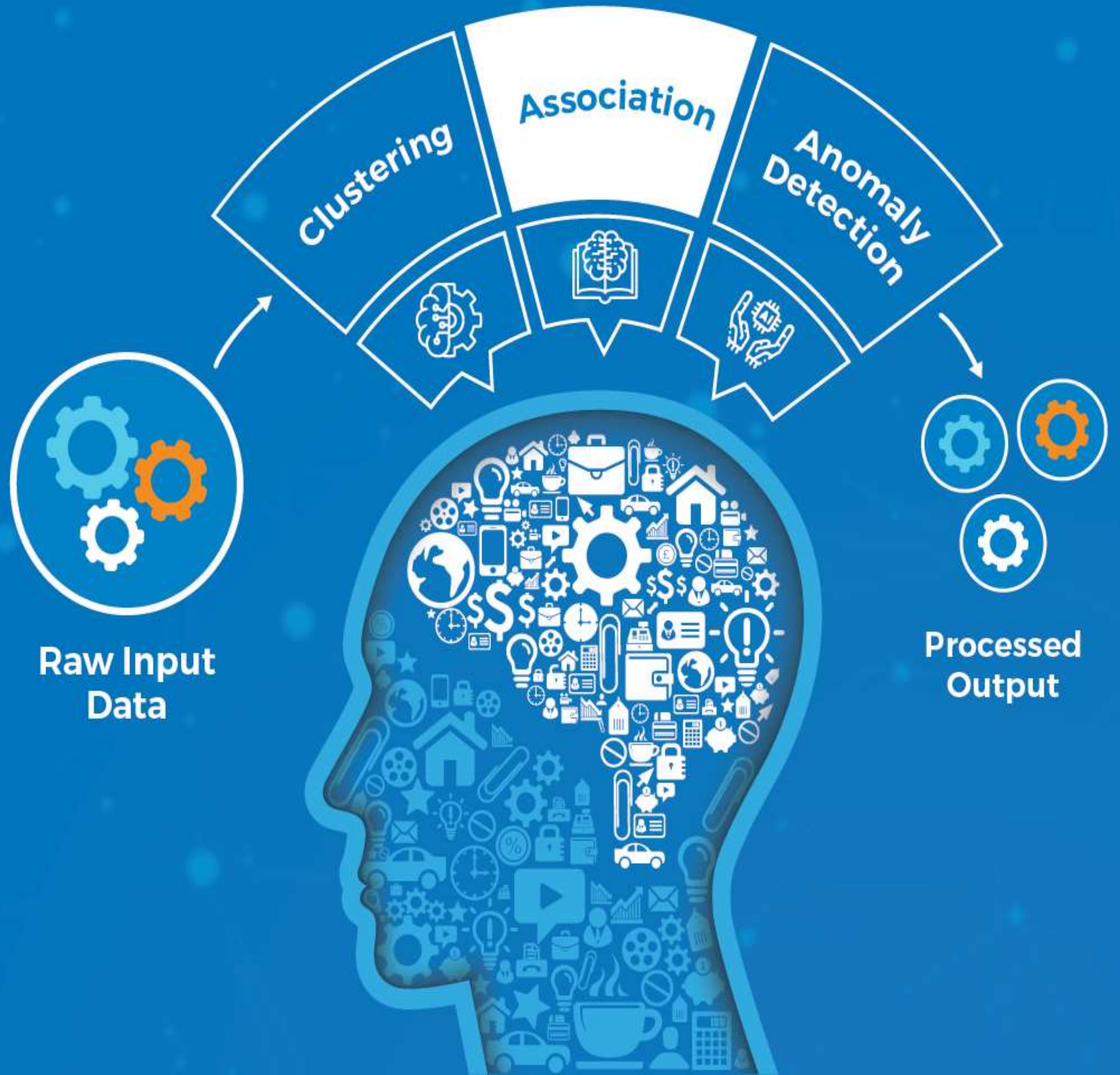
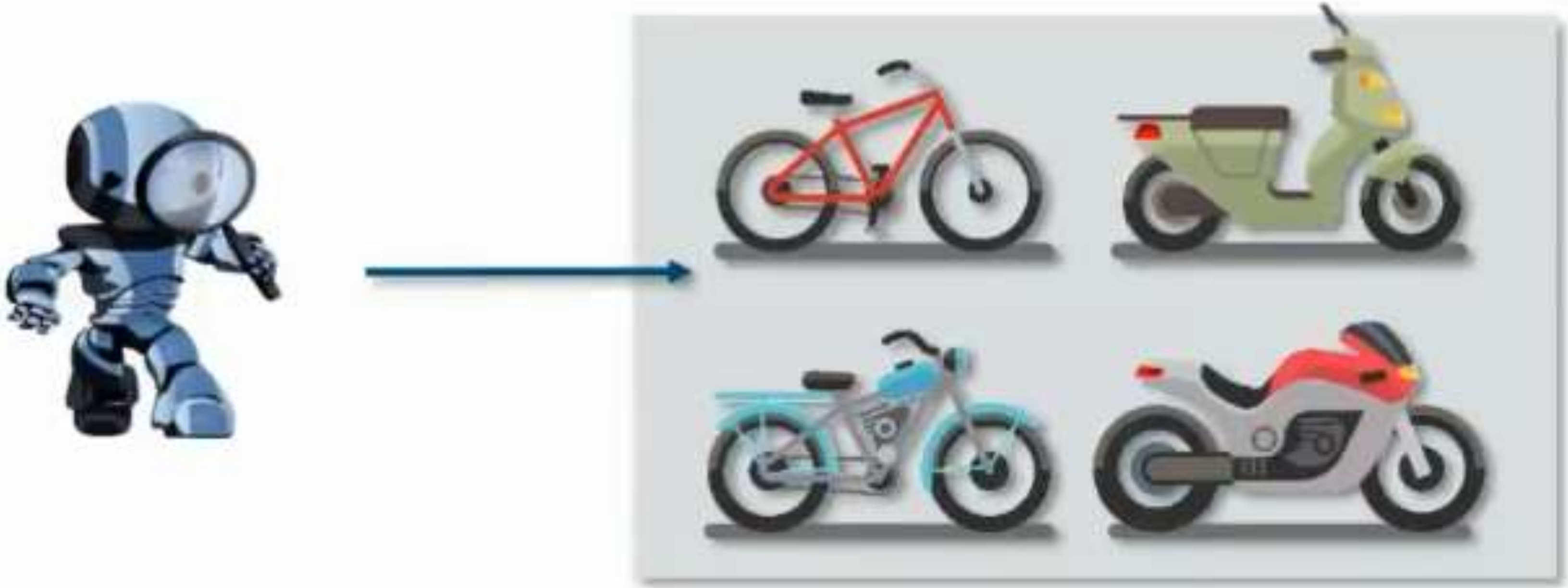
Unsupervised
Machine Learning



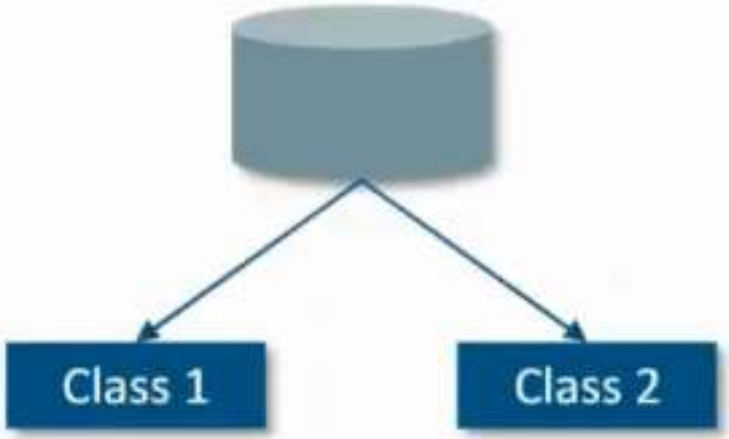
Unsupervised Learning

- Sometimes the given data is unstructured and unlabeled. So it becomes difficult to classify that data in different categories
- Unsupervised learning helps to solve this problem. This learning is used to cluster the input data in classes on the basis of their statistical properties
- Example: We can cluster different bikes based upon their speed limit, acceleration, average



What is Clustering?

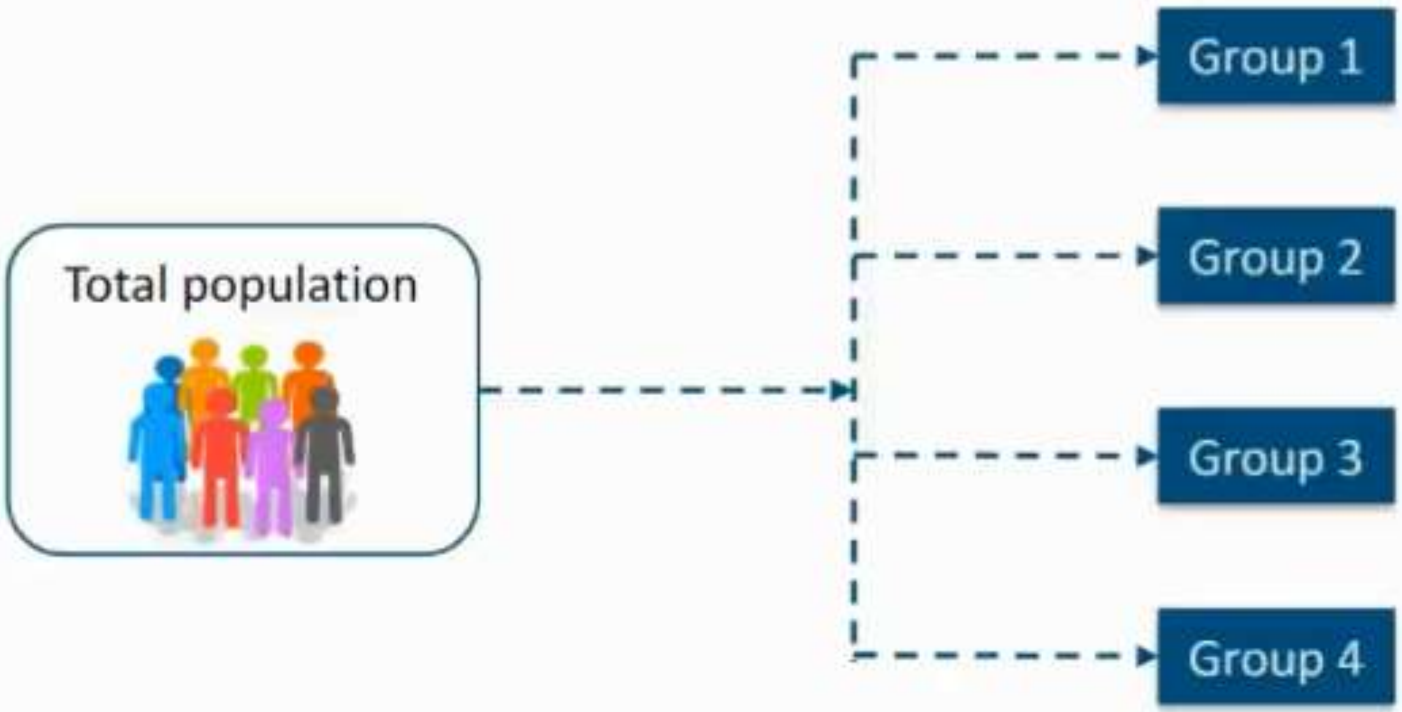
Clustering means grouping of objects based on the information found in the data, describing the objects or their relationship



The goal is that objects in one group will be similar to one other and different from objects in another group

Clustering

- The objects in group 1 should be as similar as possible
- But there should be much difference between an object in group 1 and group 2
- The attributes of the objects are allowed to determine which objects should be grouped together

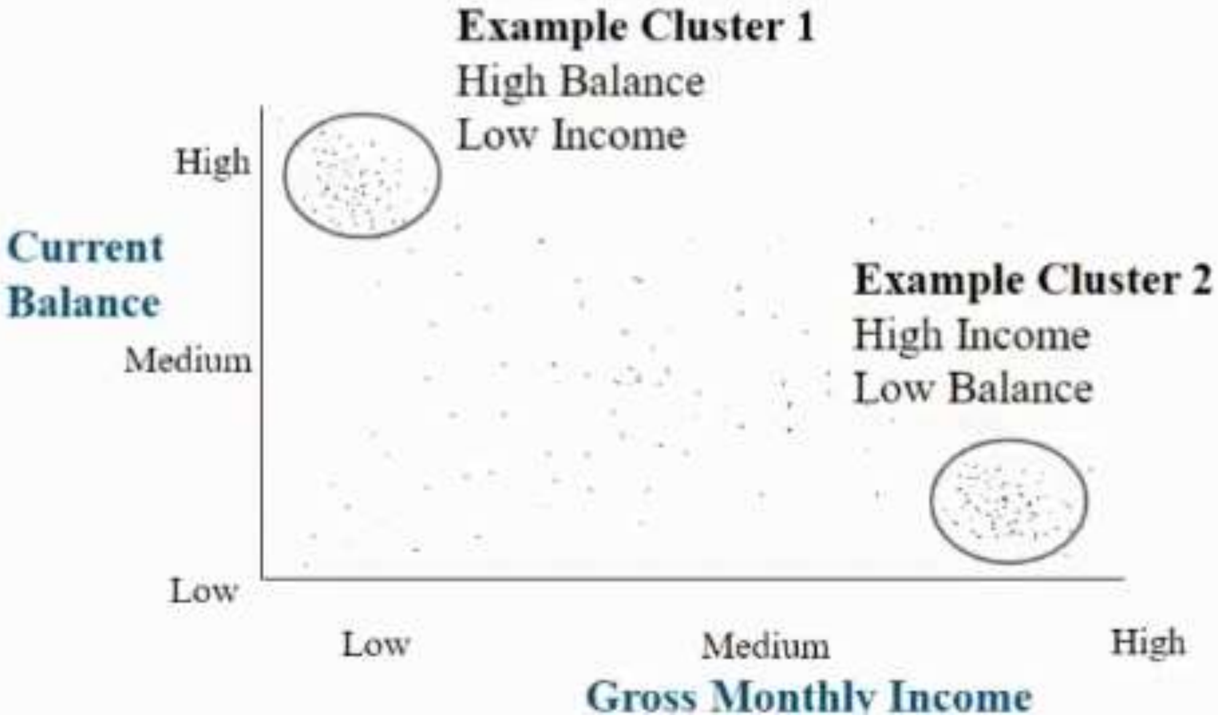


Clustering

Basic concepts of Cluster Analysis using two variables

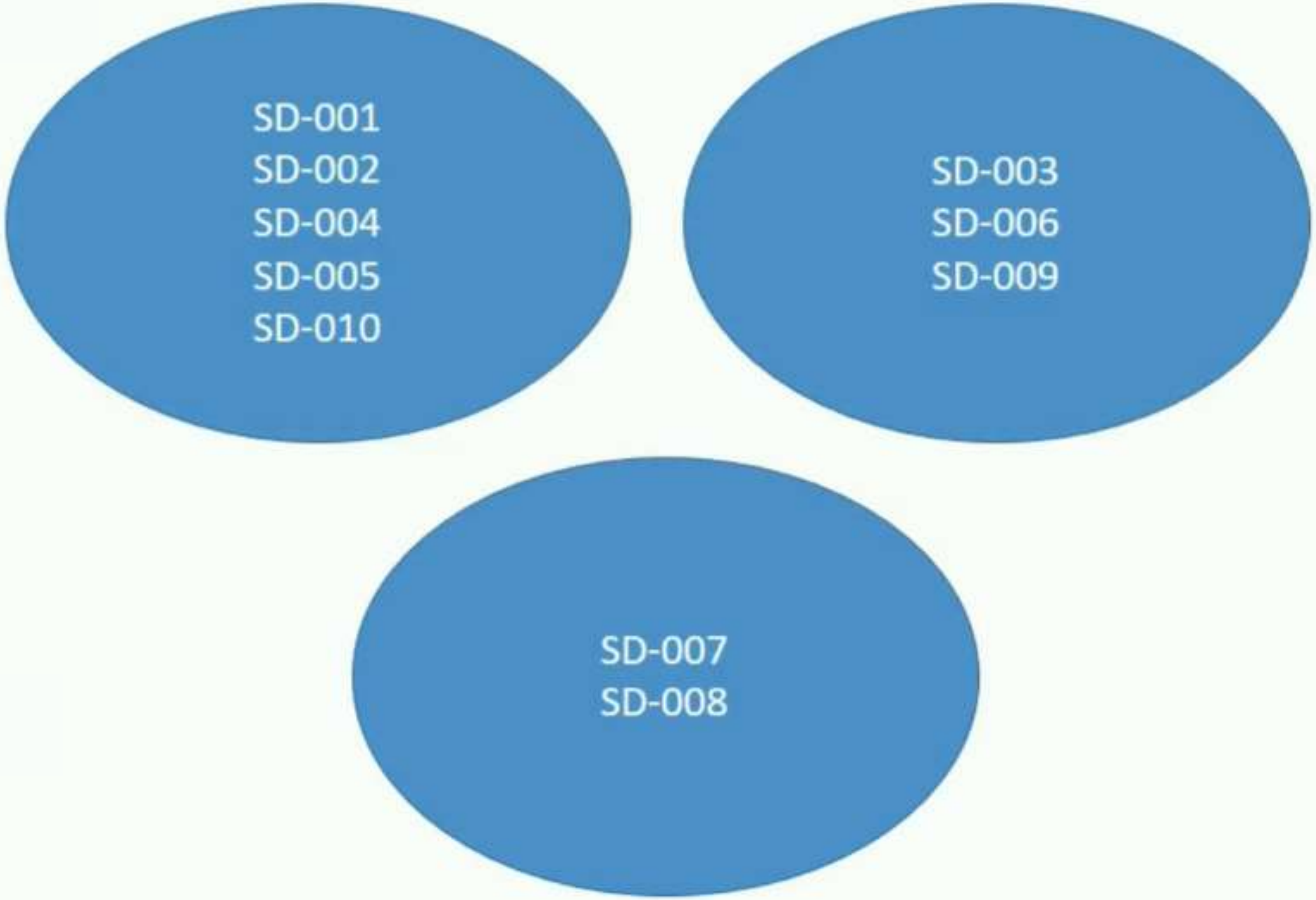
Cluster 1 and Cluster 2 are being differentiated by Income and Current Balance

- The objects in Cluster 1 have similar characteristics (High Income and Low balance)
- Also the objects in Cluster 2 have the same characteristic (High Balance and Low Income)
- But there are much differences between an object in Cluster 1 and an object in Cluster 2



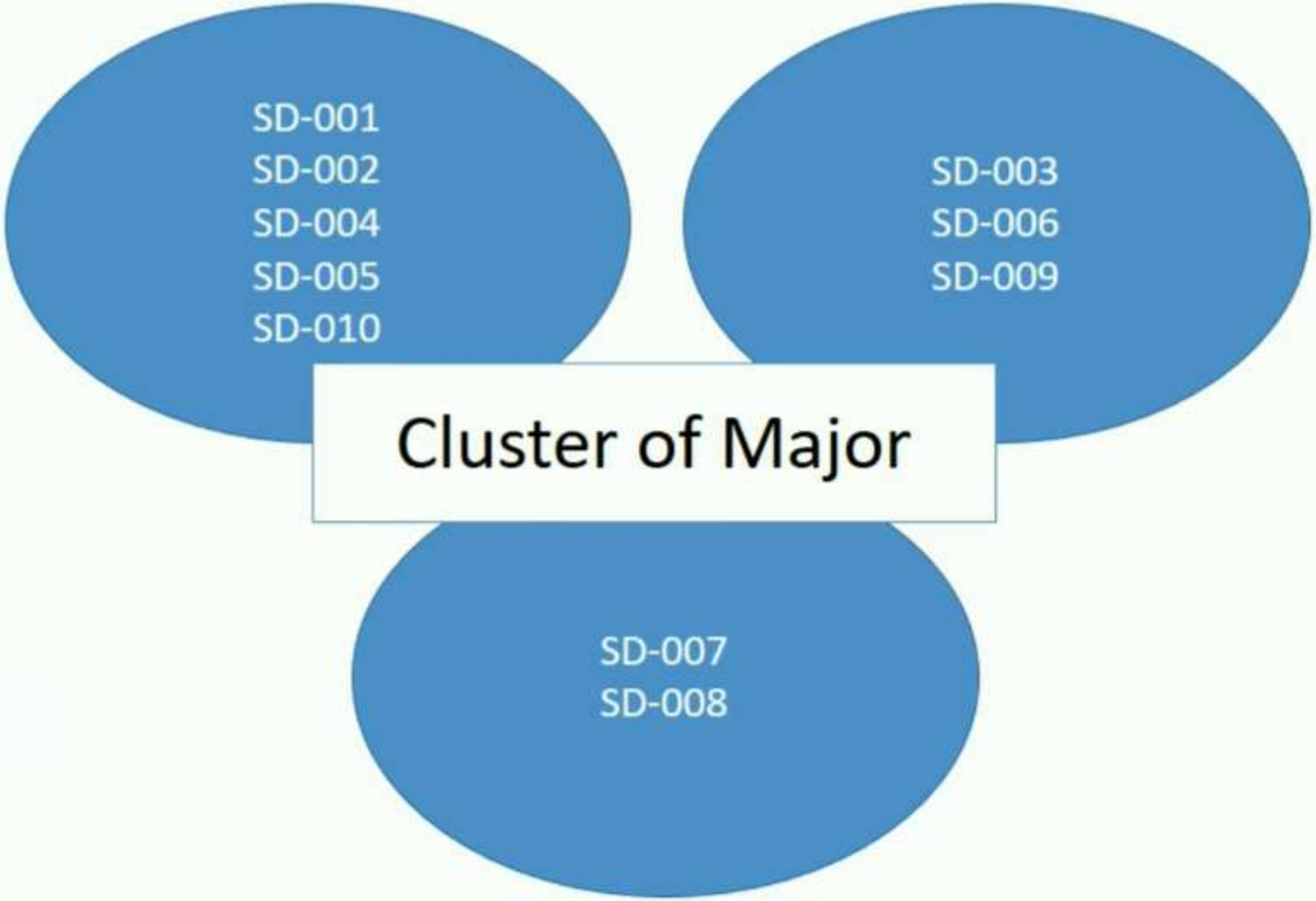
What is Clustering?

Student ID	Gender	Major	Grade
SD-001	M	Math	A+
SD-002	M	Math	A
SD-003	F	Statistics	A+
SD-004	F	Math	A
SD-005	F	Math	B
SD-006	M	Statistics	B
SD-007	F	Physics	A+
SD-008	F	Physics	A
SD-009	M	Statistics	B+



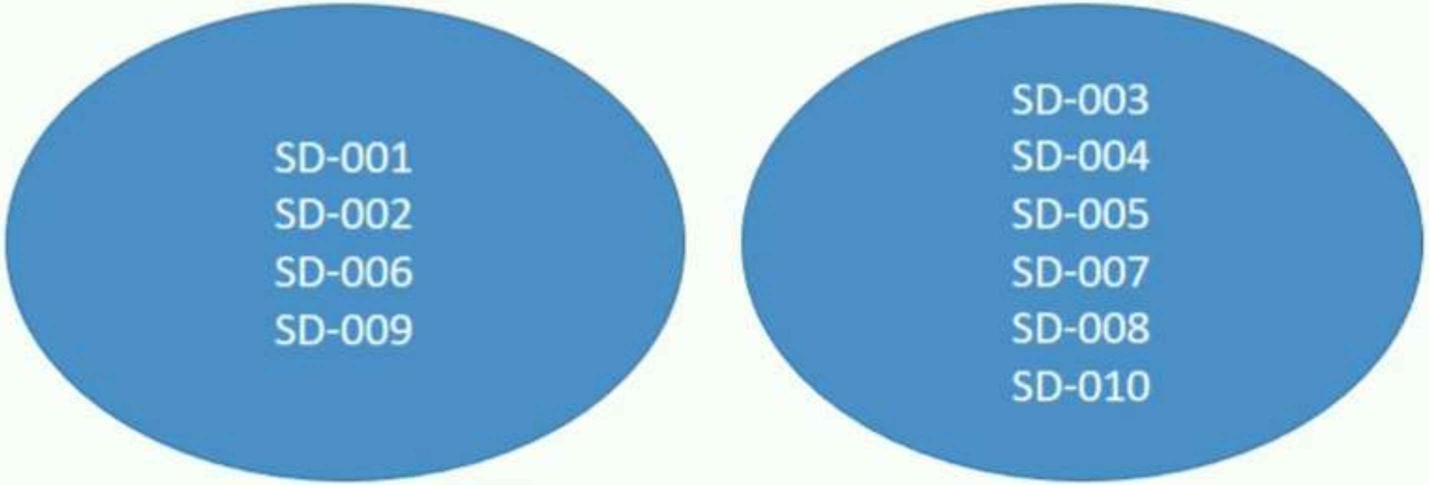
What is Clustering?

Student ID	Gender	Major	Grade
SD-001	M	Math	A+
SD-002	M	Math	A
SD-003	F	Statistics	A+
SD-004	F	Math	A
SD-005	F	Math	B
SD-006	M	Statistics	B
SD-007	F	Physics	A+
SD-008	F	Physics	A
SD-009	M	Statistics	B+



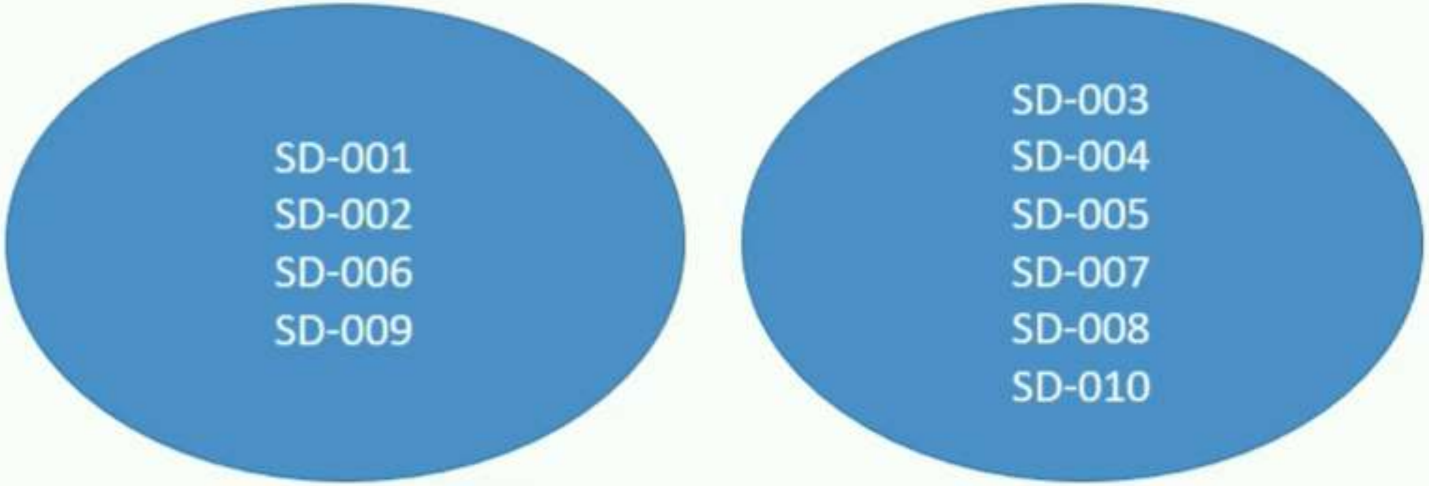
What is Clustering?

Student ID	Gender	Major	Grade
SD-001	M	Math	A+
SD-002	M	Math	A
SD-003	F	Statistics	A+
SD-004	F	Math	A
SD-005	F	Math	B
SD-006	M	Statistics	B
SD-007	F	Physics	A+
SD-008	F	Physics	A
SD-009	M	Statistics	B+



What is Clustering?

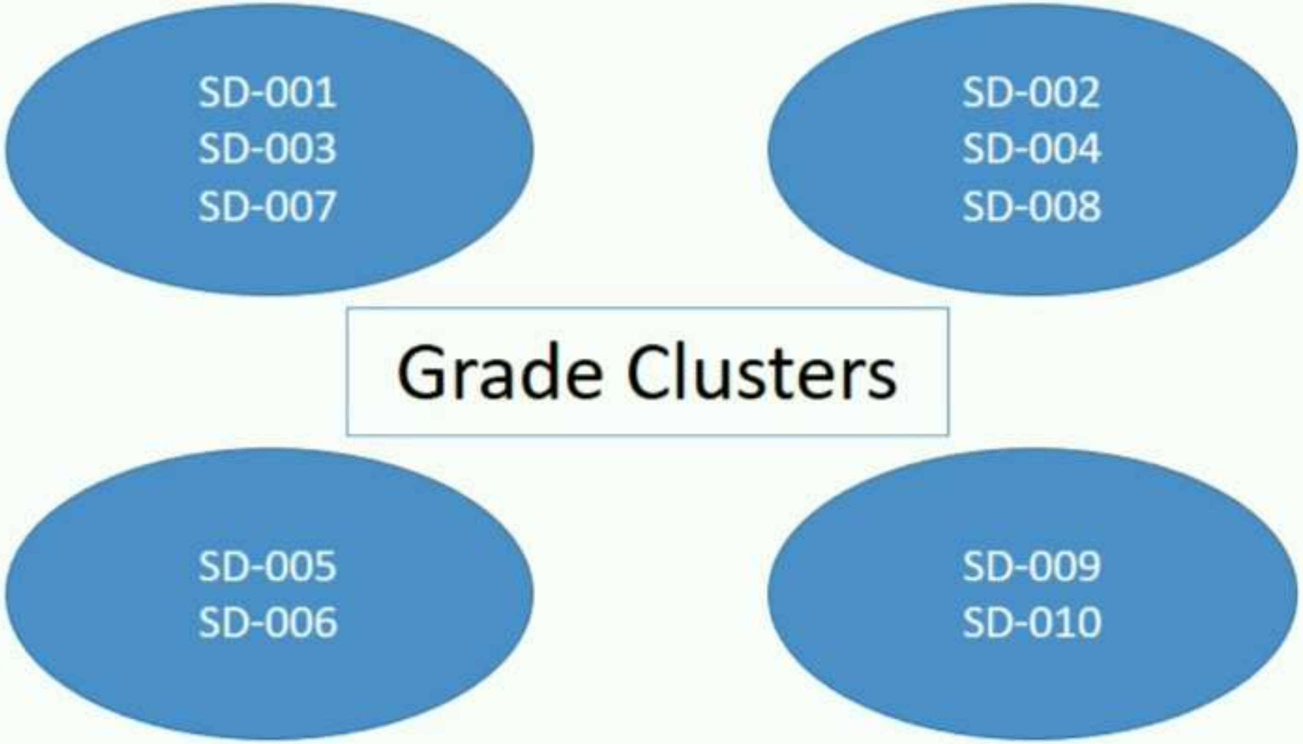
Student ID	Gender	Major	Grade
SD-001	M	Math	A+
SD-002	M	Math	A
SD-003	F	Statistics	A+
SD-004	F	Math	A
SD-005	F	Math	B
SD-006	M	Statistics	B
SD-007	F	Physics	A+
SD-008	F	Physics	A
SD-009	M	Statistics	B+



Gender Specific Clusters

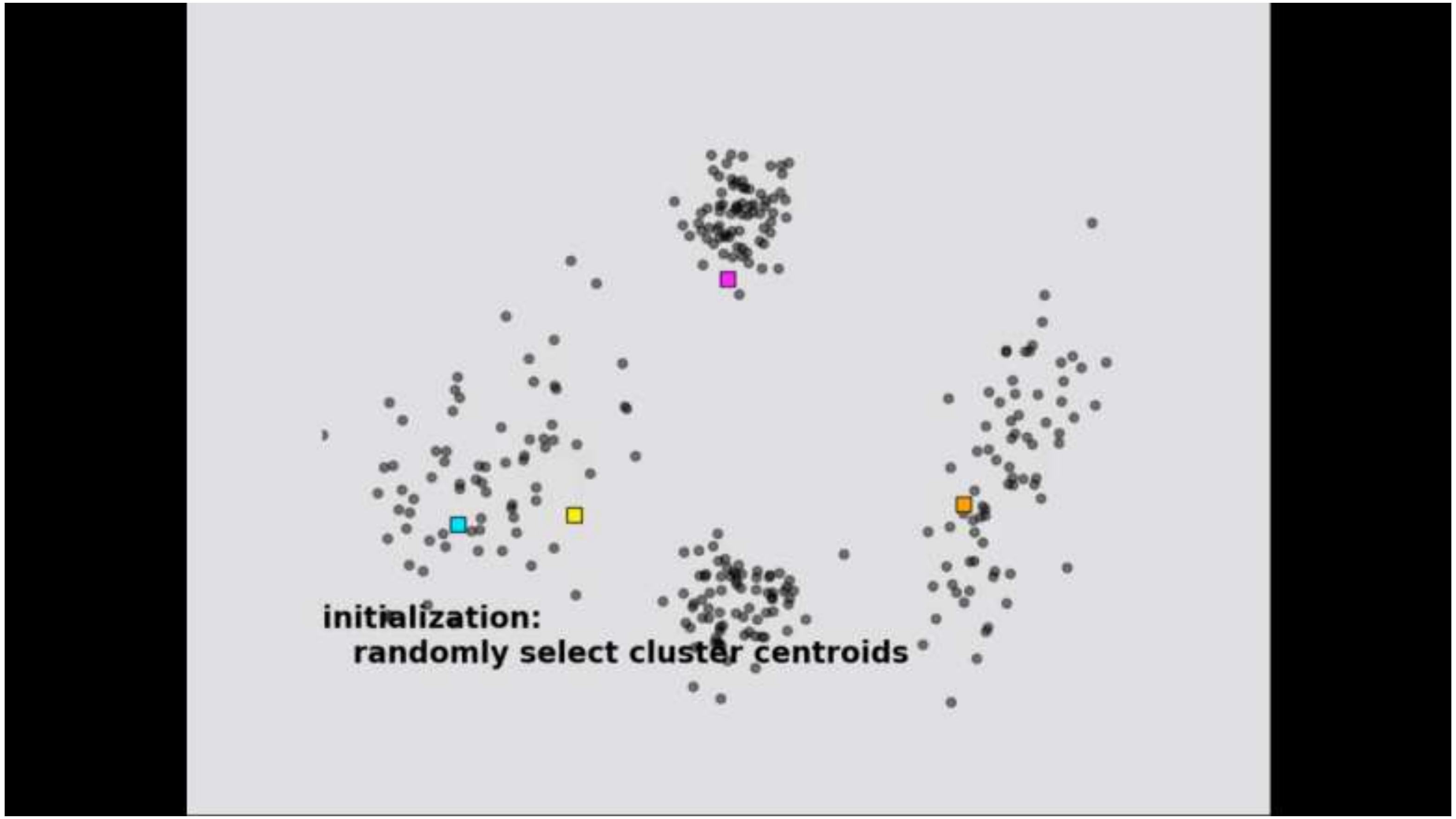
What is Clustering?

Student ID	Gender	Major	Grade
SD-001	M	Math	A+
SD-002	M	Math	A
SD-003	F	Statistics	A+
SD-004	F	Math	A
SD-005	F	Math	B
SD-006	M	Statistics	B
SD-007	F	Physics	A+
SD-008	F	Physics	A
SD-009	M	Statistics	B+



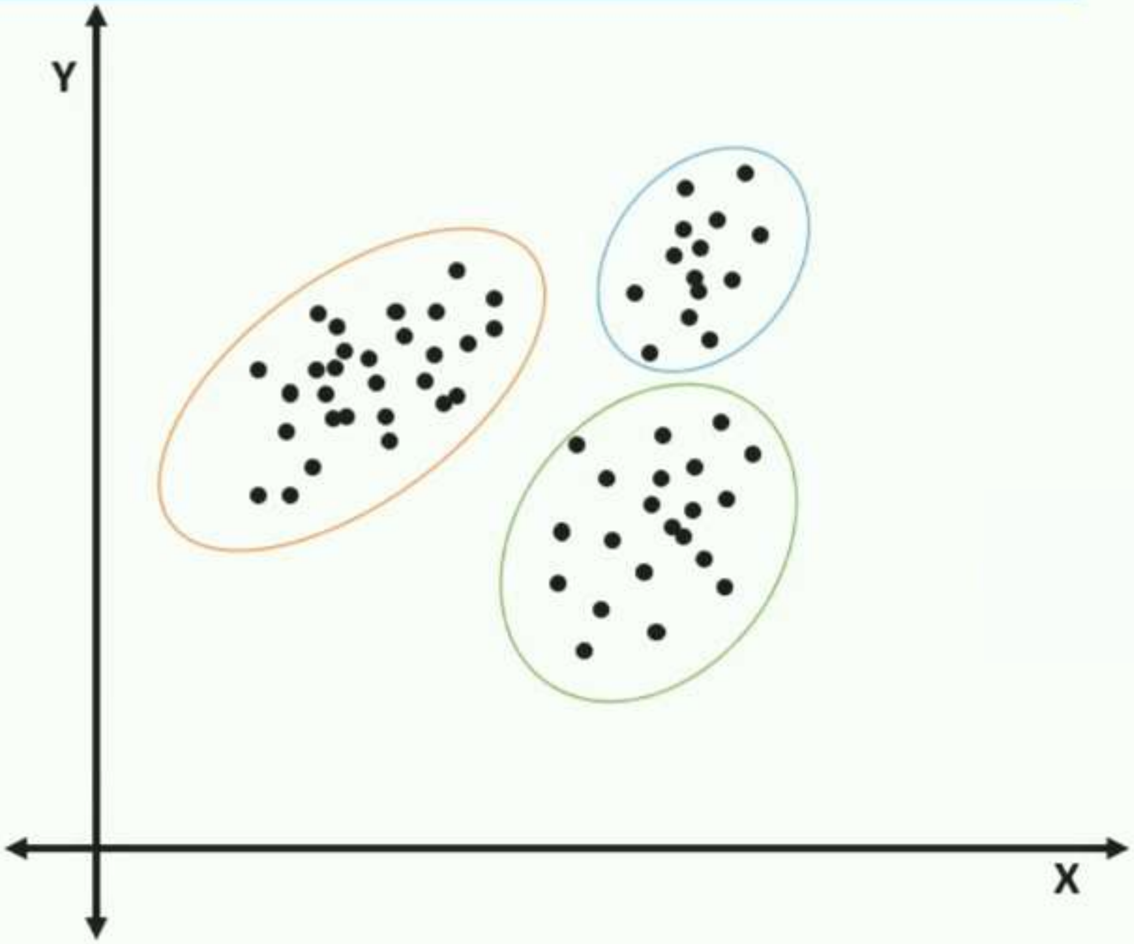
K-Means Clustering

https://www.youtube.com/watch?v=5I3Ei69I40s



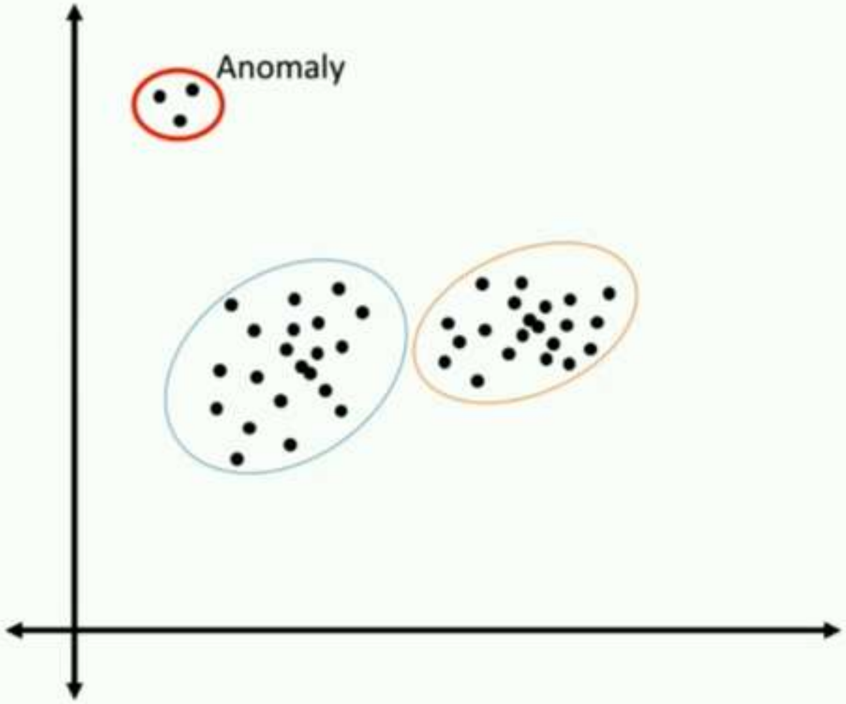
What is Clustering Analysis ?

- Clustering is the task of grouping a set of objects
- Unsupervised Learning model
- Discovering distinct groups in customer databases
- Used for creating strategies to adopt for certain segments



Examples of Clustering

- Recommendation engines
- Market segmentation
- Social network analysis
- Medical/Health
- Image segmentation
- Anomaly detection

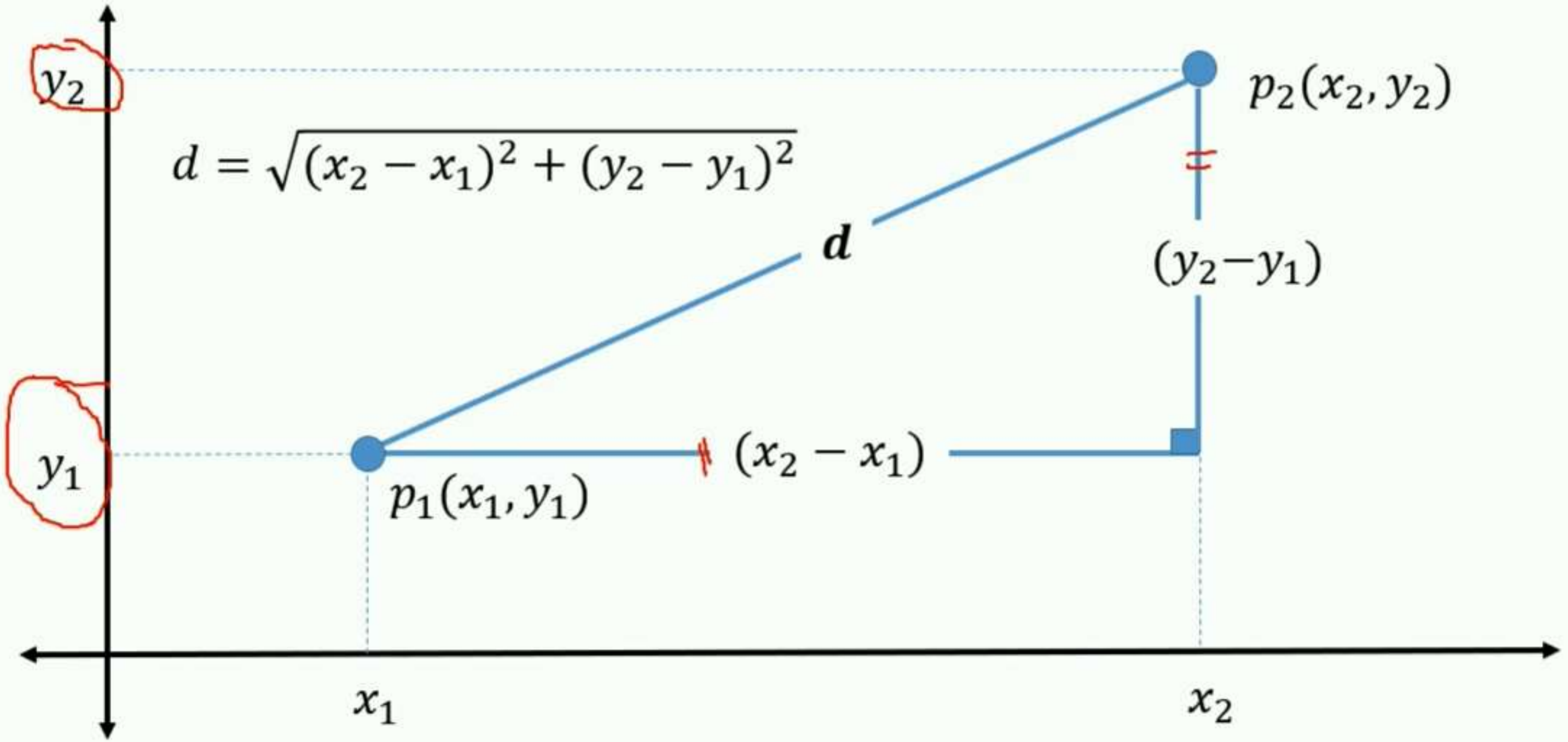


Euclidean Distance

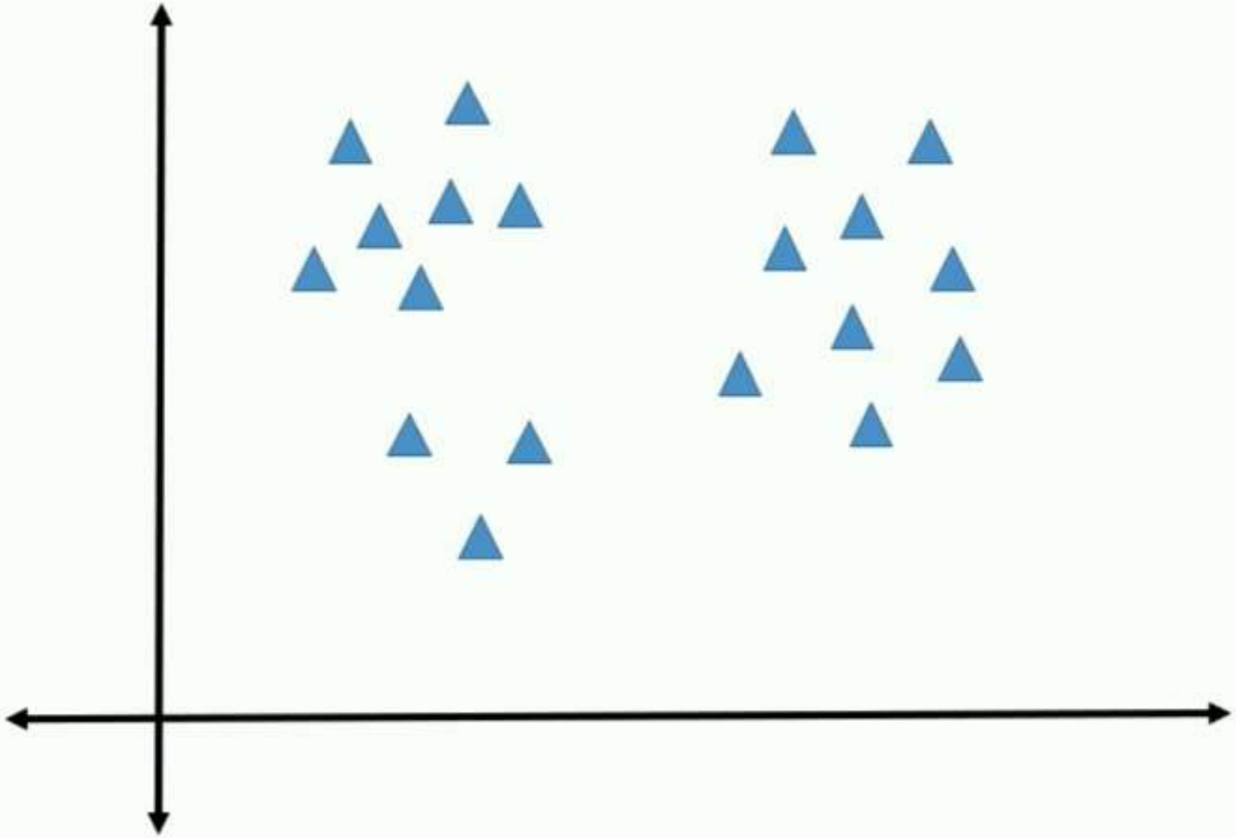
dist(.) is the Euclidean distance. the **Euclidean distance** or **Euclidean metric** is the "ordinary" straight-line distance between two points in Euclidean space

$$dist(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ip} - x_{jp})^2}$$

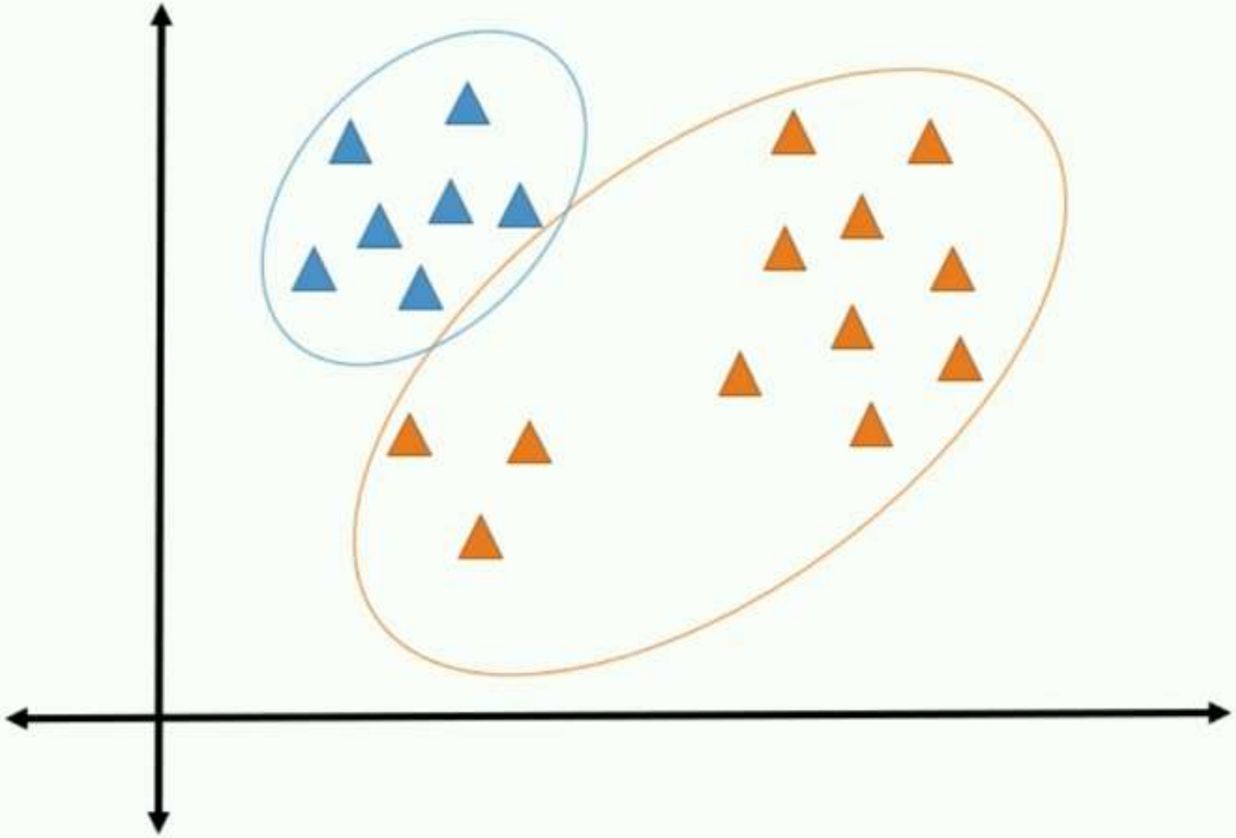
Euclidean Distance



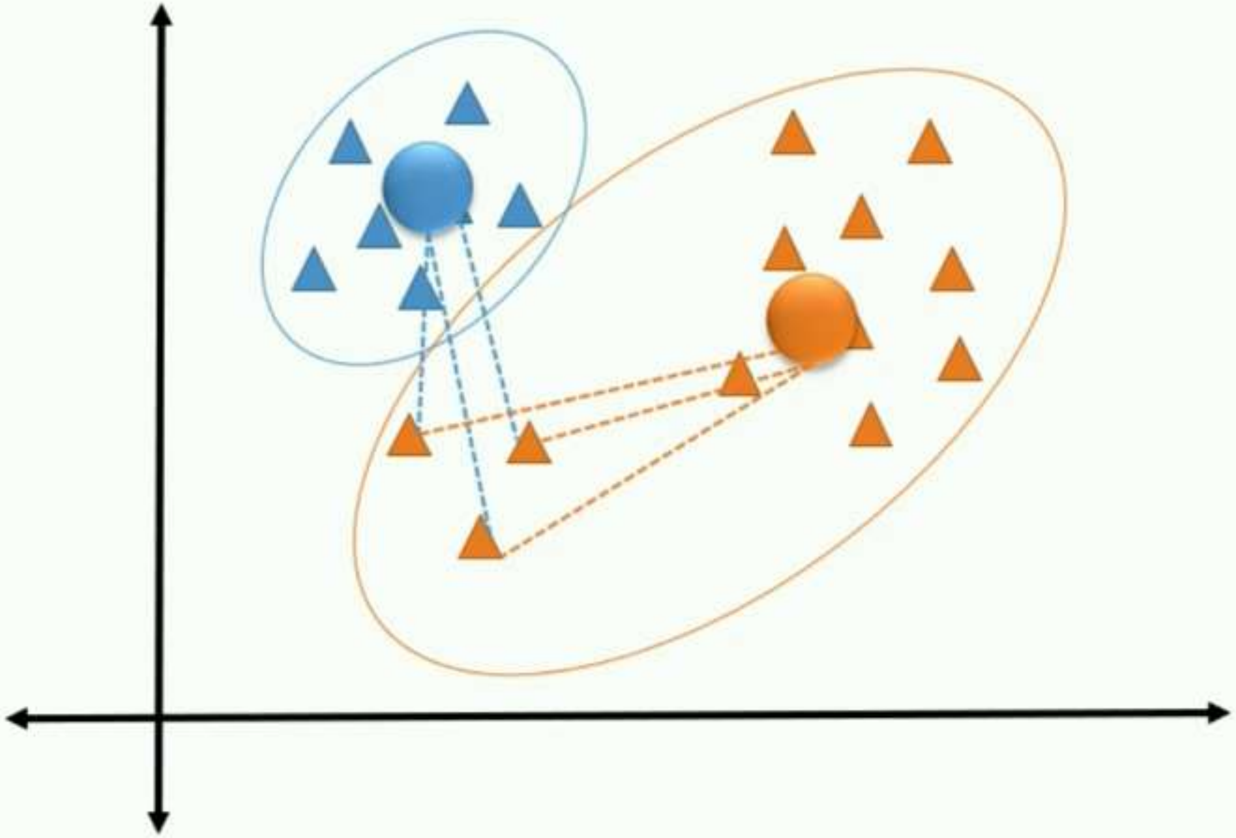
How Clusters are formed?



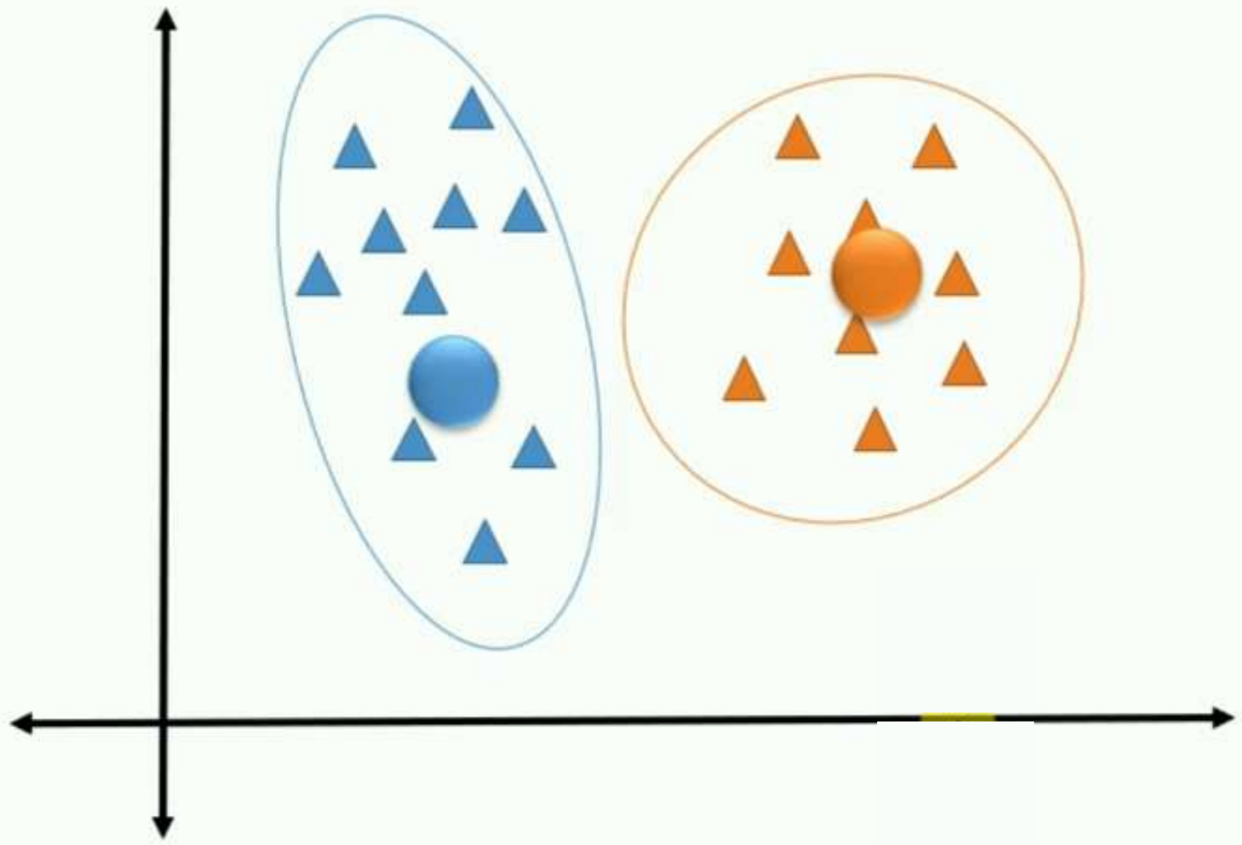
How Clusters are formed?



How Clusters are formed?



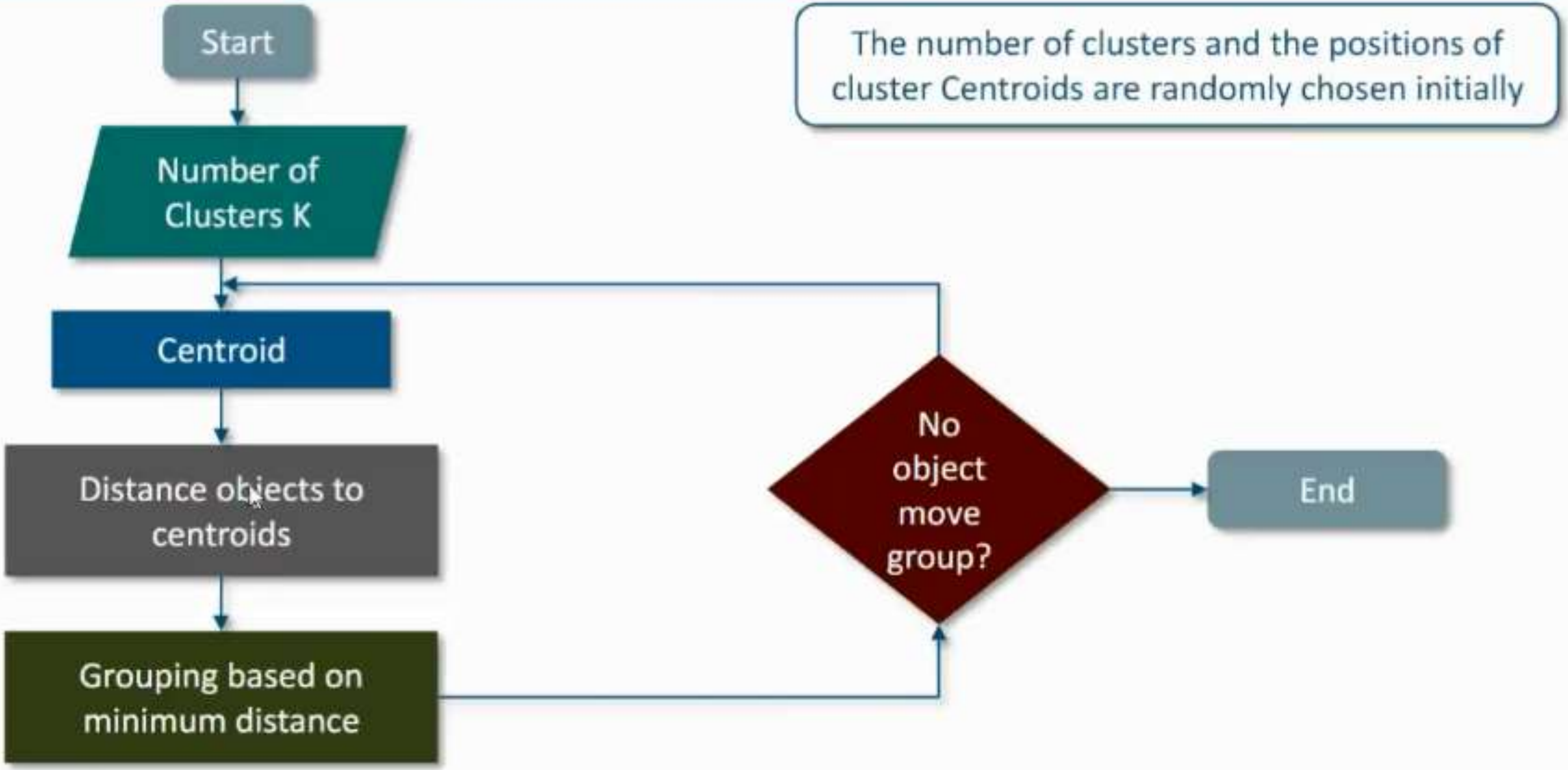
How Clusters are formed?



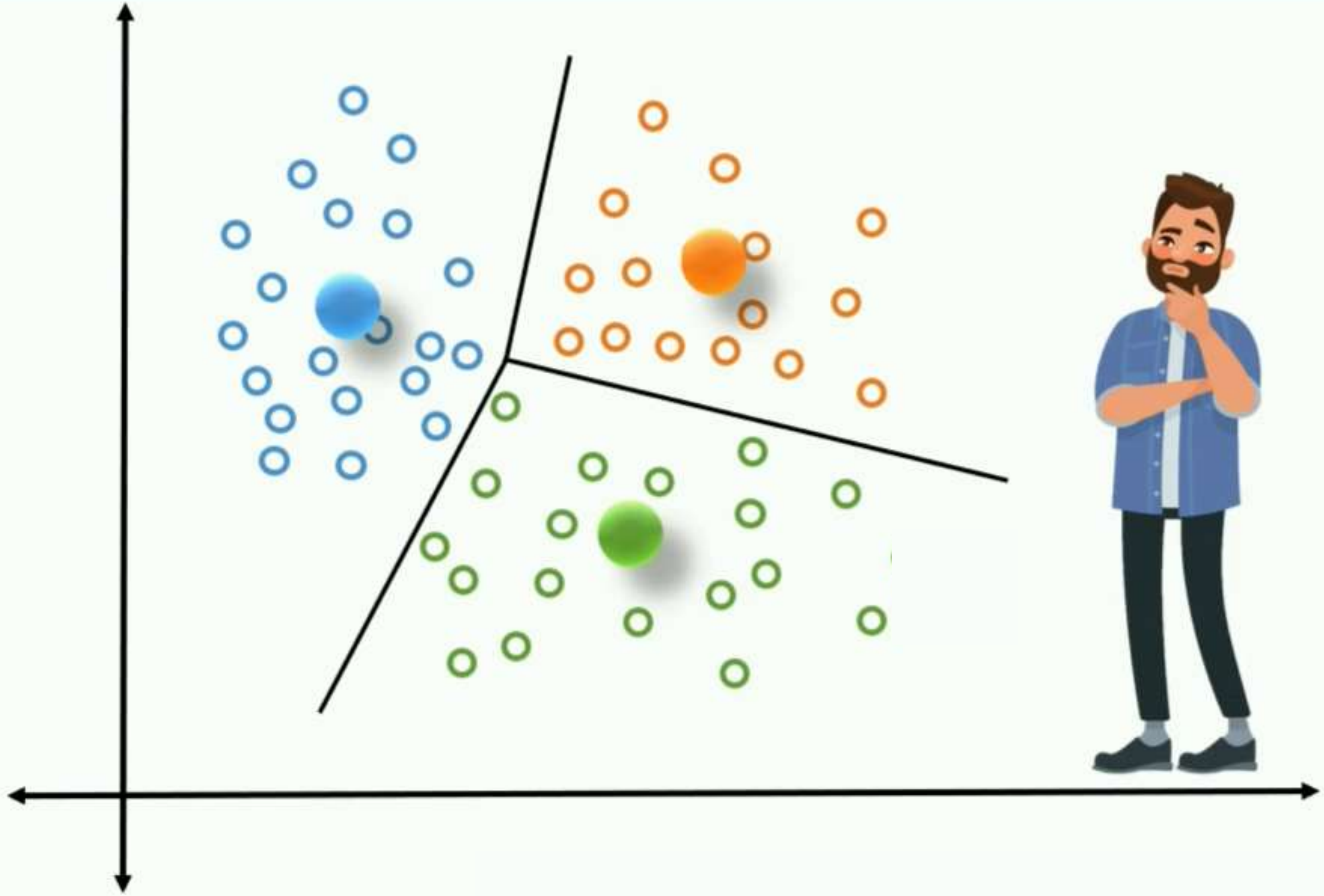
K-Means Clustering - Algorithm

1. Randomly choose k data points(seeds) to be the initial centroids, cluster centres
2. Assign each data point to the closet centroid
3. Re-compute the centroids using the current cluster memberships
4. If a convergence criterion is not meet, go to step 2

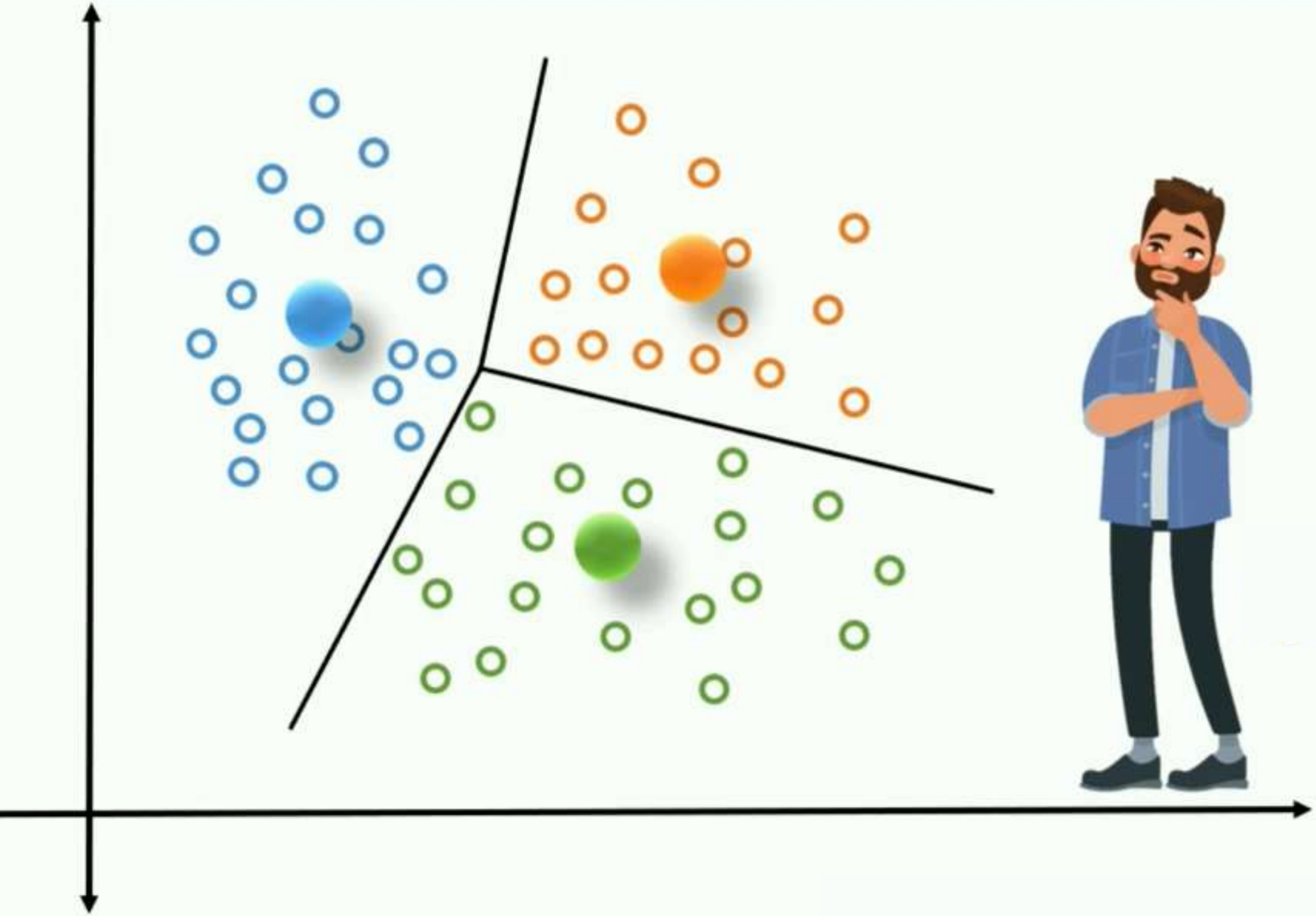
K-Means Clustering – Flow Chart



Good Clusters?



Good Clusters?



Similar characteristics

Proportionate number of observations

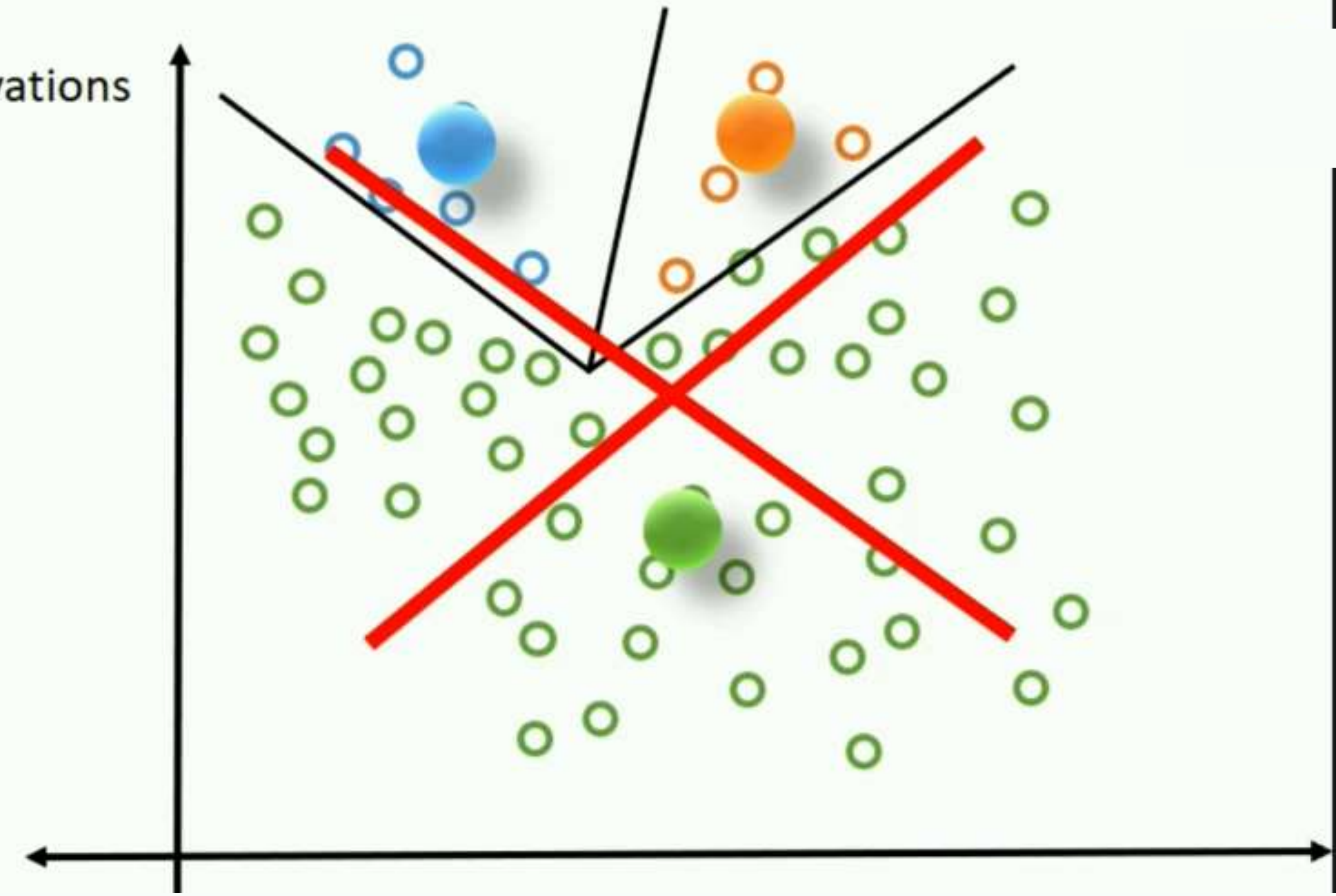
Good Cluster Analysis

- Observations in the same group share similar characteristics

Good Cluster Analysis

- Observations in the same group share similar characteristics

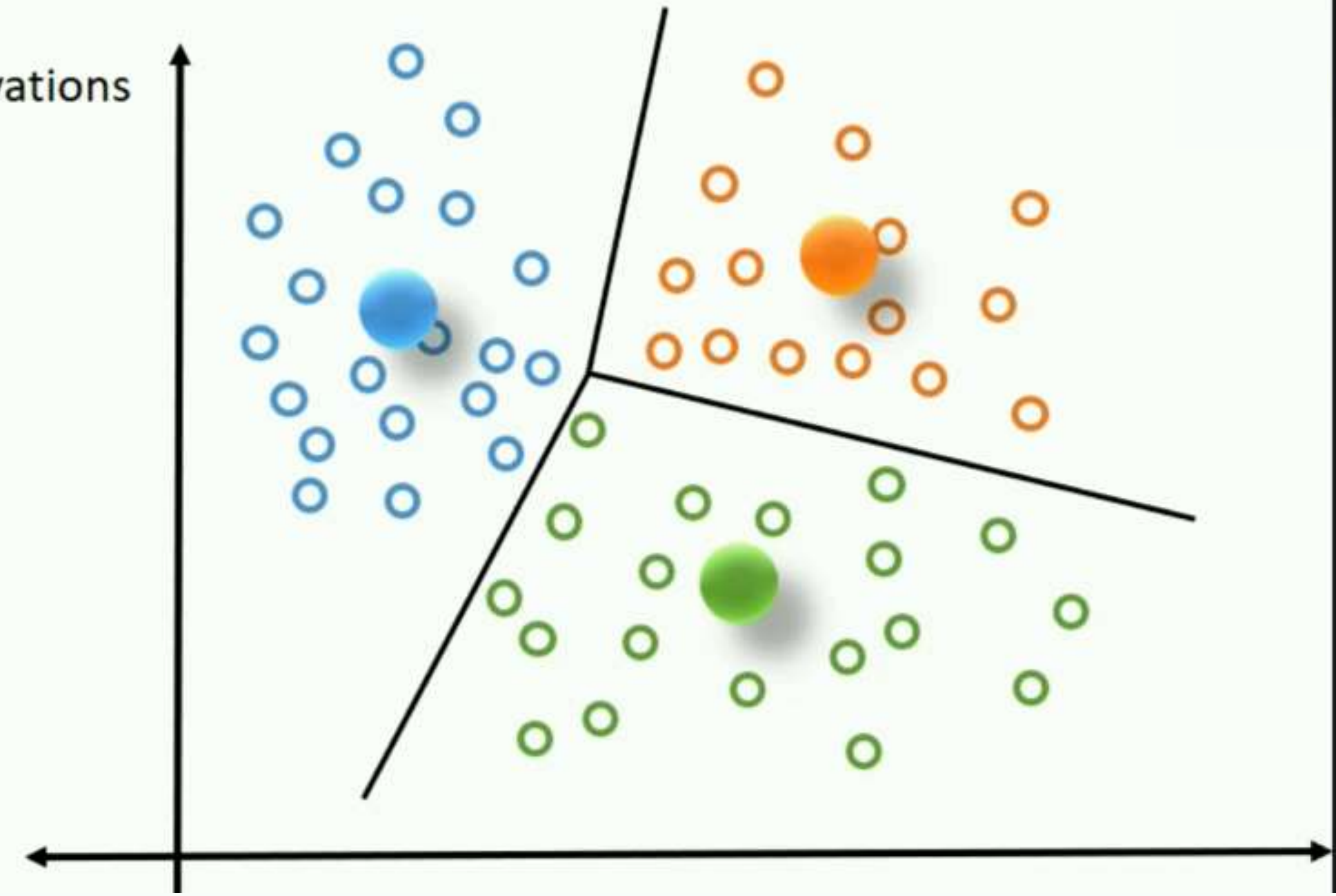
- Clusters have proportionate number observations



Good Cluster Analysis

- Observations in the same group share similar characteristics

- Clusters have proportionate number observations

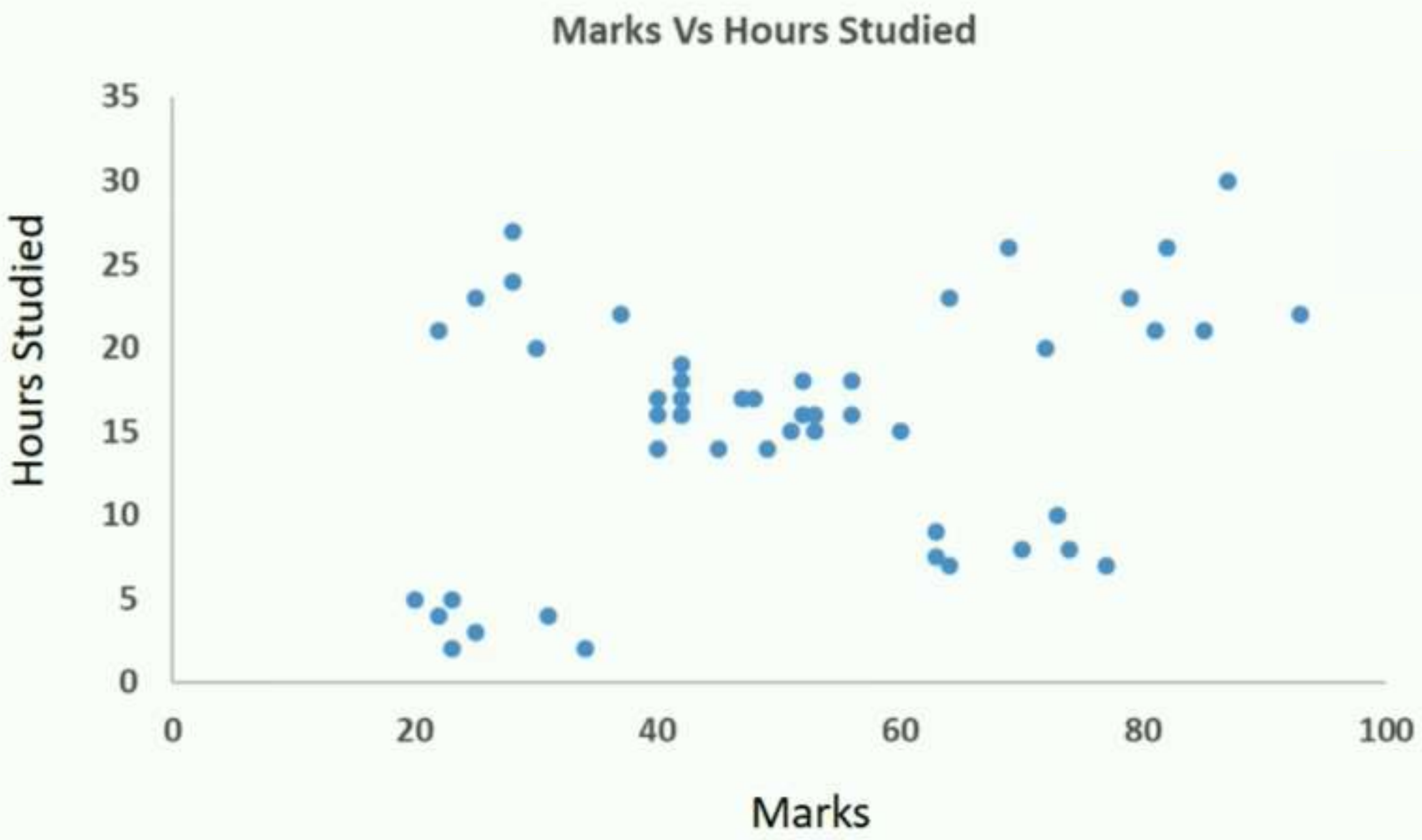


Cluster of Students

Marks Obtained	Hours Studied
72	20
42	19
77	7
93	22
30	20
53	15
74	8
28	24
69	26
64	7
87	30
70	8
42	18
79	23
37	22
52	16
51	15

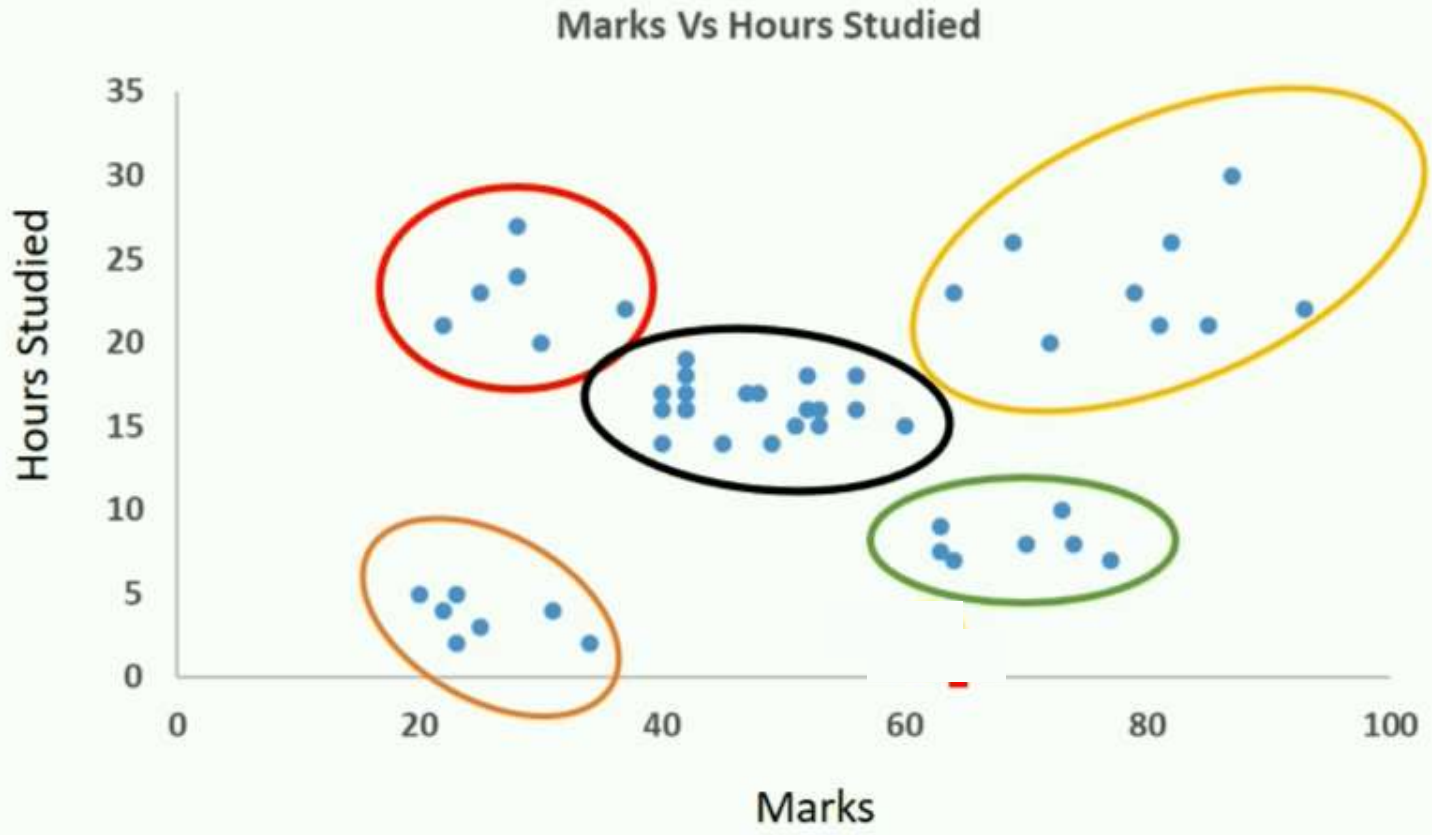
Cluster of Students

Marks Obtained	Hours Studied
72	20
42	19
77	7
93	22
30	20
53	15
74	8
28	24
69	26
64	7
87	30
70	8
42	18
79	23
37	22
52	16
51	15

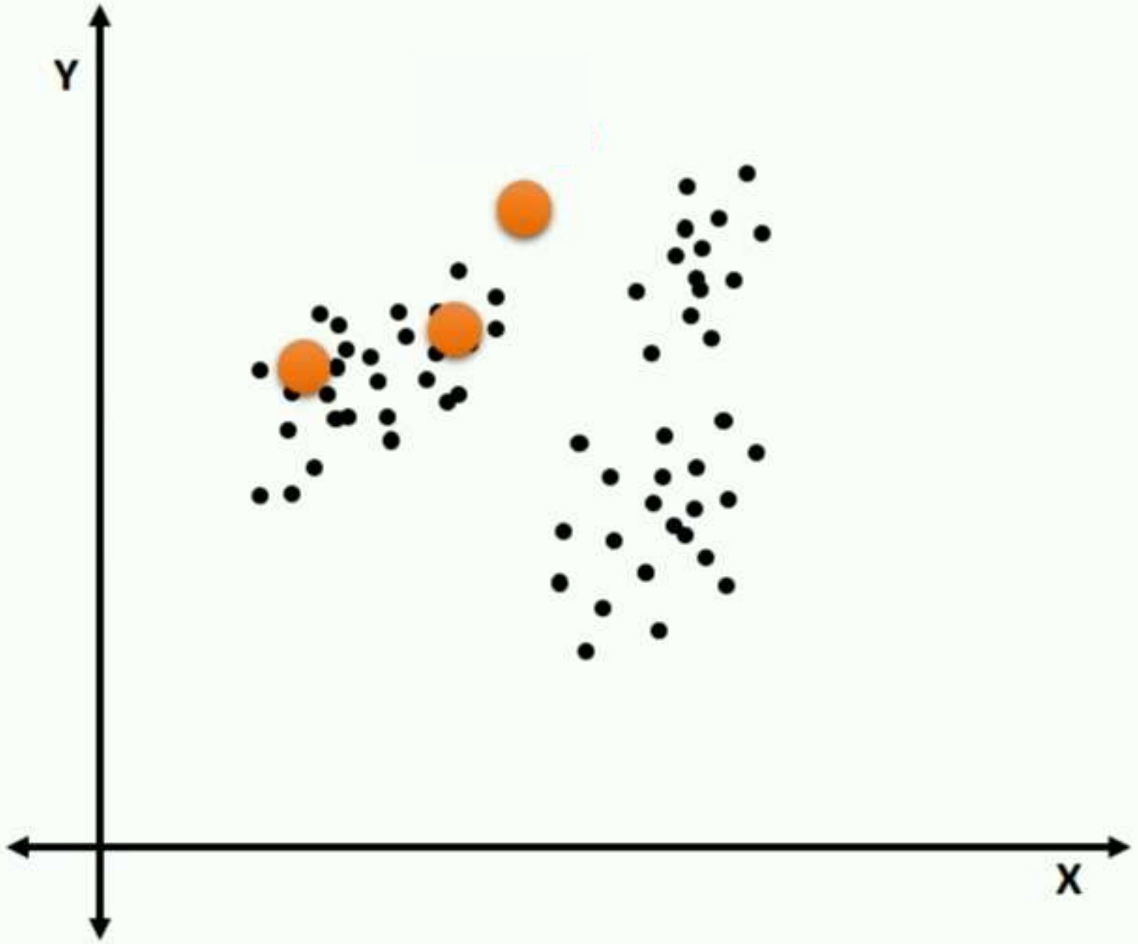


Cluster of Students

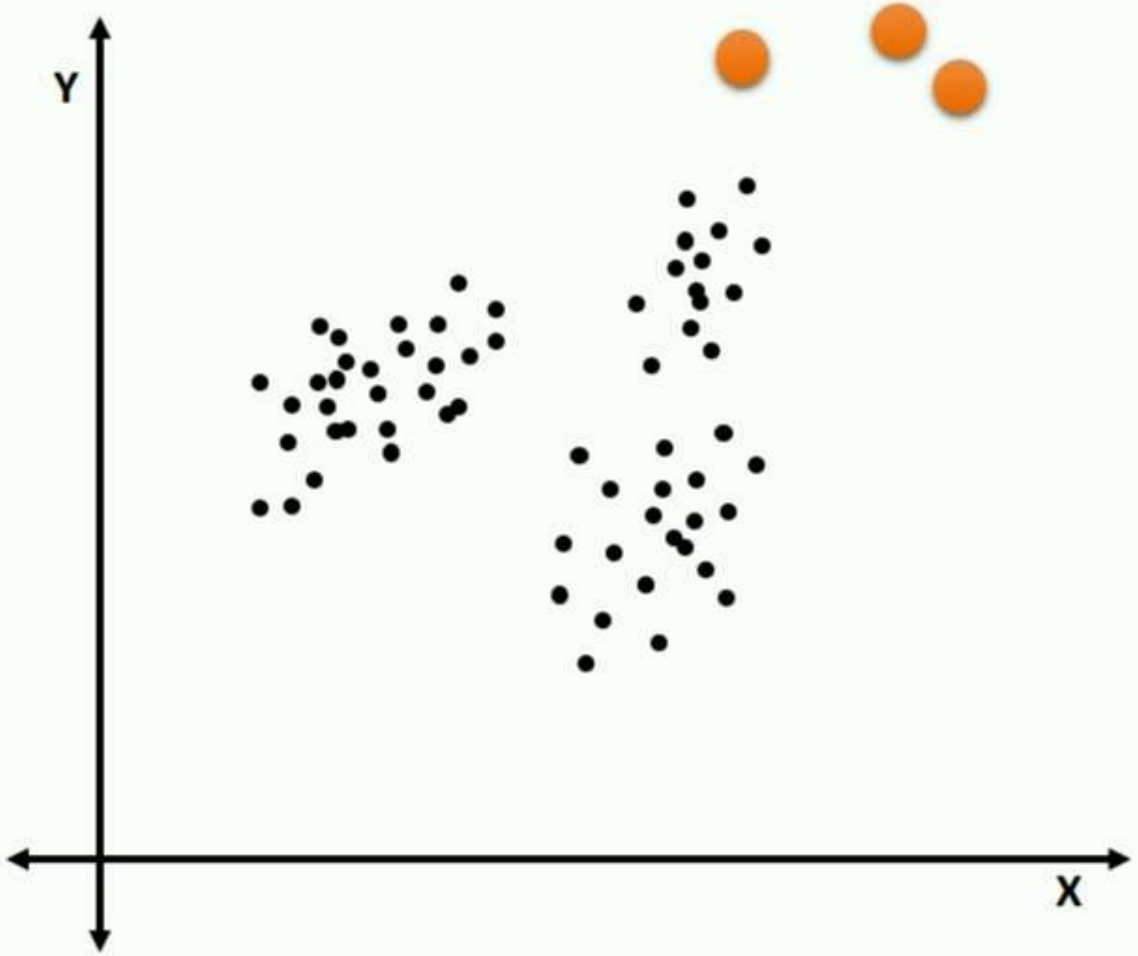
Marks Obtained	Hours Studied
72	20
42	19
77	7
93	22
30	20
53	15
74	8
28	24
69	26
64	7
87	30
70	8
42	18
79	23
37	22
52	16
51	15



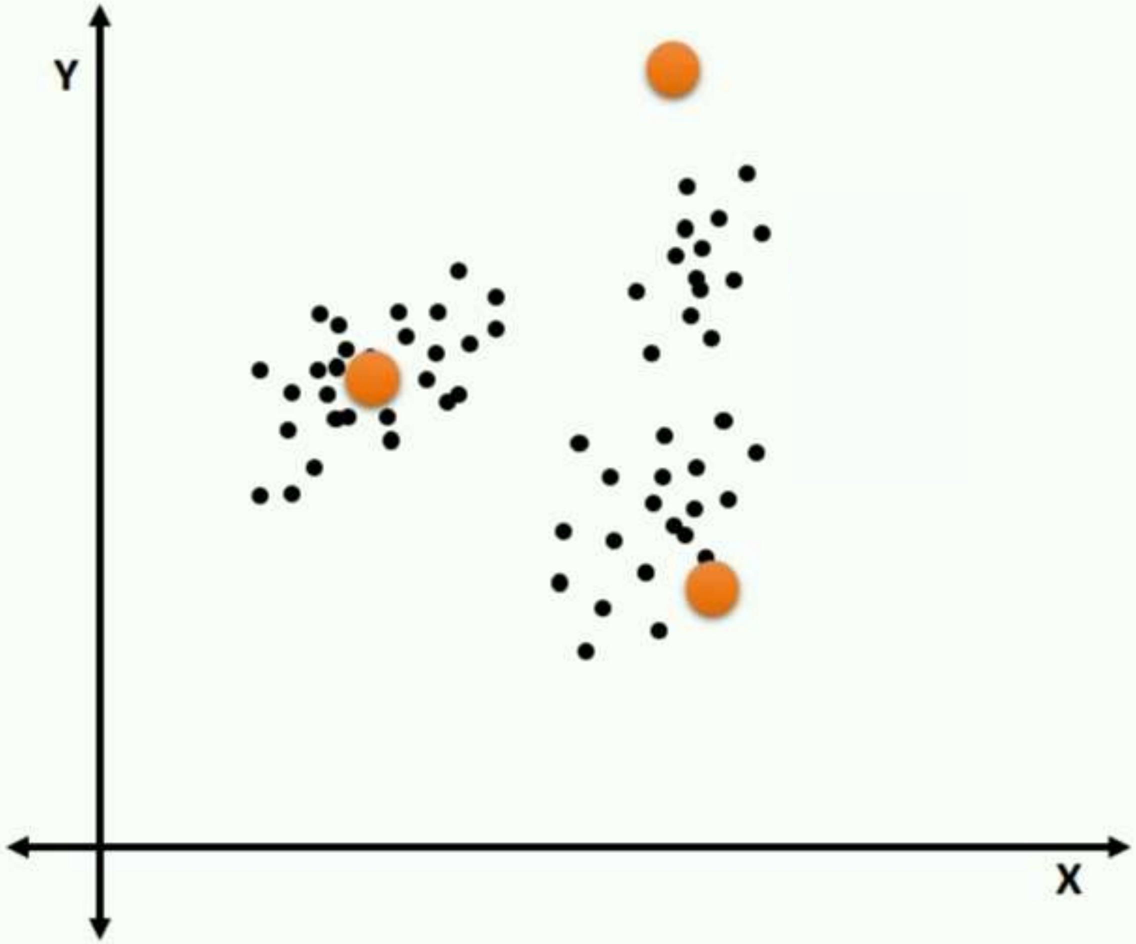
Demo: Apply cluster analysis on student marks dataset



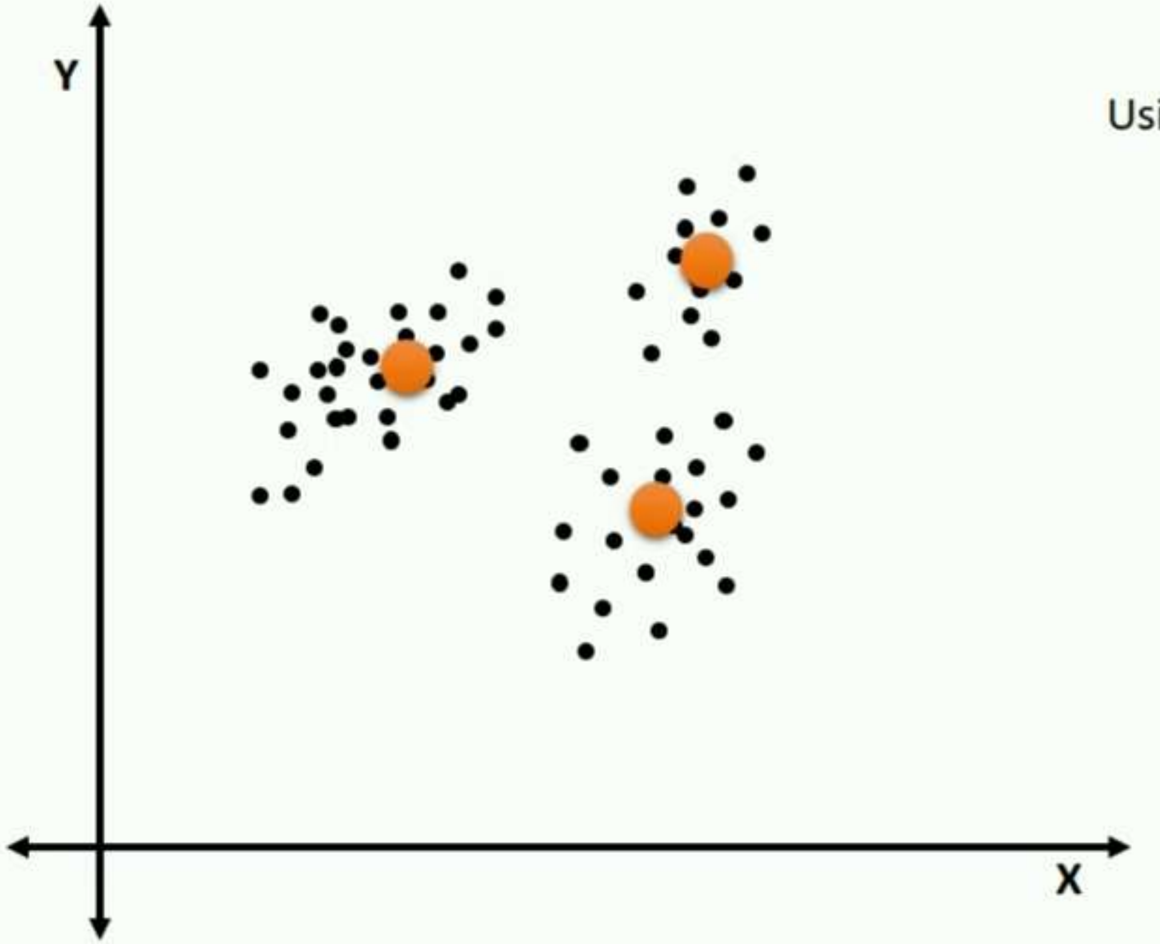
Using only Random method



Using only Random method

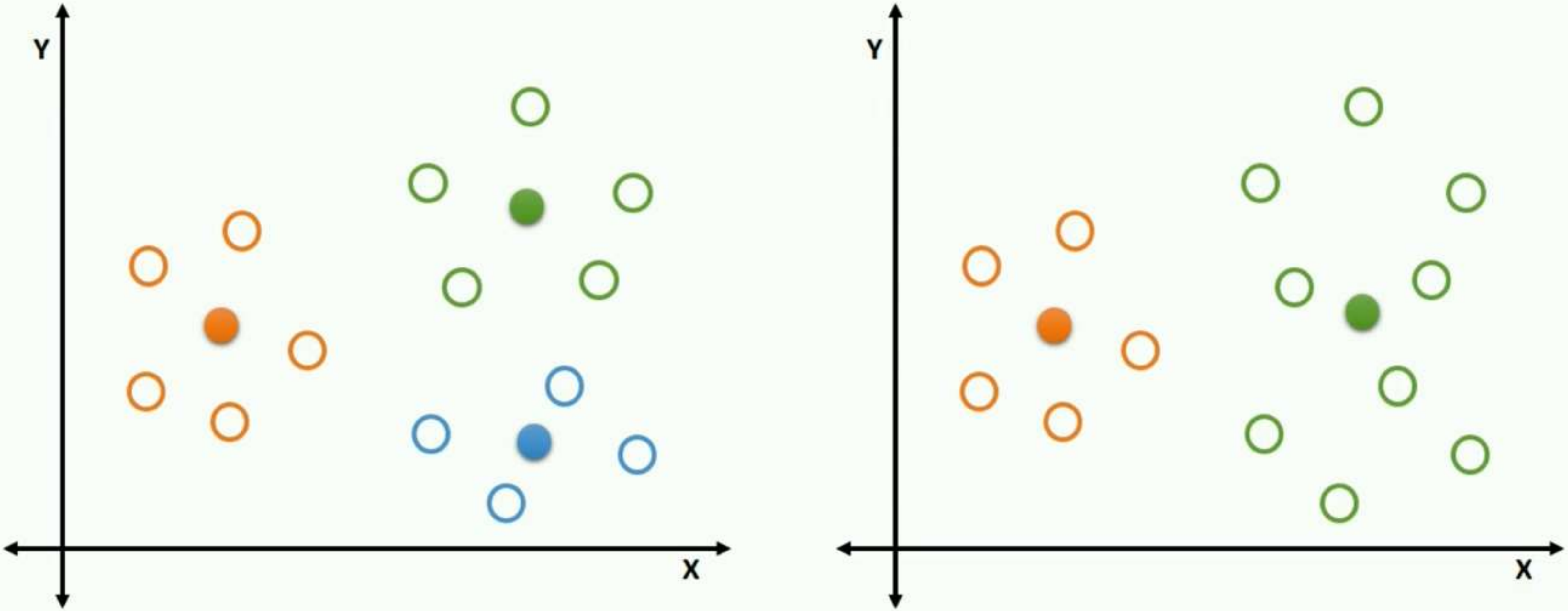


Using Kmeans++

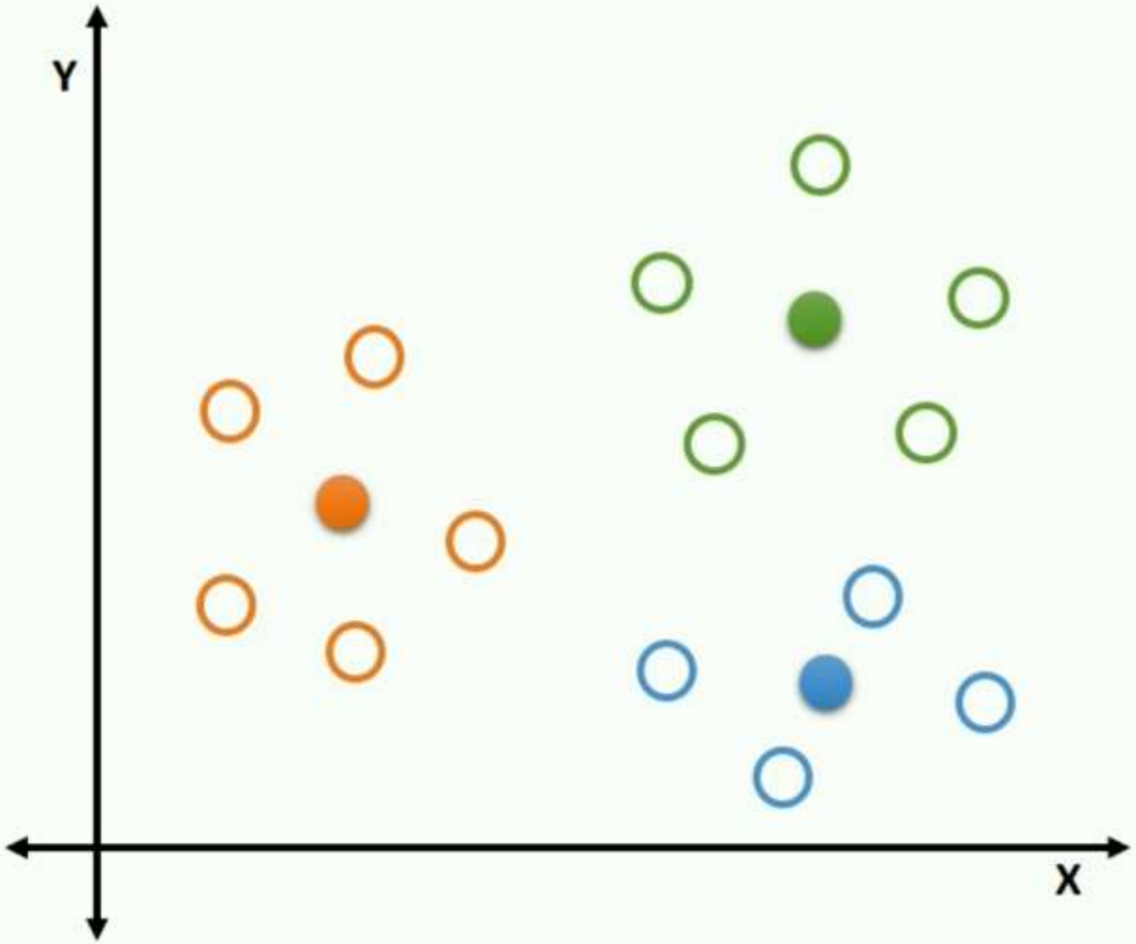


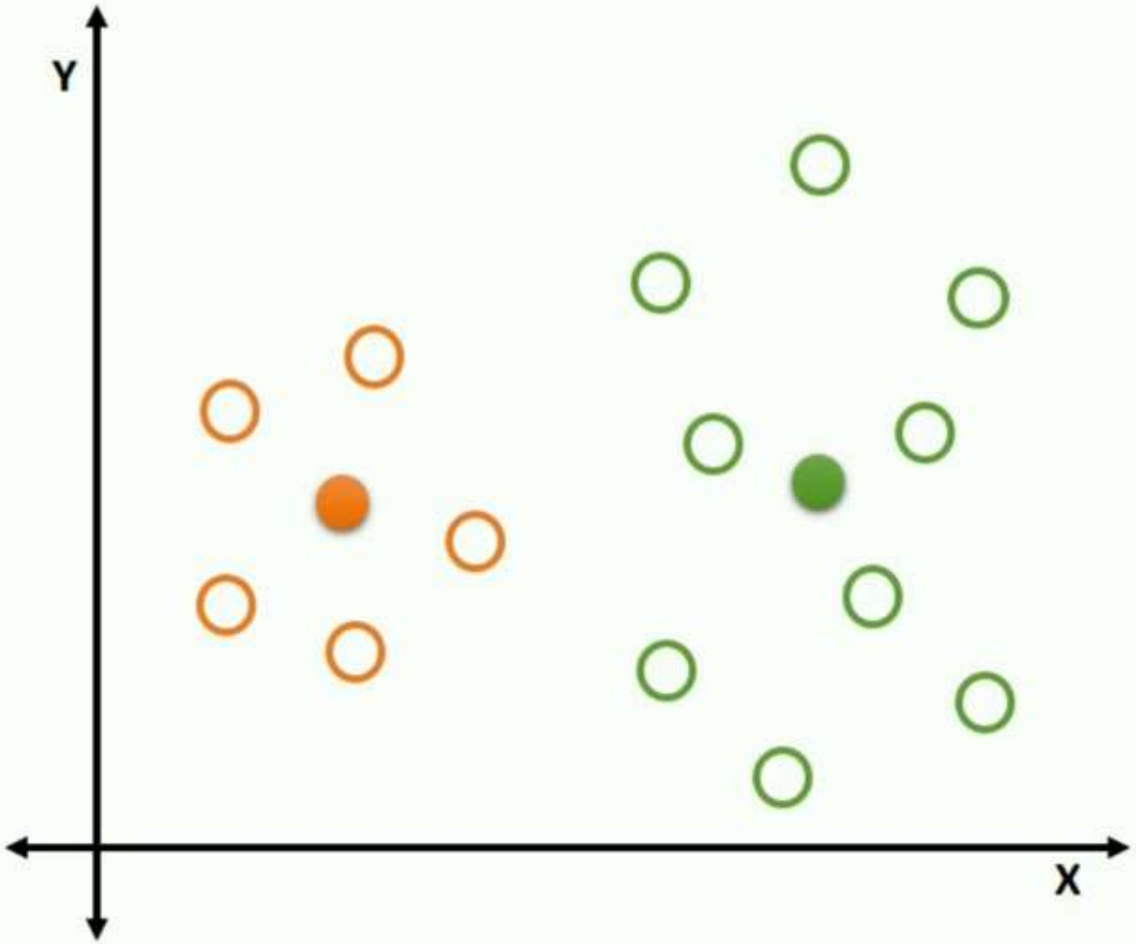
Using Kmeans++ with KMeans

How to Decide Number of Clusters?


Y
X
Y
X

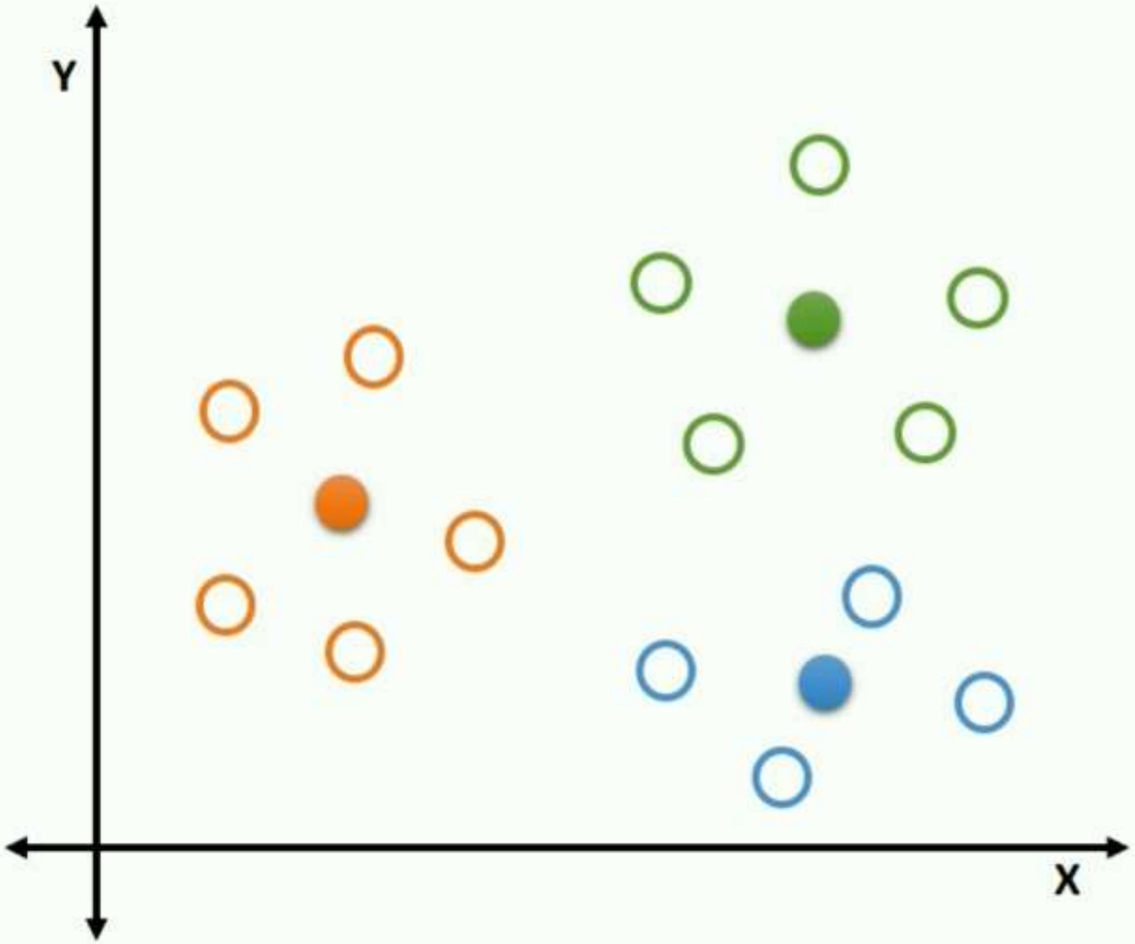
Observations in the same clusters are related to each other.

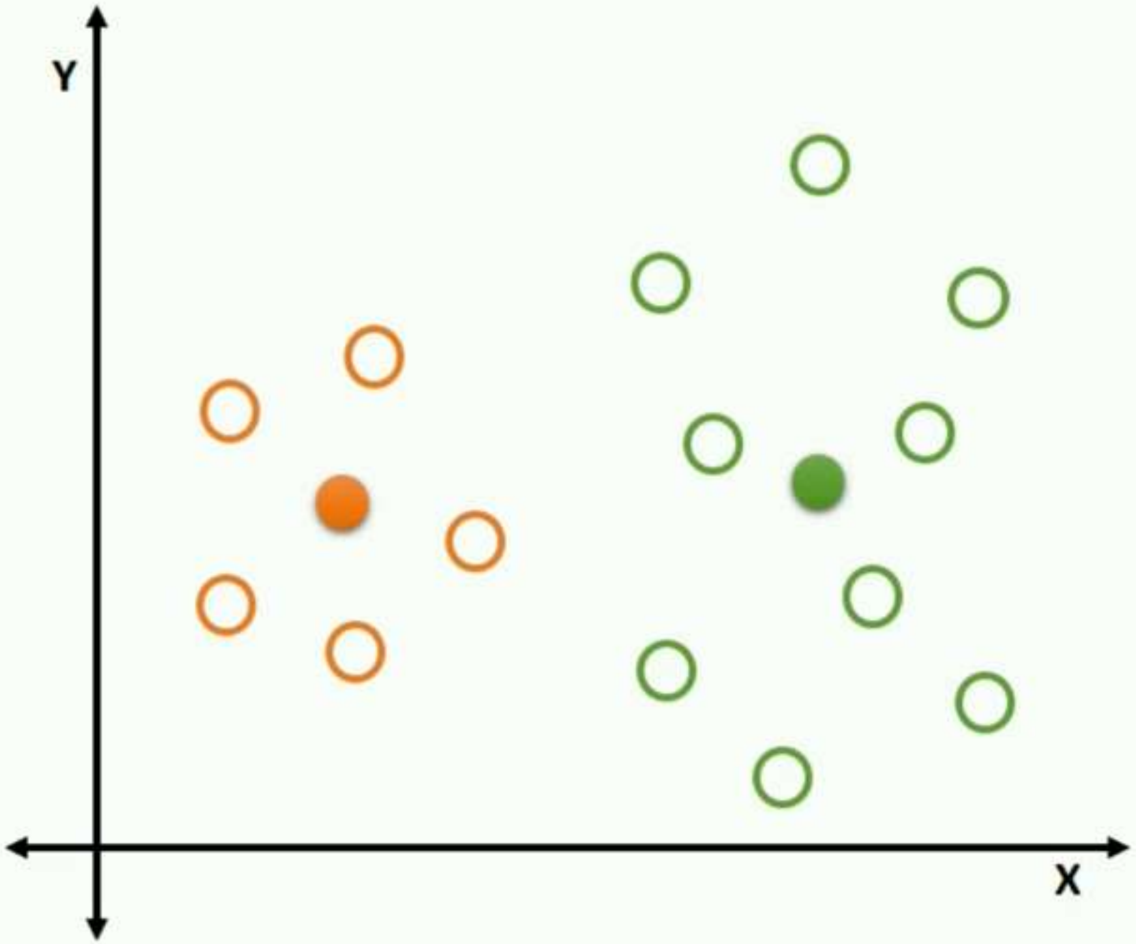




Observations in the same clusters are related to each other.

Distance of the observations from centroid decides their cluster assignment.

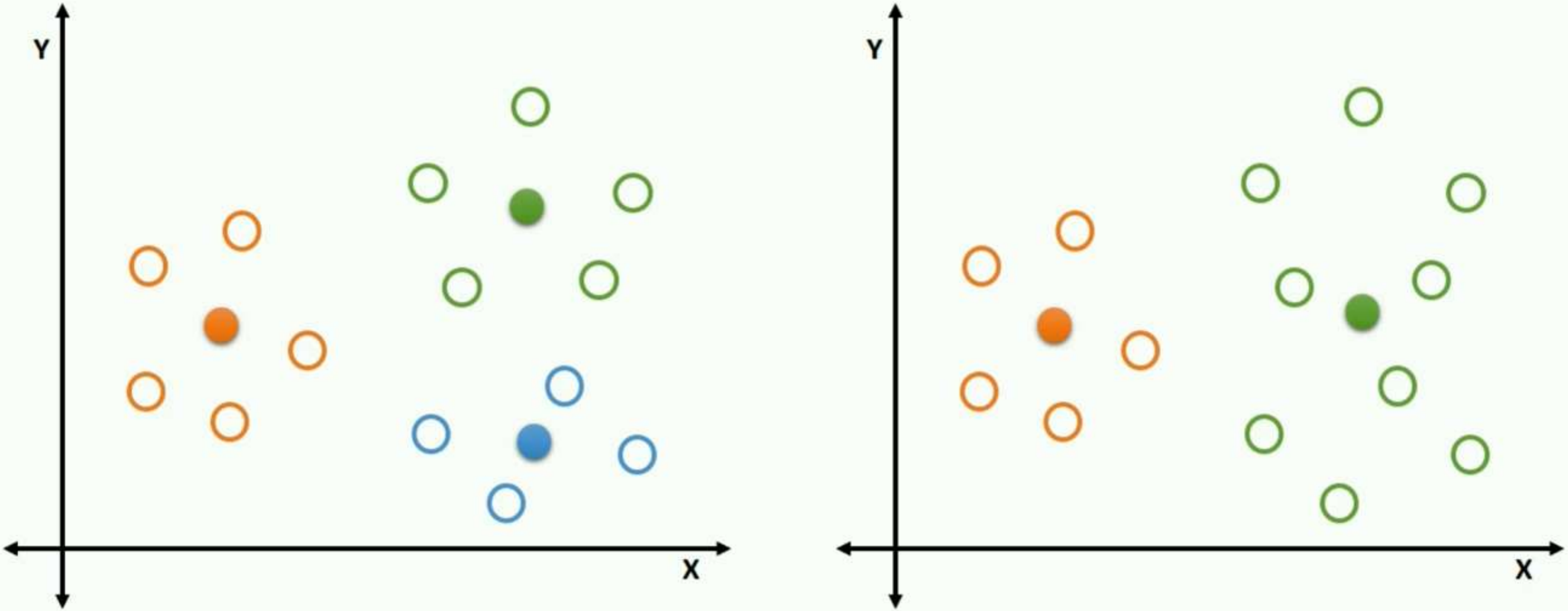




Observations in the same clusters are related to each other.

Distance of the observations from centroid decides their cluster assignment.

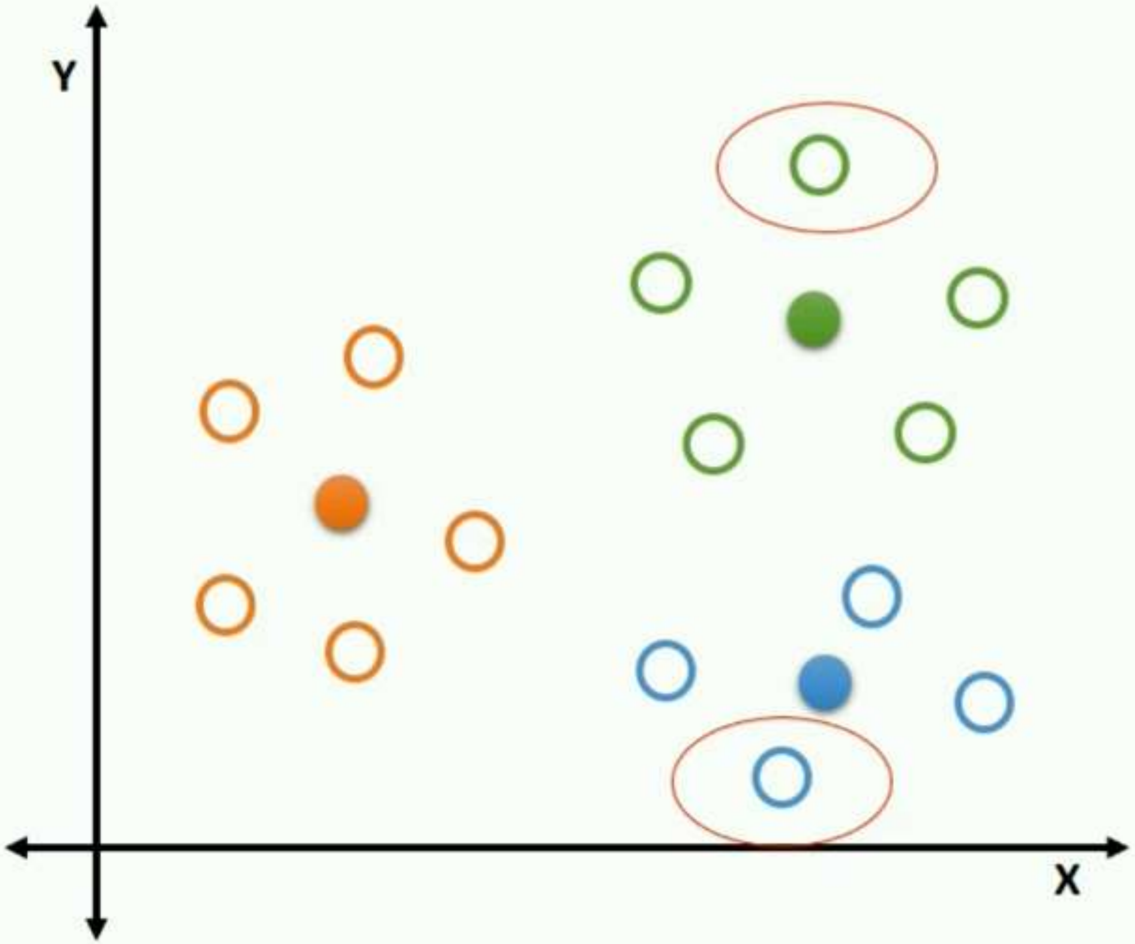
Lesser the distance,
Better the relationship

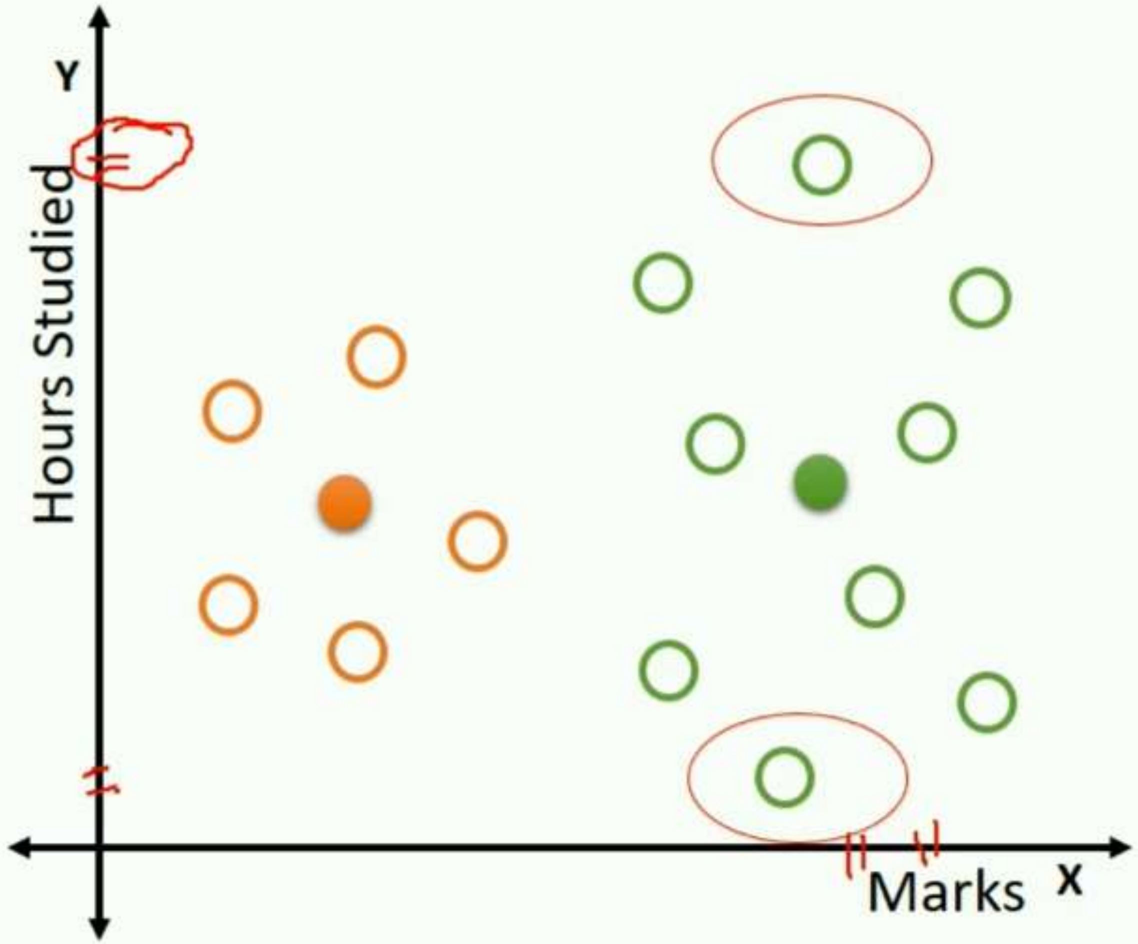


Observations in the same clusters are related to each other.

Distance of the observations from centroid decides their cluster assignment.

Lesser the distance,
Better the relationship

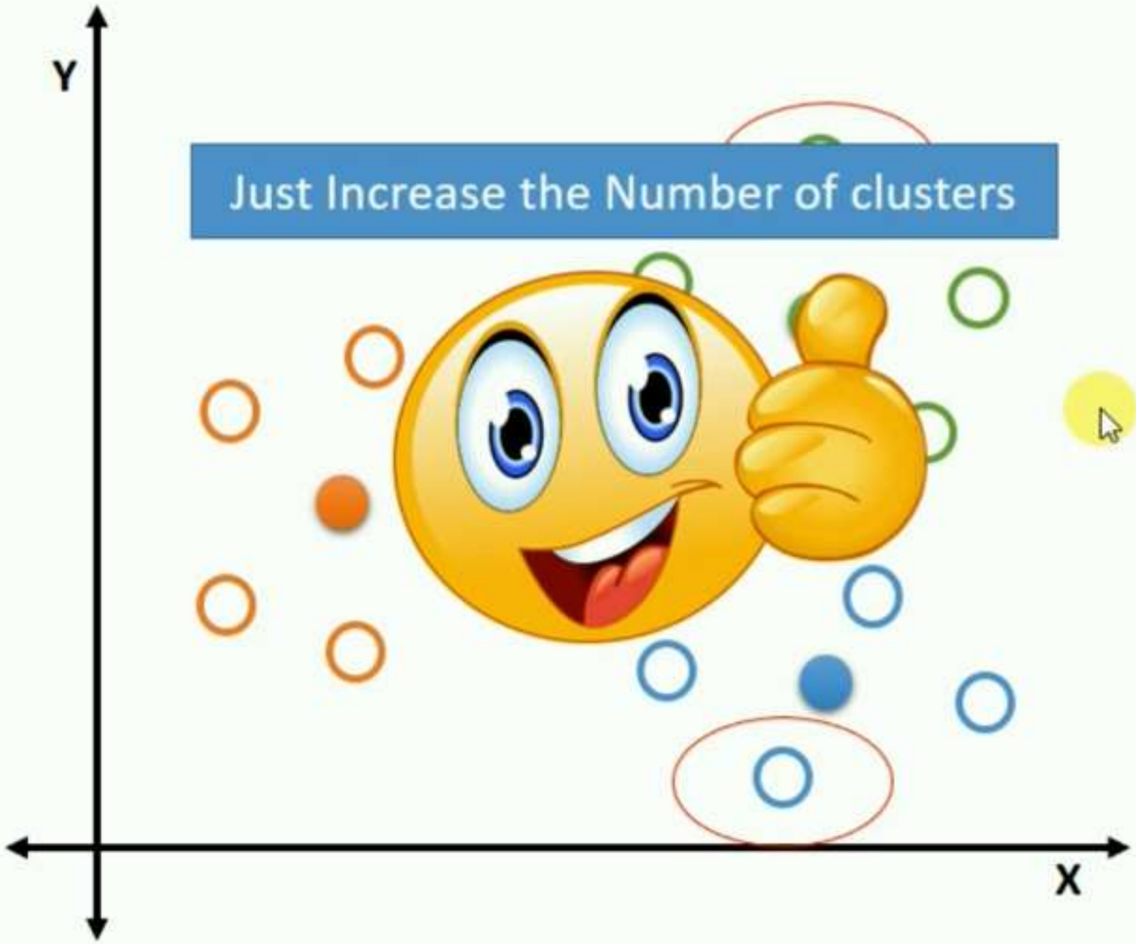


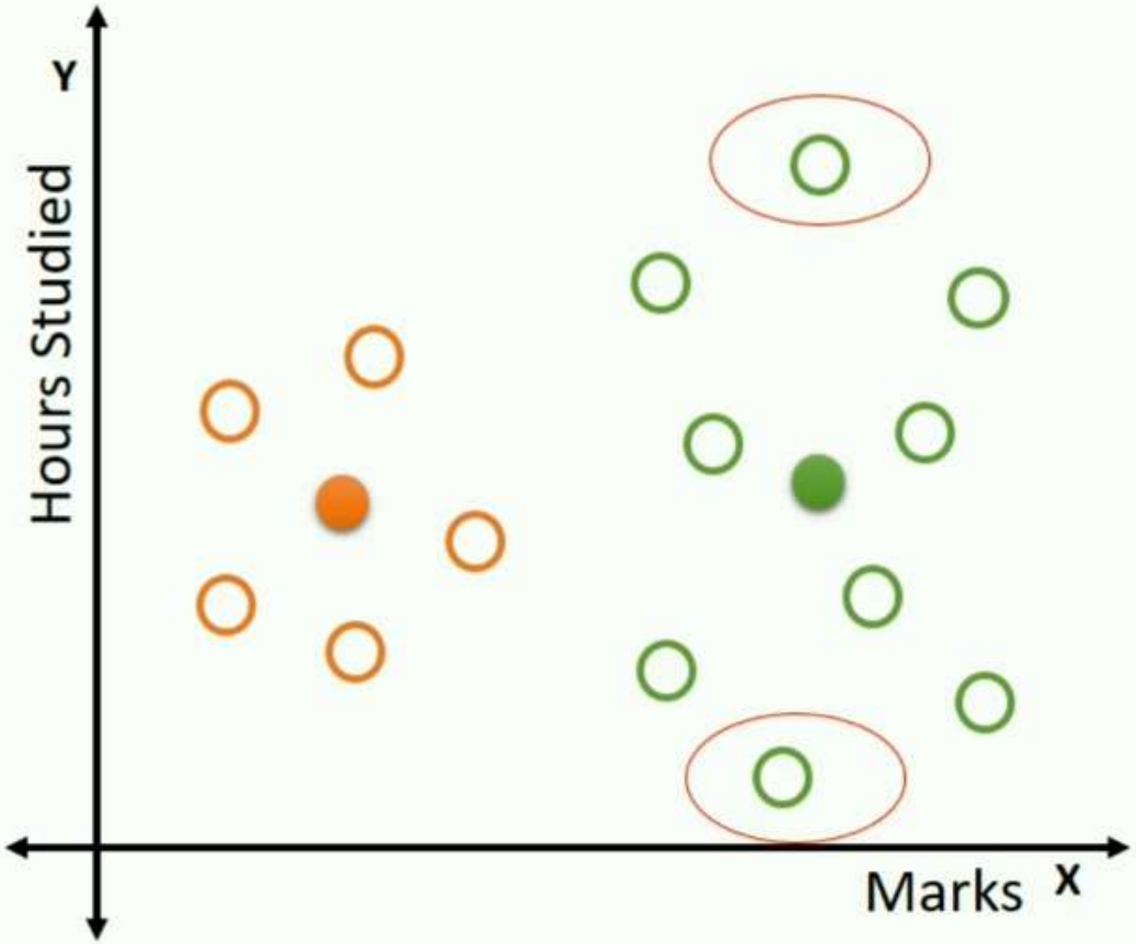


Observations in the same clusters are related to each other.

Distance of the observations from centroid decides their cluster assignment.

Lesser the distance,
Better the relationship

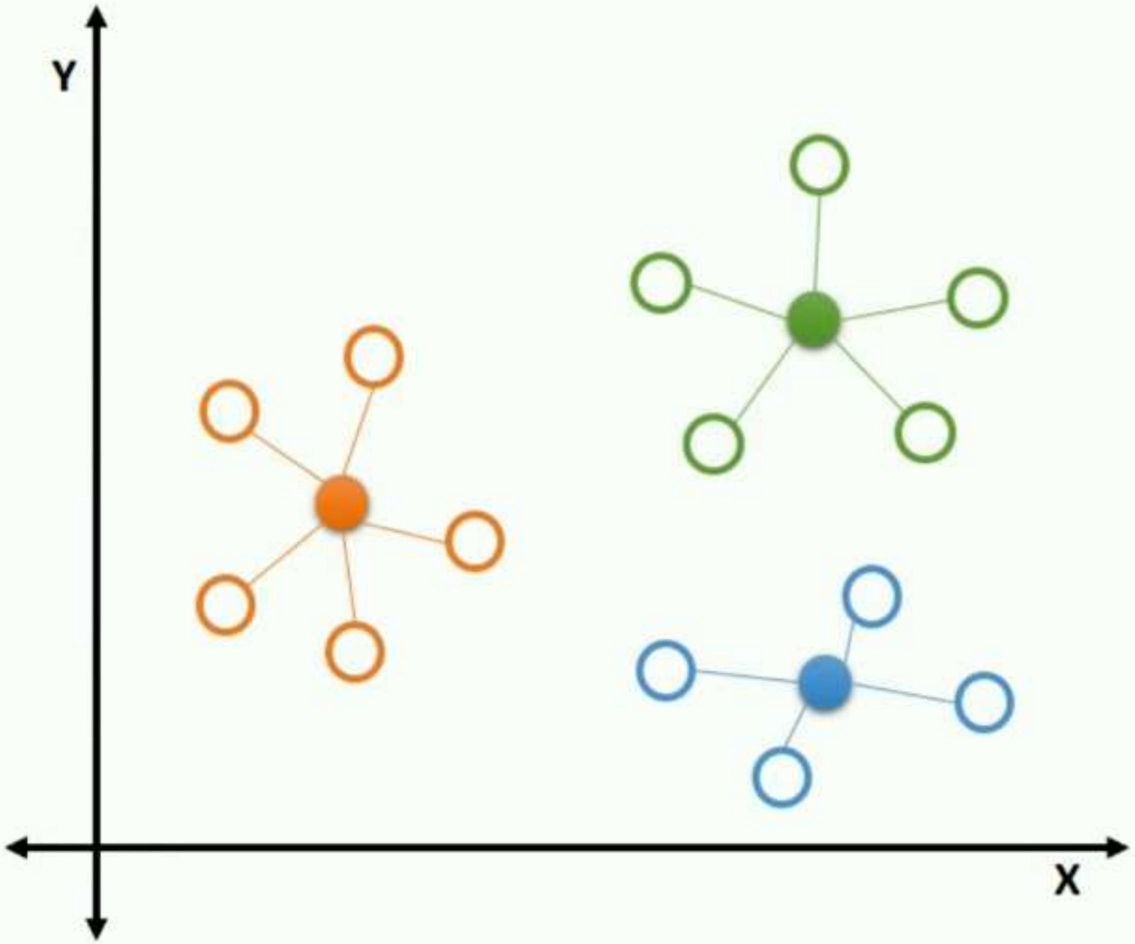


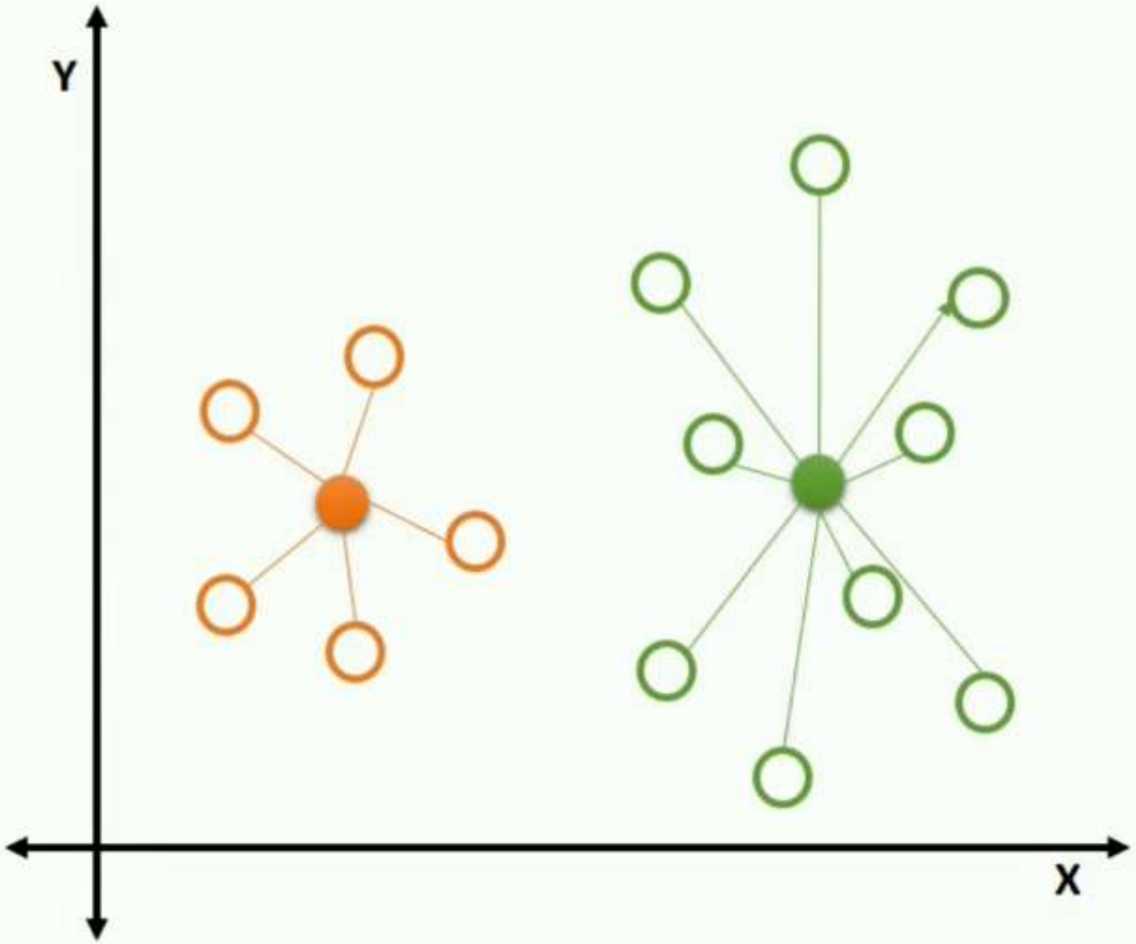


Observations in the same clusters are related to each other.

Distance of the observations from centroid decides their cluster assignment.

Lesser the distance,
Better the relationship



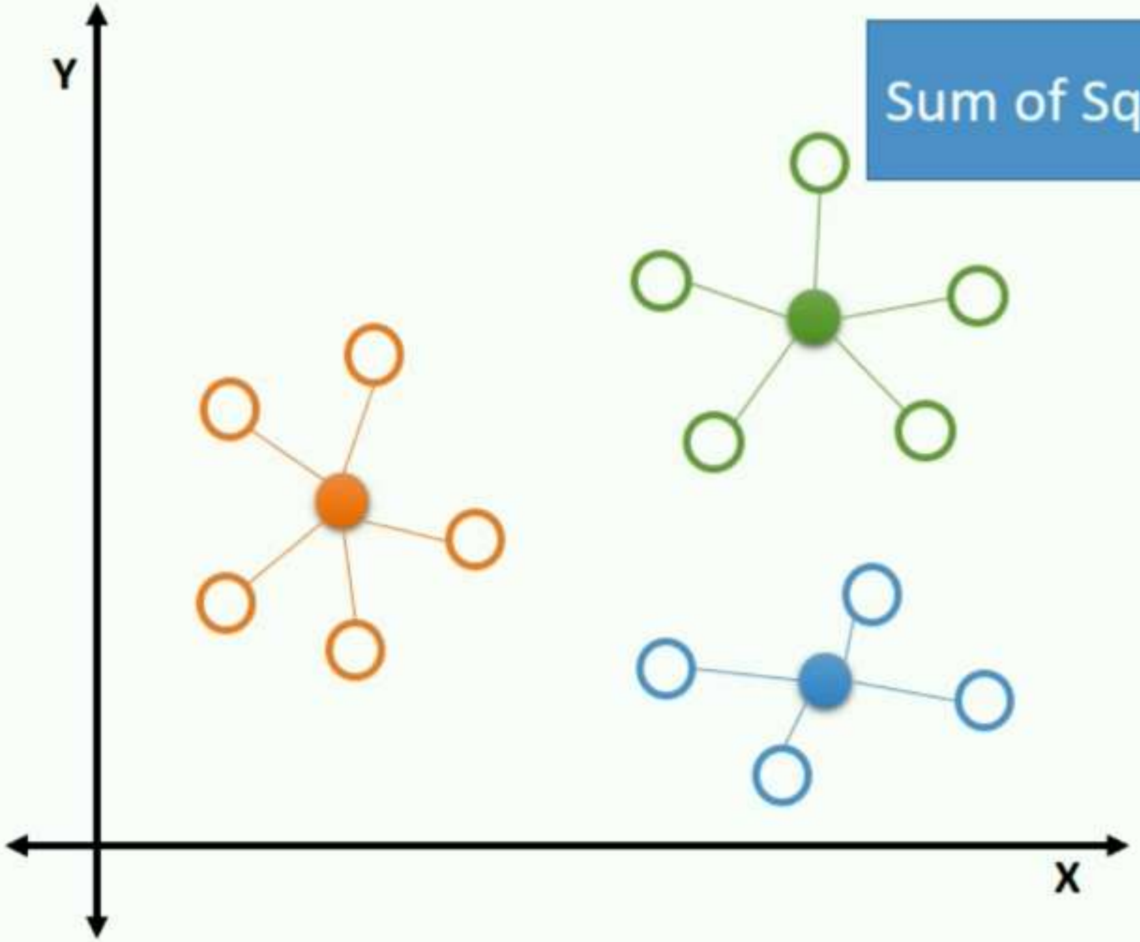


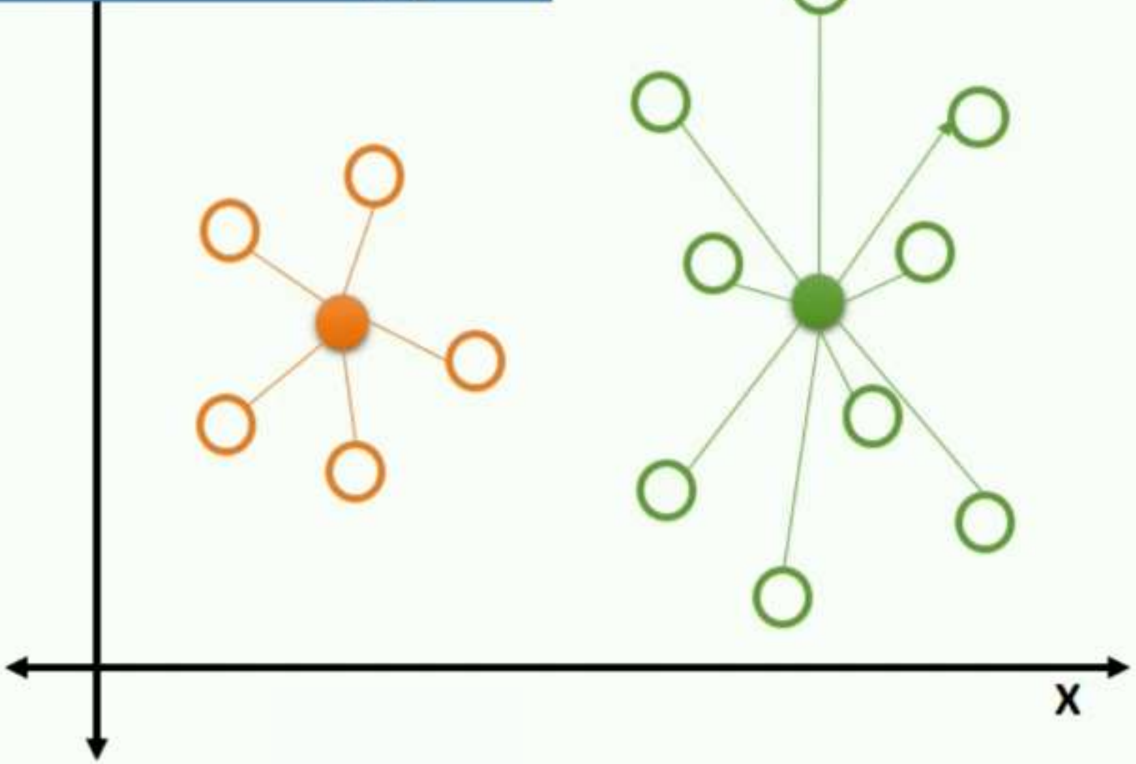
Observations in the same clusters are related to each other.

Distance of the observations from centroid decides their cluster assignment.

Lesser the distance,
Better the relationship

Sum of Squared Distance $\sum_{i=1}^{n} d_i^2$



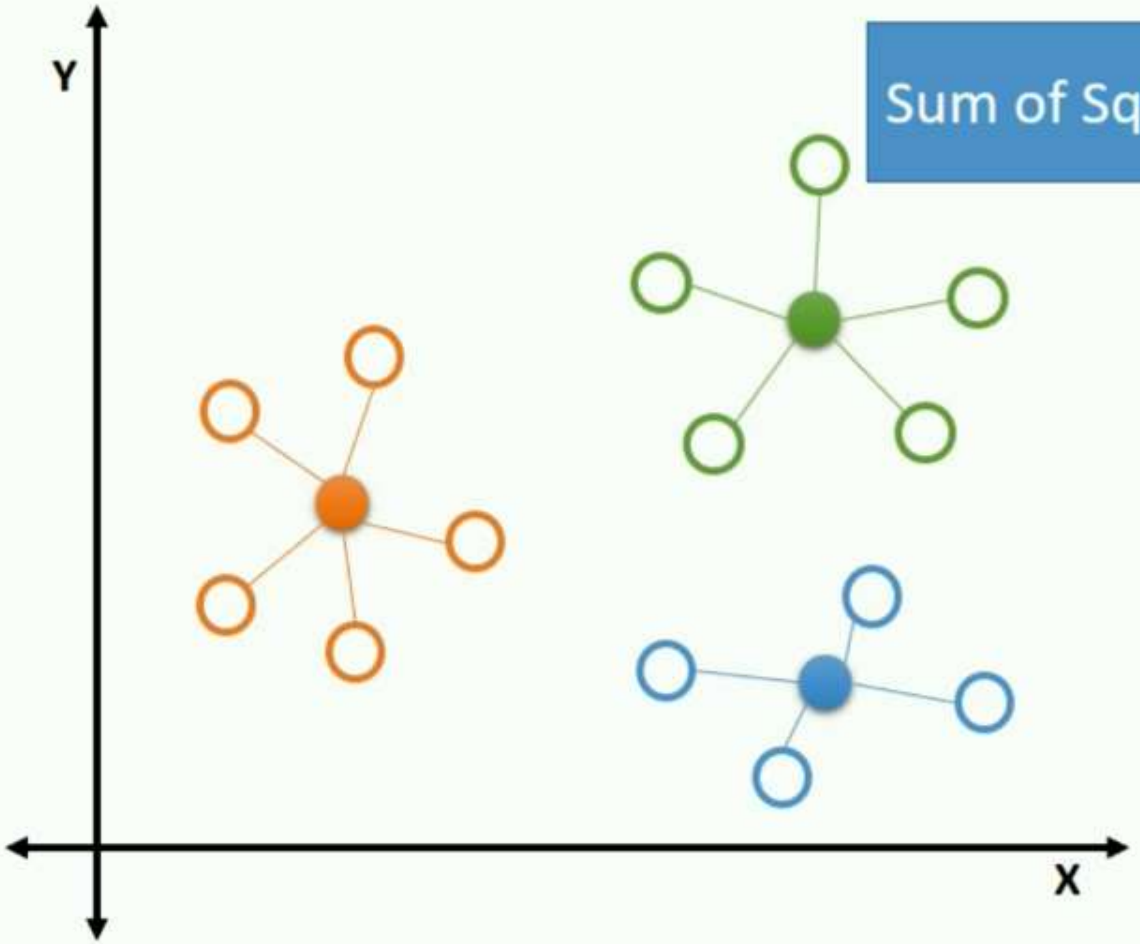


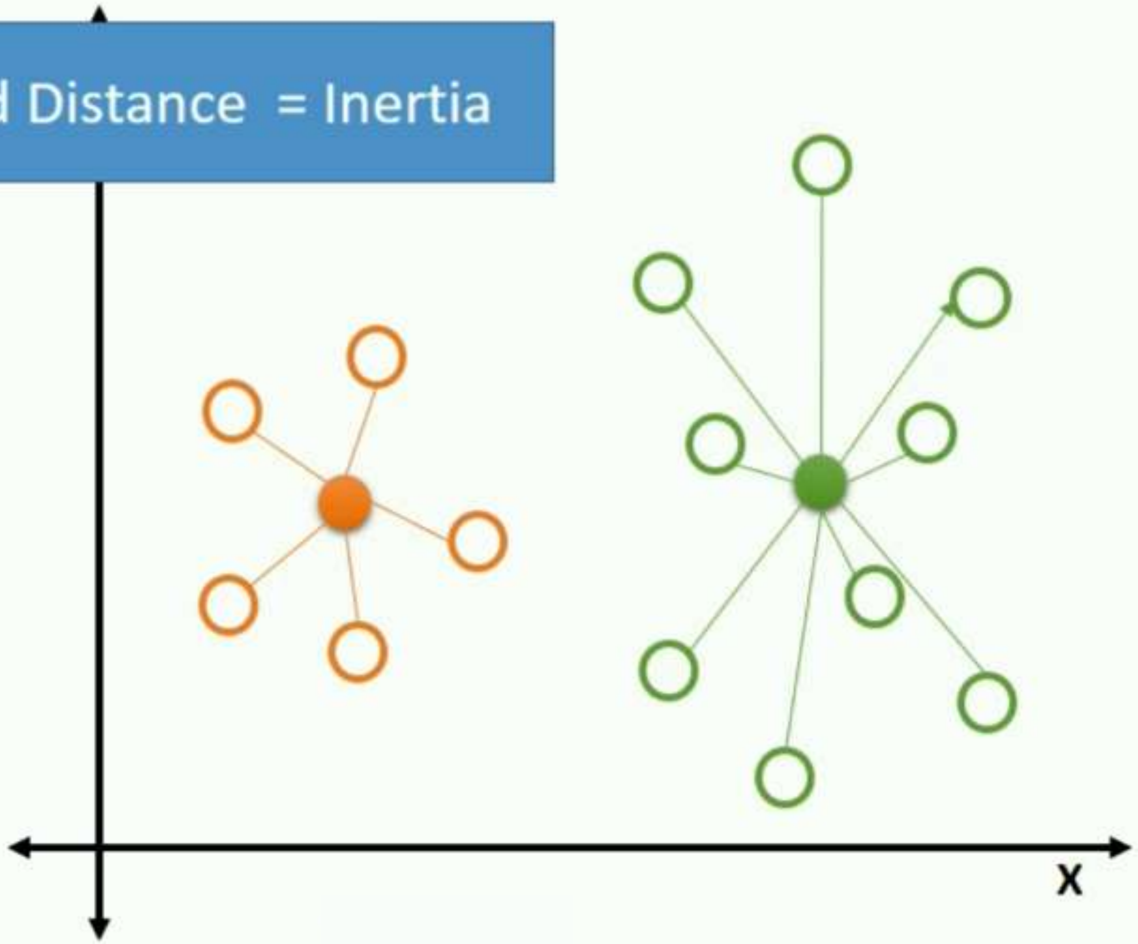
Observations in the same clusters are related to each other.

Distance of the observations from centroid decides their cluster assignment.

Lesser the distance,
Better the relationship

Sum of Squared Distance = Inertia





Source Editor Object sklearn.cluster.k_means_.KMeans

Attributes

cluster_centers_array, [n_clusters, n_features]

Coordinates of cluster centers. If the algorithm stops before fully converging (see `tol` and `max_iter`), these will not be consistent with `labels_`.

labels_ :

Labels of each point

inertia_float

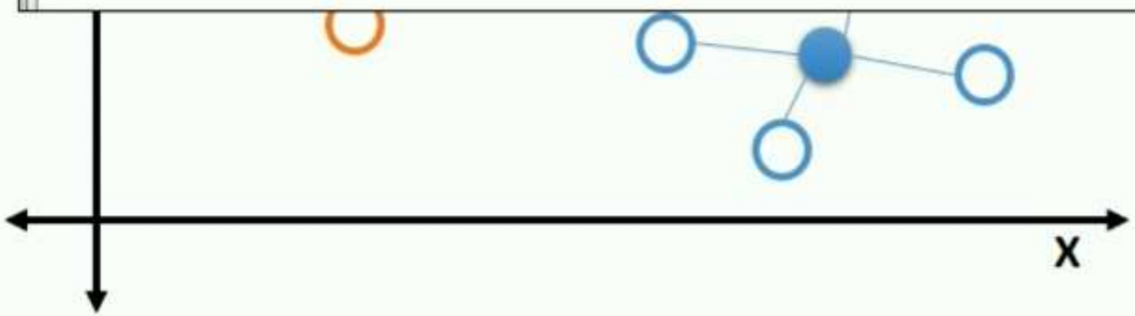
Sum of squared distances of samples to their closest cluster center.

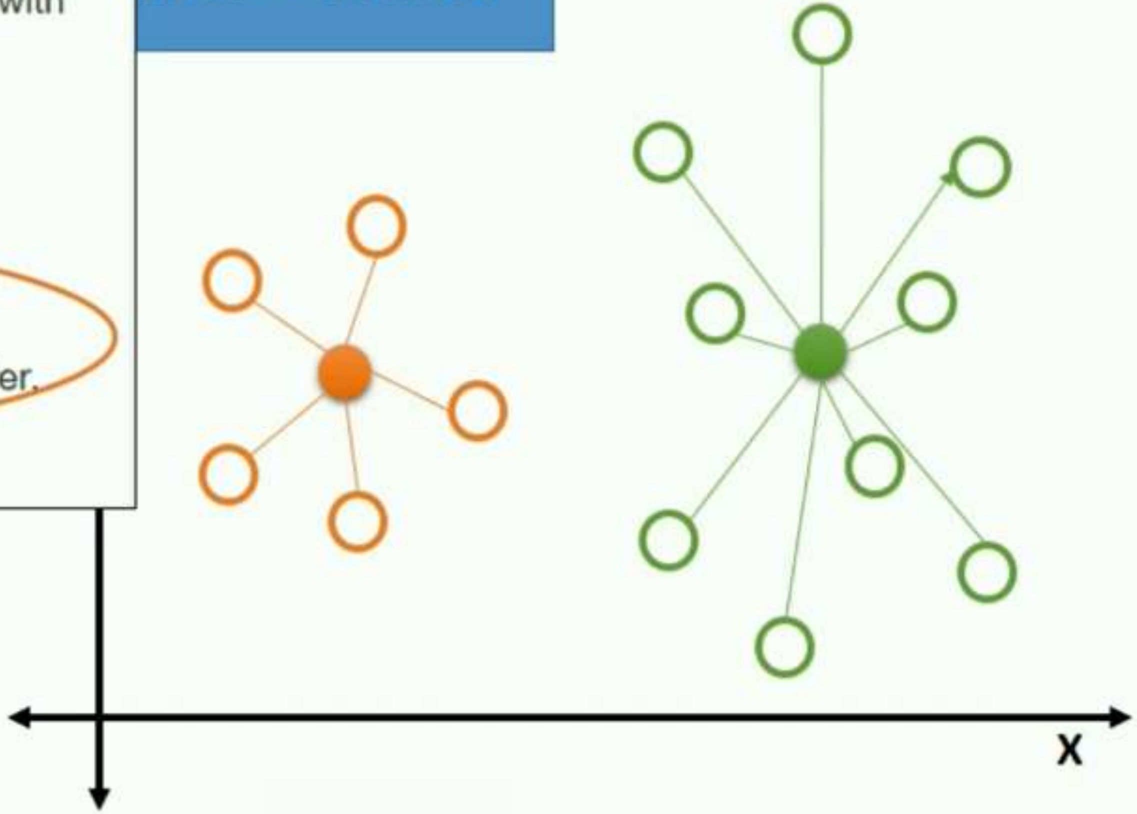
n_iter_int

ignment.

Lesser the distance,
Better the relationship

ance = Inertia



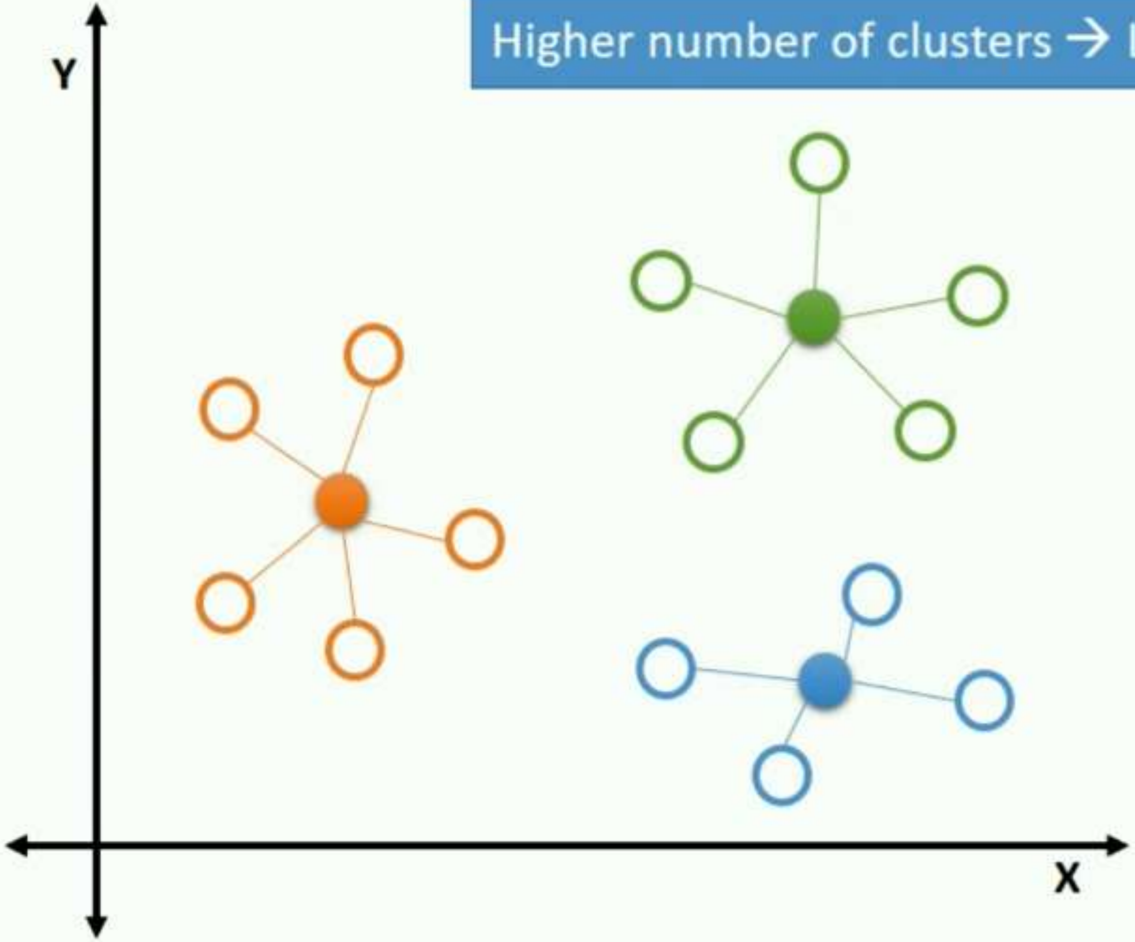


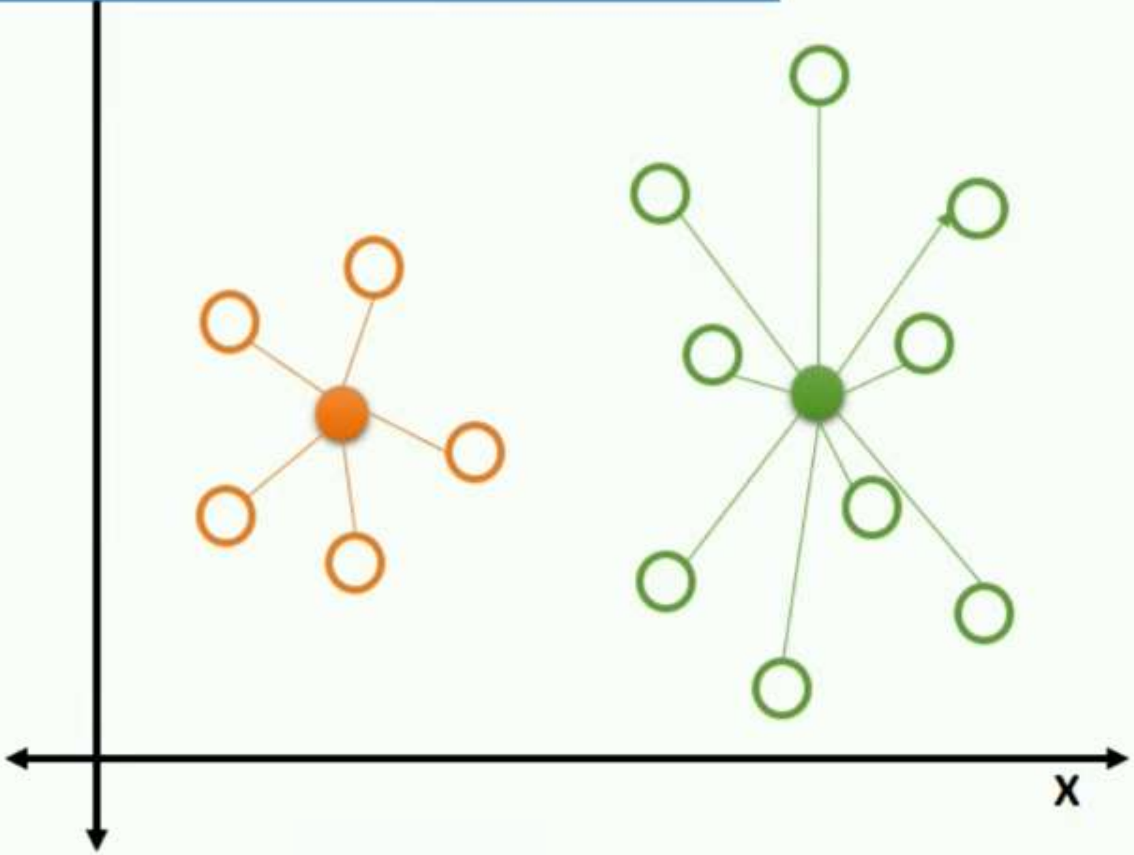
Observations in the same clusters are related to each other.

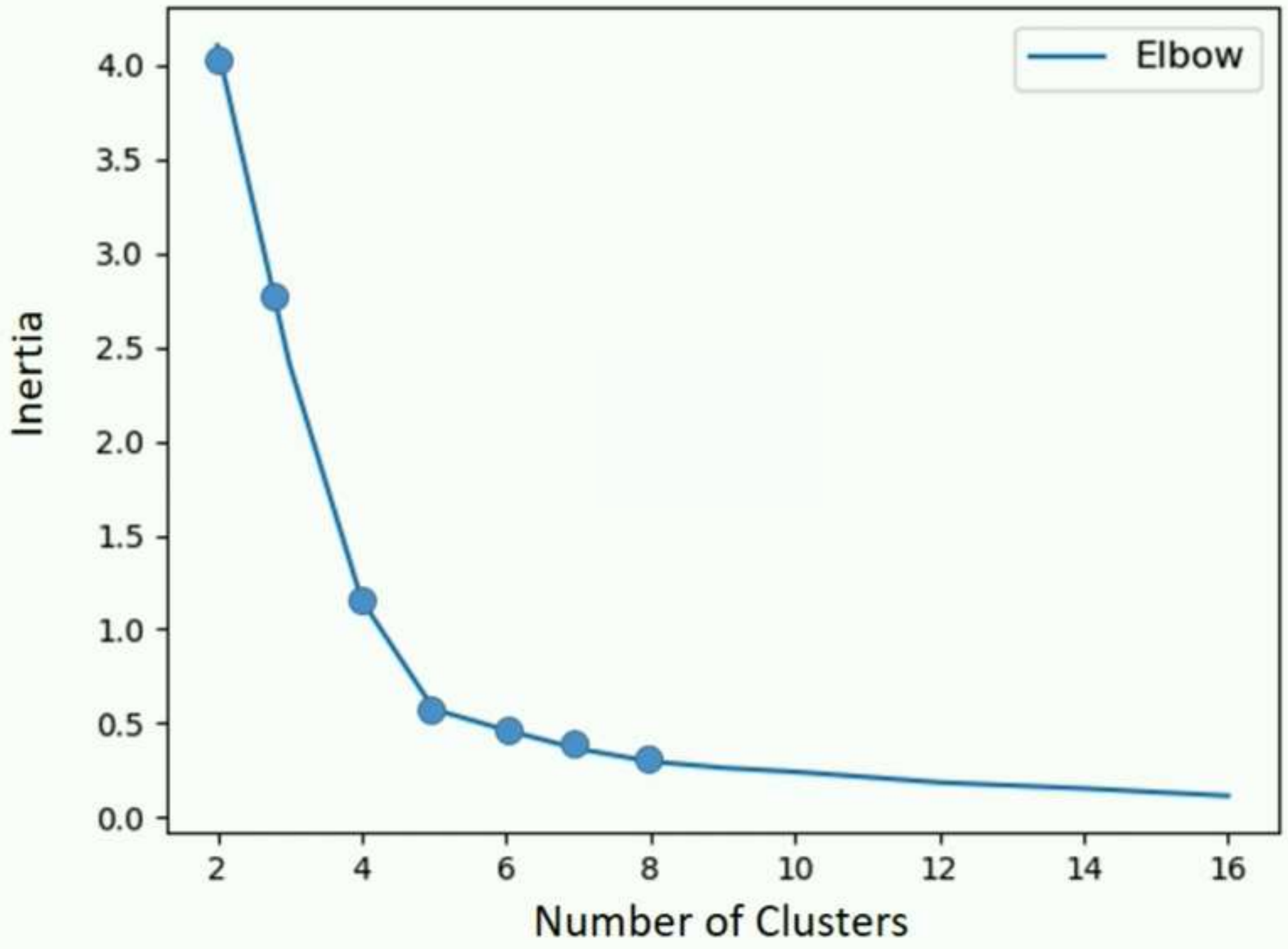
Distance of the observations from centroid decides their cluster assignment.

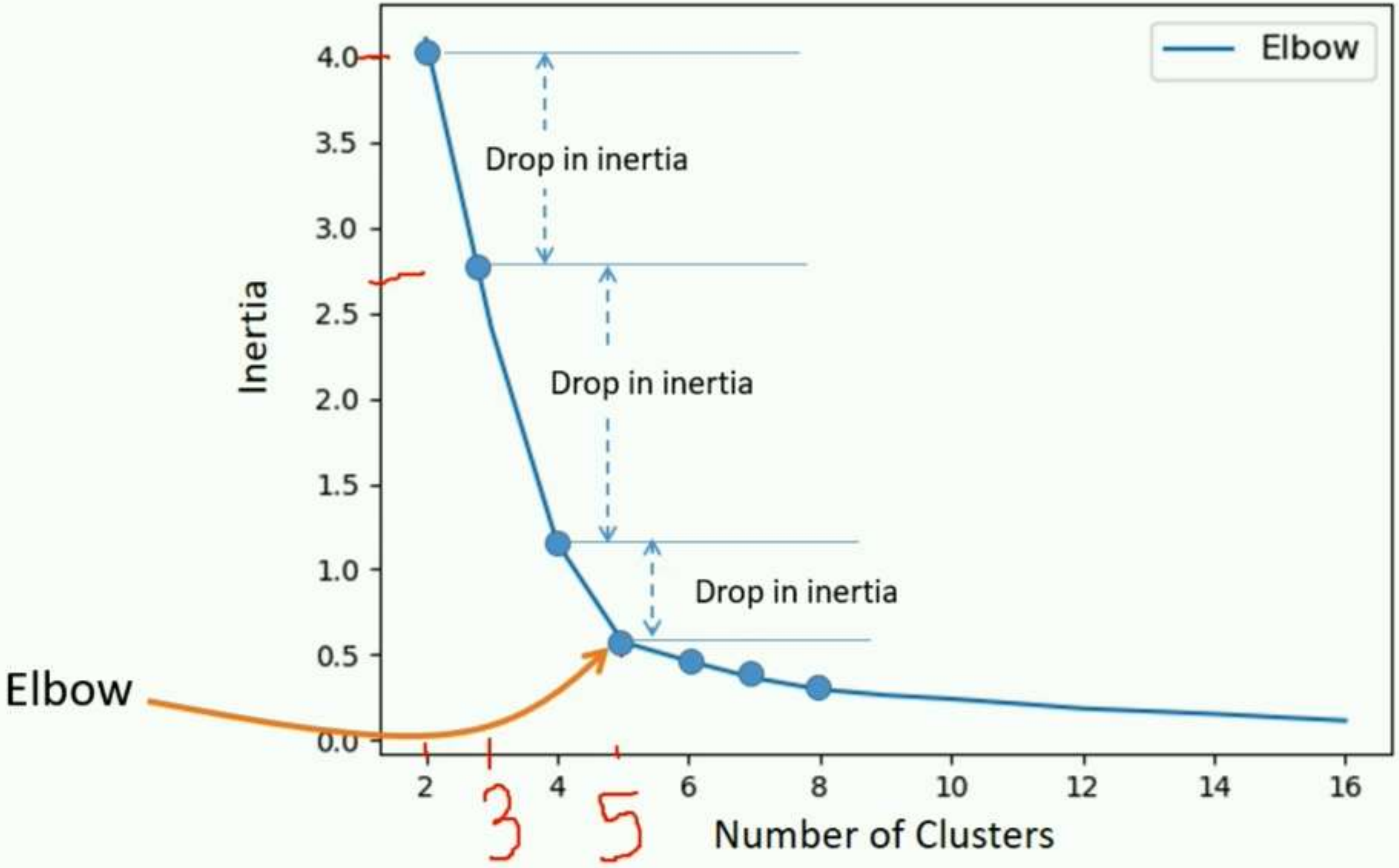
Lesser the distance,
Better the relationship

Higher number of clusters → Lesser Sum of Squared Distance (Intertia)

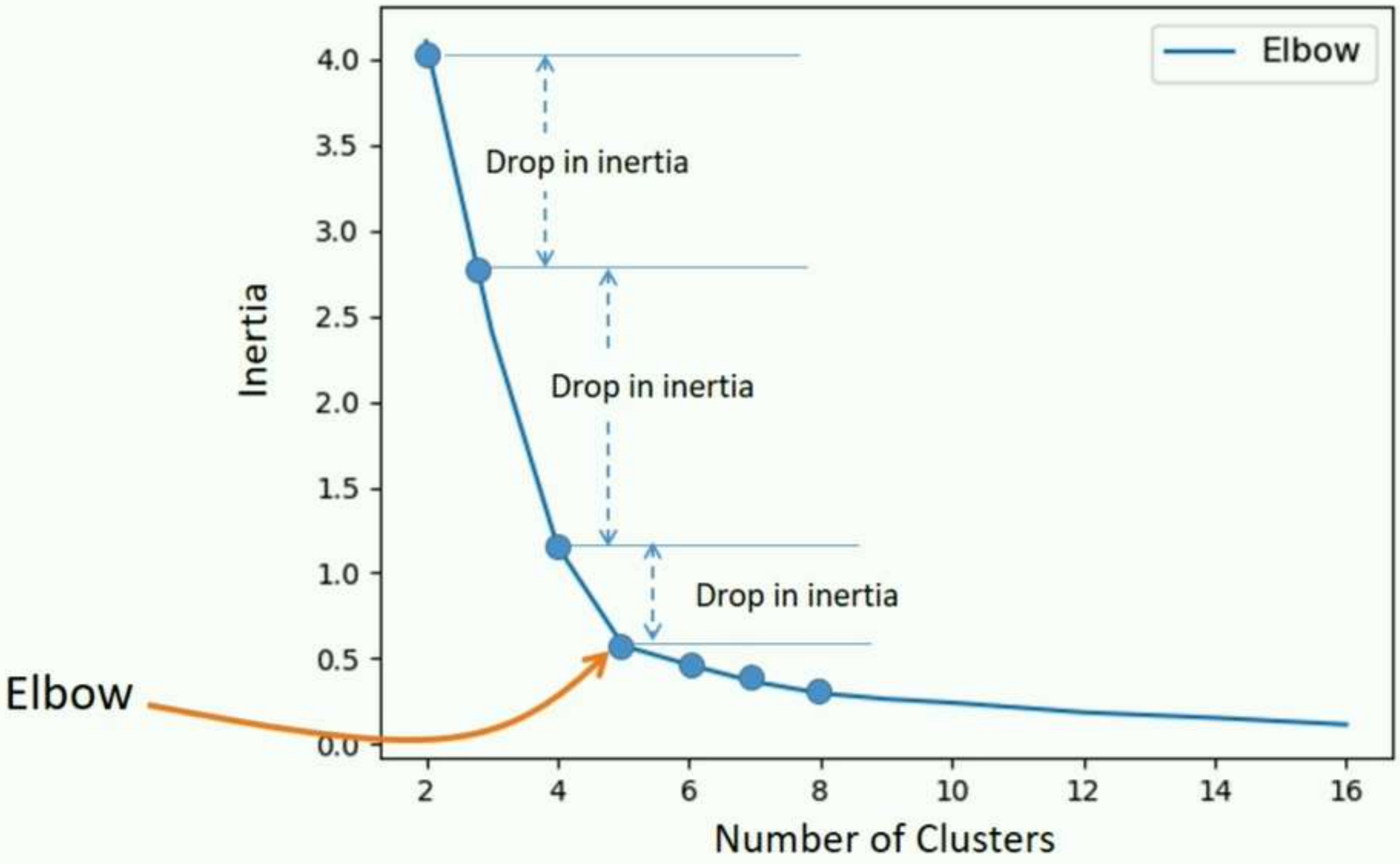





Elbow
Inertia
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2
4
6
8
10
12
14
16
Number of Clusters



Elbow
Drop in inertia
Drop in inertia
Drop in inertia
Inertia
Number of Clusters
Elbow
3
5


Elbow
Drop in inertia
Drop in inertia
Drop in inertia
Elbow
Inertia
Number of Clusters
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2
4
6
8
10
12
14
16

Demo: Use Elbow method to choose the right number of clusters.

Demo: Visualize Elbow graph

? Quiz

Question 1:

In the K-Means algorithm, we have to specify the number of clusters.

- True
- False

Question 2:

What metric can be used to find an optimal number of clusters ?

- ◯ R Squared
- ◯ MSE
- ◯ WCSS

Question 3:

We can choose any random initial centroids at the beginning of K-Means.

- True
- False

Question 4:

In Python, what is the recommended init parameter to input ?

- ◯ random
- ◯ k-means++
- ◯ inertia
- ◯ boost

Hierarchical Clustering

Agglomerative HC

STEP 1: Make each data point a single-point cluster ➡ That forms N clusters

⬇

STEP 2: Take the two closest data points and make them one cluster ➡ That forms N-1 clusters

⬇

STEP 3: Take the two closest clusters and make them one cluster ➡ That forms N - 2 clusters

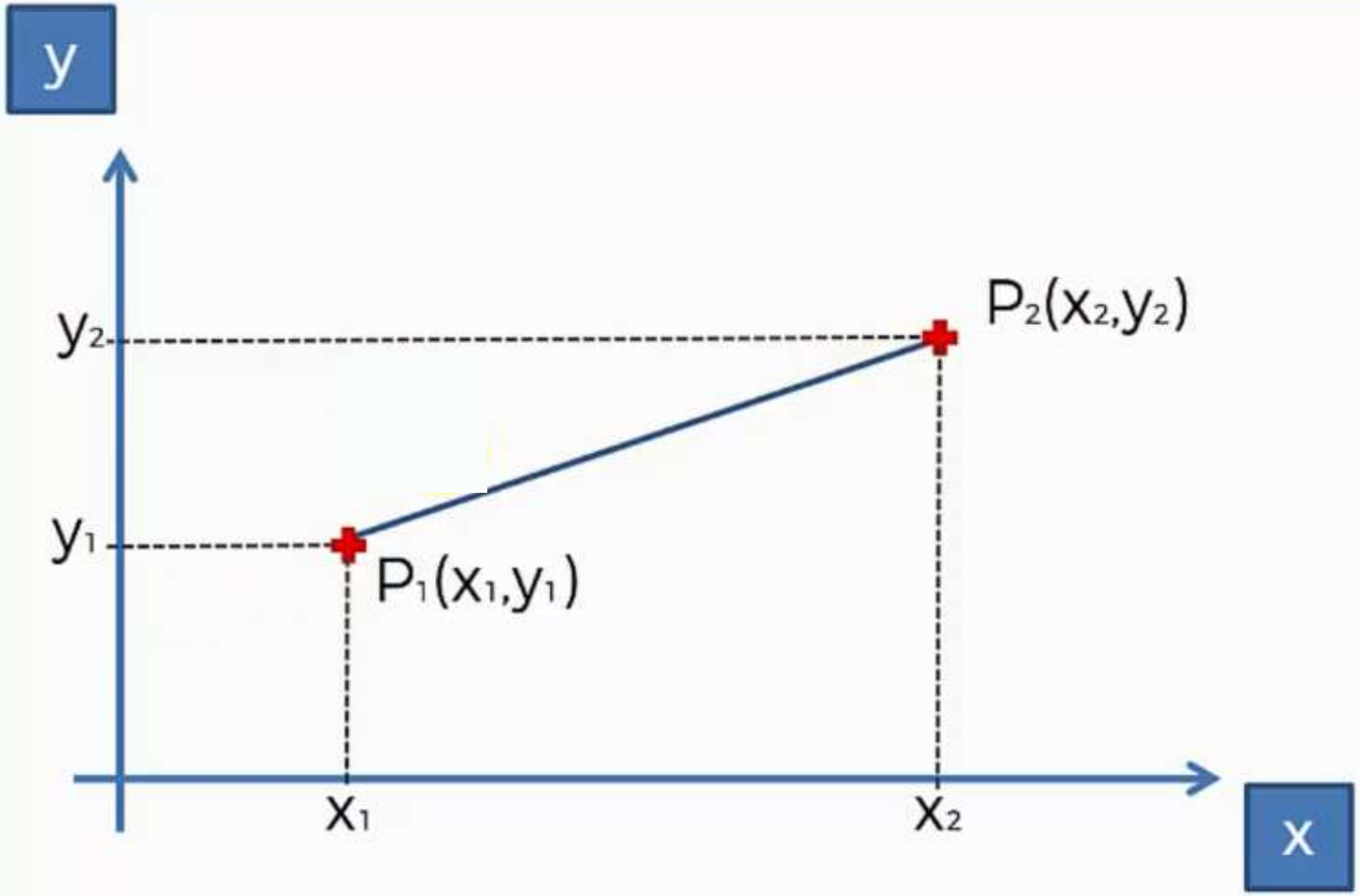
⬇

STEP 4: Repeat **STEP 3** until there is only one cluster

⬇

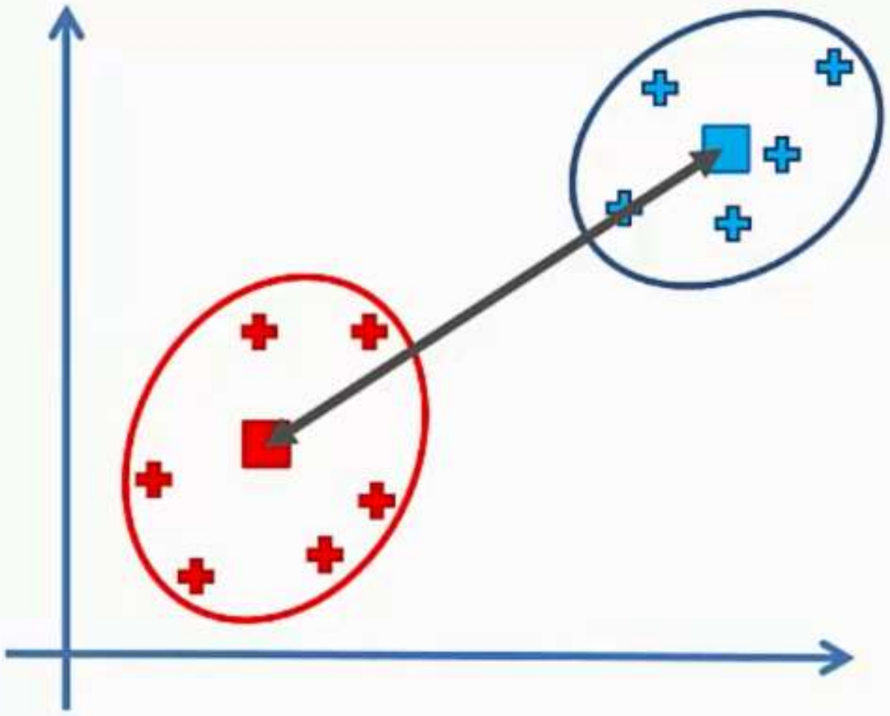
FIN

Euclidean Distance



$$\text{Euclidean Distance between } \mathrm{P}_1 \text{ and } \mathrm{P}_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Distance Between Clusters

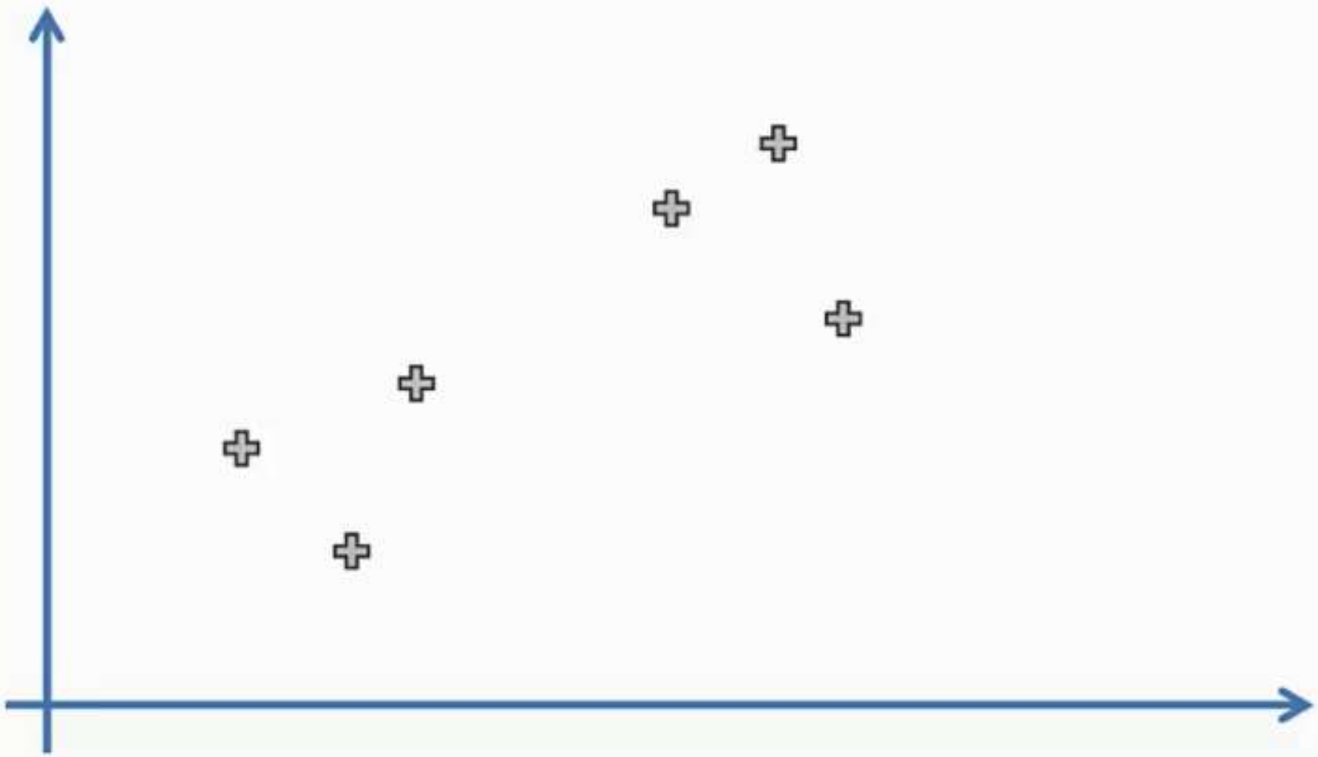


Distance Between Two Clusters:

- Option 1: Closest Points
- Option 2: Furthest Points
- Option 3: Average Distance
- Option 4: Distance Between Centroids

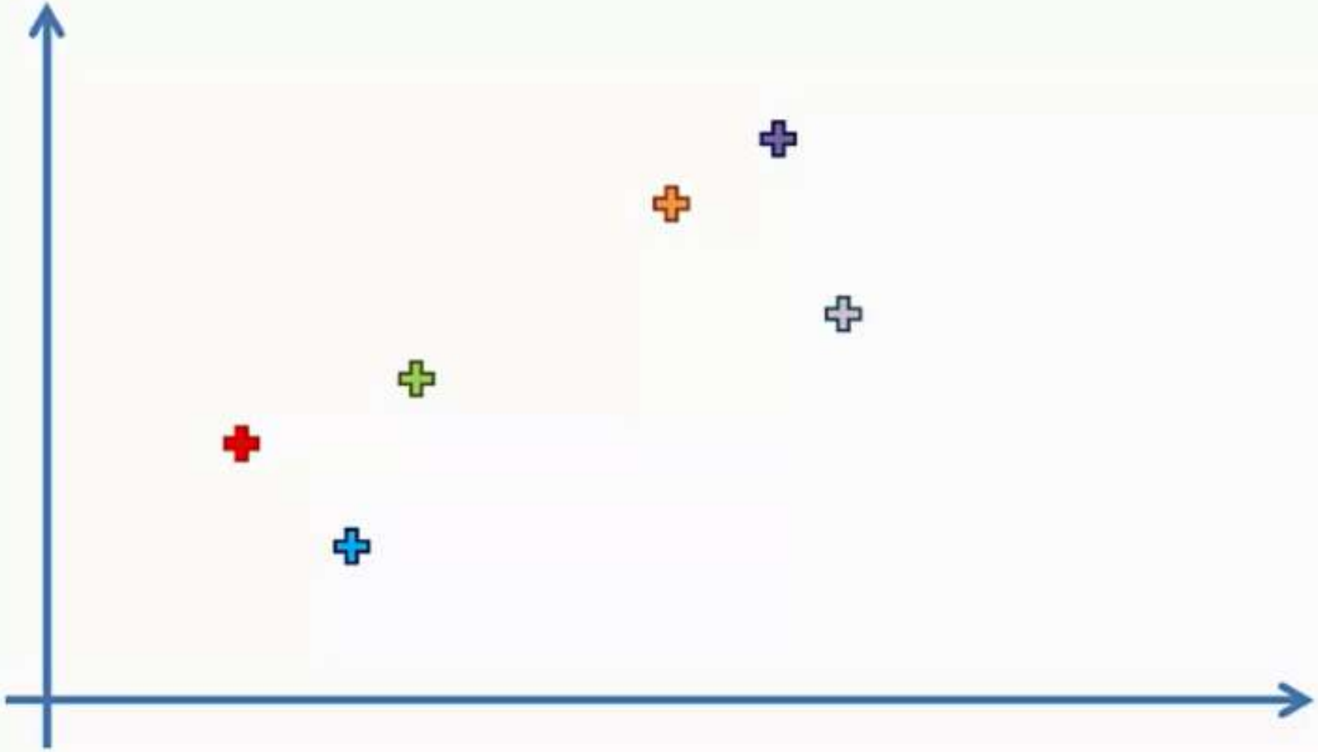
Agglomerative HC

STEP 1: Make each data point a single-point cluster ➡ That forms 6 clusters



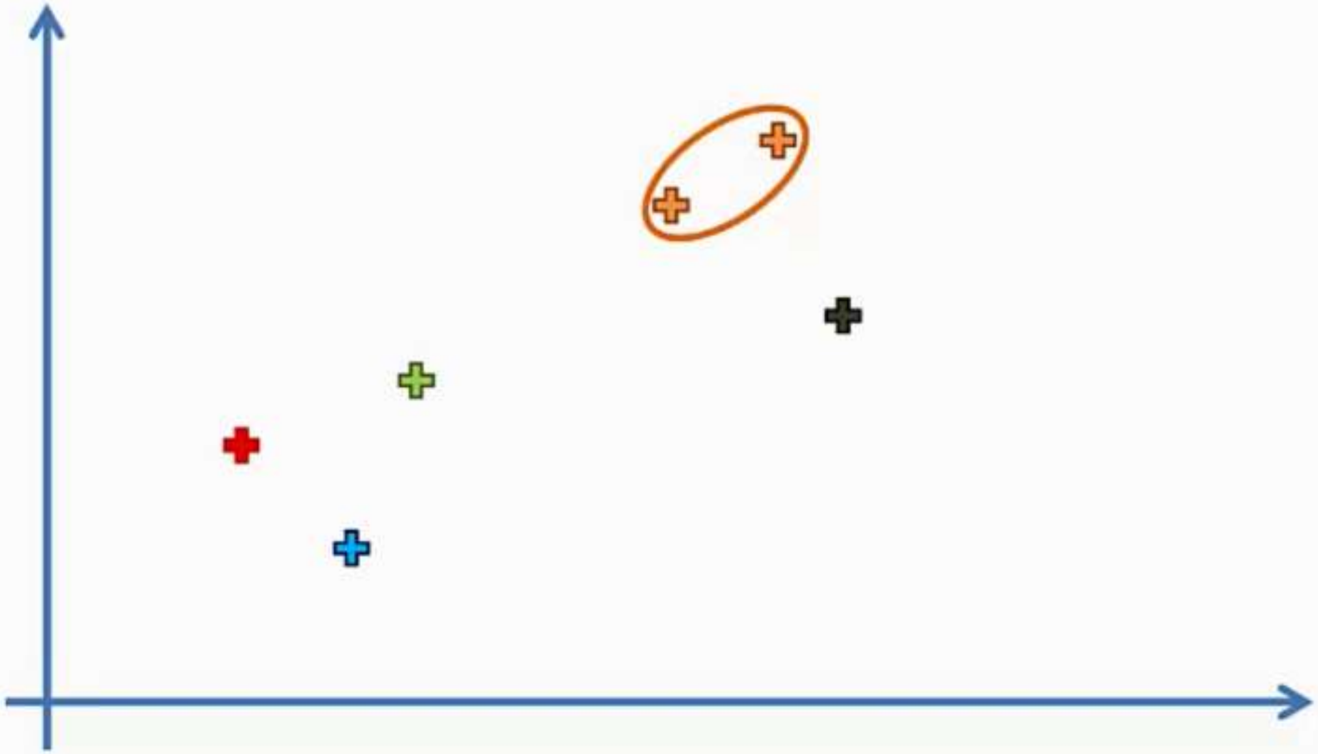
Agglomerative HC

STEP 1: Make each data point a single-point cluster ➡ That forms 6 clusters



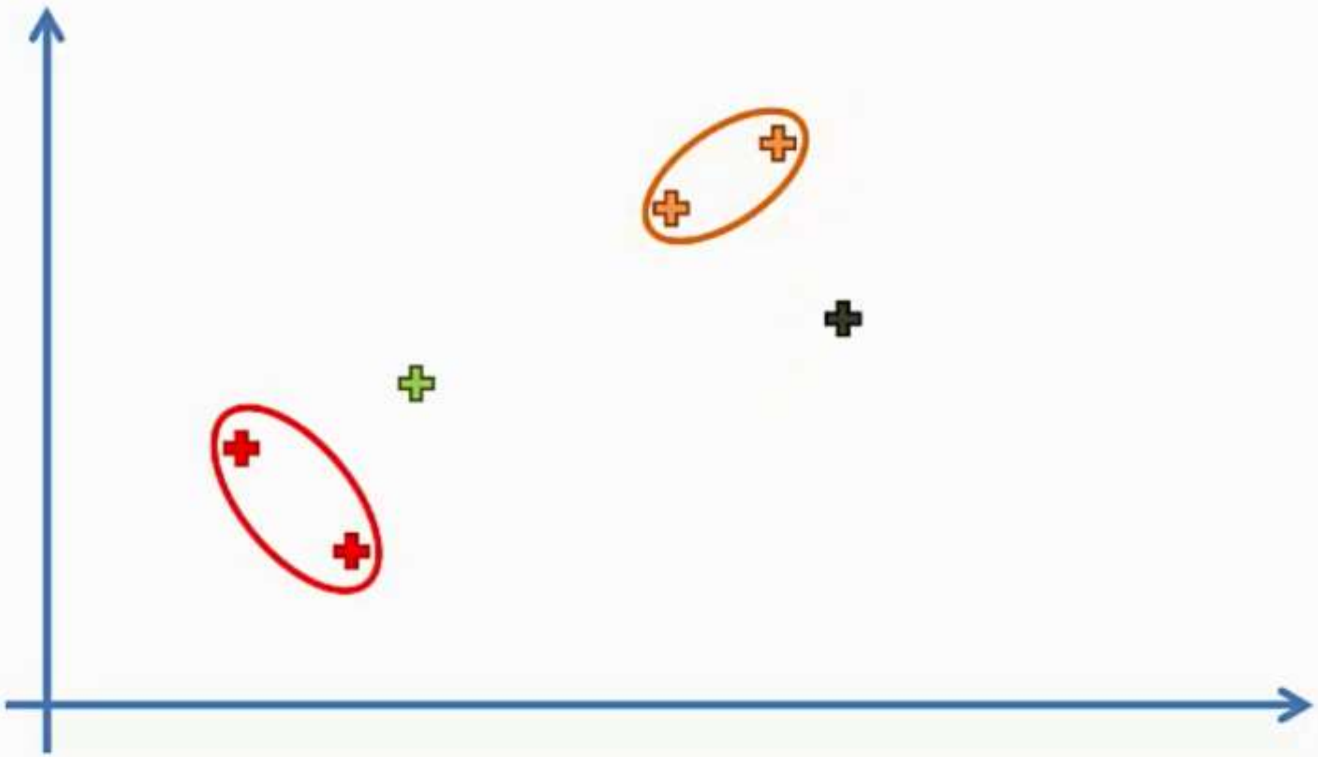
Agglomerative HC

STEP 2: Take the two closest data points and make them one cluster
➡ That forms 5 clusters



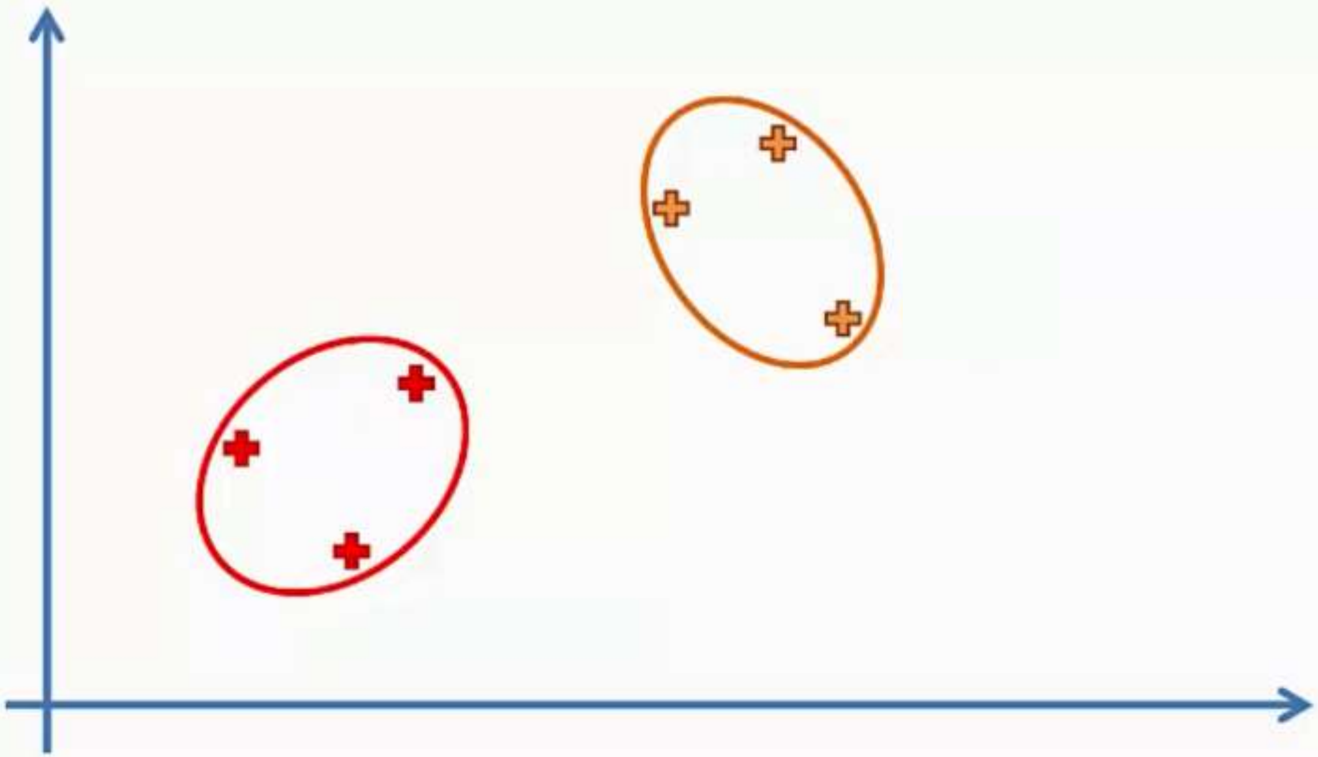
Agglomerative HC

STEP 3: Take the two closest clusters and make them one cluster
➡ That forms 4 clusters



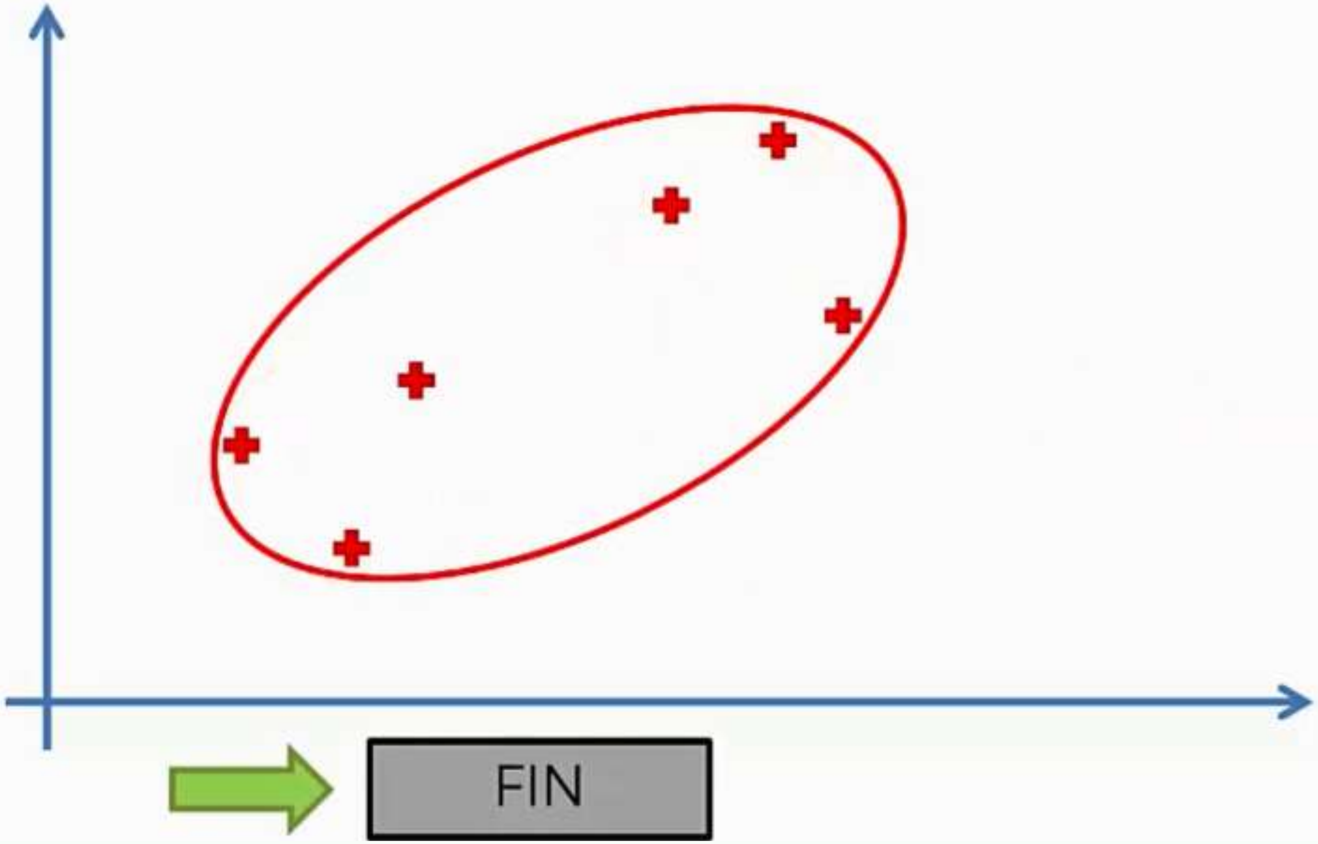
Agglomerative HC

STEP 4: Repeat STEP 3 until there is only one cluster

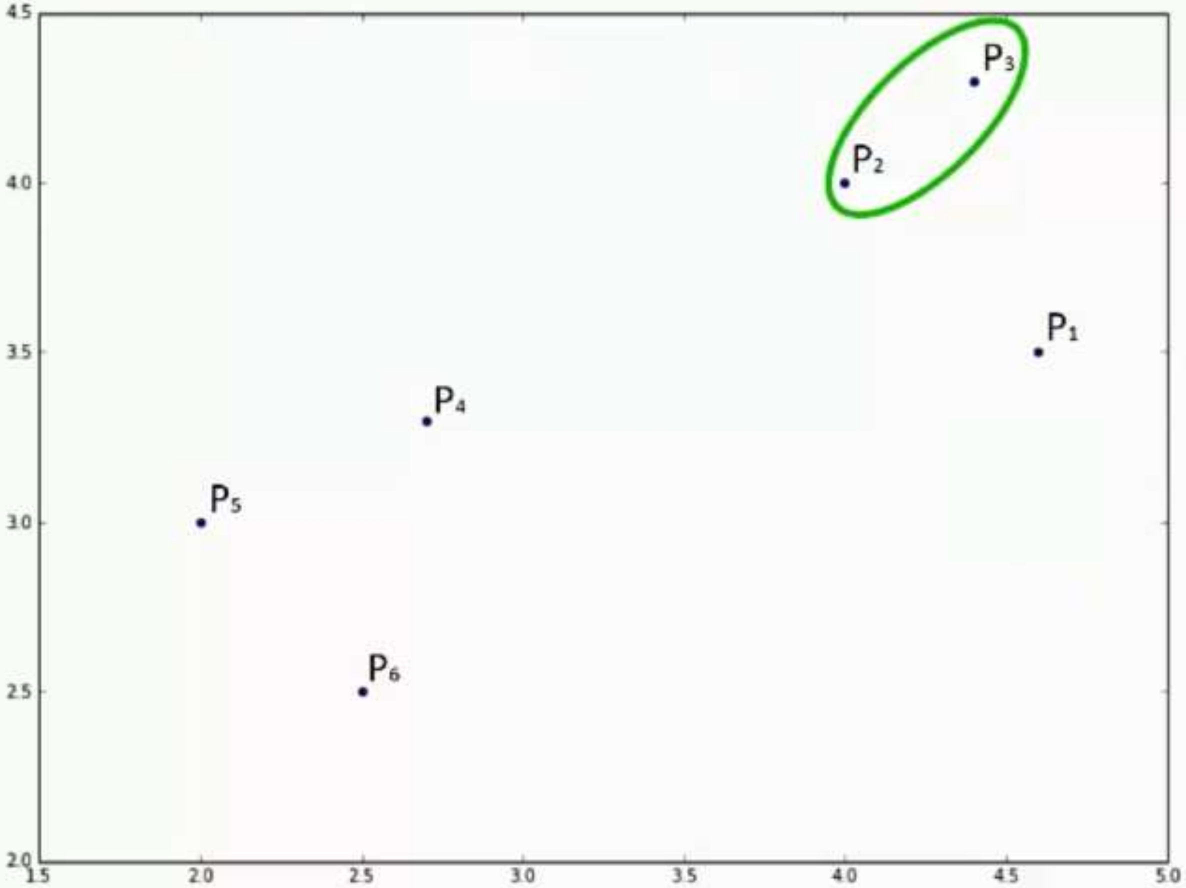


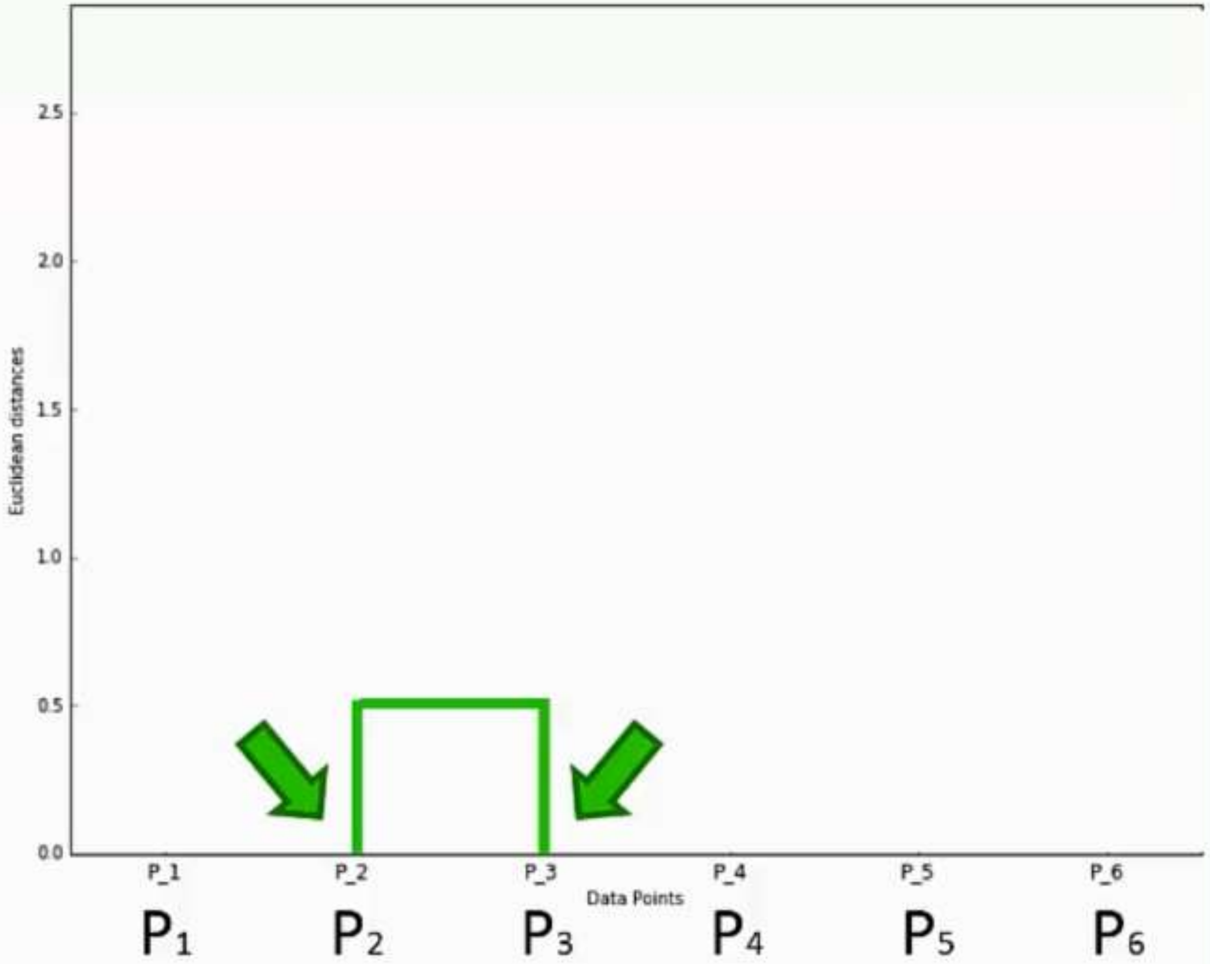
Agglomerative HC

STEP 4: Repeat STEP 3 until there is only one cluster

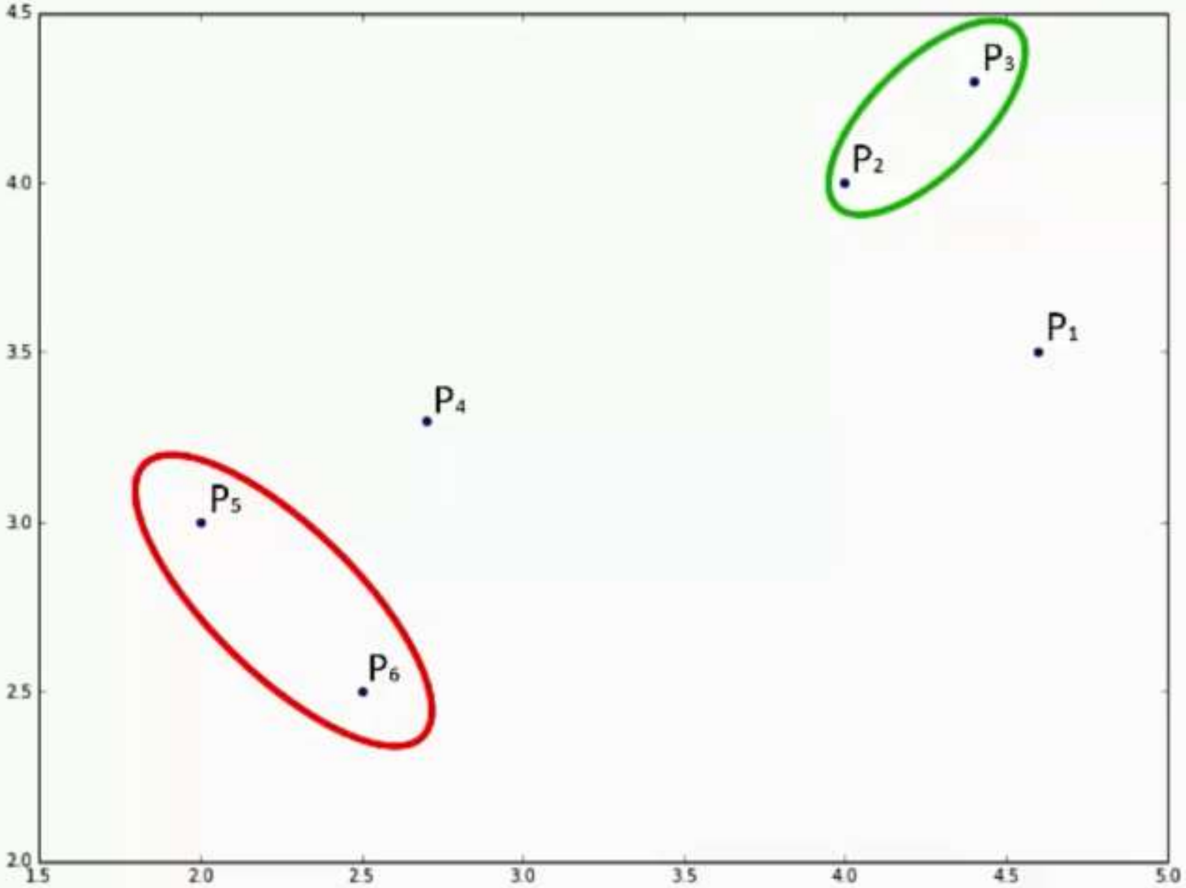


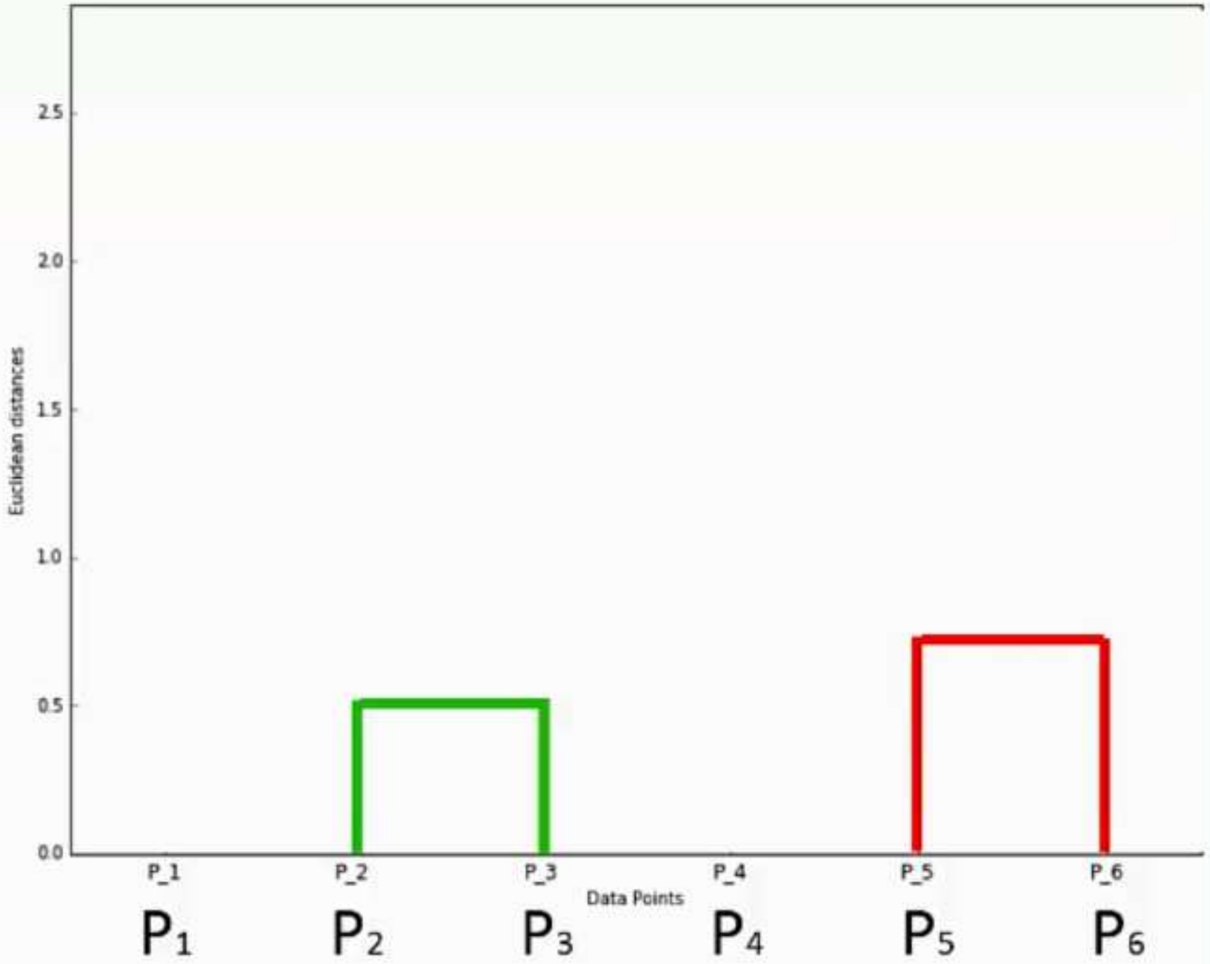
How Do Dendograms Work?



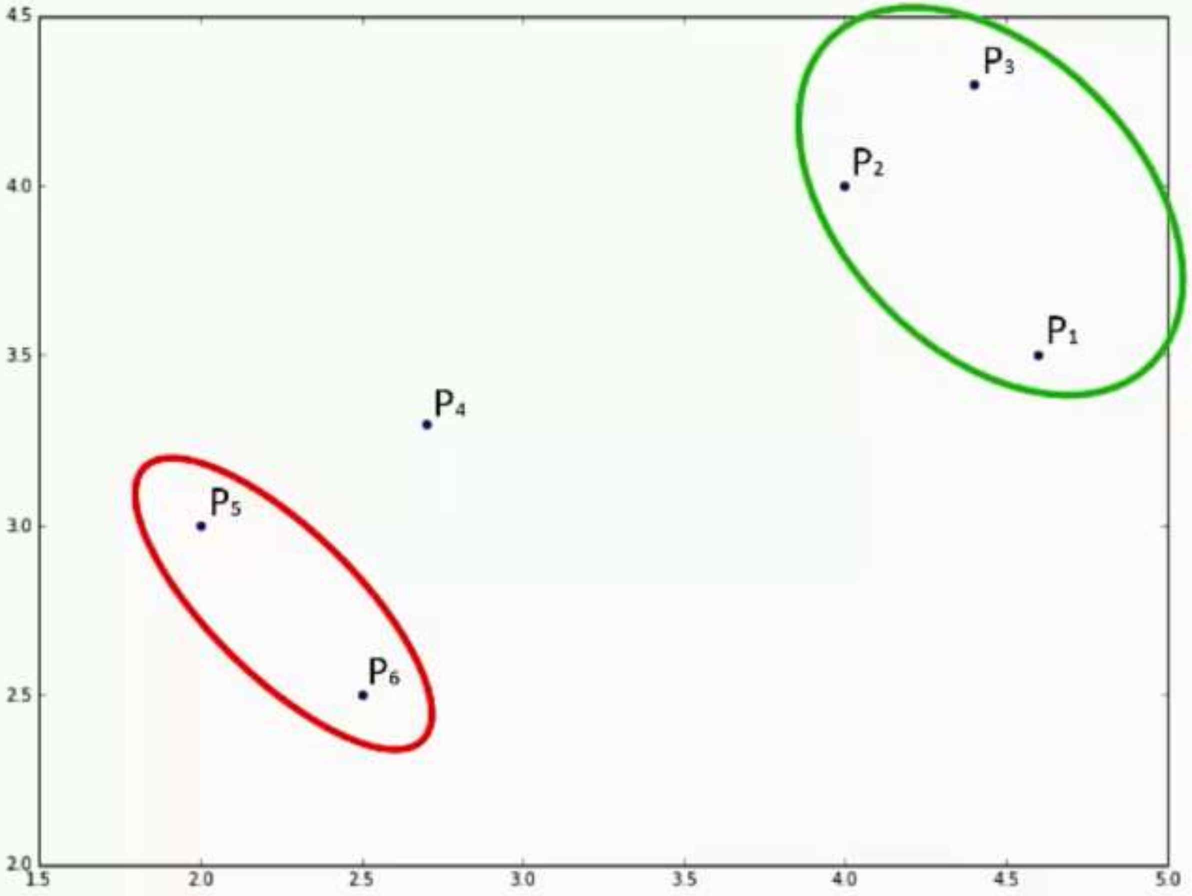


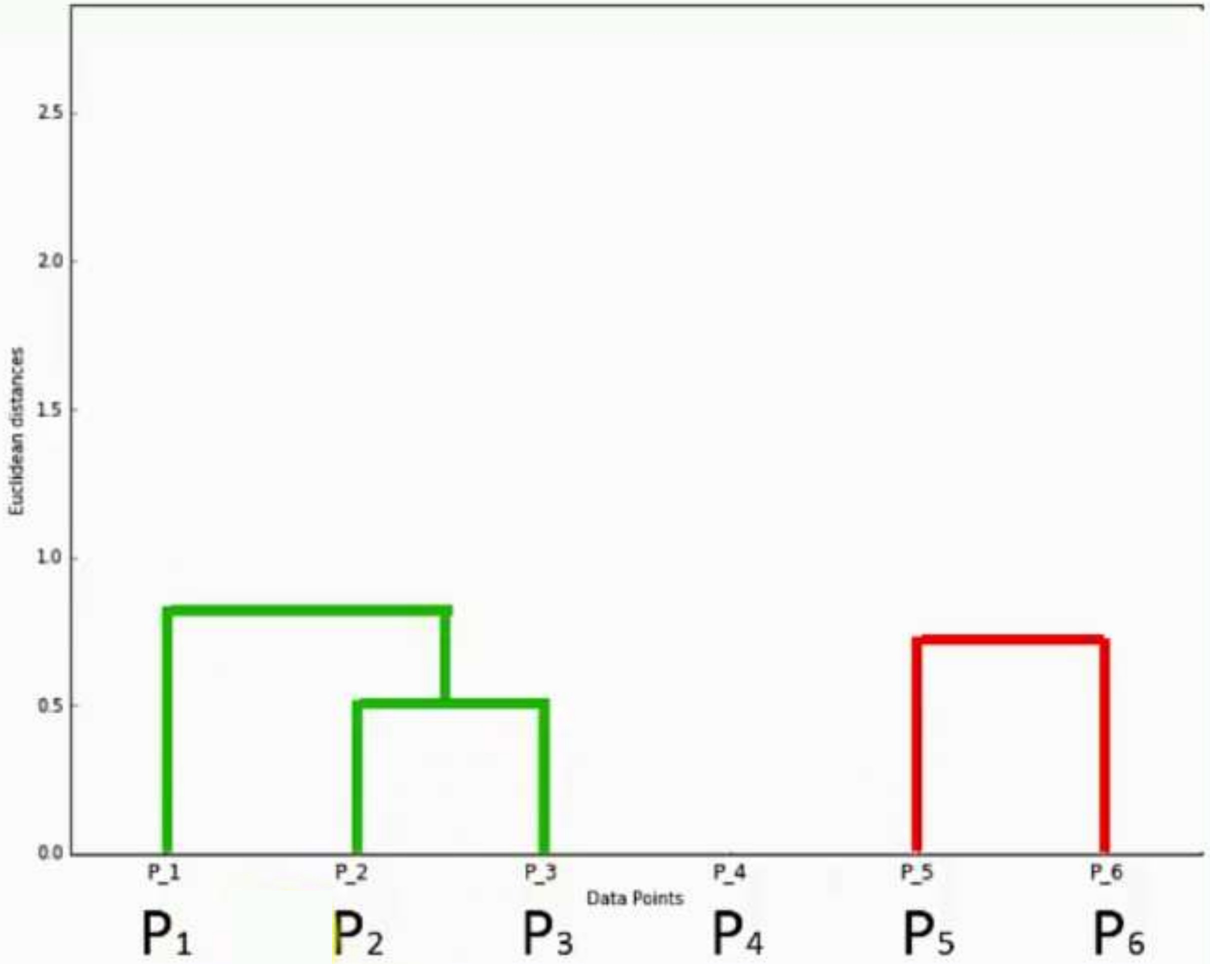
How Do Dendograms Work?



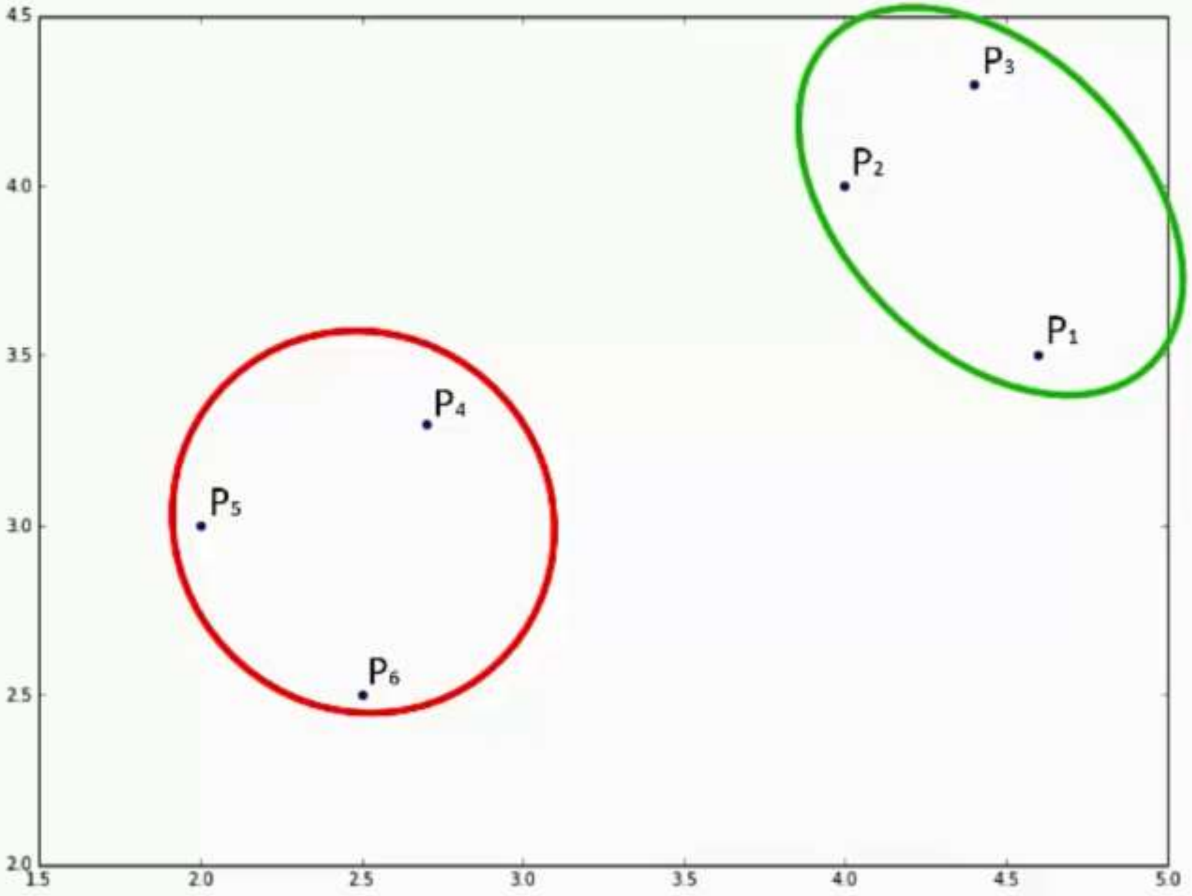


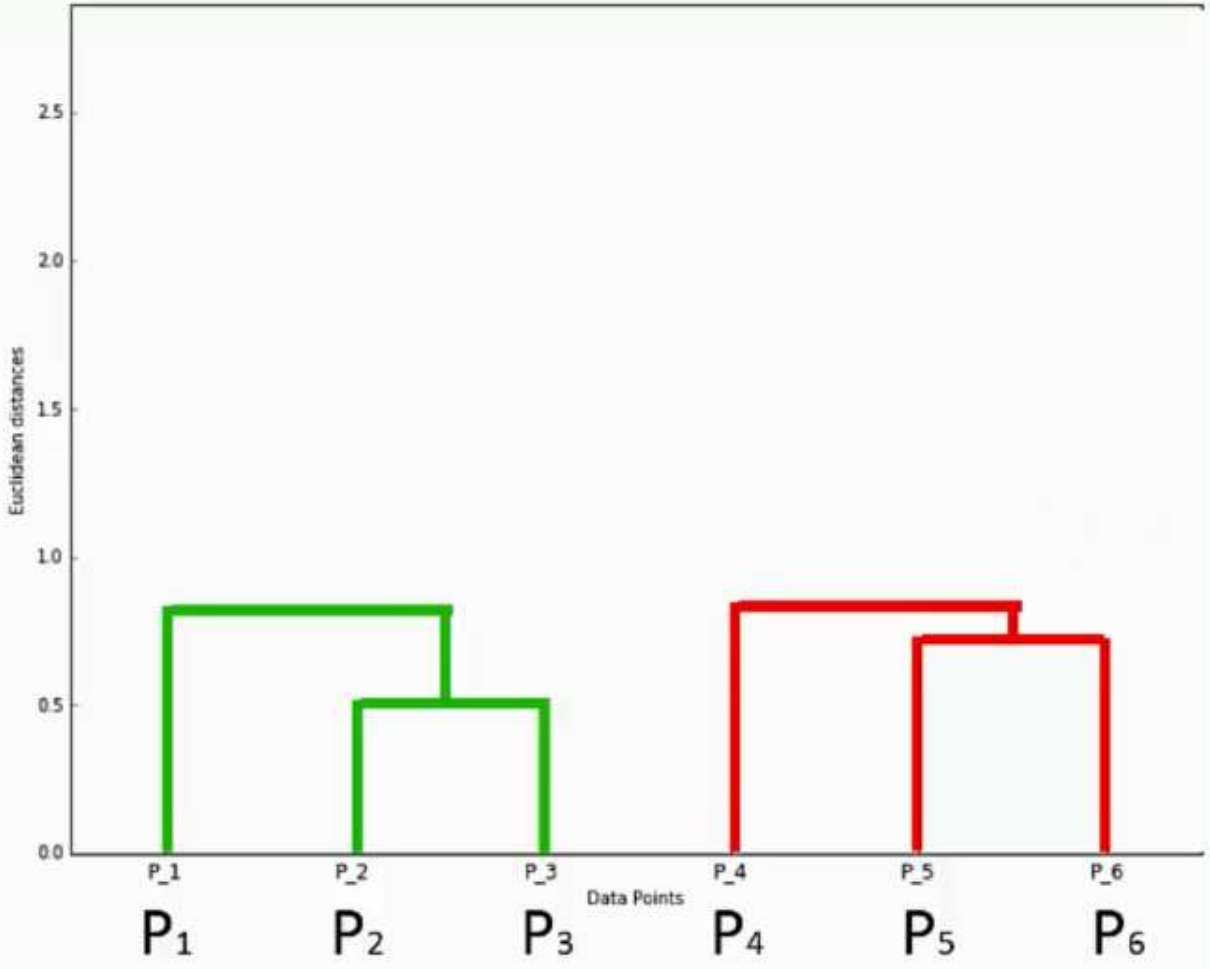
How Do Dendograms Work?



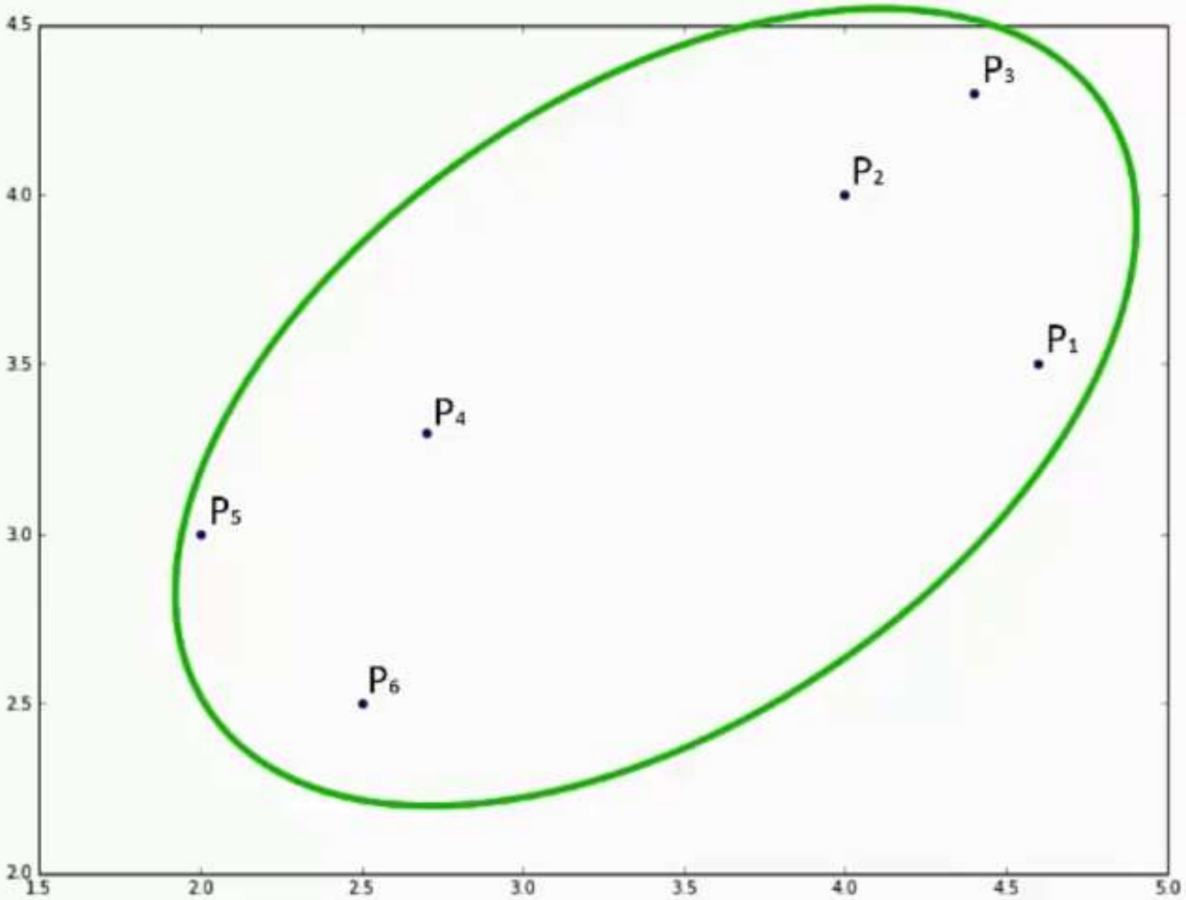


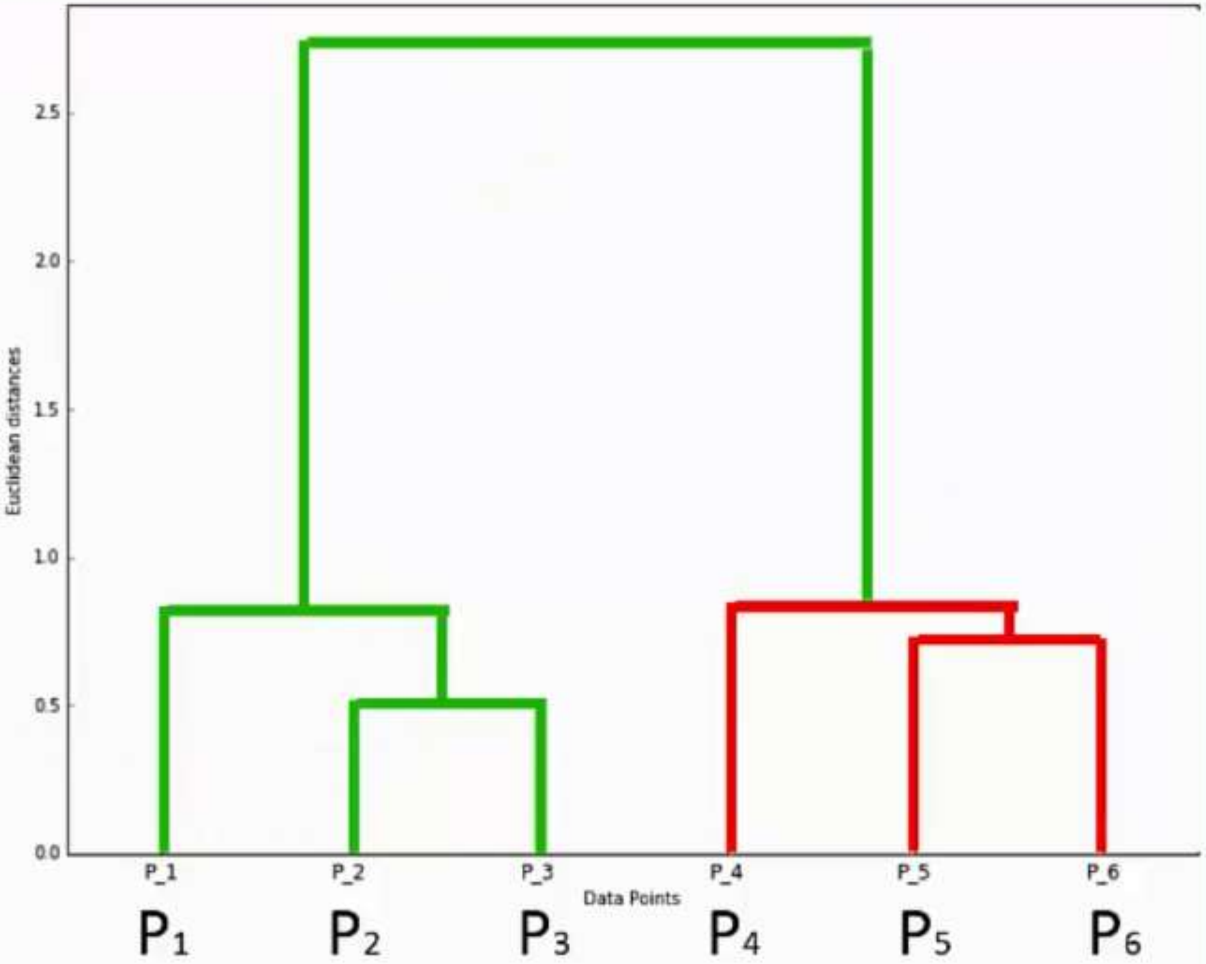
How Do Dendograms Work?



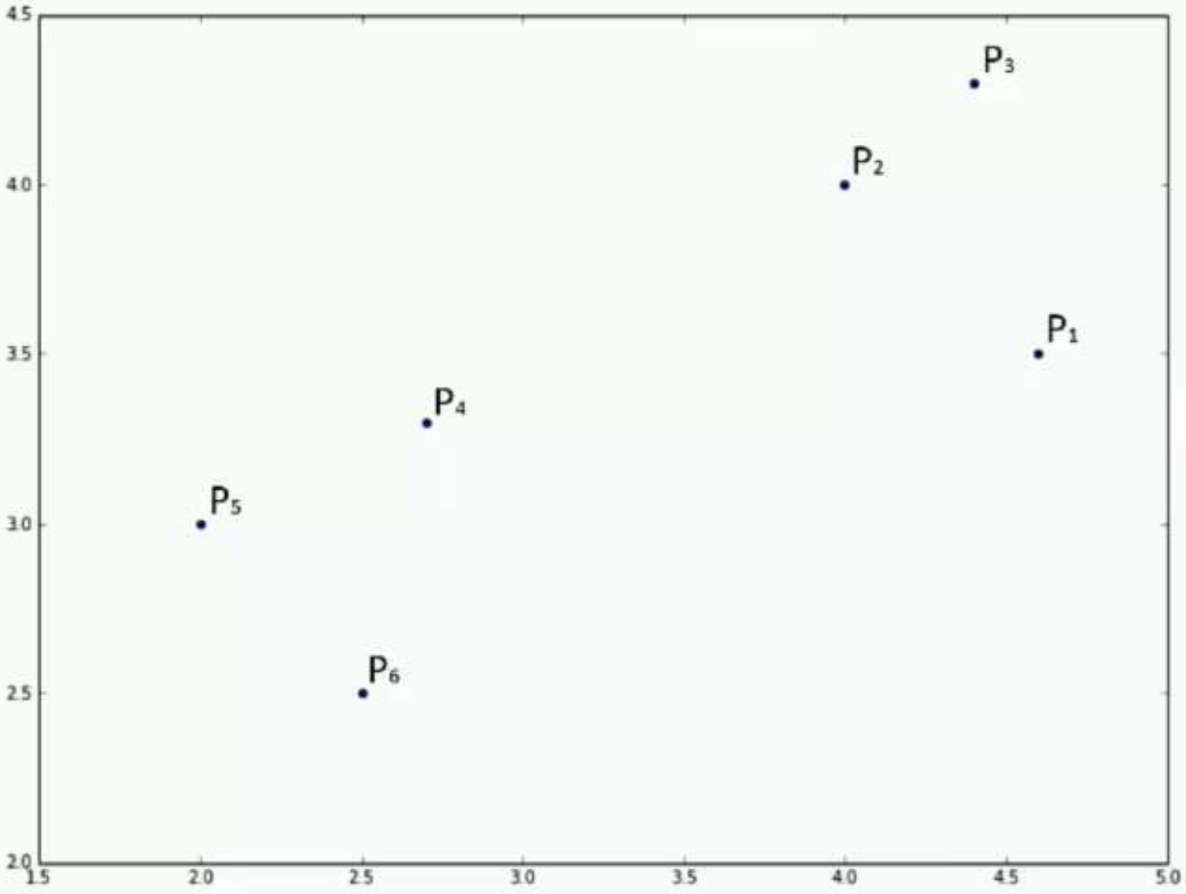


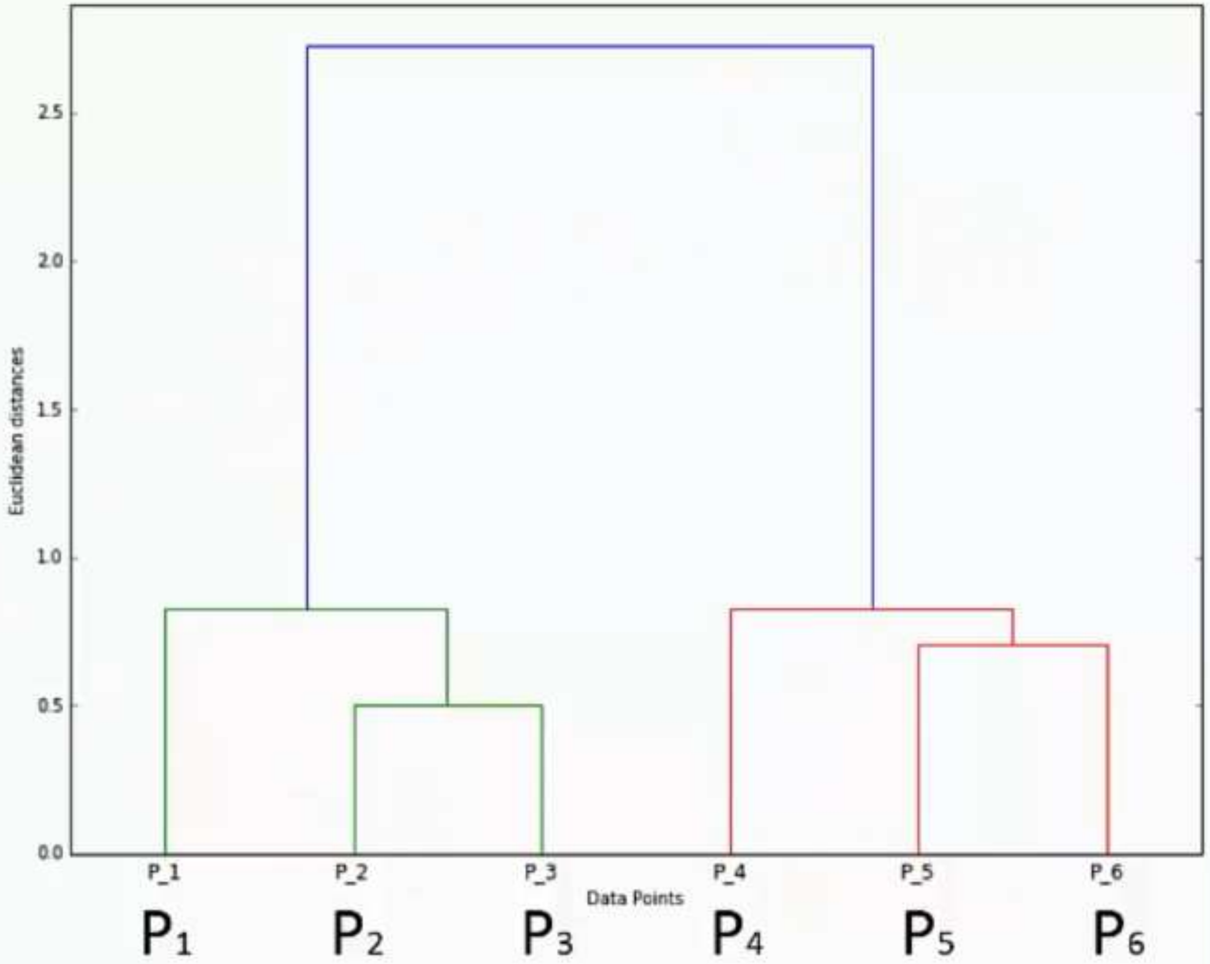
How Do Dendograms Work?



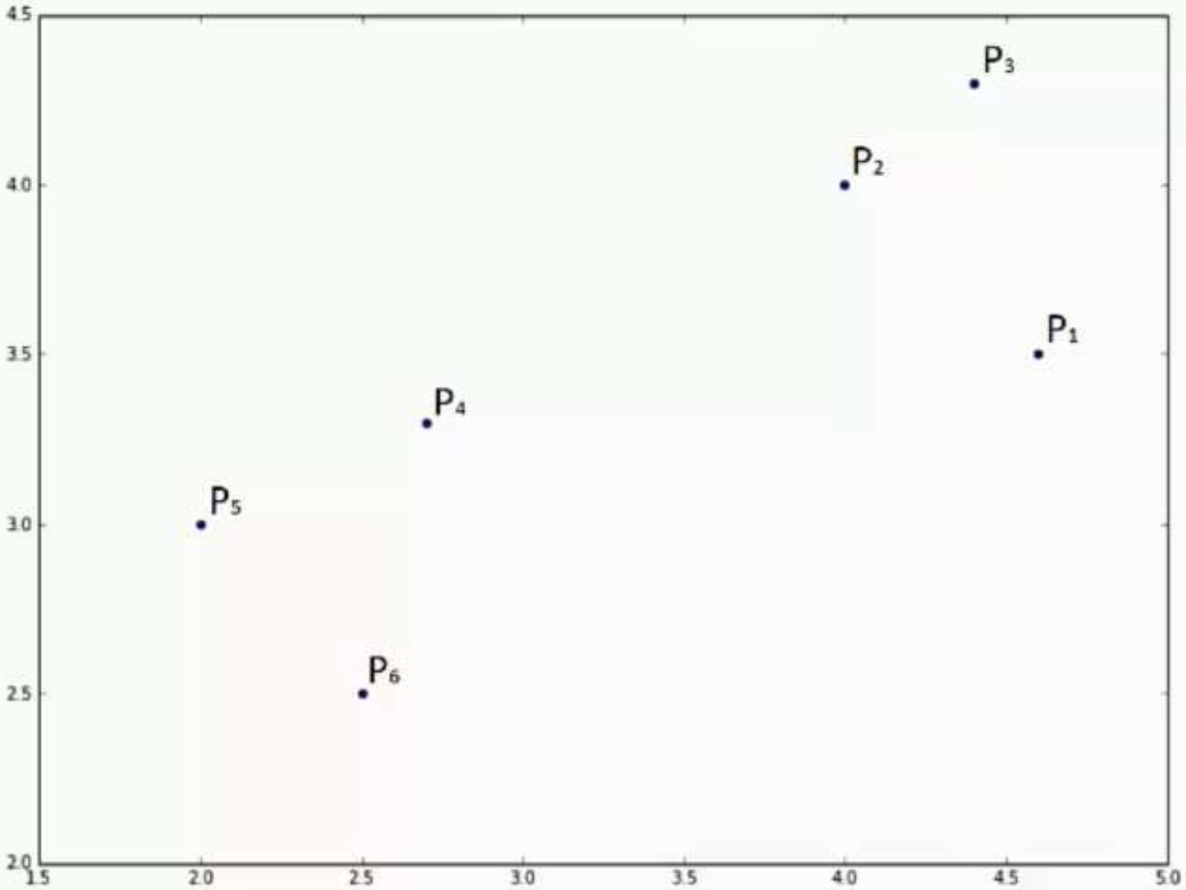


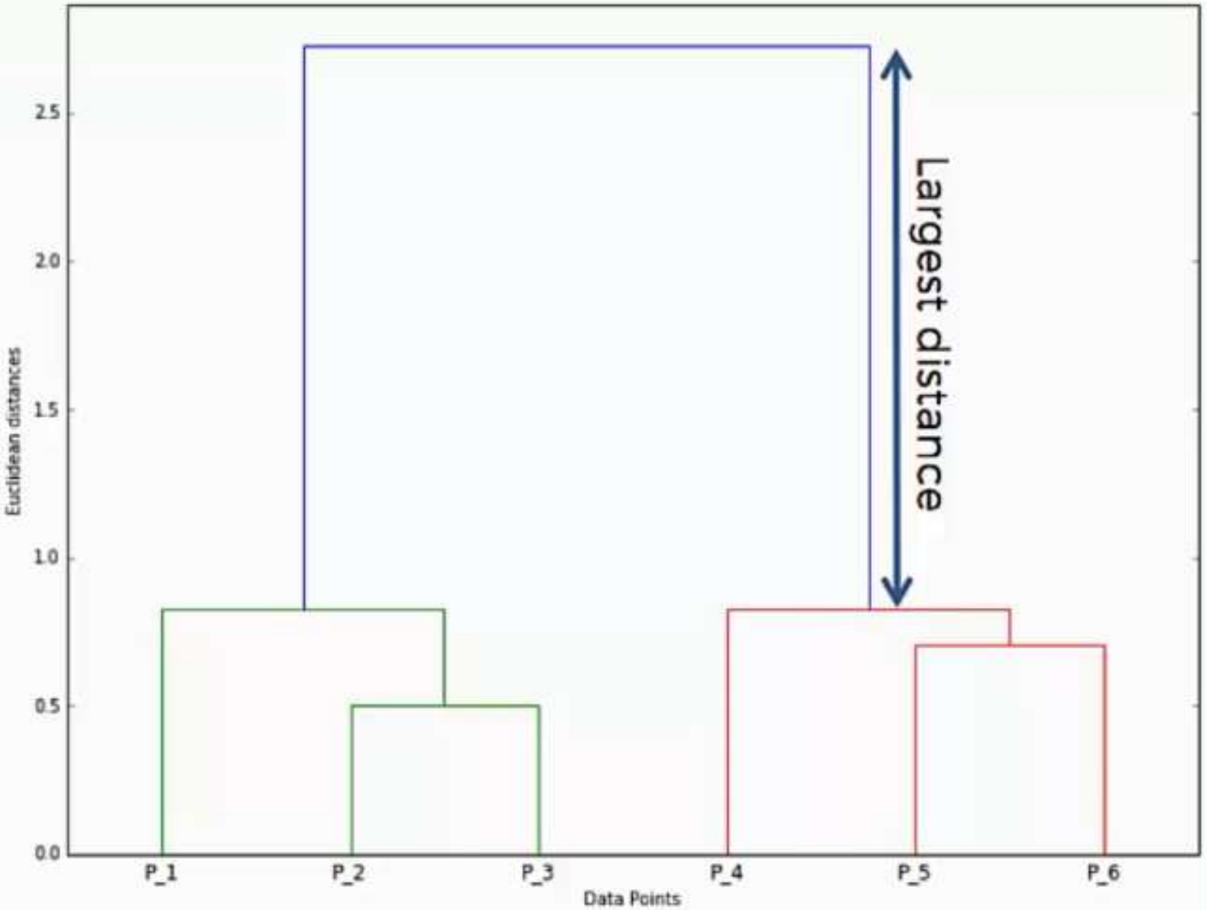
How Do Dendograms Work?



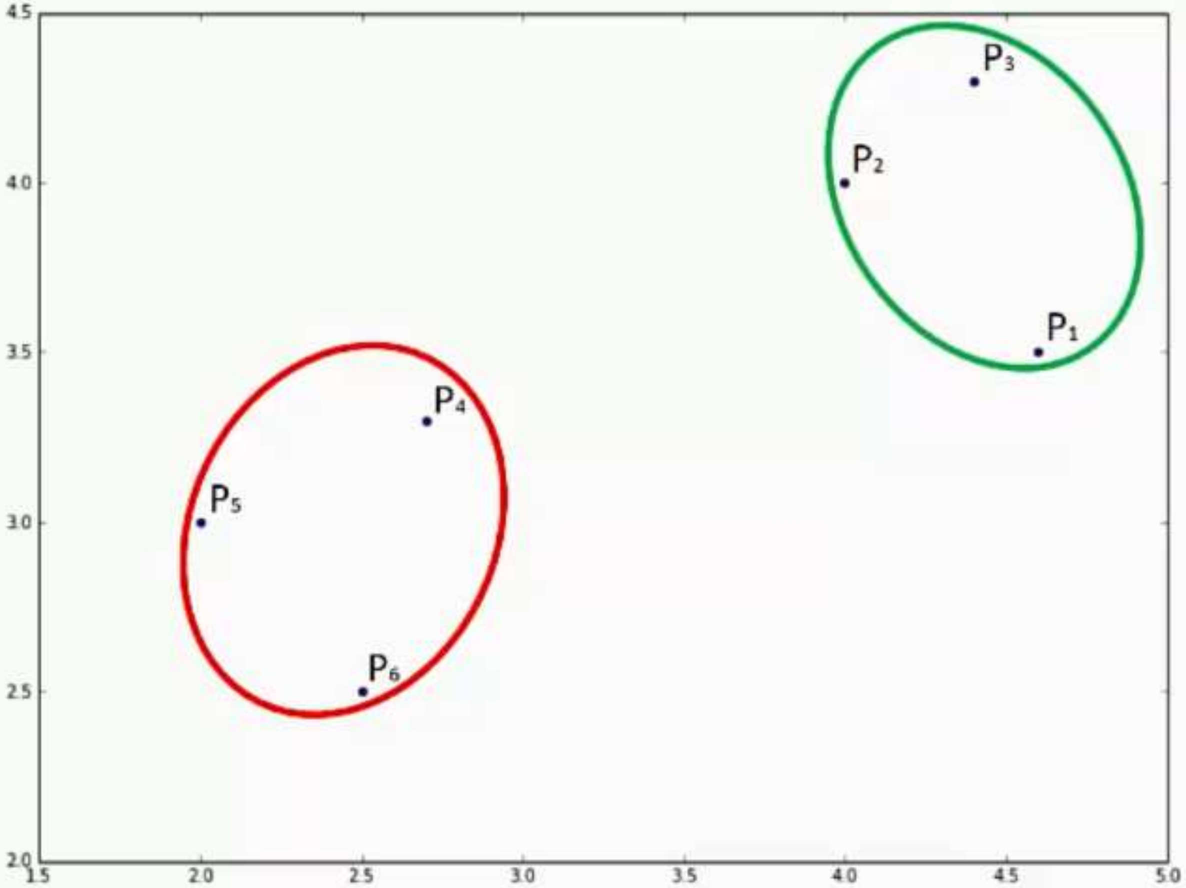


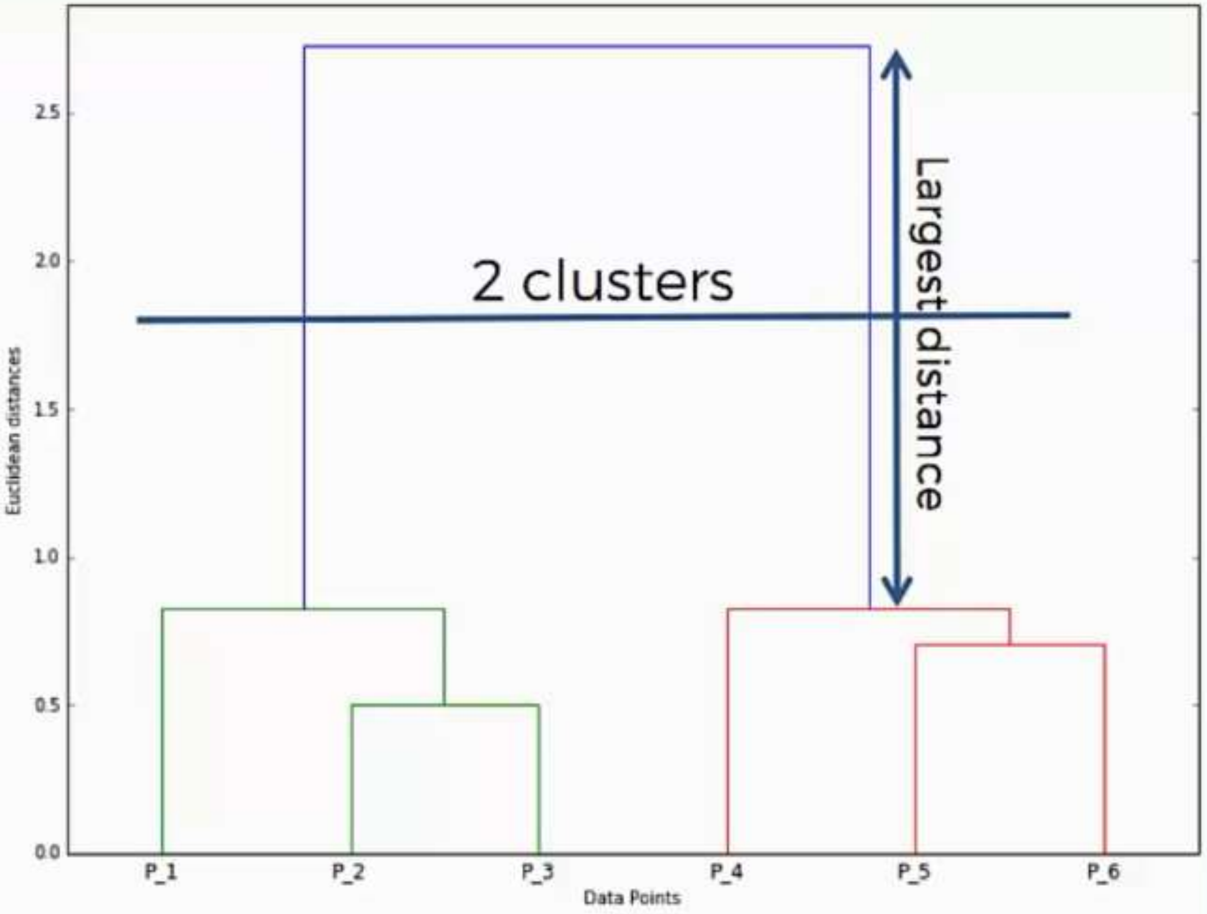
Dendrograms – Optimal # of Clusters



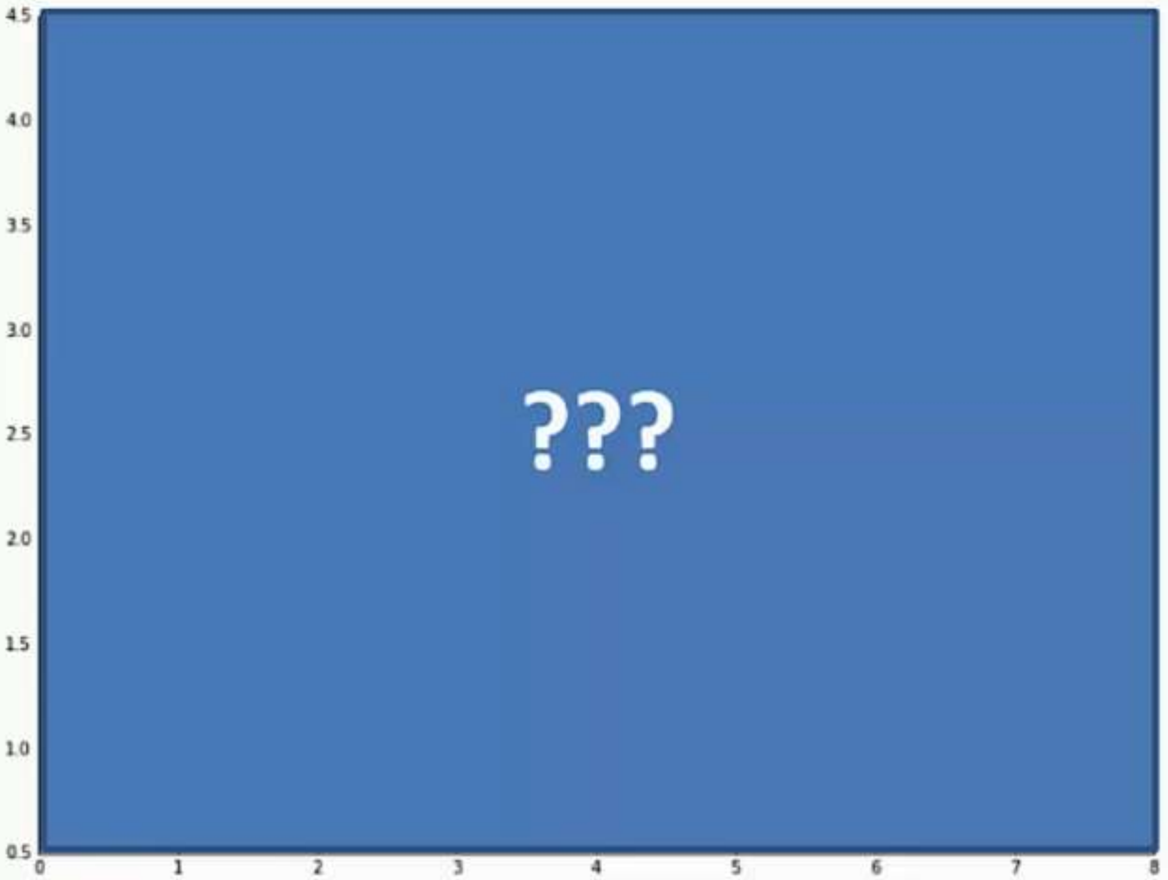


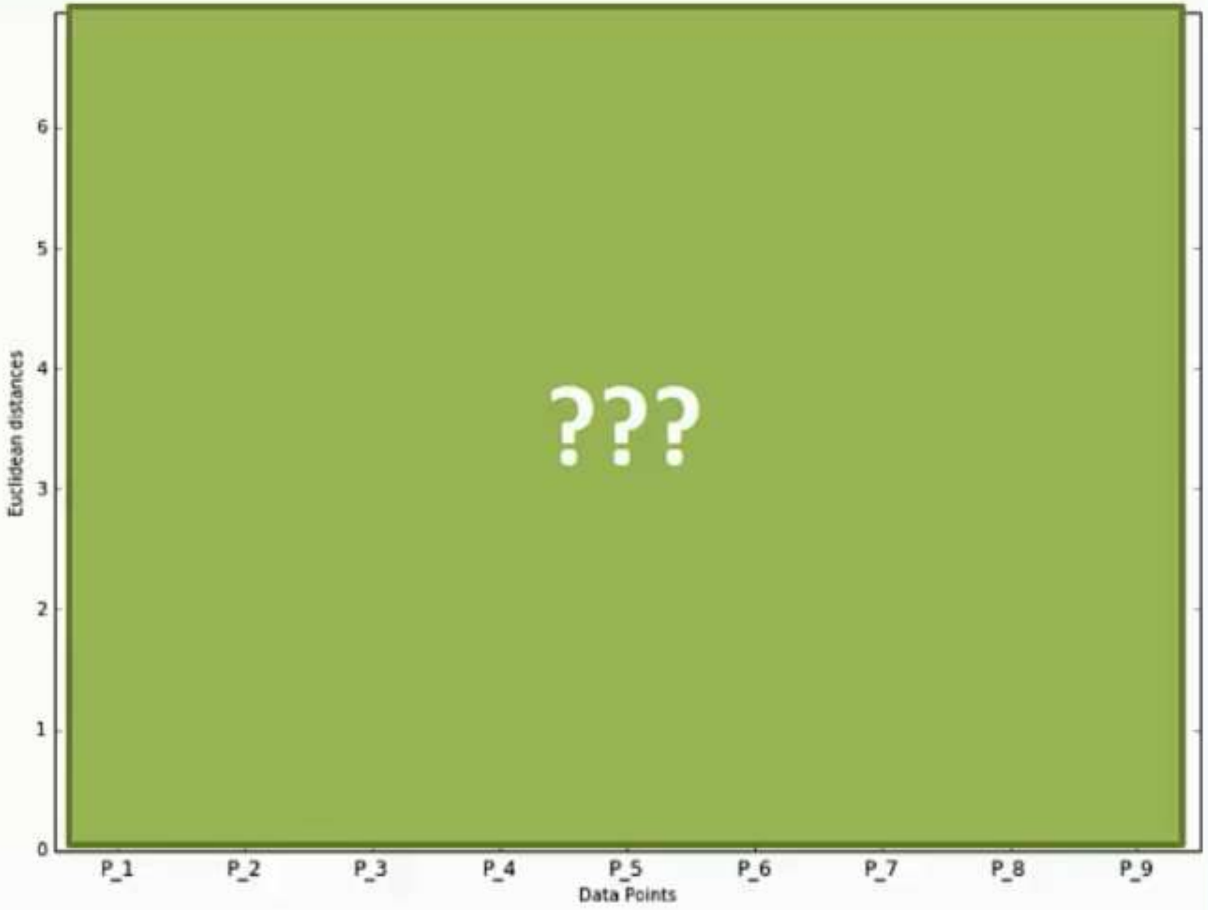
Dendrograms – Optimal # of Clusters



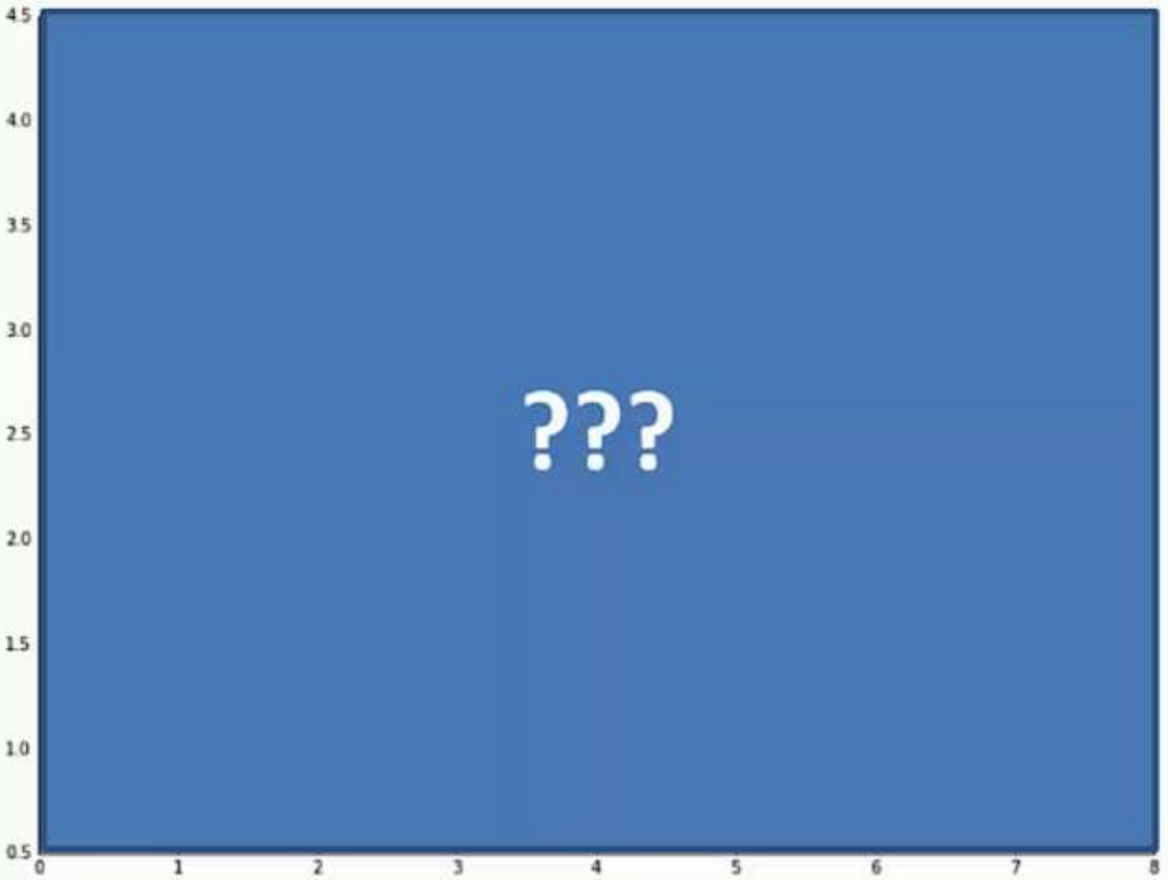


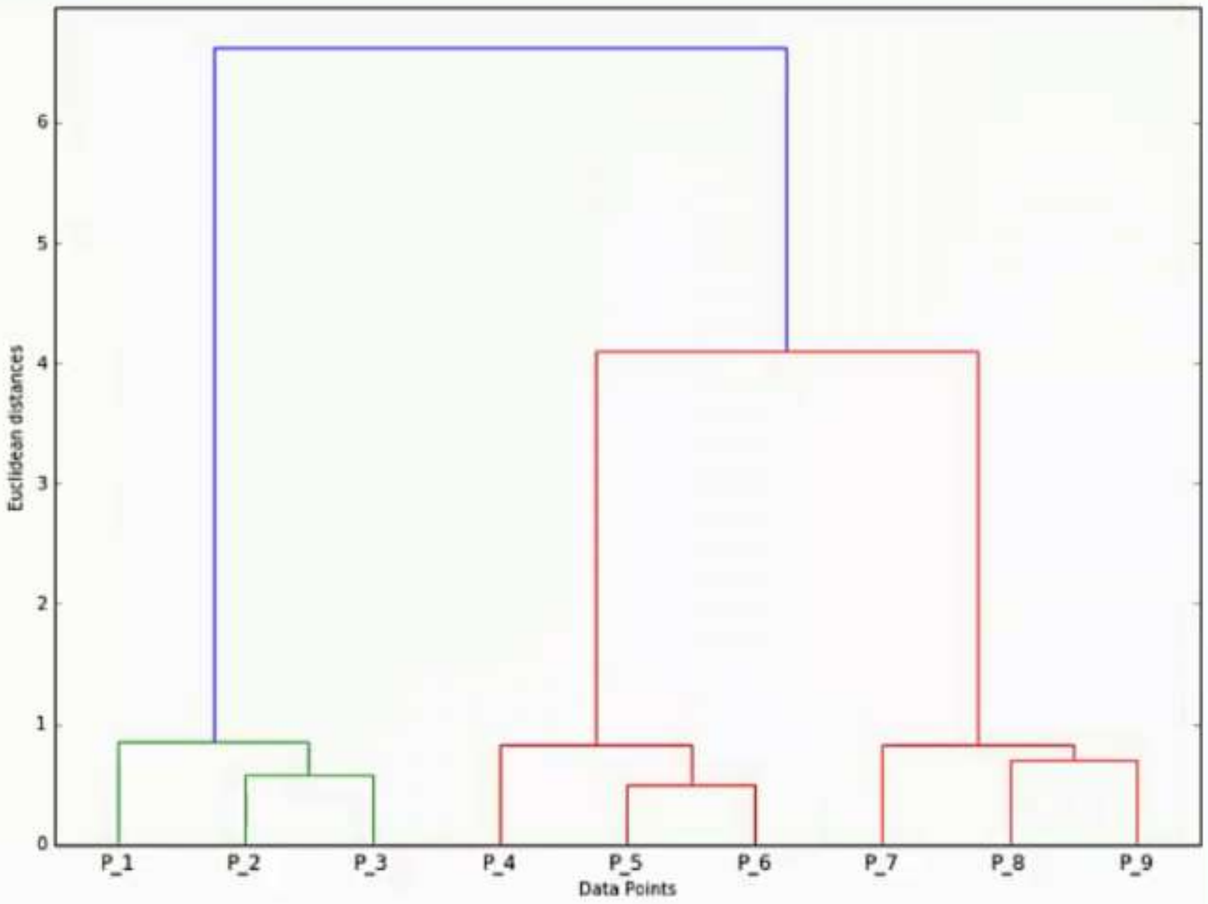
Dendrograms – Knowledge Test



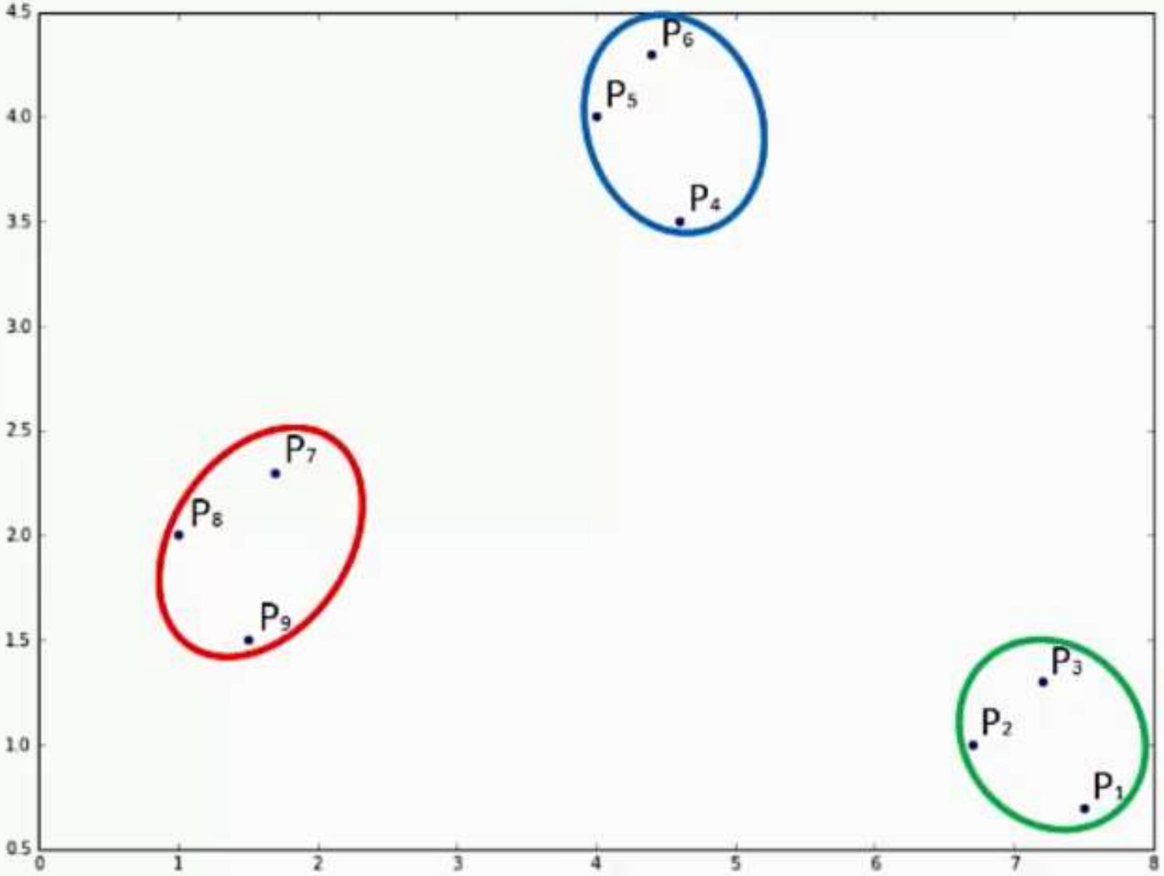


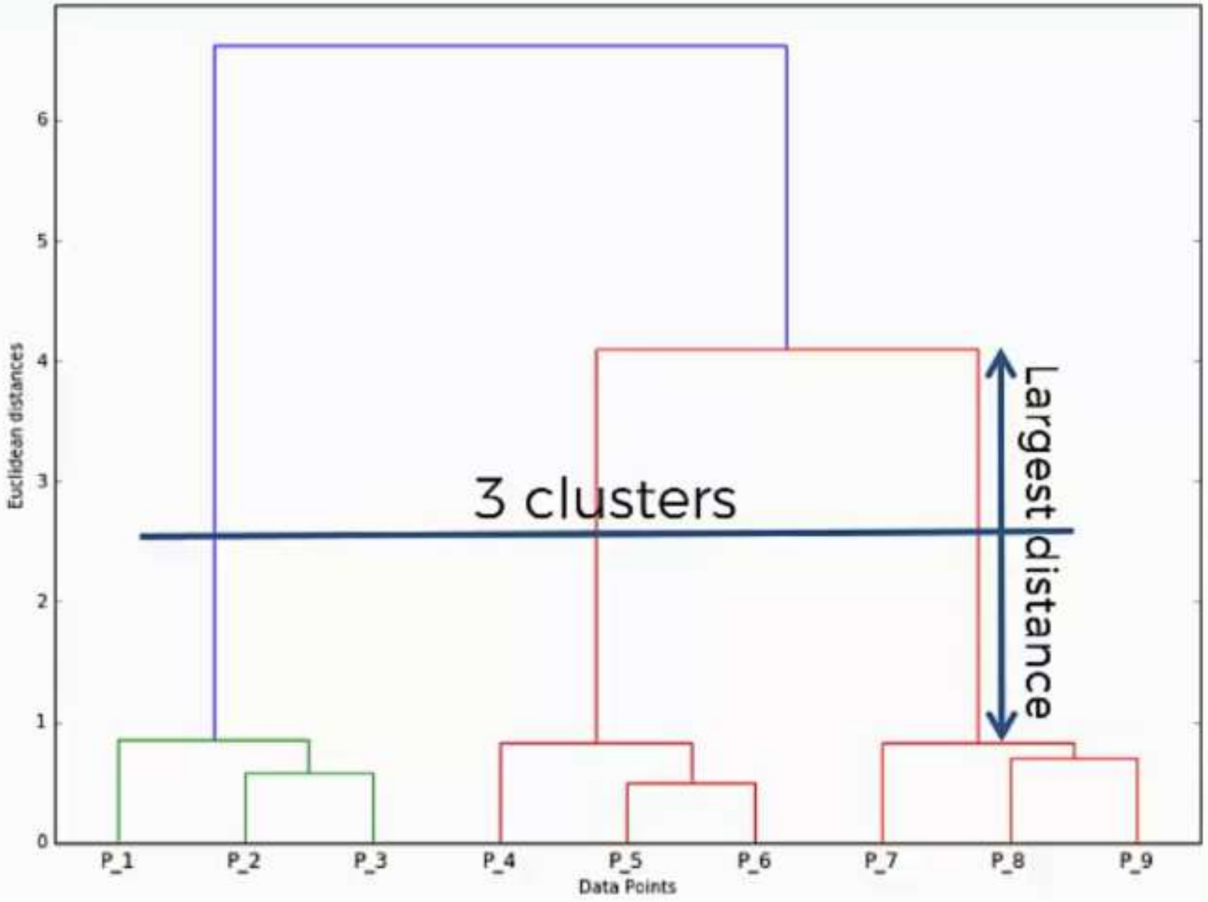
Dendrograms – Knowledge Test





Dendrograms – Knowledge Test





Demo: Use HC to do a clustering analysis for mall customer data.