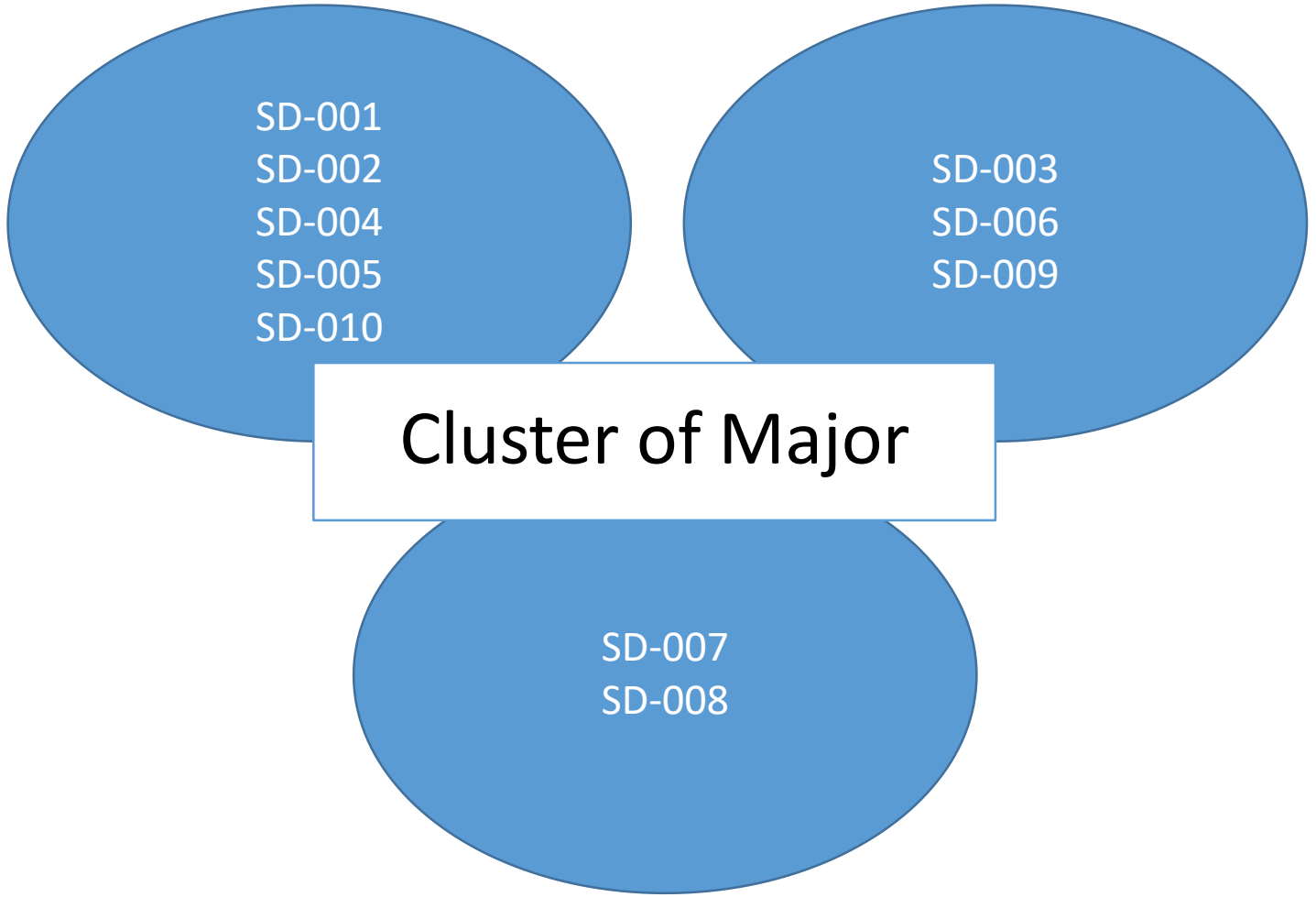
Clustering

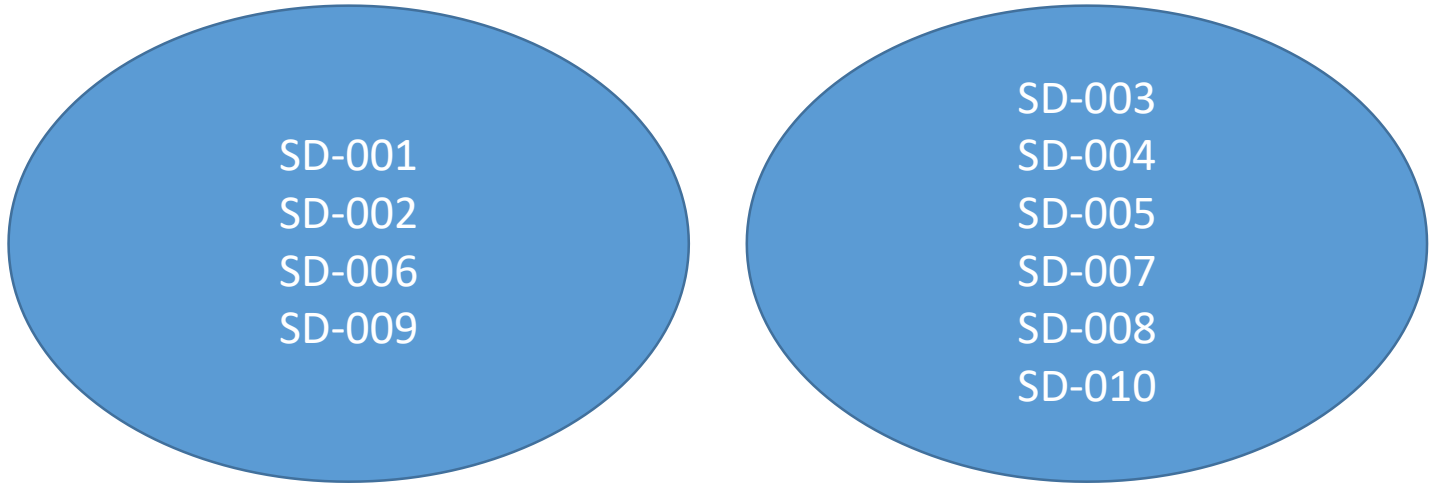
What is Clustering?

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+
SD-010	F	Math	B+



What is Clustering?

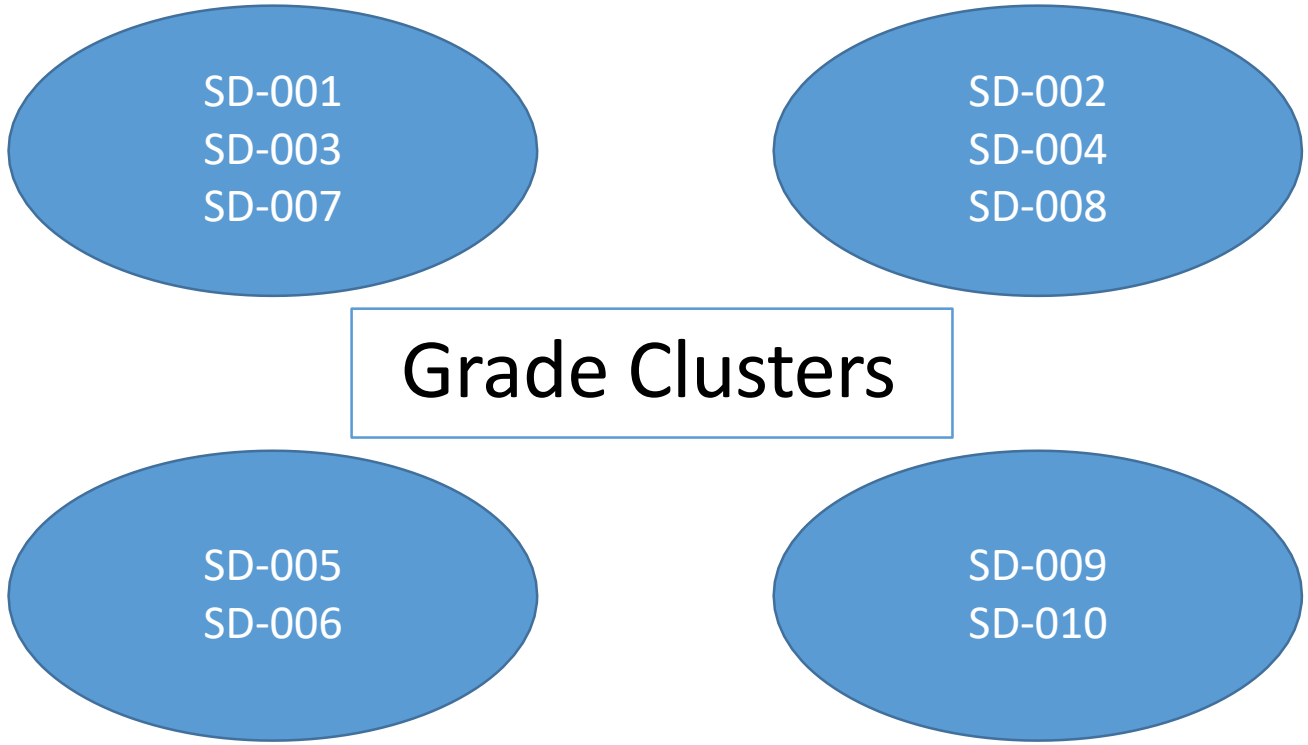
Student ID	Gender	Major	Grade
SD-001	M	Math	A+
SD-002	M	Math	A
SD-003	F	Statistics	A+
SD-004	F	Math	A
SD-005	F	Math	B
SD-006	M	Statistics	B
SD-007	F	Physics	A+
SD-008	F	Physics	A
SD-009	M	Statistics	B+
SD-010	F	Math	B+



Gender Specific Clusters

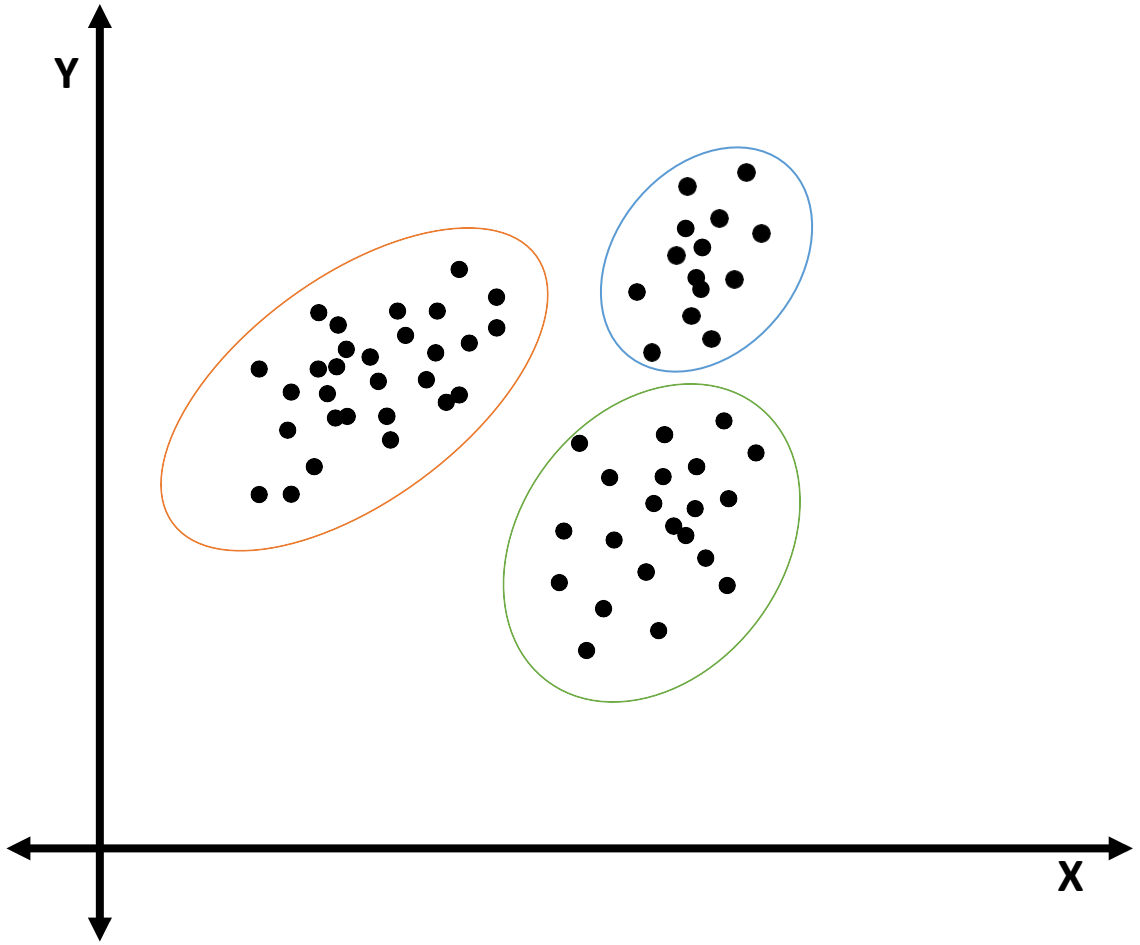
What is Clustering?

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+
SD-010	F	Math	B+



What is Clustering or Cluster 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

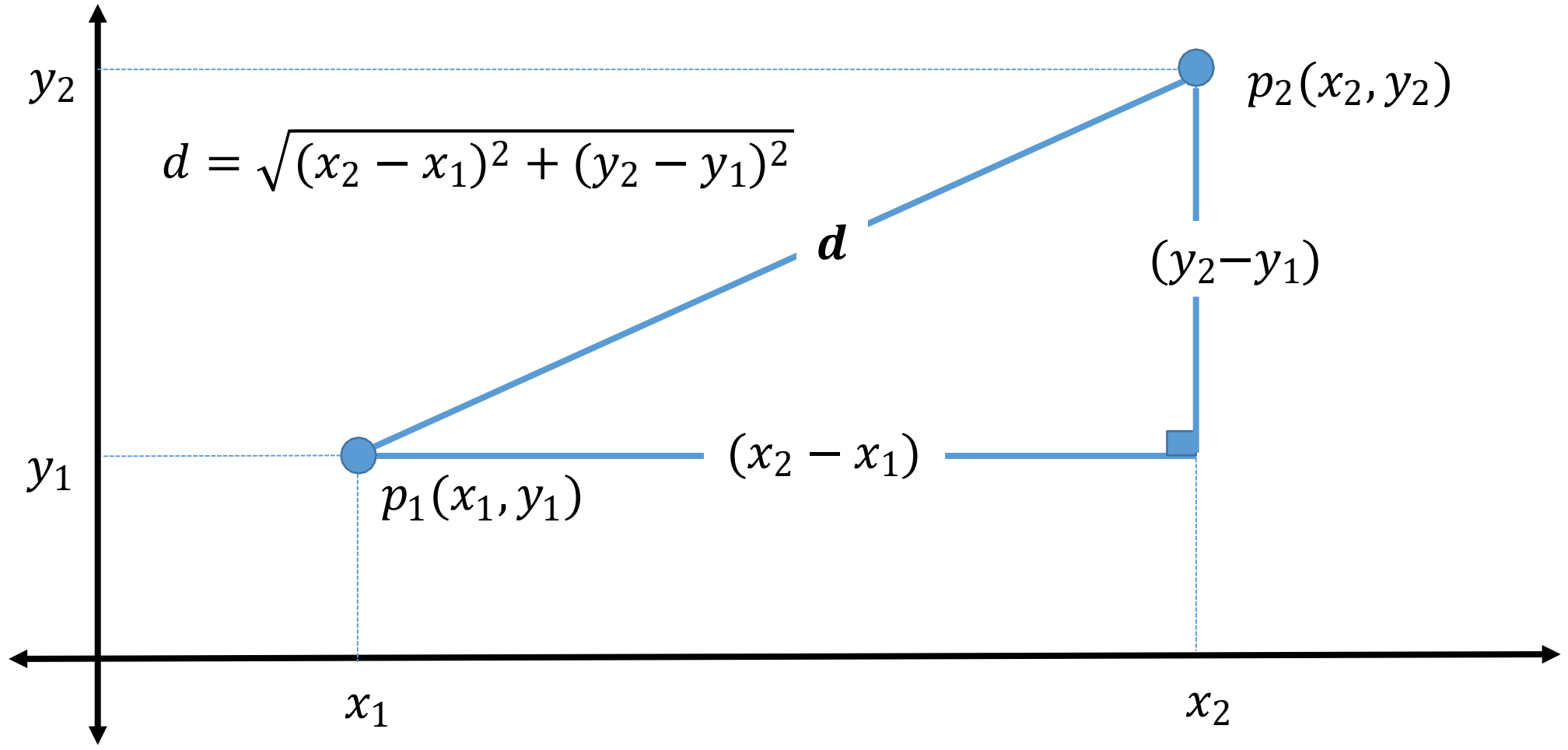


Examples of Clustering

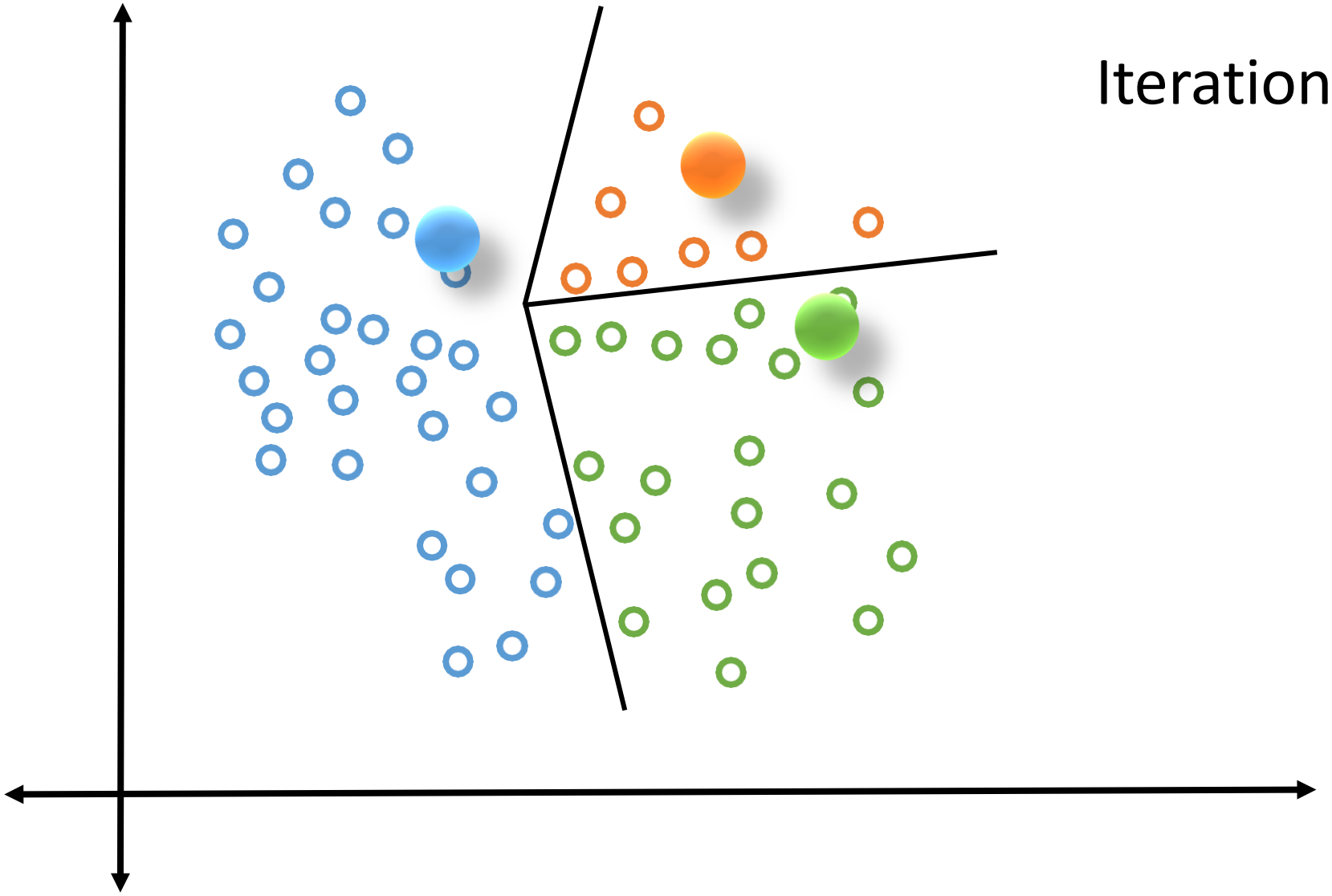
- Recommendation engines
- Market segmentation
- Social network analysis
- Medical/Health
- Image segmentation
- Anomaly detection

How Clustering works?

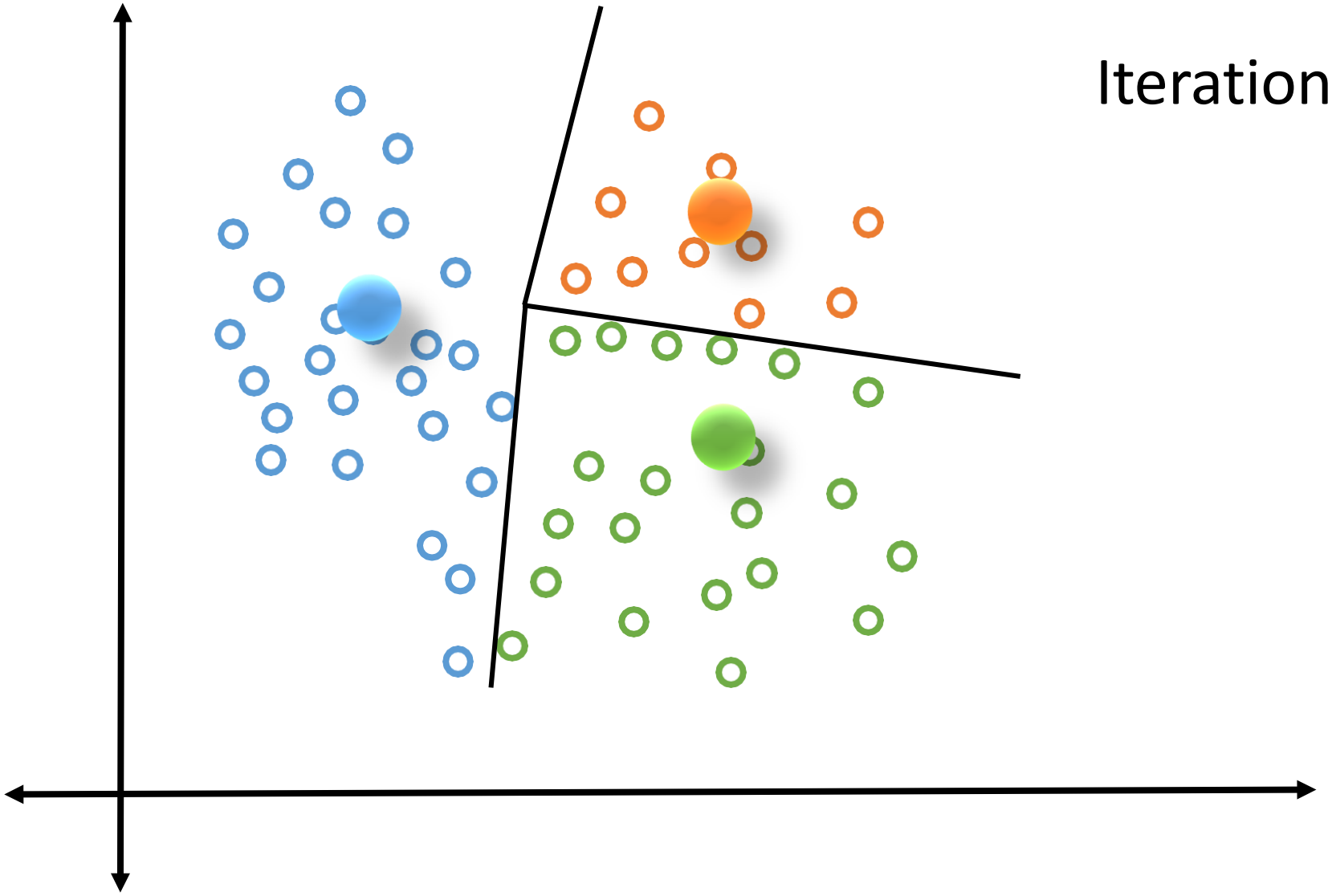
Euclidean Distance



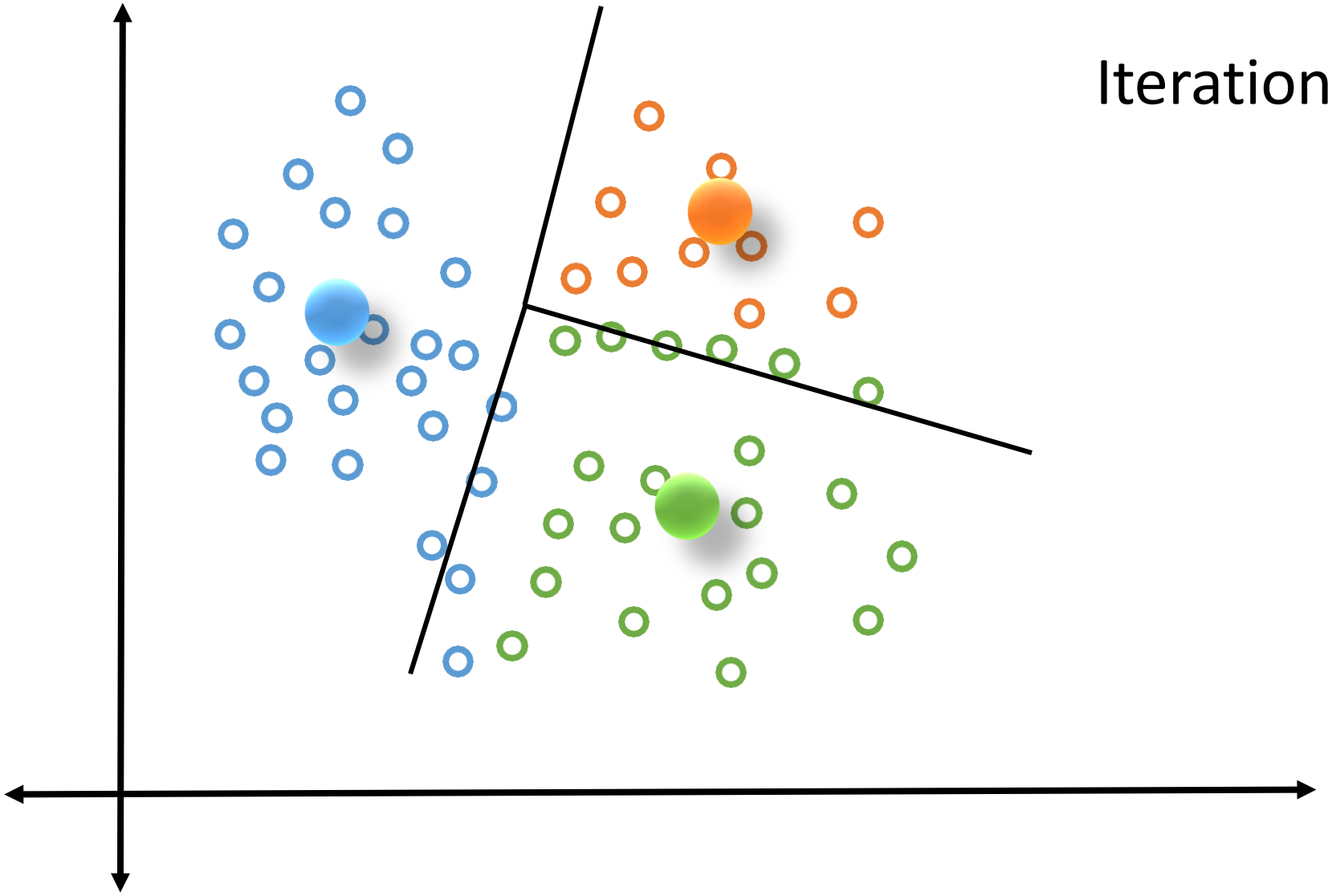
How Clusters are formed?



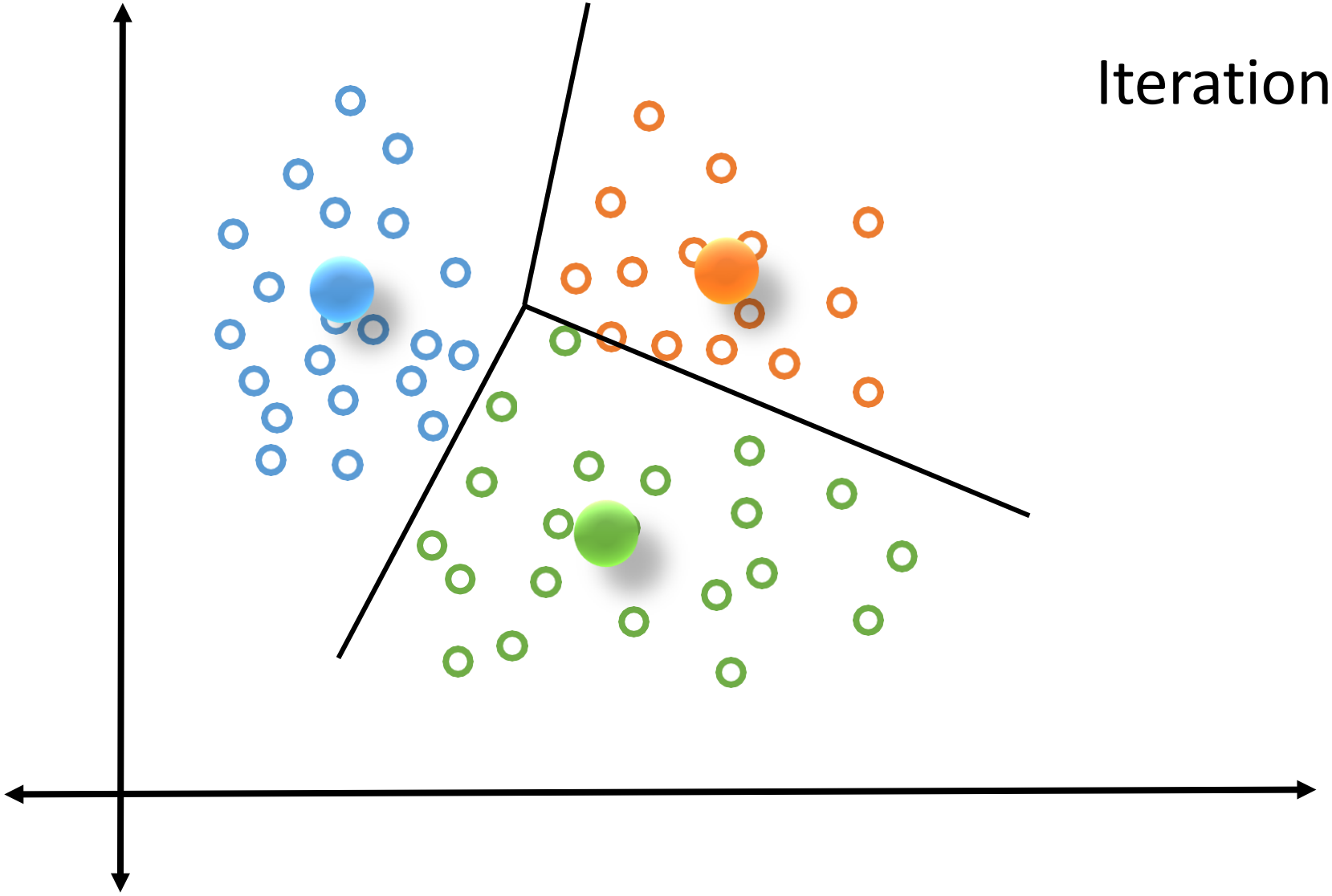
How Clusters are formed?



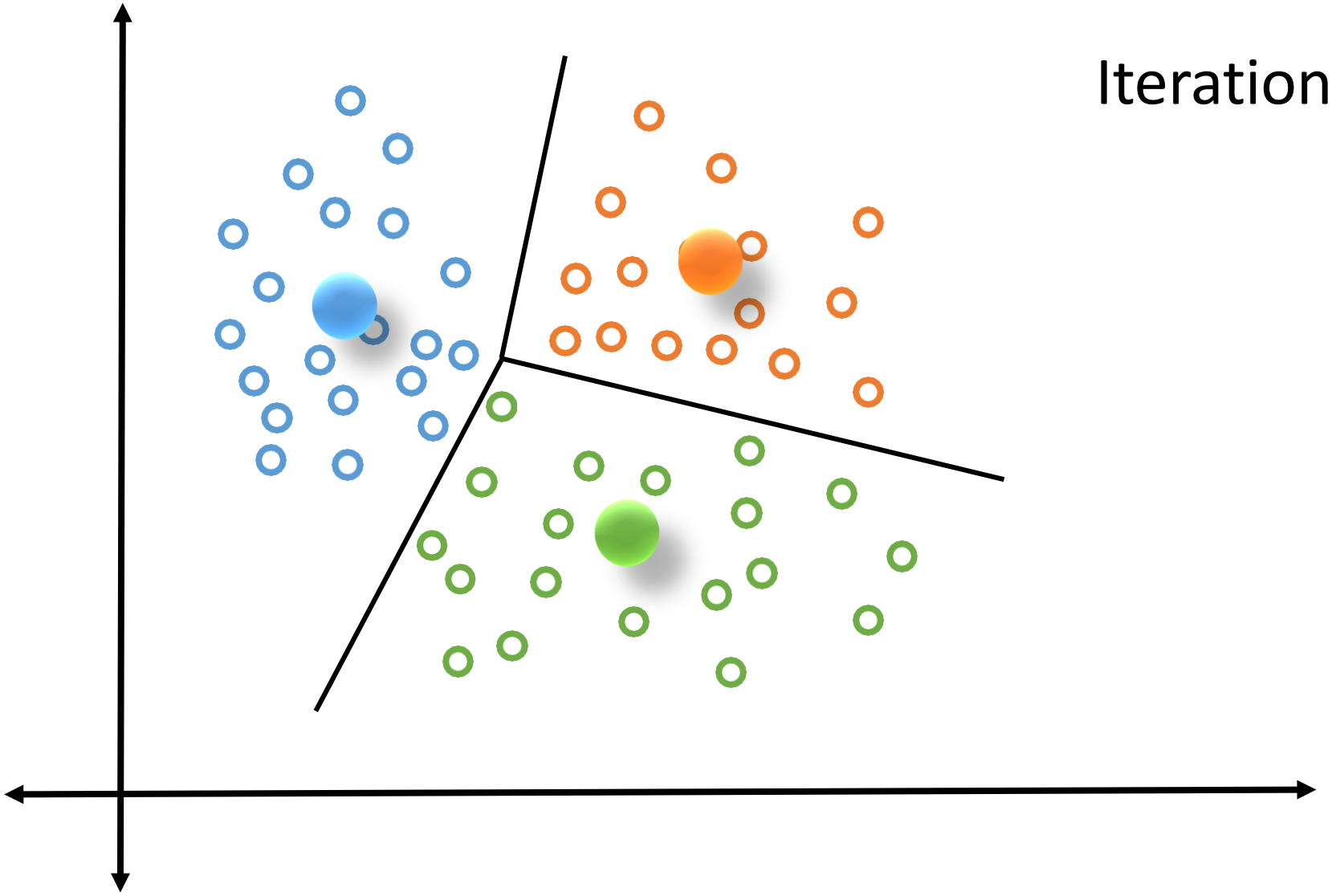
How Clusters are formed?



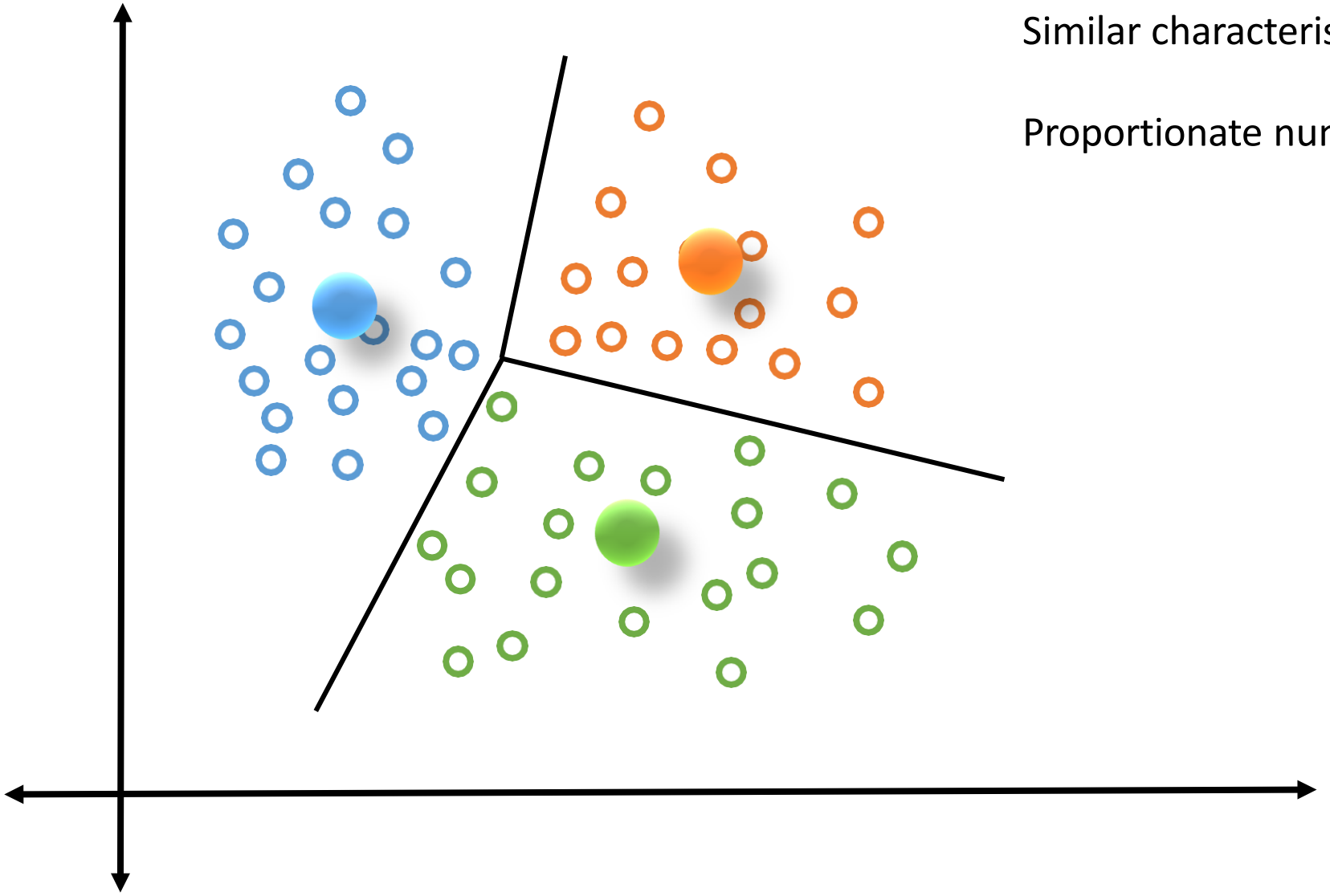
How Clusters are formed?



How Clusters are formed?



How Clusters are formed?

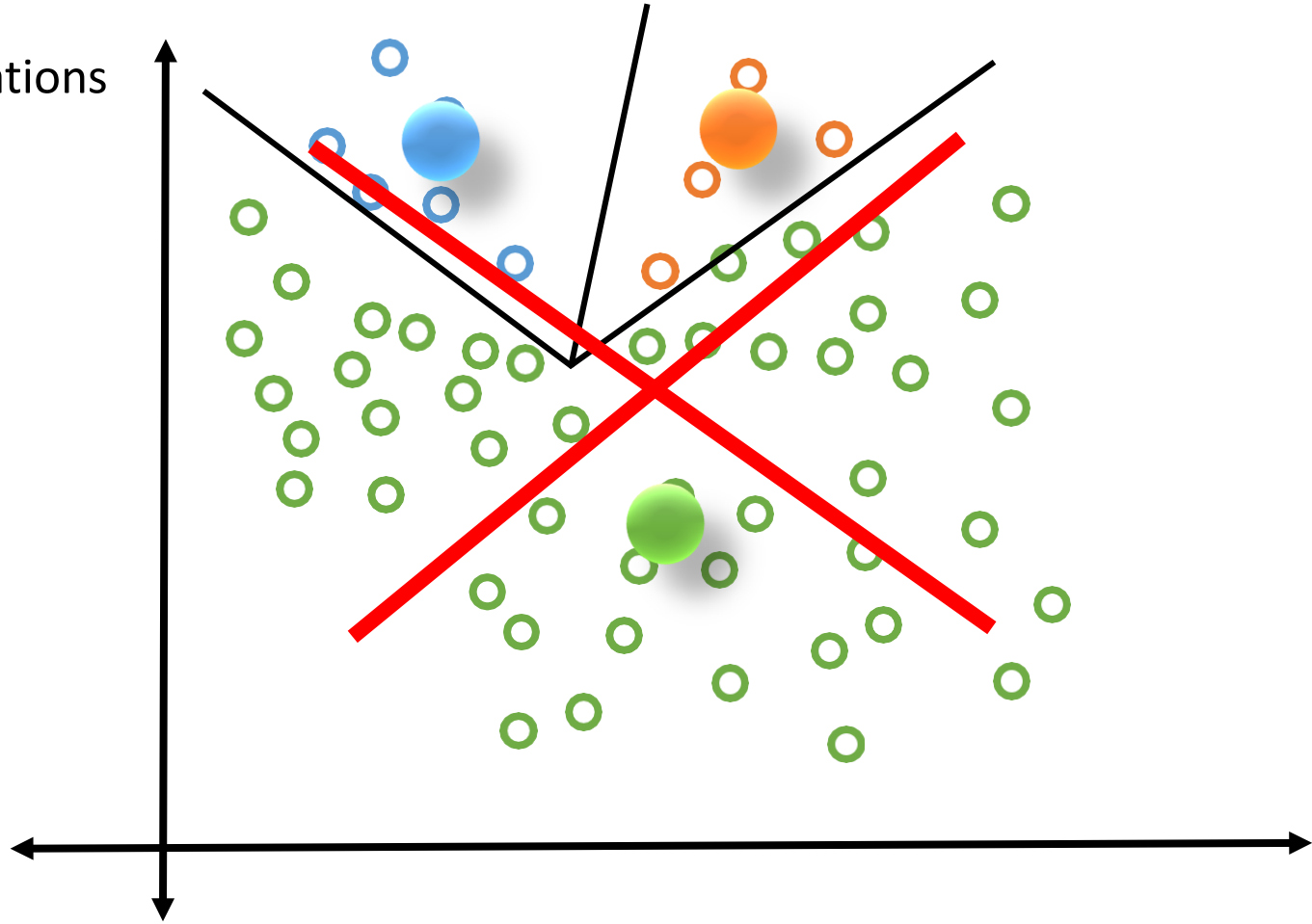


Similar characteristics

Proportionate number of observations

Good Cluster Analysis

- Observations in the same group share similar characteristics
- Clusters have proportionate number observations

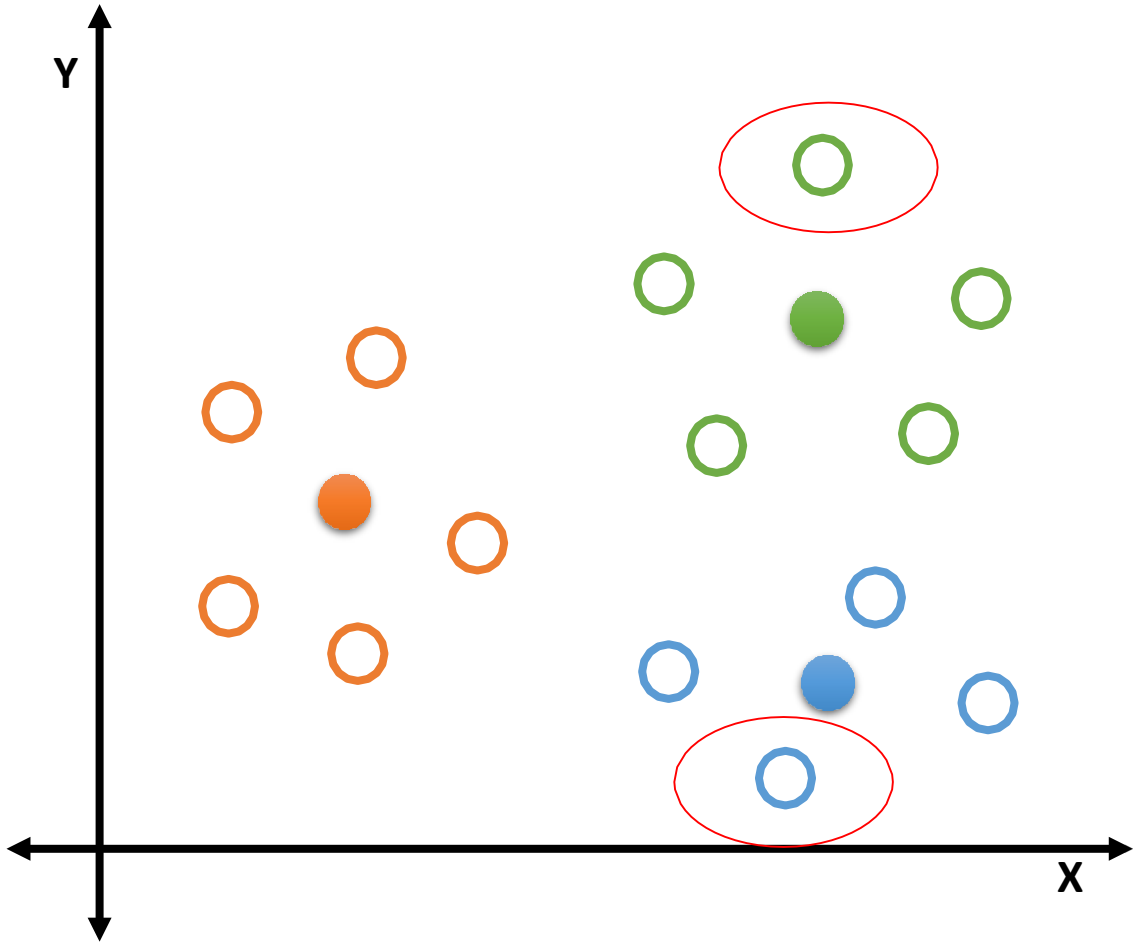


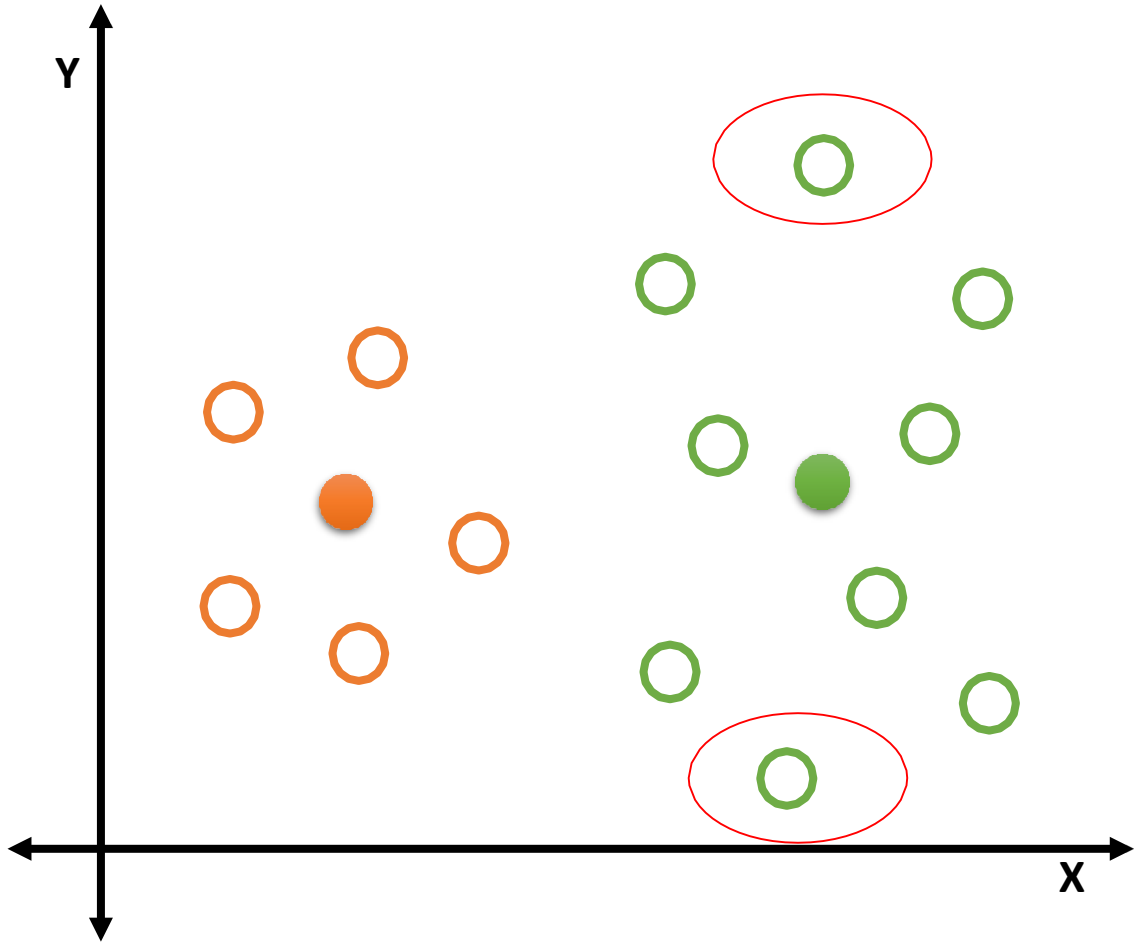
How to Decide Number of Clusters?

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

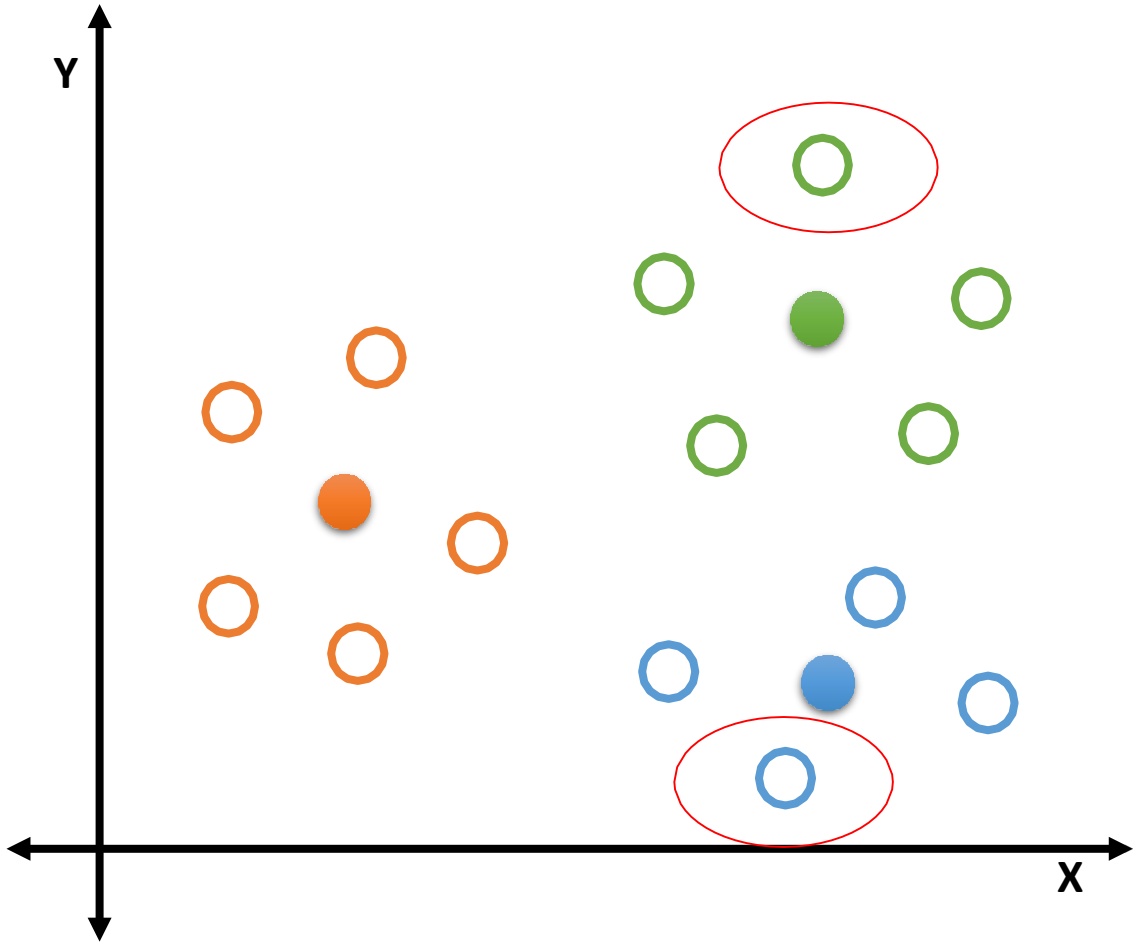


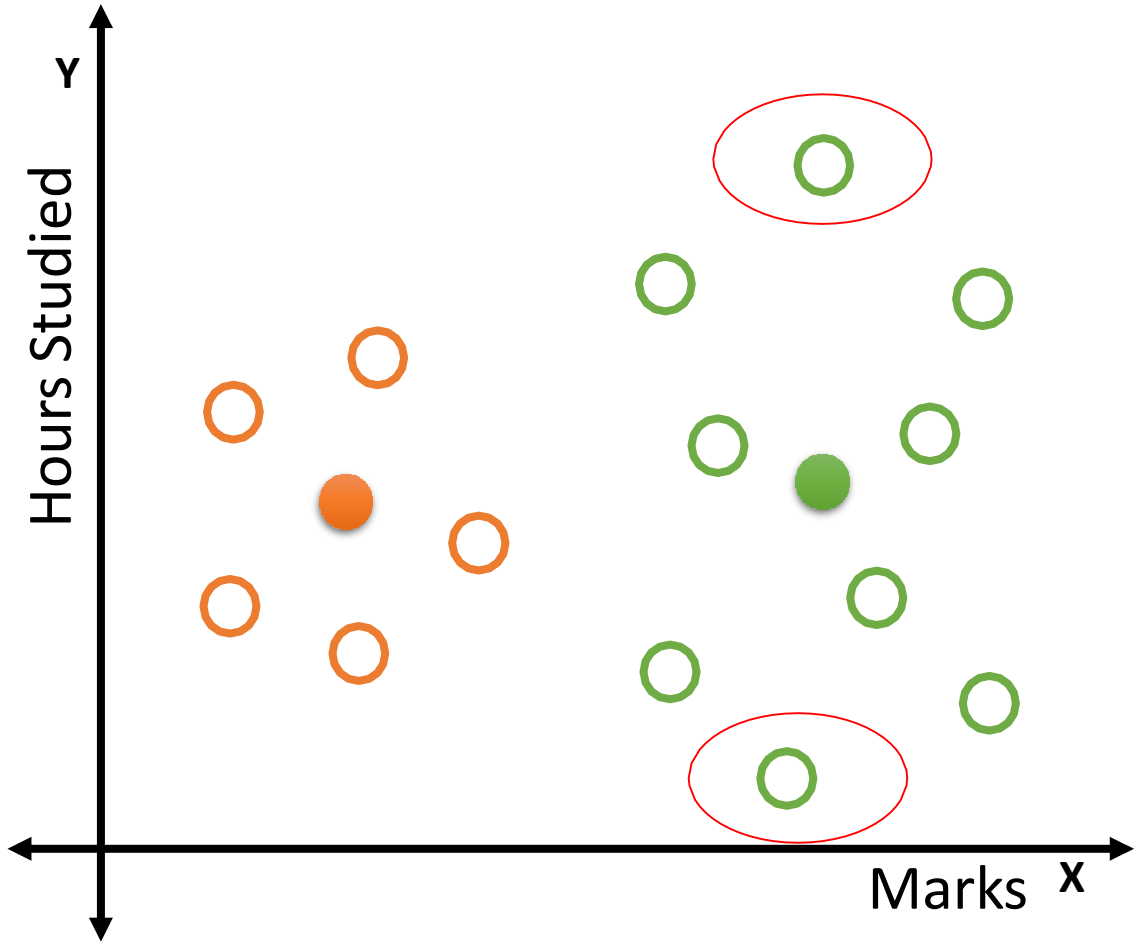


Observations in the same clusters are related to each other.

Distance of the observations from centroid decides their cluster assignment.

Lesser the distance,
Better the relationship

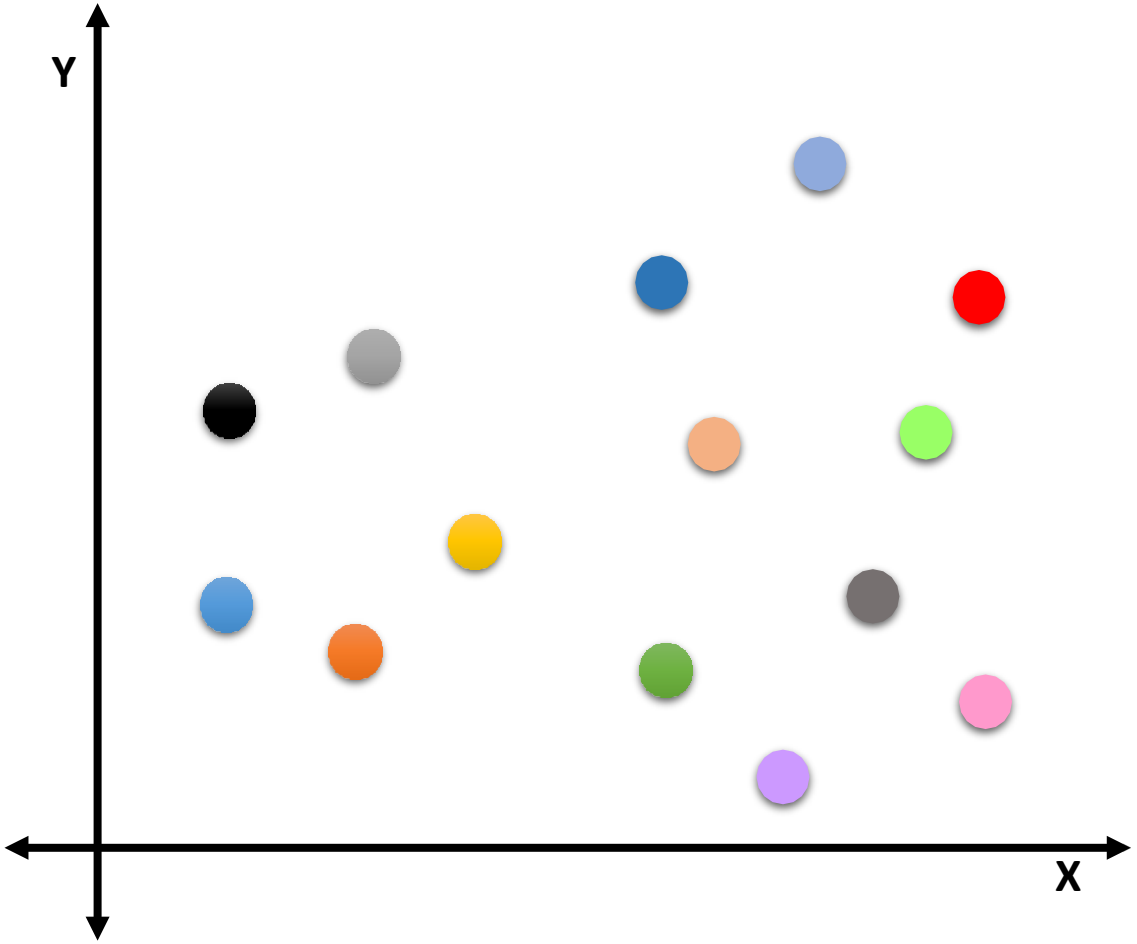


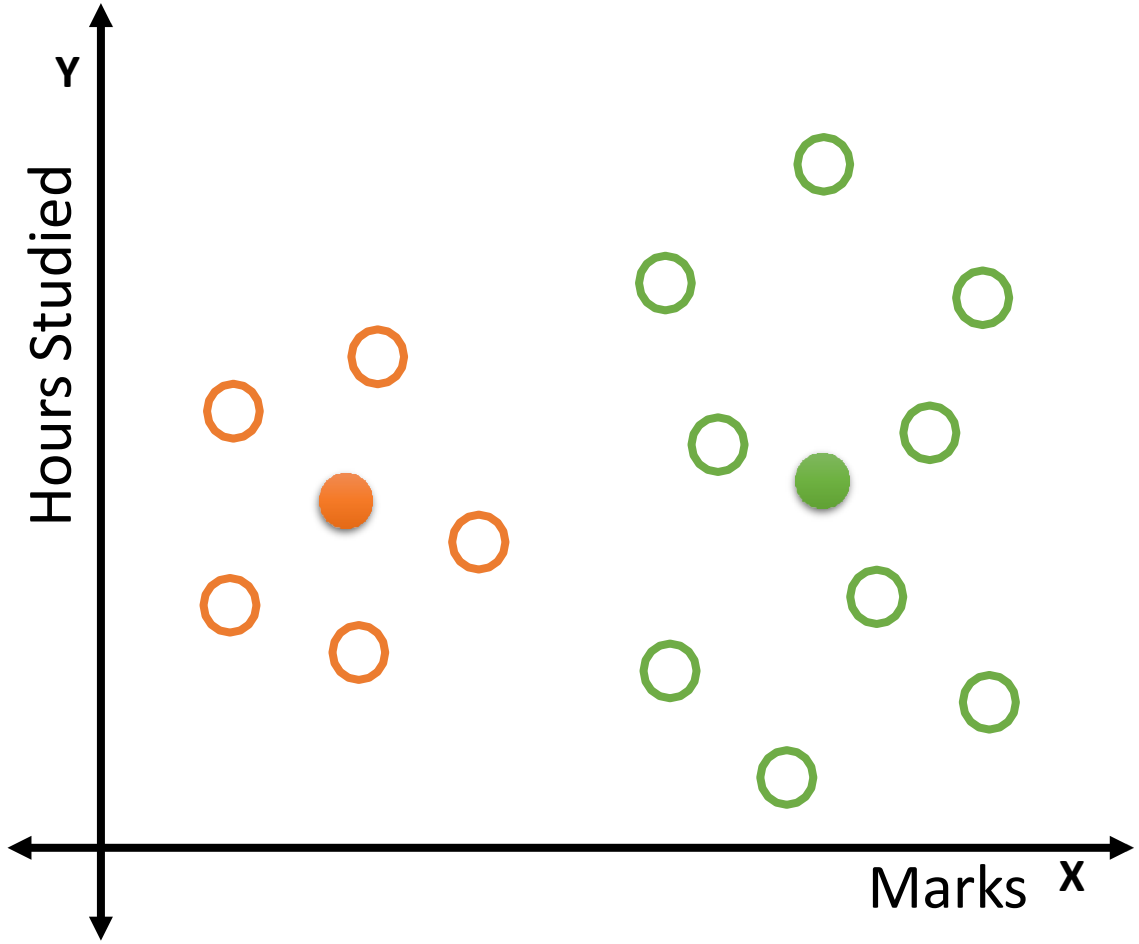


Observations in the same clusters are related to each other.

Distance of the observations from centroid decides their cluster assignment.

Lesser the distance,
Better the relationship



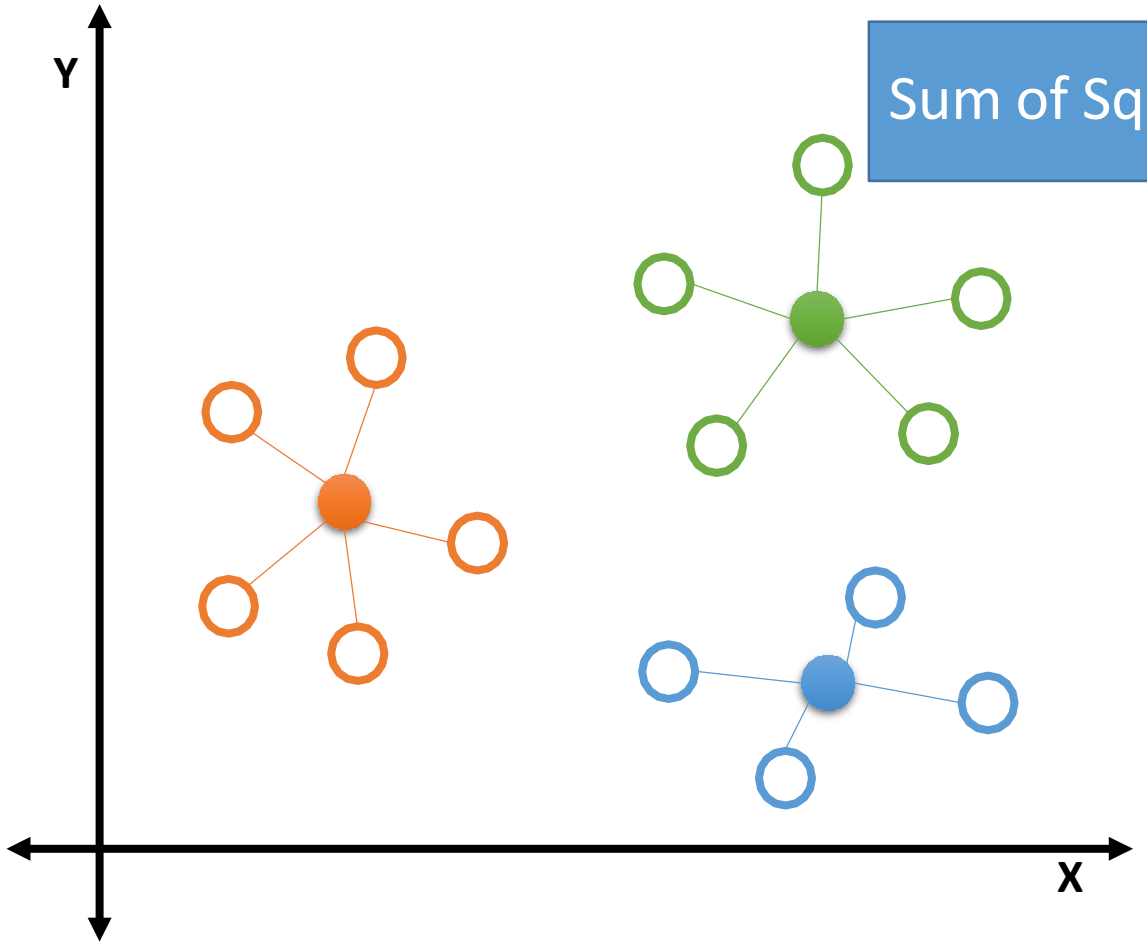


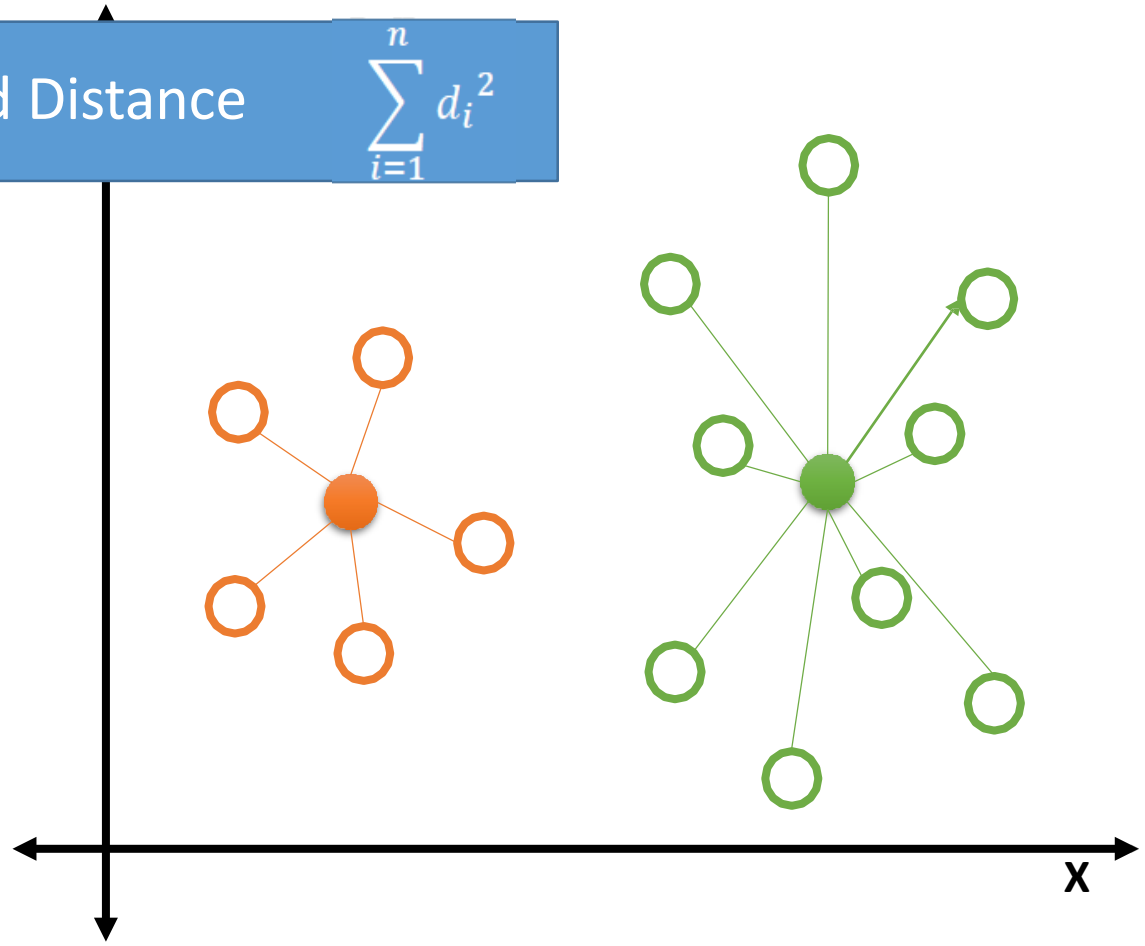
Observations in the same clusters are related to each other.

Distance of the observations from centroid decides their cluster assignment.

Lesser the distance,
Better the relationship

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



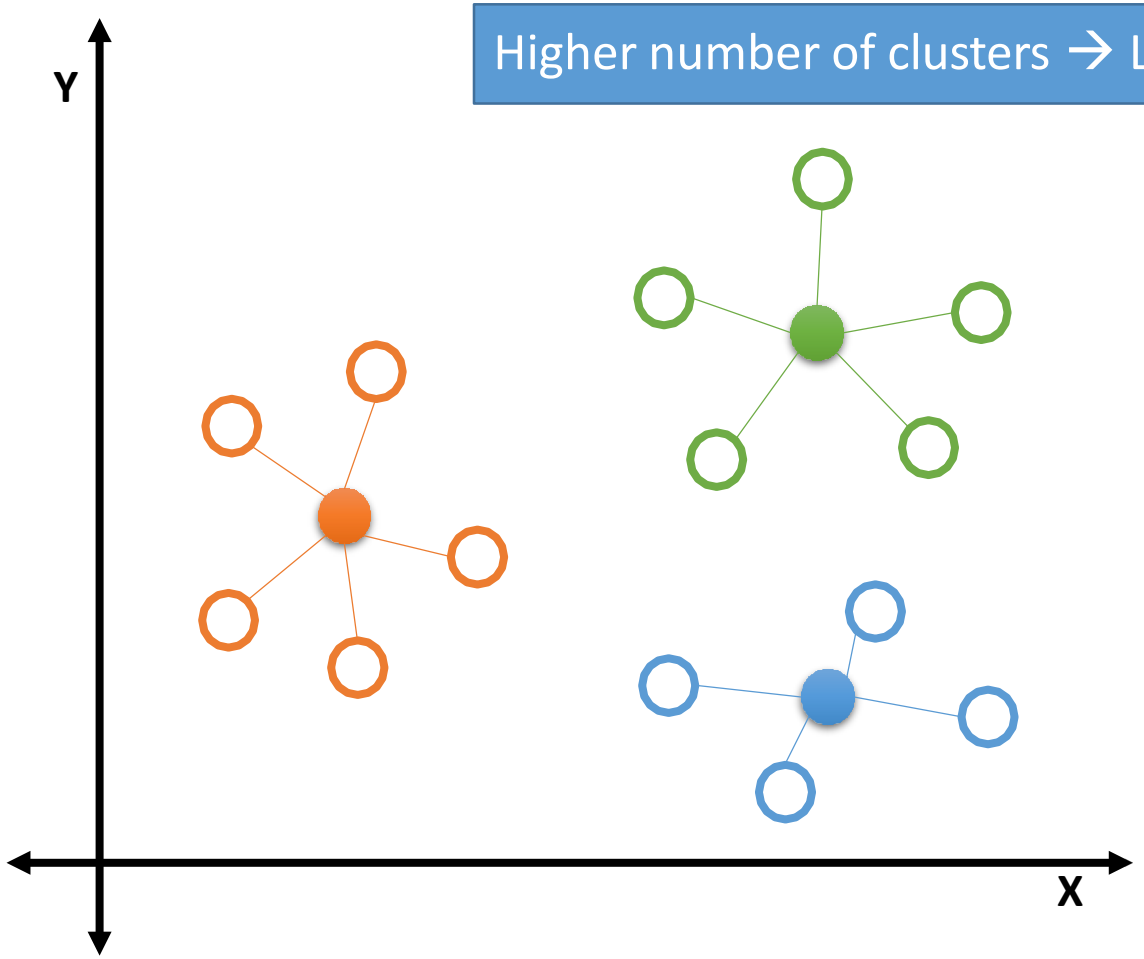


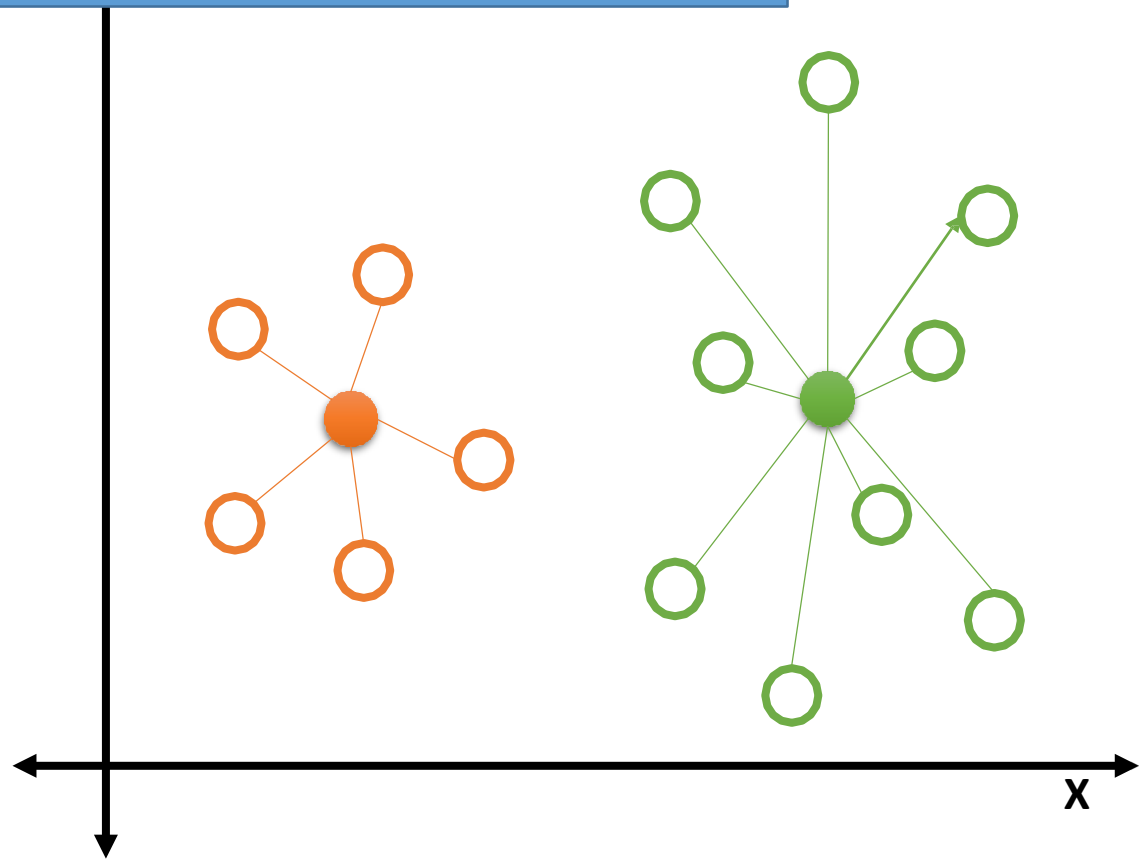
Observations in the same clusters are related to each other.

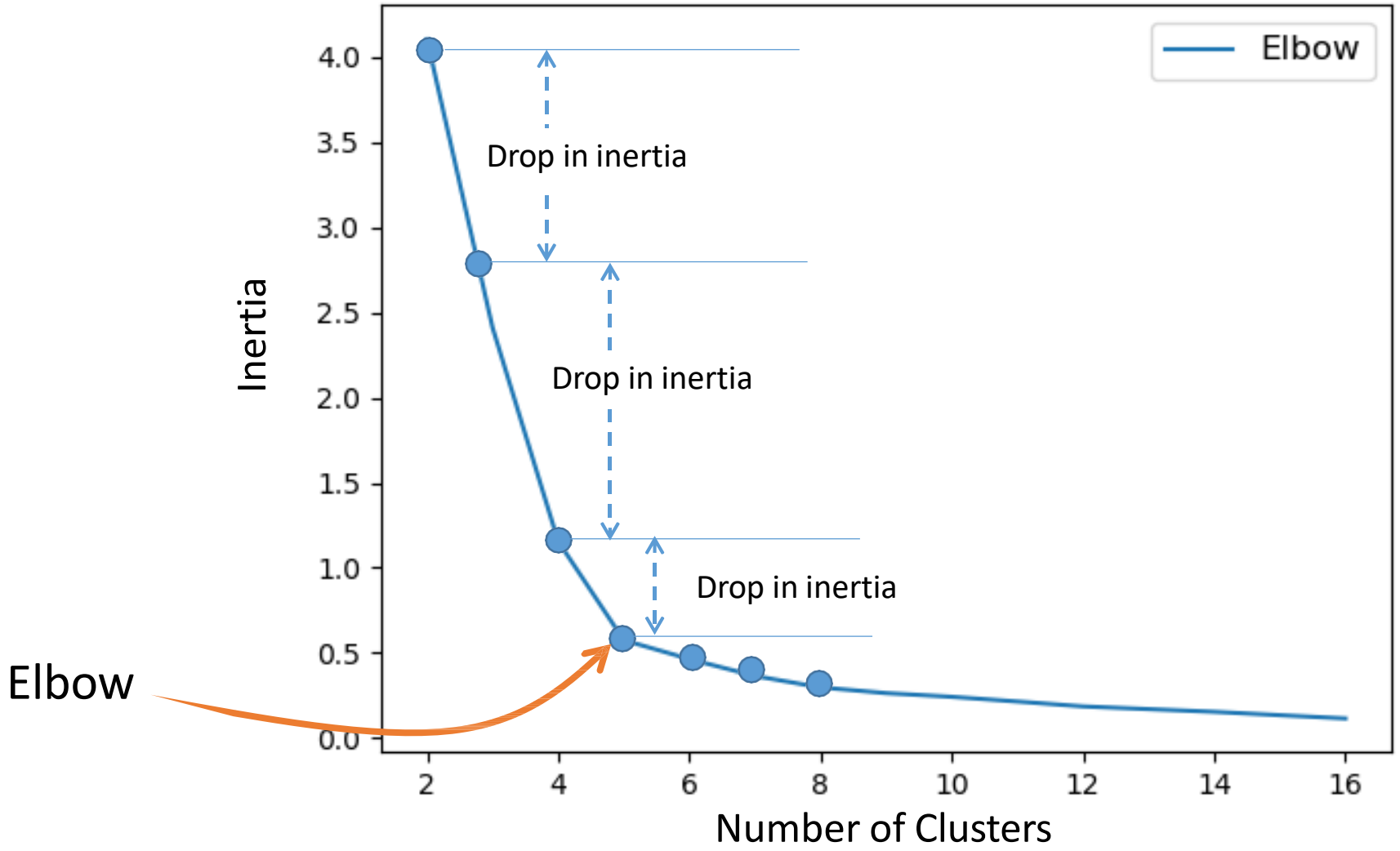
Distance of the observations from centroid decides their cluster assignment.

Lesser the distance,
Better the relationship

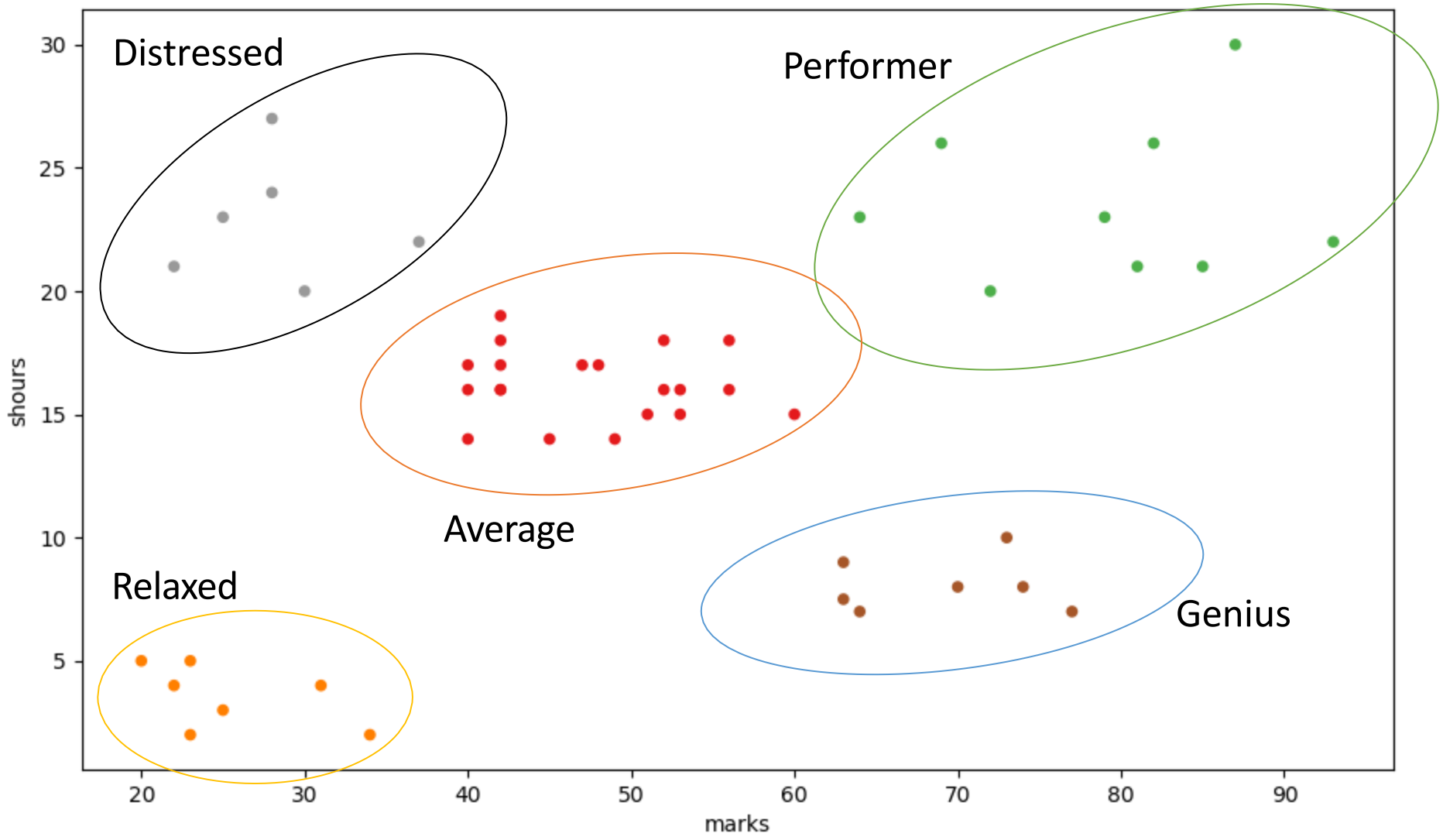
Higher number of clusters → Lesser Sum of Squared Distance (Intertia)

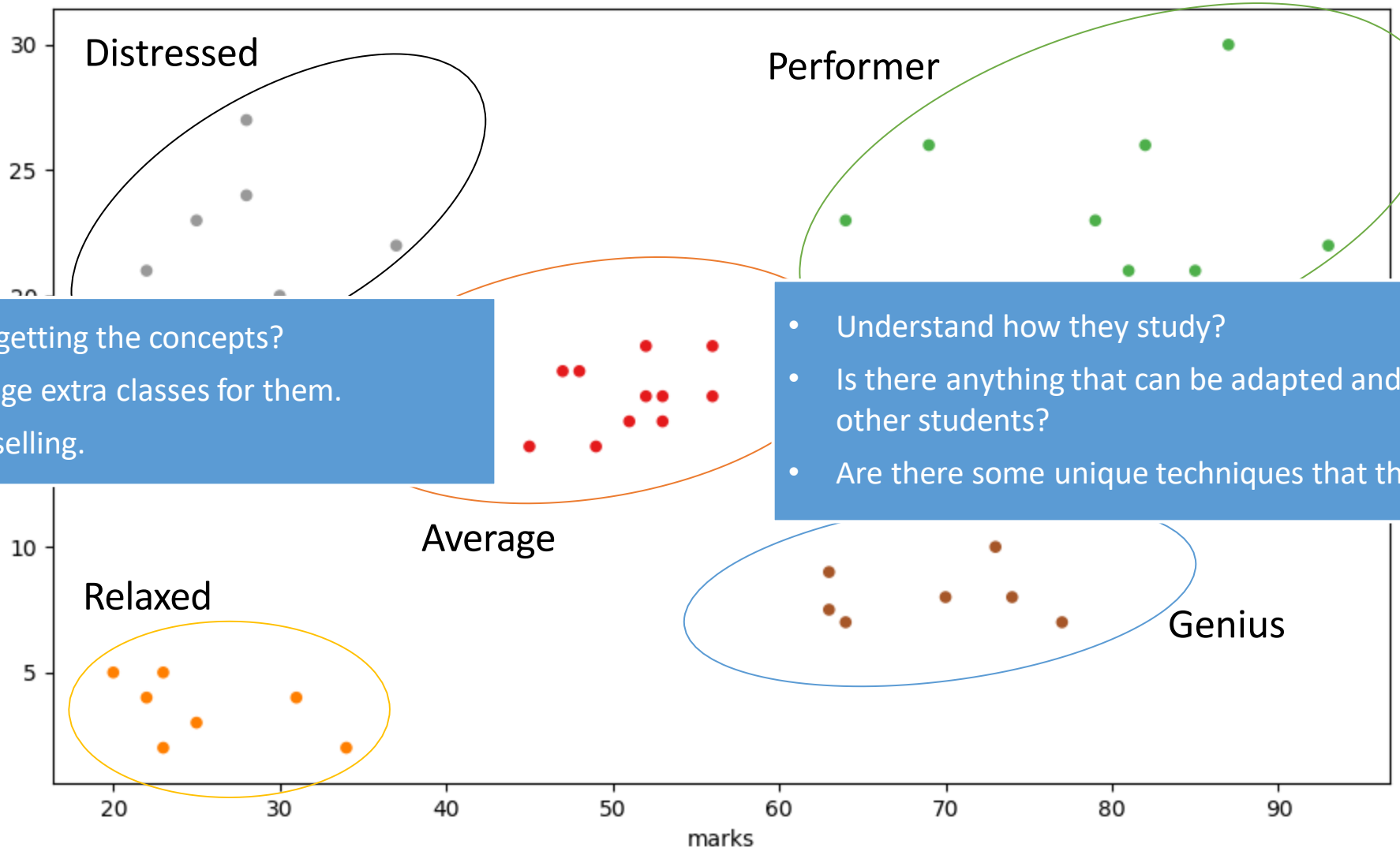





Elbow
Drop in inertia
Drop in inertia
Drop in inertia
Inertia
Number of Clusters
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2
4
6
8
10
12
14
16

Elbow


Distressed
Performer
Average
Relaxed
Genius
shours
marks
30
25
20
15
10
5
20
30
40
50
60
70
80
90



- Are they not getting the concepts?
- May be Arrange extra classes for them.
- May be counselling.

- Understand how they study?
- Is there anything that can be adapted and provided to other students?
- Are there some unique techniques that they are following?