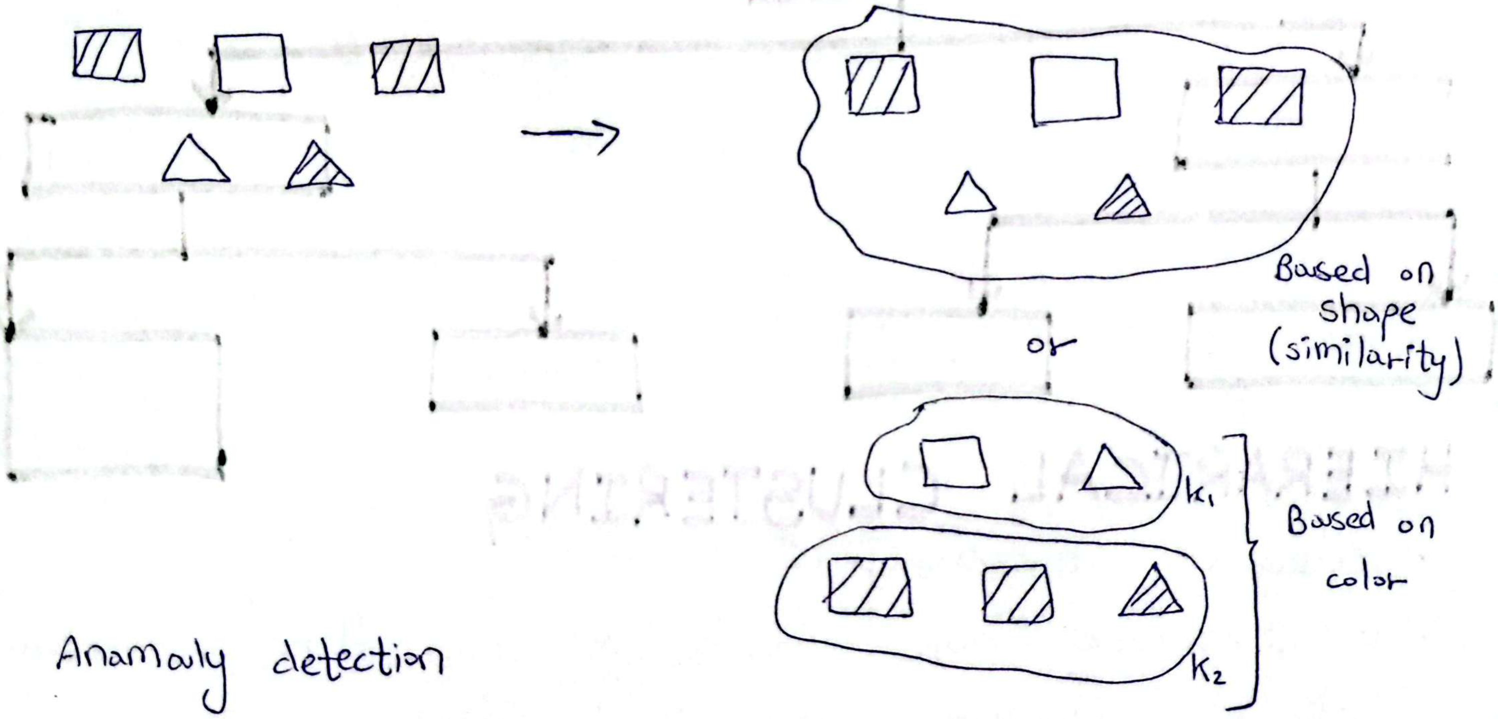
UNSUPERVISED LEARNING

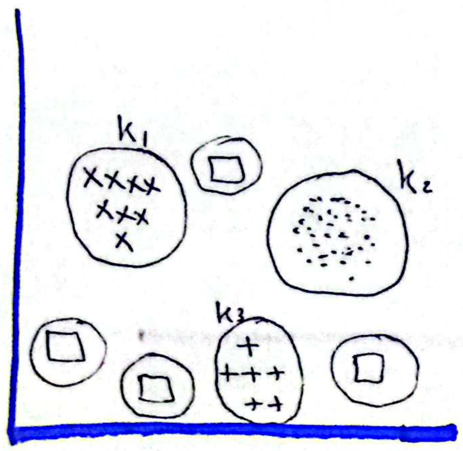
→ For data that is
- Unclassified
- Unlabelled
- More complex
- Moderately accurate but reliable results.

→ Used for
- finding patterns (clustering)
- Anamaly detection

Example



Anamaly detection



→ Fault detection

→ Intrusion detection

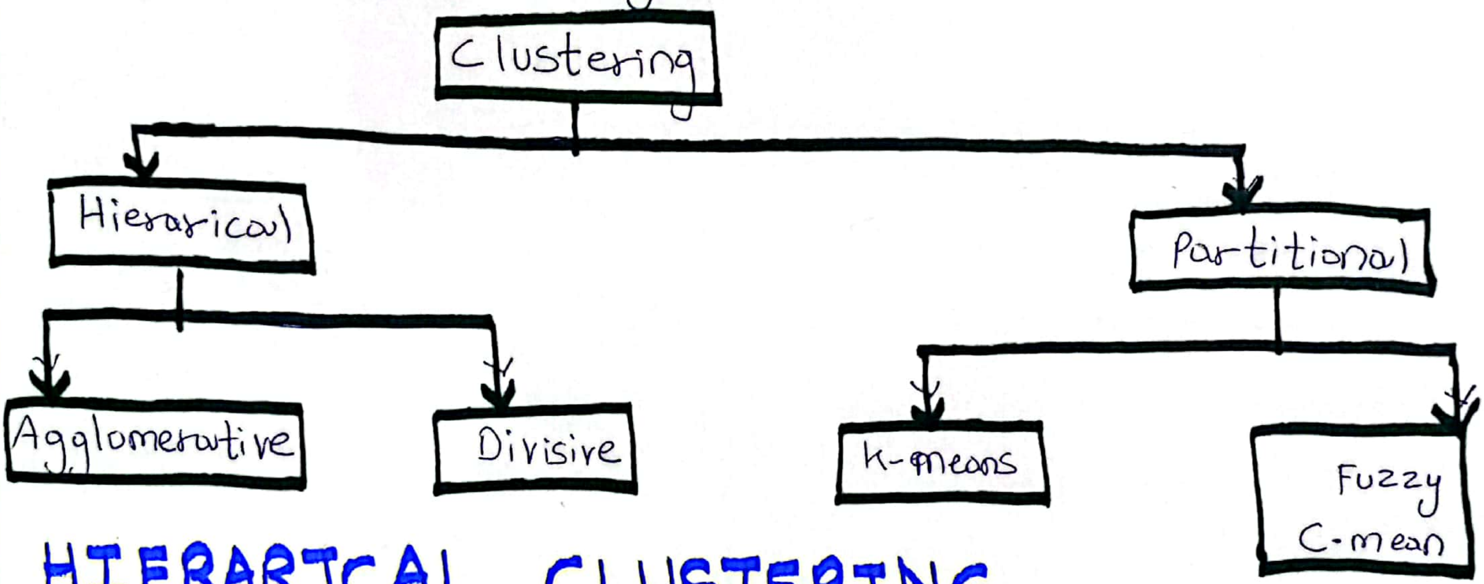
→ System fault detection.

→ Need for clustering

in unlabelled, structured & unstructured data.

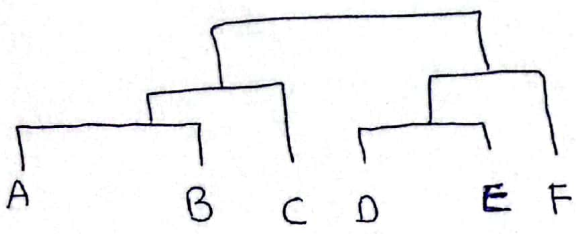
- To determine intrinsic grouping.
- To organize data into clustering showing the internal structure of data.
- To partition the data points.
- To understand and exhibit value from large sets of structured & unstructured data.

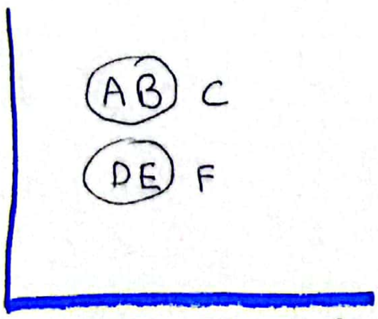
→ Types of clustering



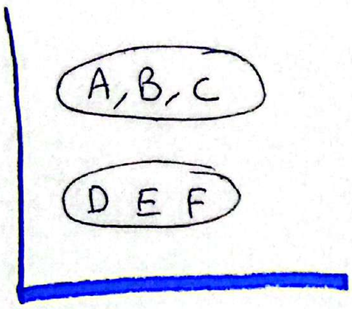
HIERARICAL CLUSTERING

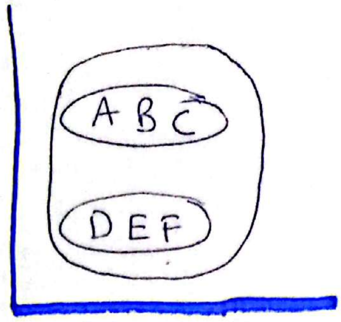
- Occupies hierarchy
- A structure more informative than the unstructured set of clusters returned by flat clustering.





Combine A and B,
D and E





STEPS

1. Assign each item to its own clusters (e.g if there are N items, you will have N clusters)
2. Find the closest (most similar pair) of clusters & combine them.
3. Compute similarities (distance) between the new clusters & every old cluster, then combine.
4. Repeat step 2 & 3 till all "N" items are in single cluster.

PARTITIONAL CLUSTERING

- Division of data into non-over-lapping clusters, where a data object is only in one set (cluster).

DISTANCE MEASURE

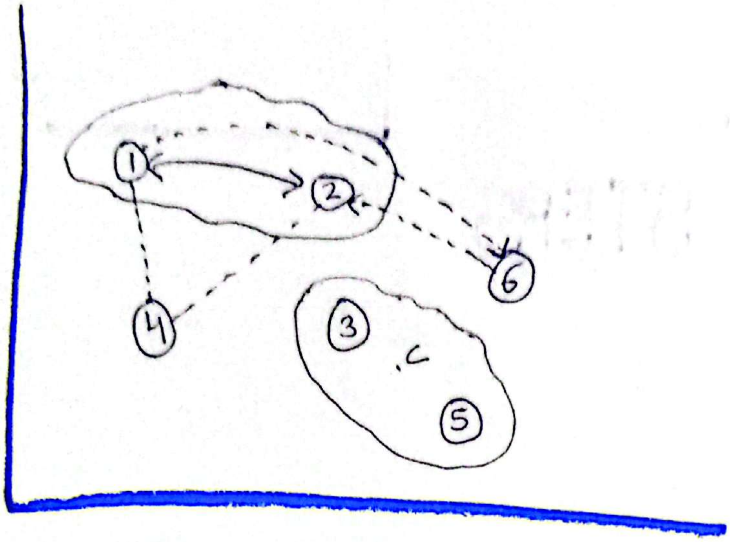
1) Complete Linkage Clustering
↳ maximum possible distance between points.

2) Single Linkage Clustering
↳ minimum possible distance between points.

3) Mean Linkage Clustering
↳ find all possible pair-wise distance between pair-wise two clusters & then calculate the average distance.

4) Centroid Linkage clustering

↳ find centroids of each cluster & calculate the distance between them.



K MEAN CLUSTERING

Step 1 → Choose cluster's ($k=2$ e.g. k_1, k_2) centroids

Step 2 → Calculate Euclidean Distance of each point (item)

$$ED = \sqrt{(x_p - \hat{x}_c)^2 + (y_p - y_c)^2}$$

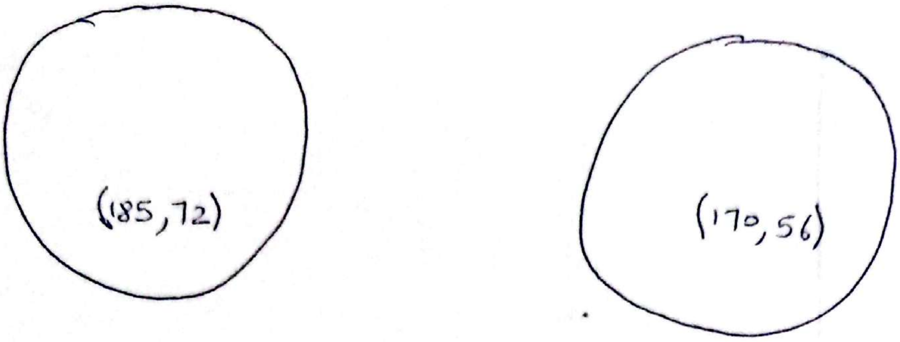
Step 3 → Put the point (item) with smallest (nearest) ED in respective cluster.

Step 4 → Recalculate the respective cluster's centroid with new addition.

Step 5 → Repeat step 2~4.

SNo	Height	weight
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72

Step 2: $k_1 = \{1, 4, 5\}$ $k_2 = \{2, 3\}$



Step 3:

$$ED = k_1 \rightarrow \sqrt{(168-185)^2 + (60-72)^2}$$

$$\rightarrow 20.8$$

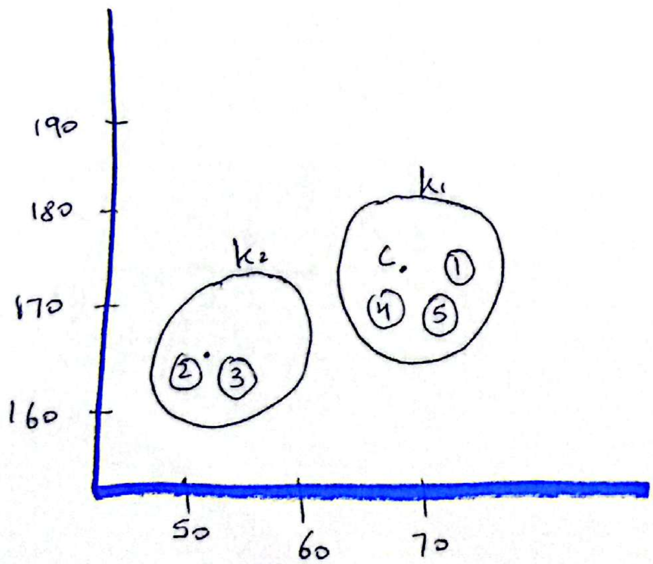
$$k_2 \rightarrow \sqrt{(168-170)^2 + (60-56)^2}$$

$$\rightarrow 4.48$$

Step 4:

$$k_2 \text{ centroid} = \left(\frac{170+168}{2}, \frac{60+56}{2}\right)$$

$$= (169, 58)$$

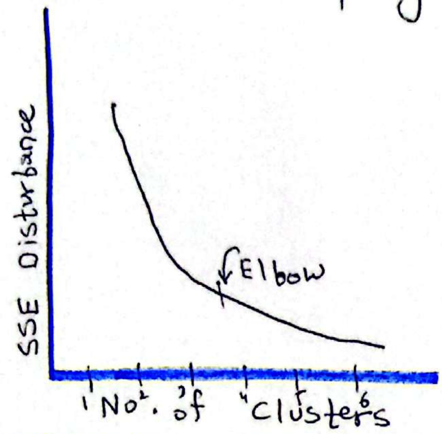


How Many Clusters?

- It is a fundamental issue in k-mean clustering.
- If sum of square error (SSE), you will see the error decreases as K increases because their size decreases

& hence distortion is also small.

- The goal of the Elbow method is to choose k, where SSE decreases abruptly.



SILHOUETTE COEFFICIENT (SC)

We have to compute

Step 1: SC of each point

$$1 - \frac{a}{b}$$

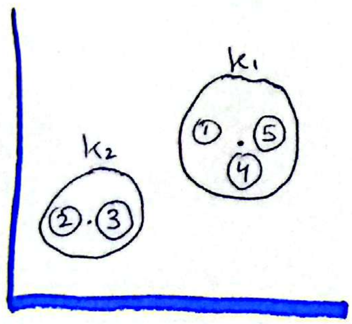
a (avg distance of a point to all other points in ~~another~~ a cluster)

b (Minimum avg distance of a point to all points in another cluster)

Step 2: SC of each cluster.

Step 3: of all clusters

EXAMPLE



Step 1:

$$a = \frac{\{(1 \to 5) + (4 \to 5)\}}{2 \text{ (No. of Points)}}$$

$$b = \frac{\{(1 \to 2) + (1 \to 3)\}}{2}$$

$$\text{SC of } 1 = 1 - \frac{a}{b}$$

Step 2: SC of each cluster.

Let's suppose SC of 2 & 3 is x & y respectively

$$\text{SC of } k_2 = \frac{x+y}{2}$$

Step 3: Overall SC

$$SC = \frac{(\text{SC of } k_1) + (\text{SC of } k_2)}{2}$$