# K-Means Clustering

Presented by ~

Abhisek Sarkar

(as20ms091@iiserkol.ac.in)

# What is Clustering?

Clustering is an unsupervised learning technique that groups data points based on their inherient similarities, aiming to discover underlying patterns and structures within the data.

## But What is Unsupervised learning?

# Types of Machine Learning :- 3 main types :-

i) Supervised Learning :- model is trained on labeled data, meaning each training example is paired with an output label. The model learns to map input to the correct outputs. Ex :- Classification

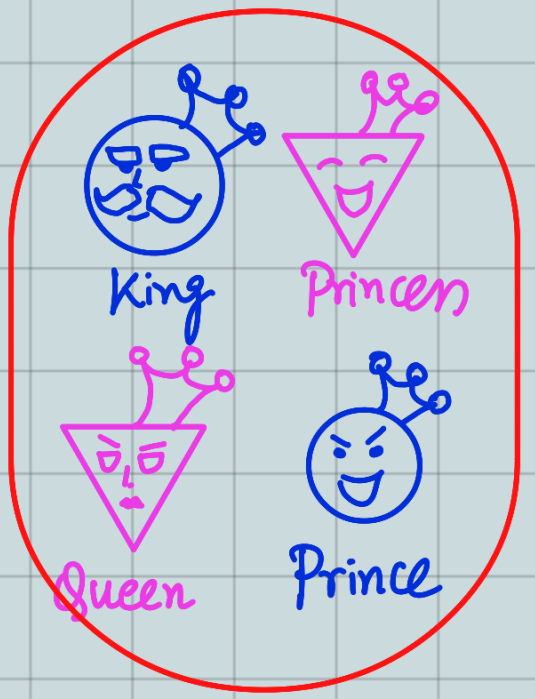
ii) Unsupervised Learning :- model deals with unlabeled data. Model tries to identify patterns or groupings within the data without any explicit labels. Ex :- Clustering

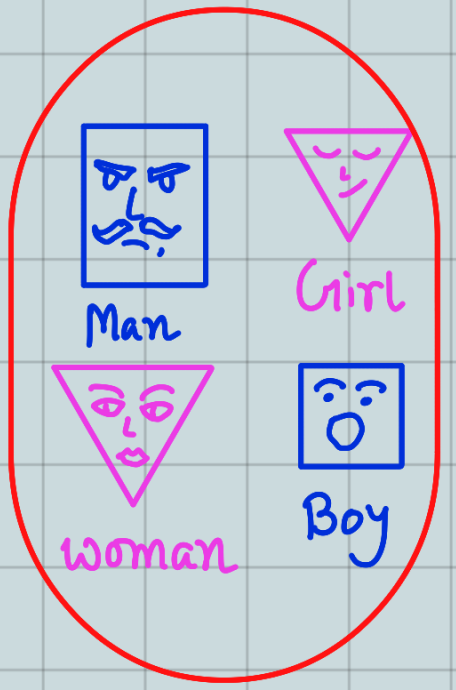
iii) Reinforcement Learning :- Agent learns to make decisions by taking actions in an environment to maximize a cumulative reward. The agent receives feedback in the form of rewards or penalties based on its actions. Ex :- Navigate a maze

What is natural grading among them?

# Clustering is subjective!

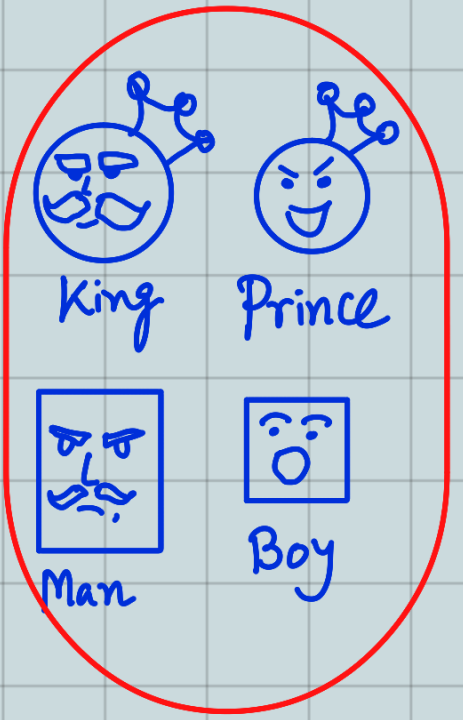
## Royality

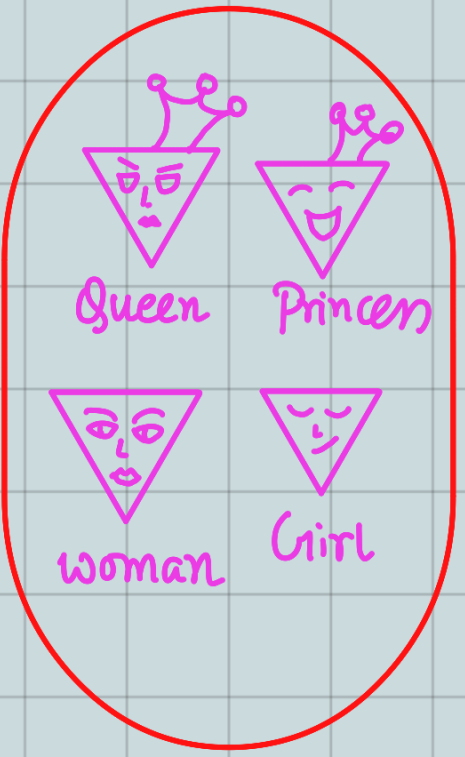



Based on Royality

## Grender




Based on Grender

# Aim of Clustering!

- high intra-class similarity
- low inter-class similarity

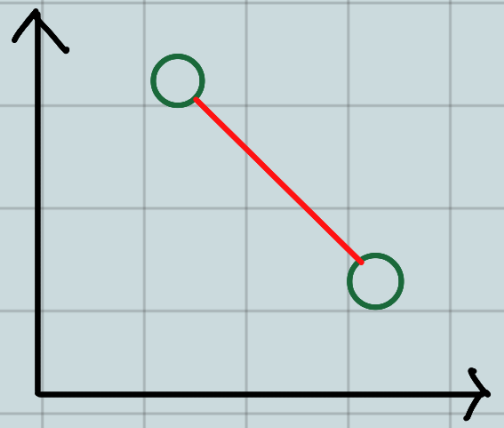
But what does determine the similarity?

The answer is Distance Measure.

Distance measure will determine how the similarity of two elements is calculated and it will influence the shape of the clusters.

Now we will discuss some distance measure.

# 1. The Euclidian distance :- (also called 2-norm distance)



measures the shortest distance between two real valued vectors.

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

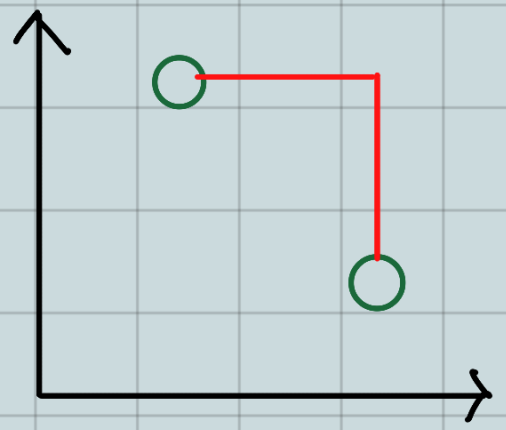
Code :-

```
from Scipy.Spatial import distance
distance.euclidean (vector_1, vector_2)
```

- Does not work well on higher dimensional data than 2D & 3D

- If we don't normalize the features, the distance might be skewed due to different units.

## 2. Manhattan distance :- (also called taxicab norm or 1-norm)



The distance betⁿ two real valued vectors is calculated as if one could only move at right angles.

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

Code :- from scipy.spatial import distance
distance.cityblock (vector_1, vector_2)

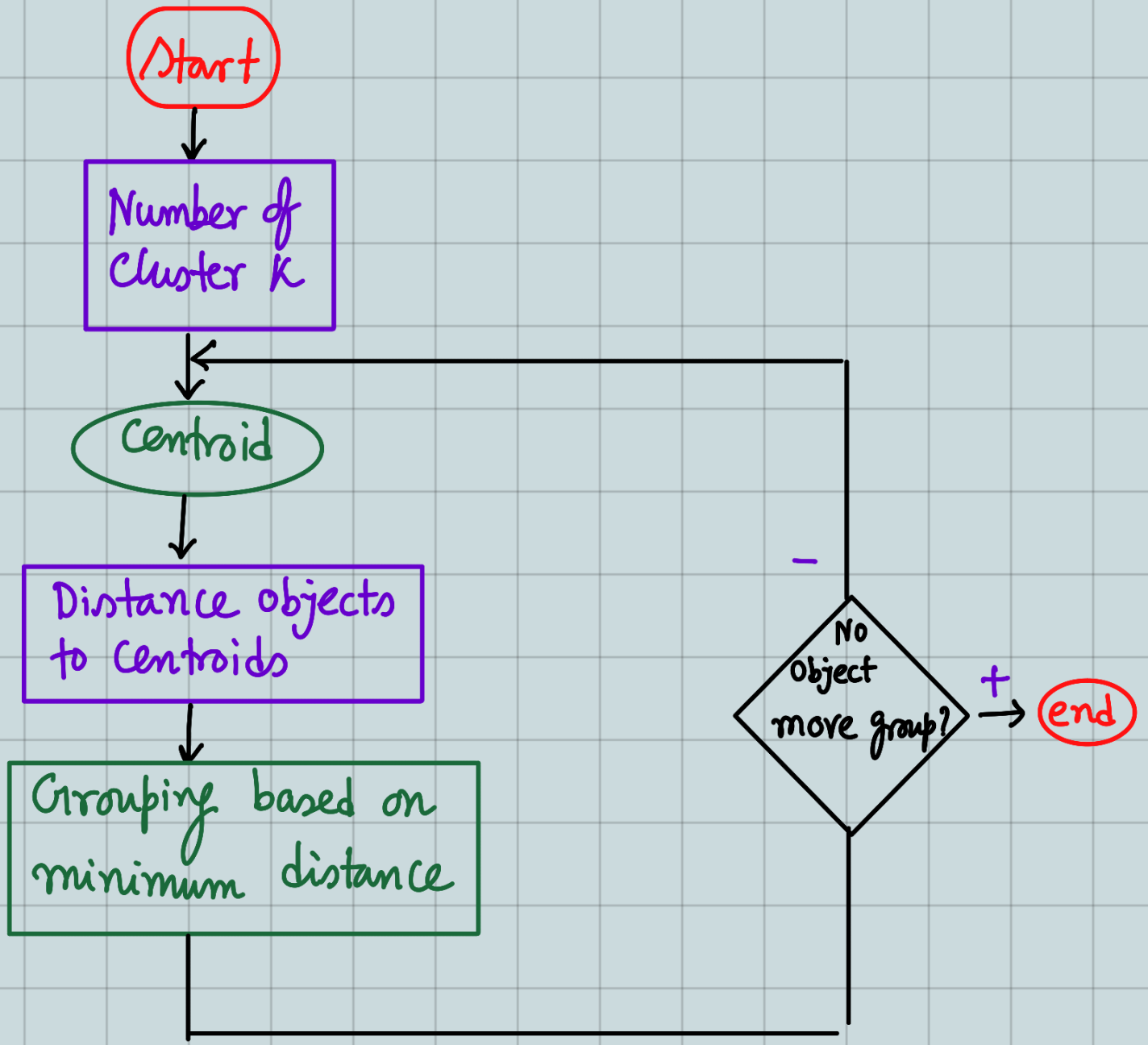
- It does not show the shortest path possible.
- Less intuitive than Eucledian distance in high dimensional space.

**K-means Clustering:-** The K-means clustering is an algorithm to cluster $n$ objects based on attributes into $k$ partitions where $k < n$

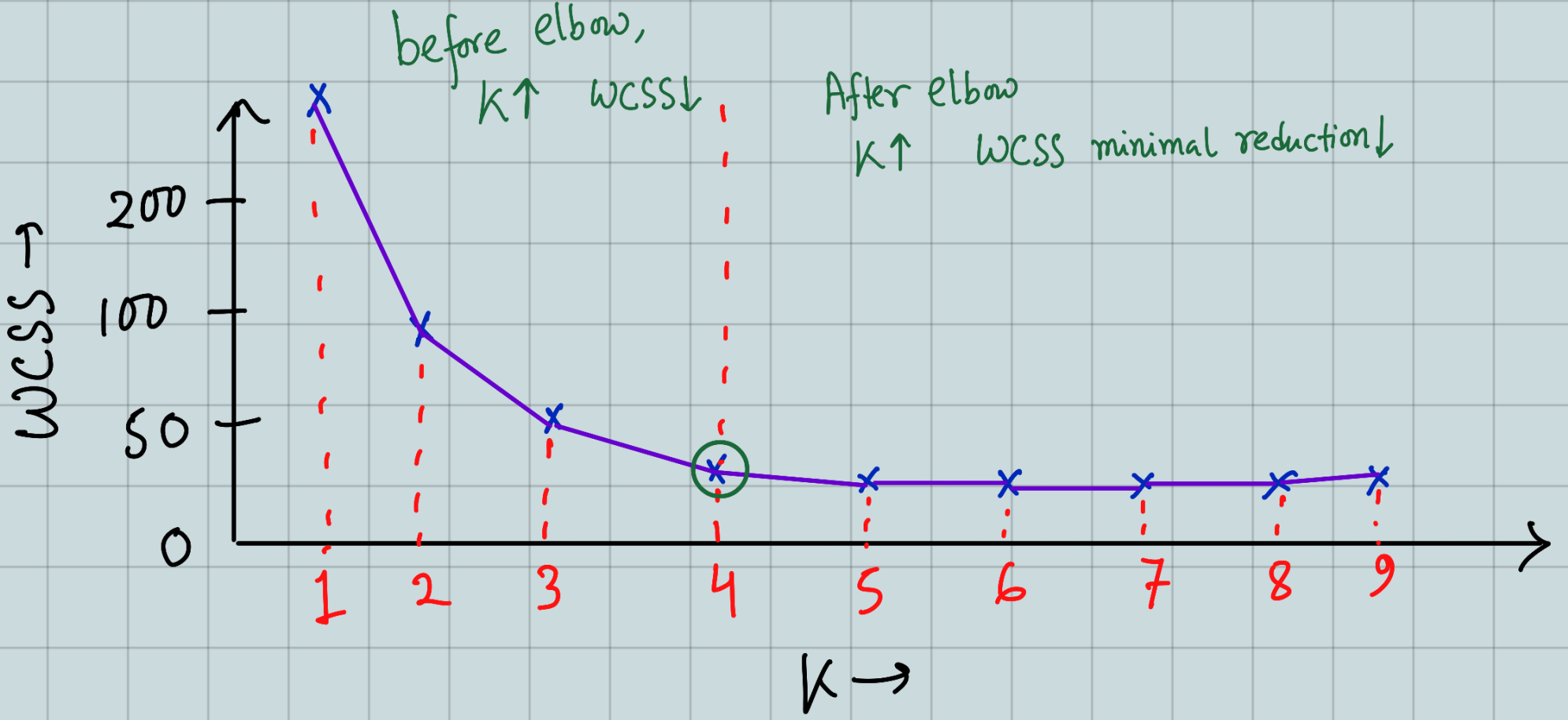
- $k$ is a positive integer number

- The grouping is done by minimizing the sum of squares of distances betⁿ data and the corresponding Cluster Centroid.

But how does it work?

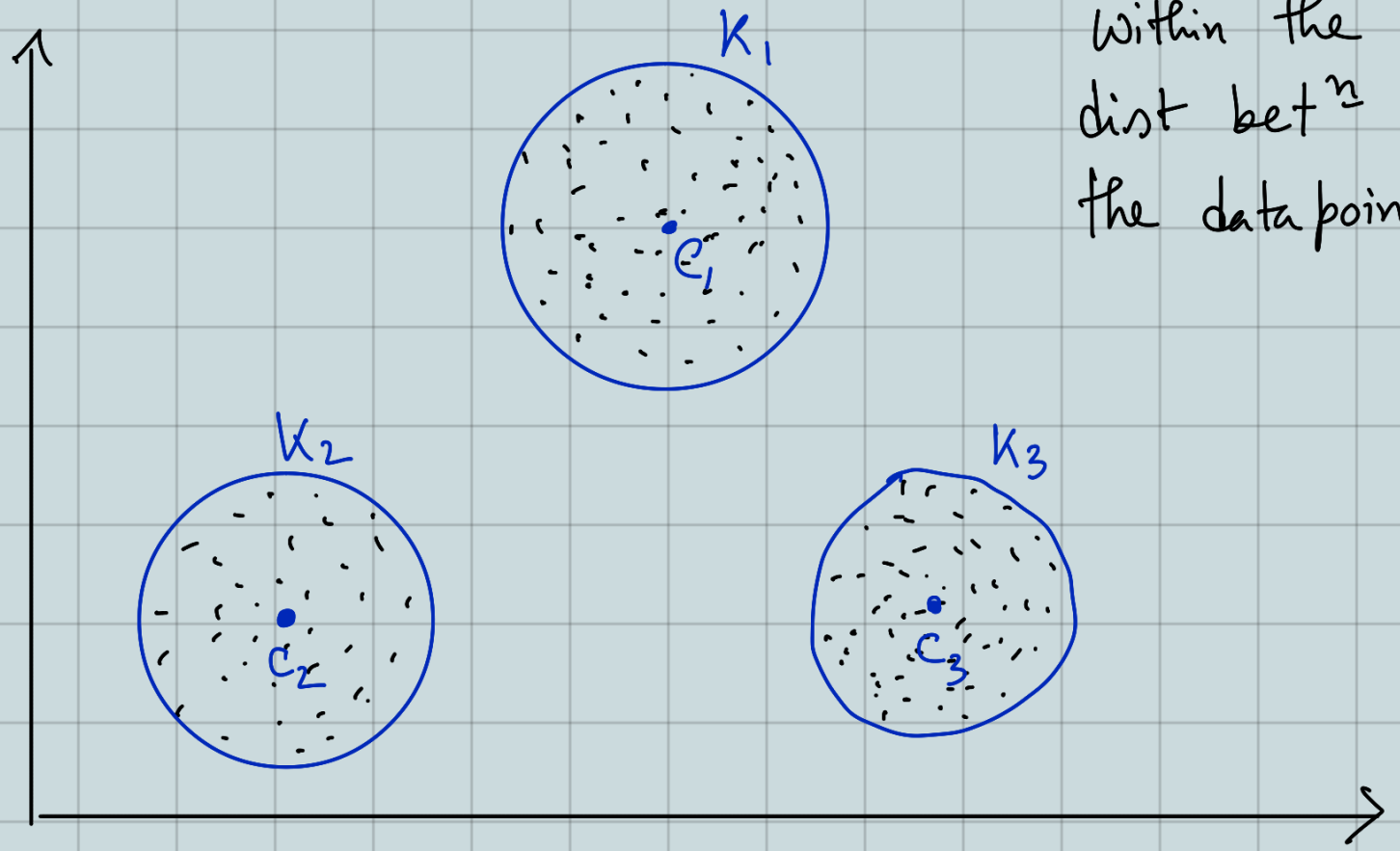

Start
Number of Cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No Object move group?
-
+
end

**Elbow Method:-** Choosing optimal number of clusters is a crucial step. Since we don't have predefined cluster counts in unsupervised learning, we need a symantic approach to determine the K- value.

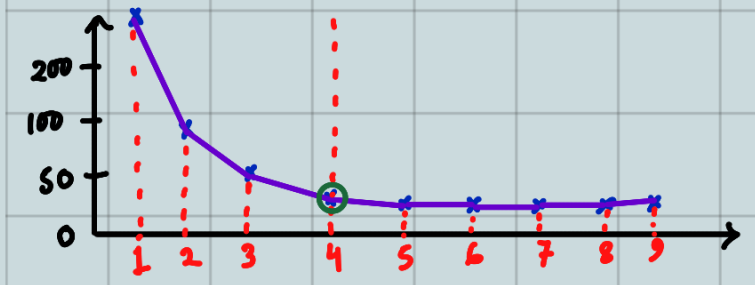


# WCSS (Within Cluster Sum of Squares)

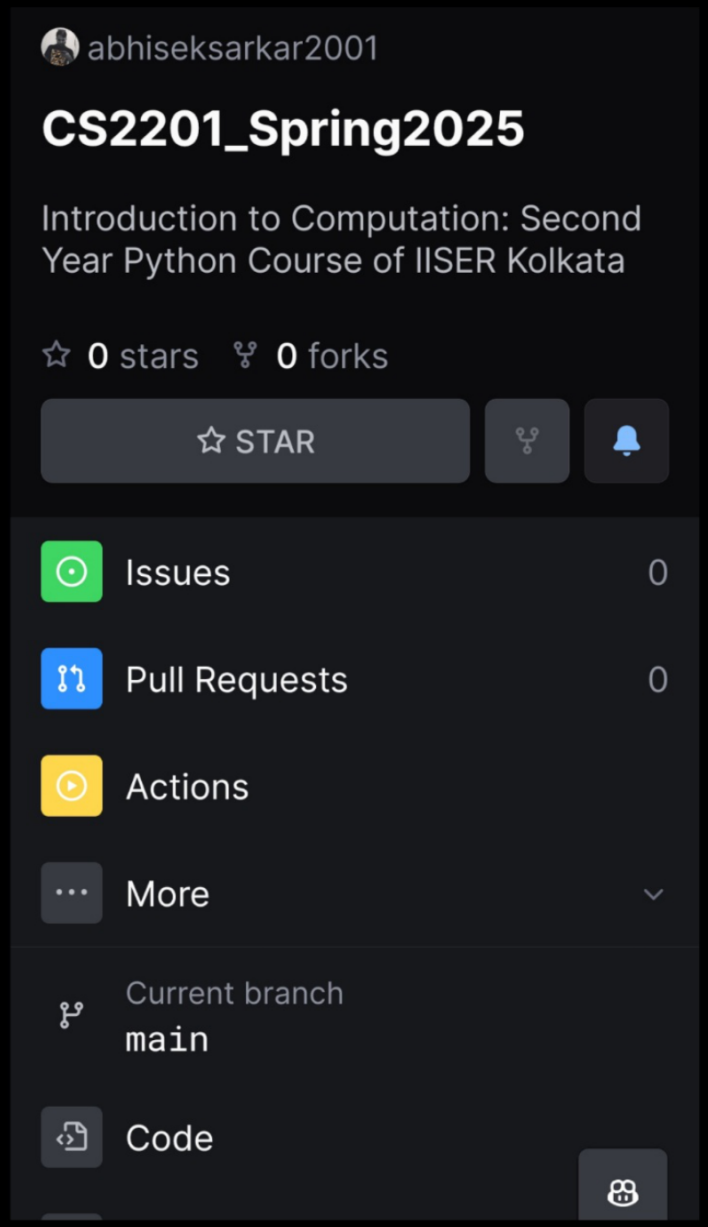
$$WCSS = \sum_{C_k}^{C_n} \left( \sum_{d_i \text{ in } C_i}^{d_m} \text{distance}(d_i, C_k)^2 \right)$$




within the cluster the dist betⁿ centeroids and the datapoints.




- We Calculate a distance measure called WCSS (within-Cluster-Sum of Squares). This tells us how spread out the data points are.
- We try different K-values and run KMeans and Calculate the WCSS.
- We plot a graph with K- on x axis and WCSS on y axis.
- We increase K, the WCSS typically decreases because we are creating more clusters, which tend to capture more data variations. However, there comes a point where adding more clusters results in only a marginal decrease in WCSS. This is where we observe "Elbow" shape in the graph.




# Visit The GitHub Repo for all the Slides, Notes and Codes for this Course

https://github.com/abhiseksarkar2001/CS2201_Spring2025