Oriental Institute of Science and Technology, Bhopal

Department of Computer Science and Engineering

Progress Seminar - I

on

FULL STACK SCRUTINY

(Minor Project II)

Session 2022

Guided By : Prof. Goldi Jarbais

Presented by:

1. Anshul Verma (0105CS191023) [Team Leader]
2. Ayush Waghmare (0105CS191028)
3. Harshit Shrivastava (0105CS191048)
4. Manish Nathrani (0105CS191062)

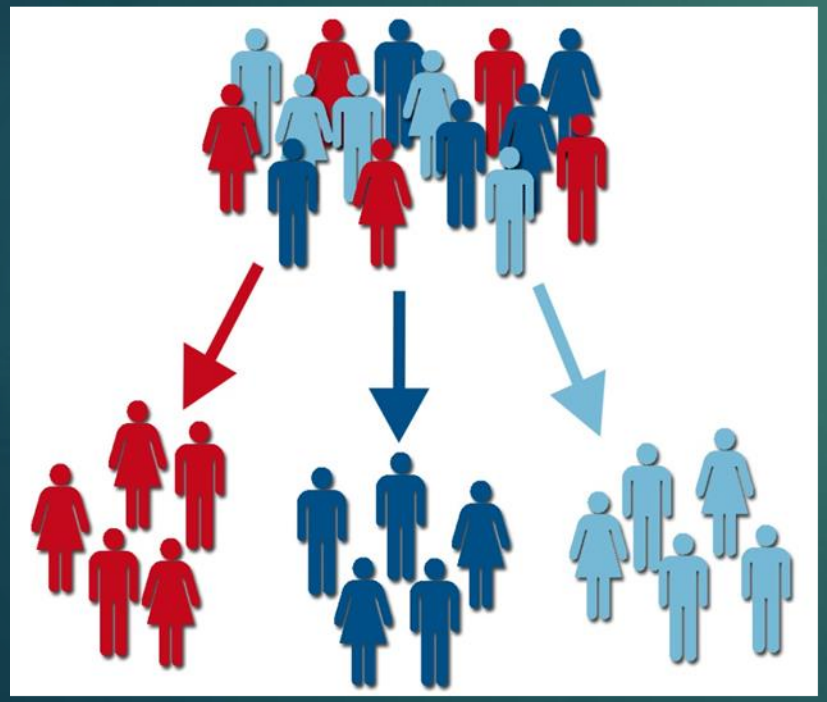
OVERVIEW

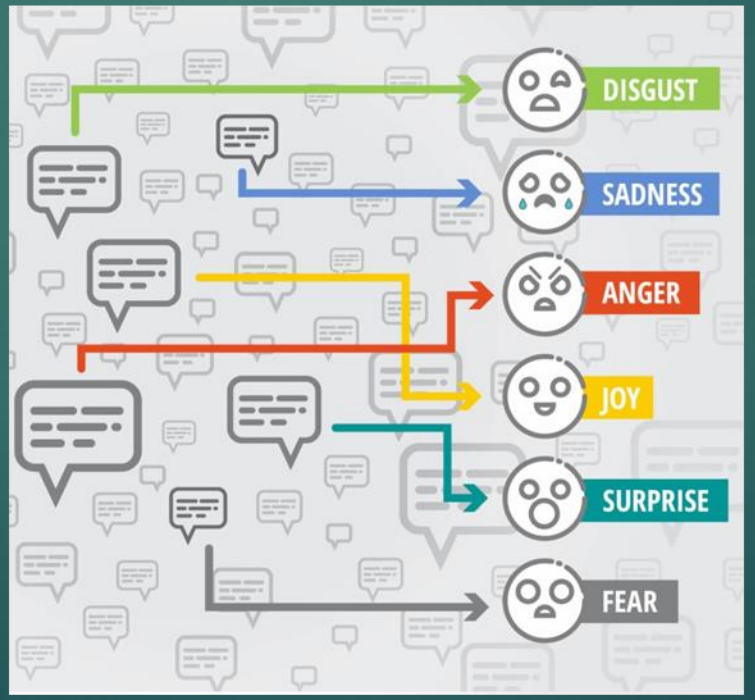
- Introduction and Detail Idea
- Design Process
- Flowchart
- UML Diagrams
- Description of Modules
- Module 1
- References

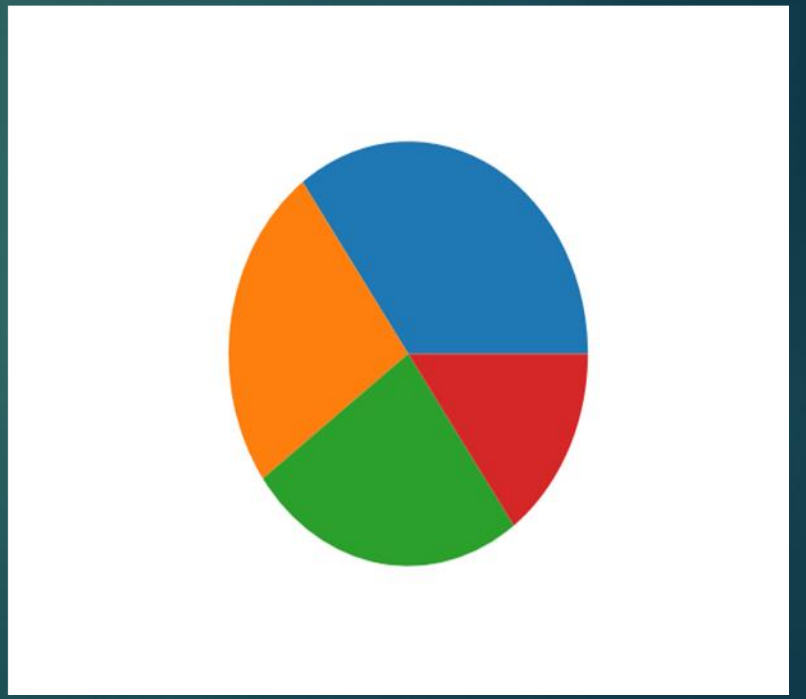
INTRODUCTION & DETAIL IDEA OF PROJECT

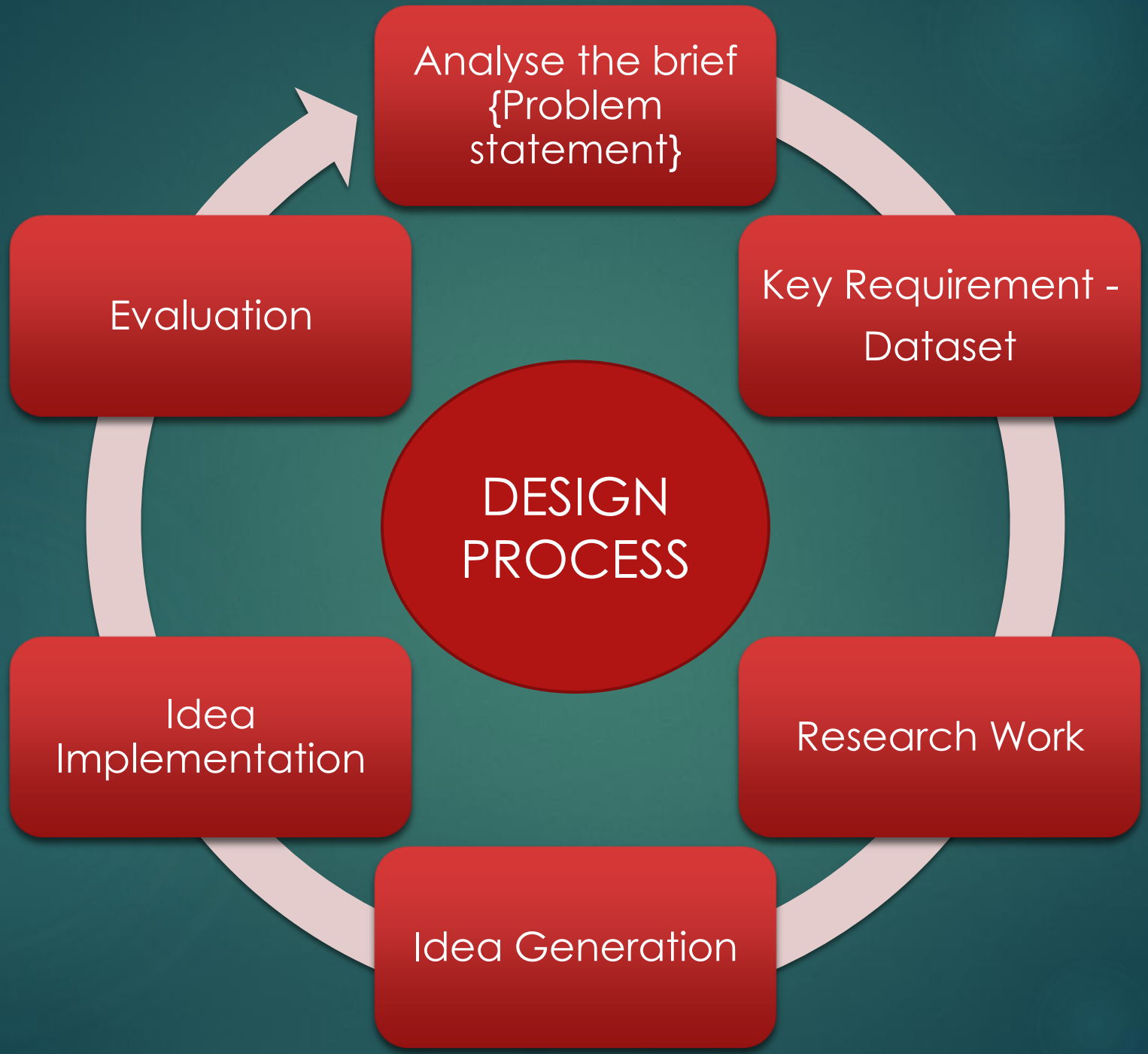
All the three things at the same platform.

Customer Segmentation + Sentiment Analysis + Exploratory Data Analysis

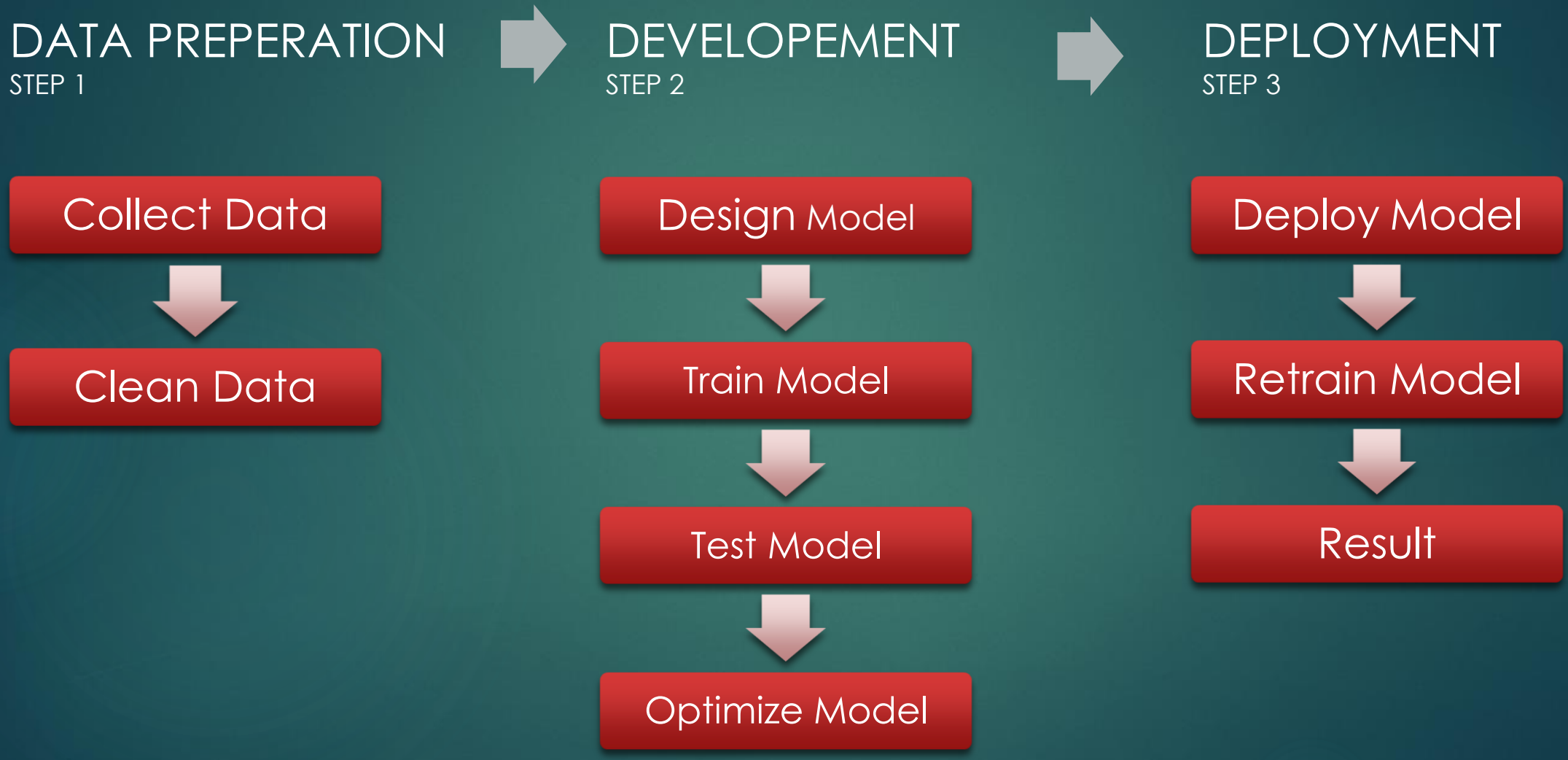







Analyse the brief {Problem statement}
Key Requirement - Dataset
Research Work
Idea Generation
Idea Implementation
Evaluation
DESIGN PROCESS

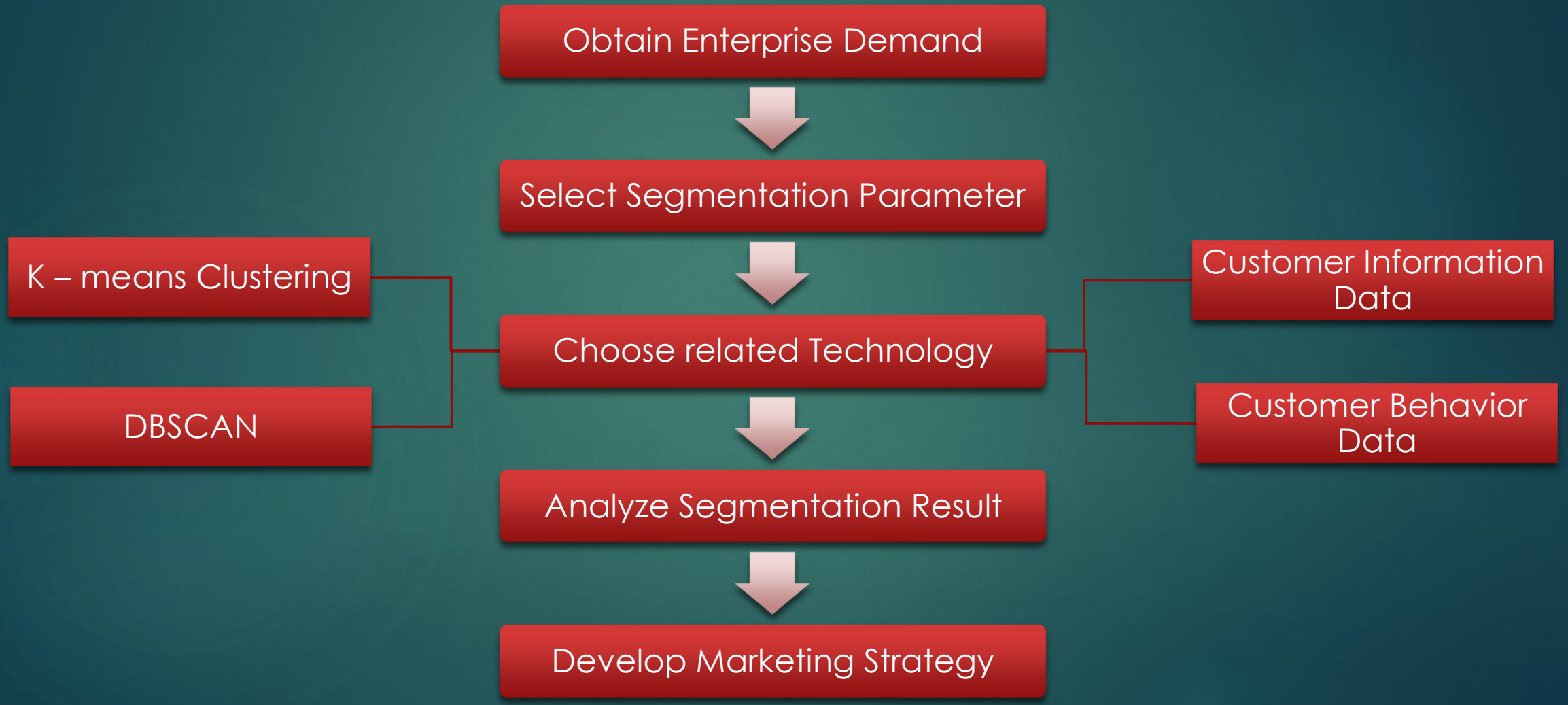
FLOWCHART


DATA PREPERATION
STEP 1
DEVELOPEMENT
STEP 2
DEPLOYMENT
STEP 3
Collect Data
Clean Data
Design Model
Train Model
Test Model
Optimize Model
Deploy Model
Retrain Model
Result

UML DIAGRAMS

- Customer Segmentation
- Sentiment Analysis

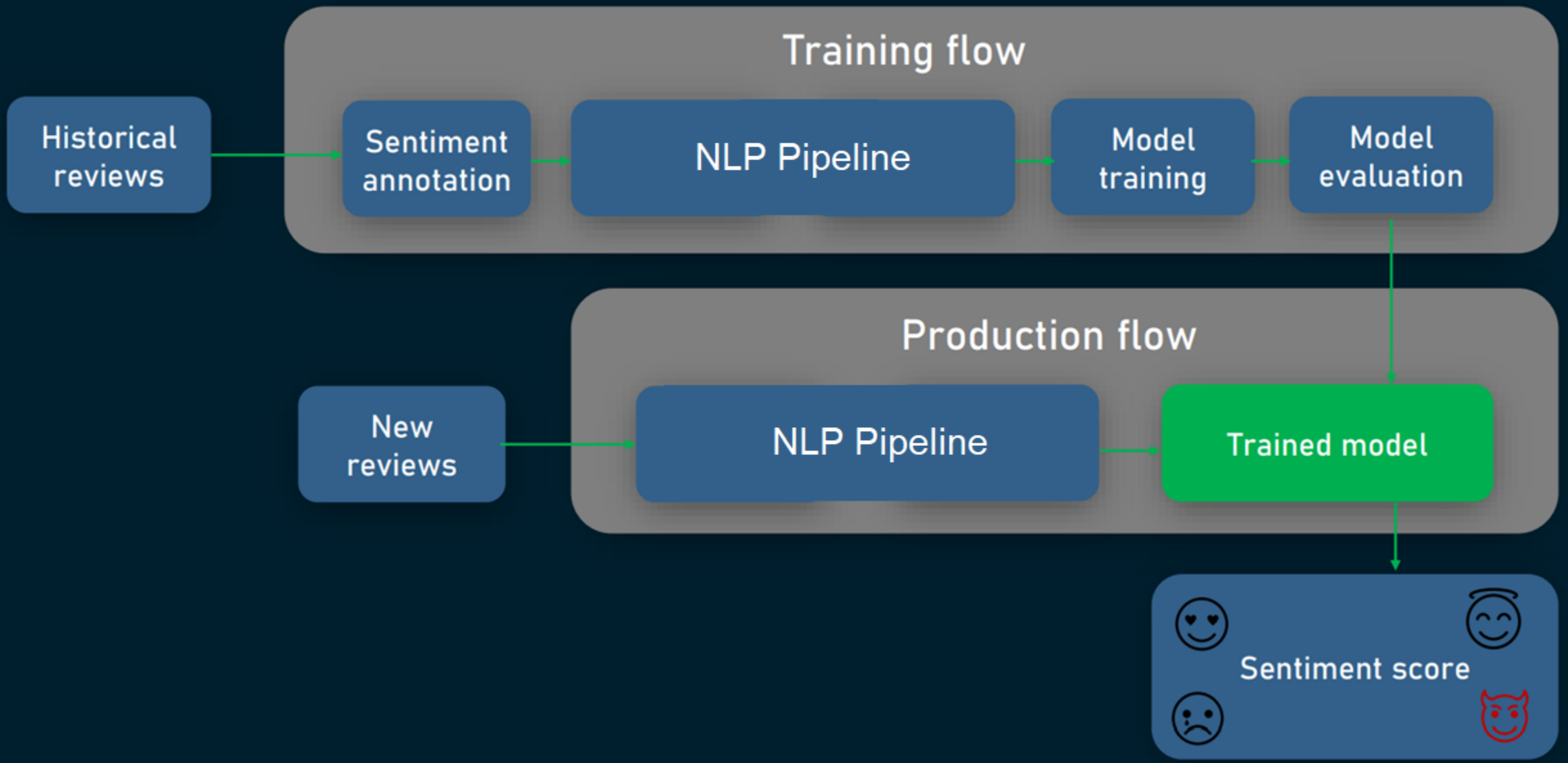
CUSTOMER SEGMENTATION



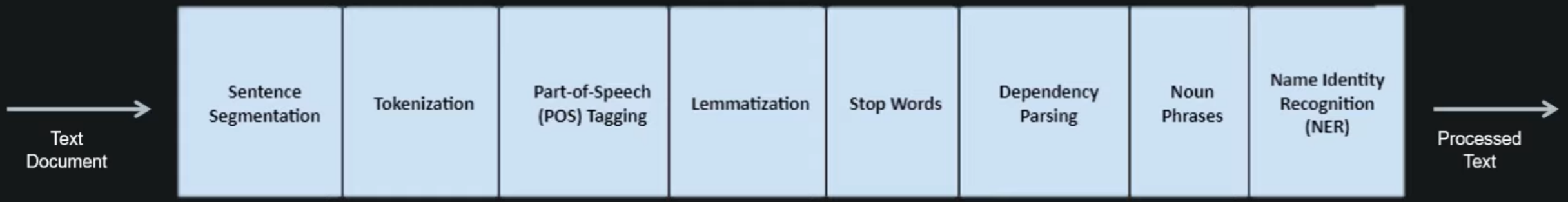
SENTIMENT ANALYSIS

A typical Sentiment Analysis Model

SENTIMENT ANALYSIS WITH MACHINE LEARNING



NLP PIPELINE

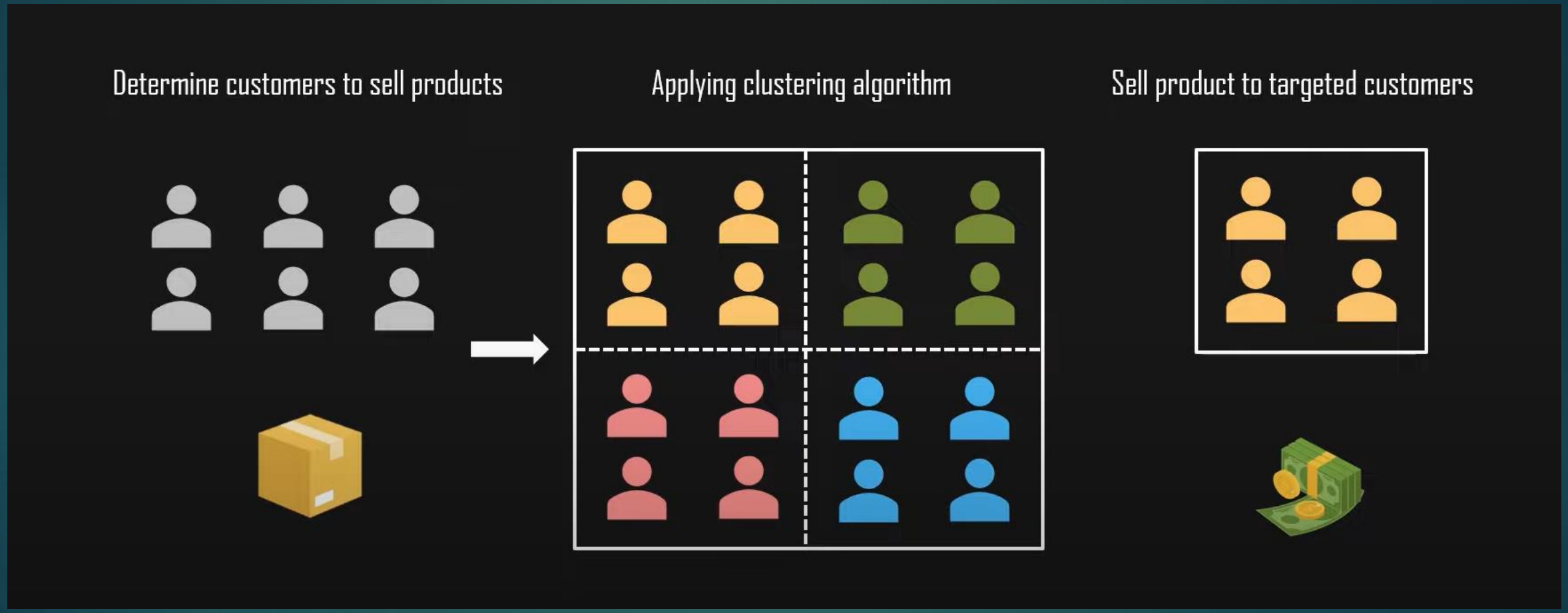


A pipeline is just a way to design a program where the output of one module feeds to the input of the next. For example, Linux shells feature a pipeline where the output of a command can be fed to the next using the pipe character

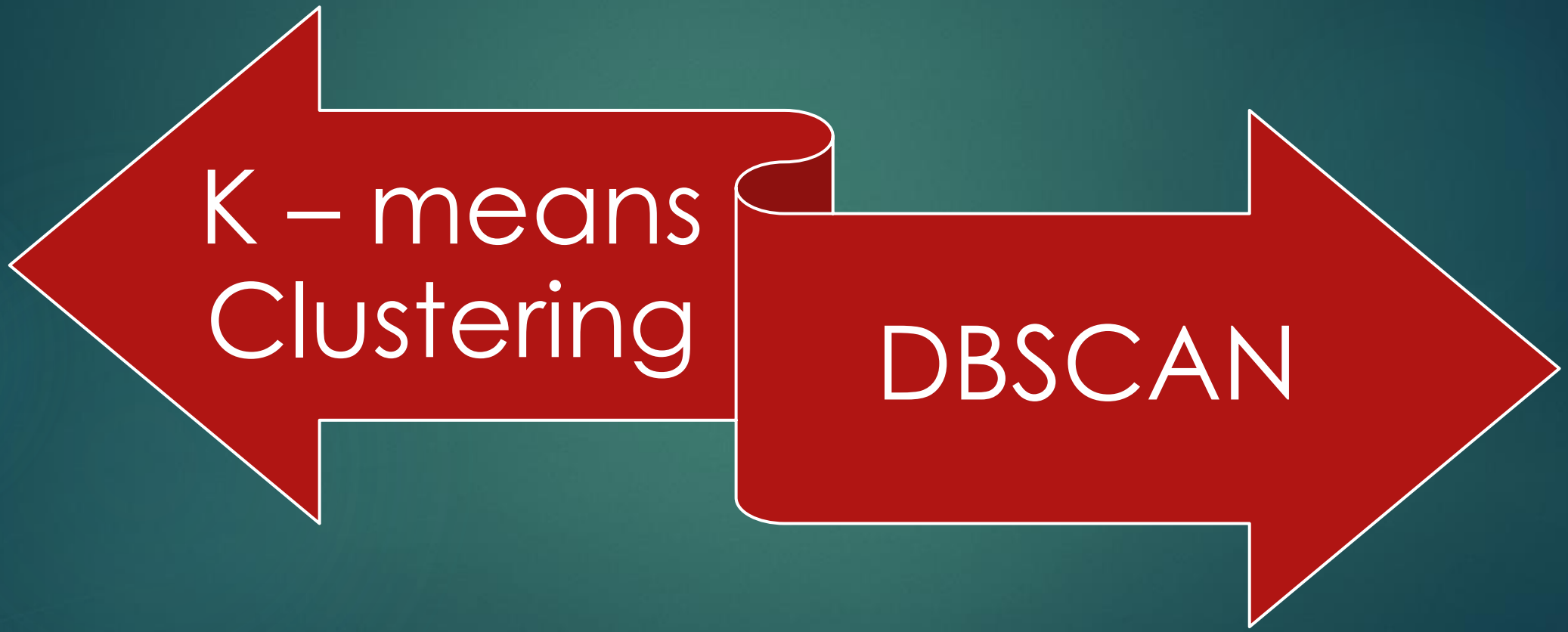
DESCRIPTION OF MODULES

- M 1 — Customer Segmentation
- M 2 — Sentiment Analysis
- M 3 — Exploratory Data Analysis
- M 4 — Data Visual Representation & Hypothesis Generation

MODULE 1 – CUSTOMER SEGMENTATION

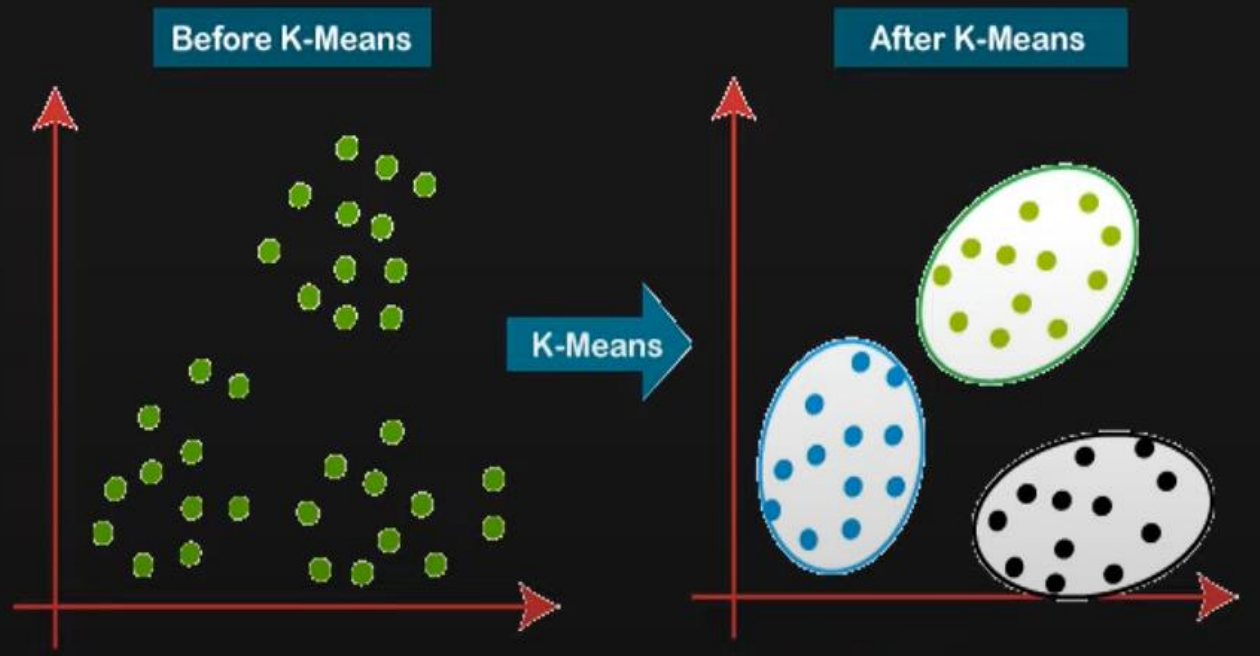


CLUSTERING OPTIONS



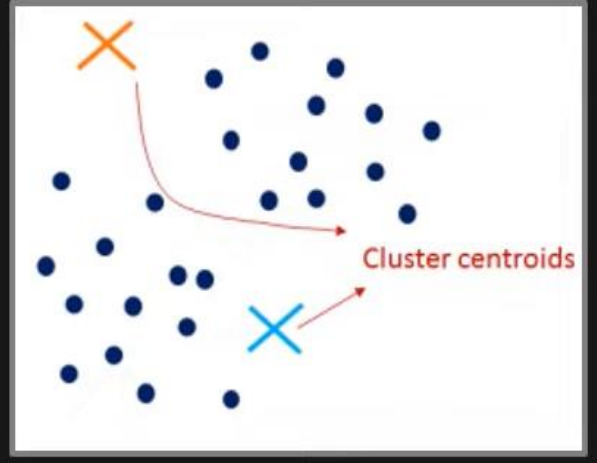
K-means Algorithm

It is an iterative algorithm that divides the unlabelled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

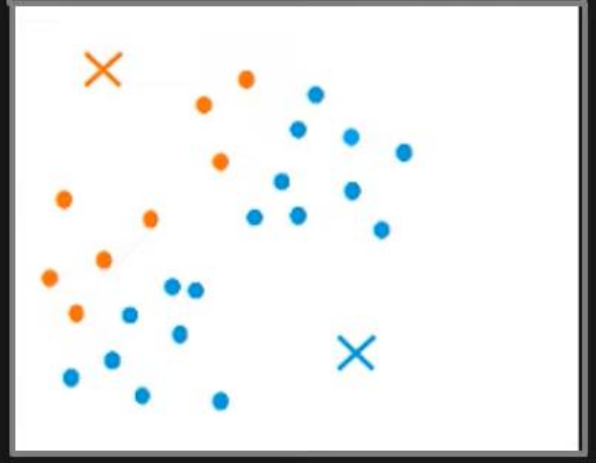


K-means Algorithm

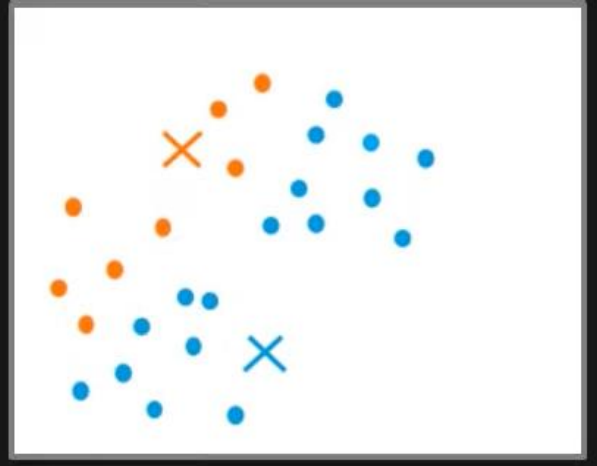
Initialization



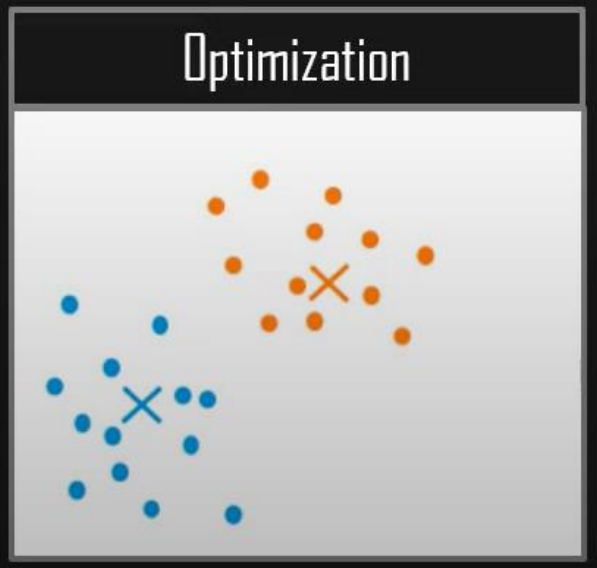
Cluster Assignment



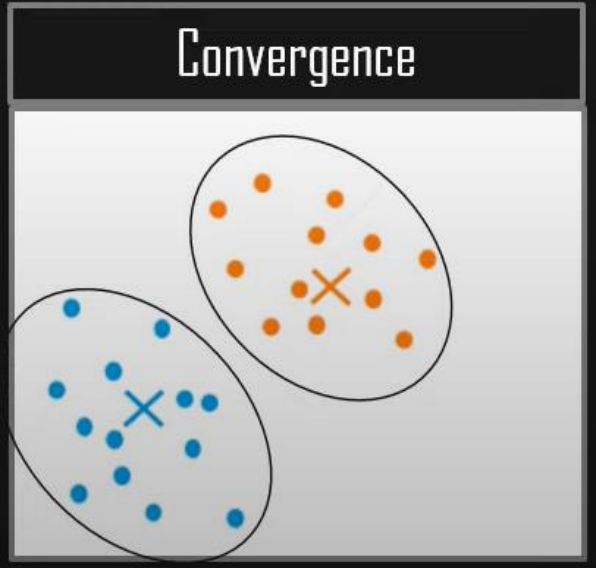
Move Centroid



Optimization



Convergence

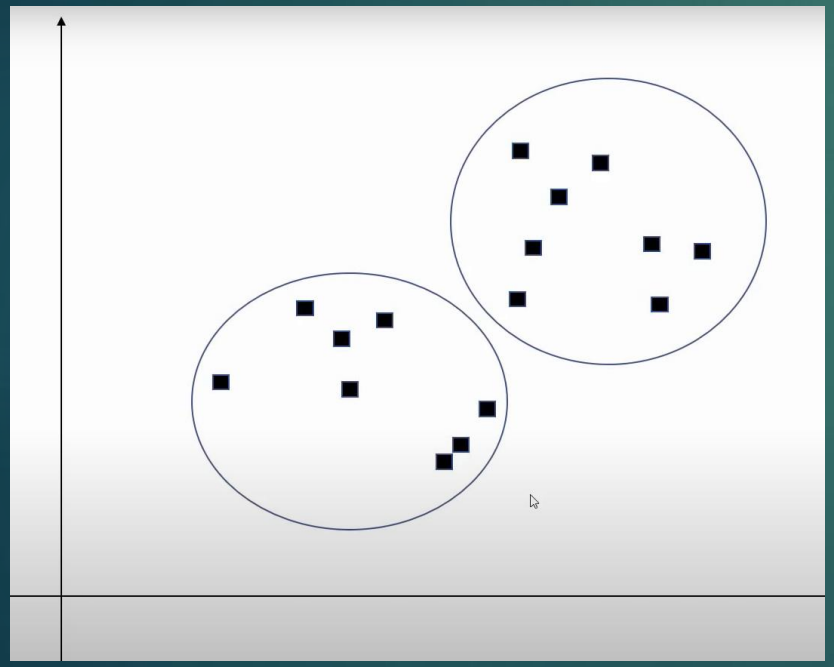


But there is a Catch…

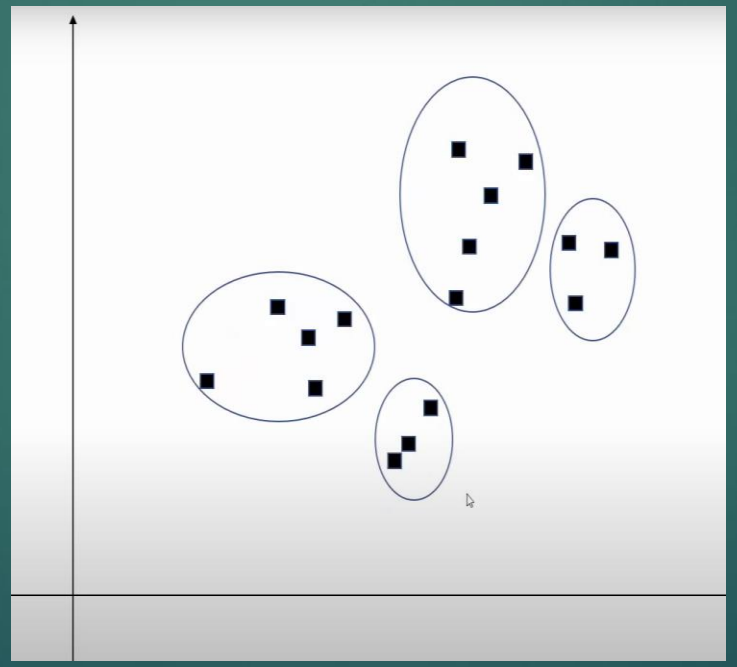
How to determine the Number of Clusters (K)?

The Problem is…

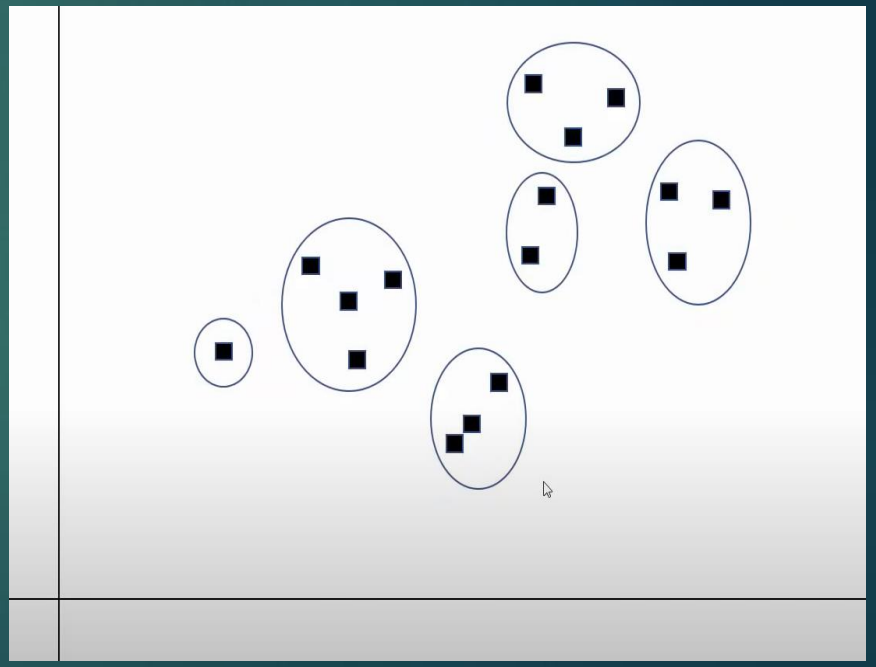
K = 2



K = 4



K = 6

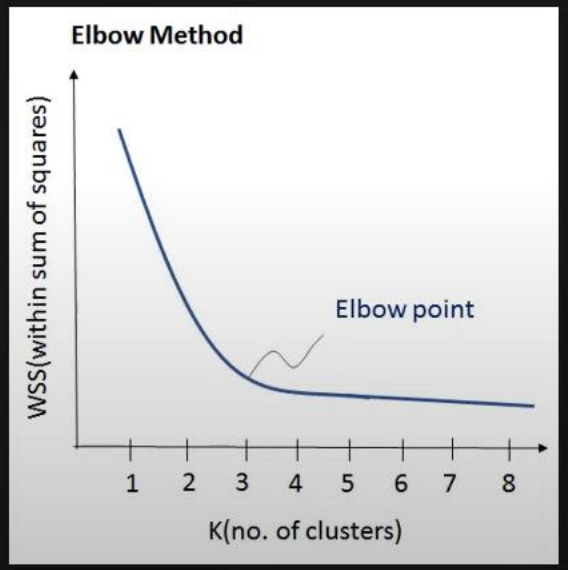


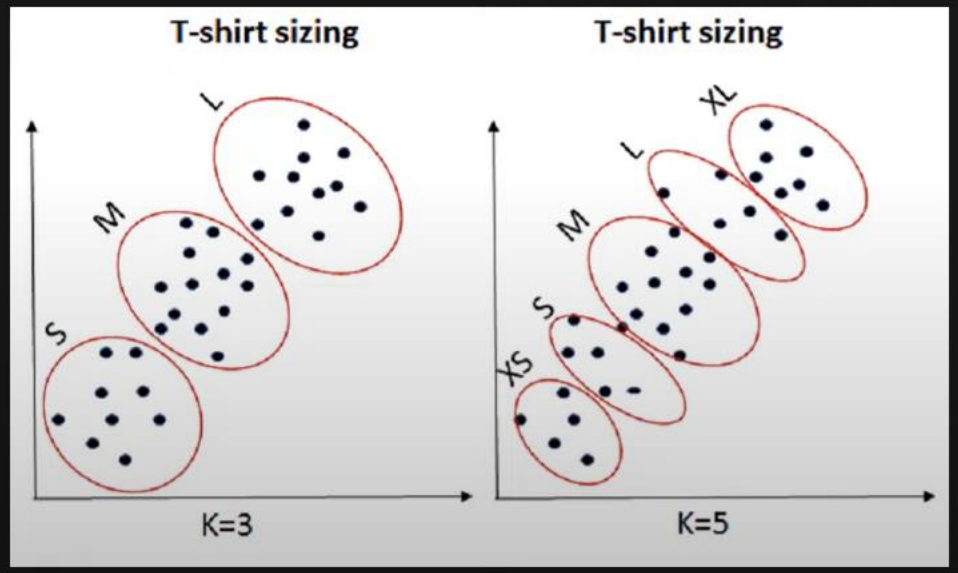
Solution is…

Ways to choose the optimum K value

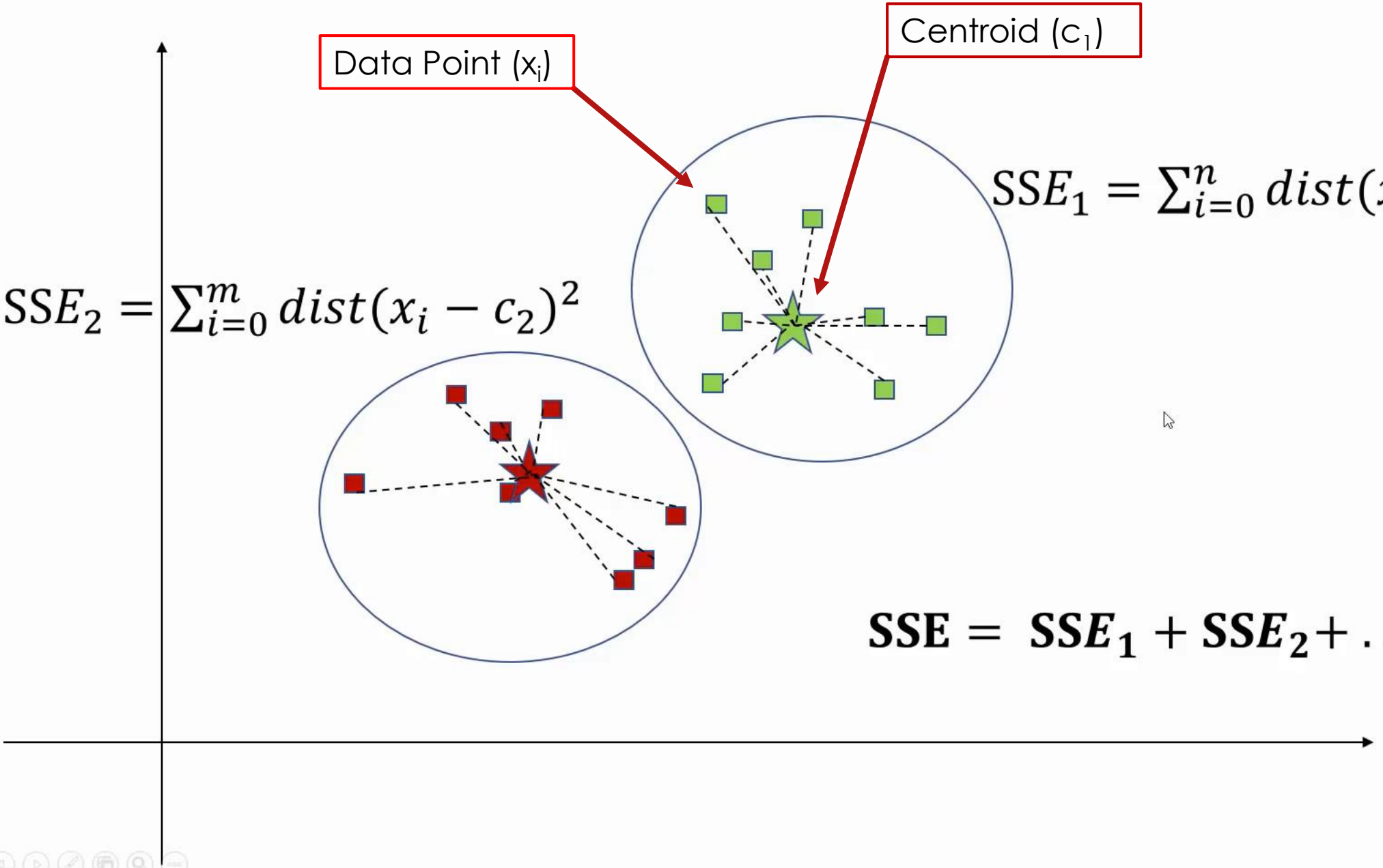
Elbow Method

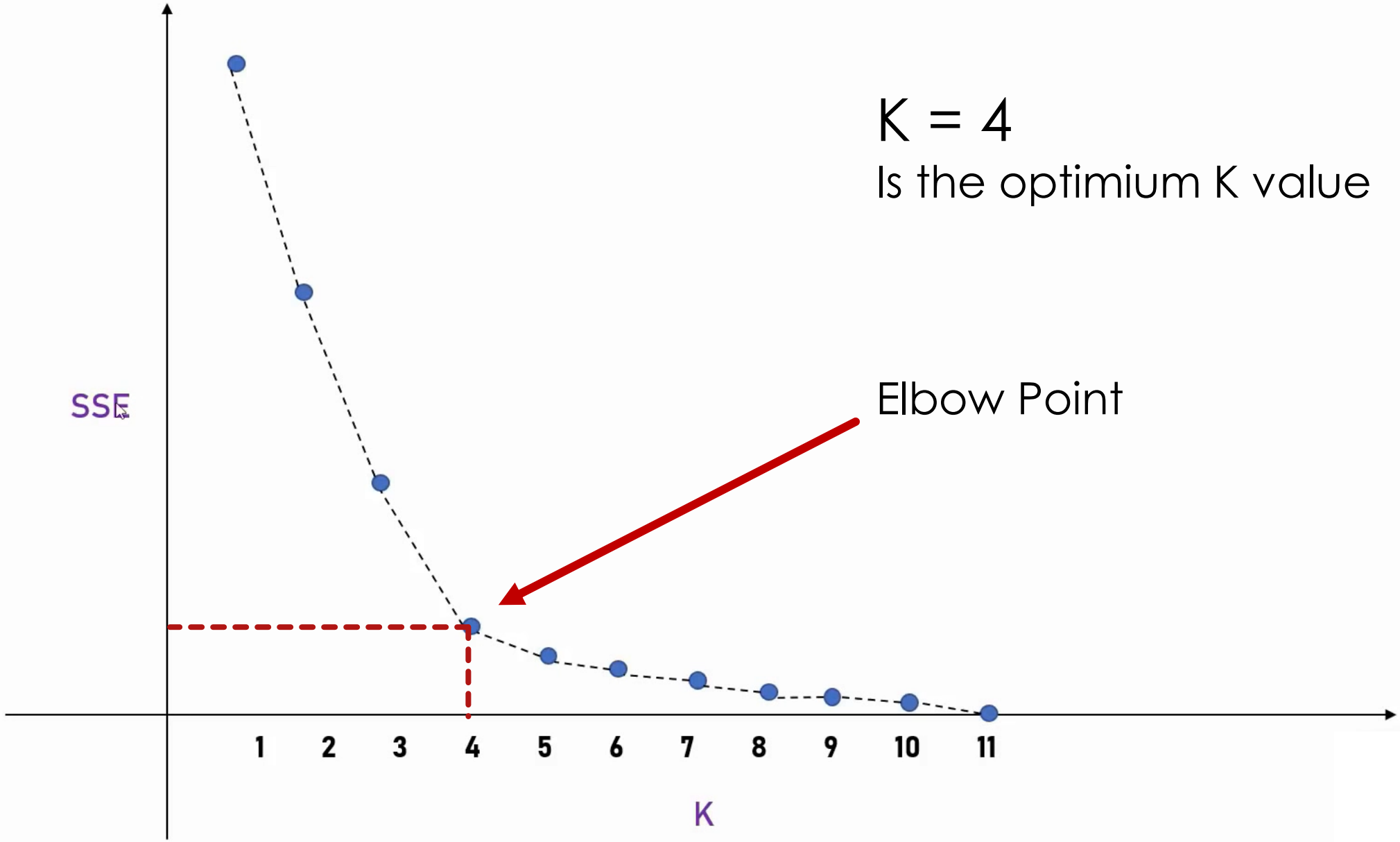
Purpose Based





SSE = Sum of Squared Errors



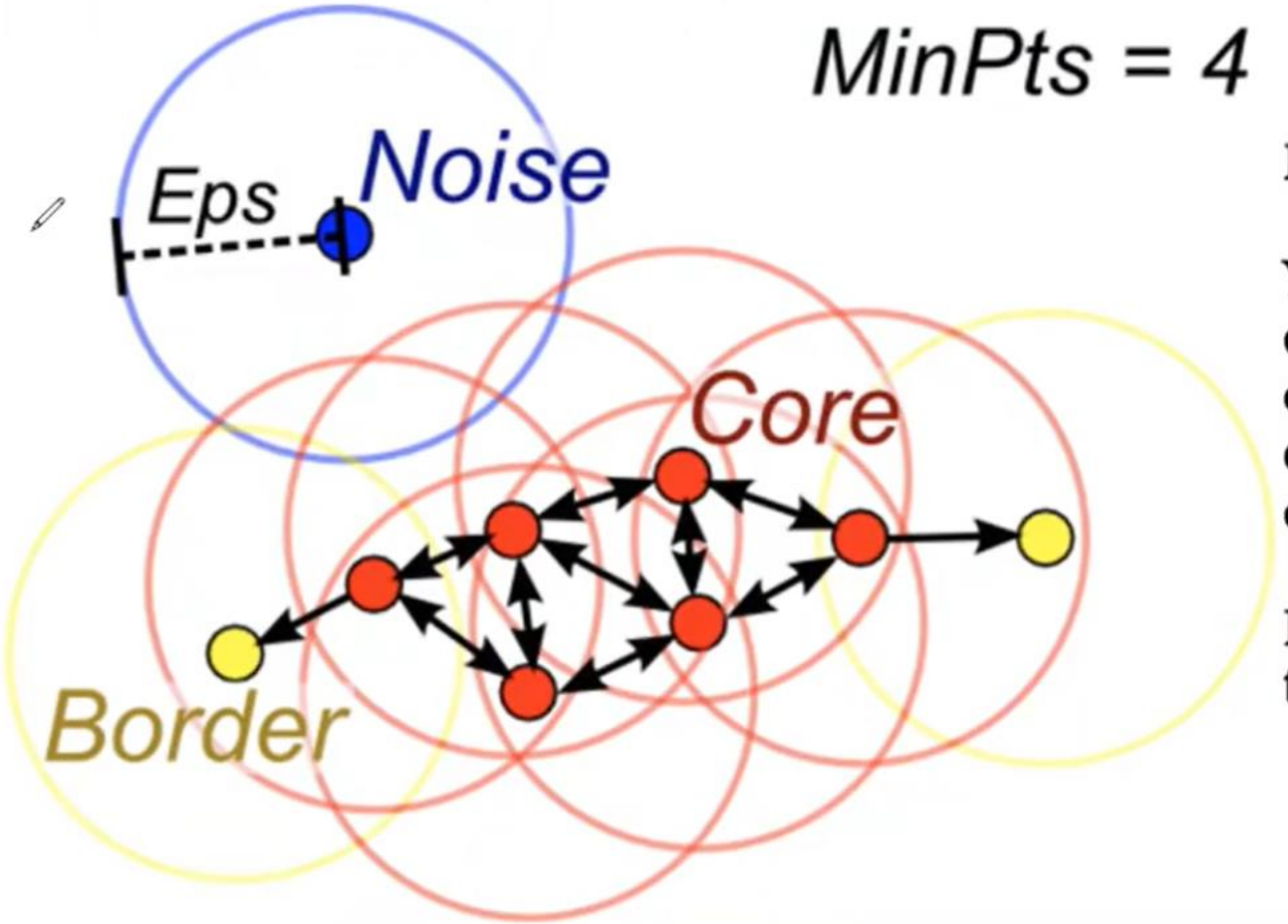


K = 4

Is the optimium K value

Elbow Point

Density-Based Spatial Clustering of Applications with Noise(DBSCAN)

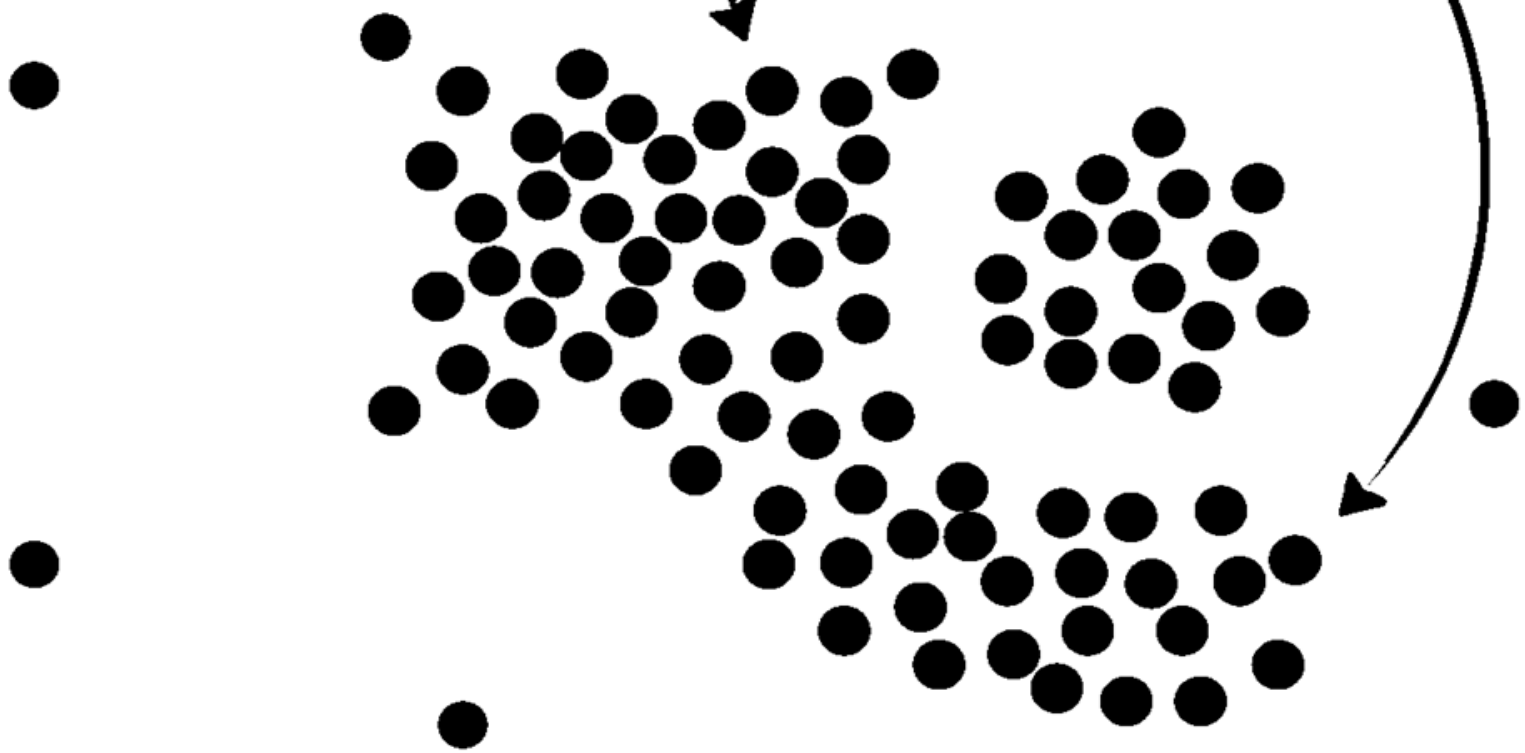


Red: Core Points

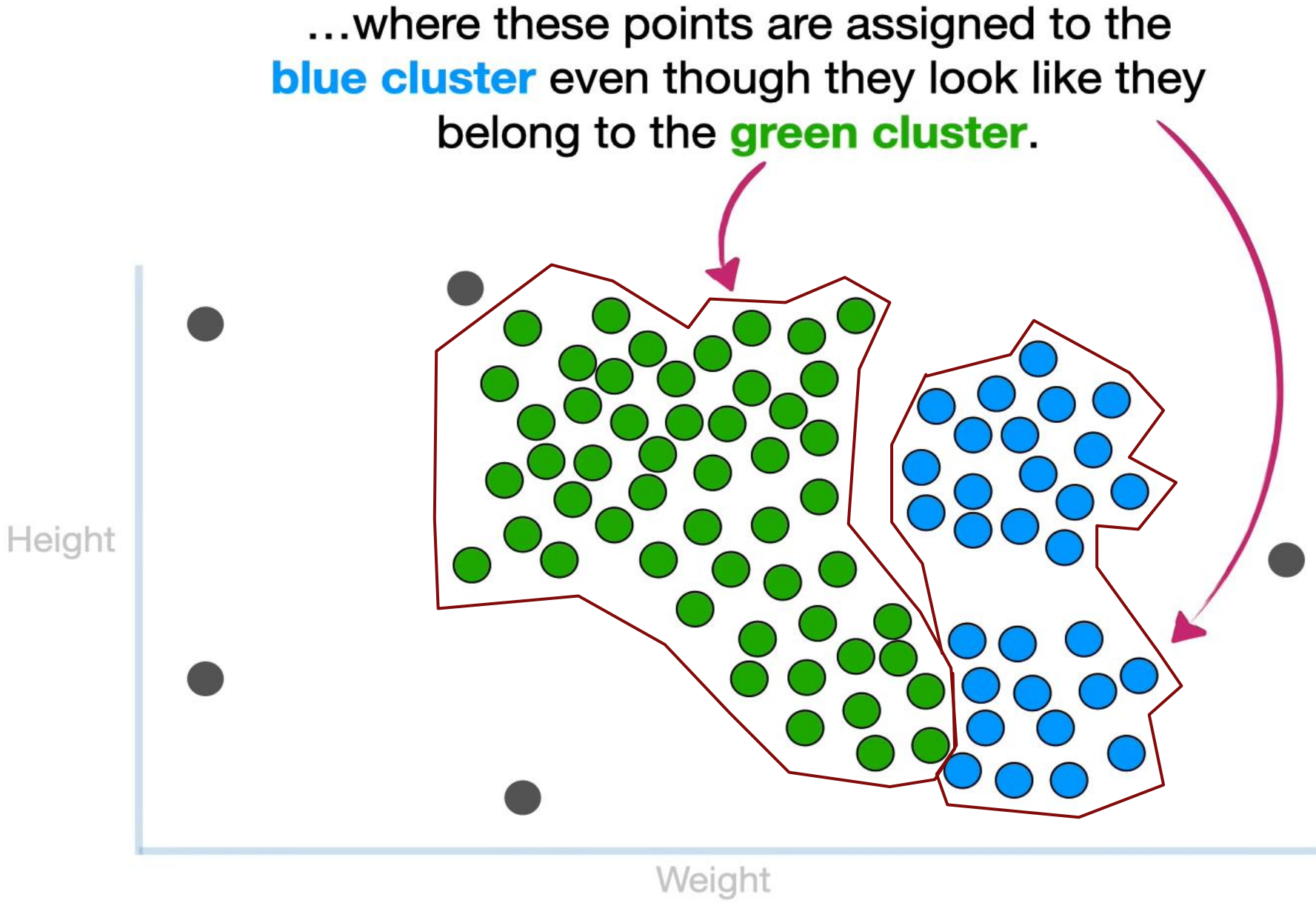
Yellow: Border points. Still part of the cluster because it's within epsilon of a core point, but not does not meet the min_points criteria

Blue: Noise point. Not assigned to a cluster

…where these points are assigned to the **blue cluster** even though they look like they belong to the **green cluster.**

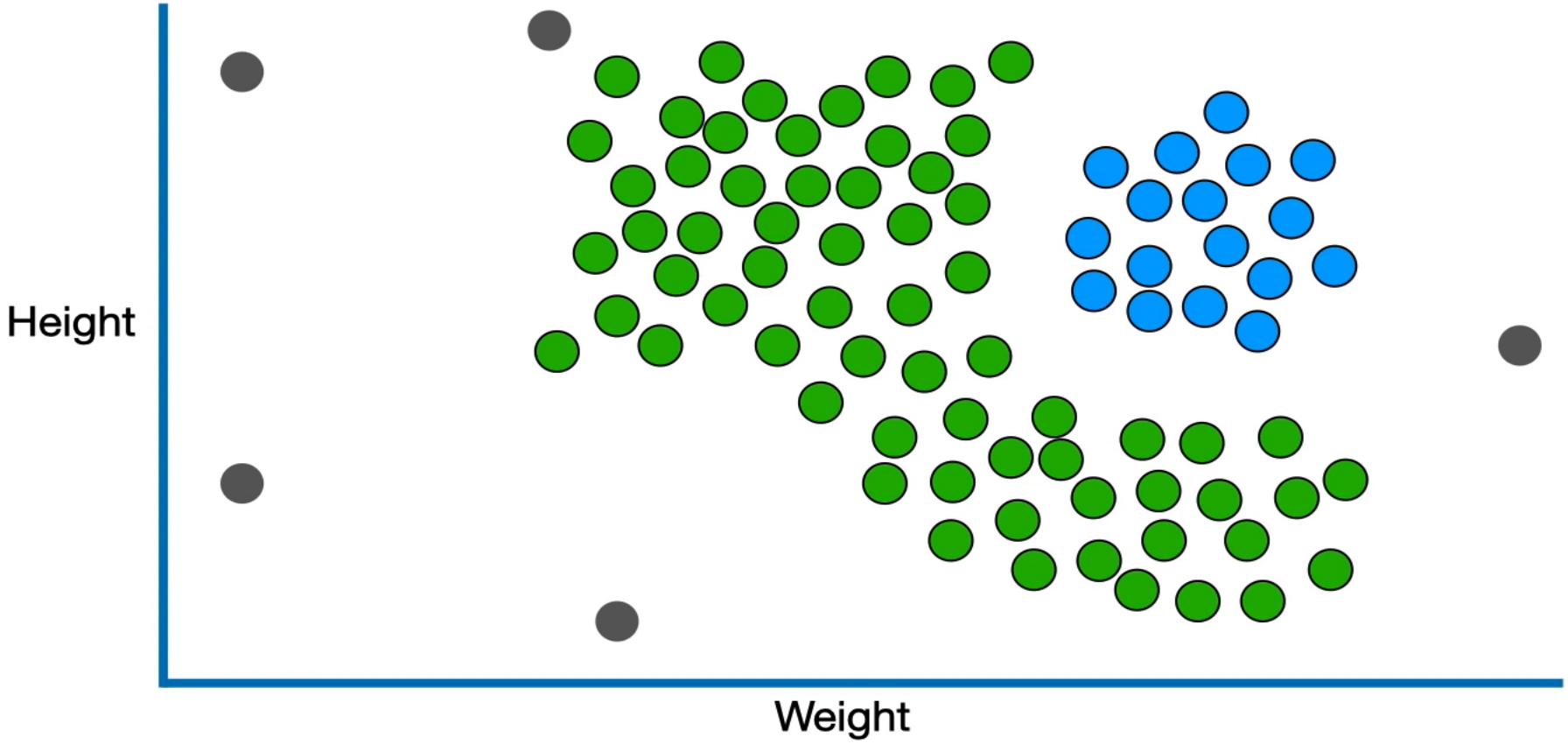


Clustering using K – means Clustering



Clustering using DBSCAN Clustering

So let's go back to the original **2**-dimensional graph and see how **DBSCAN** tries to mimic what we can easily do by eye.



Conclusion

- It has a capacity of analyzing the needs of the Customer.
- It has a target of reaching the Products & services for the particular group of Customers.
- For Customer Segmentation we are using DBSCAN over K-Means Algorithm as DBSCAN is more efficient & K-means is a Show stopper.

References

- https://technologyadvice.com/blog/marketing/customer-segmentation-methods/
- https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17#:~:text=Sentiment%20analysis%20is%20contextual%20mining,service%20while%20monitoring%20online%20conversations.
- https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15
- https://thesai.org/Downloads/Volume10No2/Paper_48-A_Study_on_Sentiment_Analysis_Techniques.pdf
- https://www.researchgate.net/publication/313737530_Review_on_Customer_Segmentation_Technique_on_Ecommerce#:~:text=This%20paper%20will%20review%20customer%20segmentation%20using%20data%2C,and%20survey%20data%20were%20as%20the%20external%20data.

Thank You!

Technologies

- Python Libraries –
  - matplotlib.
  - seaborn.
  - plotly
  - sklearn.
  - Pandas.
  - Numpy

- Machine Learning Algorithms –
  - Customer Segmentation :
    - DBSCAN (Density Based Spatial Clustering of Application with Noise)
    - K-means Clustering
  - Sentiment Analysis :
    - Natural Language Processing (NLP)
    - Using different classifiers
      - SVM (Support Vector Machine)
      - Random Forest Classifier
      - Multinomial Naive Bayes
      - Naive Bayes
      - Combination of SVM and Multinomial Naive Bayes

Jens Martensson

Work Flow

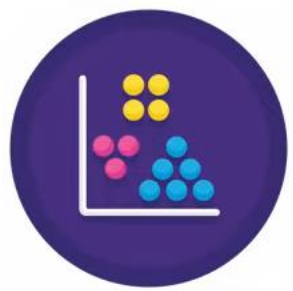
Customer Data

Data pre processing

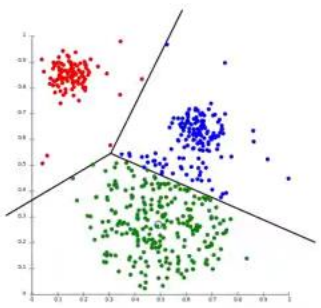
Data Analysis

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$

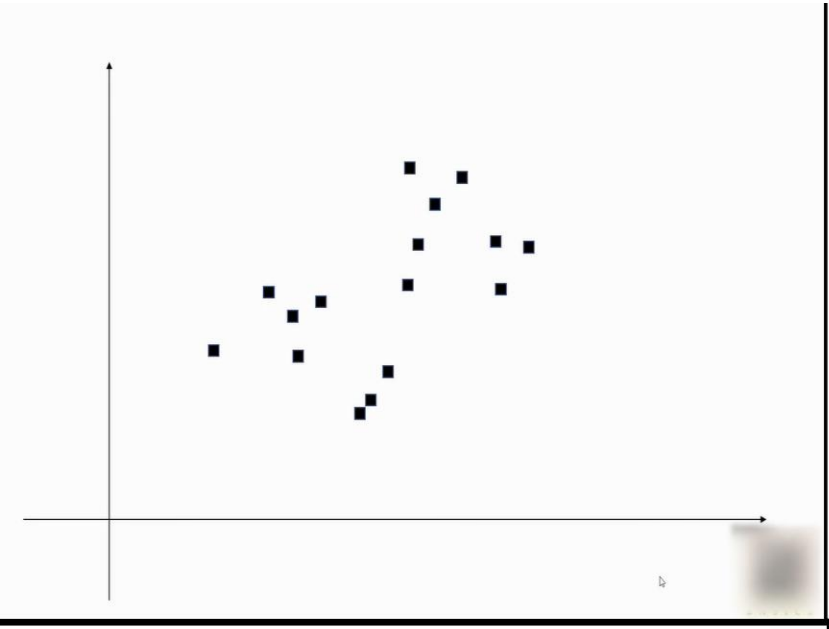
Optimum number of Clusters

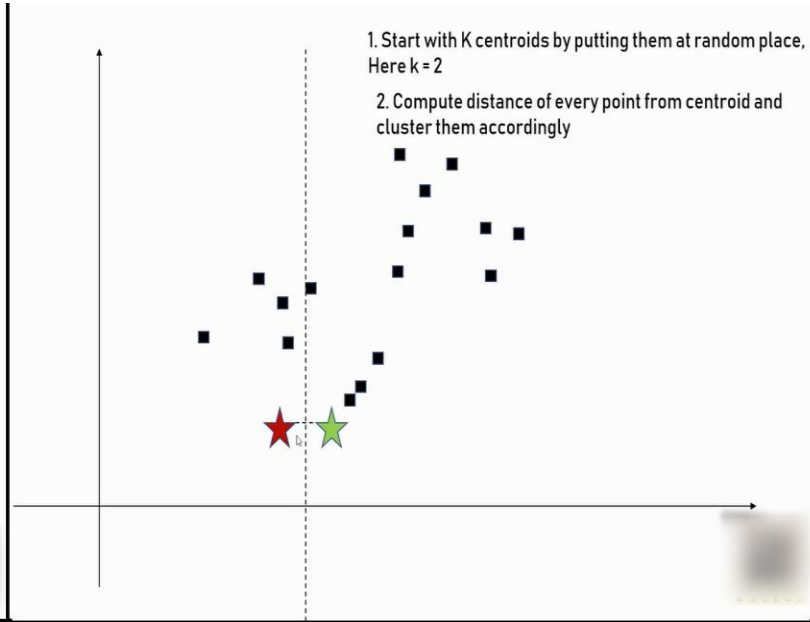


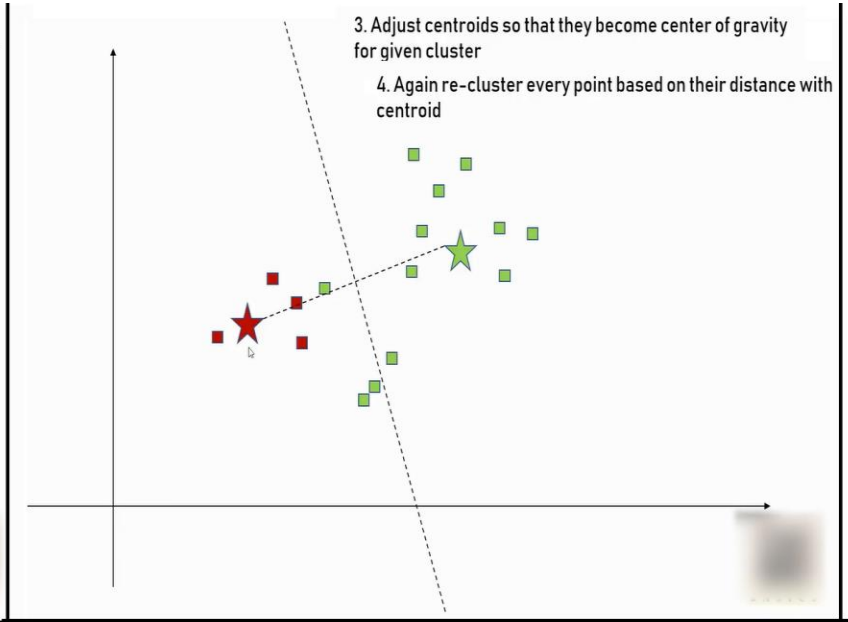
K-Means Clustering

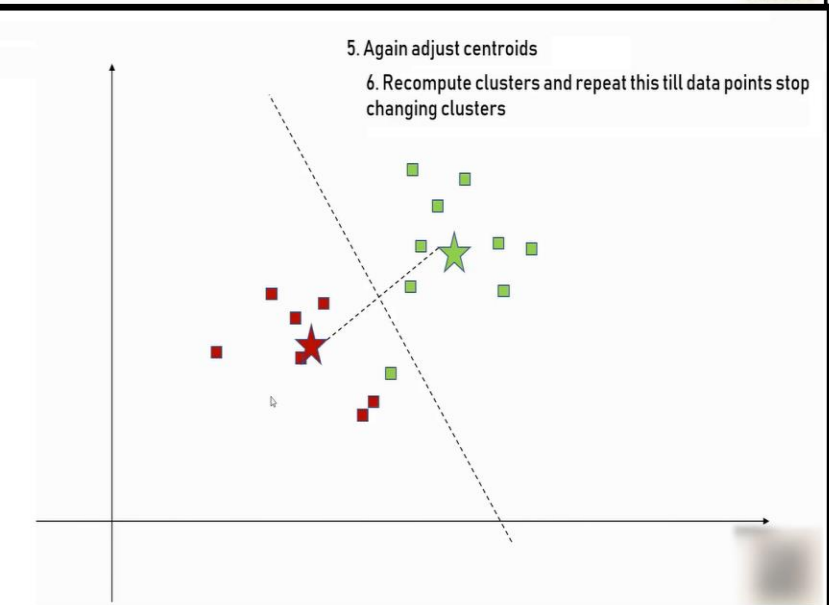


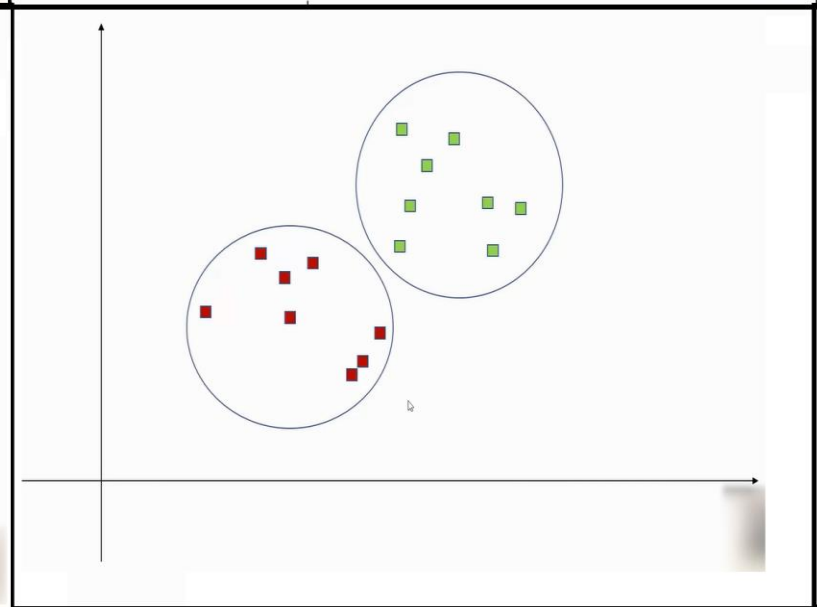
Visualizing the Clusters

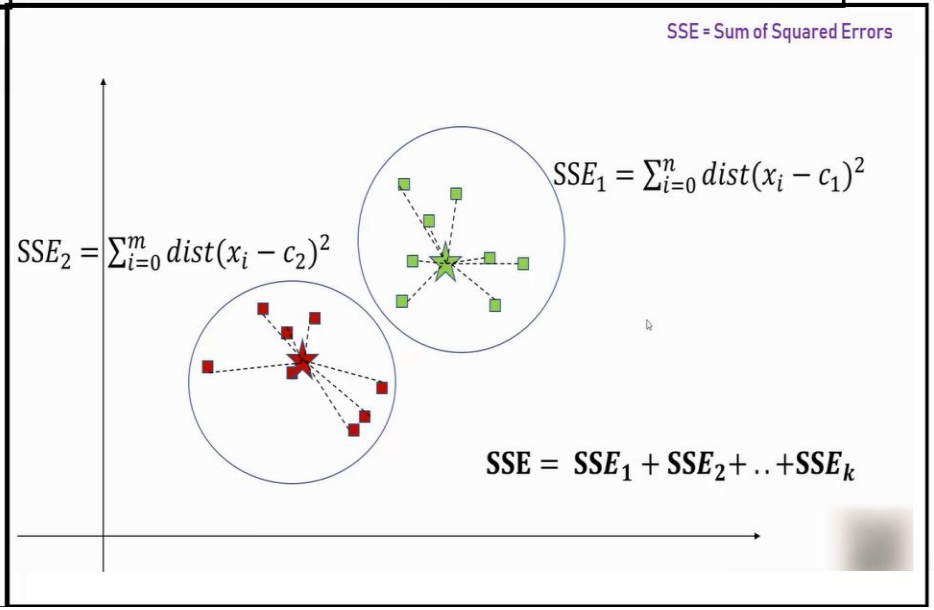



1. Start with K centroids by putting them at random place, Here k = 2
2. Compute distance of every point from centroid and cluster them accordingly


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




SSE = Sum of Squared Errors
$SSE_1 = \sum_{i=0}^{n} dist(x_i - c_1)^2$
$SSE_2 = \sum_{i=0}^{m} dist(x_i - c_2)^2$
$\mathbf{SSE = SSE_1 + SSE_2 + .. + SSE_k}$


NEG (700/1,400)
Table:
Category % n
NEG 50.0 700
POS 50.0 700
Total 100.0 1,400
Chart:
Color column: Document class
bad
<= 0.5
> 0.5
POS (665/1,117)
Table:
Category % n
NEG 40.5 452
POS 59.5 665
Total 79.8 1,117
Chart:
Color column: Document class
NEG (248/283)
Table:
Category % n
NEG 87.6 248
POS 12.4 35
Total 20.2 283
Chart:
Color column: Document class
wast
film
<= 0.5
> 0.5
<= 0.5
> 0.5
POS (665/1,000)
Table:
Category % n
NEG 33.5 335
POS 66.5 665
Total 71.4 1,000
Chart:
Color column: Document class
NEG (117/117)
Table:
Category % n
NEG 100.0 117
POS 0.0 0
Total 8.4 117
Chart:
Color column: Document class
POS (21/39)
Table:
Category % n
NEG 46.2 18
POS 53.8 21
Total 2.8 39
Chart:
Color column: Document class
NEG (230/244)
Table:
Category % n
NEG 94.3 230
POS 5.7 14
Total 17.4 244
Chart:
Color column: Document class