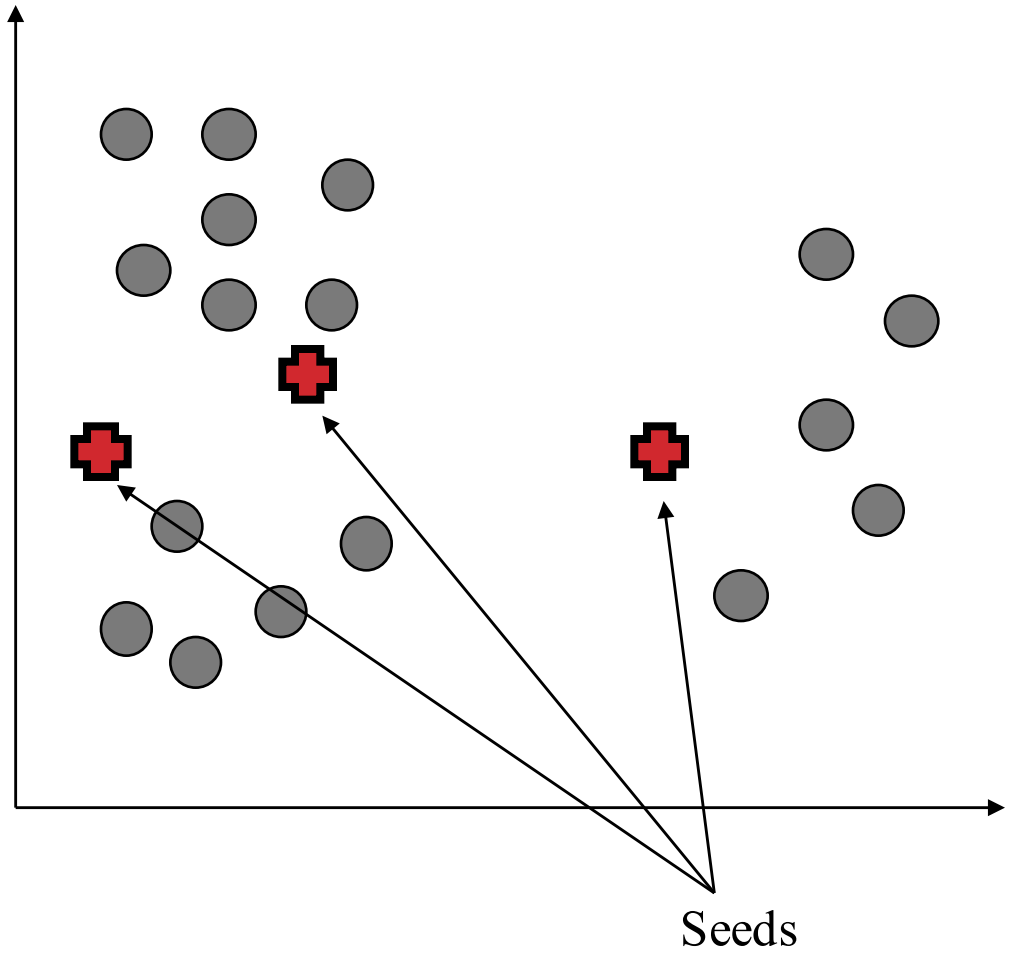
- Clusters are defined by a center point
- # groupings chosen in advance
- Each object belongs to cluster in which it has minimum distance to cluster center
- Generally cheaper, but not stable

Hierarchical

- Clusters are arranged in a nested taxonomy
- K can be chosen after the fact
- Stable but computationally expensive

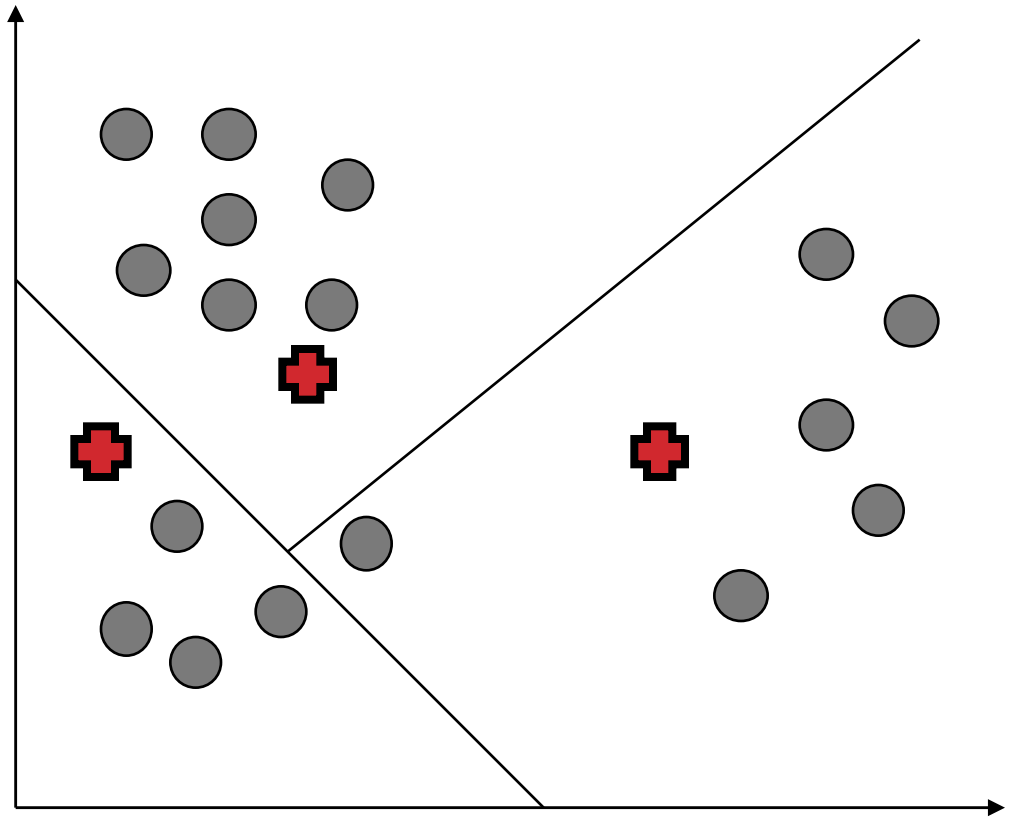
K-MEANS

1. Start with data and a predefined number of clusters, k.
2. Choose k seeds randomly.



ASSIGN INSTANCES TO CLUSTERS

3. Assign each instance to the cluster for which it is closest to the cluster center.



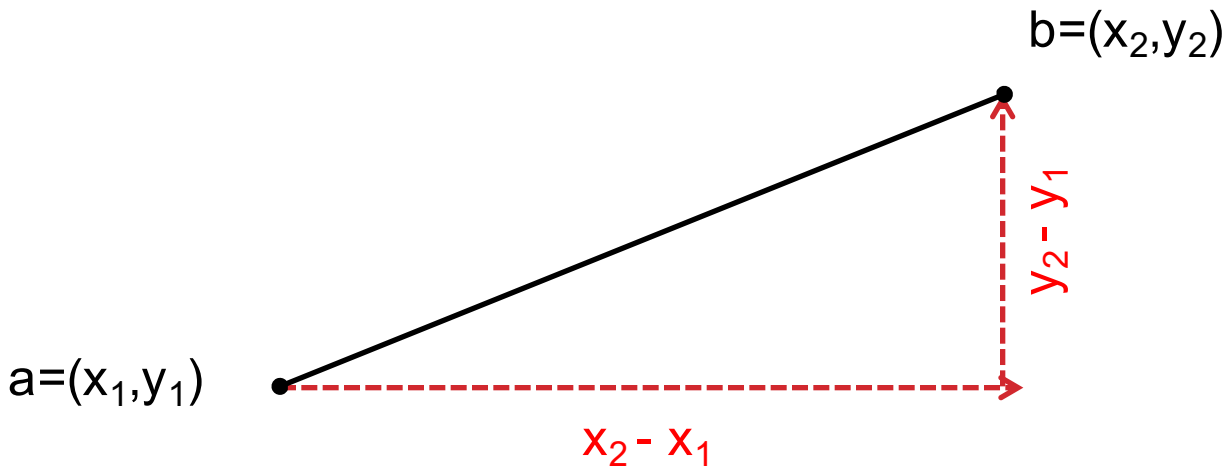
EUCLIDEAN DISTANCE

This metric derives from basic geometry, and is the way distance is often defined in physical coordinate systems.

Let ***a*** and ***b*** be two k-dimensional vectors in Euclidean space. The Euclidean distance between them is defined as:

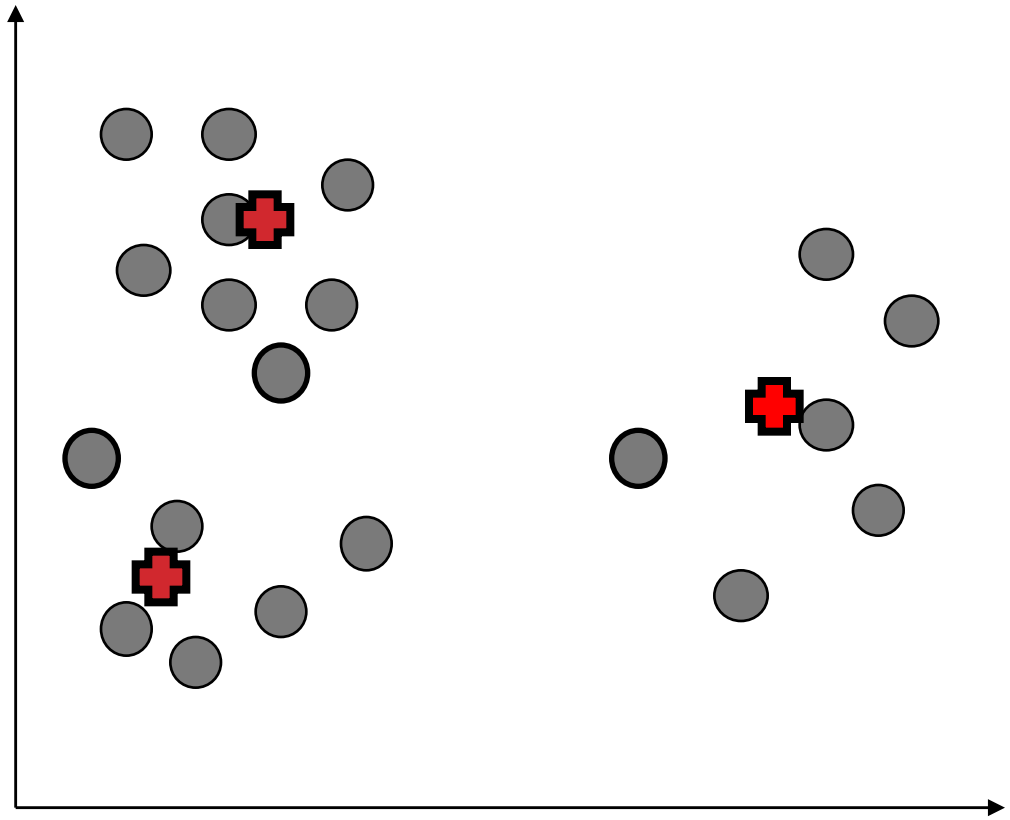
$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_k - b_k)^2} = \sqrt{\sum_{i=1}^{k} (a_i - b_i)^2}$$

The two dimensional case is the famous Pythagorean Theorem:



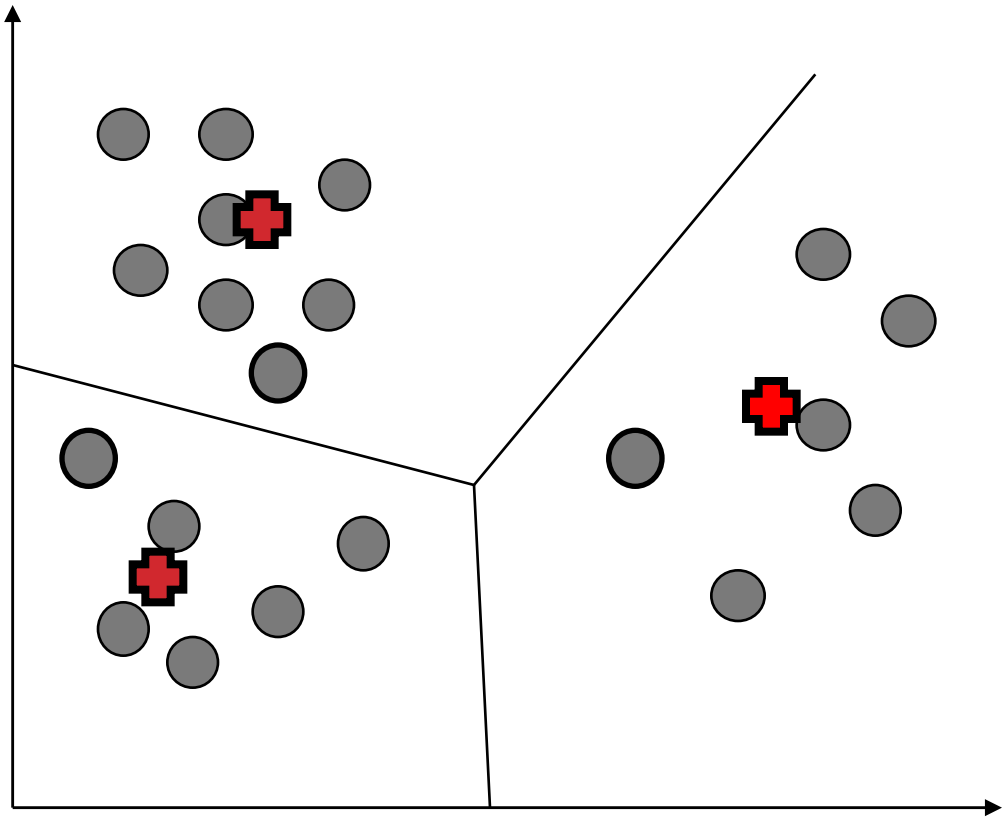
FIND NEW CENTROIDS

4. Compute new centroids per cluster, which is the avg coordinate for all objects within the cluster.



DEFINE NEW CLUSTER

5. Repeat steps 3-4 until cluster centroids converge



Demo:
http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html

KMEANS DISCUSSION

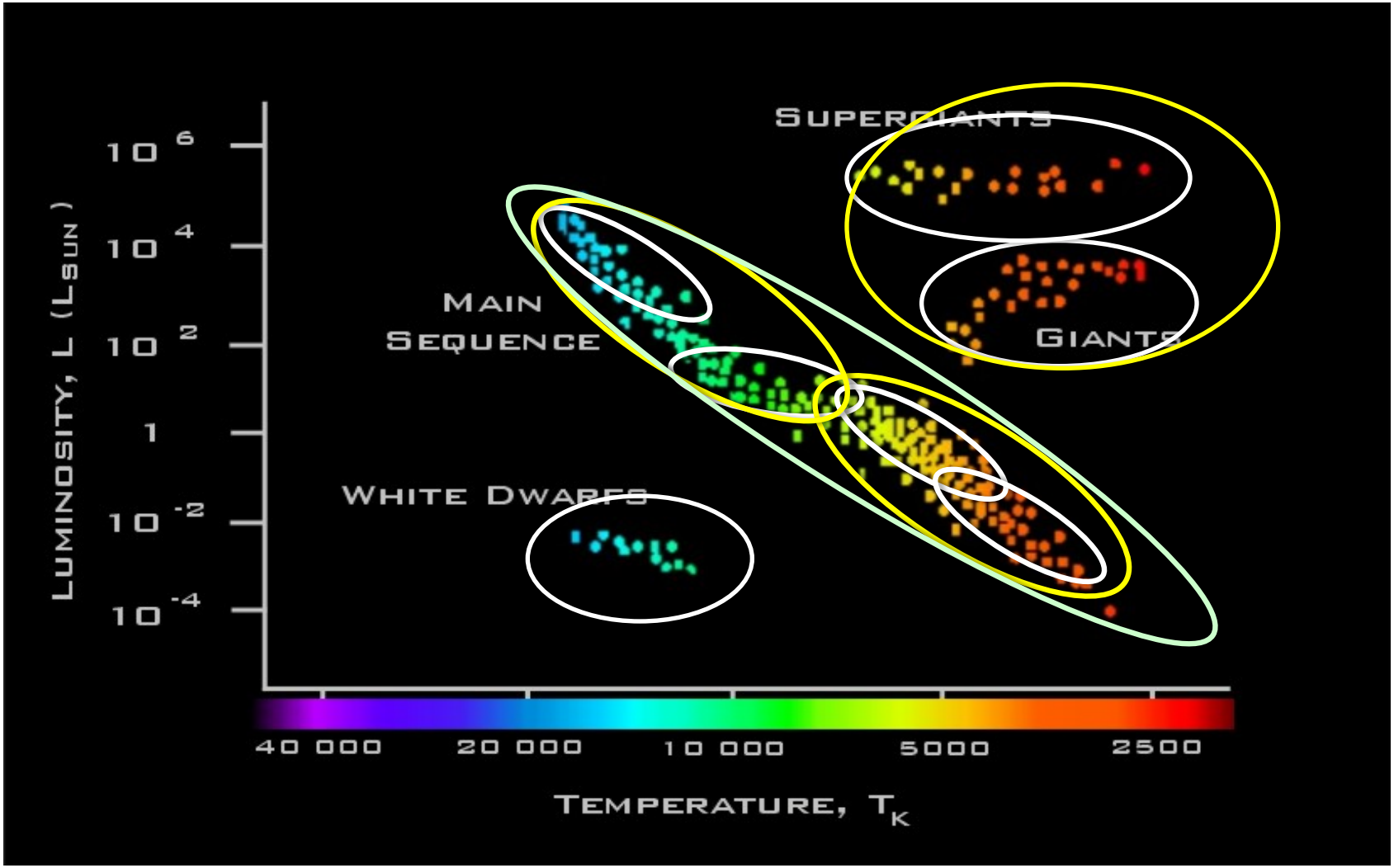
- Most common distance metric is Euclidean distance, but others are appropriate

- Many rules from k-NN apply here
  - Scaling of features matters
  - Can up-weight important features
  - Features need to be numeric

- Assessment
  - What is best k?
  - Is this a good fit?
  - What do the clusters mean?

CLUSTER HIERARCHY

Clusters can be embedded in other clusters. Hierarchical clustering attempts to uncover this embedding.



GENERIC ALGORITHM

This is the generic algorithm for agglomerative clustering methods.

- Compute all pairwise similarities
- Place each instance into its own cluster
- Merge the two most similar clusters into one
  - Replace two clusters into the new cluster
  - Recompute intercluster similarity scores

- Repeat until there are only k clusters left

EXAMPLE

In our first homework all students filled out a DS profile self-assessment

6. In each box, give a ranking of 1-10 on well you think you are in the category listed. A 1 should mean you are a complete novice and a 10 means you are pretty much an expert.

Data Visualization	
Computer Science	
Mathematics	
Statistics	
Machine Learning	
Business Strategy	
Communication	

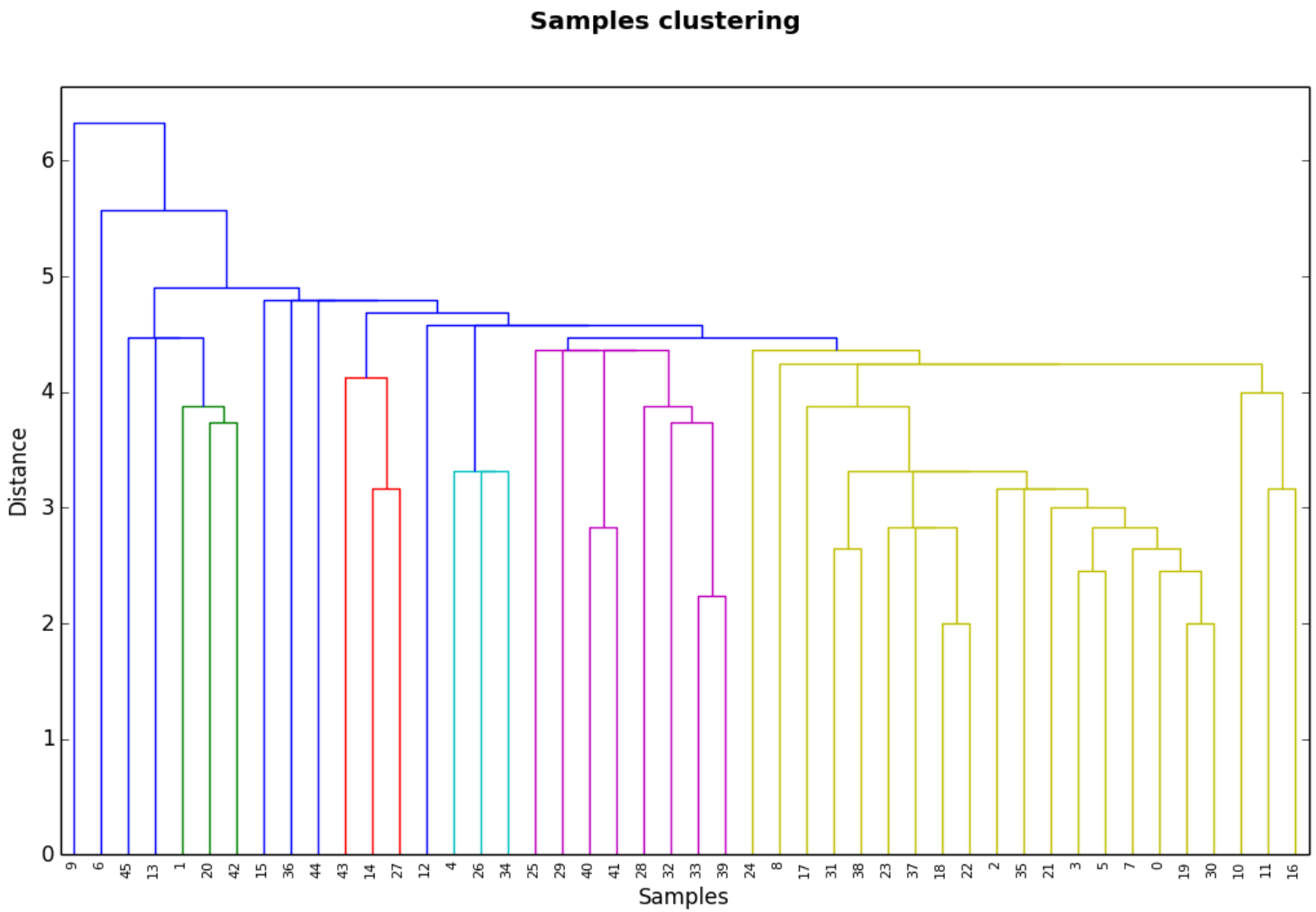
APPLICATION

Goal :assign students into study groups of size 4, where each study group is maximally diverse with respect to skill distribution

We'll use cluster analysis to segments students into roughly equal sized groups. We'll also attempt to qualify each group.

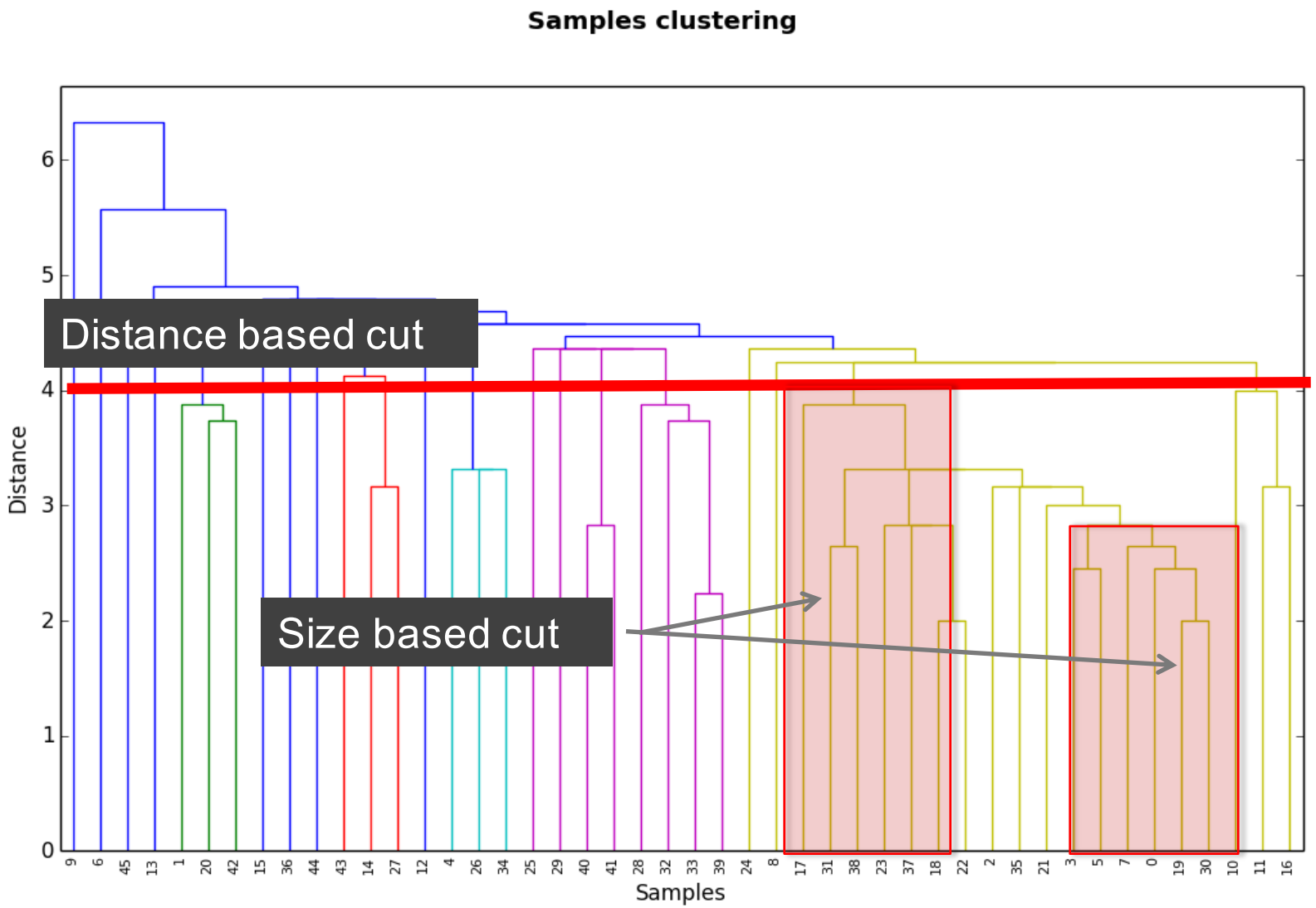
DENDROGRAM

Hierarchical clustering gives us the opportunity to plot the nested nature of the clusters. We can use the dendrogram to spot outliers or determine k.



USING THE DENDROGRAM

We can use the dendrogram to cut the space into k clusters based on: distance or size



EXPLAINING CLUSTERS

We look at the mean adjusted centroid of each cluster to get a sense of what/who the cluster describes.