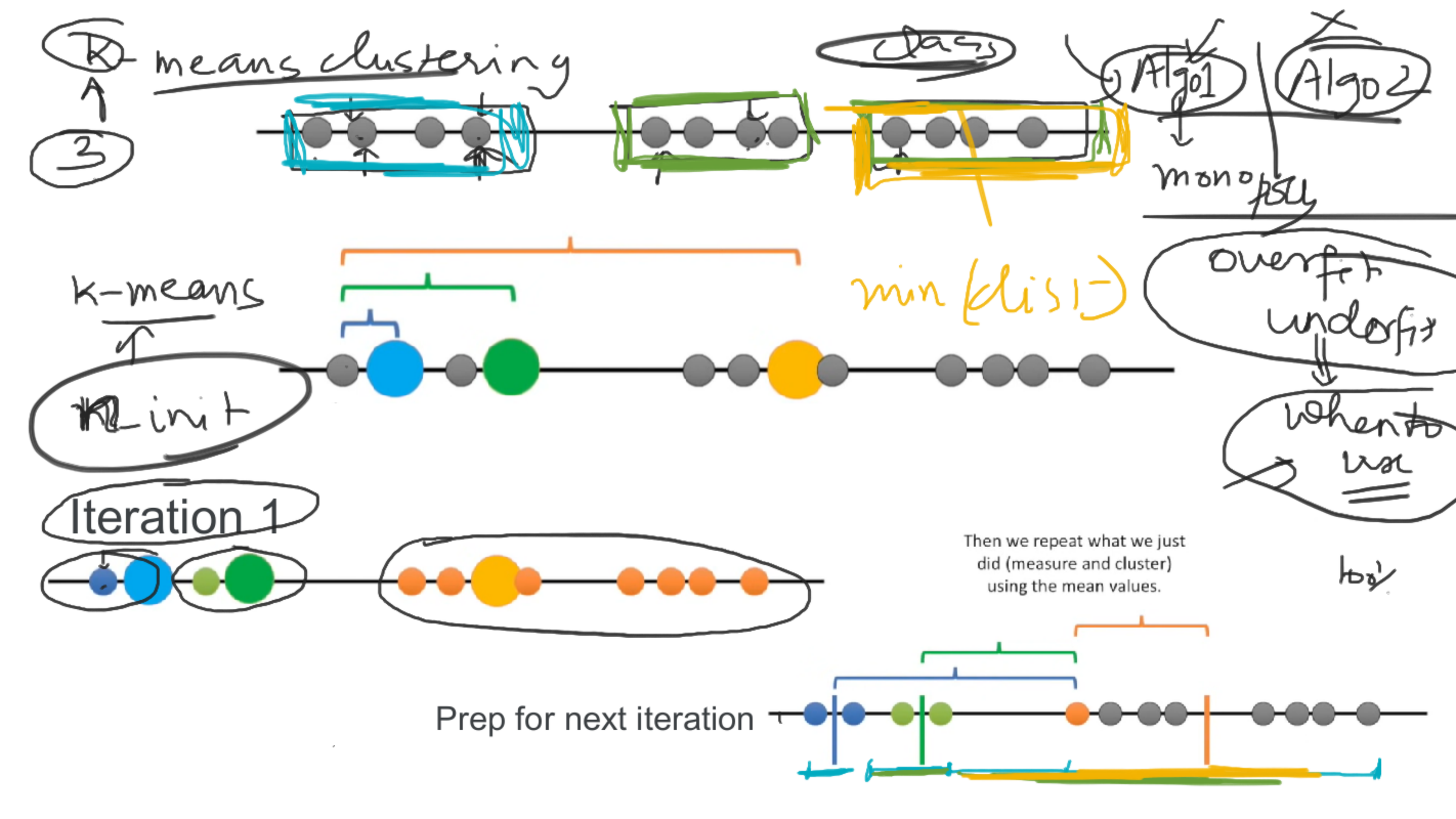
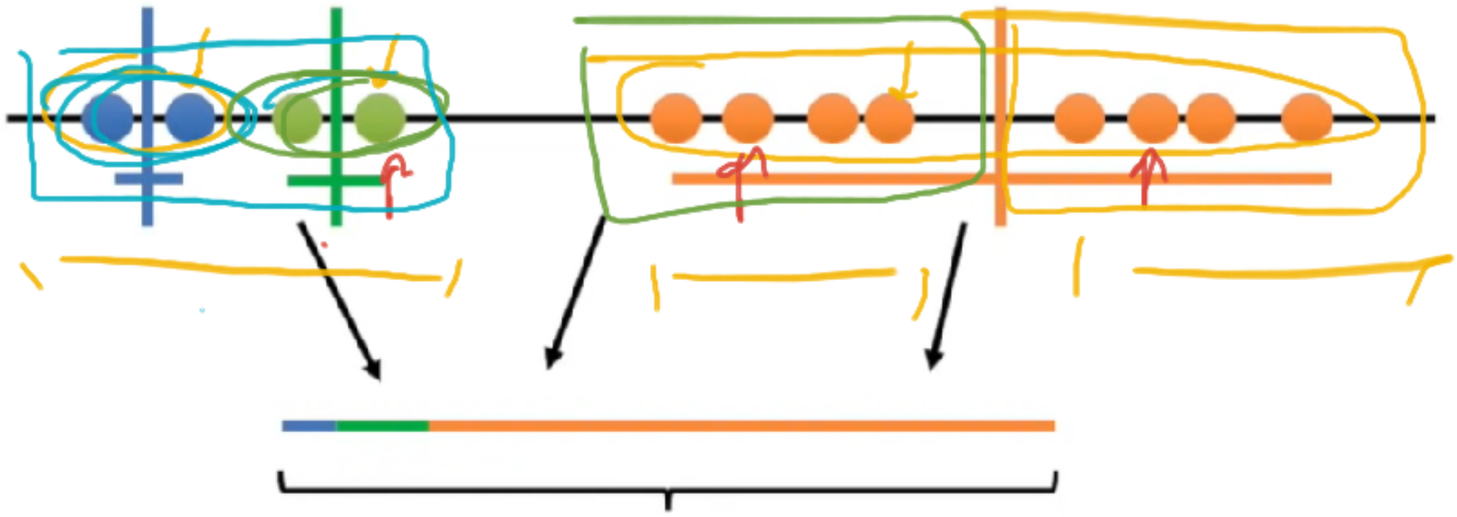

K
means clustering
3
Algo1
Algo2
monopoly
k-means
init
min dist
overfit
underfit
when to use
Iteration 1
Then we repeat what we just did (measure and cluster) using the mean values.
Prep for next iteration

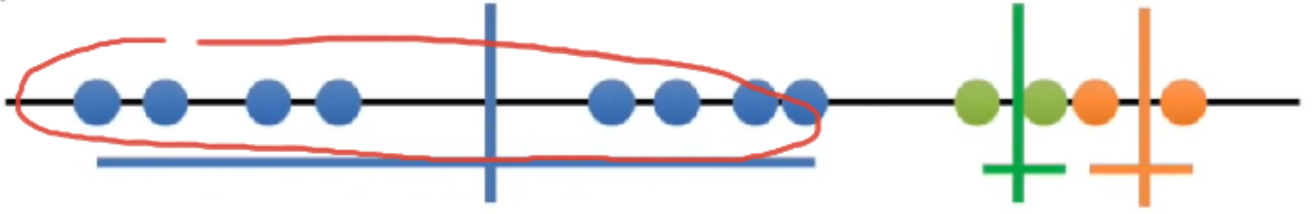
We can assess the quality of the clustering by adding up the variation within each cluster.

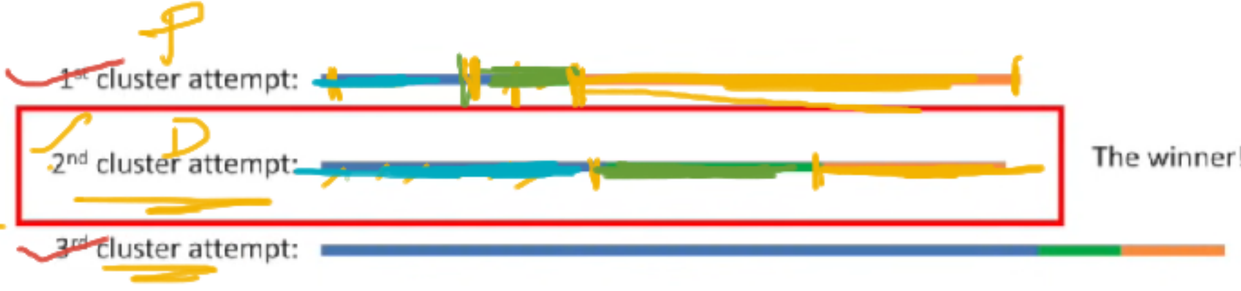


Total variation within the clusters

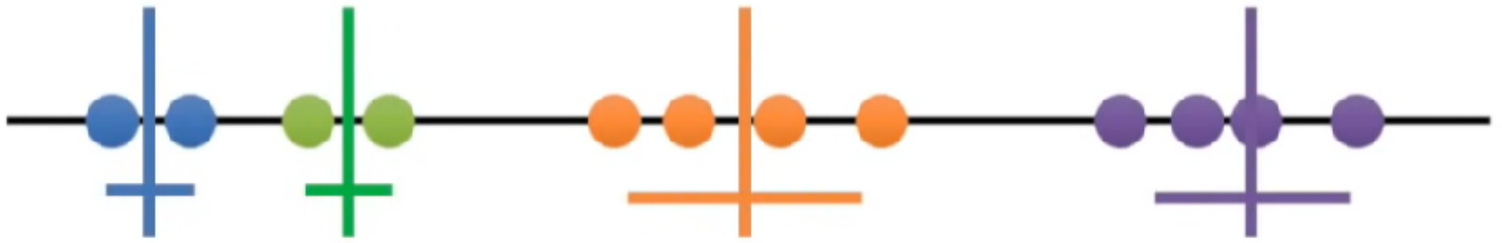
Since K-means clustering can't "see" the best clustering, its only option is to keep track of these clusters, and their total variance, and do the whole thing over again with different starting points.

At this point, K-means clustering knows that *the 2nd clustering is the best clustering so far*. But it doesn't know if it's *the best overall*, so it will do a few more clusters (it does as many as you tell it to do) and then come back and return that one if it is still the best.





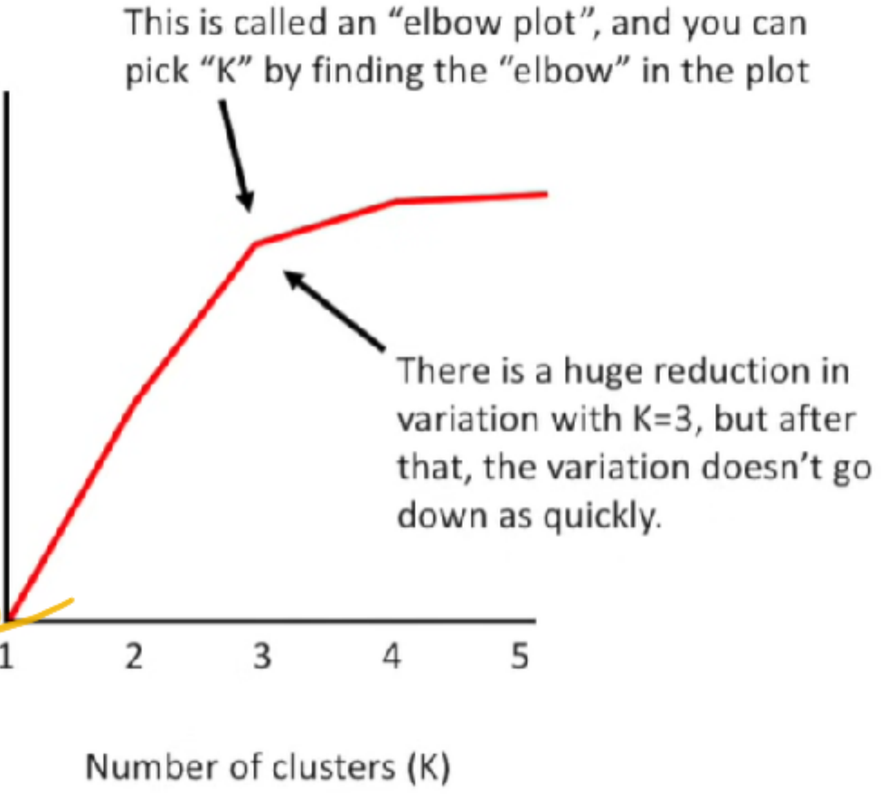
How to get the best k



The total variation within each cluster is less than when K=3

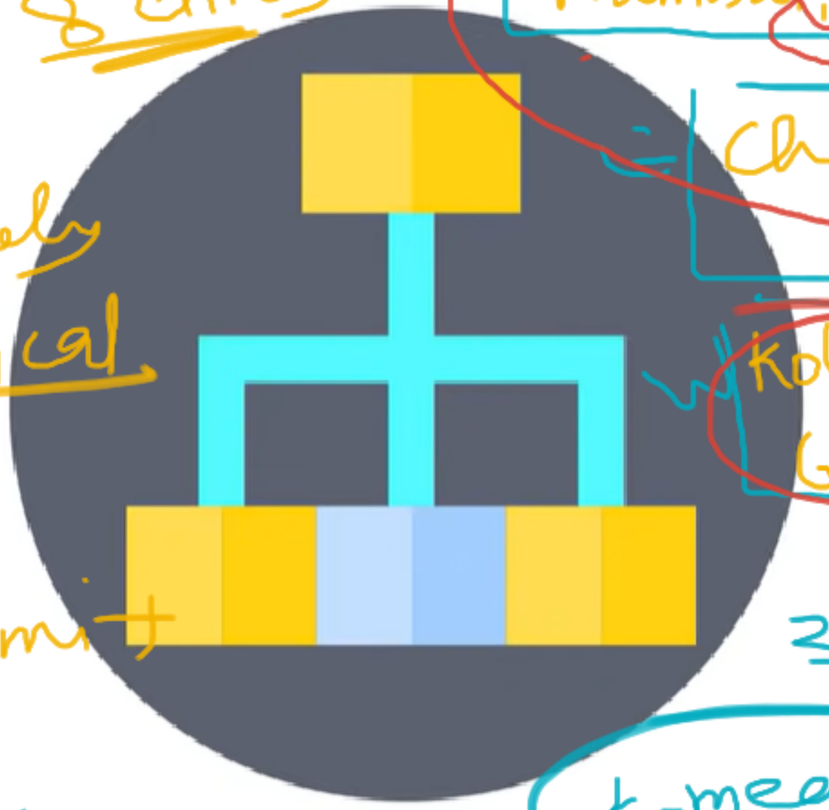
Each time we add a new cluster, the total variation within each cluster is smaller than before. And when there is only one point per cluster, the variation = 0.

However, if we plot the reduction in variance per value for K...



What's in it for you?

- What is Clustering?
- What is Hierarchical Clustering?
- How Hierarchical Clustering works?
- Distance Measure
- What is Agglomerative Clustering?
- What is Divisive Clustering?

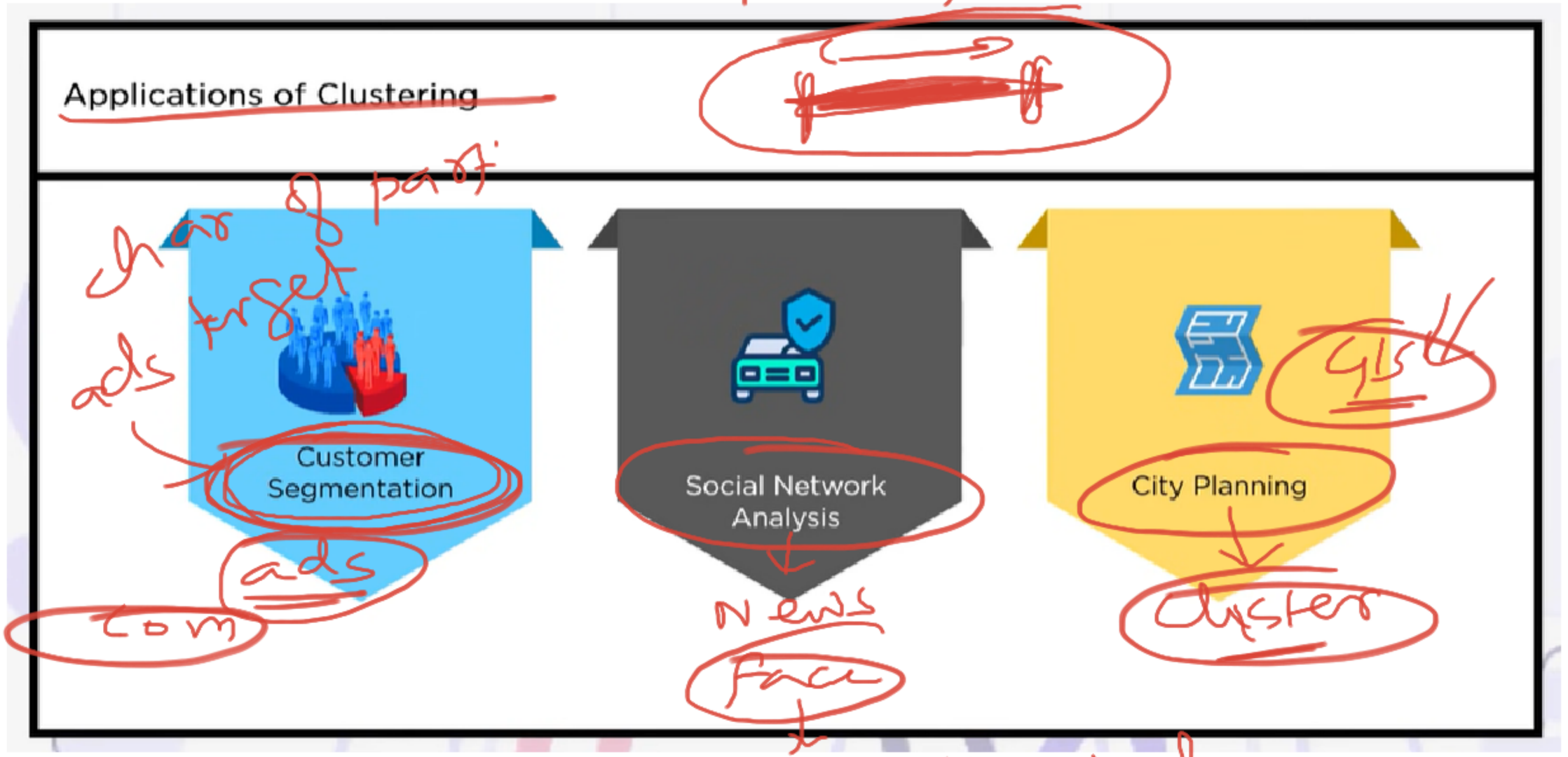


What is Clustering?

The method of dividing the objects into clusters which are similar between them and are dissimilar to the objects belonging to another cluster

What is Hierarchical Clustering?

Let's consider that we have a set of cars and we have to group similar one's together

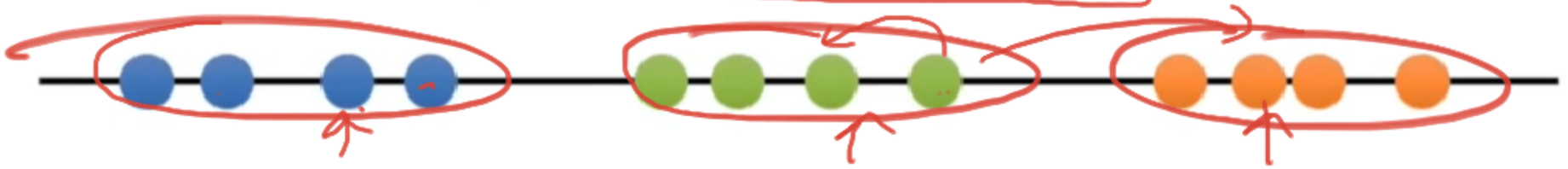

Applications of Clustering
Customer Segmentation
Social Network Analysis
City Planning

sparse → n-dimensional → EDA

Question: How is K-means clustering different from hierarchical clustering?

K-means clustering specifically tries to put the data into the number of clusters you tell it to.

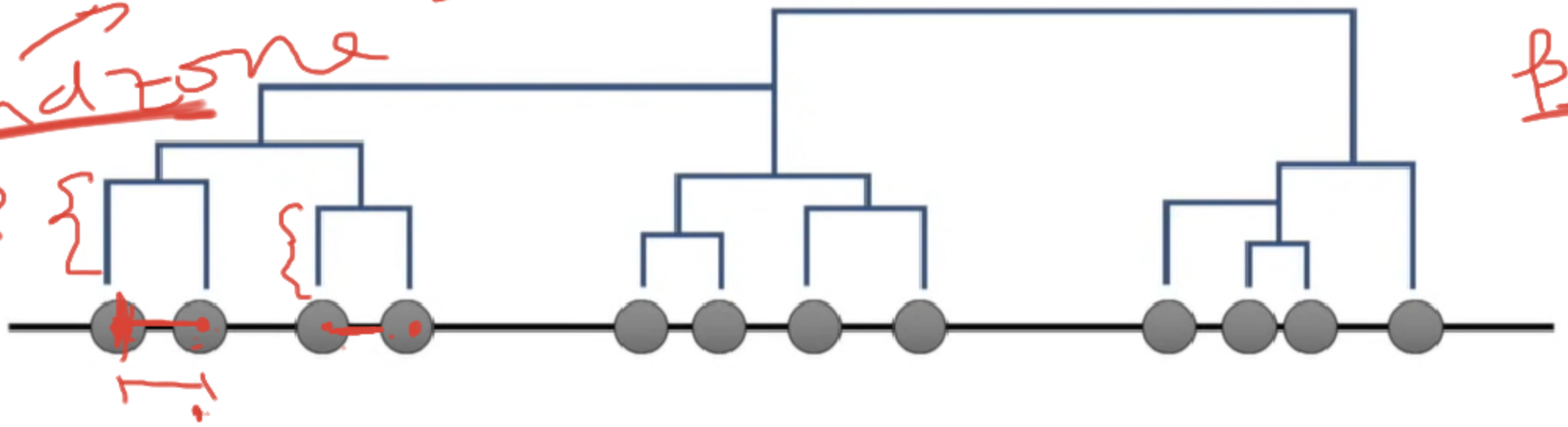
3 - Force fit



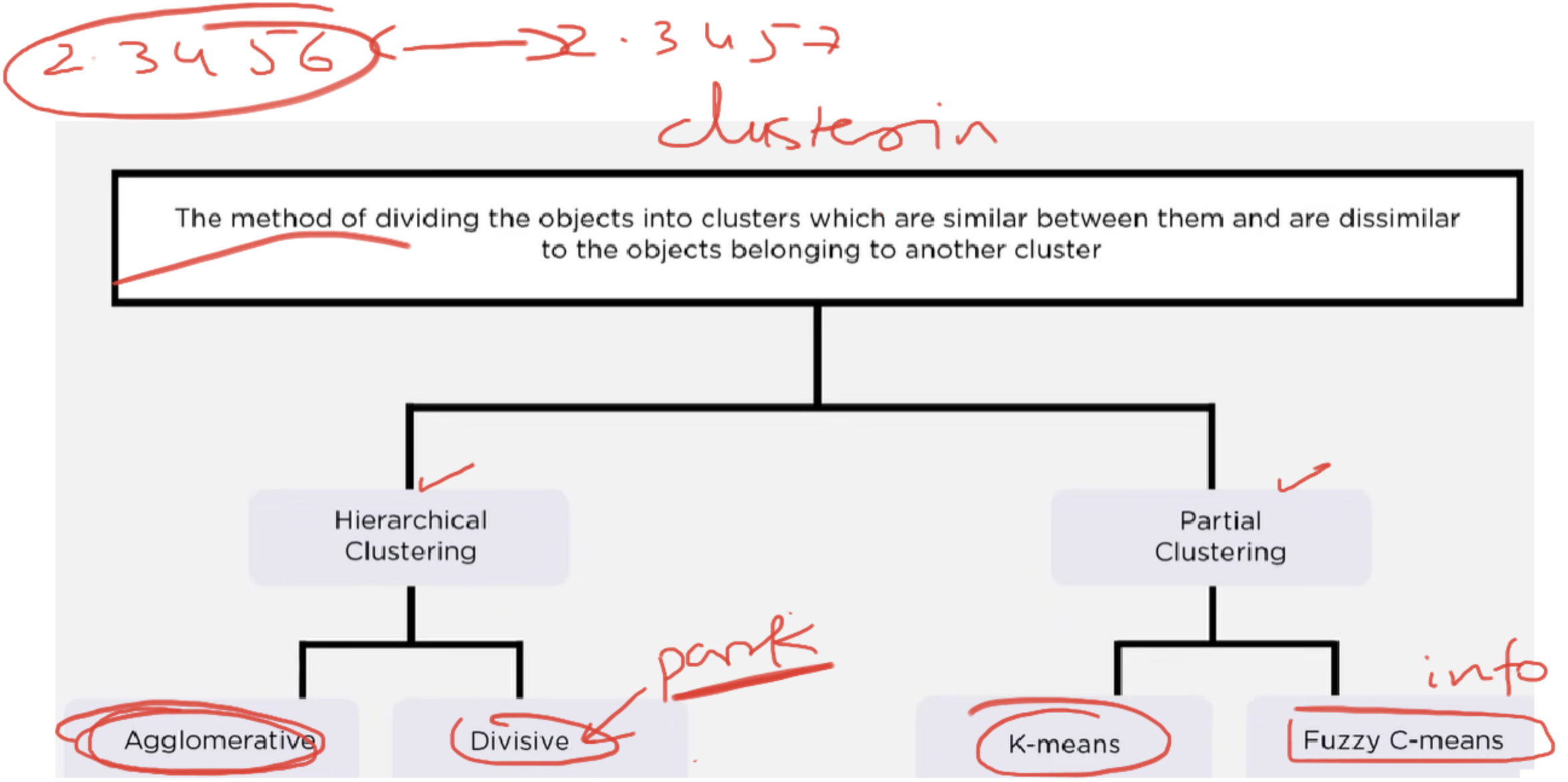
Hot

Hierarchical clustering just tells you, pairwise, what two things are most similar.

Friend zone algo

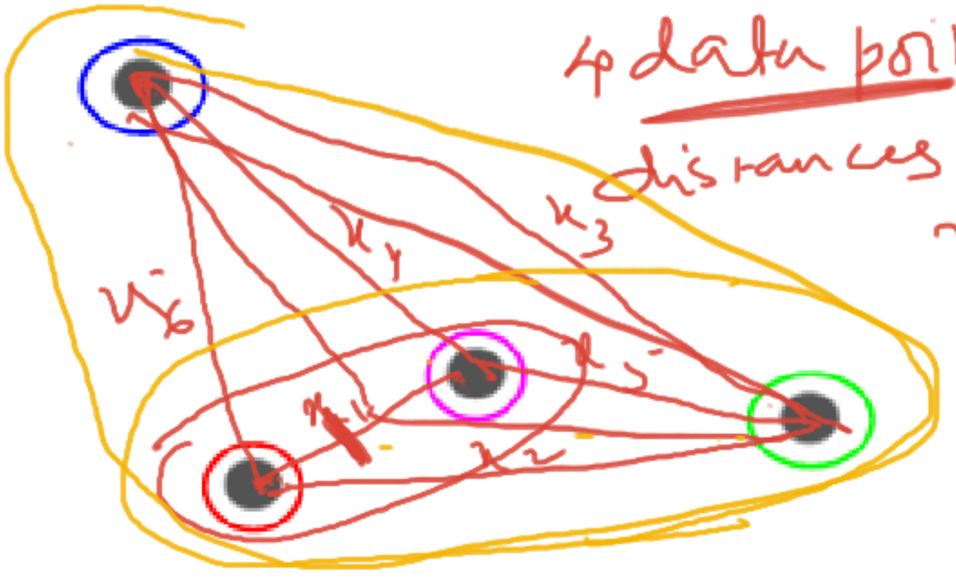


Pair friend


The method of dividing the objects into clusters which are similar between them and are dissimilar to the objects belonging to another cluster
Hierarchical Clustering
Partial Clustering
Agglomerative
Divisive
K-means
Fuzzy C-means

Types of Hierarchical Clustering

Agglomerative Clustering is known as bottom-up approach



How does hierarchical clustering work?

Video

How to measure distance?

Euclidean distance measure

Manhattan distance measure

Distance measure will determine the similarity between two elements and it will influence the shape of the clusters

Squared Euclidean distance measure

Cosine distance measure

01 **Euclidean distance measure**

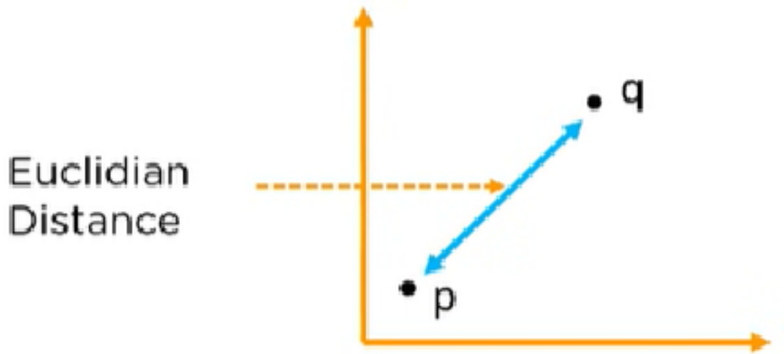
02 Squared euclidean distance measure

03 Manhattan distance measure

04 Cosine distance measure

- The Euclidean distance is the "ordinary" straight line
- It is the distance between two points in Euclidean space

$$d=\sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}$$



01 Euclidean distance measure

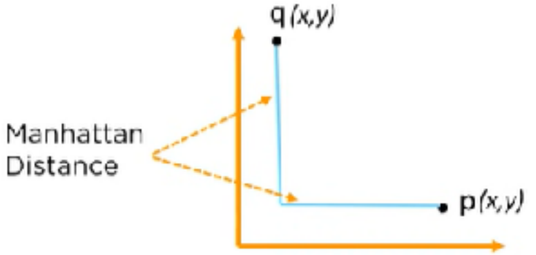
02 Squared euclidean distance measure

03 Manhattan distance measure

04 Cosine distance measure

The Manhattan distance is the simple sum of the horizontal and vertical components or the distance between two points measured along axes at right angles

$$\mathbf{d} = \sum_{i=1}^{n} | q_x - p_x | + | q_y - p_y |$$



01 Euclidean distance measure

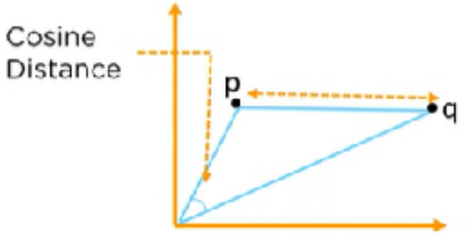
02 Squared euclidean distance measure

03 Manhattan distance measure

04 Cosine distance measure

The cosine distance similarity measures the angle between the two vectors

$$\mathbf{d} = \frac{\sum_{i=0}^{n-1} q_i - p_x}{\sum_{i=0}^{n-1} (q_i)^2 \times \sum_{i=0}^{n-1} (p_i)^2}$$



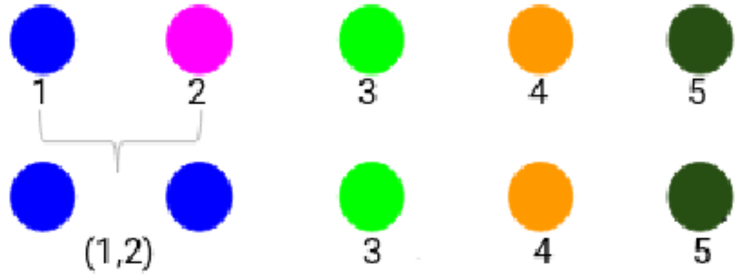
Student_ID	Marks
1	10
2	7
3	28
4	20
5	35

complete

min

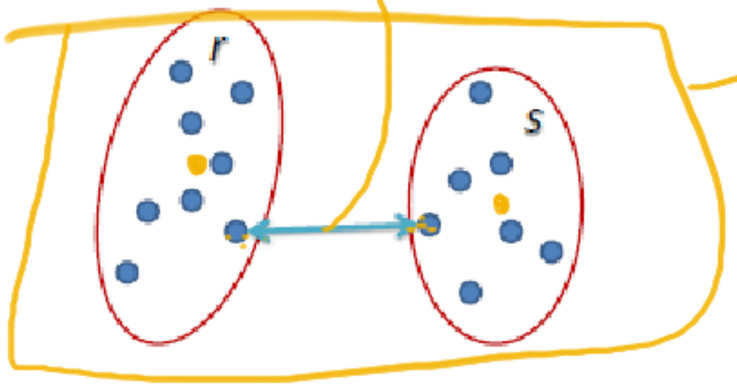
ID	1	2	3	4	5
1	0	3	18	10	25
2	3	0	21	13	28
3	18	21	0	8	7
4	10	13	8	0	15
5	25	28	7	15	0

Student_ID	Marks
(1,2)	10
3	28
4	20
5	35



Student_ID	Marks
(1,2)	10
3	28
4	20
5	35

ID	(1,2)	3	4	5
(1,2)	0	18	10	25
3	18	0	8	7
4	10	8	0	15
5	25	7	15	0

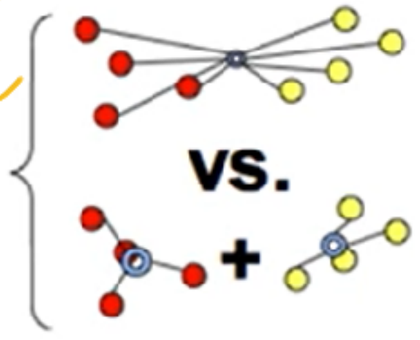


$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

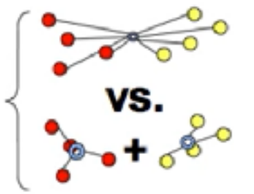
Technique to measure the distance between clusters

- Single
- Complete
- Average
- Centroid
- Ward

- **Ward's method:** $TD_{c_1 \cup c_2} = \sum_{x \in c_1 \cup c_2} D(x, \mu_{c_1 \cup c_2})^2$
  - consider joining two clusters, how does it change the total distance (TD) from centroids?



- Ward's method: $TD_{c_1 \cup c_2} = \sum_{x \in c_1 \cup c_2} D(x, \mu_{c_1 \cup c_2})^2$
  - consider joining two clusters, how does it change the total distance (TD) from centroids?

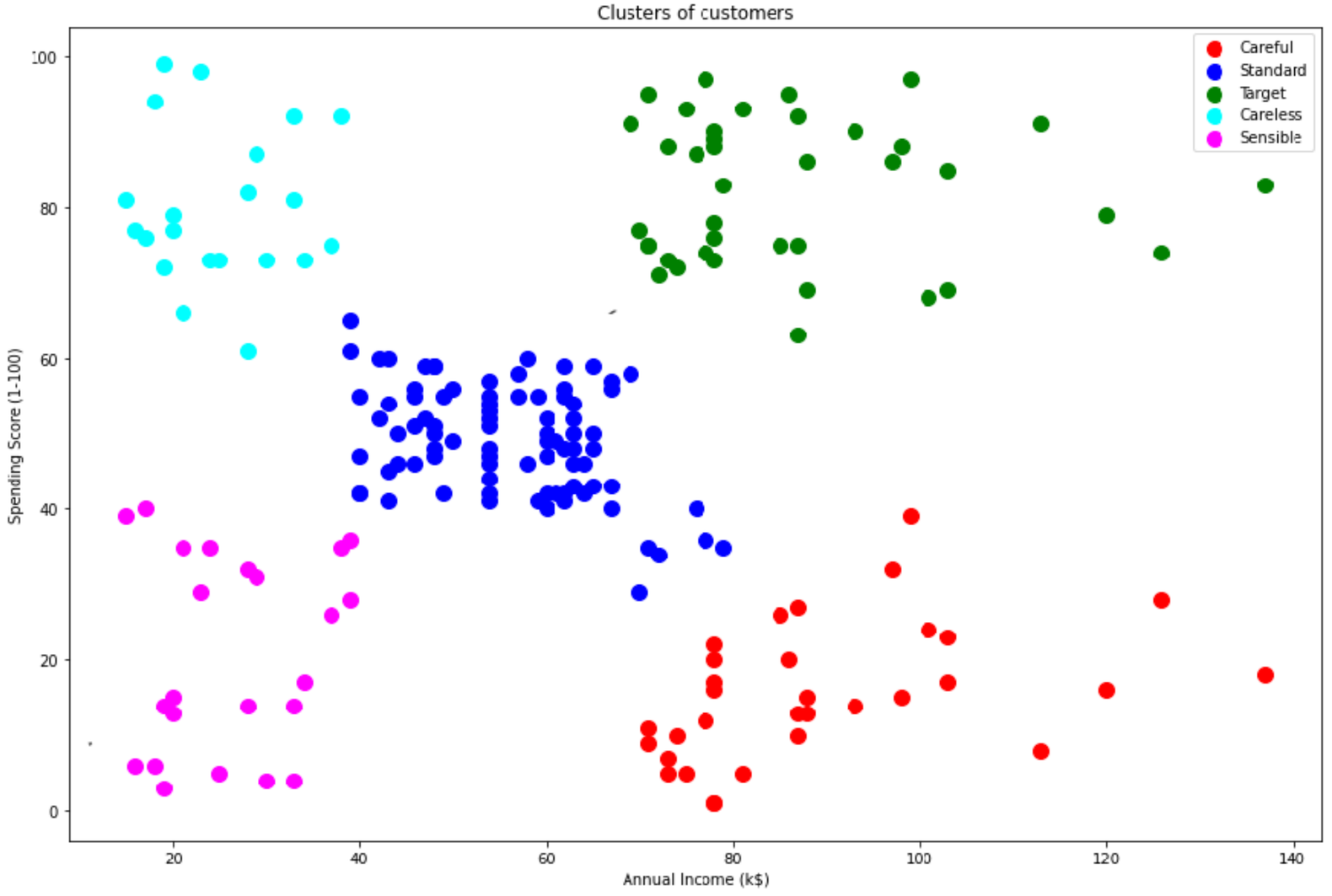


3 Steps of Agglomerative clustering

How do we represent a cluster of more than one point?

How do we determine the nearness of clusters?

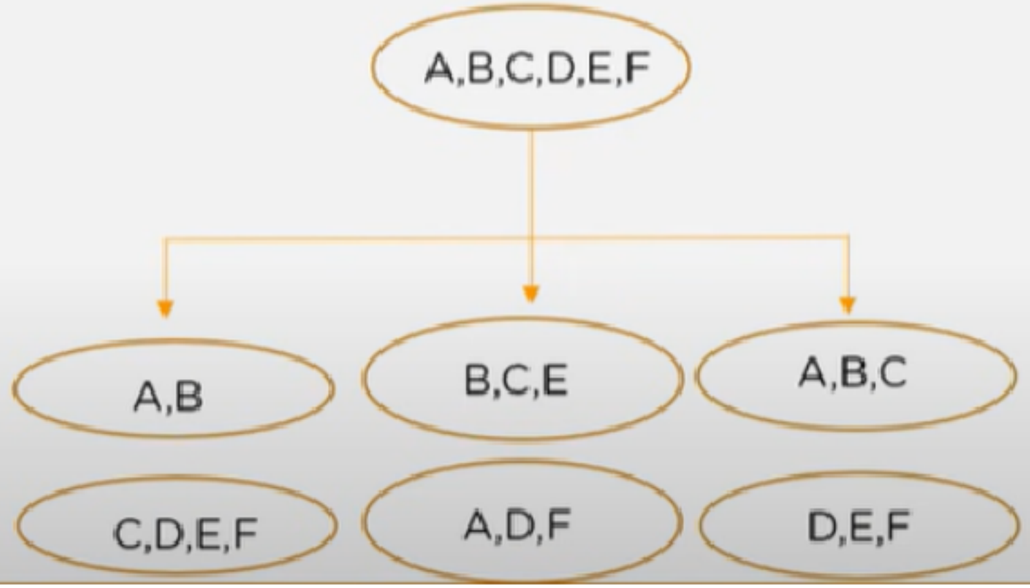
When to stop combining clusters?

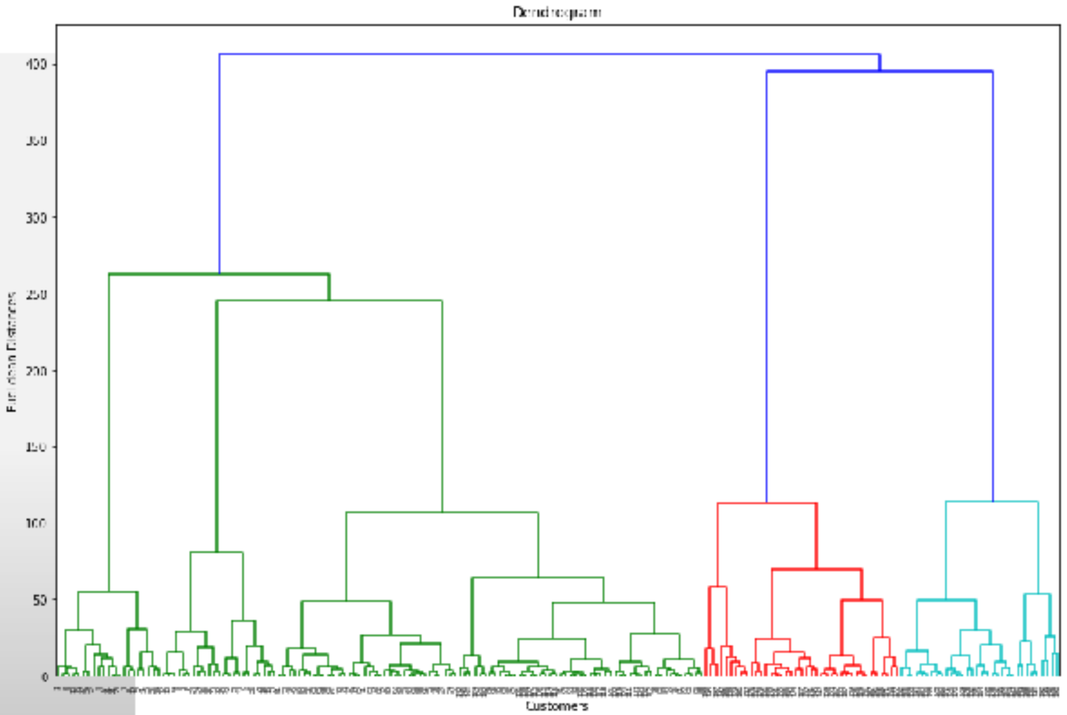


Divisive Clustering

Info

- Obtain all possible splits into two clusters





- For each split, compute cluster sum of squares

$$B_{j12} = n_1(\bar{x}_{j1} - \bar{x}_j)^2 + n_2(\bar{x}_{j2} - \bar{x}_j)^2$$

B_j	Between cluster 1 and 2	X_{j1}	Mean of the cluster
n	Number of members in cluster	X_j	Grand mean

We select the cluster with the largest sum of squares