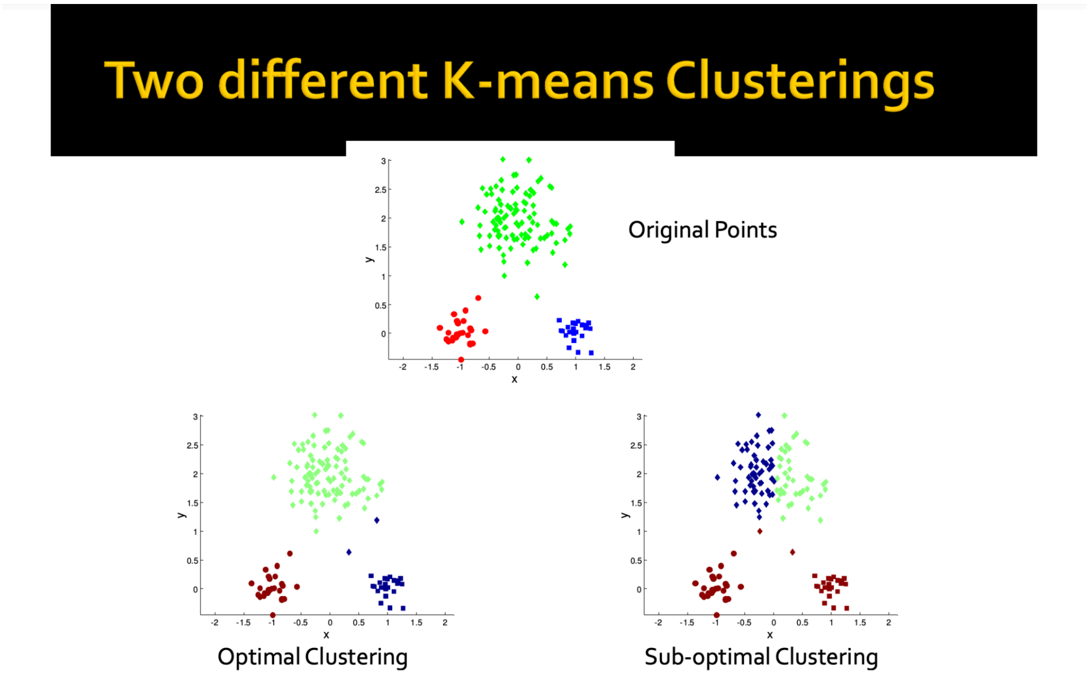
K-Means++

Drawbacks of K-Means

1. K-means is initialization dependent.
    a. The same data, with different initialization, will get different results (different clusters).



Two different K-means Clusterings

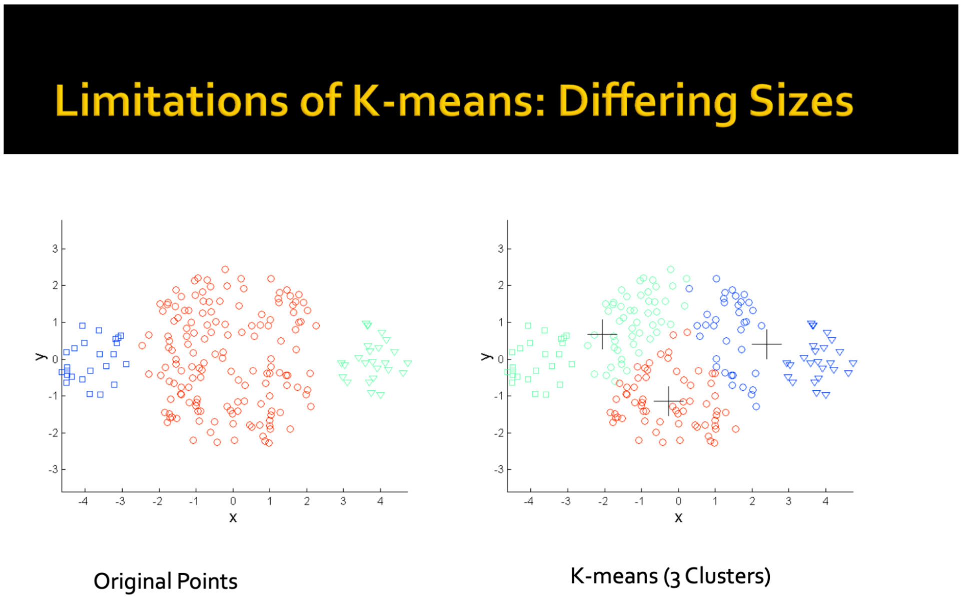
Original Points

Optimal Clustering

Sub-optimal Clustering

Link: visualization tool to see this problem

2. The k-means algorithm may not give the best results for data where the clusters are of varying size or density.

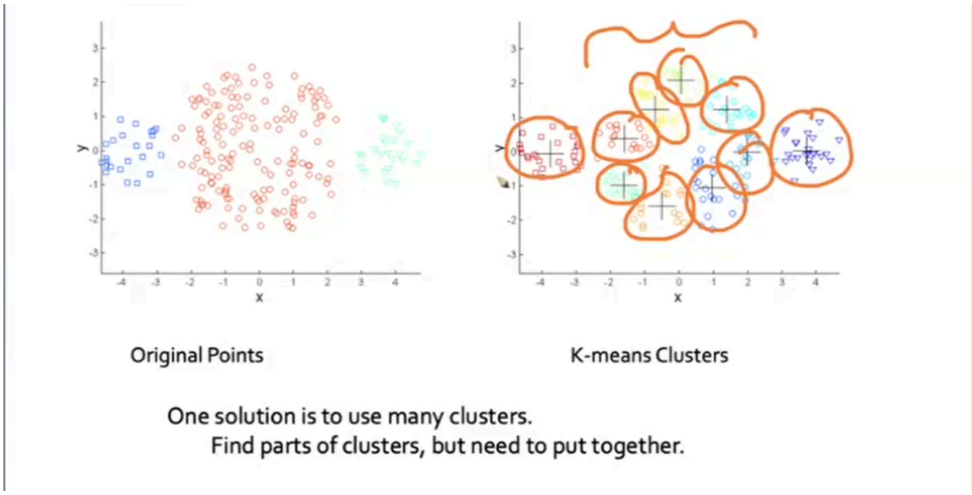


Limitations of K-means: Differing Sizes

Original Points

K-means (3 Clusters)

- **How to solve this problem?** increase the value of K.
- Once clusters are formed, similar clusters can be grouped to form a mega cluster.



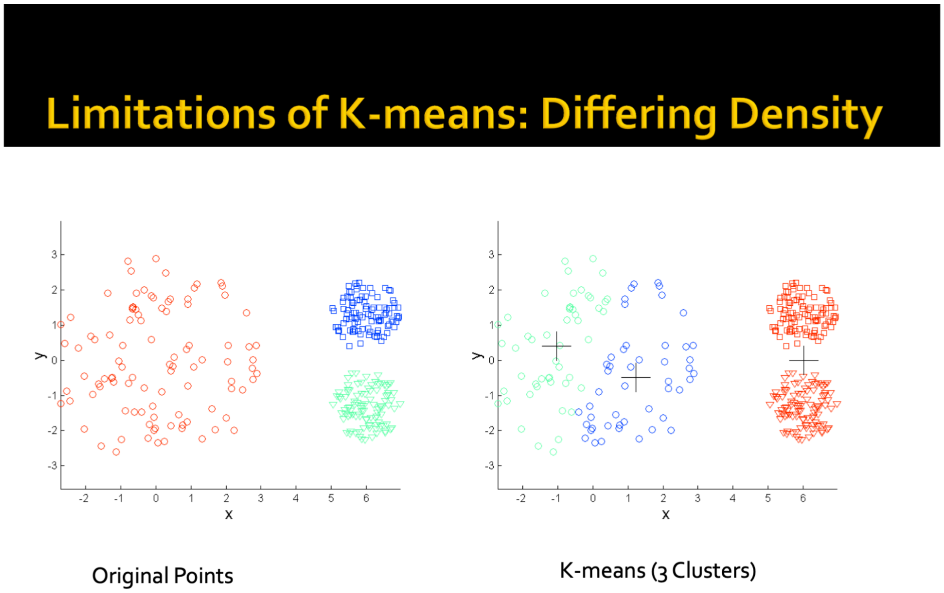
Original Points

K-means Clusters

One solution is to use many clusters.
Find parts of clusters, but need to put together.

- The problem with this approach is the grouping of similar clusters is not easy

3. The number of clusters (k) needs to be defined prior to clustering.

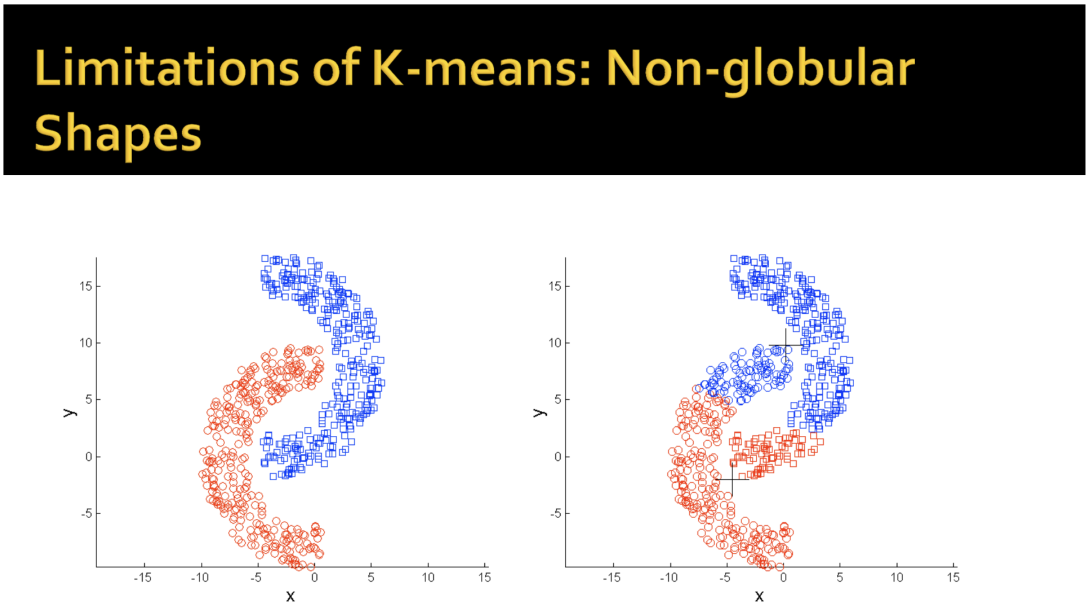


Limitations of K-means: Differing Density

Original Points

K-means (3 Clusters)

4. It does not work well with non-globular clusters.



Limitations of K-means: Non-globular Shapes

K-Means++

- It uses a smarter way to initialize the centroids to improve the clustering algorithm.

- Consider data where we want to initialize 3 centroids.
  - We pick the first centroid at random
  - Now, to pick the second centroid, we want to pick a point that is as far away as possible
- We would want to pick a point that is far away because if two centroids are closer to each other, two clusters for that region of data points will be formed

- We compute the distance from the centroid C1 of all the data points present in our dataset D such as D - {C_1}

- **Risk**: If we select a datapoint as a second centroid with the farthest distance, then an outlier might be picked as a centroid, and we might have a cluster with the centroid C_2 only.

- **Solution**: Pick a centroid **probabilistically**, instead of picking it deterministically.
  - I.e. The probability of picking a centroid is proportional to the distance from the first centroid C_1.

- The steps involved in the initialization of centroids are:
  - ➔ Select the first centroid randomly from the data points.
  - ➔ Choose the next center as the farthest point (probabilistically) from the first center.
  - ➔ The next center would be a data point farthest from both the first and second centers.

- Repeat steps 2 and 3 until **k** centroids have been sampled.

- If there are **outliers** in our data, then instead of choosing them as centroid, we can choose the farthest point as the centroid with a **probability proportional to the distance. (** Default implementation of Sklearn)