Clustering Prework

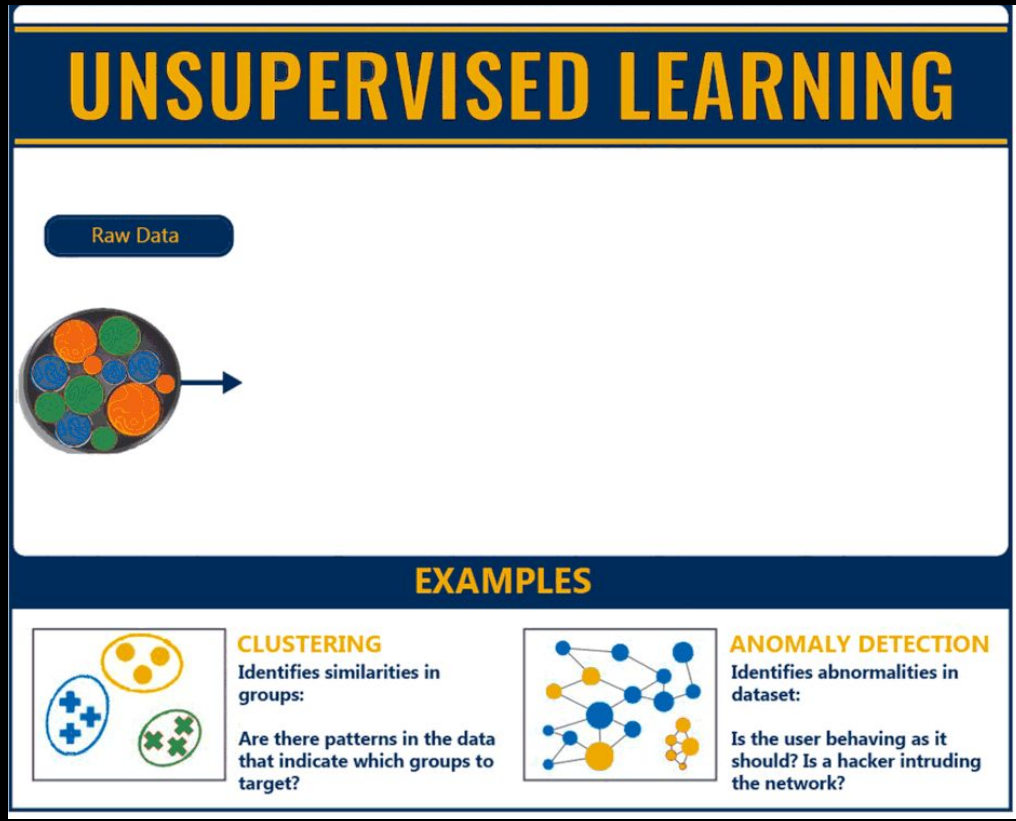
David Moste, Sam Lojacono, Kiana Herr,
Joel Bianchi, Maxwell Yearwood, Seth Adams

Please Complete this Pre-Survey
By TUESDAY MAY 2nd, 11:59PM

MANDATORY PRE-SURVEY LINK

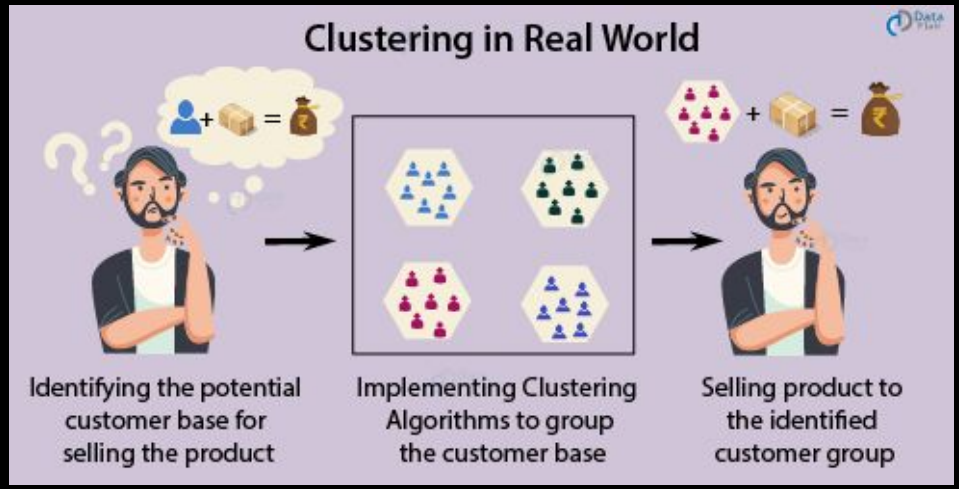
What is Clustering?

- An Unsupervised method in machine learning that groups unlabeled examples according to a criteria.
- Data must be unlabeled for clustering to work
- Many different clustering algorithms can be used to group data
  - We will focus of K-Means



What is Clustering? Part 2

- Common applications for clustering include the following:
  - market segmentation
  - social network analysis
  - search result grouping
  - medical imaging
  - image segmentation
  - anomaly detection
- After clustering, each cluster is assigned a number called a cluster ID. Now, you can condense the entire feature set for an example into its cluster ID. Representing a complex example by a simple cluster ID makes clustering powerful. Extending the idea, clustering data can simplify large datasets.

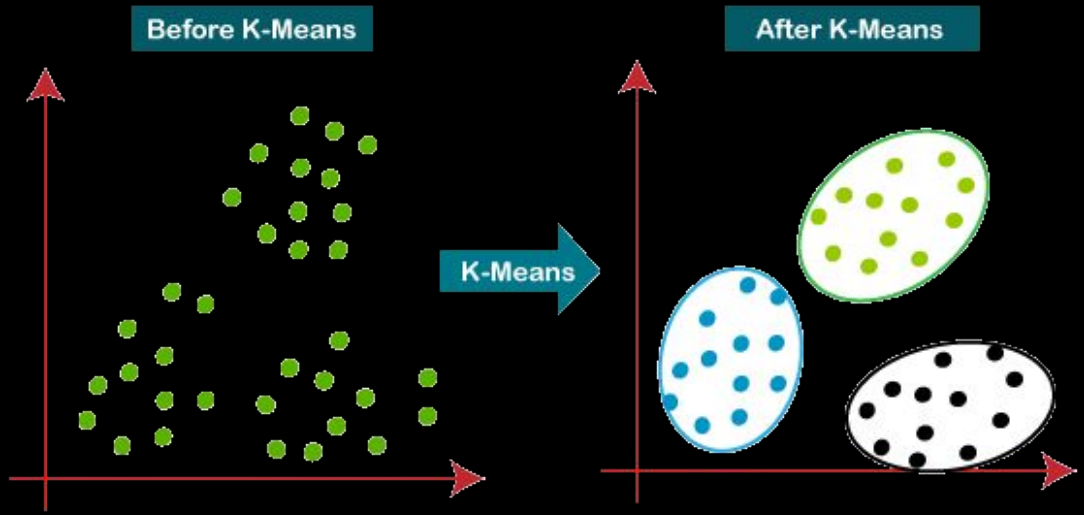


Applications of k-means clustering in the real world

- There are many uses of the k-means algorithm
- It helps creates classification models
  - Useful for if data may be simple yes/no or pass/fail
- Text categorization
- Delivery store optimization
  - Where should companies place the location of a new store?
  - What are the number of optimal launch points for deliveries?
  - Use traveling salesman problem to work out optimal truck routes
- Customer discovery and identification
  - What kind of customers does a company have? How should they appeal to these people? What were their past purchases?
- Crime classification
  - Where types of crimes may occur in specific given areas
- Fantasy league and draft picks
  - Identify players based on similar statistics, features or properties

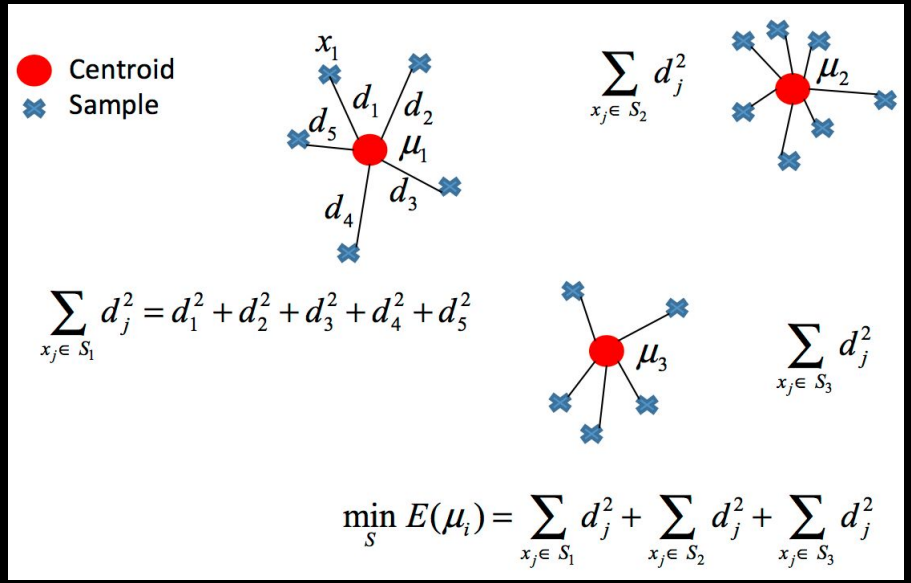
What is k-means clustering?

- So what is k-means **clustering** exactly?
- Clustering is a method used in machine learning that organizes a large set of data into different groups, based on given similarities



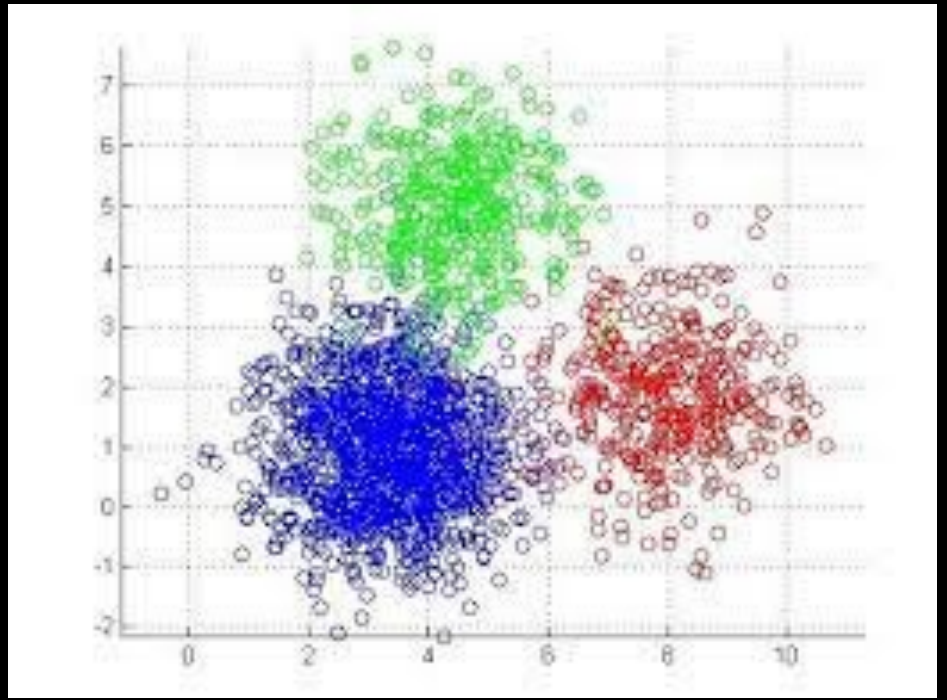
Why do we call it ***k means*** though?

- K-means *means* is that we are looking for a certain amount, k, of centers or centroids we are looking for in the dataset
- Each of these centroids represent the center of each group we are looking for. It can be a real or imaginary location in the dataset
- Each centroid (real or unreal) is found by averaging the data
- The algorithm identifies k number of centroids, assigning every data point to the nearest centroid, while keeping the centroids and the clusters themselves as small as possible



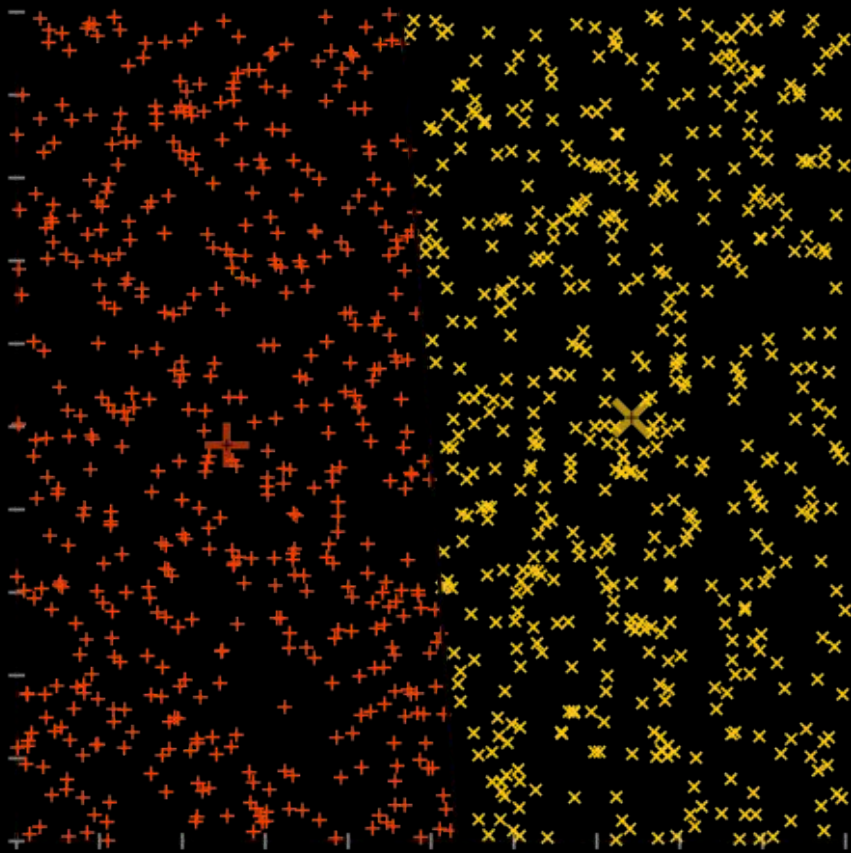
The structure and nature of K-means clusters

- The resulting clusters will be usually circular or globular in shape
- K-means tends to pick globular clusters, but depending on the data, this may not always be the case
- However for k-means to work:
  - Clear boundaries need to exist
  - It may not work if the nature of the dataset is not even close to hinting look like
- K-means will still tend to pick spherical groups



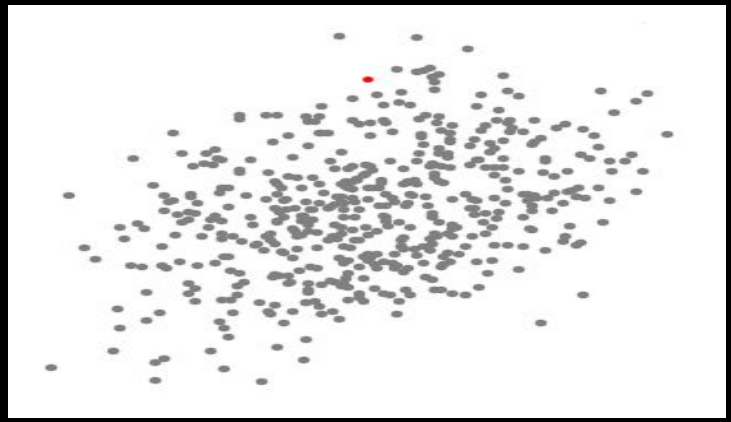
What happens if K-means doesn't work or fails?

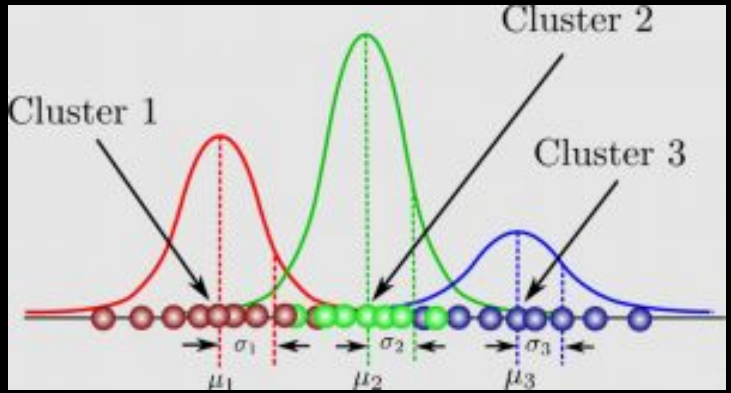
- The algorithm still clusters data, but does not show you ***where*** it doesn't cluster. Take a look at the example on the right
    - The data are uniformly distributed, and k-means did cluster, but as you can see, there is no real CLEAR boundary between the two clusters



Other kinds of clustering other than k-means

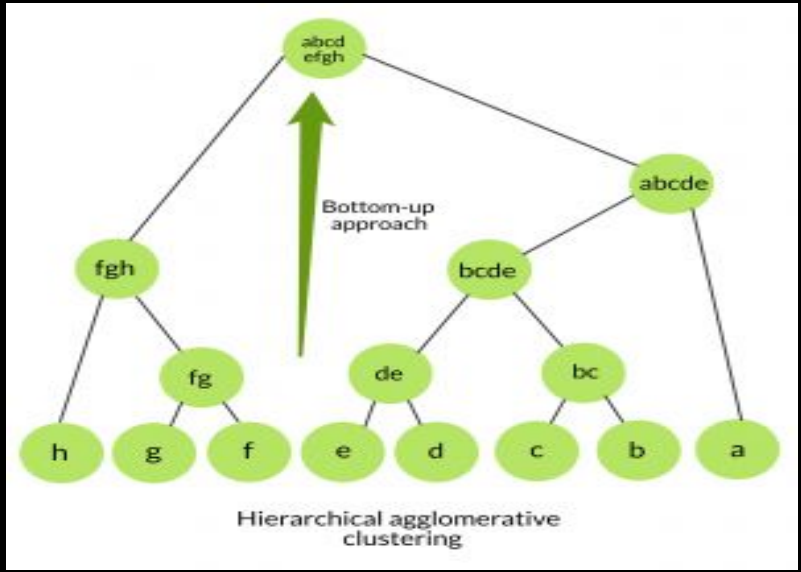
- Means-Shift Clustering
  - Sliding window based algorithm that tries to find densely populated areas of data
- Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM)
  - Perhaps a better option than K-means
  - Assumes data is Gaussian distributed
  - Two parameters to describe the shape of the cluster
    - Standard deviation
    - Mean of the cluster(s)





Other kinds of clustering other than k-means

- Agglomerative Hierarchical Clustering
  - Assumes one data point is a cluster
  - Consecutively merges pairs of clusters until the result is one large cluster
  - This is known as bottom up hierarchical clustering



Check out these videos for more information!