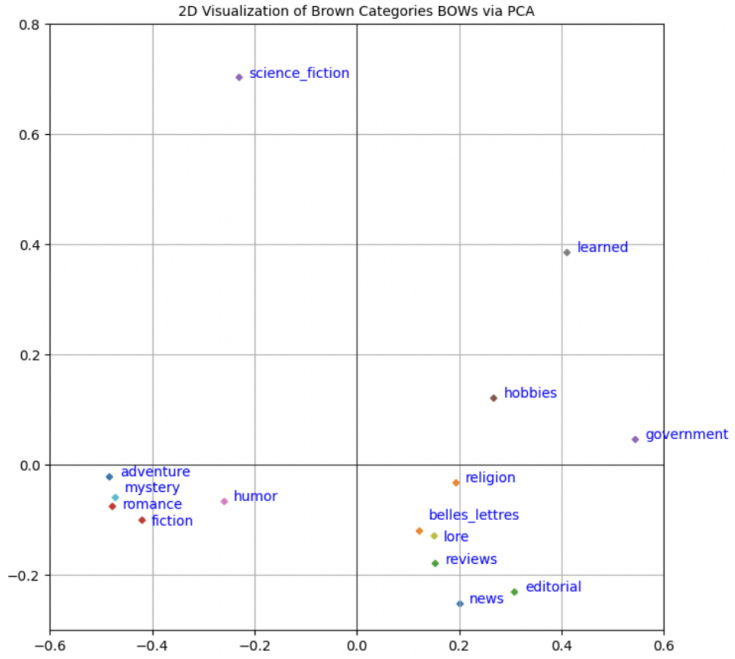
CAS CS 505
Lec 7

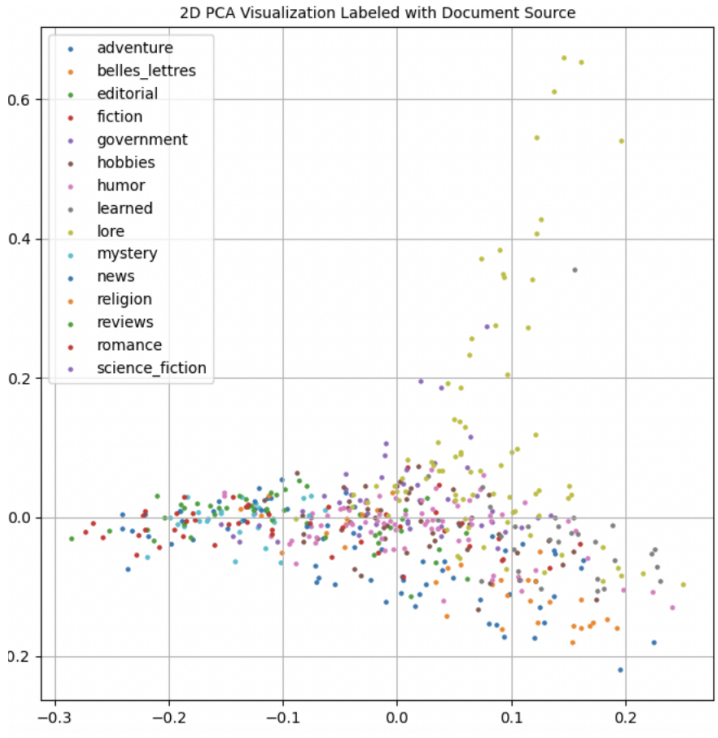
Introduction to Machine Learning; Unsupervised ML; Clustering with KMeans, Hierarchical Clustering

1. Addendum to PCA from last time

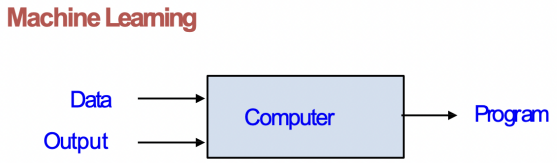
a.



b.

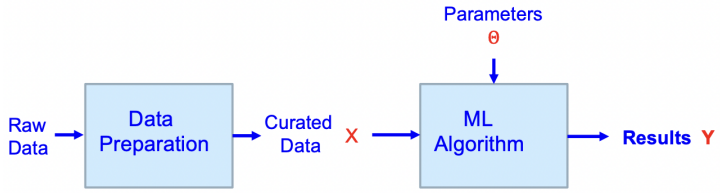


2. What is Machine Learning?
    a. "Learning is any process by which a system improves its performance from experience." - HerbertSimon (A founder of AI)
    b. "Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed." - Arthur Samuel (Creator of first checker-playing program, 1959)
    c. Machine Learning is the study of algorithms that perform
        i. Some task T (The problem/task to solve)
        ii. After some experience E (Training)
        iii. And improve in some performance metric P (Testing) A well-defined learning task is given by <P,T,E>
    d. The ways that these three parameters are defined gives rise to the variety of different approaches to Machine Learning.

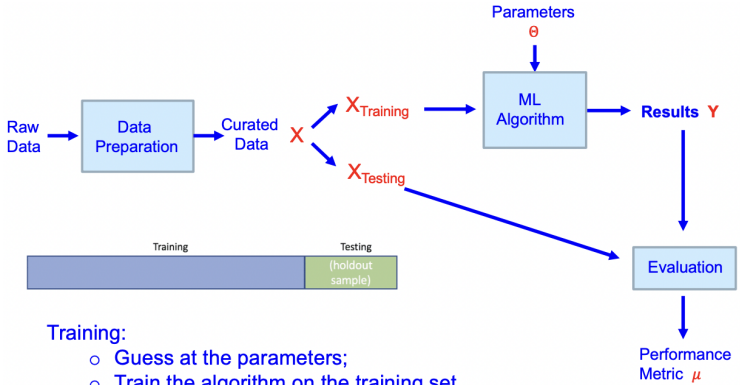


    e. → this is supervised learning

3. Introduction to Machine Learning
    a. There are several flavors of Machine Learning….
    b. Unsupervised Machine Learning Workflow:



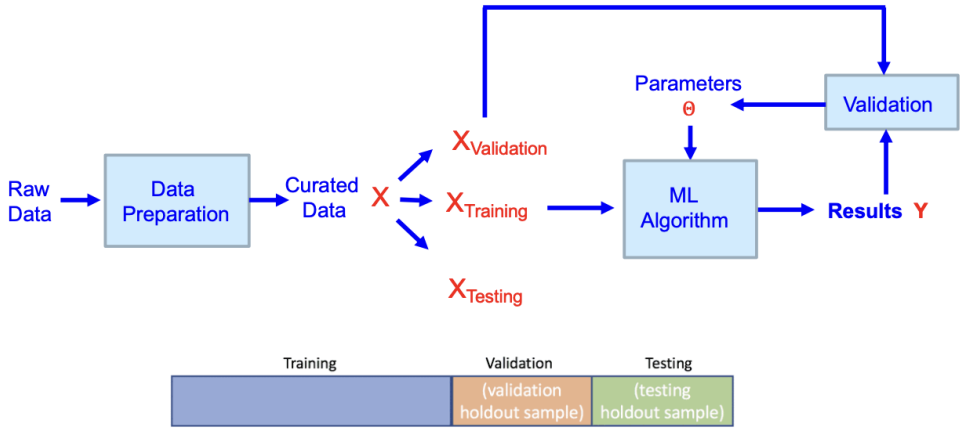
    c. Supervised Machine Learning Workflow:



i. Training:
    1. Guess at the parameters
    2. Train the algorithm on the training set
    3. Evaluate using the testing set
    4. (adjust parameters to get better evaluation)

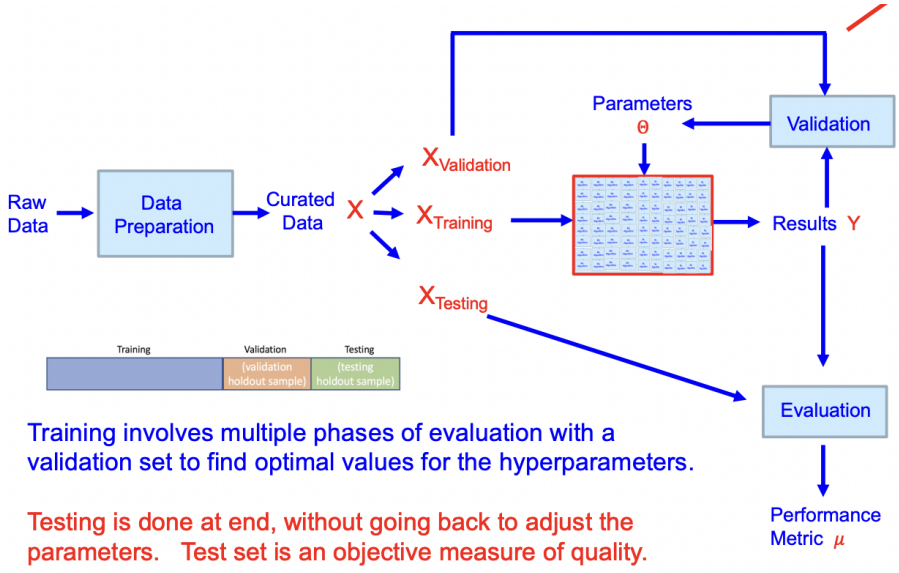
ii. What's wrong with this picture?
    1. Either you guess at the parameters, or adjust them to fit the testing set. You have no objective measure of quality!
    2. You fit your data based on the test, which is great if the test is a representation of the world, but becomes a problem if it is not

d. Supervised Machine Learning Workflow (Actual):

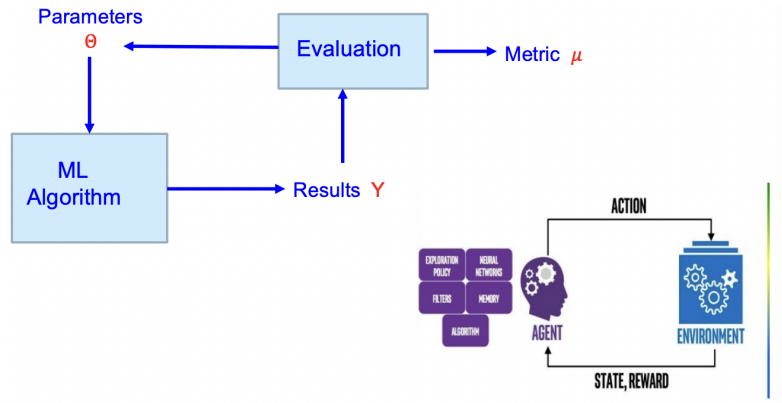


i. Take out validation set (split data into three different sets)
ii. Test it on the validation set and update the parameters (don't touch the testing set)
iii. Trained the sample based on the estimate
iv. Then you test it on the testing set (there is no step from testing set going back to the parameters to improve the testing set)
v. Training involves multiple phases of evaluation with a validation set to find optimal values for the hyperparameters.
vi. Testing is done at end, without going back to adjust the parameters. Test set is an objective measure of quality.

e. Deep Learning (Supervised ML with Neural Networks):

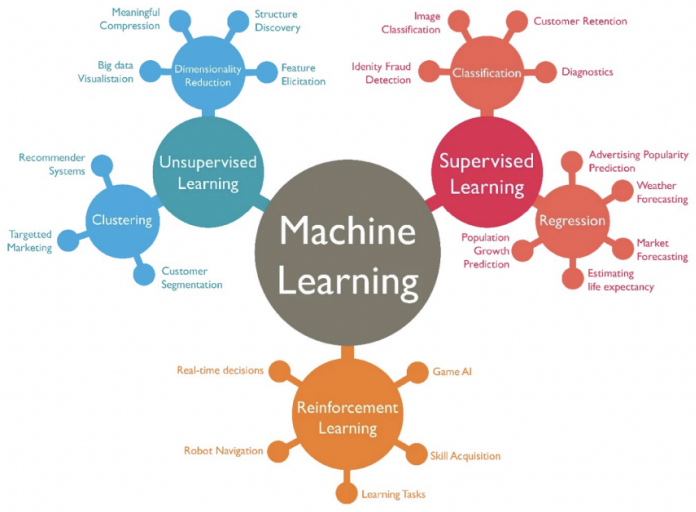


f. Reinforcement machine learning workflow:

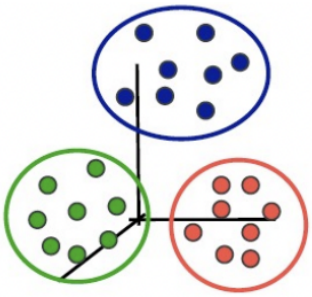


i. Put agent in an environment and the agent gets a reward based on the action it takes (learning is the cycle)

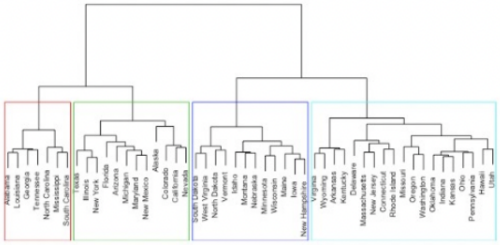
g.



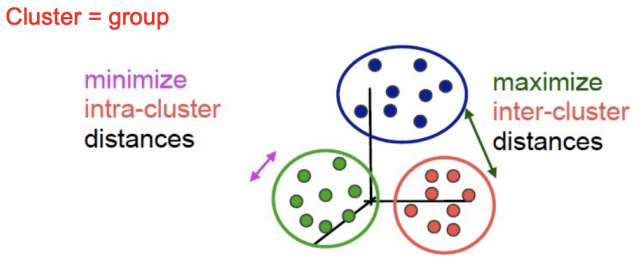
4. Unsupervised Learning: Clustering
    a. What is Clustering?
    b. Use Clustering to figure out how similar/different the data/documents are
    c. There are two basic types of clustering:
        i. Partitioning



        ii. Hierarchical



    d. For now we will only consider partitioning algorithms: each object belongs to exactly one cluster.
    e. A grouping of data objects such that the objects within a group are similar (or near) to one another and dissimilar (or far) from the objects in other groups
    f.



5. The clustering problem:
    a. Given a collection of data objects, find a grouping such that
        i. Similar objects are in the same cluster
        ii. Dissimilar objects are in a different cluster
    b. Why is this important?
        i. A stand-alone tool for visualizing and understanding the data
        ii. A preprocessing step for other algorithms
            1. Creating group labels for supervised learning
            2. Indexing or compression often relies on clustering
        iii. Classification where the only data available is unlabeled

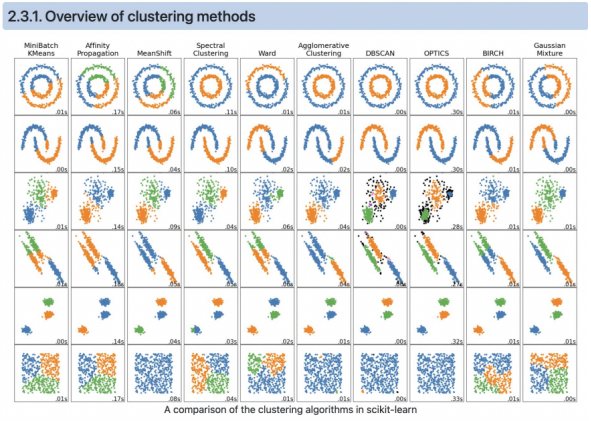
c. Given a collection of data objects, find a grouping such that
    i. Similar objects are in the same cluster
    ii. Dissimilar objects are in different cluster
d. Basic Questions:
    i. What does similar mean? (could be euclidean distance, etc.)
    ii. What are the most efficient algorithms?
    iii. How do we evaluate the quality of the resulting partition?
e. The Big problem: The notion of a cluster is ambiguous (the number of clusters are difficult to find)

6. K-Means Clustering
    a. BIG Assumption: Assume in advance you want exactly k clusters.
    b. The K-Means Algorithm:
        i. Given k and a set X = { x1, x2, ..., xn } of points in Rd (d = number of dimensions), find
            1. k points {c1, ..., ck} (called centers, means, or centroids) and
            2. a partition of X into k clusters {X1, ..., Xk} by assigning each point xi to its nearest cluster center,
        ii. such that the cost is minimized.
    c. Cost:

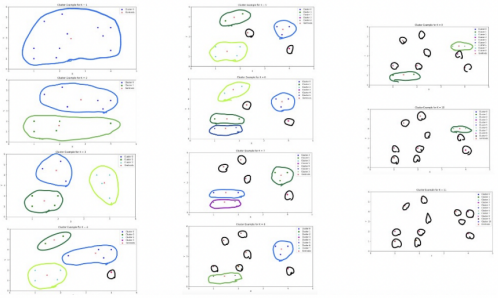
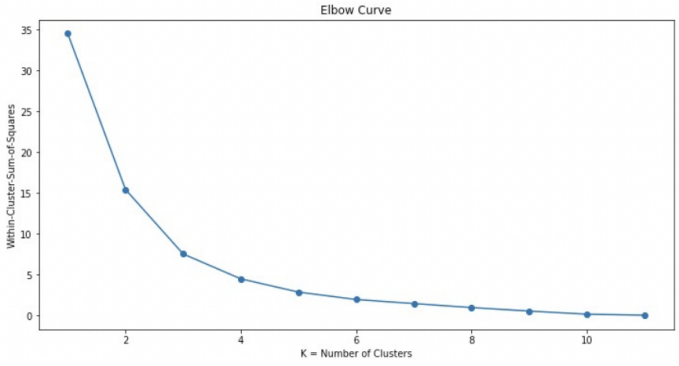
$$\sum_{j=1}^{k} \sum_{x \in X_j} \|x - c_j\|_2^2$$

L2 norm: square of distance between points x and c_j.

    d. For K = 1 (just finding away the mean distance) and K = n (each cluster has one point) the solution is trivial
    e. For other cases, it is NP-hard (probably exponential) for d > 2.
    f. But in practice, iterative greedy algorithms work quite well!
    g. There are many flavors of K-Means! We will only look at the most basic.

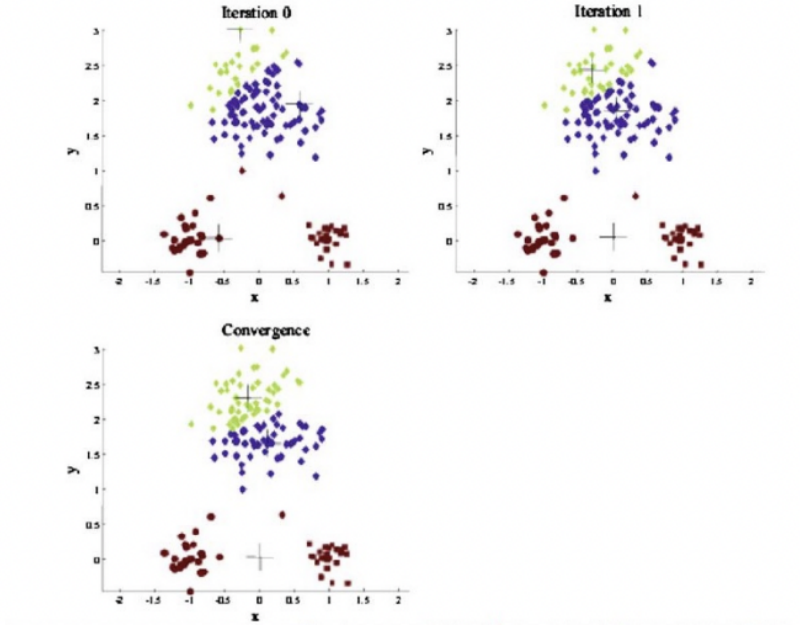


A comparison of the clustering algorithms in scikit-learn

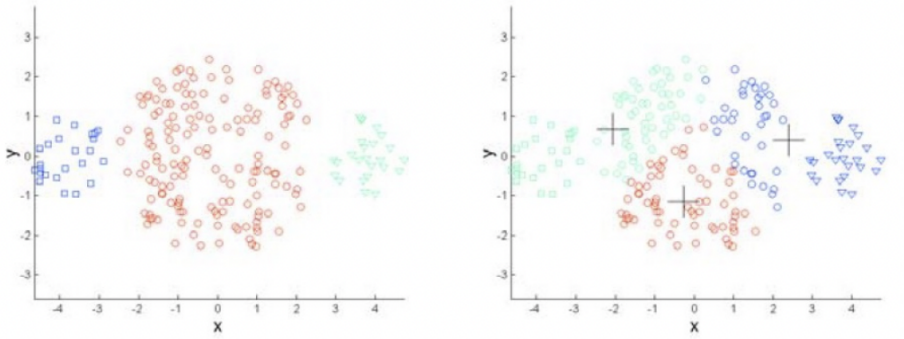
7. Lloyd's Algorithm for K-Means
    a. Repeat until some termination criterion is met: (update the centroid based on the center continuously)
        i. Randomly* choose the k centroids { c1, ..., ck };
        ii. For each $1 \leq j <= k$ set the cluster X_j to be the set of points in X which are closest to the center cj ;
        iii. For each j, update the value of cj to be the mean of the vectors in Xj . *
        iv. NOTE: This is a Hill-Climbing (search) algorithm, where the cost is the squared intra-cluster distances; it often converges quickly, but the choice of the initial set of centroids is critical, and essentially all of the refinements to this algorithm have to do with how this initialization step is done.
8. Evaluating K-Means: What should K be?
    a. The main problem is that the cost (sum of squared intra-cluster distances) decreases as number of clusters gets smaller! Which is the best one?
    b. 
    c. How to choose the right K?
        i. If you already know the ideal K for your problem, you can use that K
        ii. Well... it depends.... but a naive method is to look at the graph of K vs cost and pick an appropriate midpoint between extremes, the so-called "elbow" of the curve.
        iii. 

        iv. The goal is to minimize the number, but not 0
        v. Want to find the number that is small but tells meaningful information
        vi. Look at the angles
            1. Pick the angle that is smallest (k = 3 in this case)

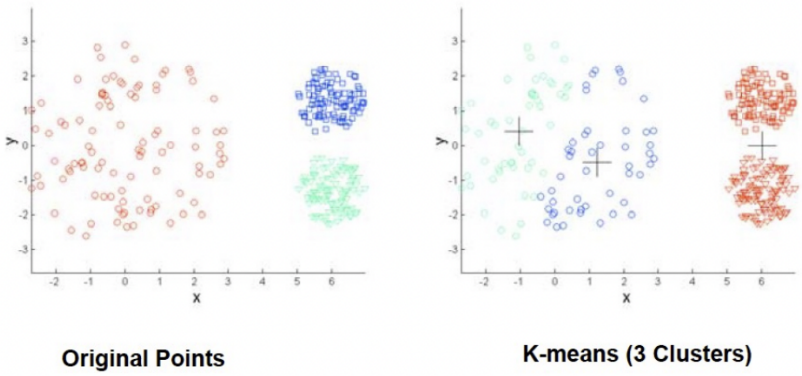
9. Effect of a Bad Initialization
    a. Once you set on k, you might want to search the best position of those k
    b. Sets depend on initialization - so repetition and finding the best cluster is good
    c.



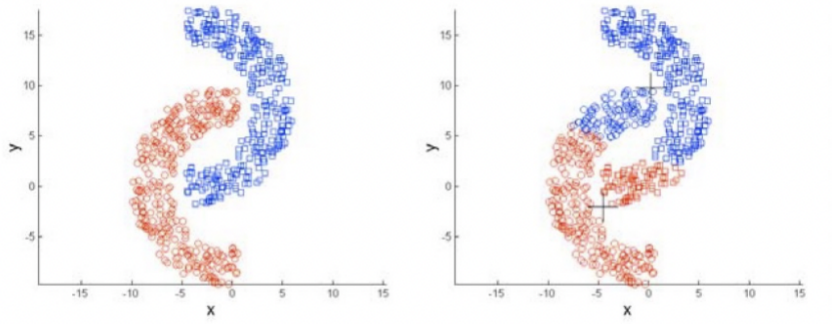
10. Limitations of K-means: Clusters of different sizes
    a.



11. Limitations of K-means: Different Cluster Densities
    a.

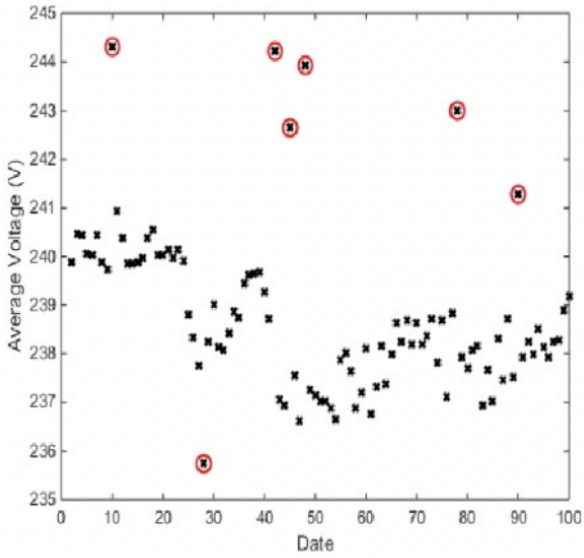


12. Limitations of K-Means: non-spherical clusters



a.

13. Limitations of K-Means: outliers are a problem!



a.

14. Improvements based on initialization
    a. Random Initialization
    b. Repeat random initialization multiple times and take the best solution
    c. Pick random points which are distant from each other
        i. Basis of K-Means++ algorithm
        ii. There are provable guarantees about quality of solutions.
    d. Pick one point in the dataset
    e. Pick one point that is farthest from the point
    f.