a. Andrew Sen
    i. K-Means is an iterative algorithm. First, random observations are chosen to be centroids. The other observations are assigned to the centroids they are closest to. The centroids are then recalculated to minimize the average distance between themselves and the observations in their respective clusters. This process is repeated until the centroids converge.

    Hierarchical clustering is a bottom-up algorithm. Each observation is first put in their own cluster. Then, the algorithm calculates the distance between each cluster and every other cluster. The formula used to define the distance between clusters varies. Single-link gives the shortest distance between any two points in each cluster and tends to result in elongated clusters. Complete-link gives the longest distance between any two points in each cluster and tends to result in compact clusters, but it is sensitive to outliers. Average-link gives the average distance between all points in both clusters. After calculating all distances, each cluster is combined with the cluster closest to them. This process is repeated until there is only a single cluster. The result of hierarchical clustering can be represented as a dendogram, and the graph can be cut at any depth in order to get the desired number of clusters.

    Model based clustering assumes many different models of the data and uses maximum likelihood estimation (MLE) and Bayesian information criterion (BIC) to select the optimal model with an optimal number of clusters.

b. Neo Zhao
    i. kNN is a method in which the algorithm finds the distance and all K numbers of datapoints. For example, we choose random observations to be the “center,” and other randomly placed observations will close in to the “centers” after every iteration. This is not an infinite loop, but rather, until all observations have clumped.

    Clustering is an unsupervised learning method. Ultimately, we would like observations to begin clustering with other observations that share a common ground. There are three methods to clustering: Single-link, Complete-link, Average-link. Within clusters, Average-Link calculates the average distance between 2 points, while Complete-Link calculates the largest distance and Single-Link calculates the shortest distance between 2 points. After this process, all of the clumps will begin to combine into one large clump.

    Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA) are used for data reduction; however, LDA considers the class while PCA does not. PCA strives to reduce dimensions of the data while reducing the number of axes.

C. - Jack Asaad

Building off of what Neo had stated before, the problem PCA seeks to find a solution to is finding a way to concatenate factors in a smart way into meaningful single factors that are correlated with one another and are ordered from most to least important. PCA looks to find a line of best fit for data points in a data set, it is important that the data is standardized and continuous so we protect from outliers and faulty data. Then PCA seeks to find covariance to identify correlations and does this via a covariance matrix, from there principal components are found by computing the eigenvectors and eigenvalues of the covariance matrix.

As Neo had stated before, LDA considers class when making computations, thus it is more practical to use than PCA in environments where the class is known. LDA looks to maximize the separation between classes and minimize the variance within a class. So first LDA finds the variance or distance between classes and computes a within-class variance for each class and constructs a reduced dimension space that maximizes class separation and minimizes internal class variance.

These two feature engineering techniques are important in machine learning for manageability, what I mean by that is that we want to reduce features into a more palatable package so that we can improve interpretability of our models. Of course this usually comes at the cost of accuracy, but such is the tradeoff between interpretability and accuracy.