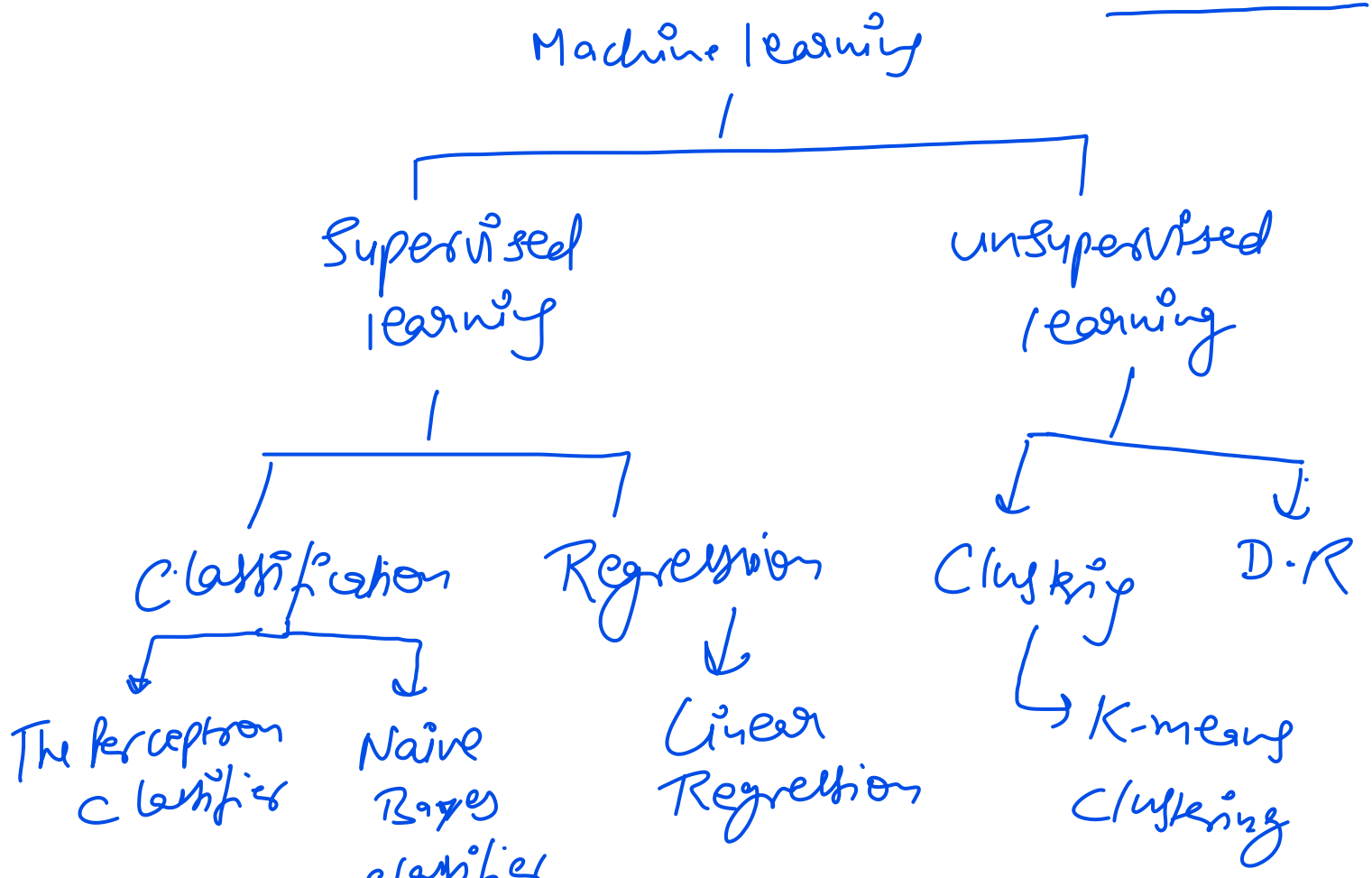
22/Nov/21



Unsupervised learning

Name	Income	Job Profile

$X \rightarrow [x_1, x_2, x_3 - \cdots - x_n]$

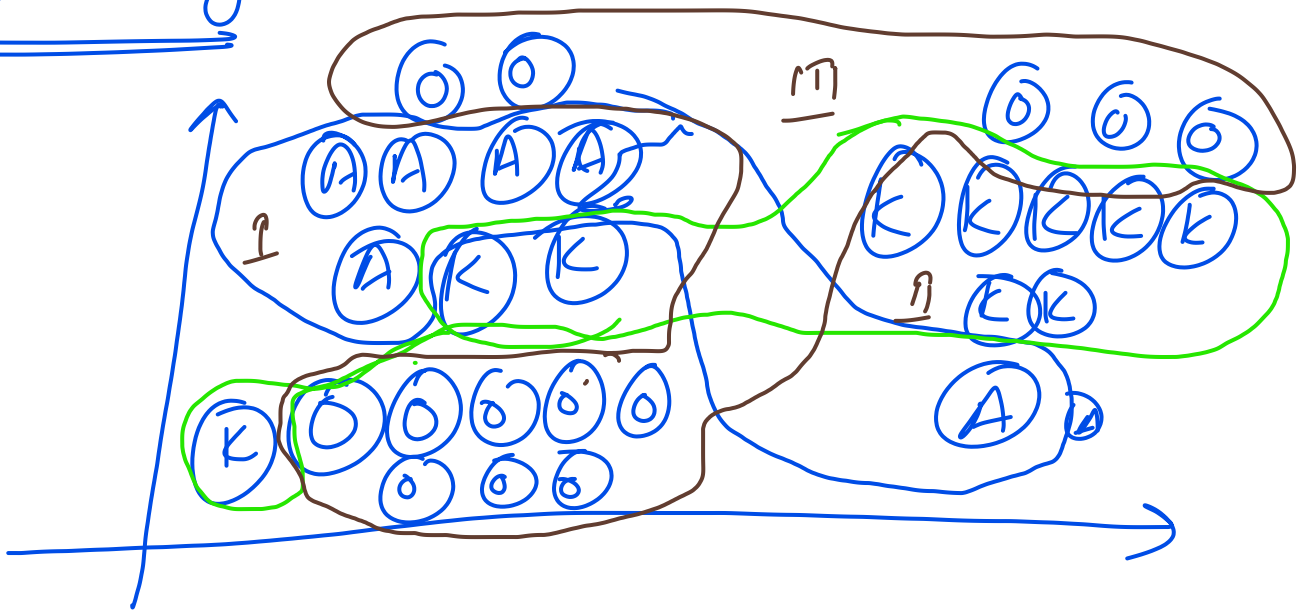
nobs features/objects

Google Search

eg manifold Google search.

eg Security breaches.

Clustering -

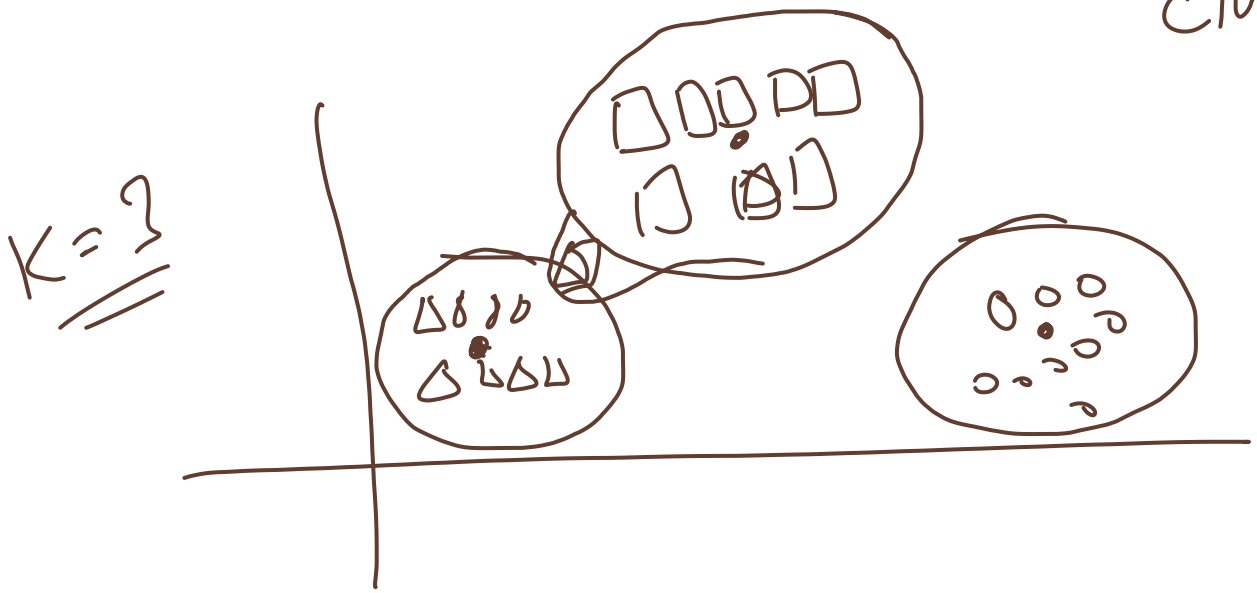


Types of clustering

- Partitional clustering
- Hierarchical clustering
  - Agglomerative
  - Divisive

① Partitional clustering - On the basis of similarity or dissimilarity, we will perform the clustering.

(unnested cluster)

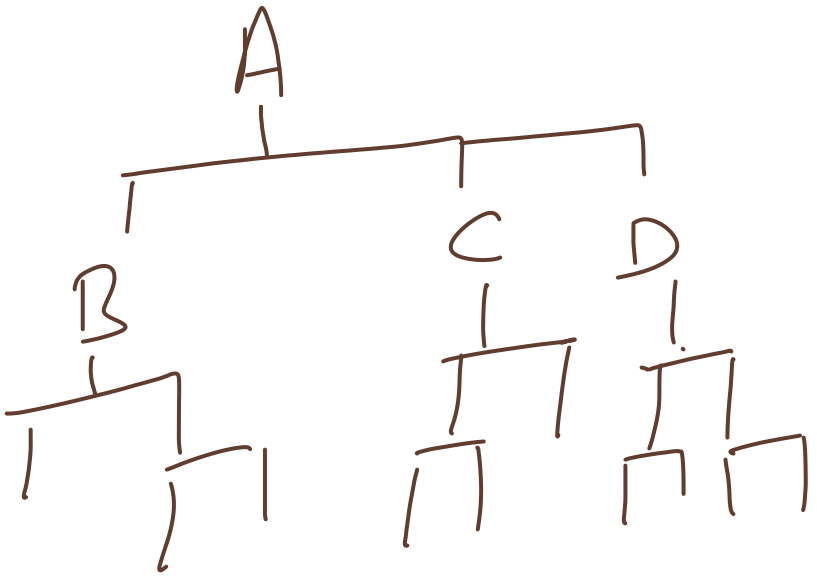


eg K-means

② Hierarchical clustering -

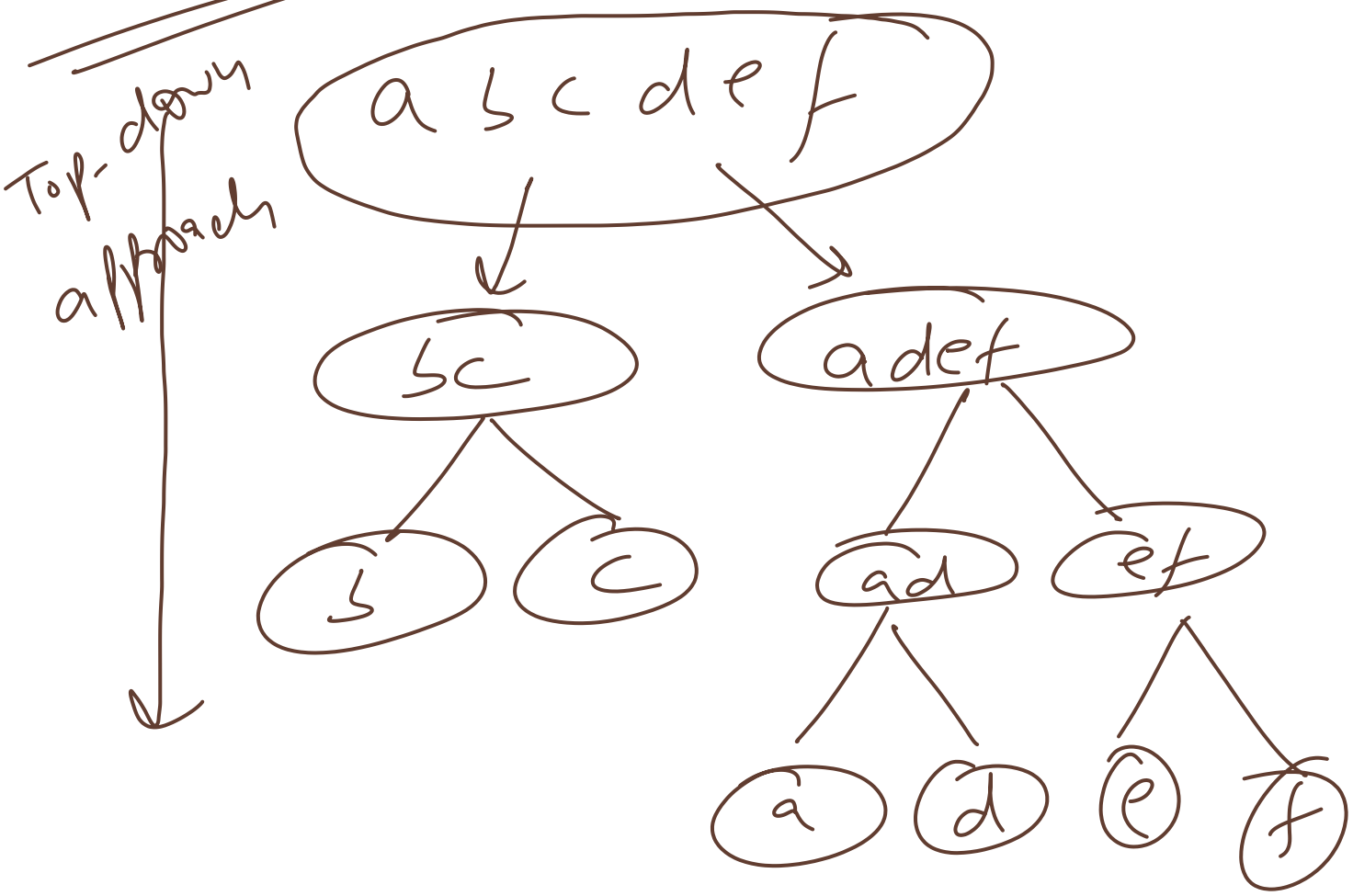
(nested clustering)

Dendogram



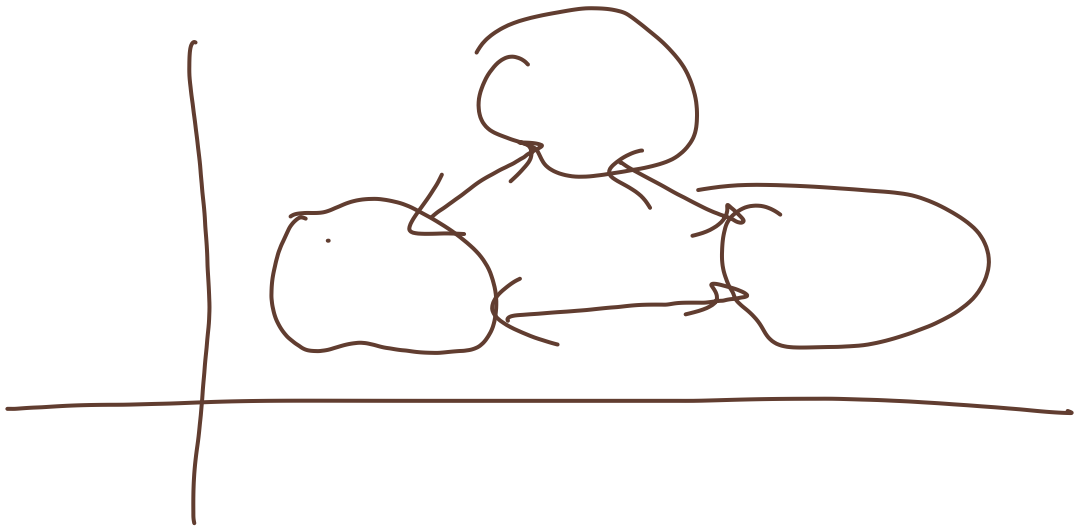
Agglomerative (Bottom - up approach)

Divisive



Distances in clustering -

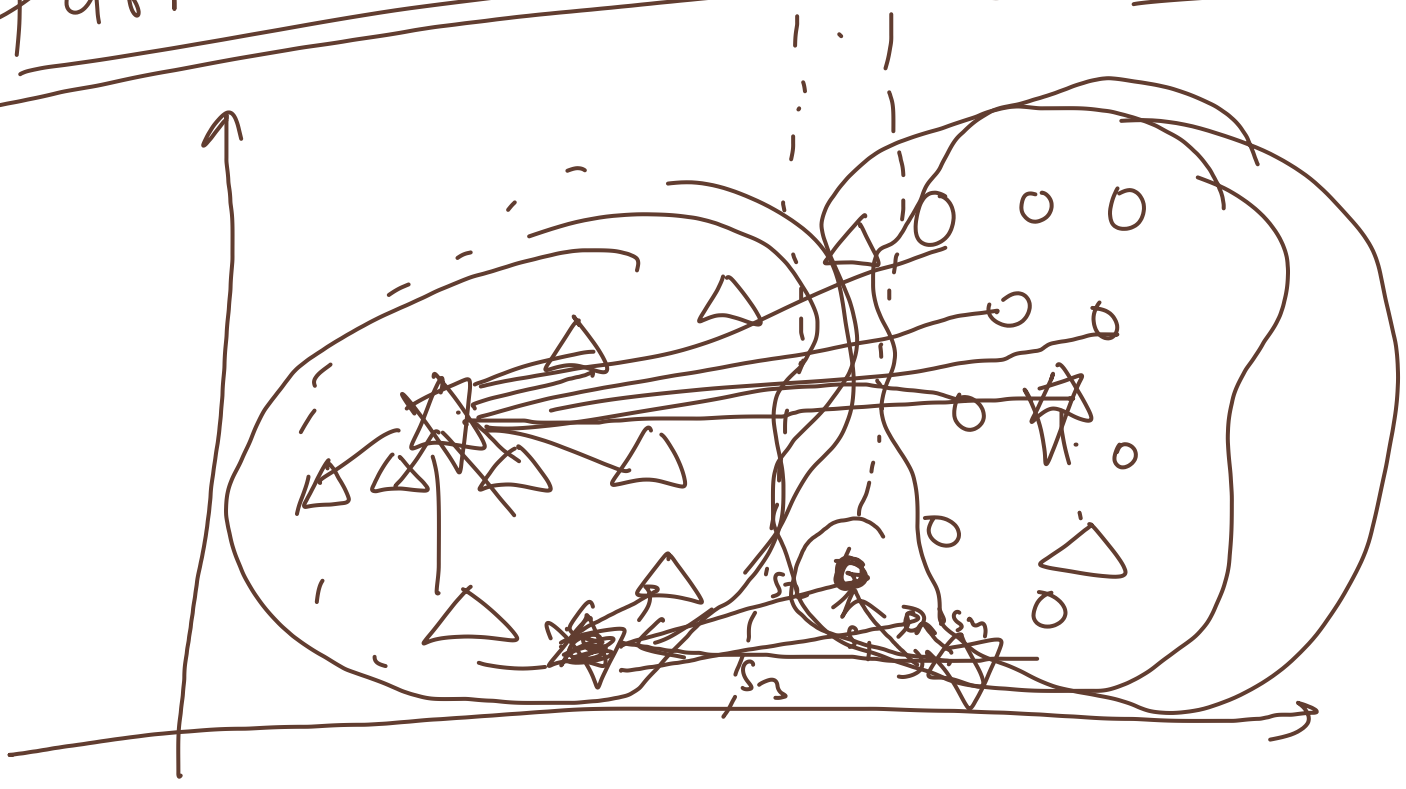
(1) Inter clustr distance

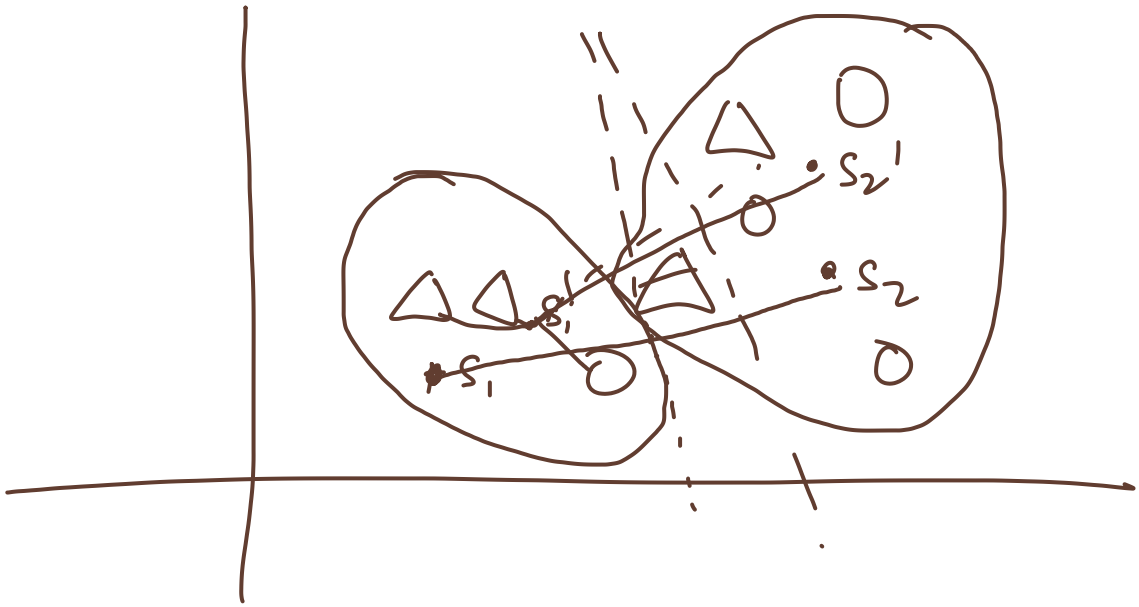


(2) Intra cluster distances

* Clustering is ambiguous in nature.

Partitioned clustering - (K-means)





Algorithm (K-means)

Step 1°:- Choose K points randomly.

Step 2°:- Calculate distance and Assign them to a choosen point (Cluster).

$$\therefore \text{Euclidean distance} = \sqrt{\sum_{s=1}^{n} (x_{is} - x_{js})^2}$$

Step 3°:- Optimization criteria (Stopping coeteria)

SSE :- Sum of squared distance (error)

$$SSE = SSE_1 + SSE_2 + SSE_3 + \cdots + SSE_K$$

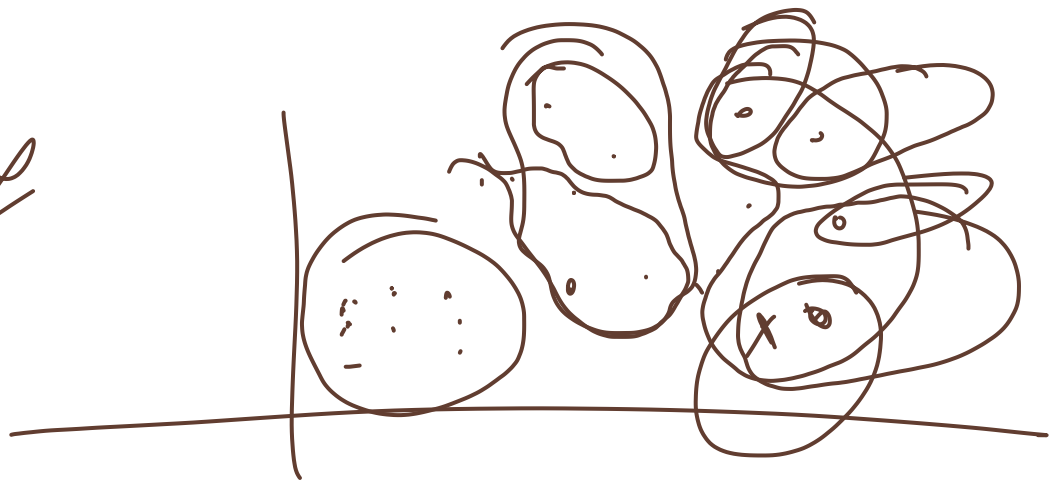
$$SSE_1 = \sum_{x \in S} \| x_i - c_i \|^2$$

c_i :- centroid

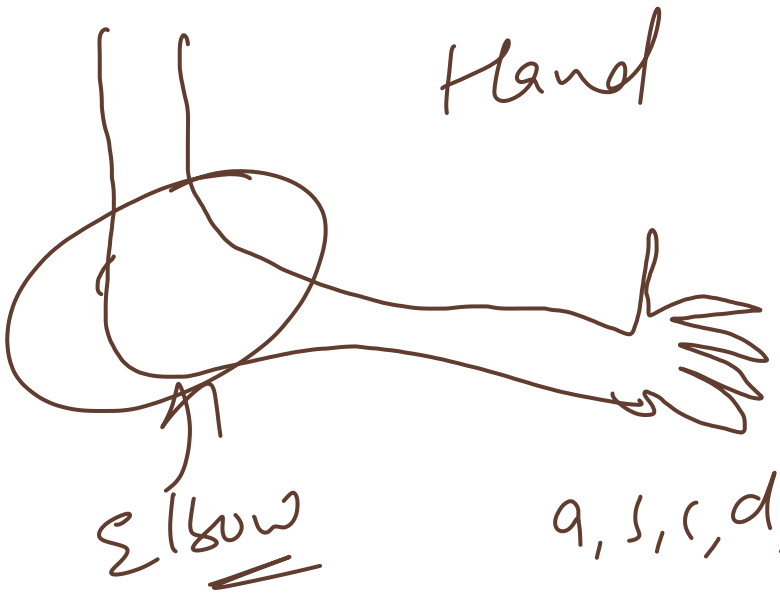
S :- cluster

Step 4 :- Recompute the cluster center and move to Step 3

eg

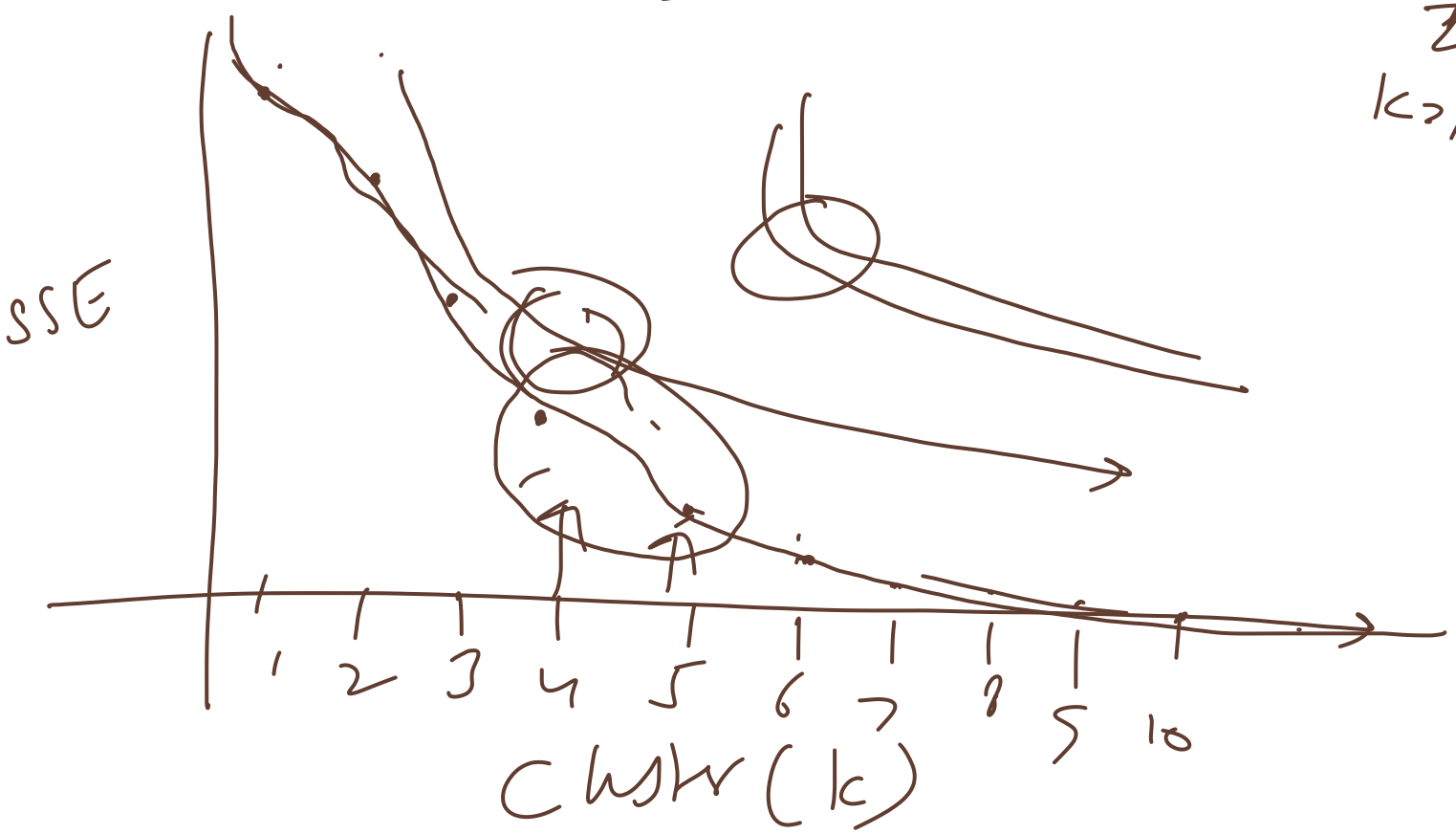


Elbow Method



a, b, c, d, e, f, g, h, i, ...

$\sum_{k=1}^{n}$



$X_2 [X_1, X_2, X_3 \ldots]$

O/P → $S \rightarrow \{S_1, S_2, S_3 \rightarrow S_k\}$

Advantage →

1. fast
2. Robust (easy always converge)
3. Efficient

Disadvantage –

1. explicitly no of clus
2. may convergence poin (local optim) missed Global optimal point