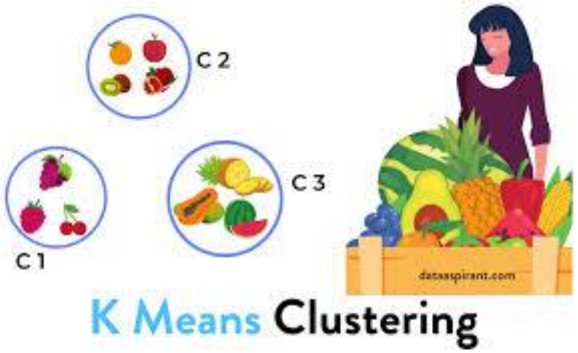
Data Science



Lecture 09:*K-means Clustering*

Dr.Fatema Nafa

Fall 2022

Learning Objectives

- **Motivating Example**
- What is clustering?
- Why would we want to cluster?
- How would you determine clusters?
- How can you do this efficiently?

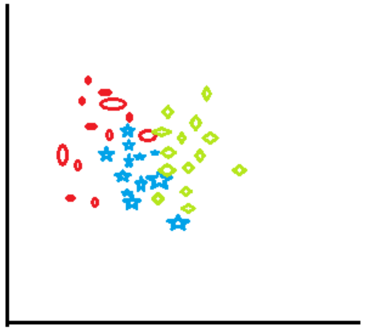


fig 1: before applying
k-means clustering

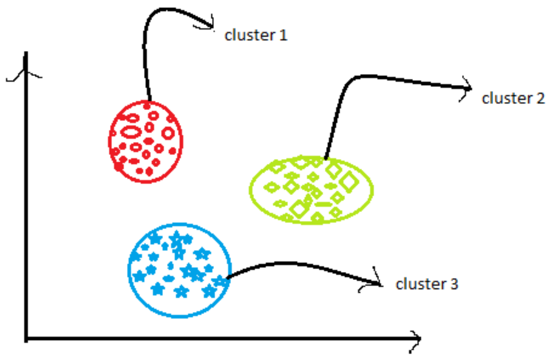
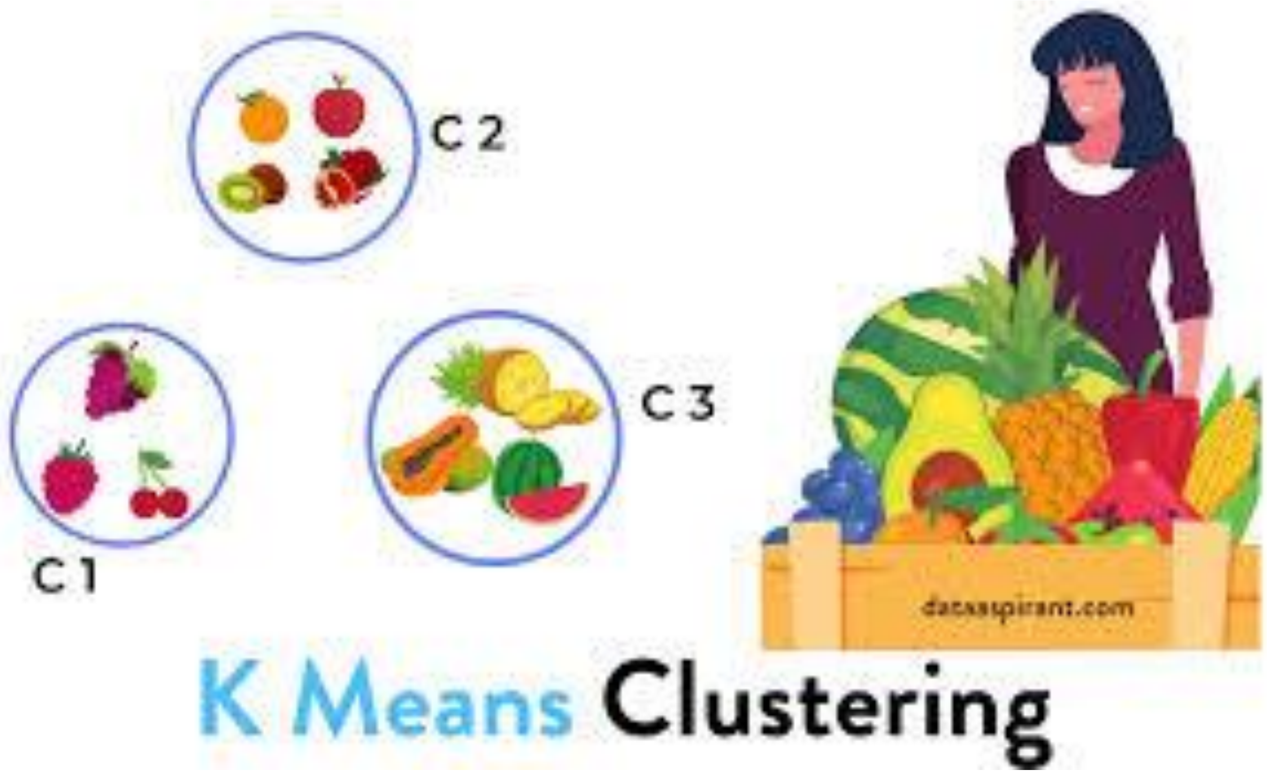


fig 2: After applying K-
means clustering


C 2
C 1
C 3
dataaspirant.com
K Means Clustering

Machine Learning Problems

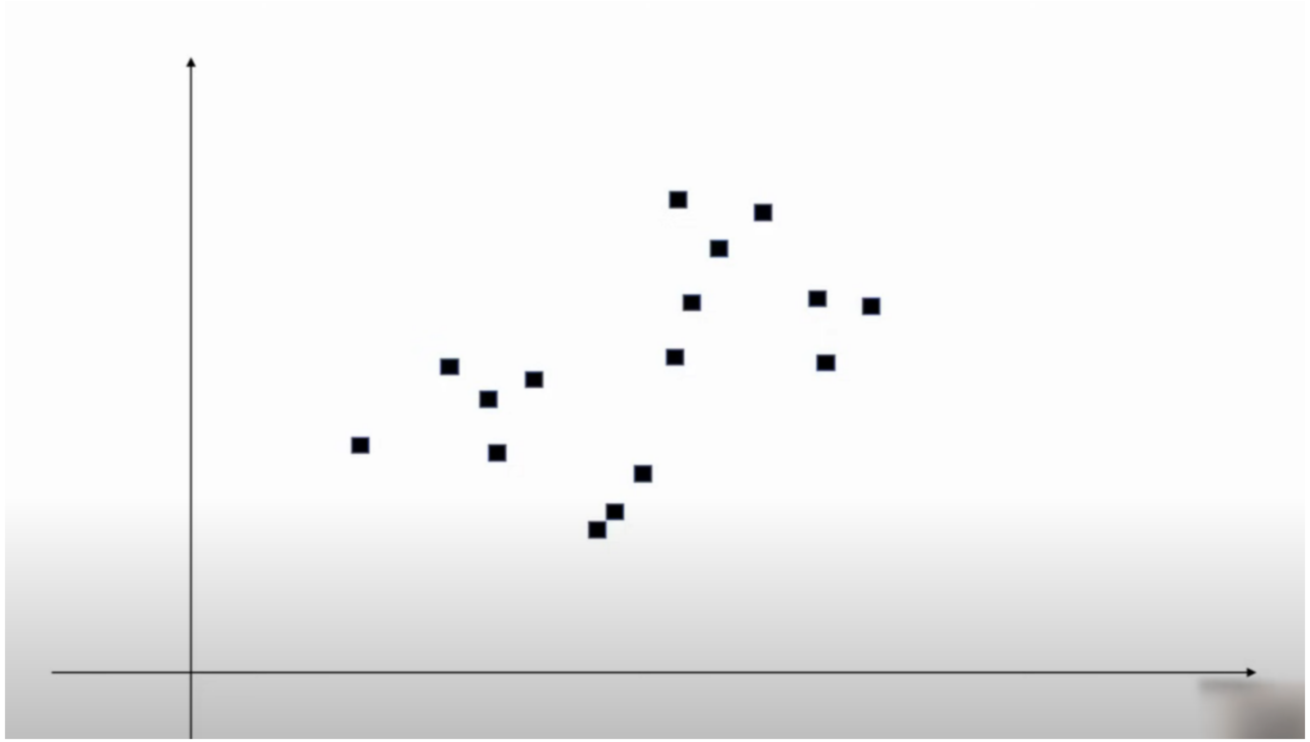
	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

Lecture Map

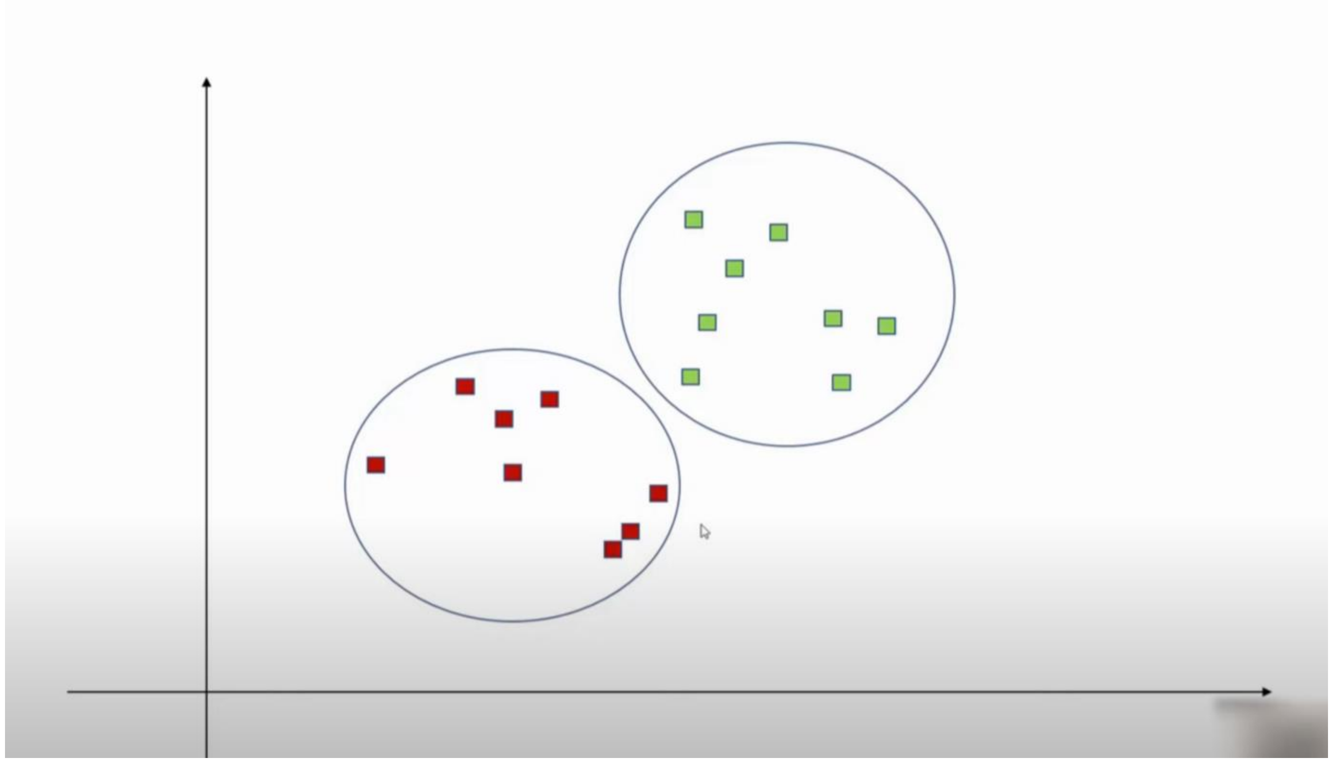
- Theory
- Coding
- Exercise

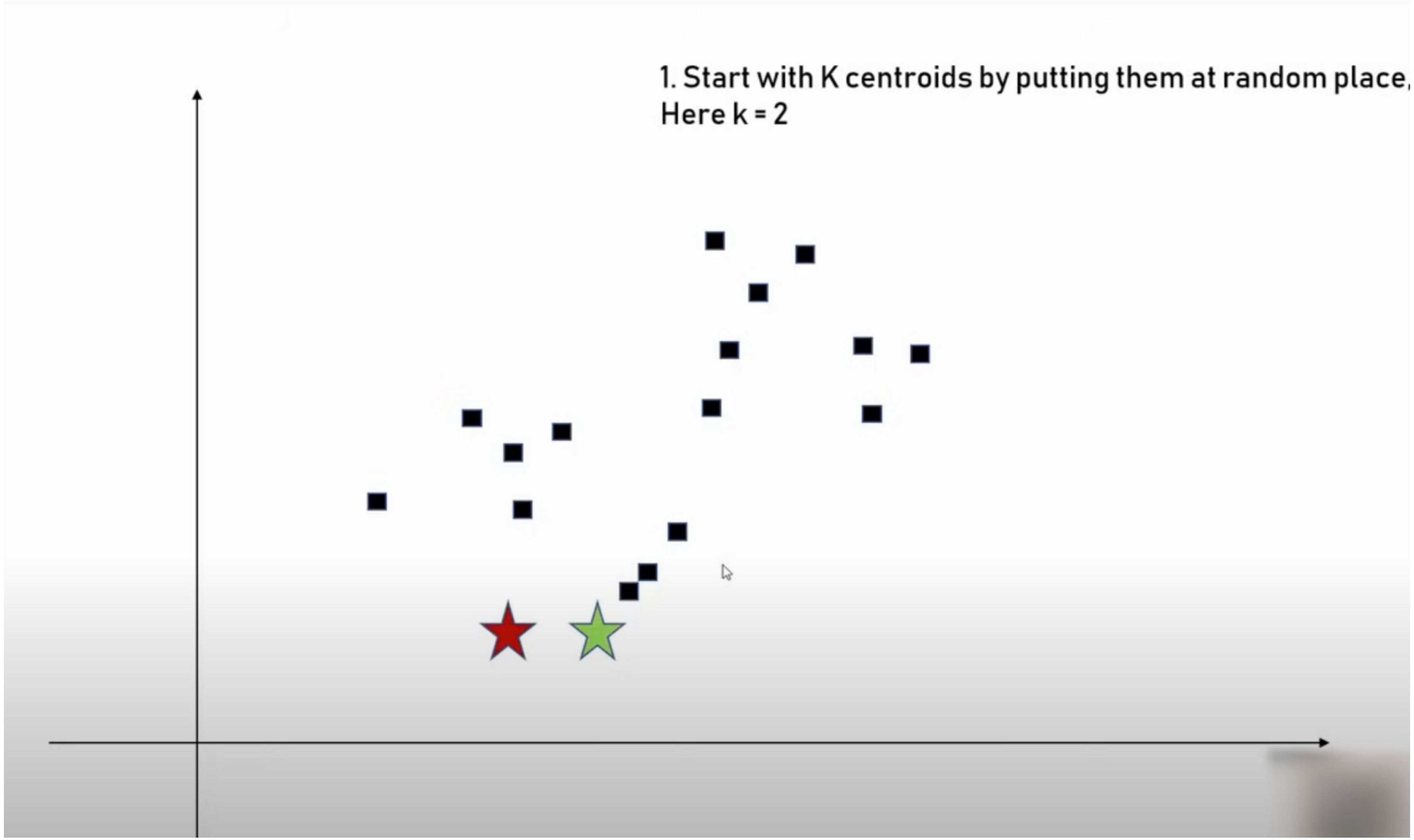
K-means Clustering

- Unsupervised learning
- Requires data, but no labels
- Detect patterns e.g. in
  - Group emails or search results
  - Customer shopping patterns
  - Regions of images
- Useful when don’t know what you’re looking for
- But: can get gibberish



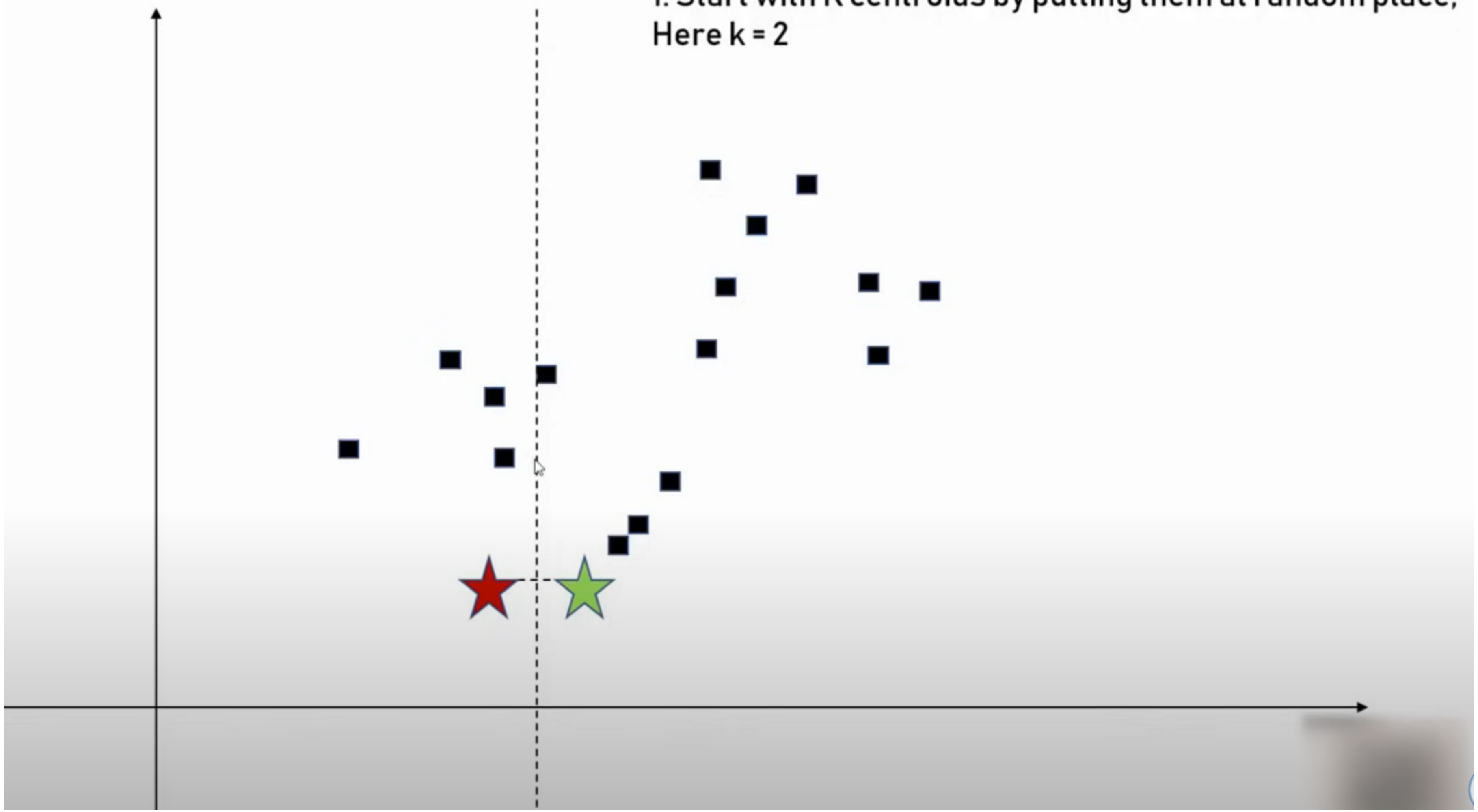
🤔

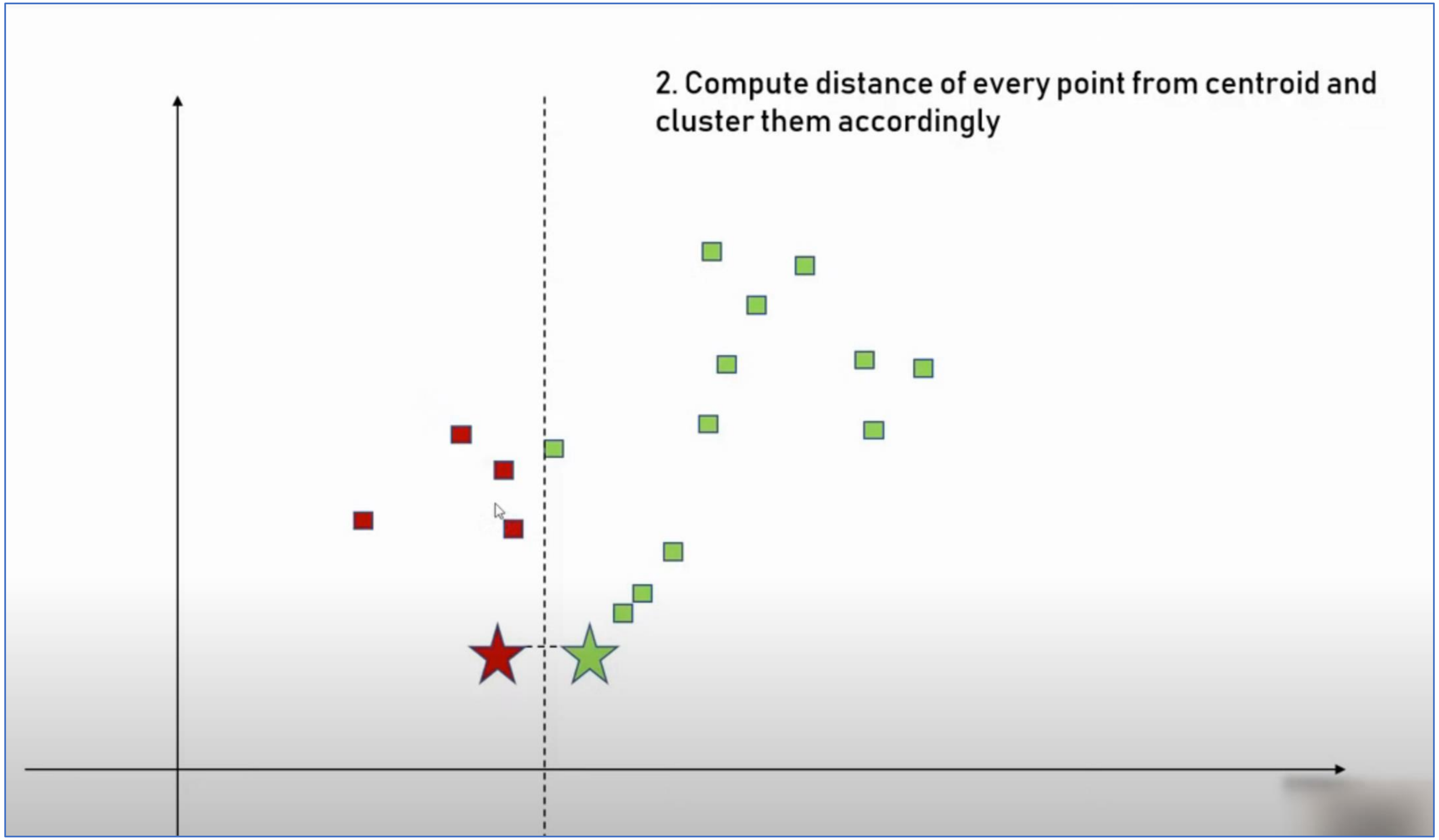




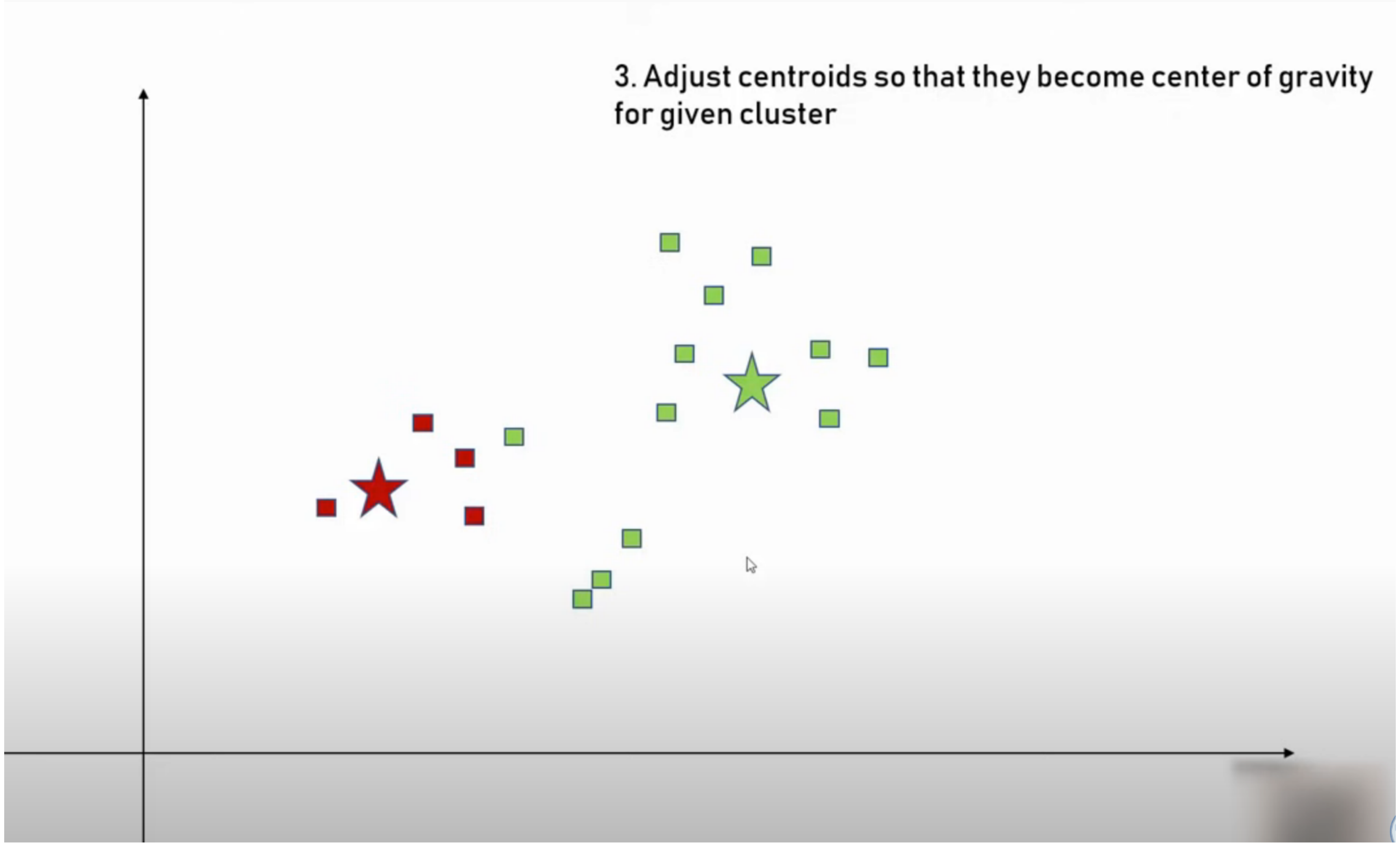
1. Start with K centroids by putting them at random place,
Here k = 2

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

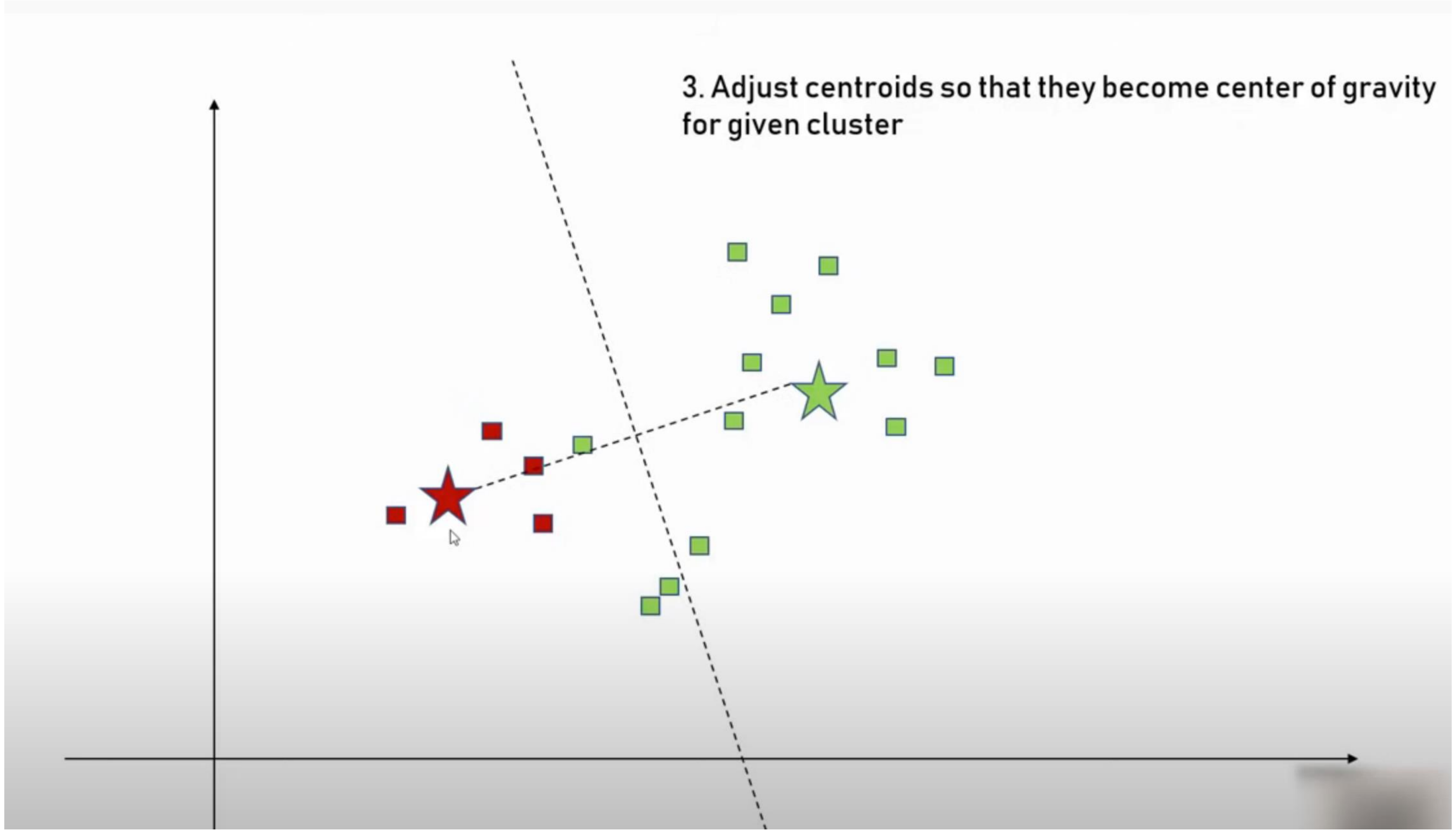




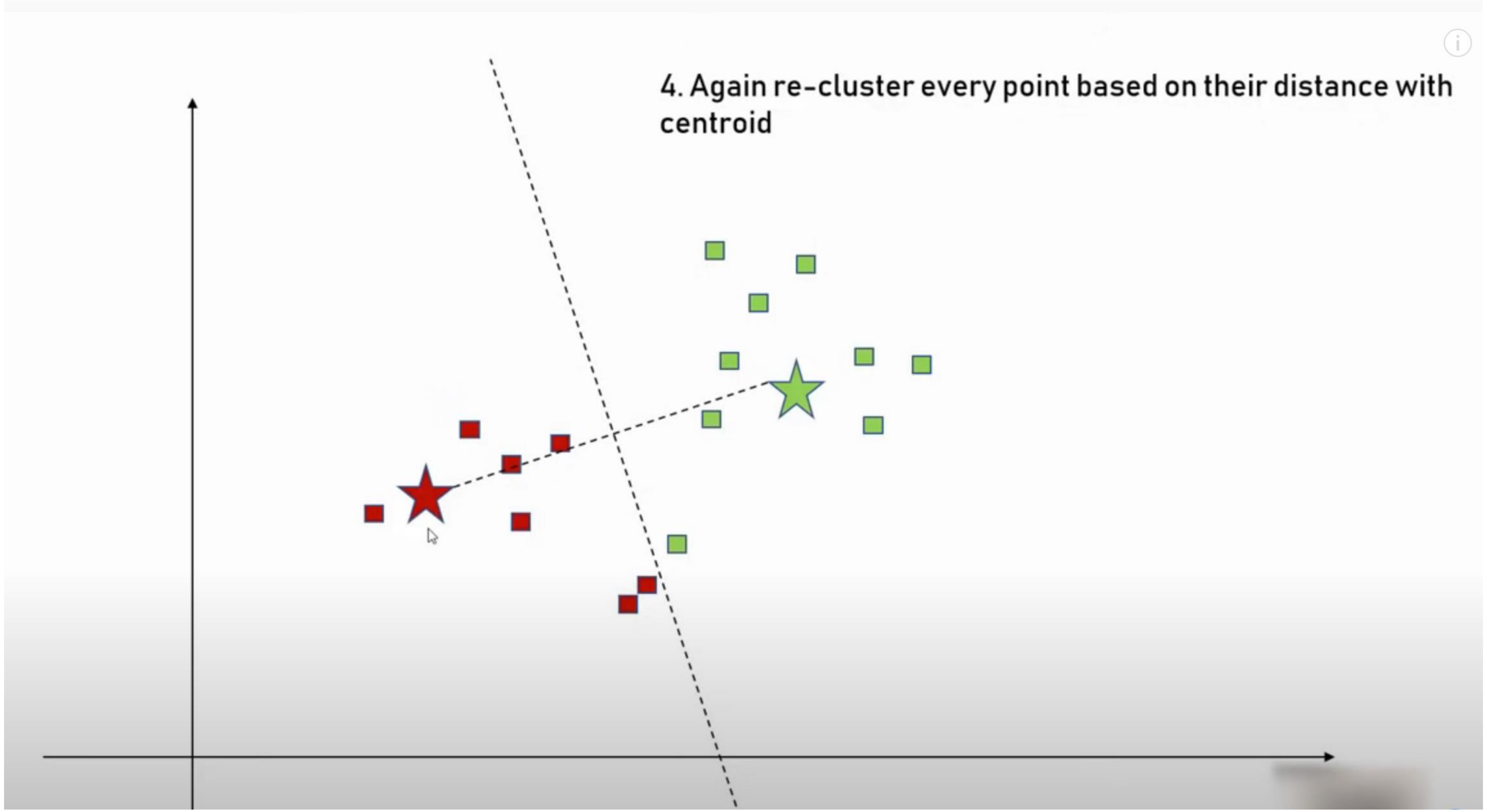
2. Compute distance of every point from centroid and cluster them accordingly



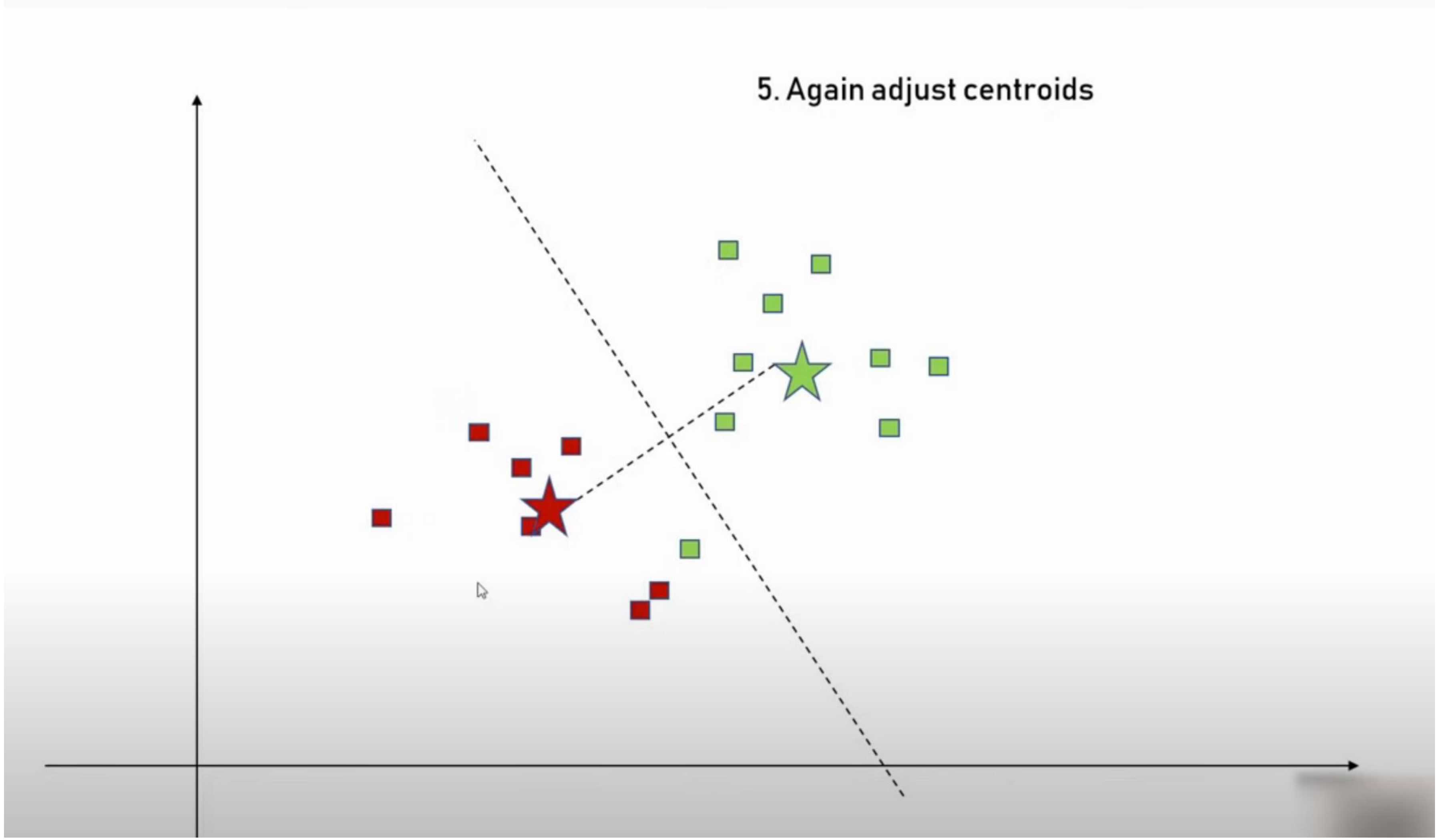
3. Adjust centroids so that they become center of gravity for given cluster



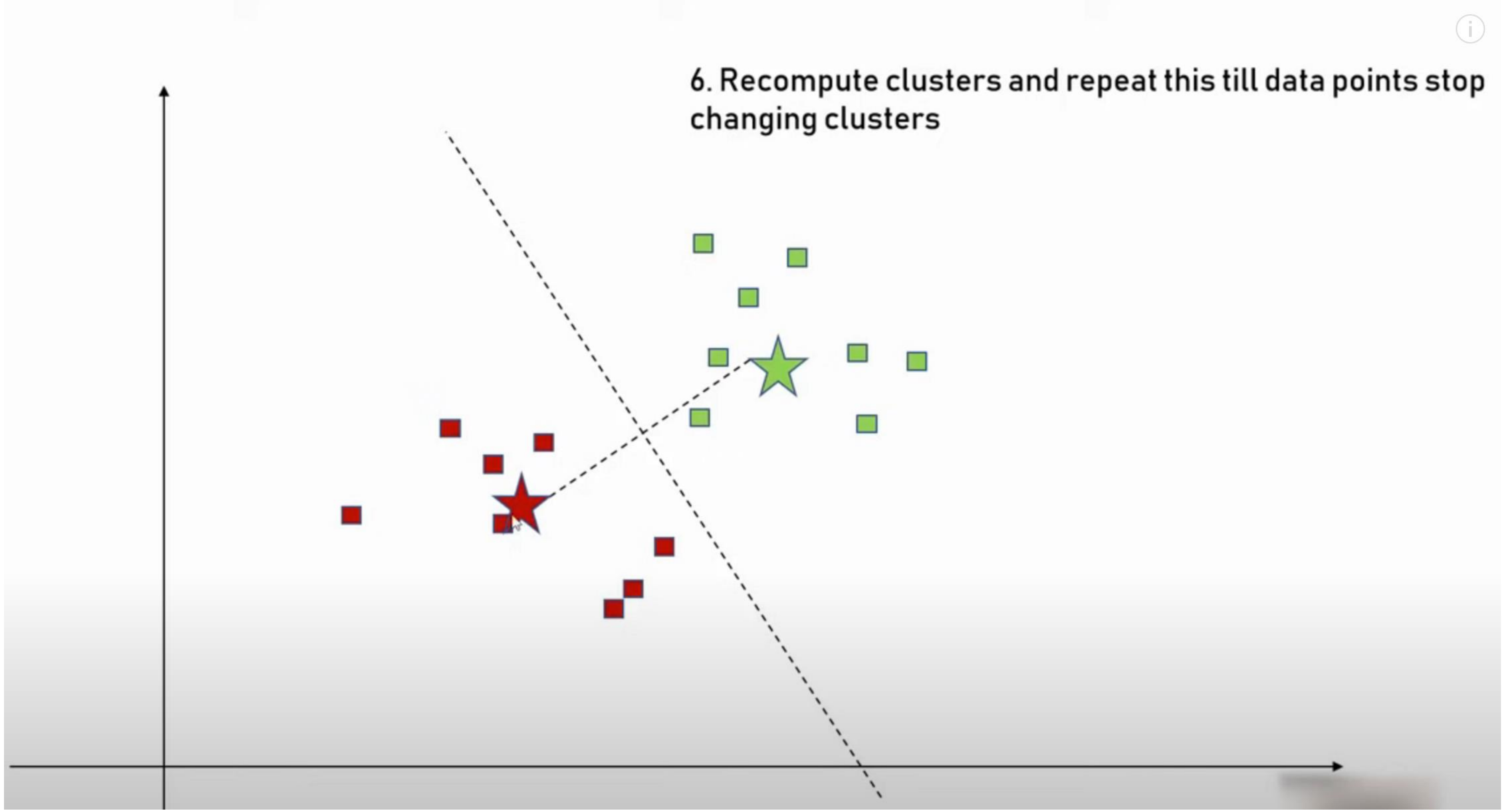
3. Adjust centroids so that they become center of gravity for given cluster



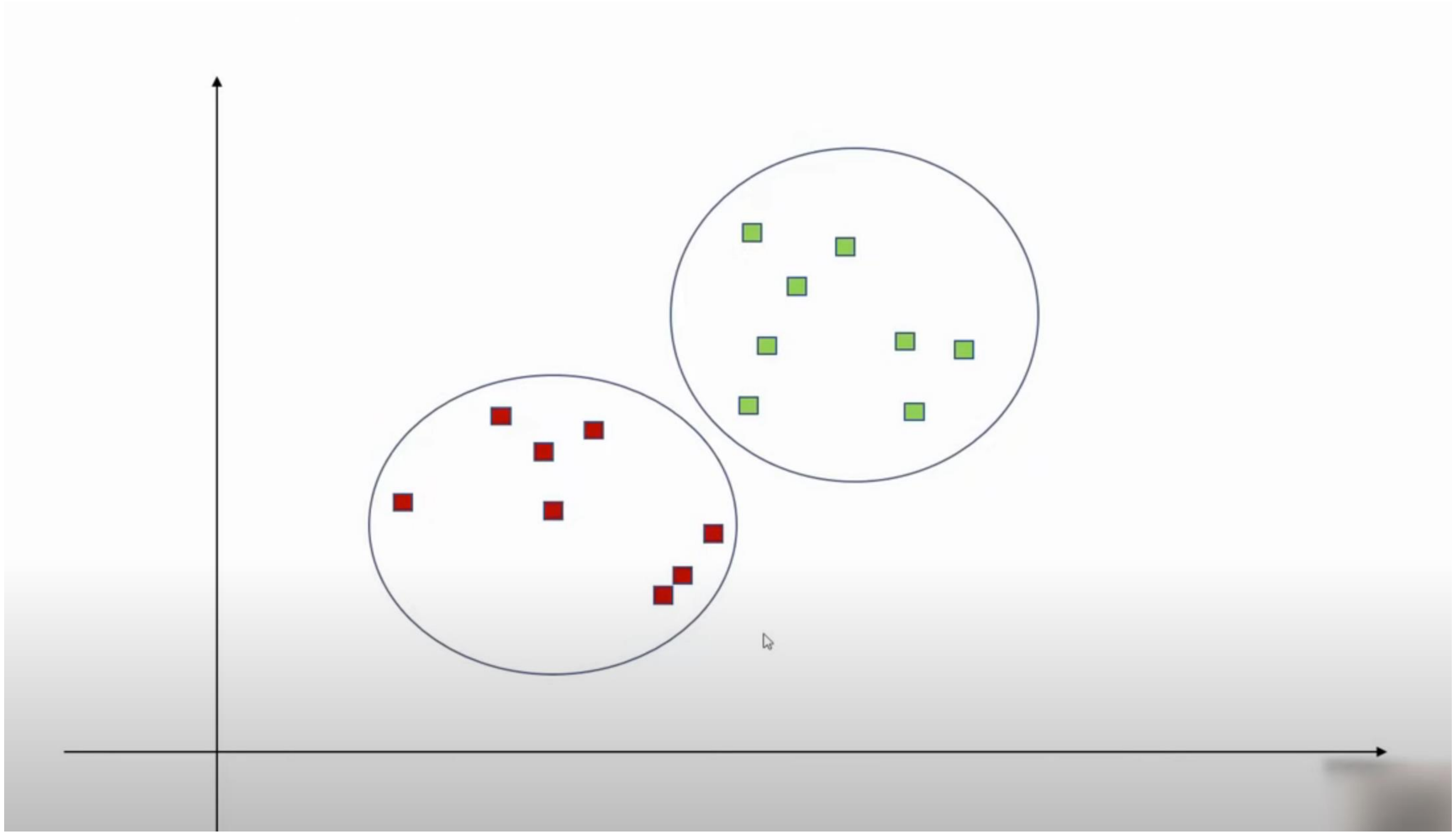
4. Again re-cluster every point based on their distance with centroid



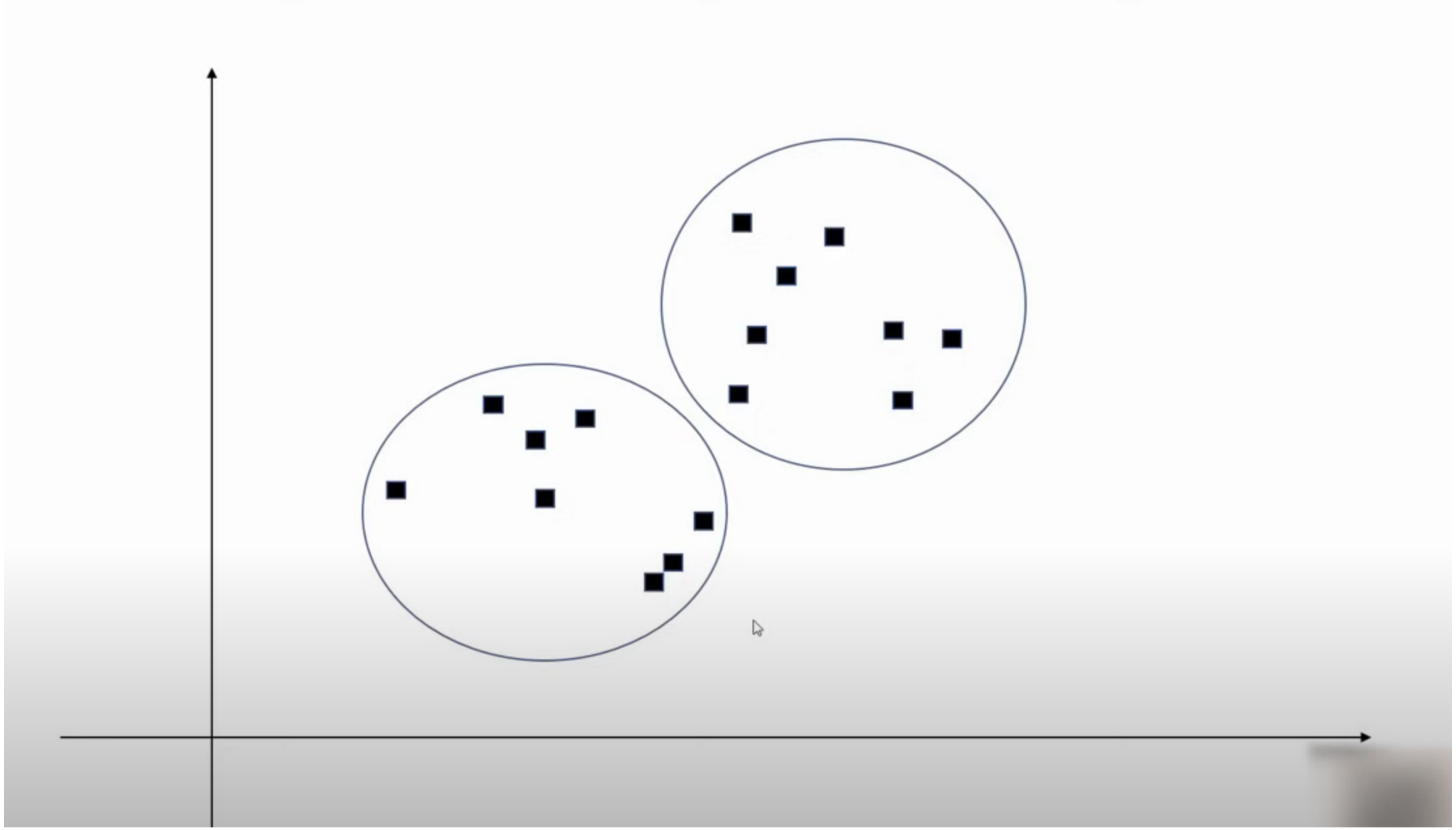
5. Again adjust centroids

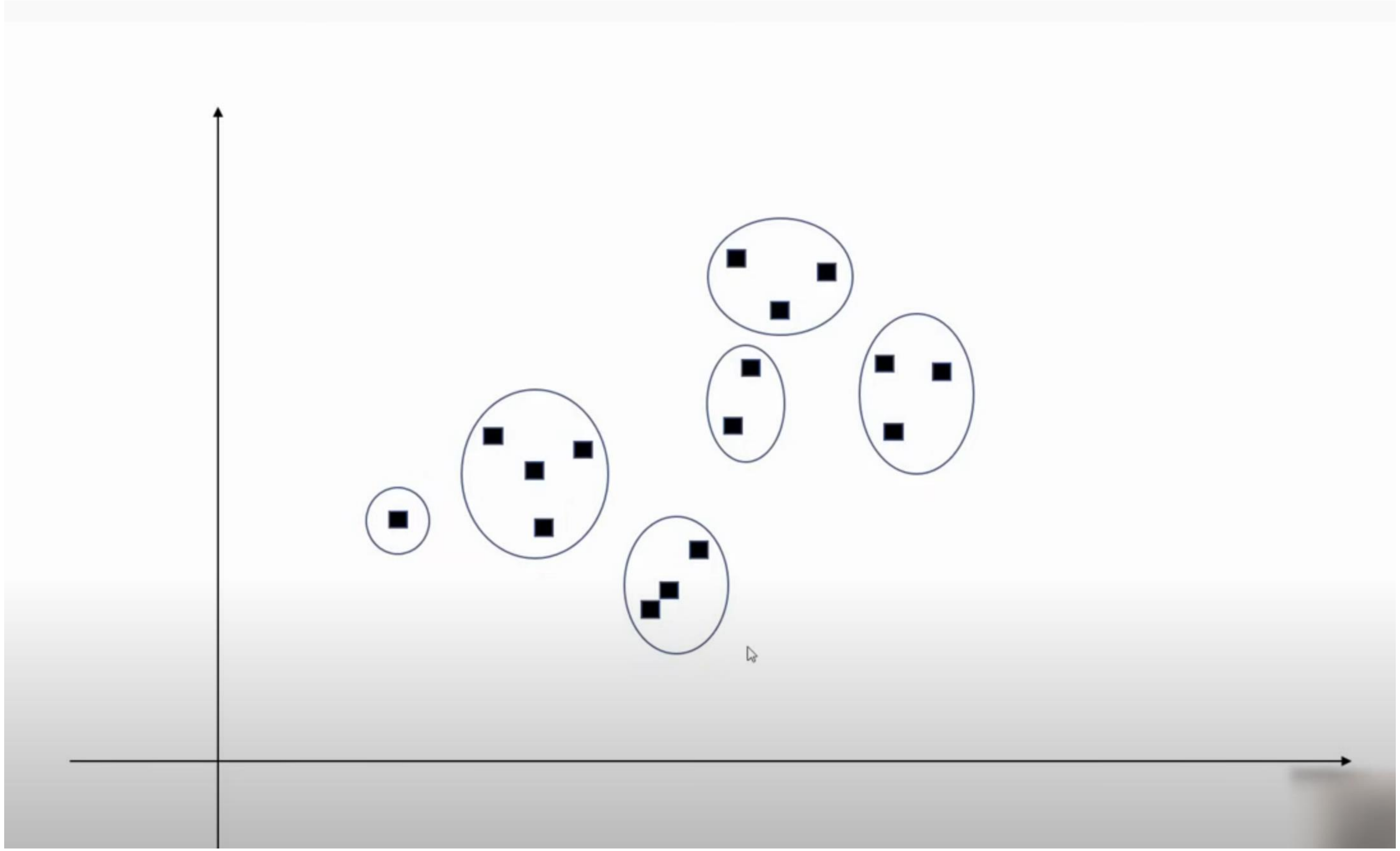


6. Recompute clusters and repeat this till data points stop changing clusters

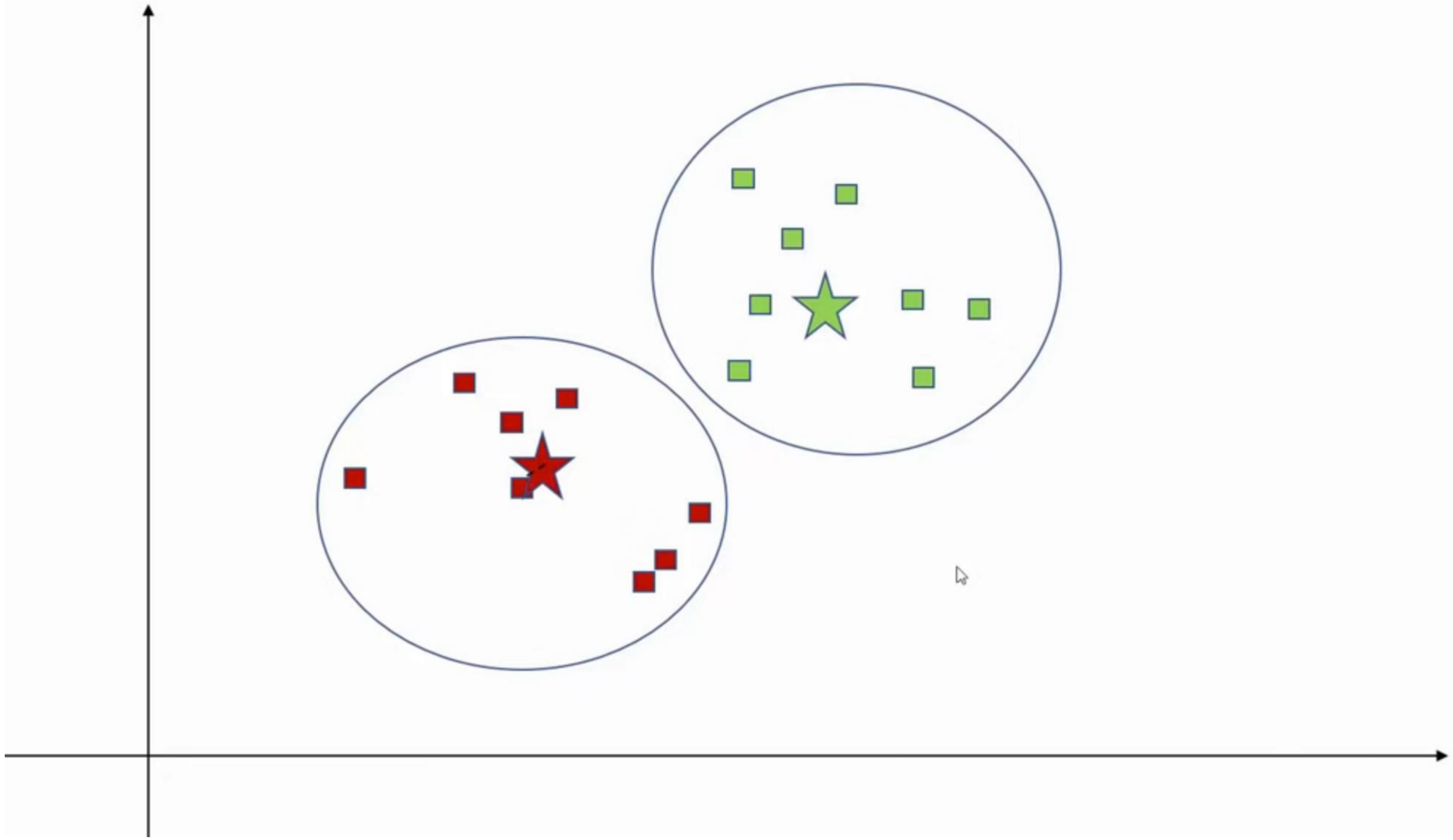


How to determine correct number of clusters (k)?

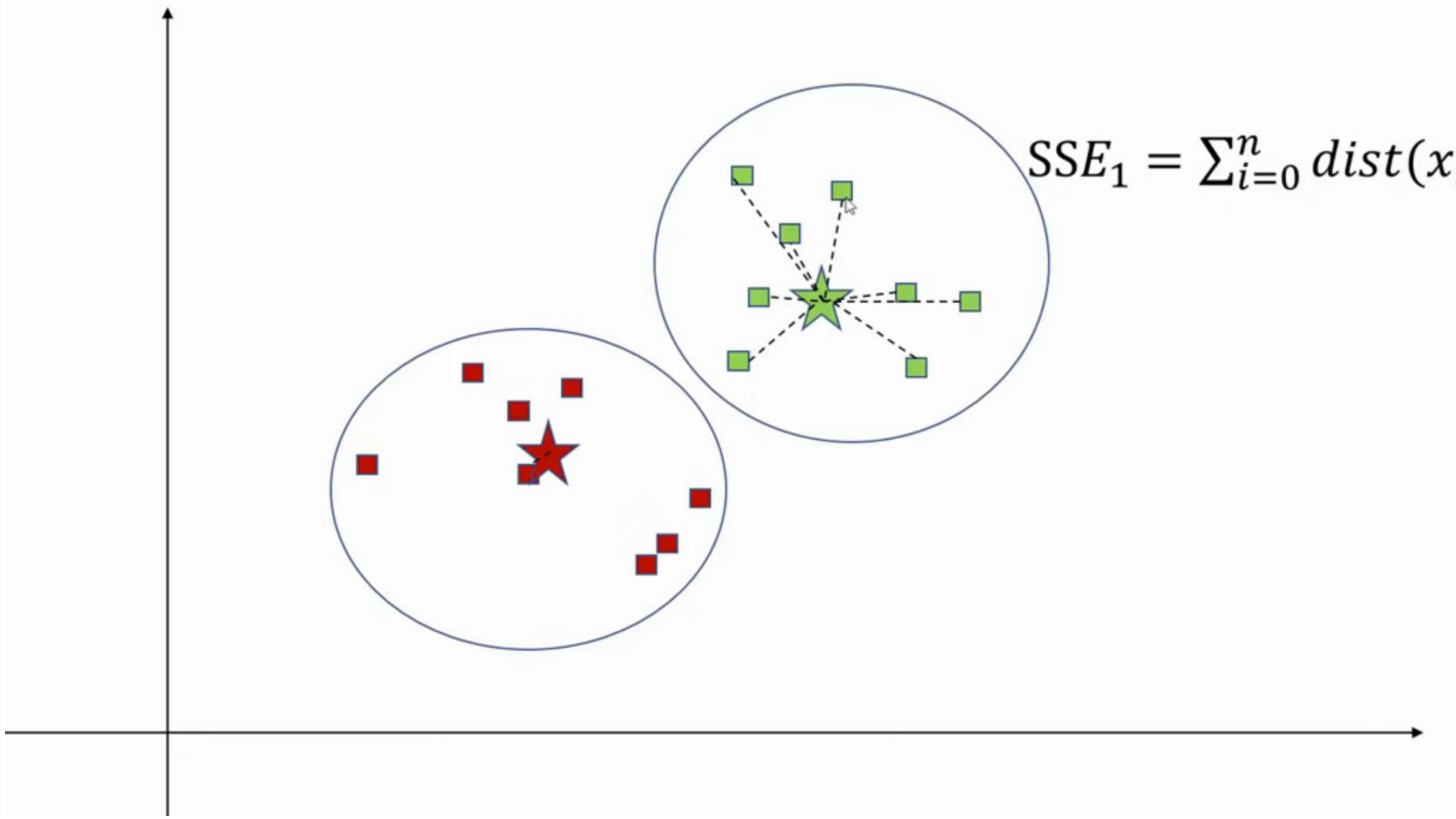




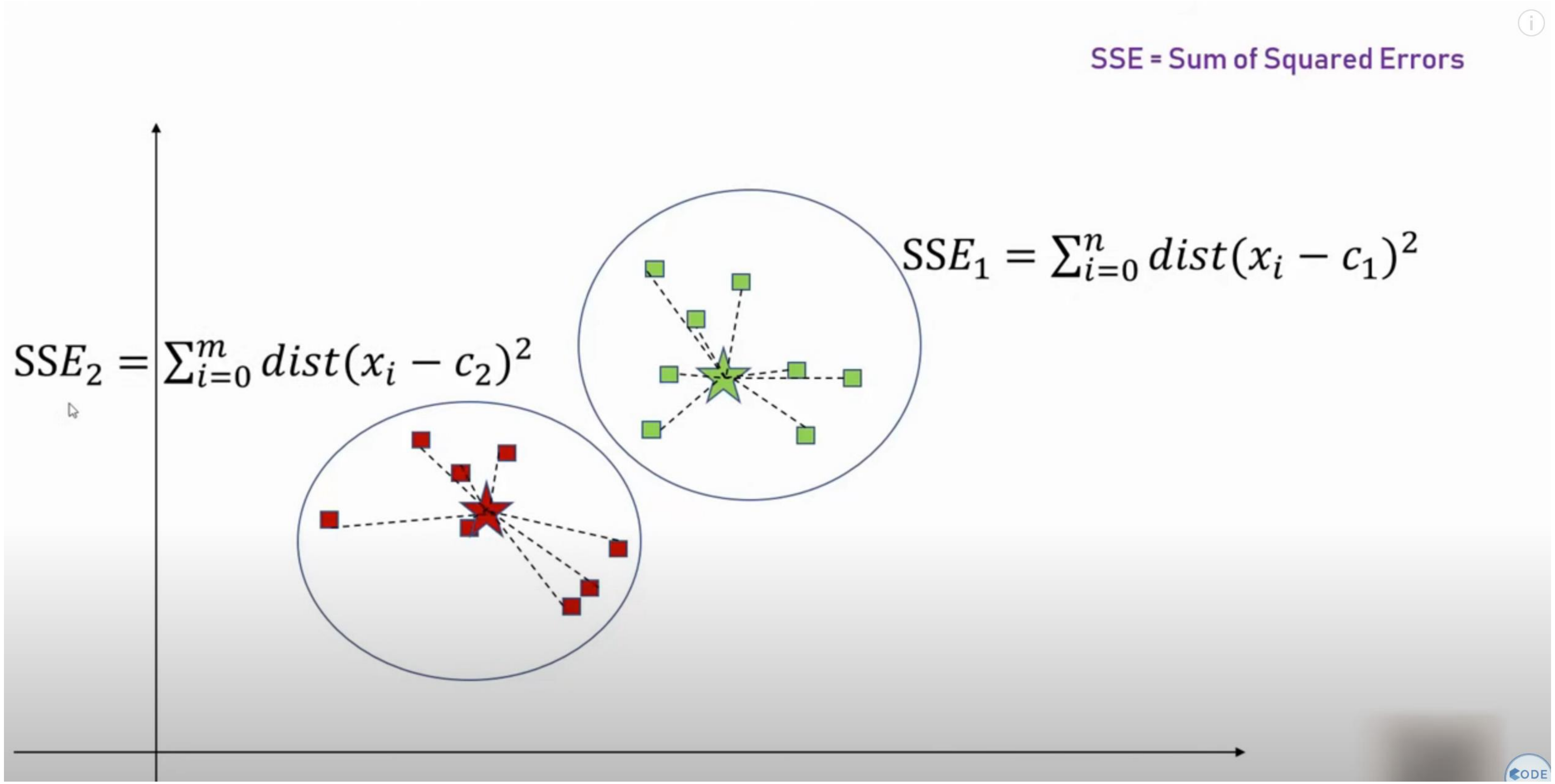
SSE = Sum of Squared Errors



SSE = Sum of Squared Errors



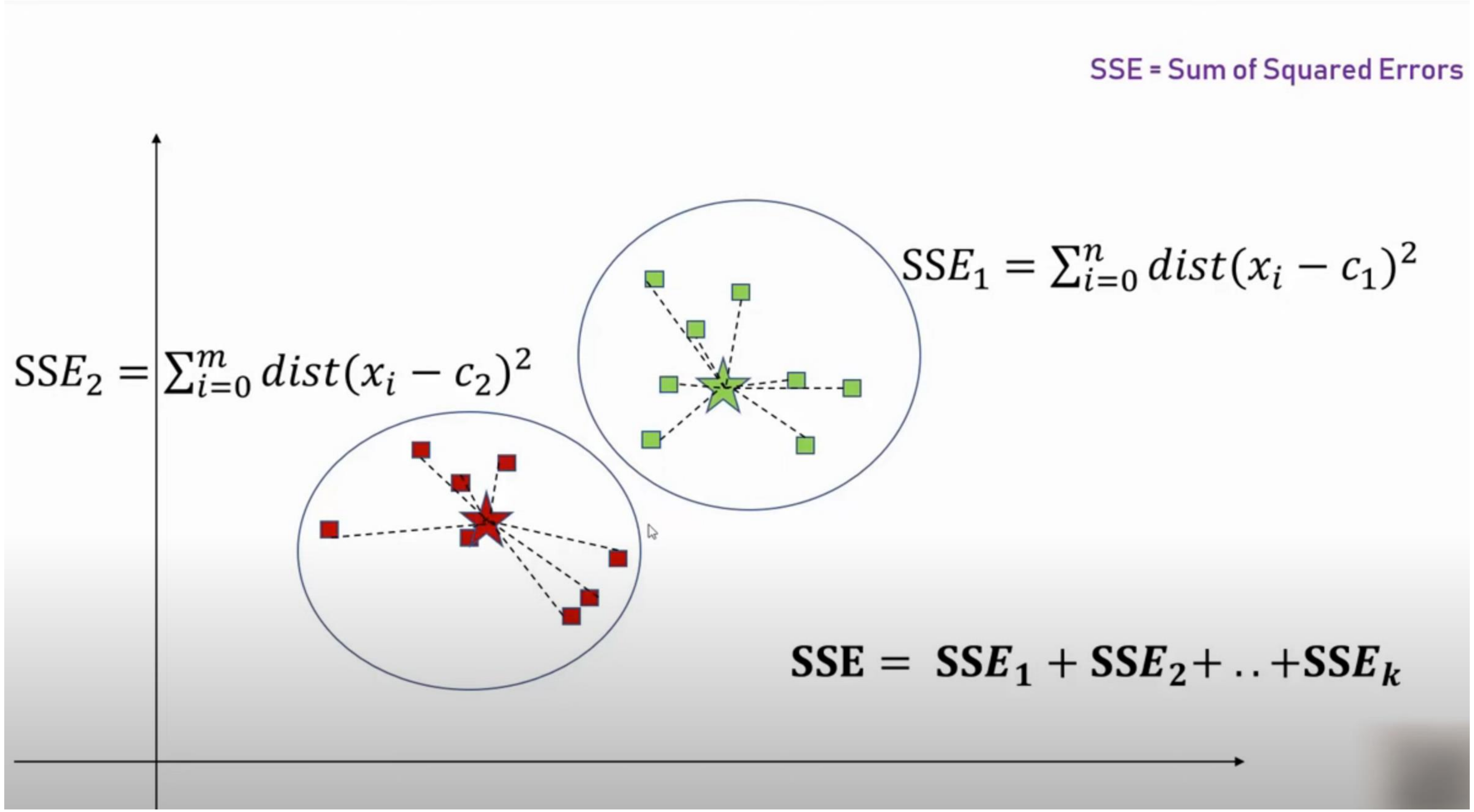
$$SSE_1 = \sum_{i=0}^{n} dist(x_i - c_1)^2$$



SSE = Sum of Squared Errors

$$SSE_1 = \sum_{i=0}^{n} dist(x_i - c_1)^2$$

$$SSE_2 = \sum_{i=0}^{m} dist(x_i - c_2)^2$$

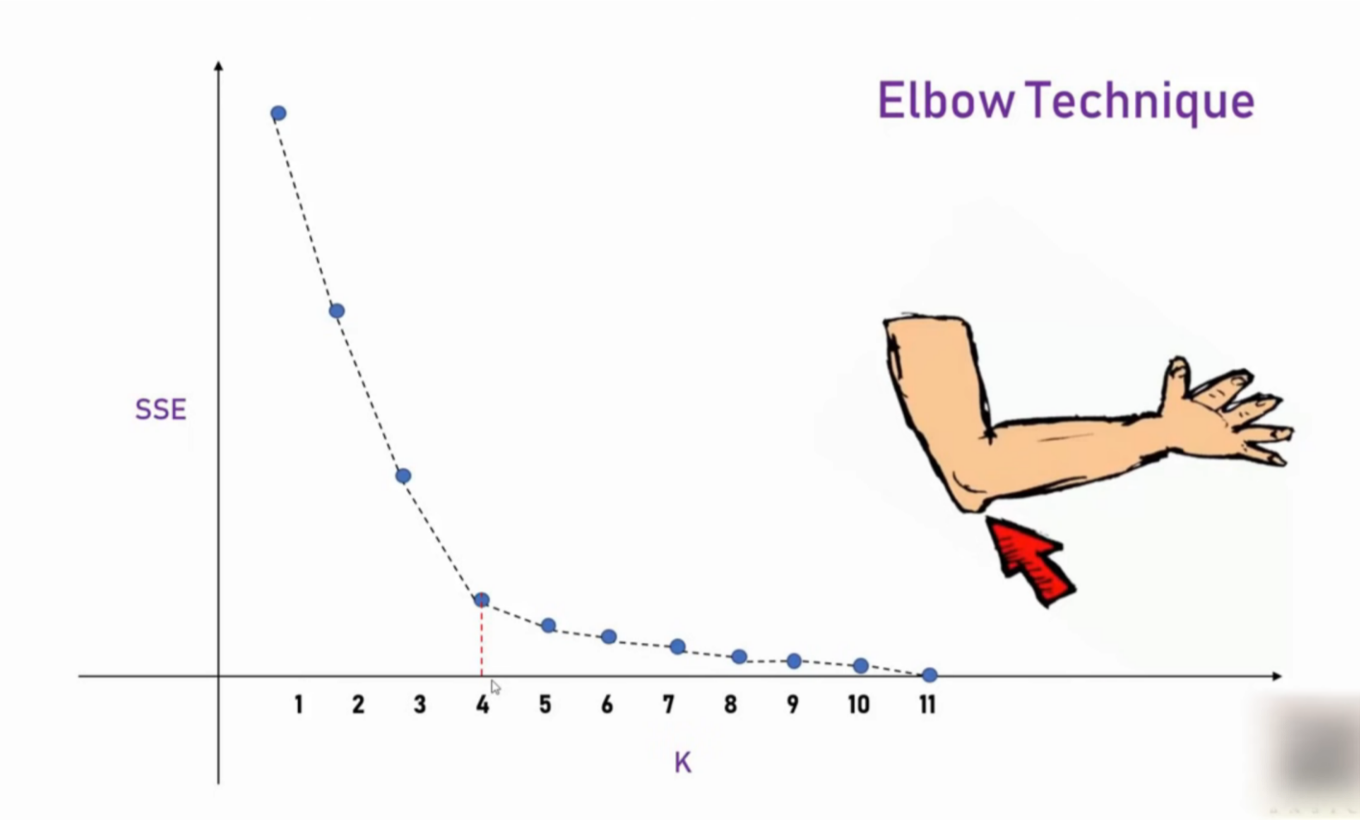


SSE = Sum of Squared Errors

$$SSE_1 = \sum_{i=0}^{n} dist(x_i - c_1)^2$$

$$SSE_2 = \sum_{i=0}^{m} dist(x_i - c_2)^2$$

$$\mathbf{SSE} = \boldsymbol{SSE_1 + SSE_2 + .. + SSE_k}$$



Elbow Technique

Lecture Map

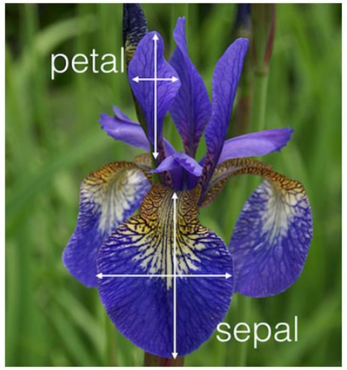
- Theory
- Coding
- Exercise

K-means Clustering

- **Strengths**
  - Simple iterative method
  - User provides “K”
- **Weaknesses**
  - Often too simple → bad results
  - Difficult to guess the correct “K”

K-means Clustering Exercise

Exercise



1. Use iris flower dataset from sklearn library and try to form clusters of flowers using petal width and length features. Drop other two features for simplicity.
2. Figure out if any preprocessing such as scaling would help here
3. Draw elbow plot and from that figure out optimal value of k

References

- http://brokerstir.com/logistic-regression-model-intuition/
- https://www.geeksforgeeks.org/implement-sigmoid-function-using-numpy/
- http://ieeexplore.ieee.org/document/6914146/
- http://www.svms.org/disadvantages.html
- https://www.mit.edu/~9.520/spring09/Classes/multiclass.pdf