Unsupervised Learning

Agenda:

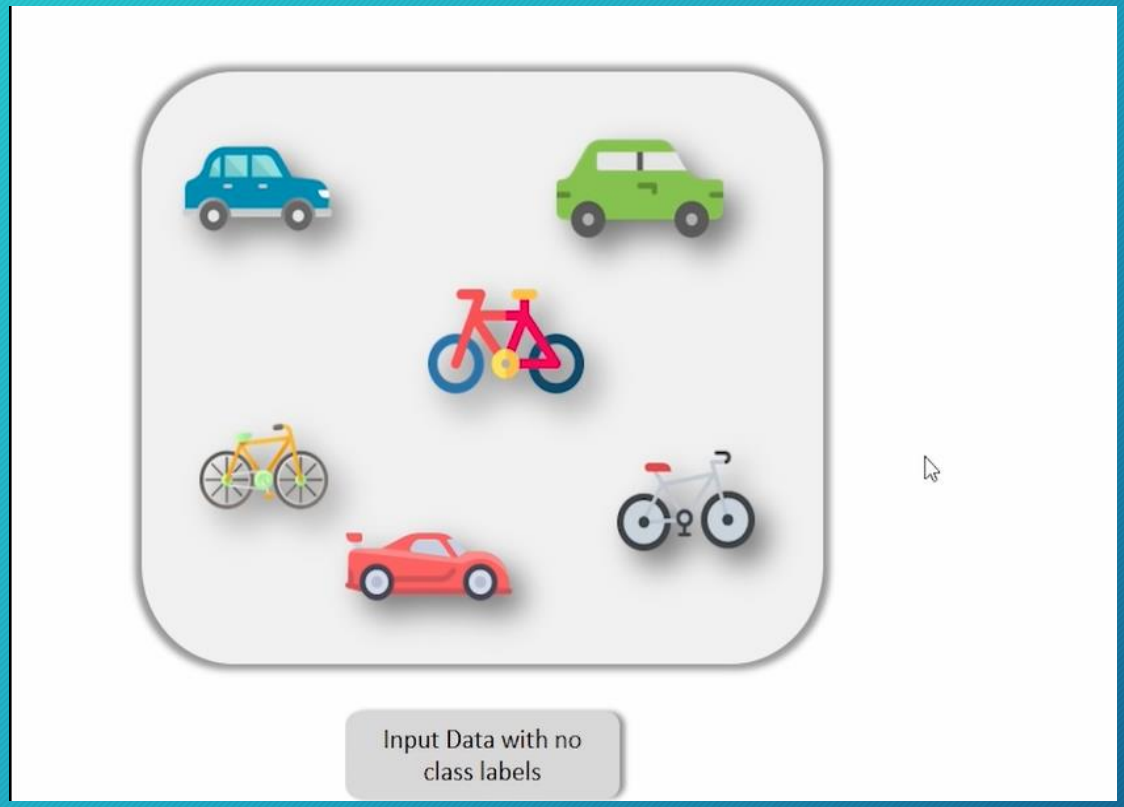
- What is clustering?
- Types of Clustering
- What is K-means Clustering?
- How does a K-Means algorithm works?
- K-means with Python

Unsupervised Learning Moto:

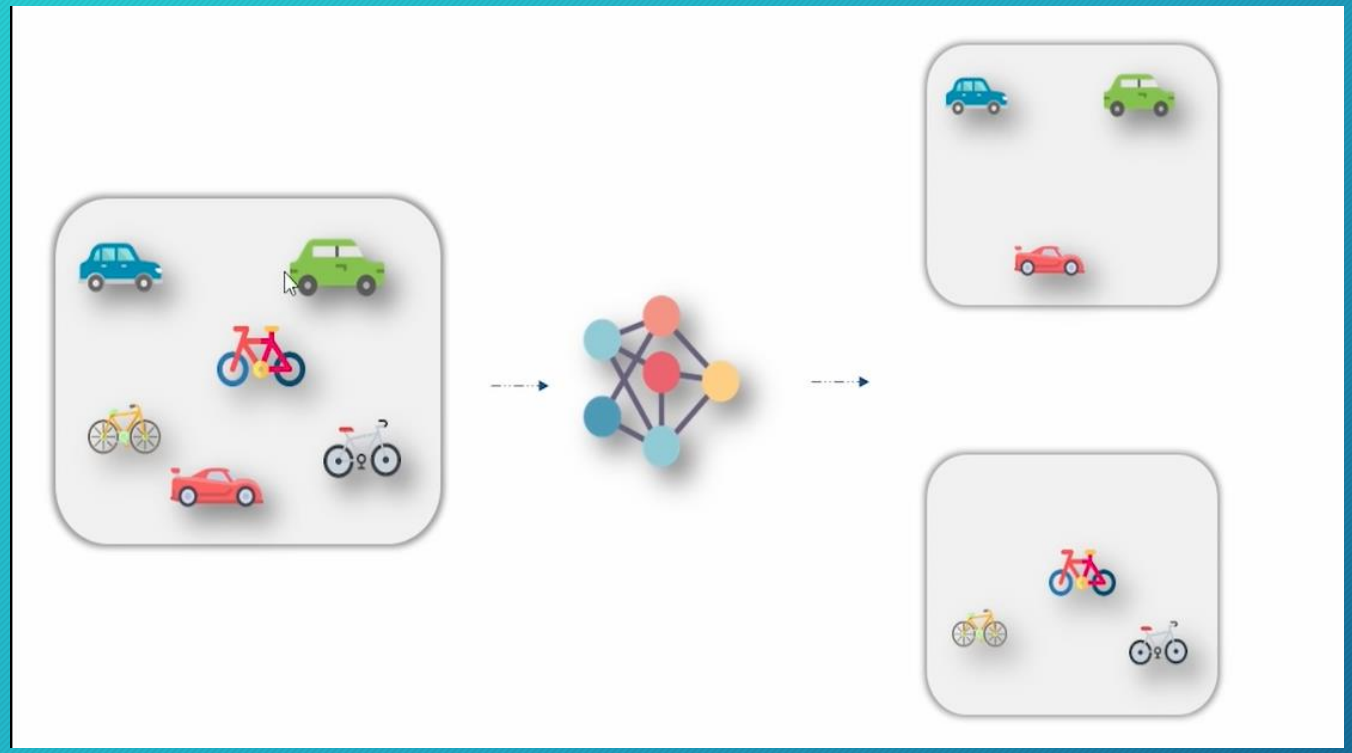
- Goal of the algorithm is to find a structure present in the data.
- Nothing apart from the input variables are present.

Unsupervised Machine Learning

Problem:



(Clustering):

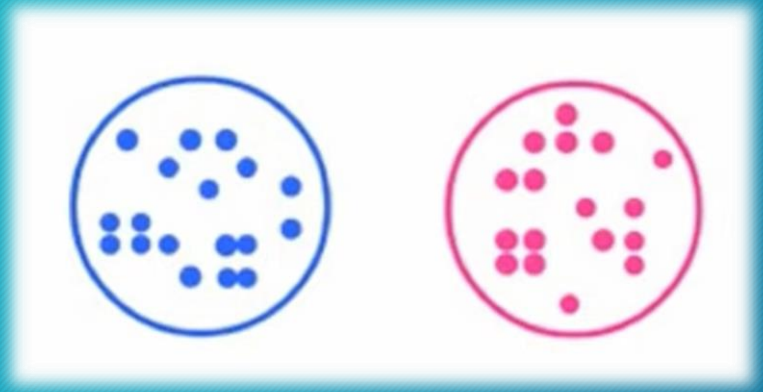


What is Clustering?

- “Clustering is the process of dividing the datasets into groups, consisting of similar data-points”
- Points in the same group are as similar as possible.
- Points in different group are as dissimilar as Possible.
- Example:
- Group of diners in a restaurant.
- Items arranged in a mall.

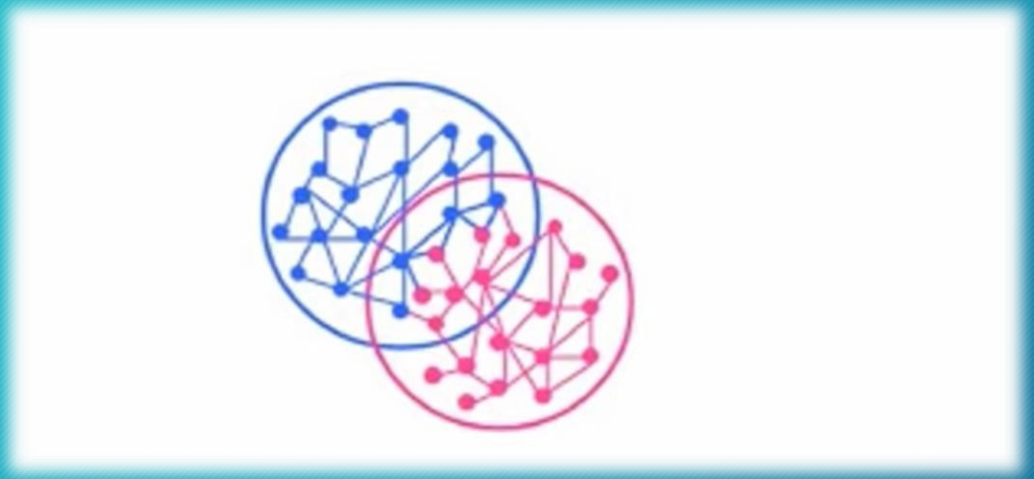
Types of Clustering?

- Exclusive Clustering
- Hard Clustering
- Data Points/Items belongs to exclusively one cluster
- For Example: K-means Clustering

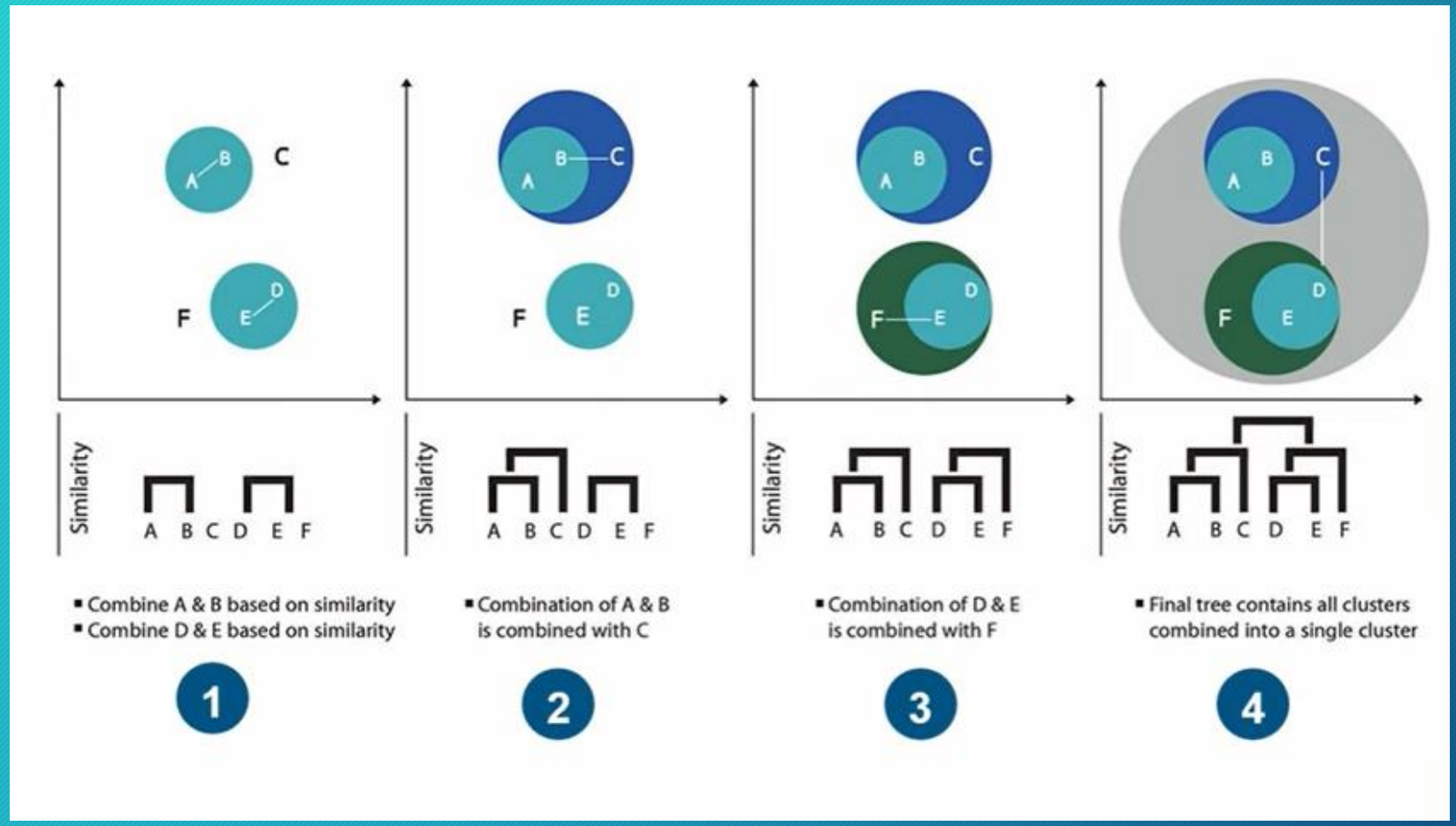


Overlapping Clustering

- Soft Clustering
- Data Point/Item belongs to multiple cluster
- For example : Fuzzy/C-means Clustering



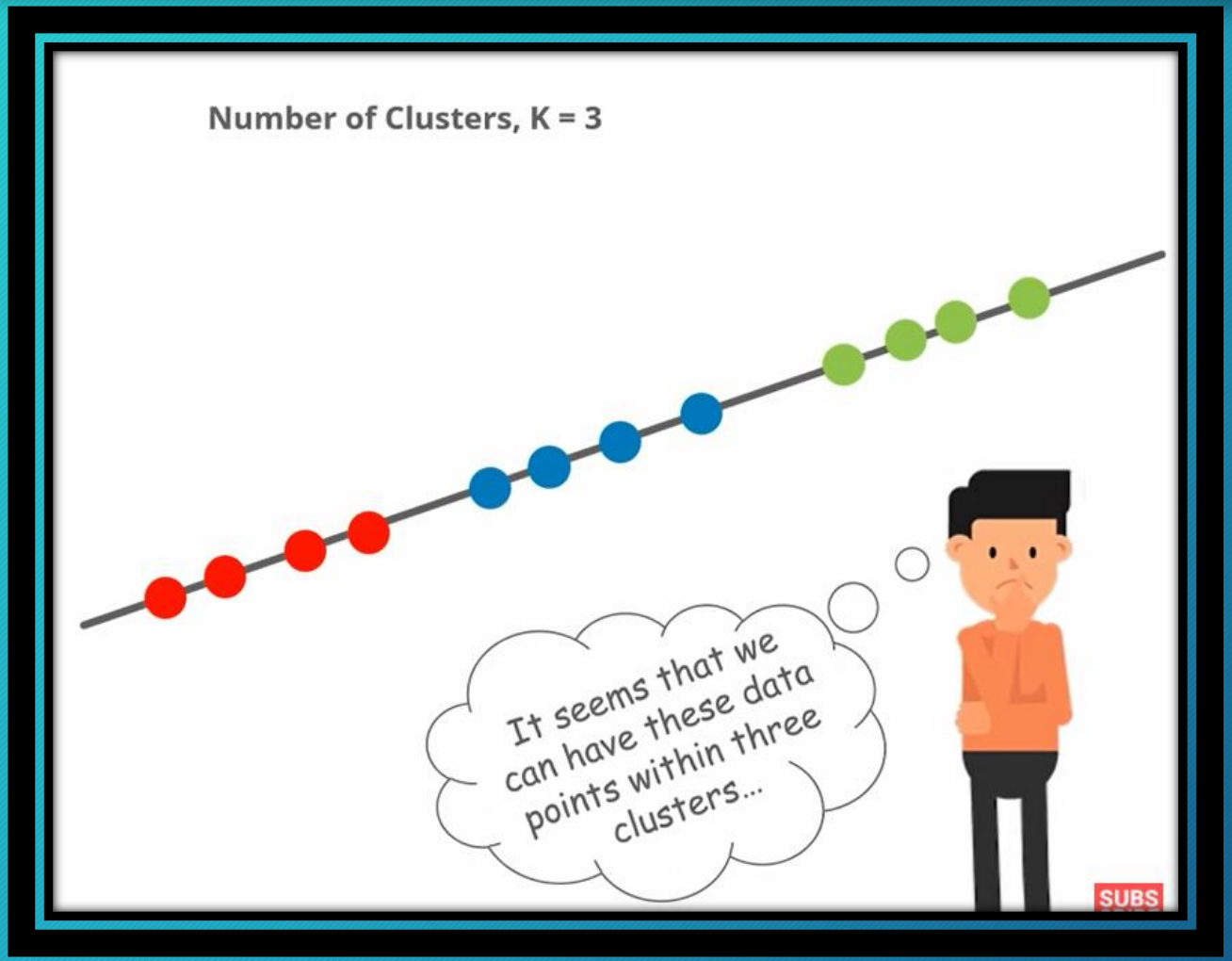
Hierarchical Clustering



What is K-mean Clustering ?

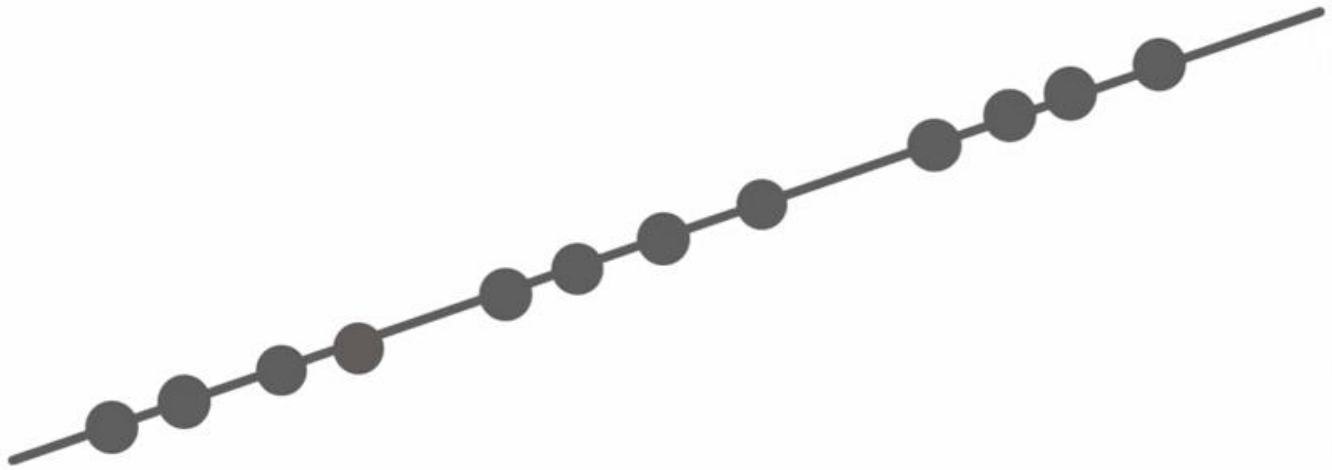
- “K-means Clustering algorithm whose main goal is to group similar elements or data points into a cluster.”
- NOTE: “K” in k-means represent the number of clusters.

Example :


Number of Clusters, K = 3
It seems that we can have these data points within three clusters...
SUBS

- **Step 1:** Select the number of clusters to be identified,

  i.e select a value for K =3 in this case

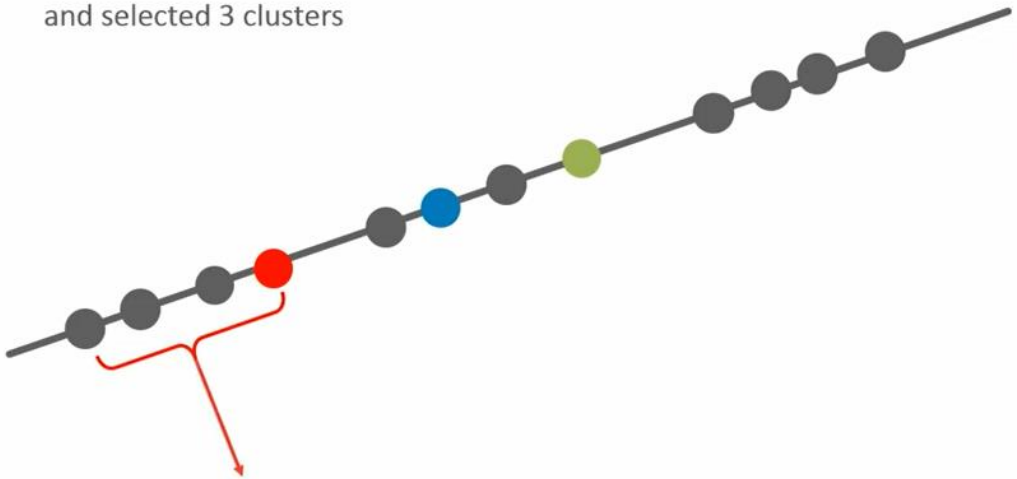


- **Step 1:** Select the number of clusters to be identified,

  i.e select a value for K =3 in this case
- **Step 2:** Randomly select 3 distinct data point

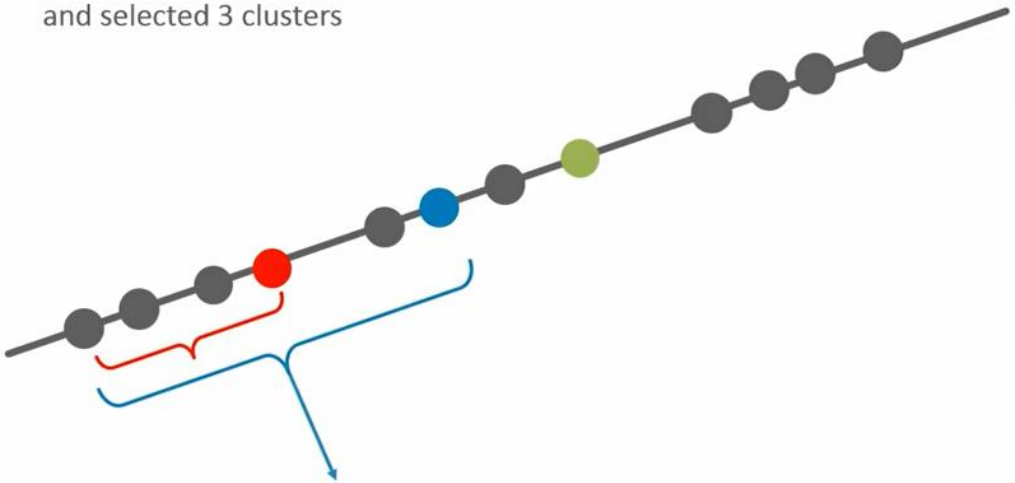
- **Step 1:** Select the number of clusters to be identified, i.e select a value for K =3 in this case
- **Step 2:** Randomly select 3 distinct data point
- **Step 3:** Measure the distance between the 1st point and selected 3 clusters

- **Step 1:** Select the number of clusters to be identified, i.e select a value for K =3 in this case
- **Step 2:** Randomly select 3 distinct data point
- **Step 3:** Measure the distance between the 1st point and selected 3 clusters



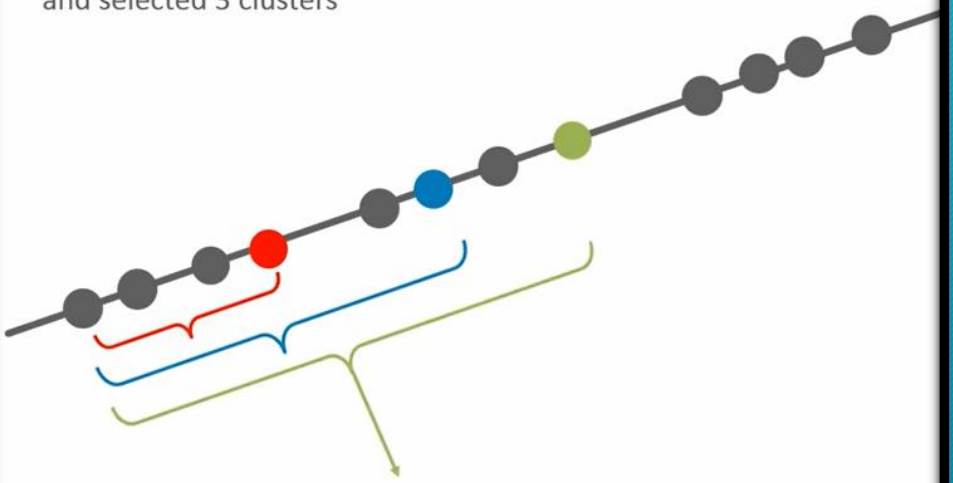
Distance from point 1 to the red cluster

- **Step 1:** Select the number of clusters to be identified, i.e select a value for K =3 in this case
- **Step 2:** Randomly select 3 distinct data point
- **Step 3:** Measure the distance between the 1st point and selected 3 clusters

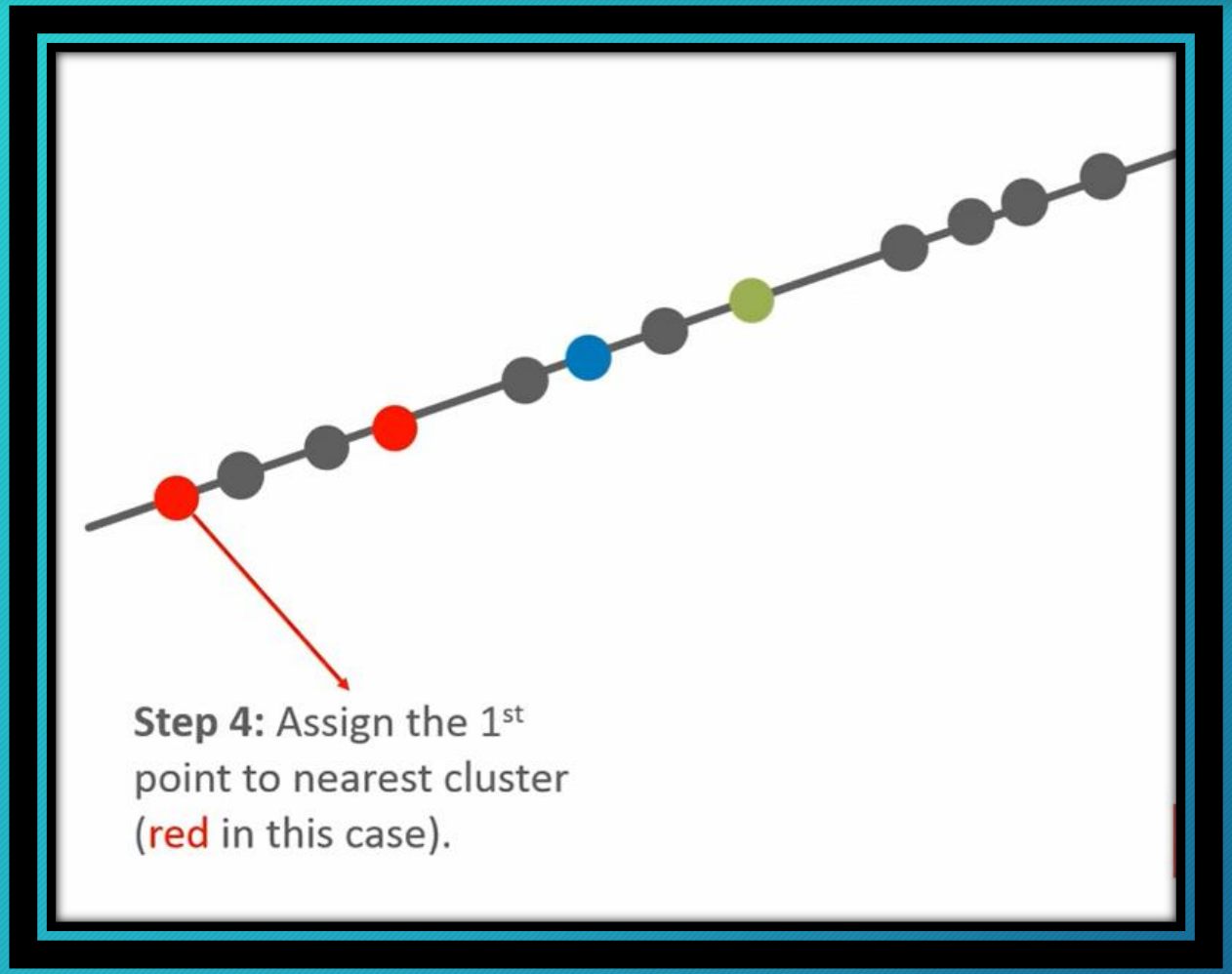


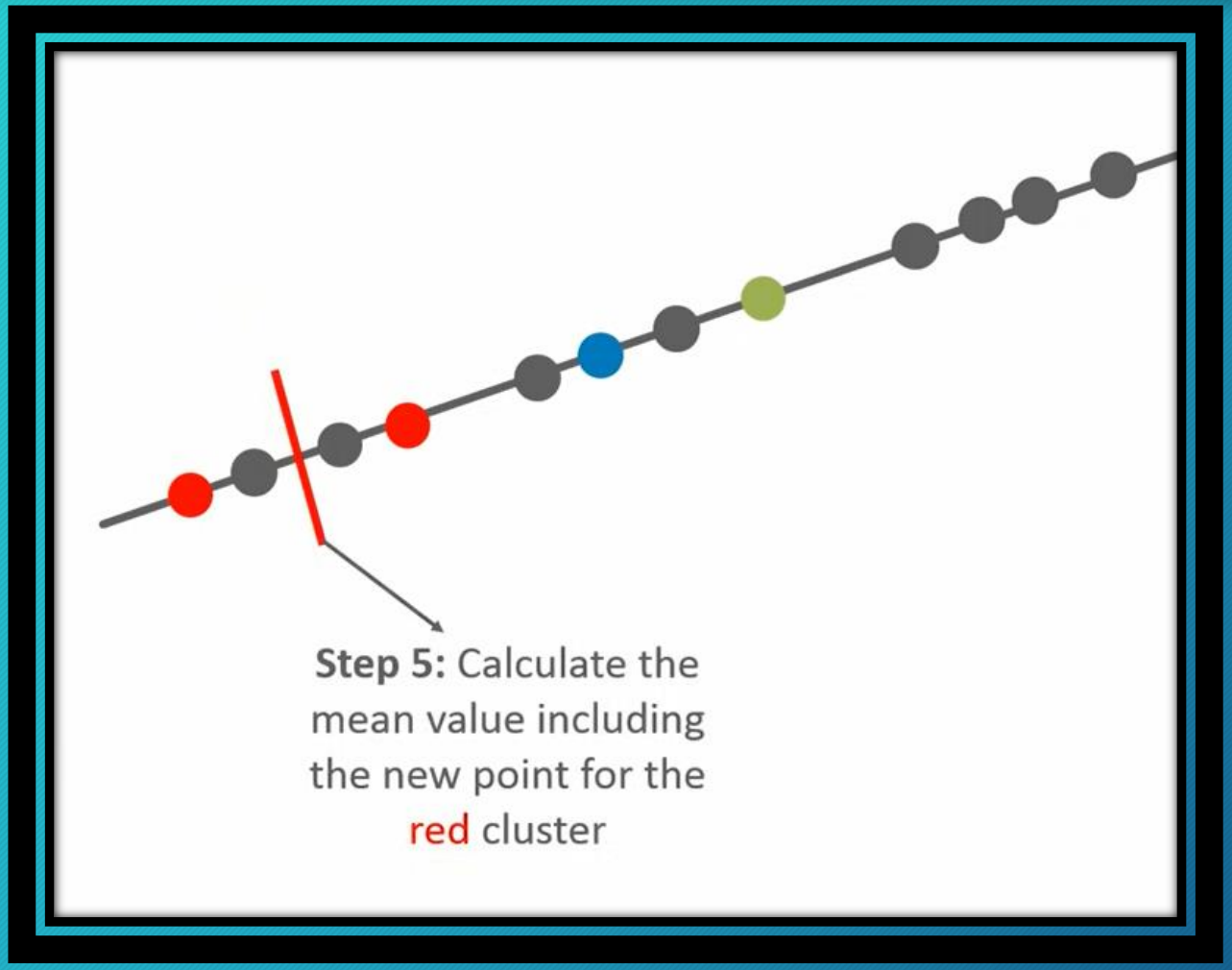
Distance from point 1 to the blue cluster

- **Step 1:** Select the number of clusters to be identified, i.e select a value for K =3 in this case
- **Step 2:** Randomly select 3 distinct data point
- **Step 3:** Measure the distance between the 1st point and selected 3 clusters

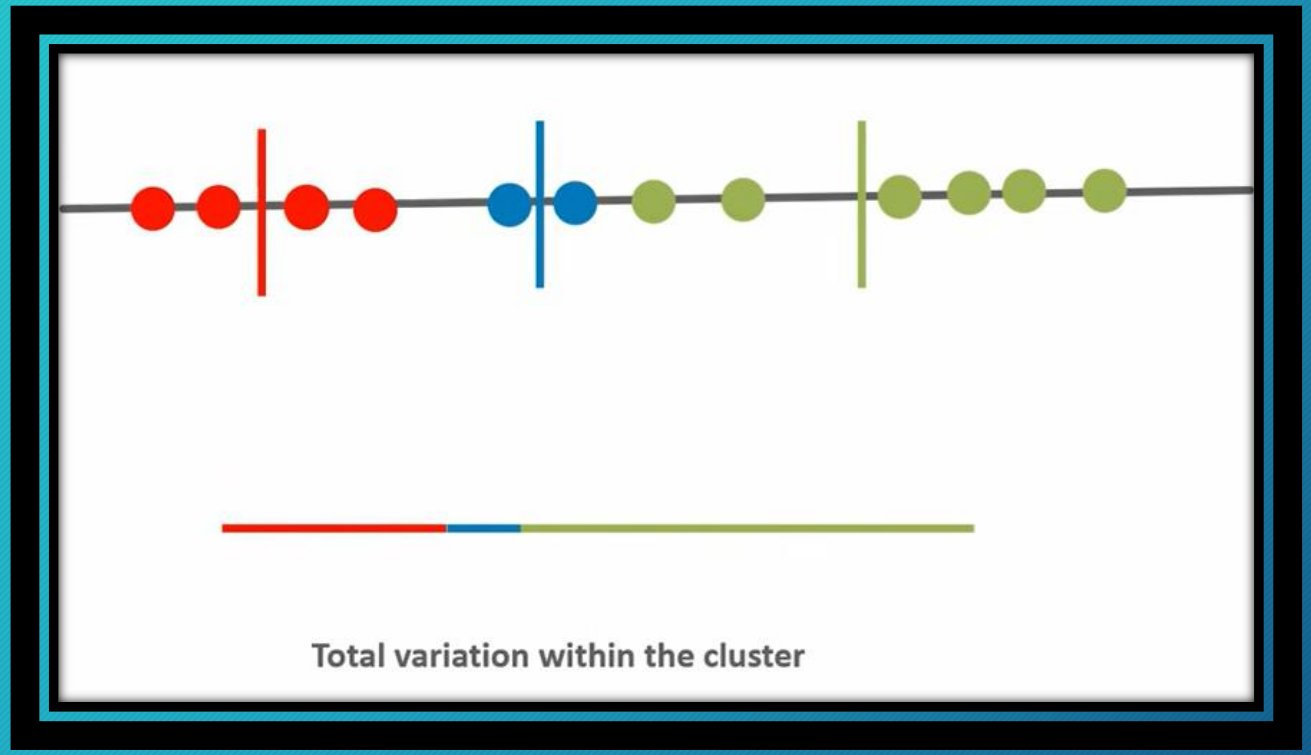


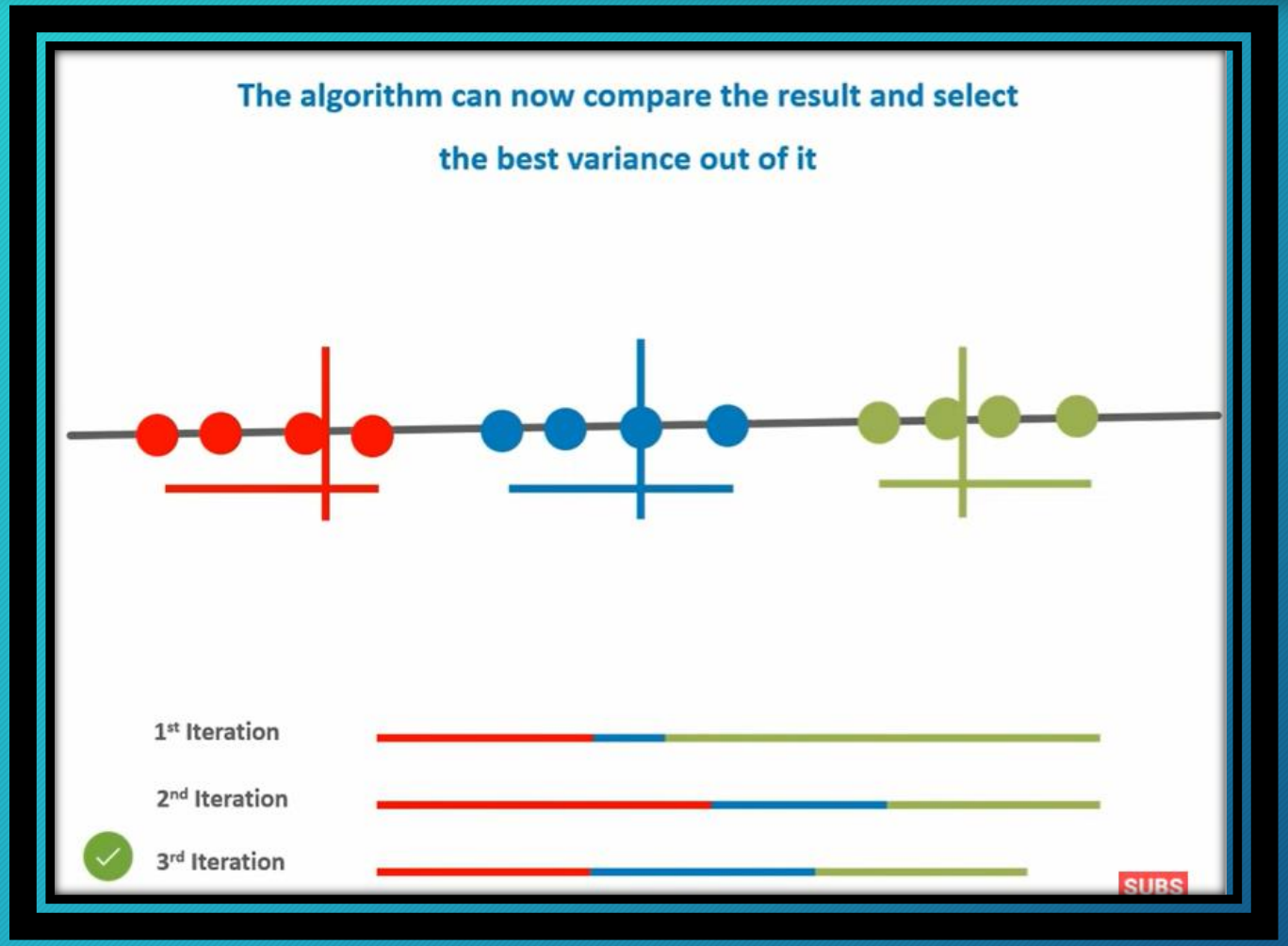
Distance from point 1 to the green cluster

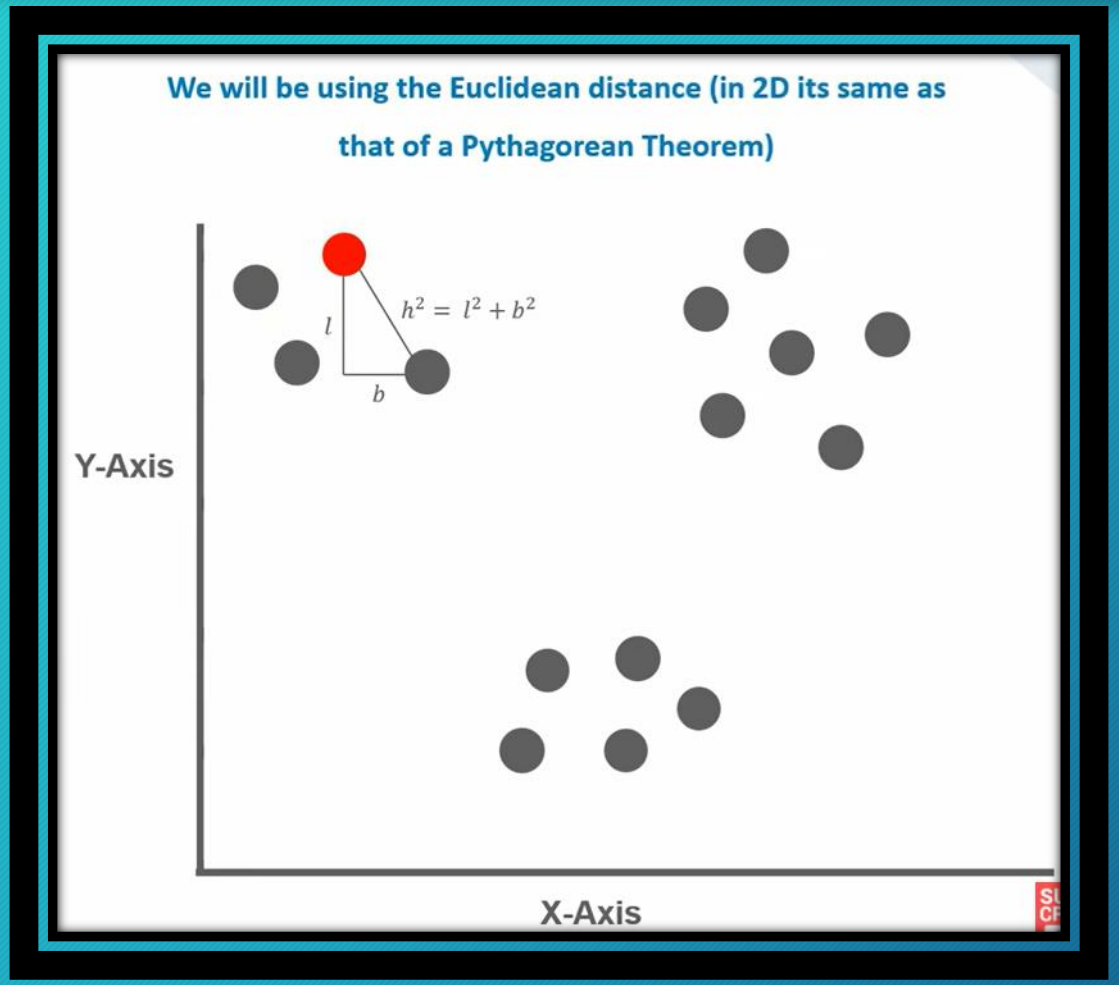

Step 4: Assign the 1st point to nearest cluster (red in this case).

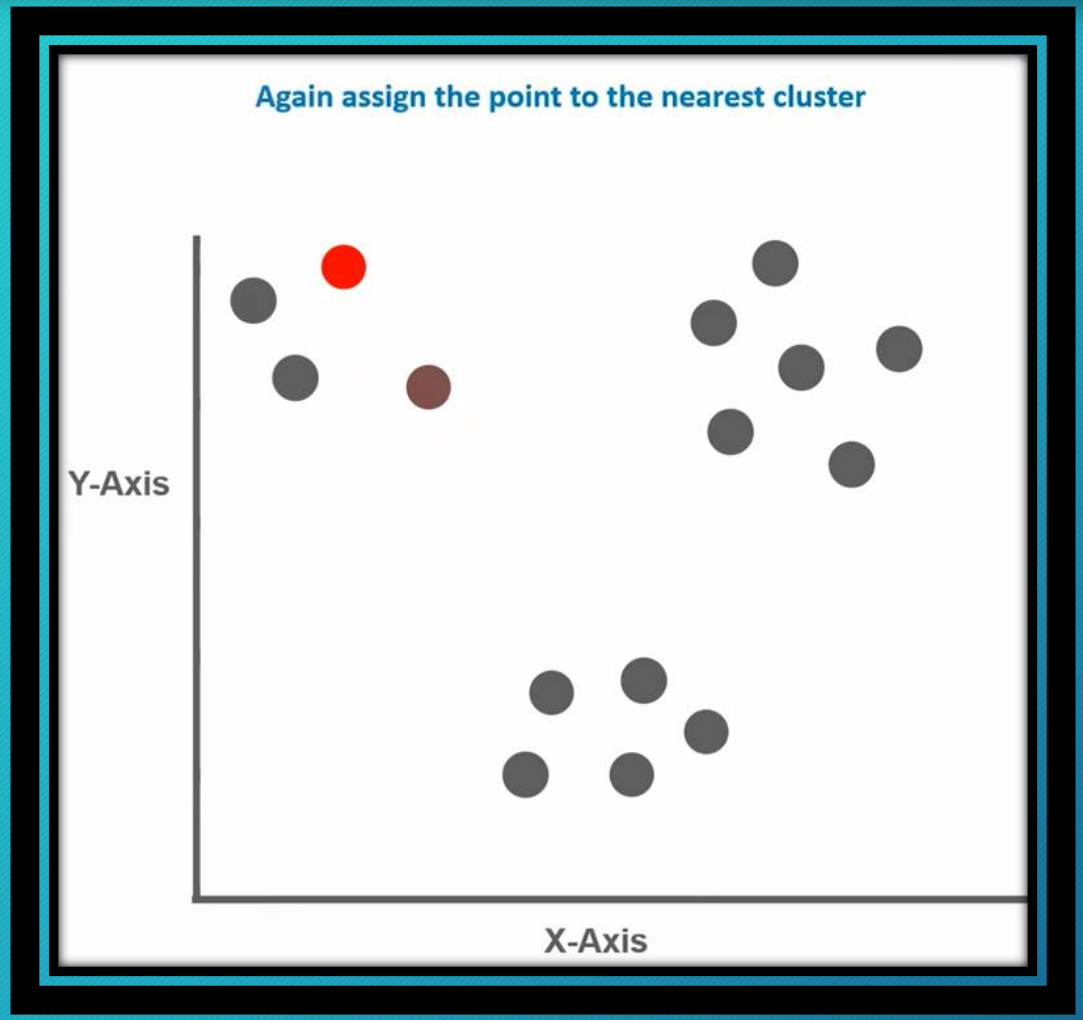

Step 5: Calculate the mean value including the new point for the red cluster

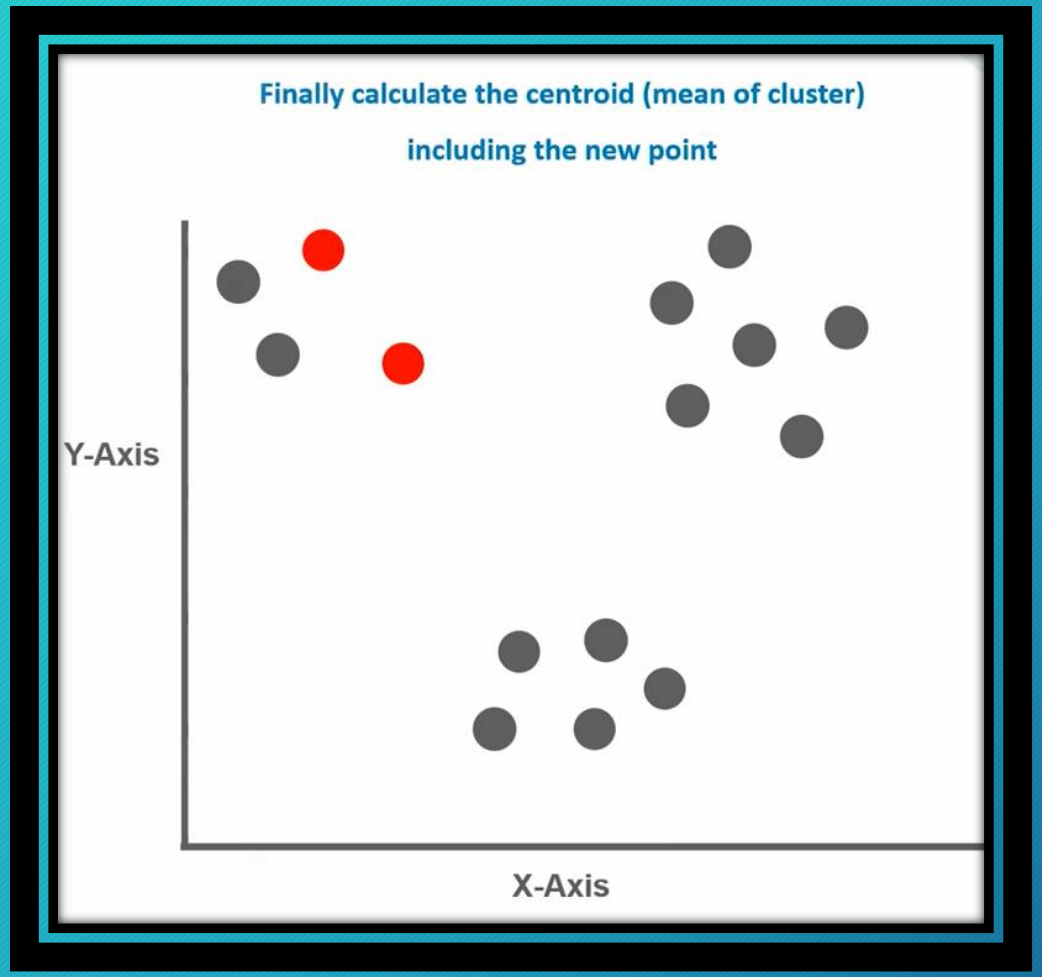
- According to the K-means algorithm it iterates over again and again.
- Unless and until the data points within each cluster stops changing.




The algorithm can now compare the result and select
the best variance out of it
1st Iteration
2nd Iteration
3rd Iteration
SUBS


We will be using the Euclidean distance (in 2D its same as
that of a Pythagorean Theorem)
$h^2 = l^2 + b^2$
l
b
Y-Axis
X-Axis


Again assign the point to the nearest cluster
Y-Axis
X-Axis


Finally calculate the centroid (mean of cluster)
including the new point
Y-Axis
X-Axis

Reinforcement Machine Learning:

- Trained based on the feedback from learning.
- Algorithms work towards the rewards.