Clustering-

Similarity within group, non similarity between groups

Depends on desired results and data, can have 4, 6 2 clusters

Types-

1. Hierarchical vs partitioned

   Nested vs non nested

   Finally, note that a
   hierarchical clustering can be viewed as a sequence of partitional clustering and a partitional clustering can be obtained by taking any member of that sequence; i.e., by cutting the hierarchical tree at a particular level.

2. Exclusive vs overlap vs fuzzy

   Exclusive- one point, one cluster

   Overlap- one point,simultaneously to more than one clusters

   Fuzzy- all points to all clusters with membership degree, 0-1

   Since in fuzzy sum of the weights/probability of a data point should be one for all clusters belonging, it often missed true multi class assignment, but its used to reduce the assignment to single clusters, however in practice point is assigned to the cluster with highest probability.

3. Completed vs Partital

   Complete- each object is assigned to a cluster
   Partial- noise/outliers are not assigned to any cluster'

Different type of clusters-

1. Well-separated , natural structure, threshold how close similar points should be, any shape
2. Prototype based- data points in a cluster are more cluster to centroid than other centroid. Mostly globular
3. Graph based- contiguity based, tends to be globular
4. Density based- higher density, when the data is irregular and intertwined

Conceptual clustering, for more different type of clusters

K-Means

Prototype based, partitioned clustering

Always converge to a solution

Distance measure
L1, L1 euclidean distance
Cosine, jaccard similarity

How to choose K

Randomly- poor initialization, sub optimal clusters

Multiple runs of random- if a cluster doesn't have one centroid initial point, then cluster might get split, and only local minimum can be found

Hierarchical clustering, get K clusters and use there centroids, but expensive, works well with less data only

Get first centroid as random, rest of them far away from existing ones

K- means additional issues-

Empty cluster-

No point is assigned to clusters, need to choose another centroid which is far from existing ones

Eliminate the centroid with highest SSE