Scalable data ⇒ Data can grow in size, systems or algos can still handle it efficiencientlly without Crashing.

K-Means (Flat Clustering)

- Clustering → organizing objects into groups whose members turn out similar.
  - Also called data segmentation (form of learning by observations rather than by example).
- Approach ⇒ collect data / find clusters / Measure quality of clusters.
- Cluster Analysis ⇒
  - requirements
    - Scalability → need high scalable algorithms which can process high data efficiently.
    - Ability to deal with different type of attributes + noisy data → Algo should perform for binary, categorical or ordinal data well.
    - Identify every shape of cluster → Most Algo tend to find only spherical clusters.
    - High dimensionality → Algo should understand all dimension and then make clusters
    - Clustering results are sensitive to input parameters. Selecting parameters is difficult.

- Issues in clusters →
  - How many clusters → You can fix K before clustering
    - Let number depend on some quality measure
    - Right choice depends on problem you want to solve.

Clustering
- Flat (Find all clusters at once)
  - Reallocate objects to improve clustering.
- Hierarchical (found new cluster using previous)
  - Agglomerative (cluster 1 and merge them)
  - Divisive (all objects in 1 cluster and then split them)

- Hard clustering (Each object belongs to exactly one cluster)
  Simple + common.
- Soft clustering (Objects have fuzzy membership in multiple clusters)

- Overall quality of clustering is measured by f
  f is closely related to measure of distance.
  Primary goals →
  - how inter-cluster similarity
  - High intra cluster similarity
- Avoid very small or large clusters
- Above all focuses on internal criteria. Compare it with external criteria (like hand crafted reference clustering).

- Naive approach & Heuristics → • Trying all possible clusterings is not practical as number of possible clustering grows exponentially with data size.
  - Use heuristic methods that provide a good solution

- Flat clustering (K-means) → • Every cluster is measured from the center.
  - K-medoids or PAM (Partition Around Medoids)
    (cluster representation is from actual data point)

- Number of clusters (K) are defined in advanced.
- Data points represented as unit vectors.

- Centroid of a cluster is defined as

$$\mu(A) = \frac{1}{m} \sum_{i=1}^{m} d_i$$

- RSS (Residual Sum of Squares) of cluster.
  - measures how far points in cluster are from centroid.
  - goal in K-mean is to minimize RSS.

$$RSS(A) = \sum ||d_i - \mu(A)||^2$$

- Quality of clustering & minimizing RSS
  - How well data points are clustered?
  - Measured by summing the RSS of each cluster.
  - K-means minimize total RSS
- K-Mean Algo (Lloyd's Algo)
  - Randomly select K data points as initial clusters centers (seeds or centroids) & create empty clusters.
  - Assign 1 centroid to each cluster.
  - Iterate over each data points & assign each data point to cluster
  - Check clustering good if not start from create empty clusters.
- When to decide K-means cluster is good?
  - Small change since previous iteration
  - Max no of iterations reached.
  - Set threshold for RSS.
- K-means is really efficient (very fast) & often terminates at local optimum.

$O(nkt)$ — clusters, iterations, objects

$K, t < n$.

Similar Approaches → ① K-mediods (use median)
② Fuzzy cmean (soft clustering)
③ Model Based Clustering

- Disadvantages →
  - cluster number are defined before.
  - only work on those whose mean is calculated.
  - not suitable for categorical data.
  - do not handle noisy data or outliers efficiately.
  - ~~do not identify non-convex clusters~~ (To ~~seed~~ this line say separate nhi kr skte).