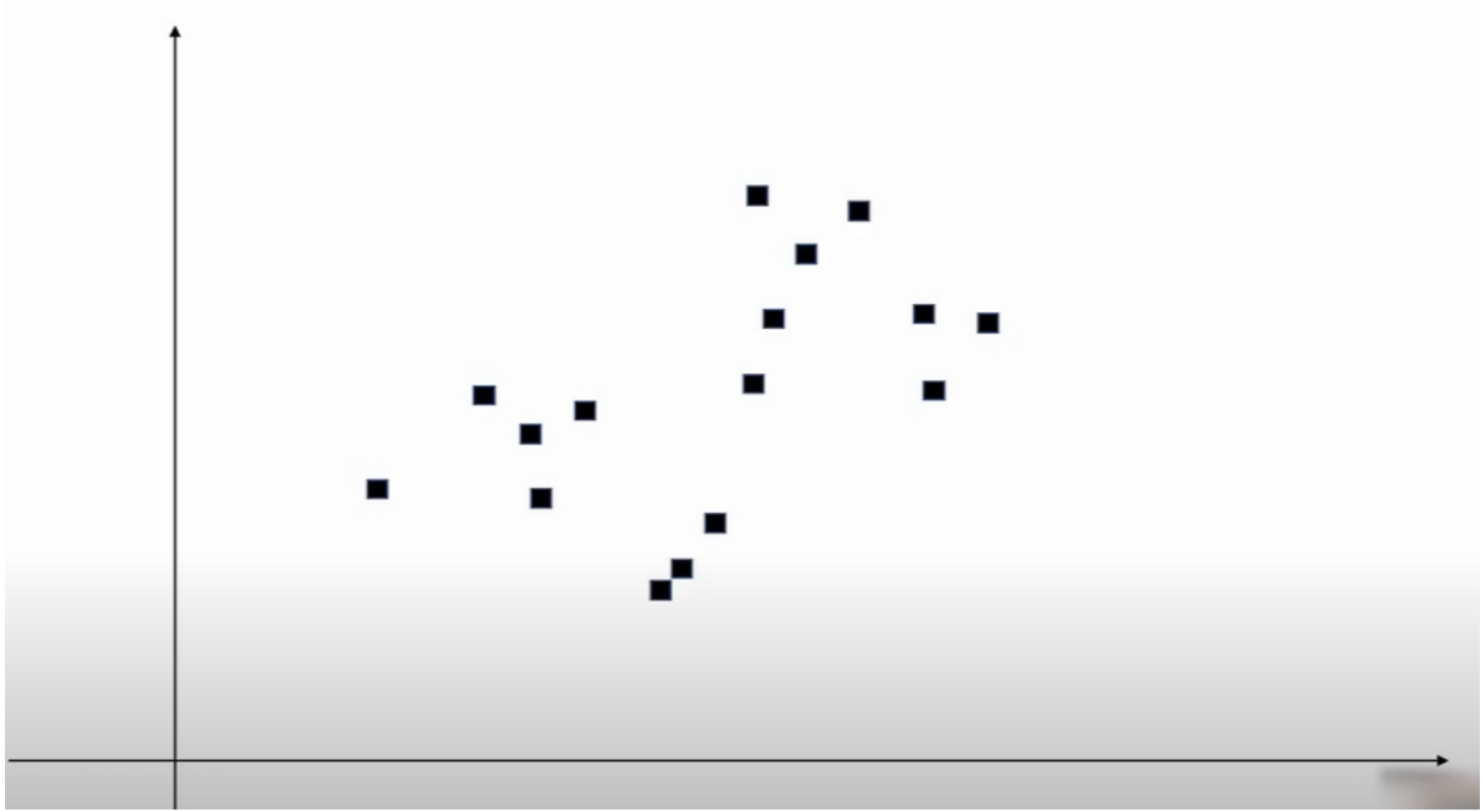
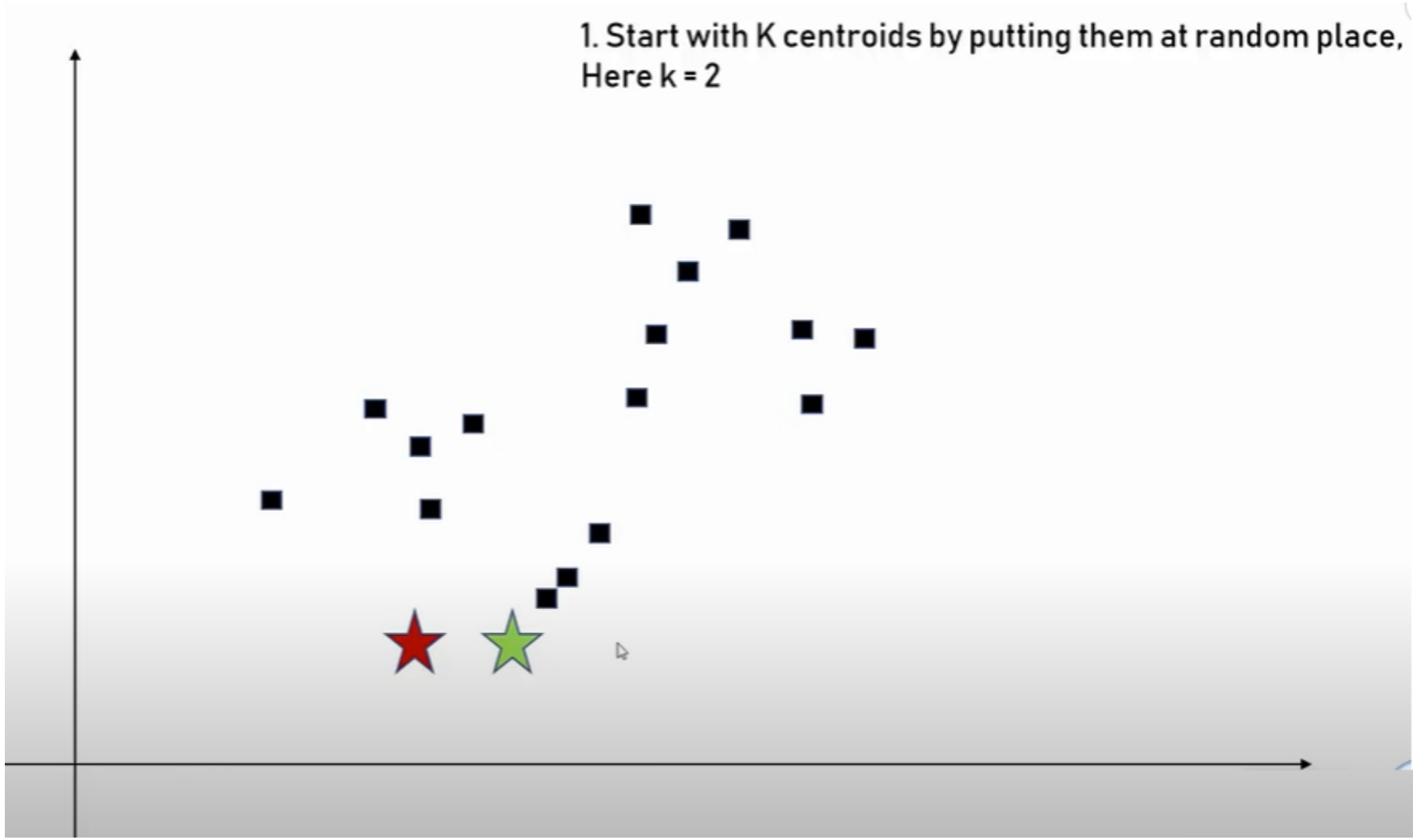
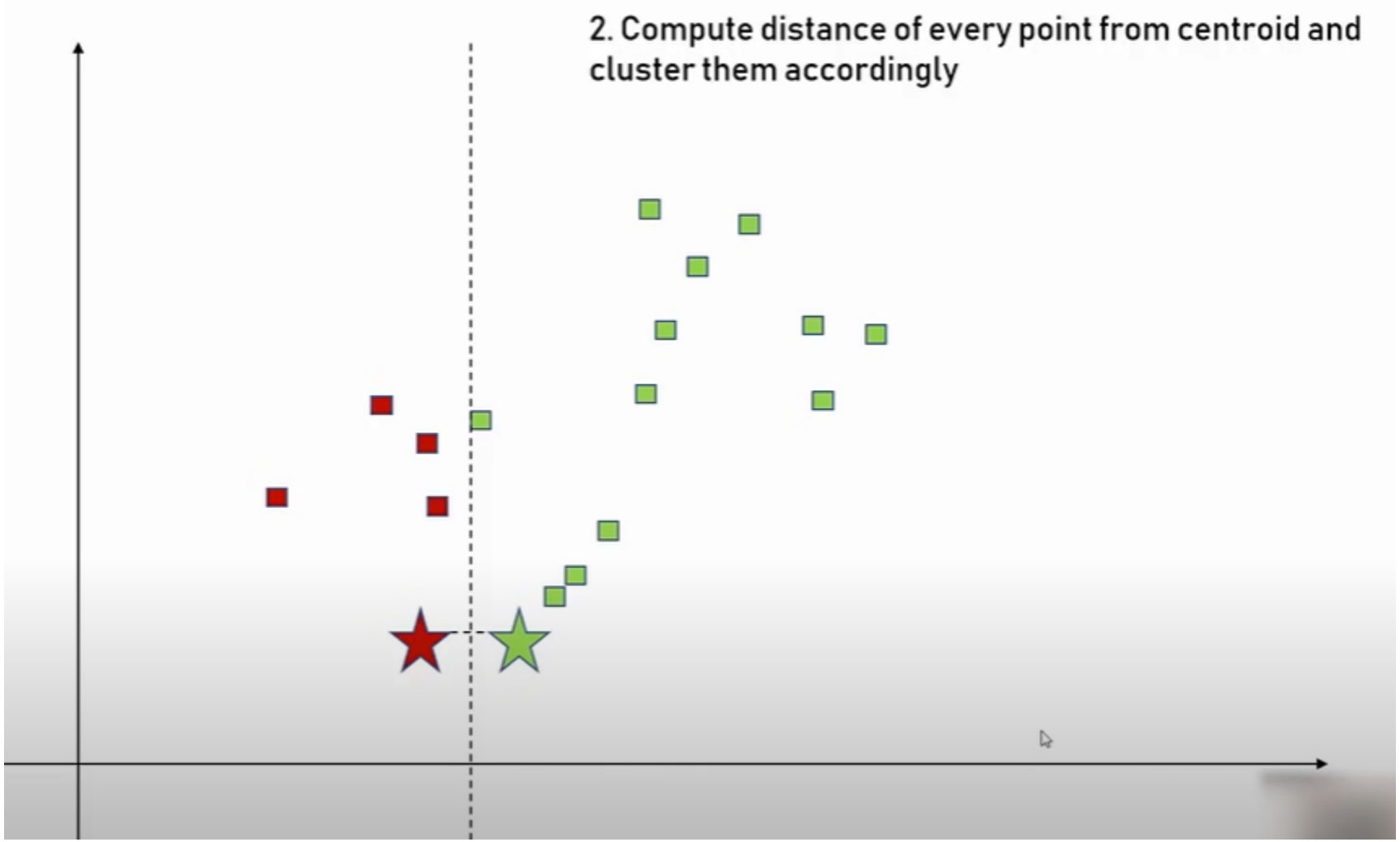
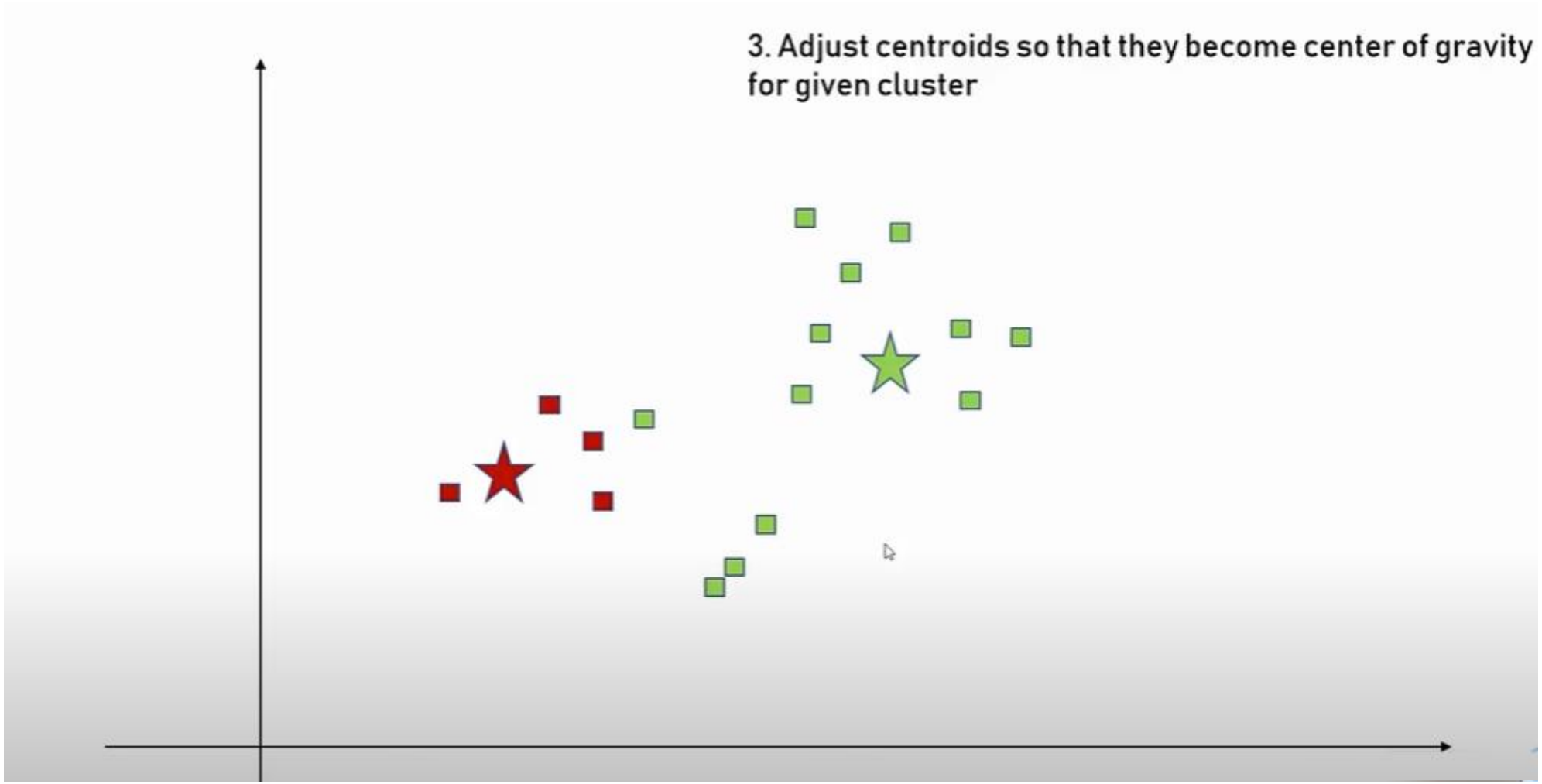
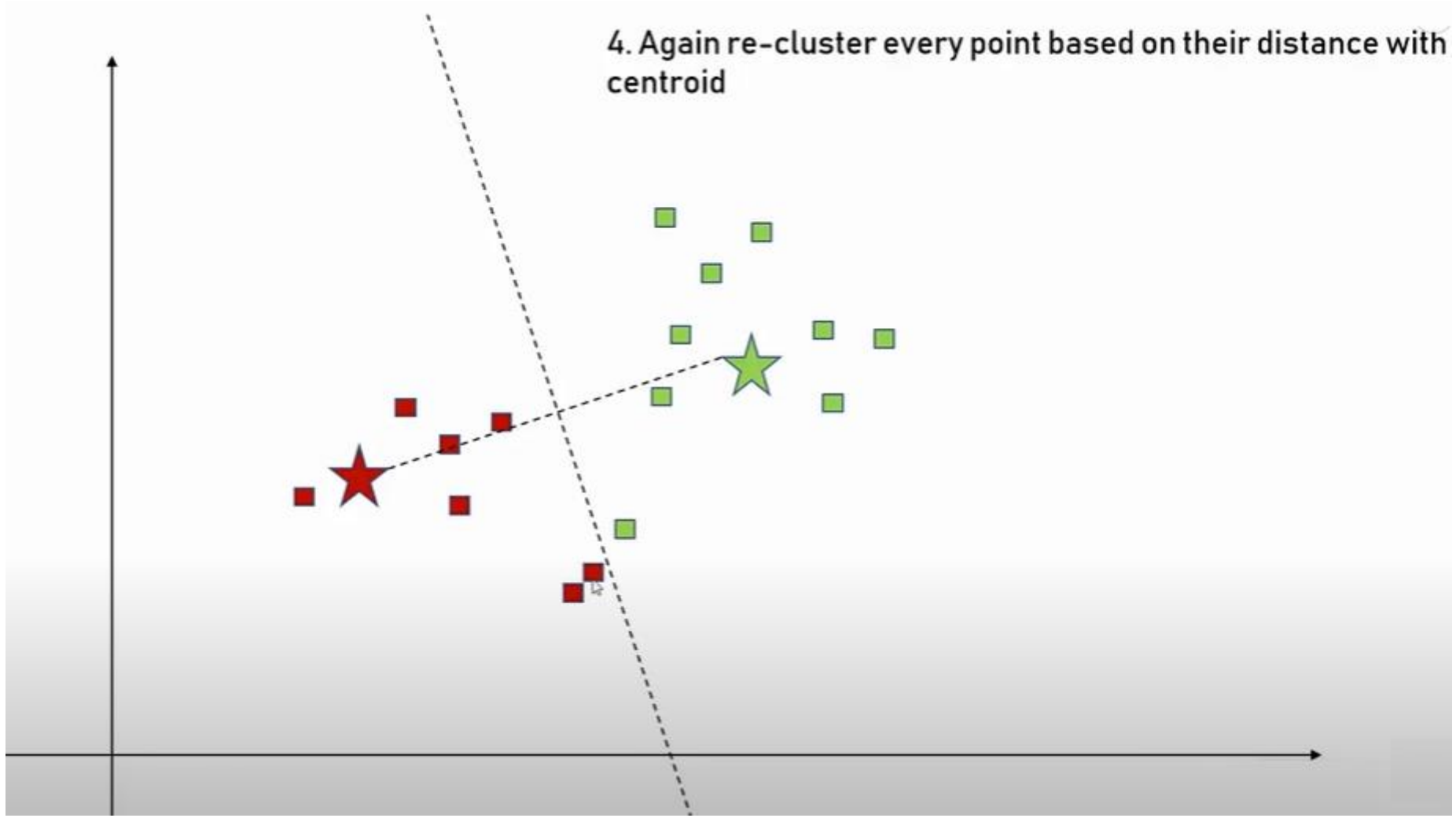
K Means Clustering Algorithm

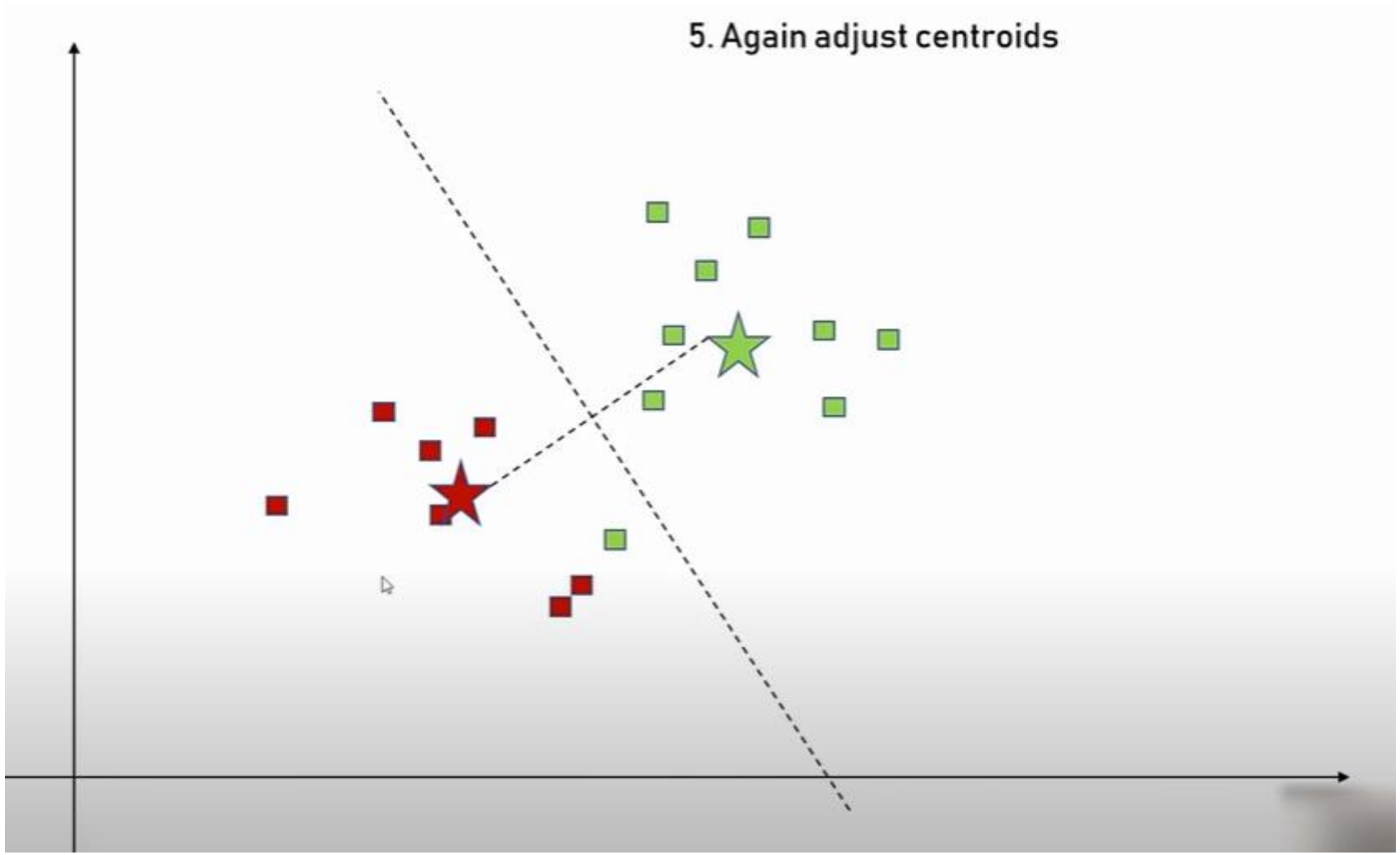


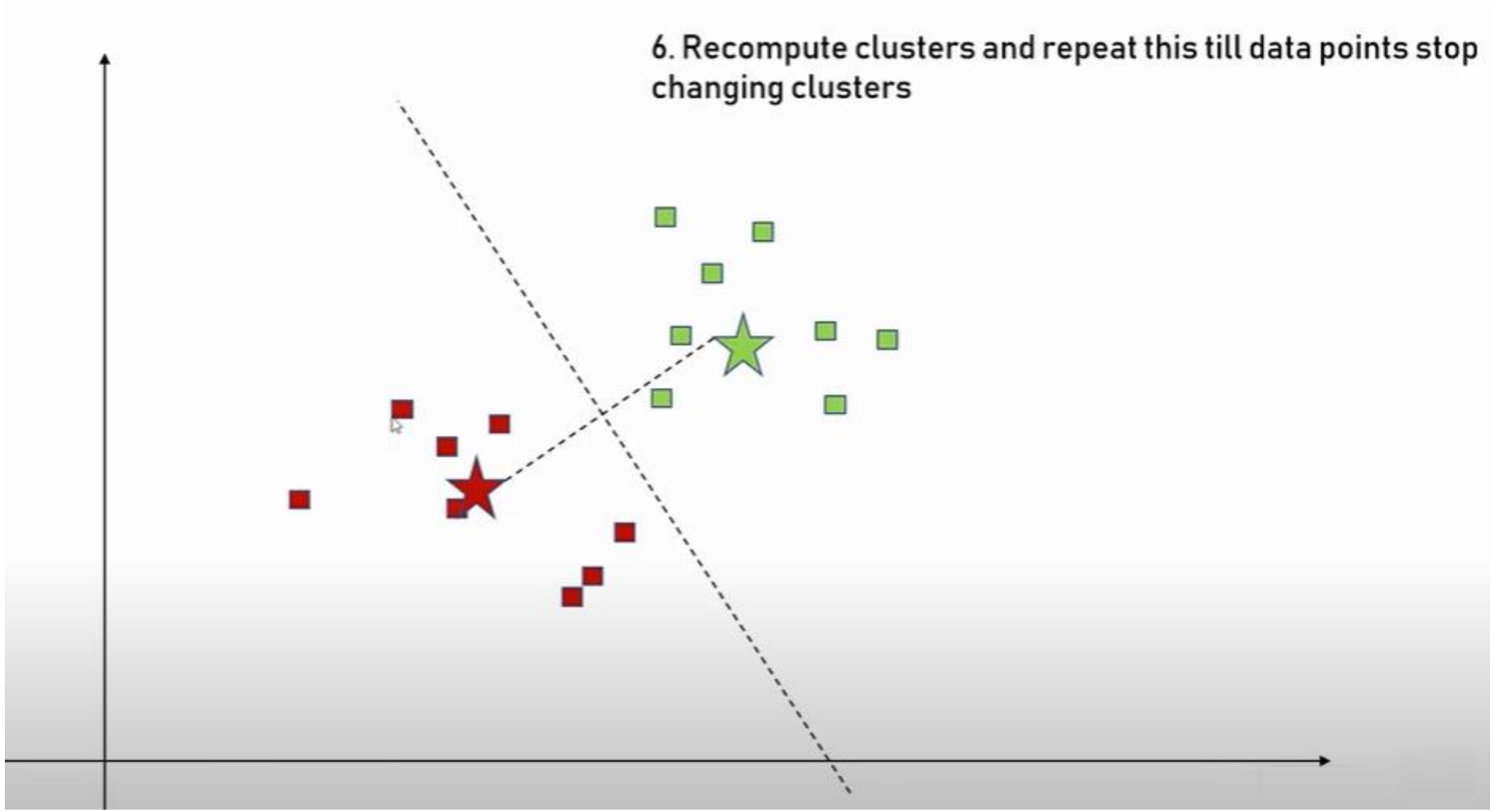

1. Start with K centroids by putting them at random place,
Here k = 2

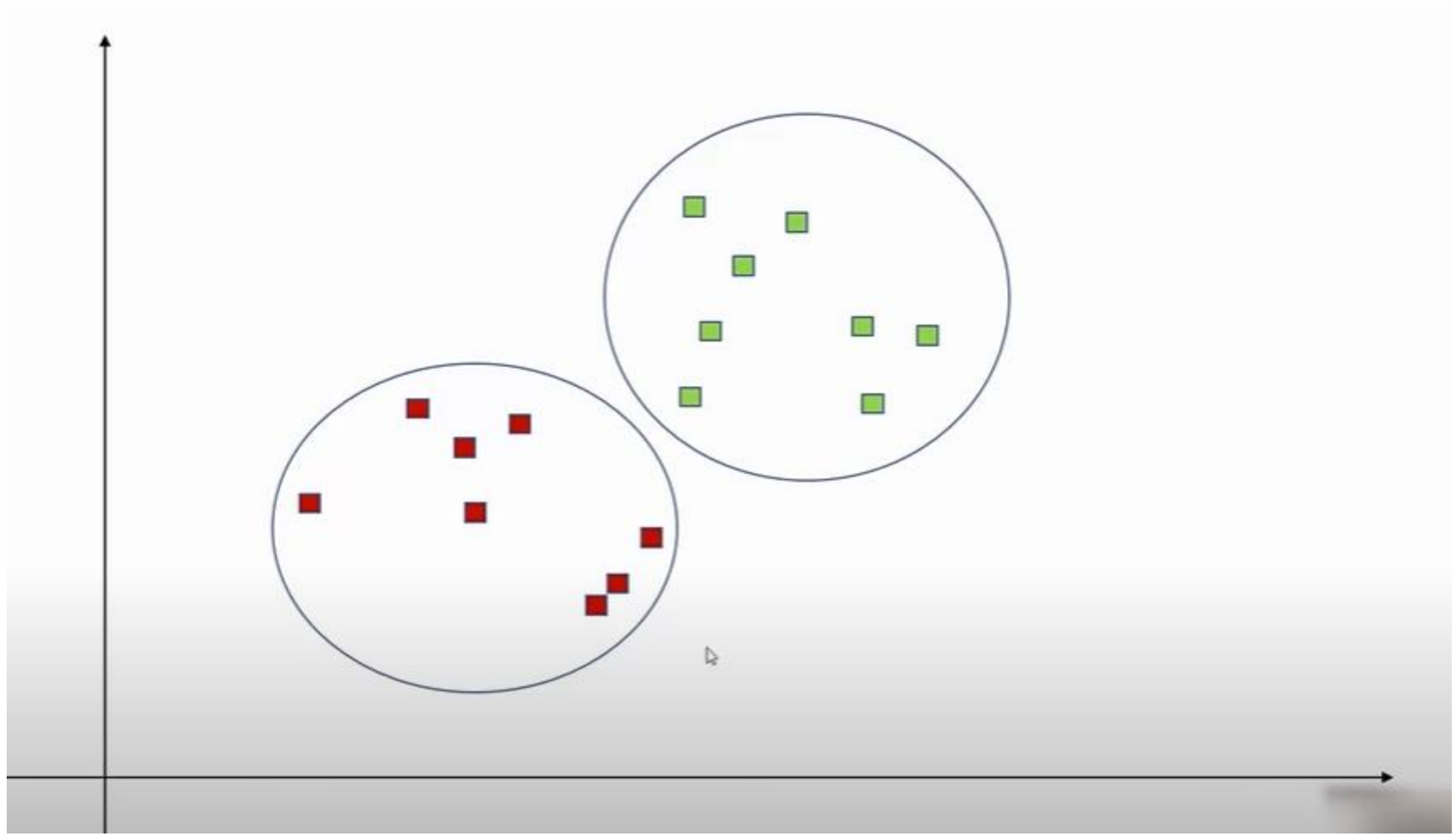

2. Compute distance of every point from centroid and cluster them accordingly


3. Adjust centroids so that they become center of gravity for given cluster


4. Again re-cluster every point based on their distance with centroid

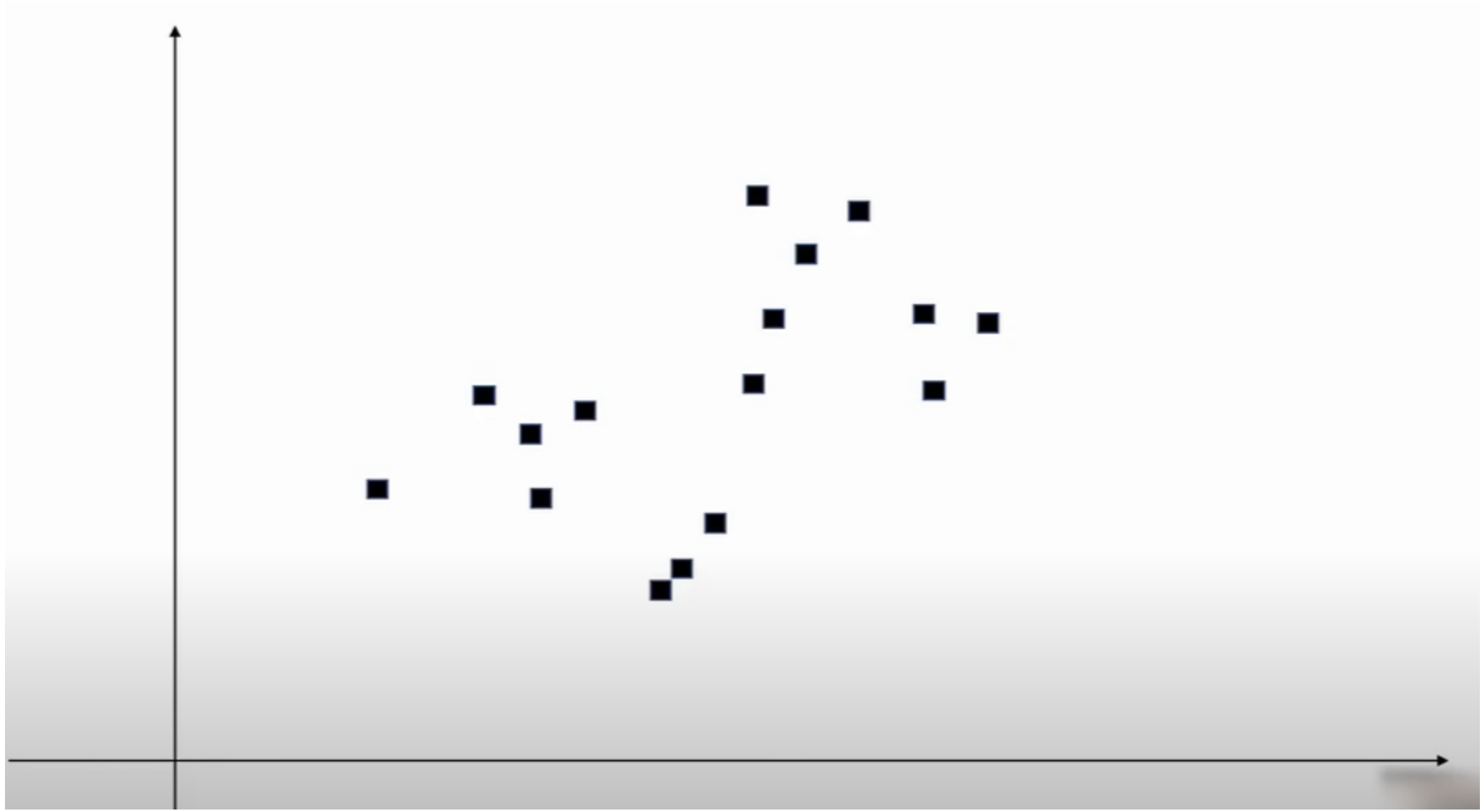

5. Again adjust centroids

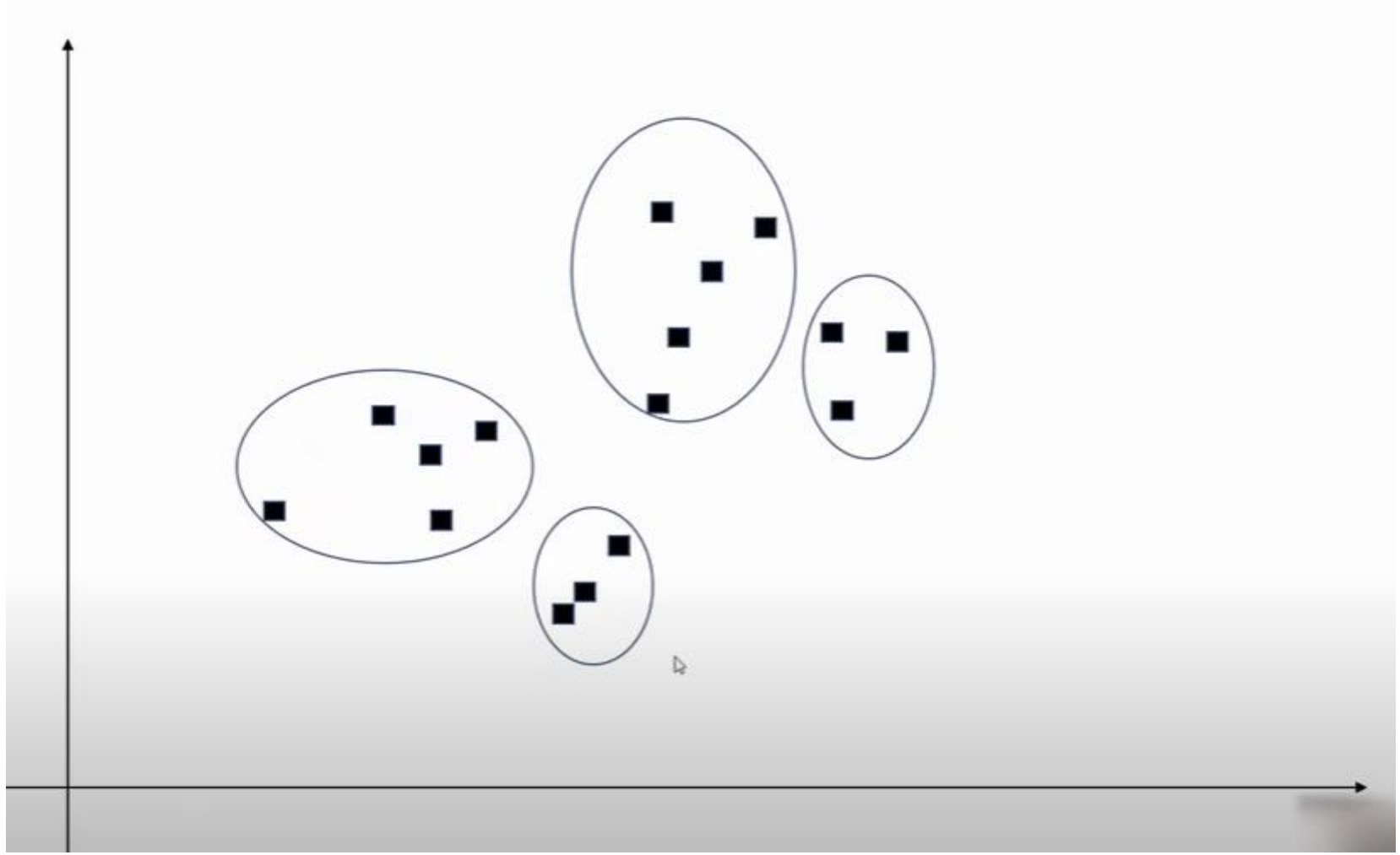

6. Recompute clusters and repeat this till data points stop changing clusters

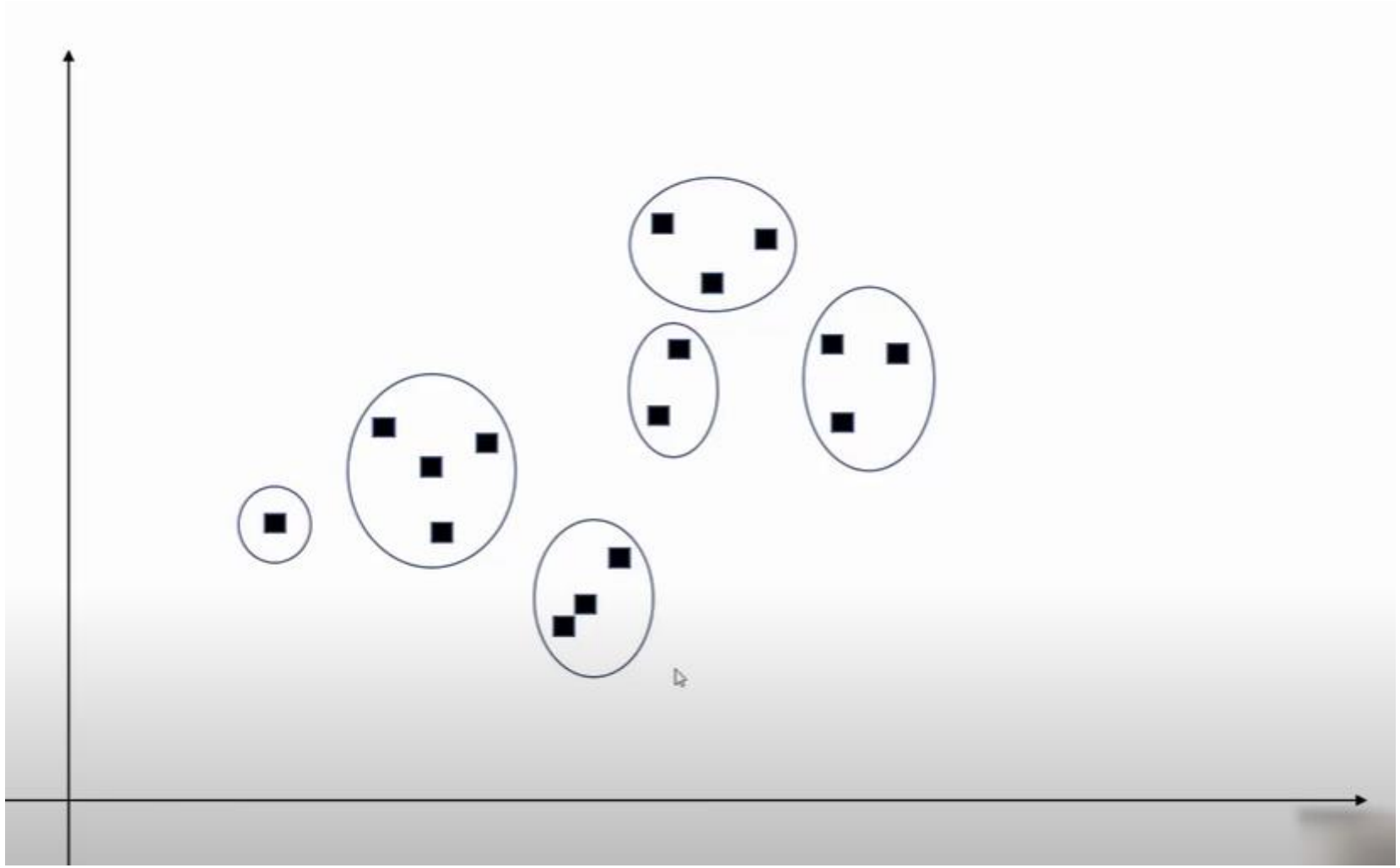


How to determine

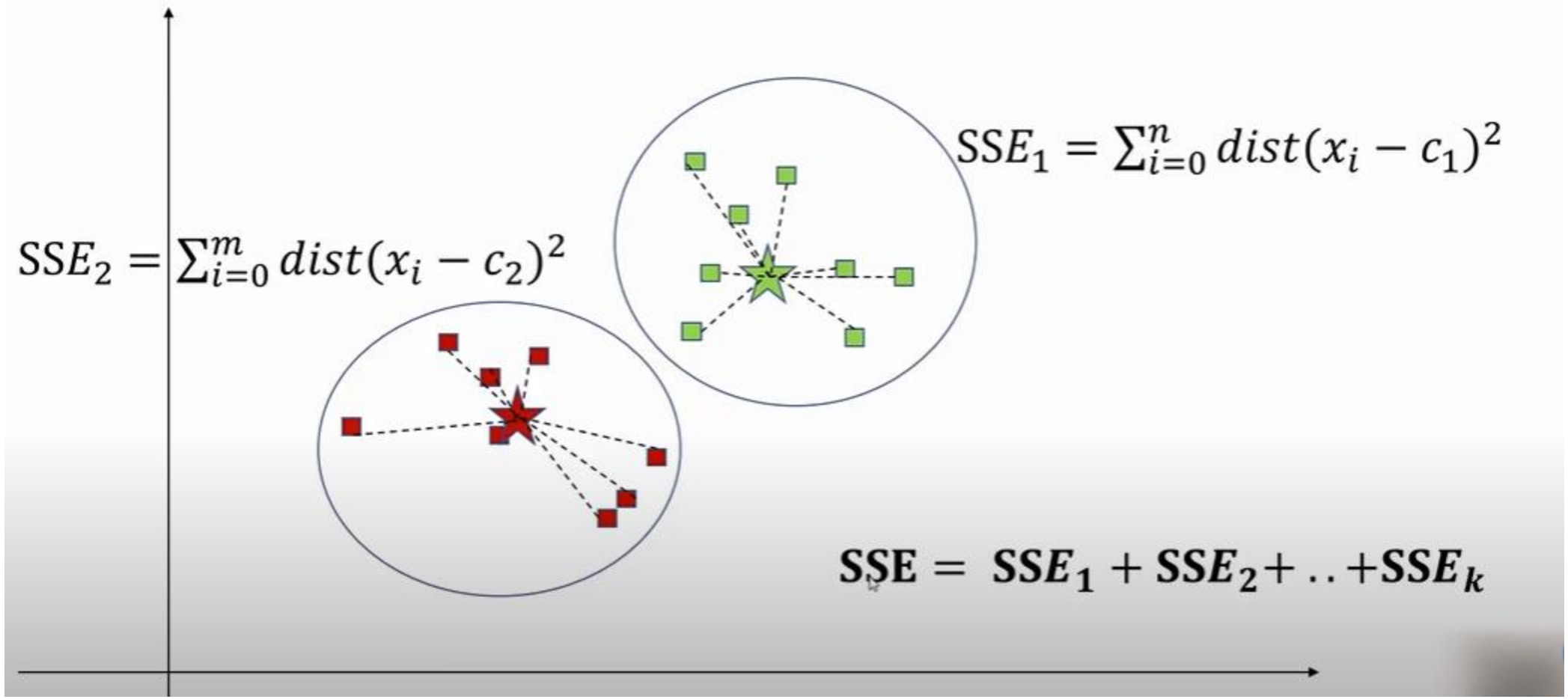
The correct number of clusters (k)?

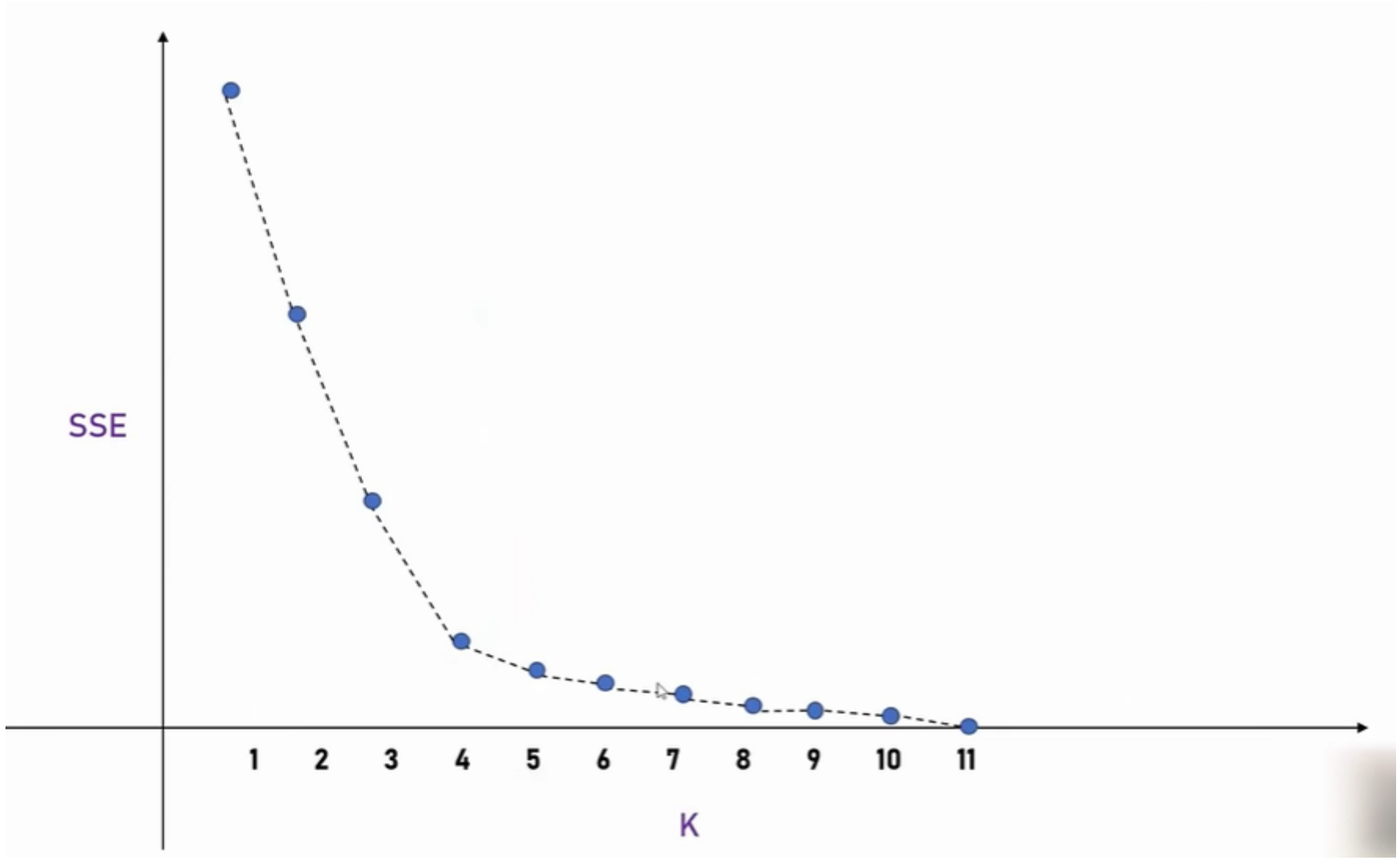




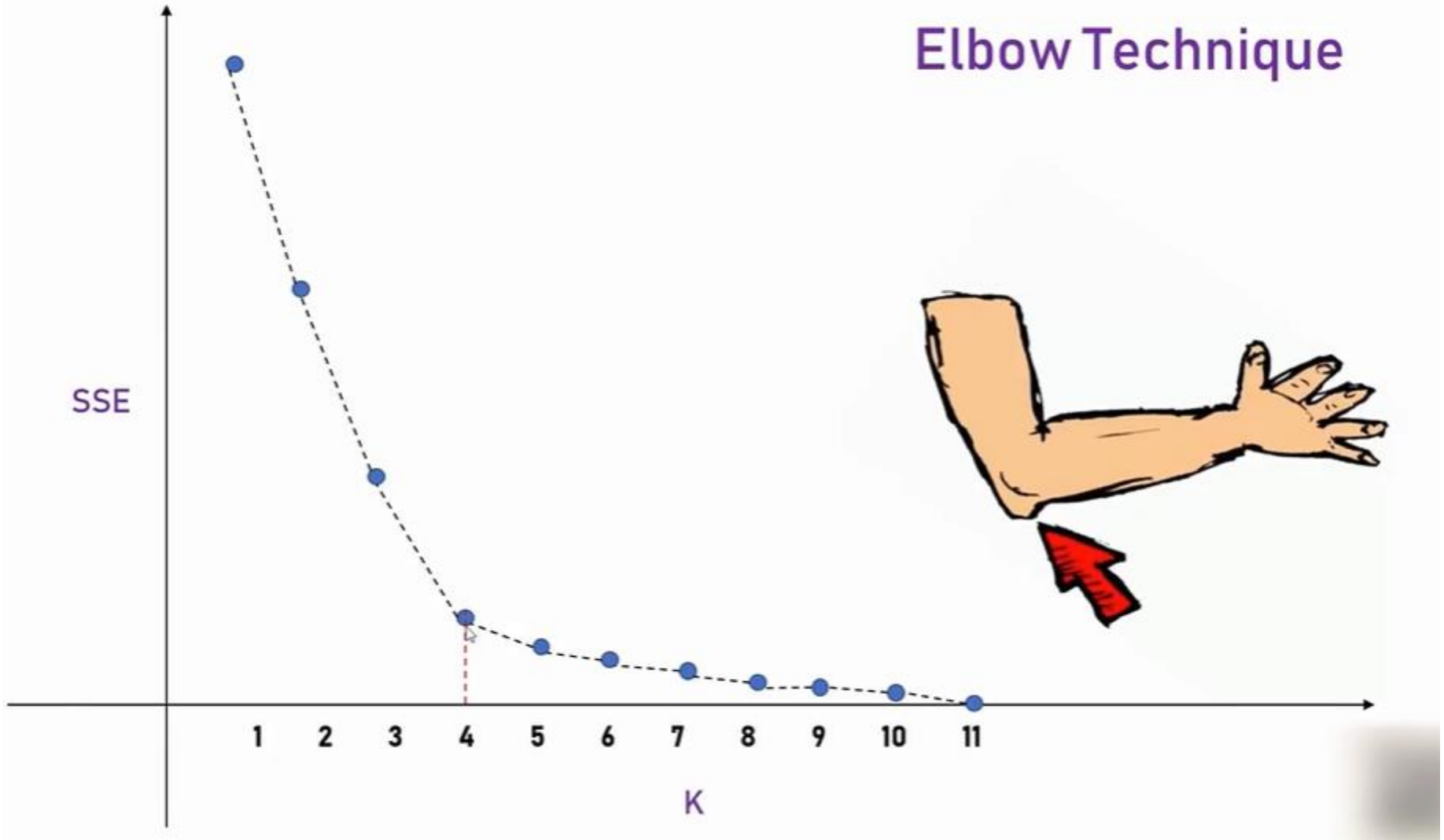


SSE = Sum of Squared Errors




SSE
1
2
3
4
5
6
7
8
9
10
11
K

Elbow Technique

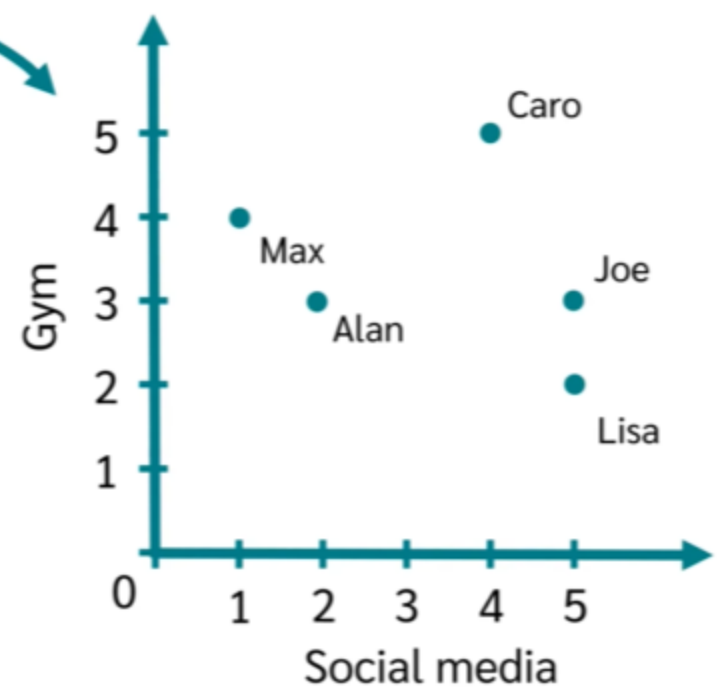


Hierarchical Clustering Algorithm

We asked **people** how many **hours a week** they spend on **social media** platforms and in the **gym**.

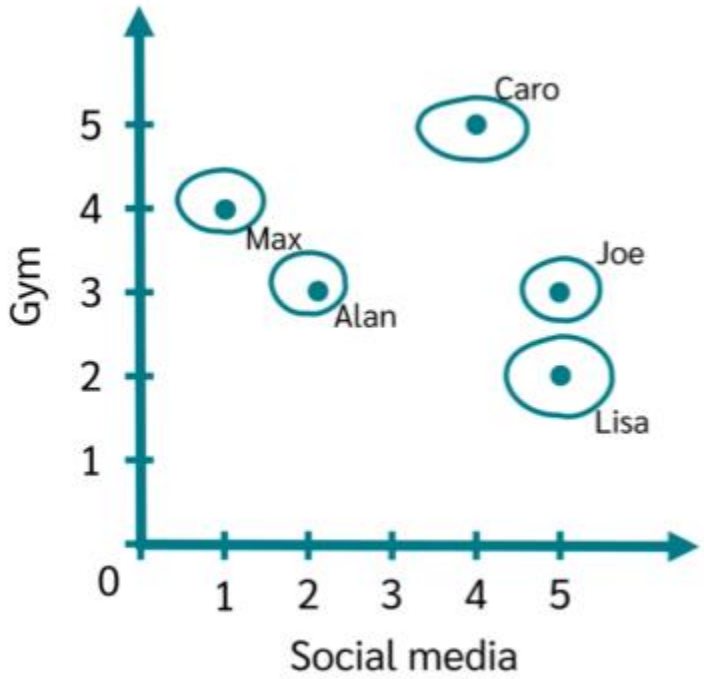
	Social media	Gym
Alan	2	3
Lisa	5	2
Joe	5	3
Max	1	4
Caro	4	5

	Social media	Gym
Alan	2	3
Lisa	5	2
Joe	5	3
Max	1	4
Caro	4	5



First step

- Assign a cluster to each individual point

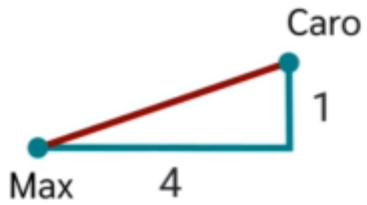


So we have as many **clusters**

Measuring the distance between two points

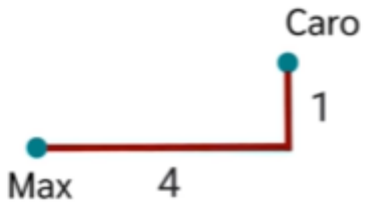
Euclidean distance

$$d = \sqrt{4^2 + 1^2} = 3{,}162$$



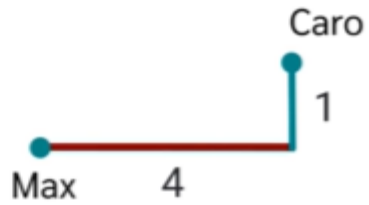
Manhattan distance

$$d = 4 + 1 = 5$$



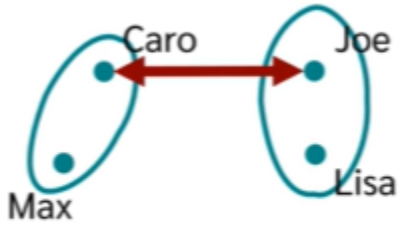
Maximum distance

$$d = max(4\,, 1) = 4$$



Measuring the distance between two clusters

Single-linkage uses the distance between the **closest elements** in the cluster.

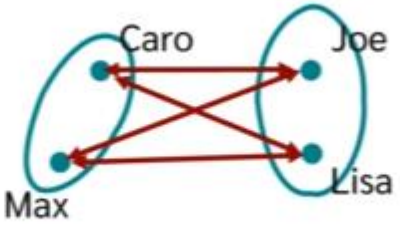


So, the distance between Caro and Joe.

Complete Linkage uses the distance between the **most distant elements** of the cluster.

So, between Max and Joe.

Average linkage uses the average of all pairwise distances.



From each combination the distance is calculated and from it the average value.

The distance matrix

The distance between **Alan and Lisa** is given by:

$$d = \sqrt{(5-2)^2+(2-3)^2} = 3{,}16$$

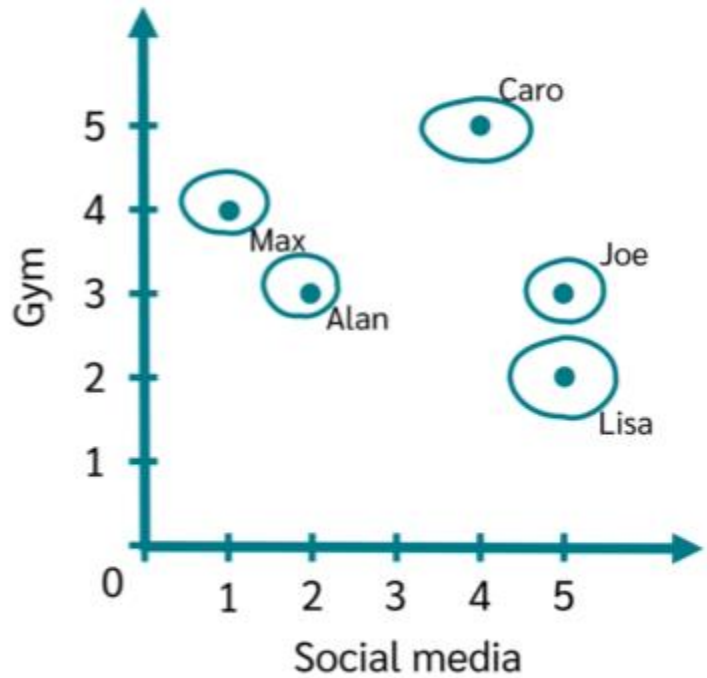
	Social media	Gym
Alan	2	3
Lisa	5	2
Joe	5	3
Max	1	4
Caro	4	5

	Alan	Lisa	Joe	Max	Caro
Alan	0				
Lisa	3,16	0			
Joe	3,00	1,00	0		
Max	1,41	4,47	4,12	0	
Caro	2,83	3,16	2,24	3,16	0

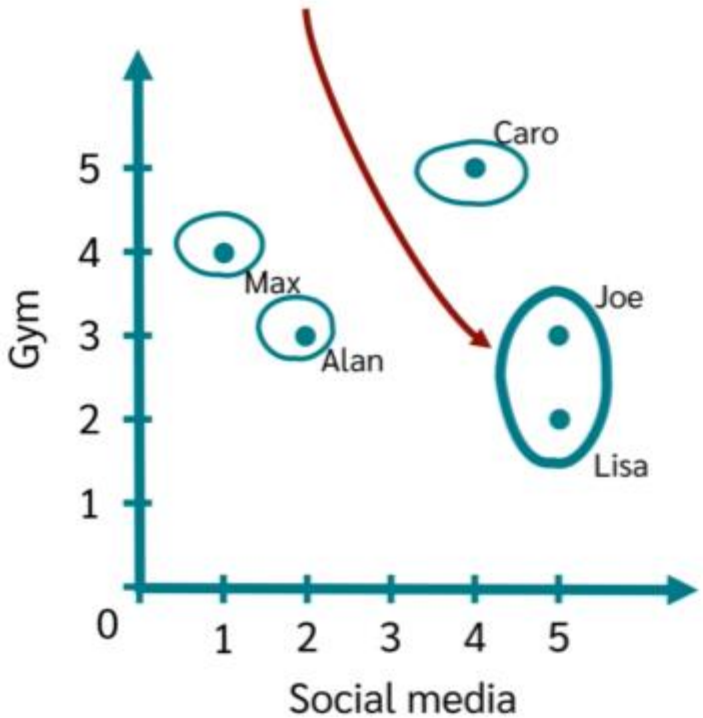
Now we can **merge** the **first clusters**.

For this, we look at which two clusters have the **smallest distance** between them.

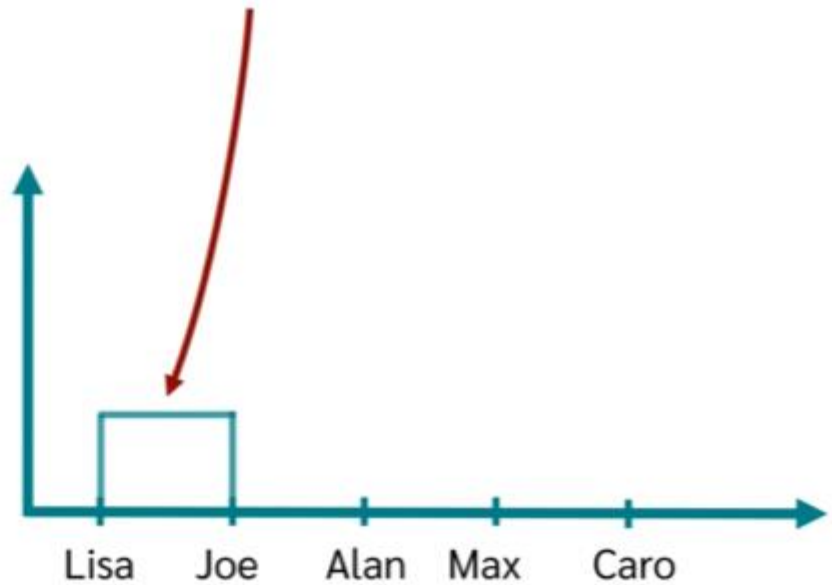
	Alan	Lisa	Joe	Max	Caro
Alan	0				
Lisa	3,16	0			
Joe	3,00	1,00	0		
Max	1,41	4,47	4,12	0	
Caro	2,83	3,16	2,24	3,16	0



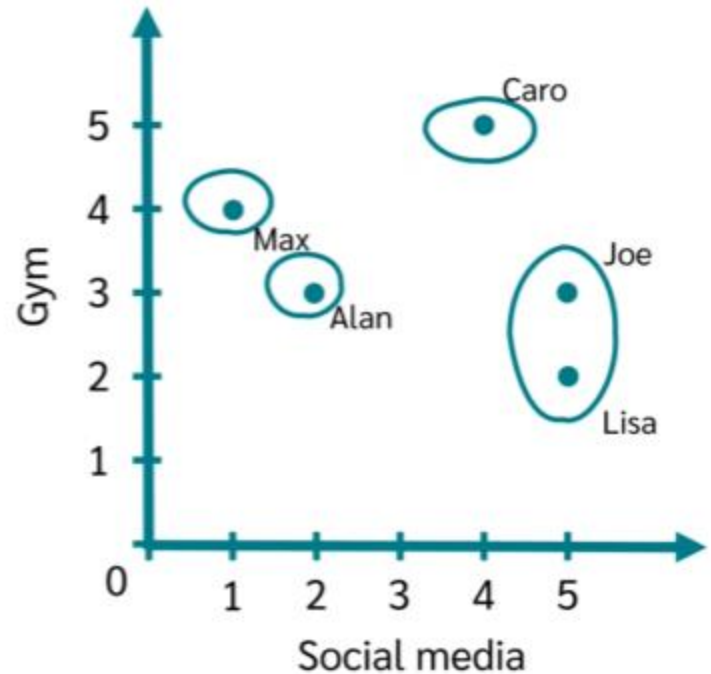
With this, we now connect Joe and Lisa to **form a cluster**.



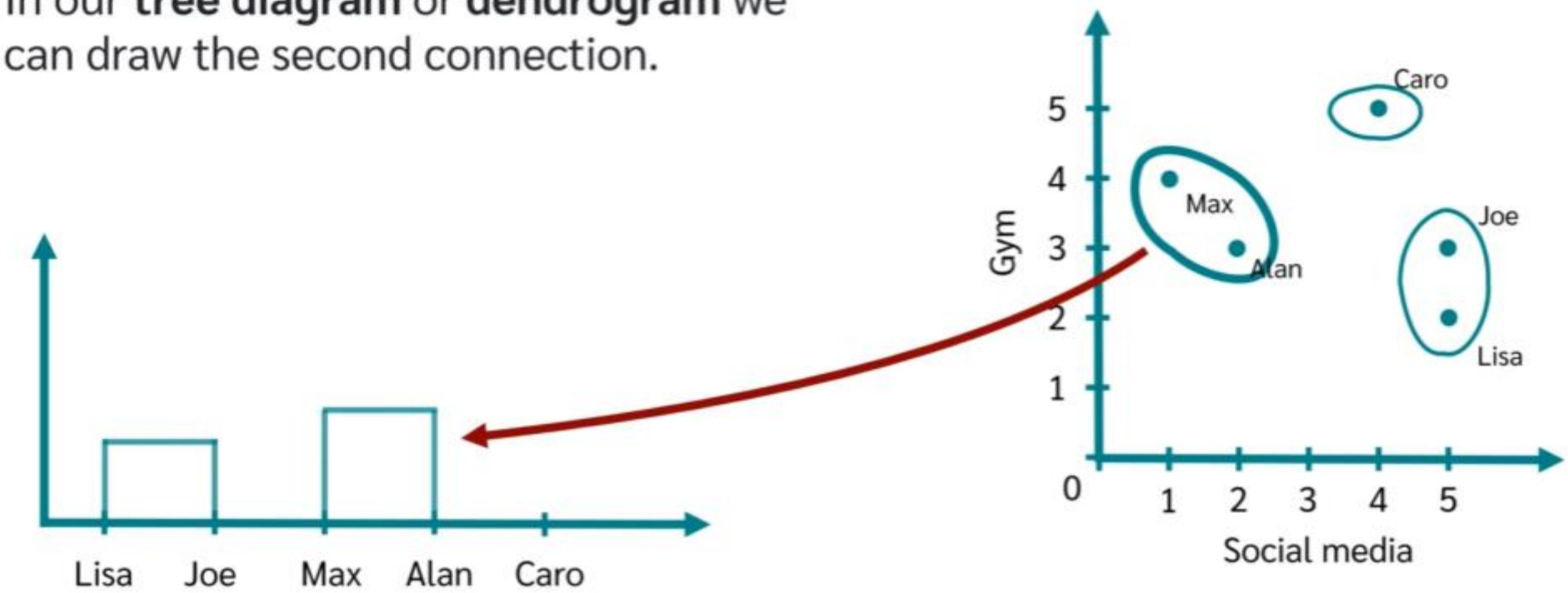
In our **tree diagram** or **dendrogram** we can draw the first connection.

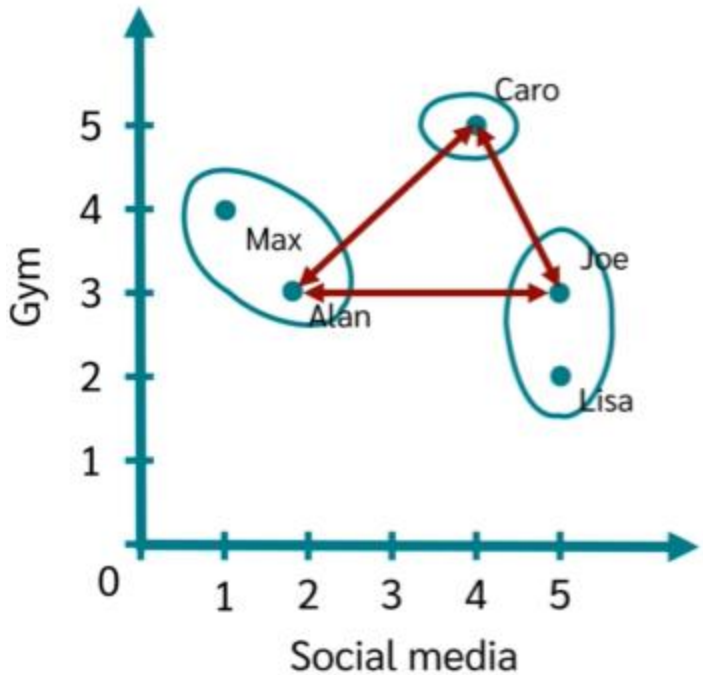


	Alan	Lisa, Joe	Max	Caro
Alan	0			
Lisa, Joe	3,00	0		
Max	1,41	4,12	0	
Caro	2,83	2,24	3,16	0



In our **tree diagram** or **dendrogram** we can draw the second connection.





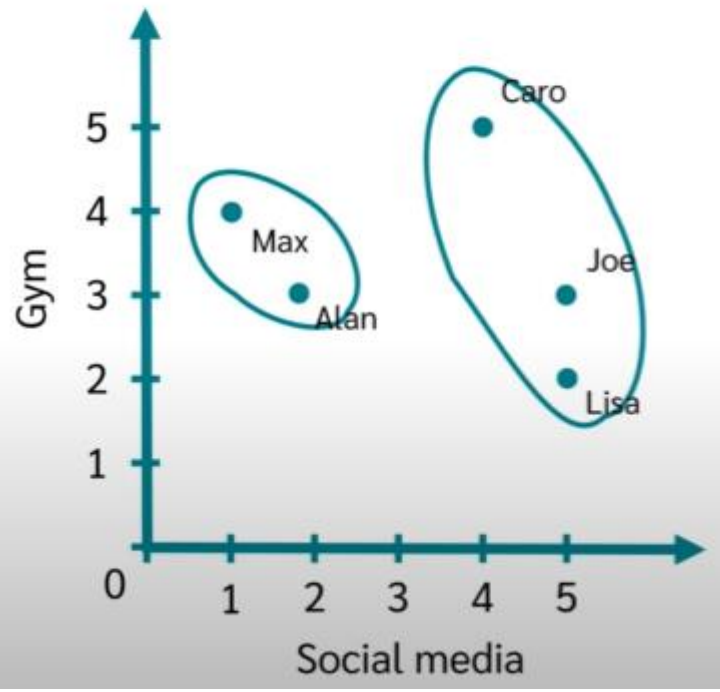
Now we **update** the

distance matrix again.

We calculate the distance between

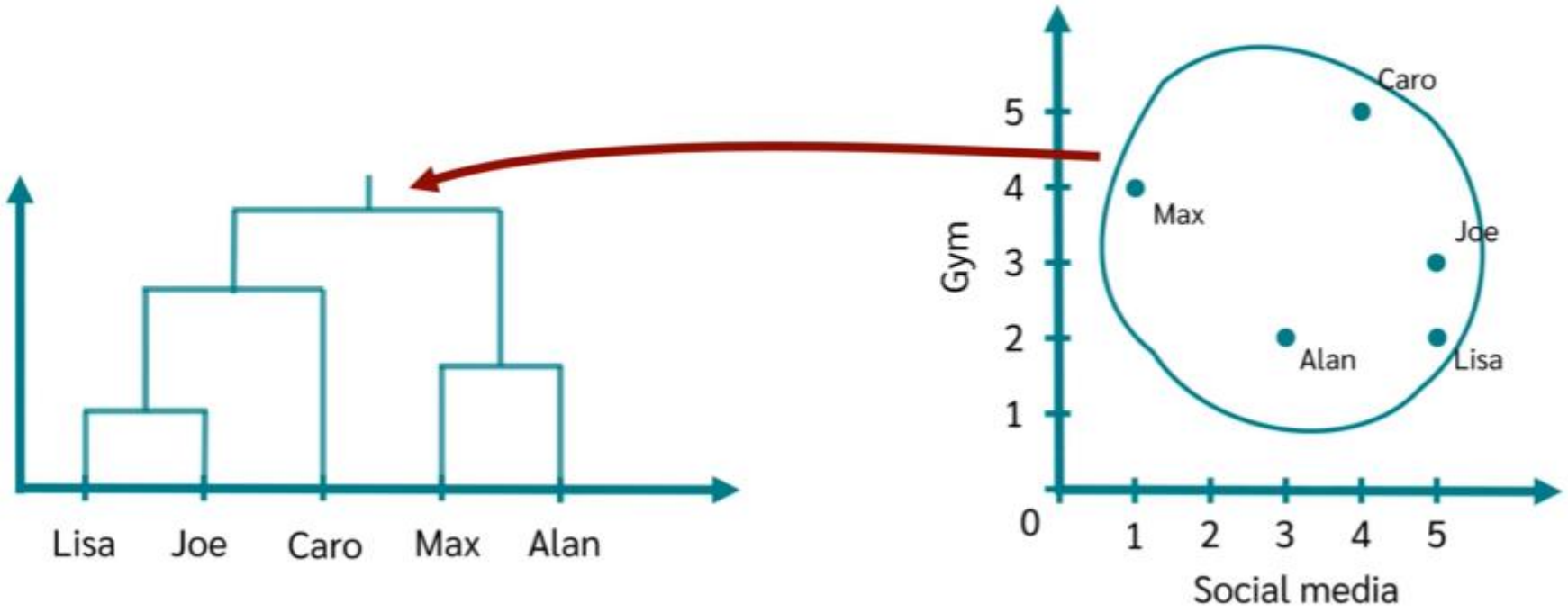
	Lisa, Joe	Max, Alan	Caro
Lisa, Joe	0		
Max, Alan	3,00	0	
Caro	2,24	2,83	0

So, we **connect** these **two clusters**

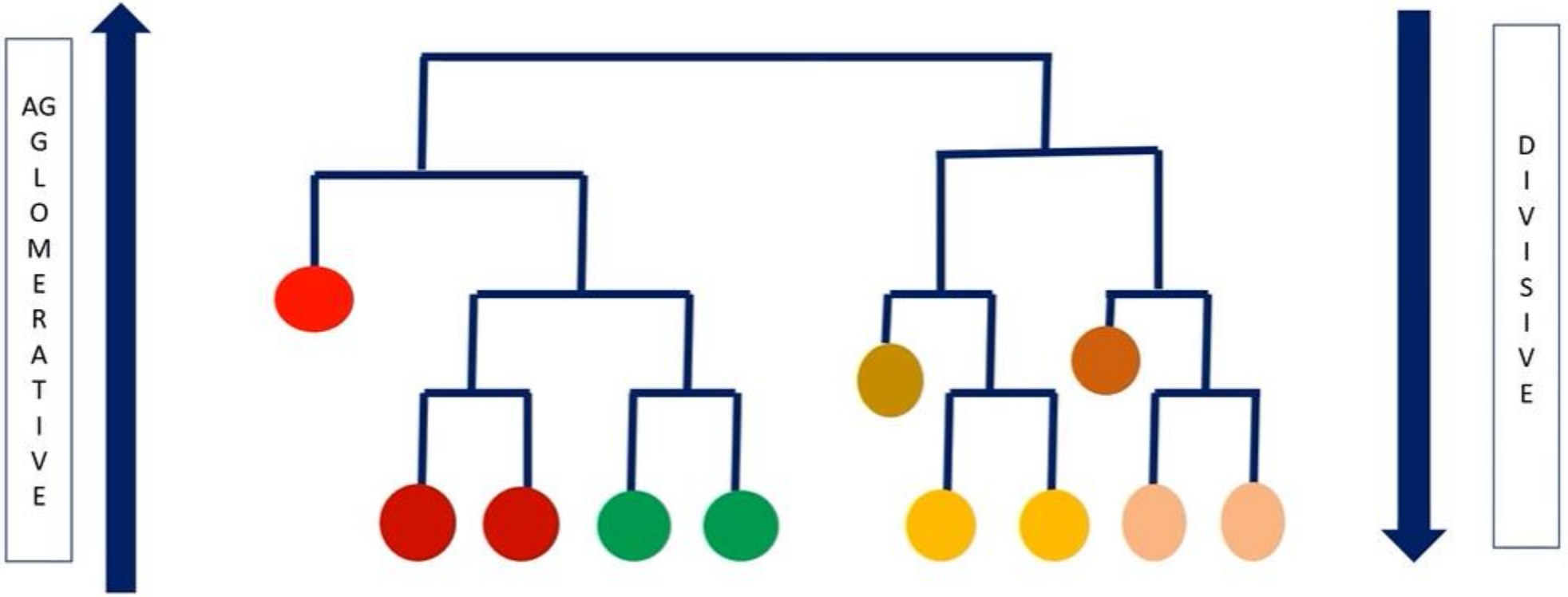


and draw the **third connection** in the **tree diagram.**

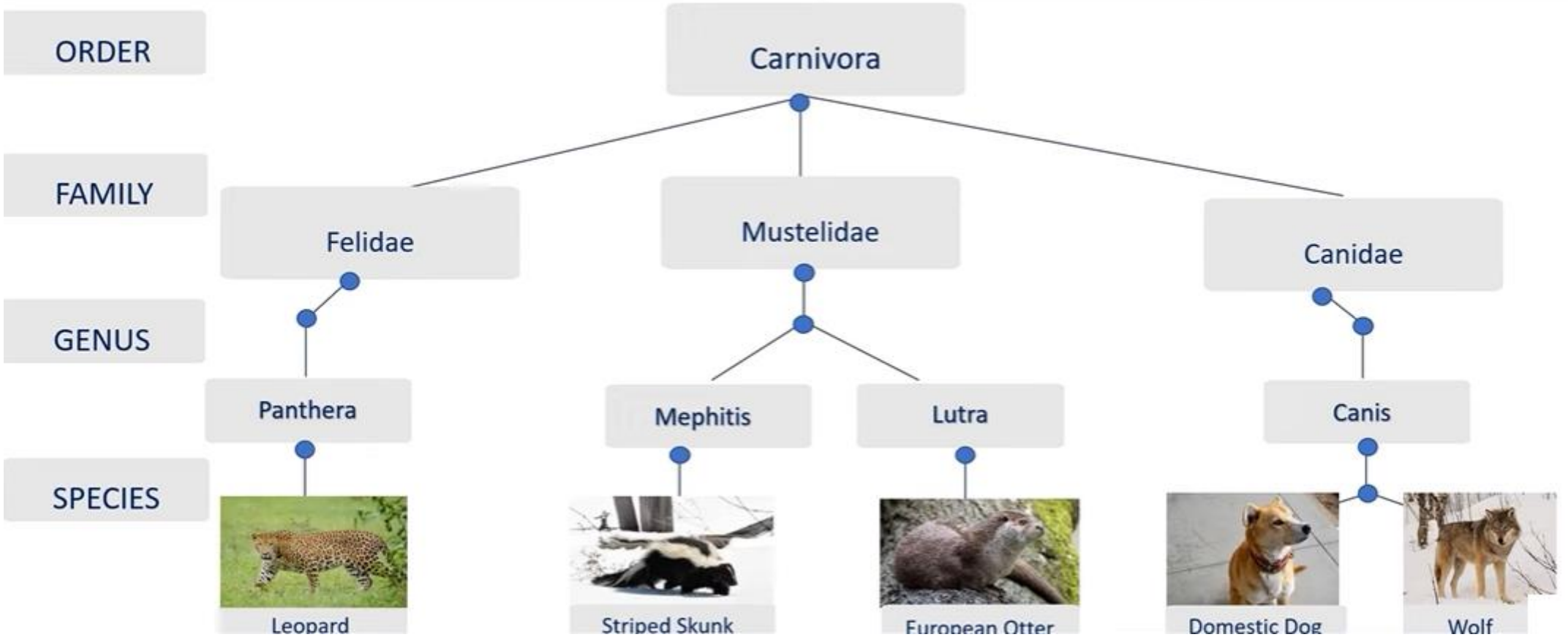
Finished dendrogram



Agglomerative and Divisive Clustering



Example


ORDER
FAMILY
GENUS
SPECIES
Carnivora
Felidae
Mustelidae
Canidae
Panthera
Mephitis
Lutra
Canis
Leopard
Striped Skunk
European Otter
Domestic Dog
Wolf