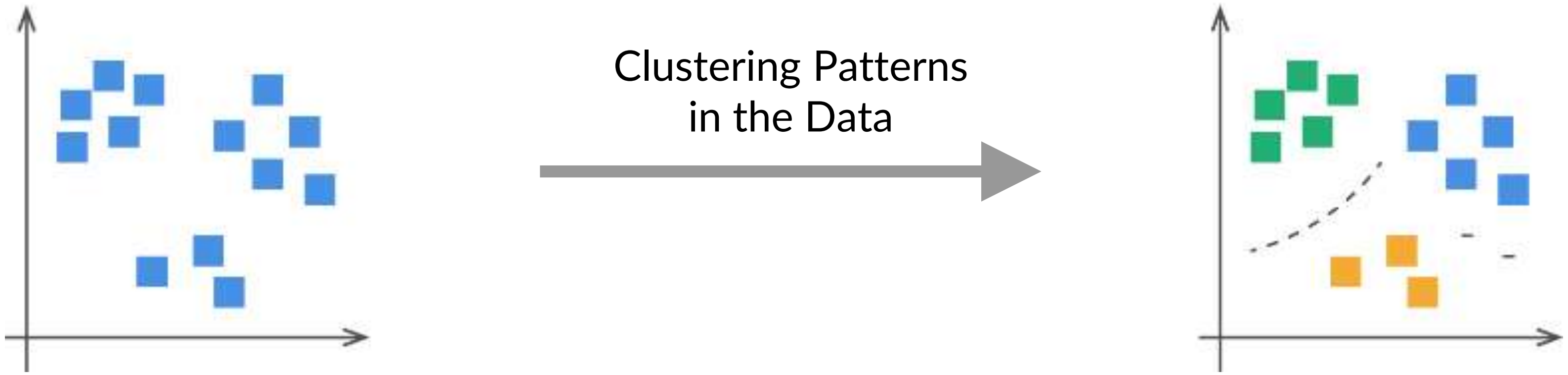
upGrad

Unsupervised Learning Clustering: K-Means

UNSUPERVISED LEARNING

- Unsupervised learning finds hidden patterns or intrinsic structures in data.
- It is used to draw inferences from data sets consisting of input data without labeled responses.
- Clustering is the most common unsupervised learning technique.



Clustering Patterns
in the Data

CLUSTERING

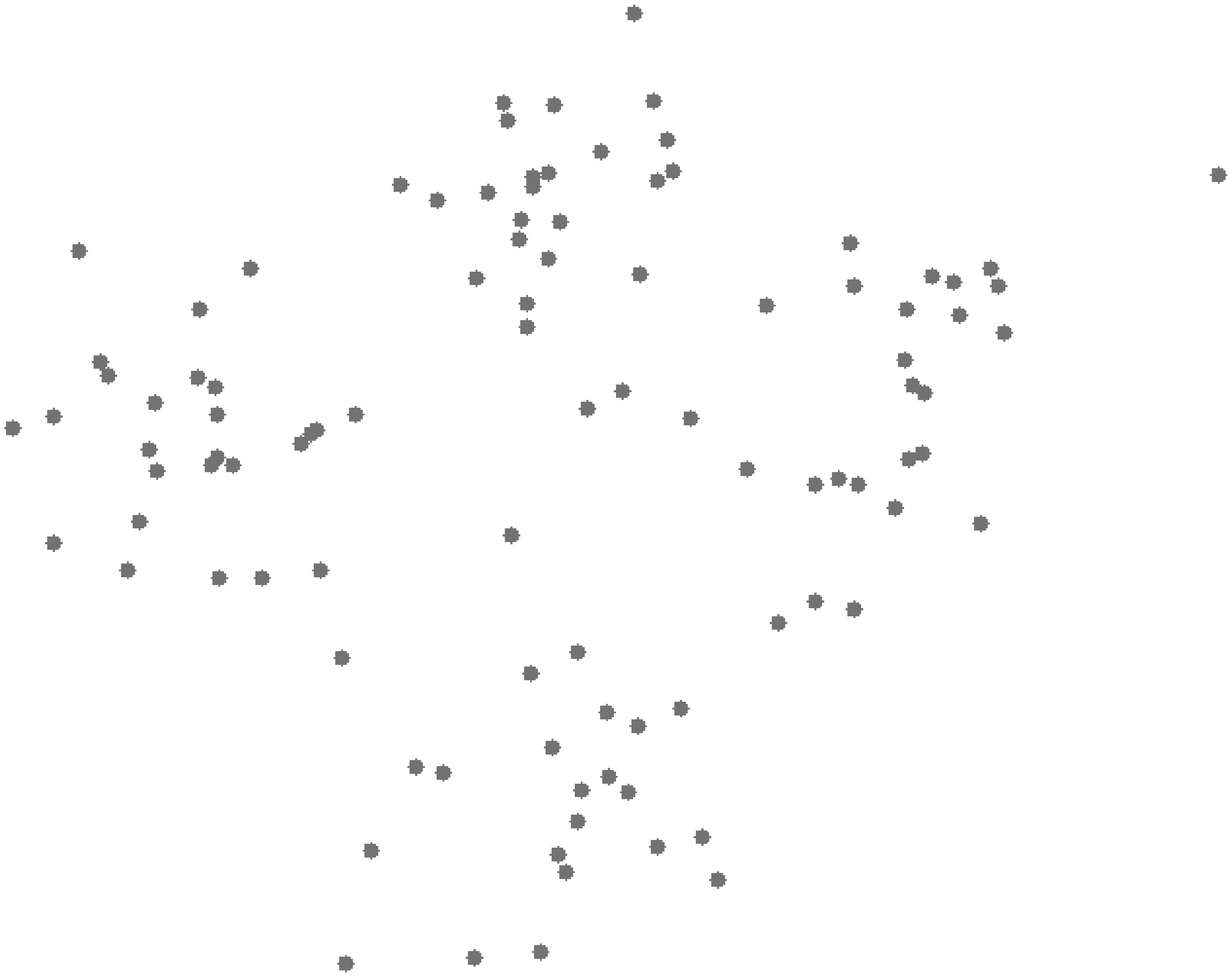
- Clustering is finding natural groups in the feature space of input data.
- Given a data set of items, with certain features, and values, clustering categorizes those items into groups of similarity (clusters).
- Clustering is the task of grouping similar objects in the same group(cluster)
- Two most commonly used types of clustering algorithms:
  - K-Means Clustering
  - Hierarchical Clustering

CLUSTERING: USE-CASE

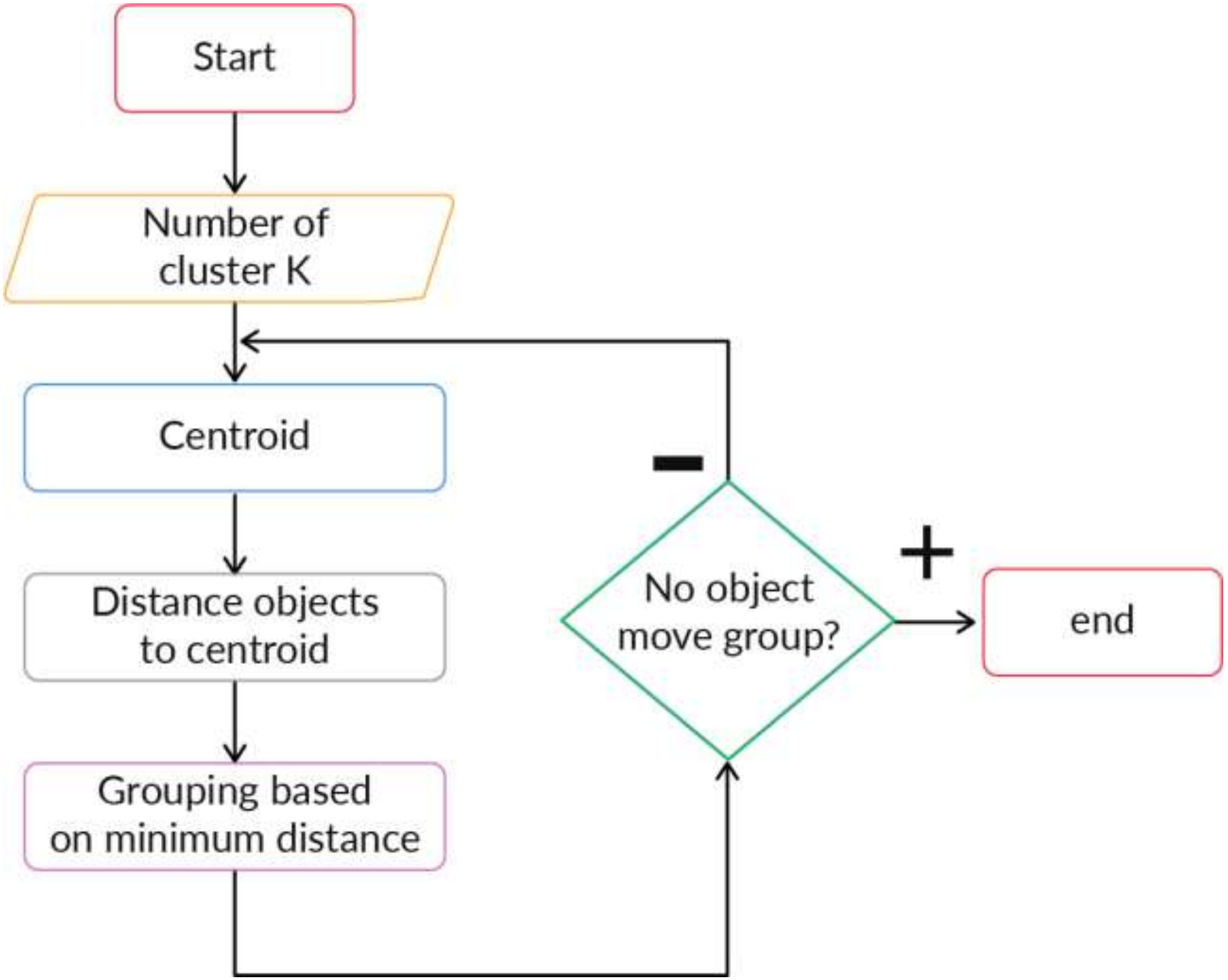
- Clustering is very widely used in the industry. Common examples:
  - Customer Segmentation
  - Marketing: Targeting set of user group
  - Document Clustering [Google News]
  - Healthcare
  - Social Media

K-MEANS

- K-Means categorizes data points into k similar groups(cluster)
- Algorithm Steps:
  - Select K, the number of clusters you want. Let’s select K=4.
  - Initialize k random points as centroid of the initial cluster
  - Measure the euclidean distance between each data point and each centroid and assign each data point to its closest centroid and corresponding cluster.
  - Recalculate the midpoint(centroid) of each cluster.
  - Repeat steps three and four to reassign data points to clusters based on the new centroid locations.
  - Stop when the centroids have been stabilized, after computing the centroid of a cluster, no data points are reassigned.
- Animation for k=4 in next slide



K-MEANS


Start
Number of cluster K
Centroid
Distance objects to centroid
Grouping based on minimum distance
No object move group?
–
+
end

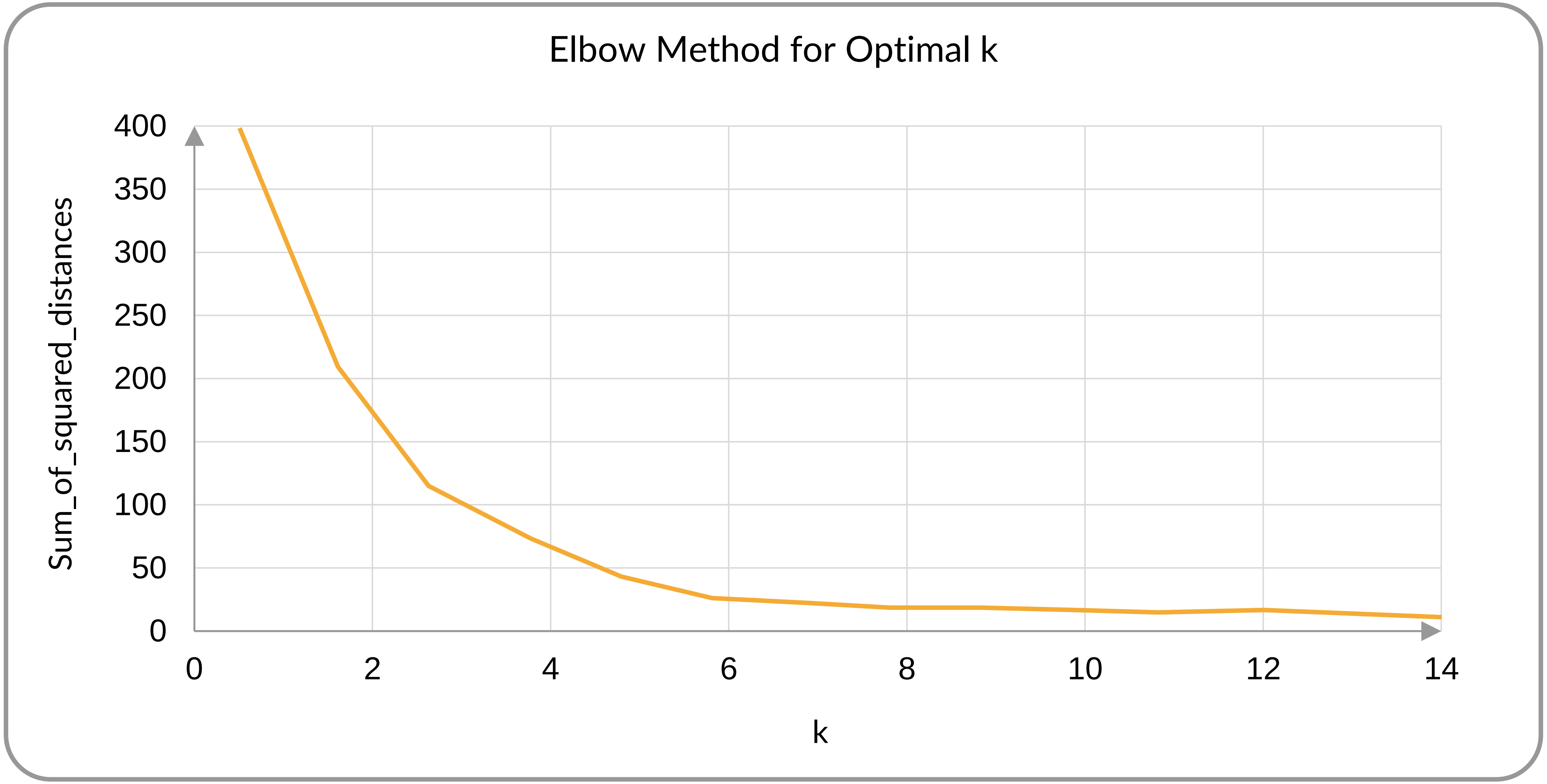
K-MEANS LOSS FUNCTION

- Loss function:

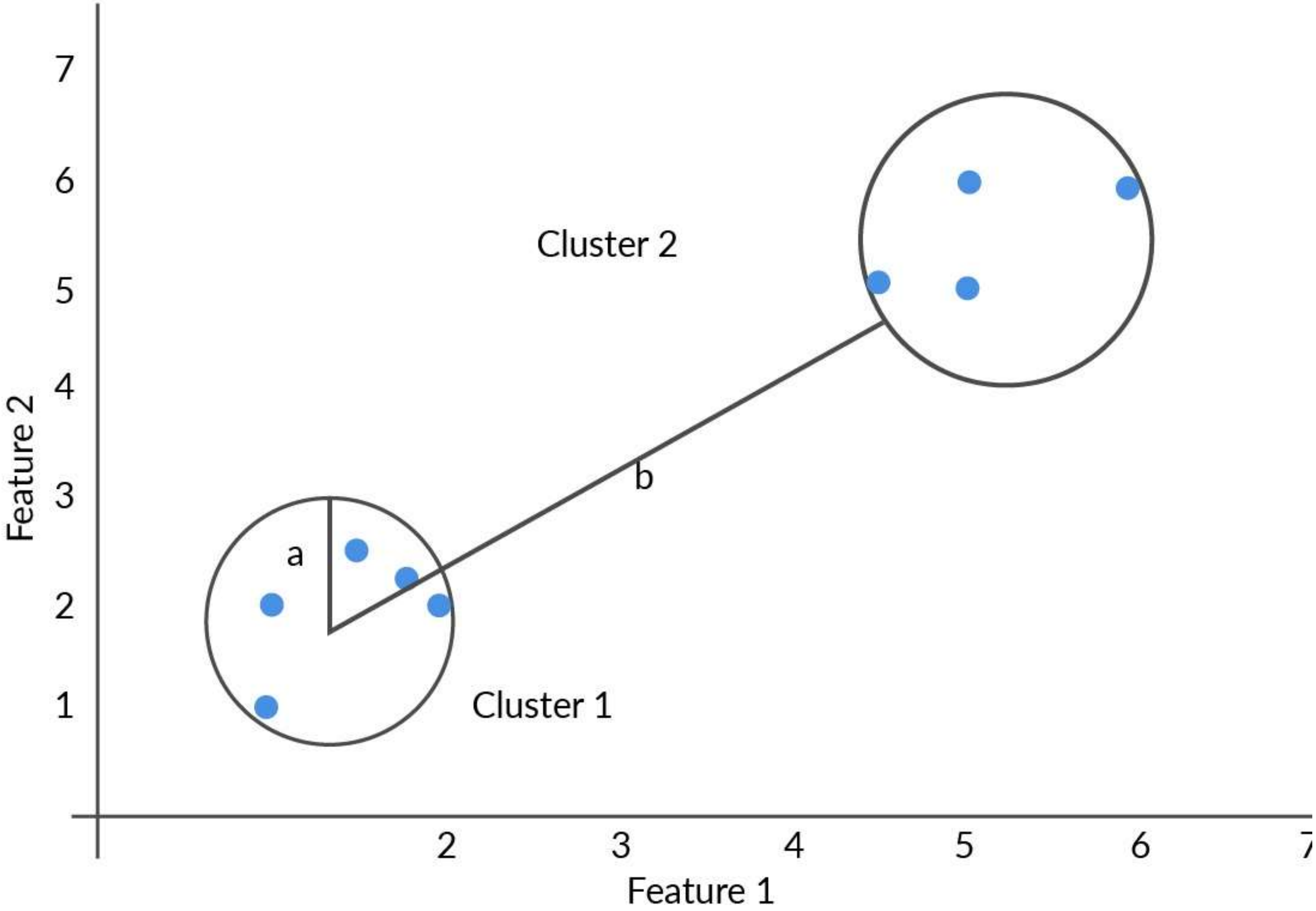
$$J = \sum_{i=1}^{n} ||X_i - \mu_{k(i)}||^2$$

- Where
  - J is loss function
  - Xi is ith data point
  - Uki:
    - Ki gives the cluster number corresponding to ith datapoint.
    - Uki is the centroid associated to that datapoint.
- Loss function is the sum of the squared distances from each point to the associated cluster center
- To get the best possible clusters, the loss function should be minimum

HOW TO SELECT K?



Silhouette Coefficient Score



upGrad

Case Study:

Case Study

- Last.fm UK based Music Company, now acquired by Spotify
- Problem Statement:
    - Find Cluster of Similar music artist from the data based on their popularity in terms of how many times people have listened their song.Initialize k random points as centroid of the initial cluster
    - Can be used for recommendation: Similar artist song to users
    - Useful for monetization & business point of view: Exclusive launch of songs on the platform [profit sharing]
    - Solve cold start problem: Categorizing new artist songs in a cluster based on the features.

Dataset

- Dataset contains user_id, artist_id, artist_name, plays.
- Where:
  - user_id: Unique id of each user playing the songs.
  - artist_id: Unique id of each artist whose song is present on the dataset.
  - artist_name: Name of the artist.
  - plays: Total number of times user has listened to this artist song.

- Data exploration in excel.