08 K-Means

Thursday, 13 May 2021 8:38 PM

Clusterin

→ Most common EDA techniques used to get an intuition about the structure of Data.

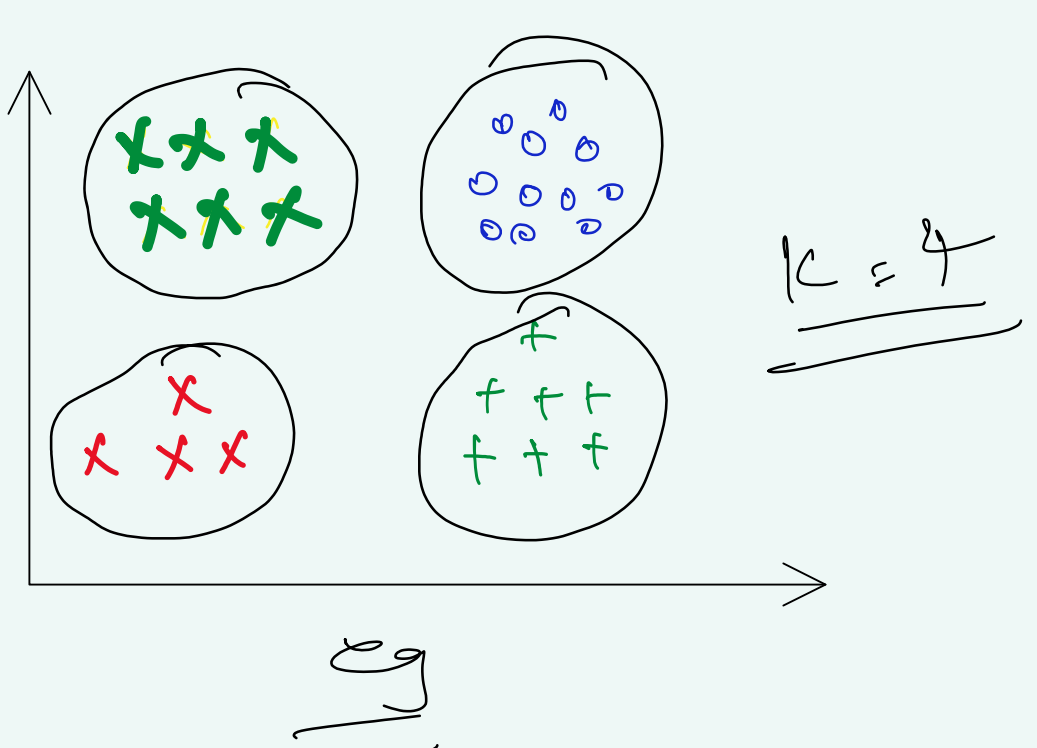
→ It can be defined as the task of identifying subgroups in the data such that point in a sub group are homogenous in the nature.

K-Means Algorithm

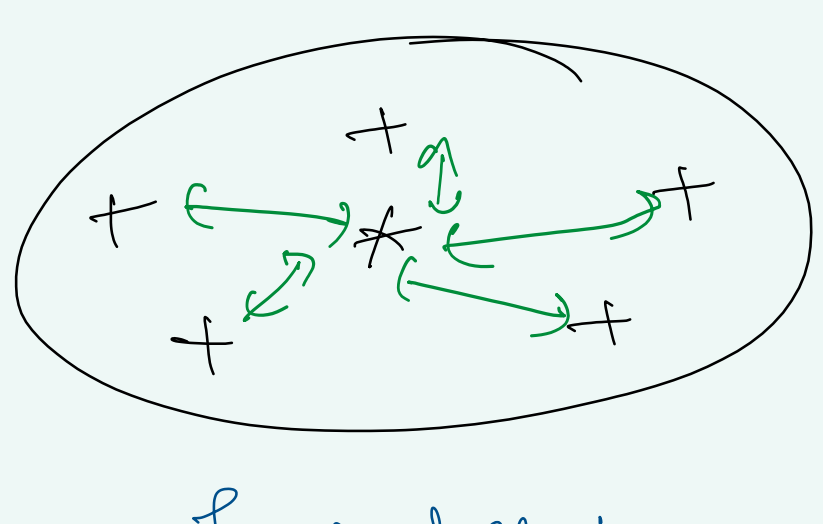
Kmeans is an iterative algorithm that tries to partition the data into K - pre-defined subgroups where each point belongs to only 1 subgroup.

It assigns data points to a cluster such that the sum of the squared distances between the data points and the cluster's centroid is at the minimum.

This approach is K/s Expectation Maximization

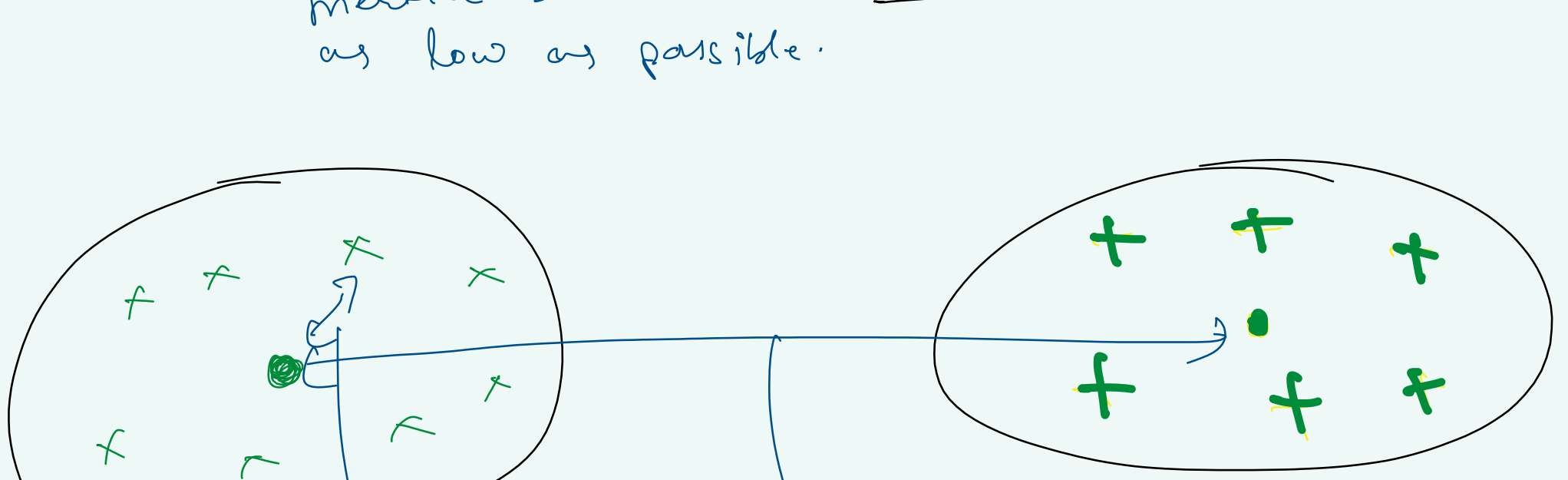


eg



The sum of distances of all the points within a cluster from the centroid of that cluster is called Inertia

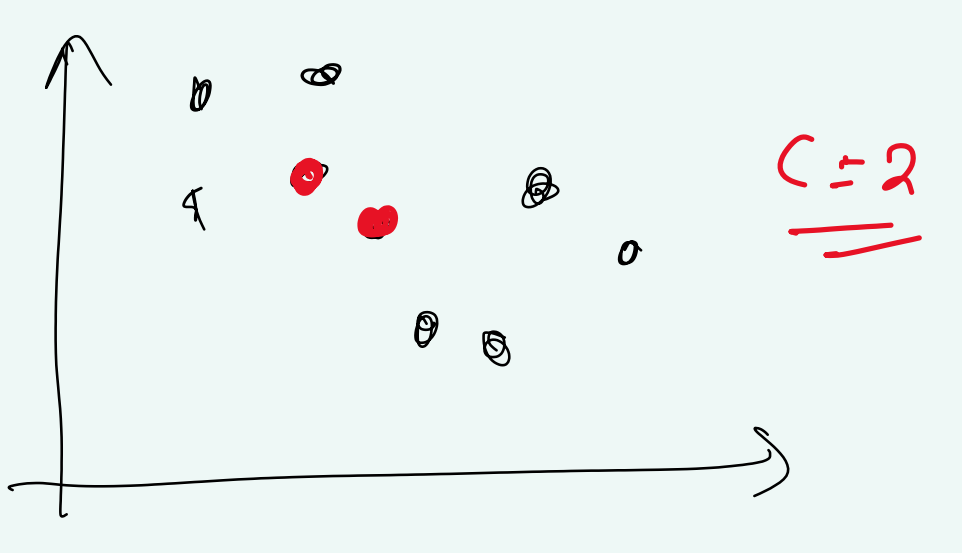
For good cluster, Inertia should be as low as possible.



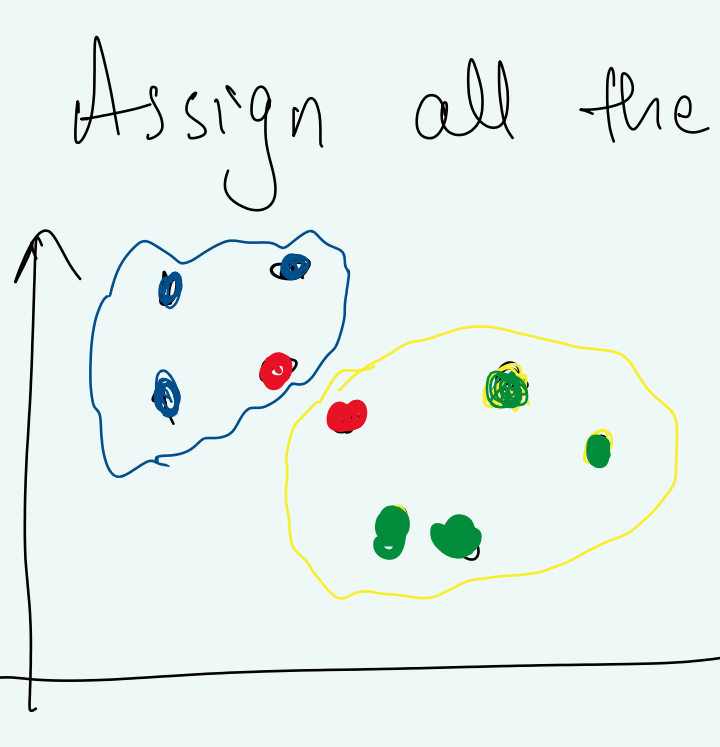
Dunn Index :→ $\frac{\min(\text{Inter cluster distance})}{\max(\text{Intra cluster distance})}$

Step1: Choose the number of cluster k

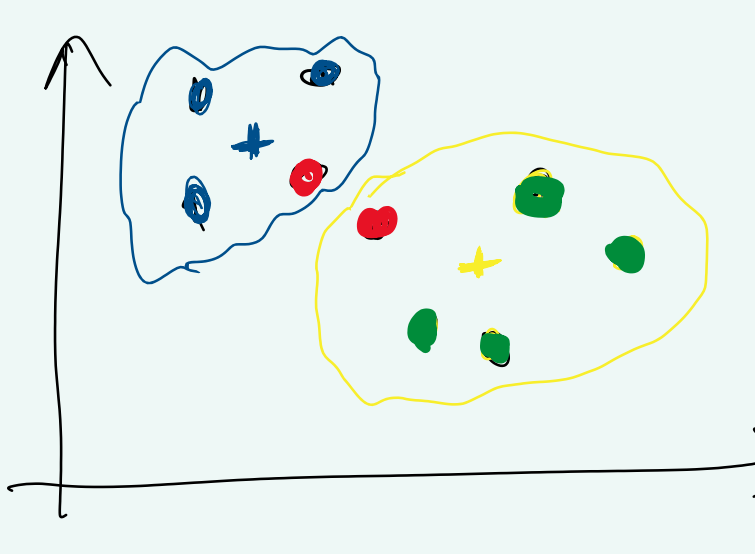
Step2: Select k random points from the data as centroid.



Step 3 :- Assign all the points to the closest centroid



Step 4 :- Recompute the centroid of newly formed cluster.

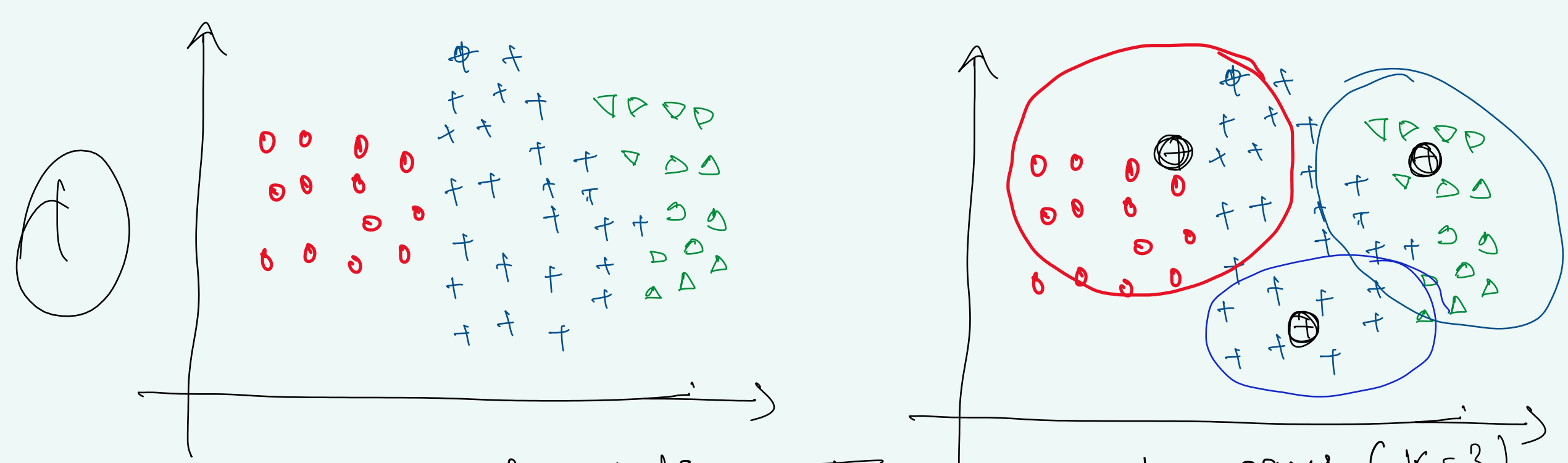


Step 5 : Repeat step 3 and step 4

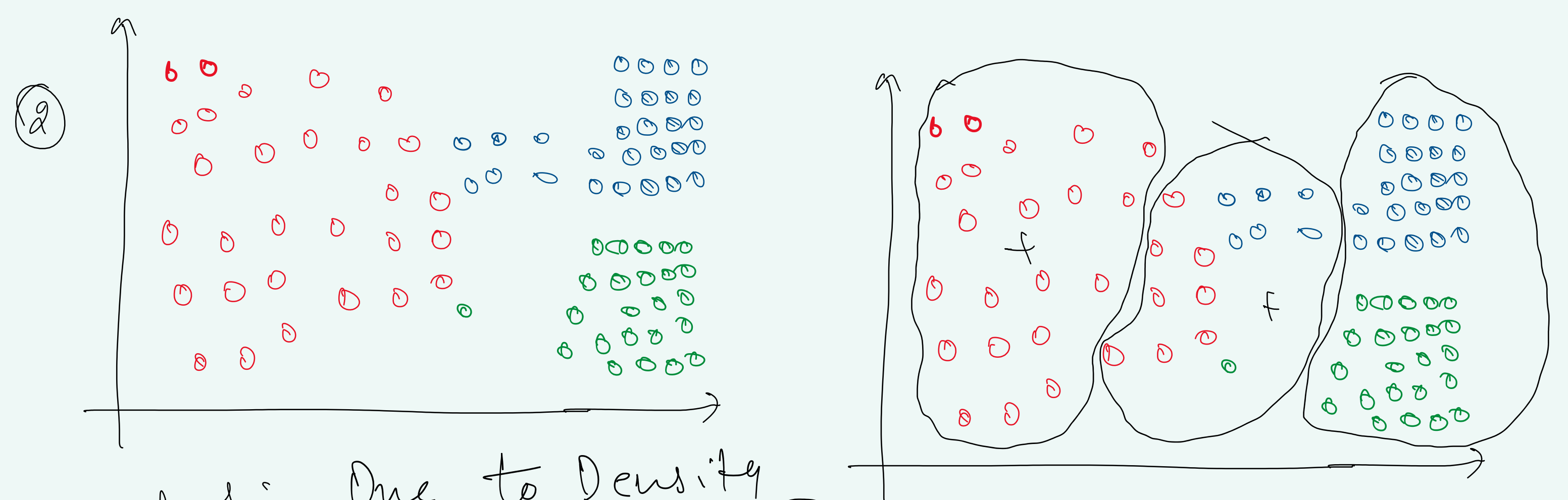
Stopping Criteria of k-means :-

1. Centroid of newly formed clusters do not change.
2. Points remain in the same cluster
3. Maximum number of iterations are reached.

Challenges with k-means



(due to size of cluster)



Prob: Due to Density

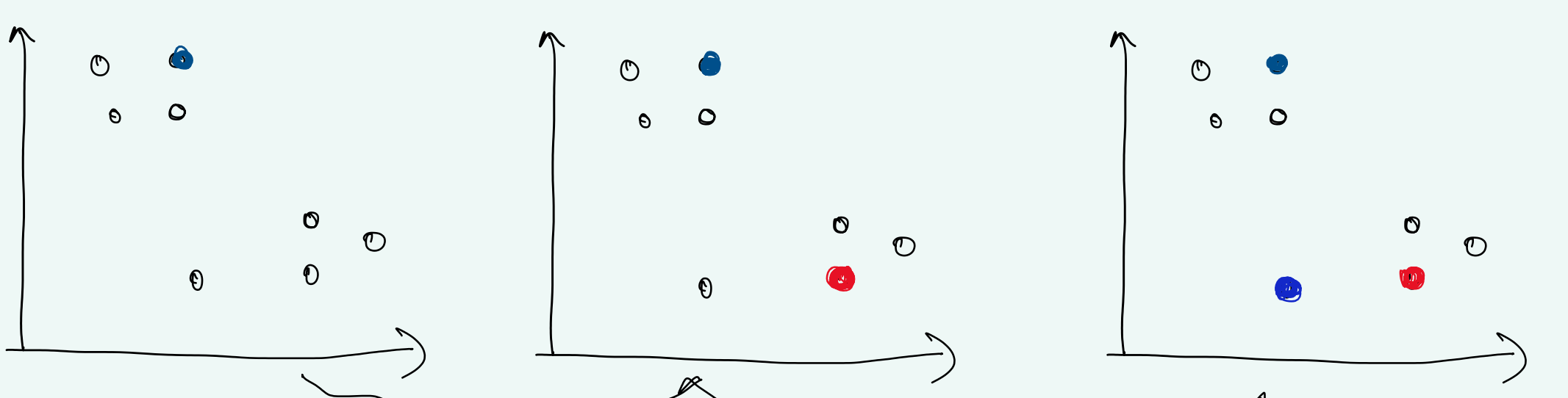
Sol^n :- Use a higher number of cluster.

K-Means + + +

→ So initializing the centroids in kmeans is kinda problematic.

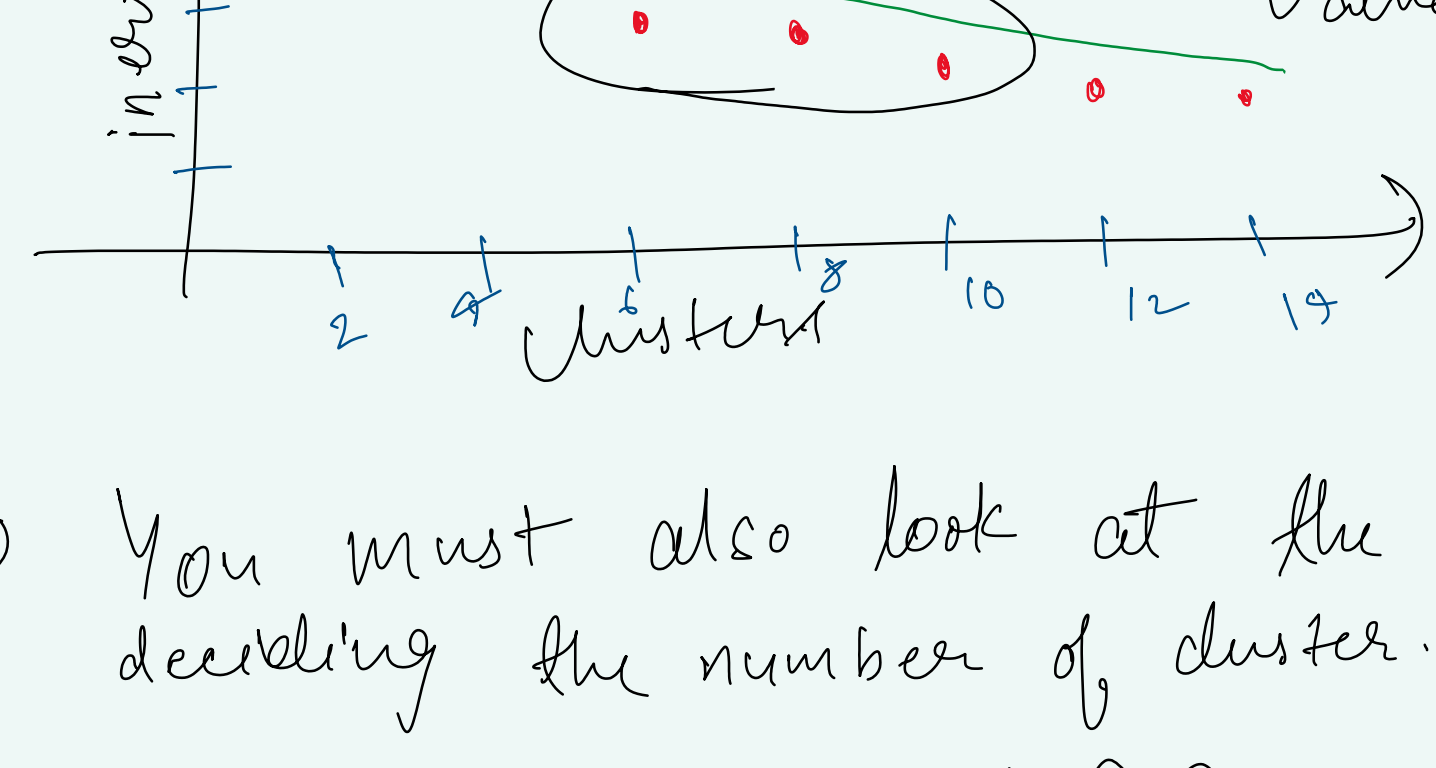
→ It specifies a procedure to initialize the cluster centers before moving forward with the standard k-means clustering algorithm.

Steps:



Q How to choose the right number of clusters in K-means?

→ Elbow curve



Cluster value where this decrease in inertia becomes constant can be choosen as the right value for our data.

→ You must also look at the computation cost while deciding the number of cluster.

K ↑↑ ∝ Computational ↑↑↑