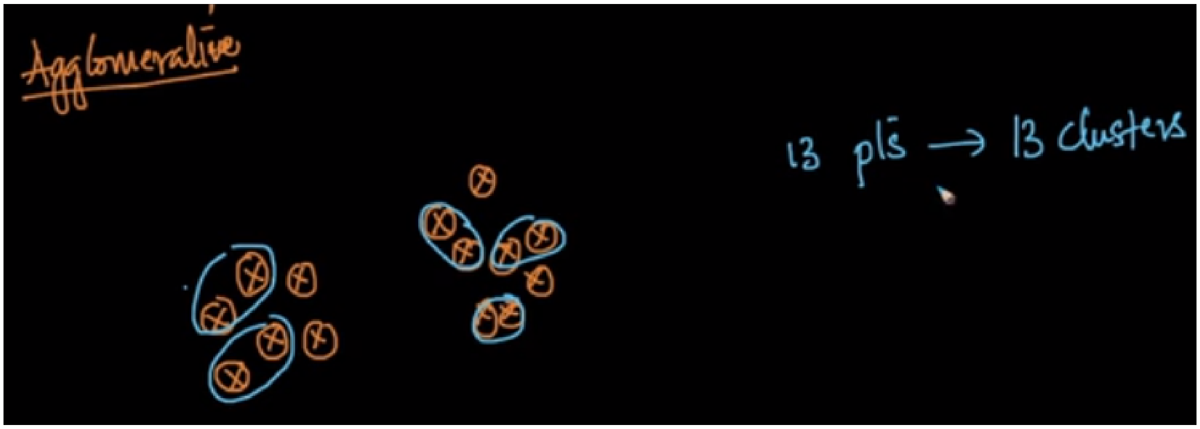
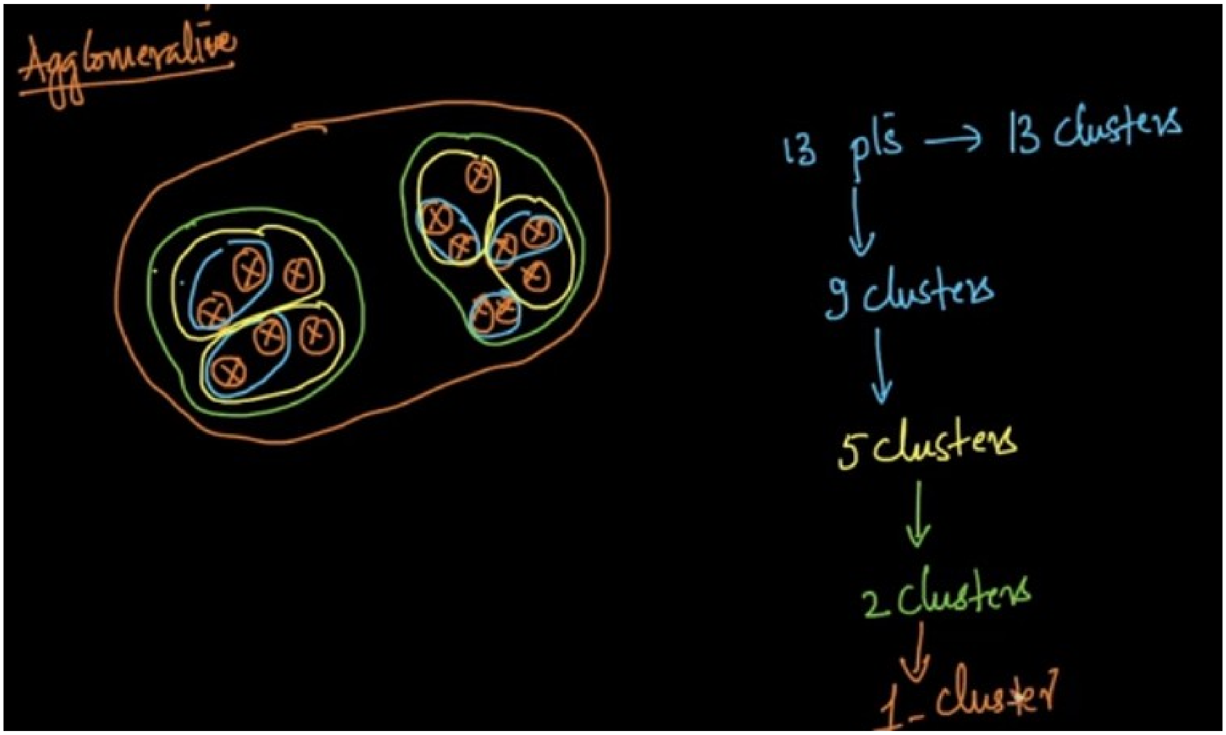
Agglomerative and divisieve, Dendograms:

Agglomerative(more popular):

It initially assumes that each point is a cluster itself. Then it takes the two clusters that are at the less distance and groups it together and make the other cluster.

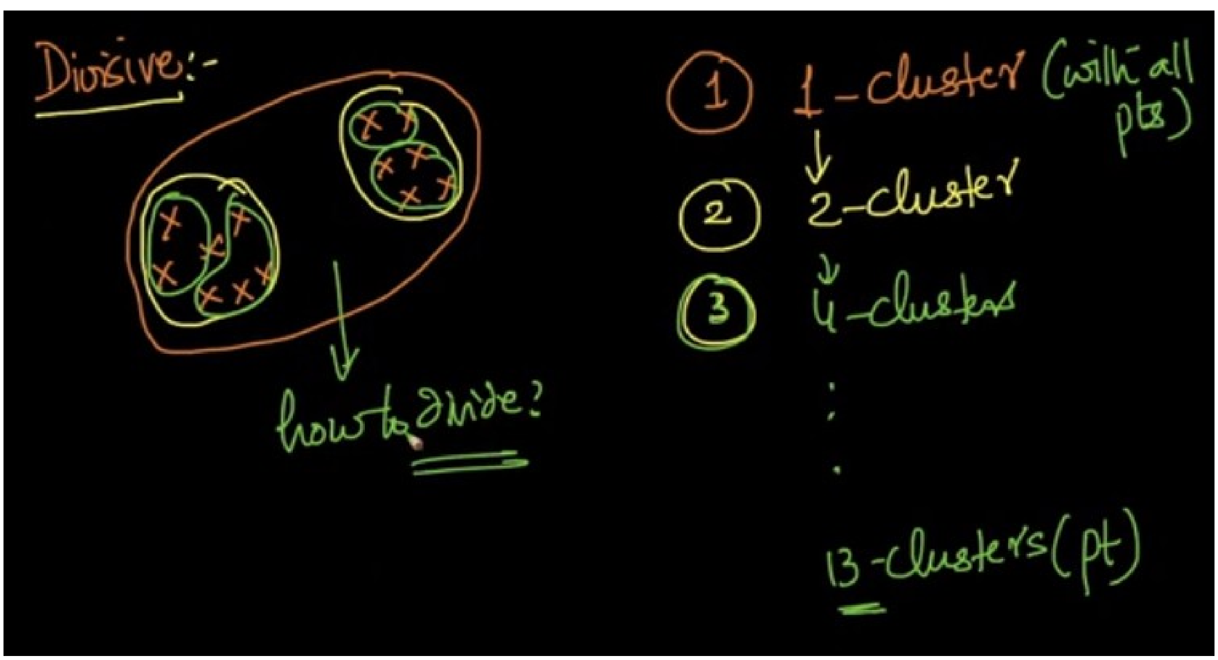


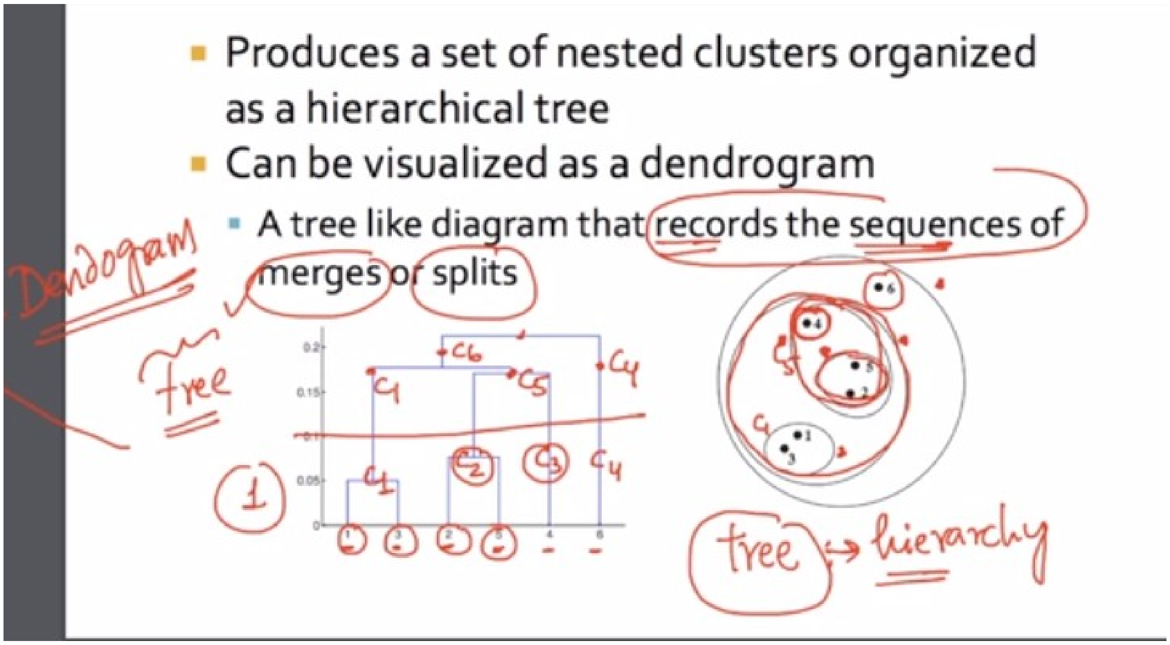
And so on…



In agglomerative clustering we group with some similarity (or) distance.

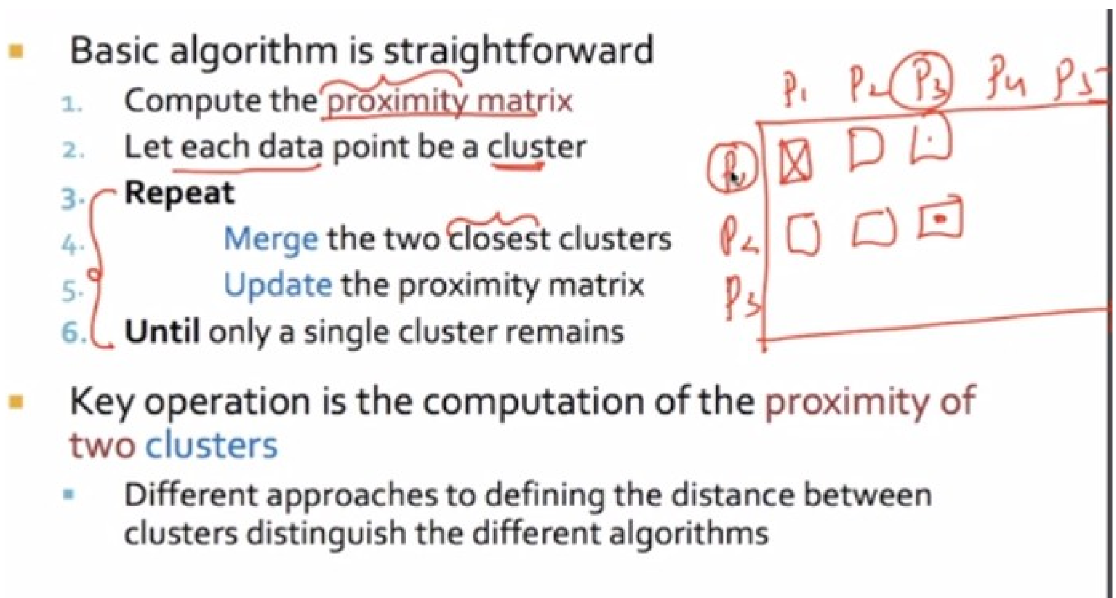
Divisive:
Works opposite to the agglomerative clustering.
How to divide?



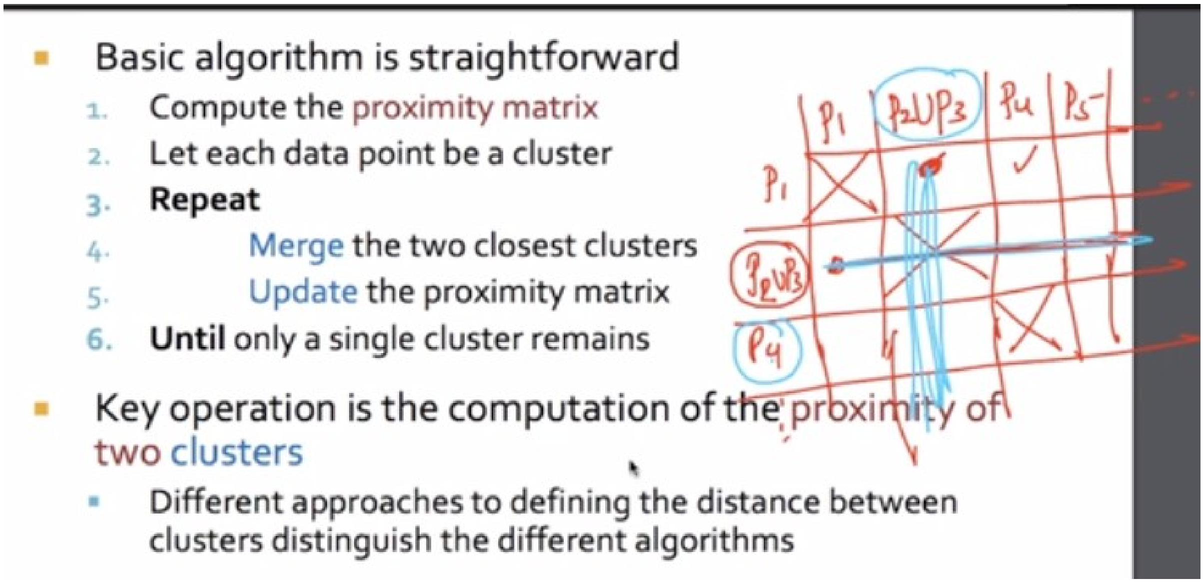


The tree gives the sequence of merges. **We never mention the number of clusters.**
We never give the number of clusters as the hyper parameter. The cluster is chosen as needed.
We can go deeper and deeper of clustering.
Agglomerative clustering:

The proximity matrix or distance between the points is the measure for clustering.

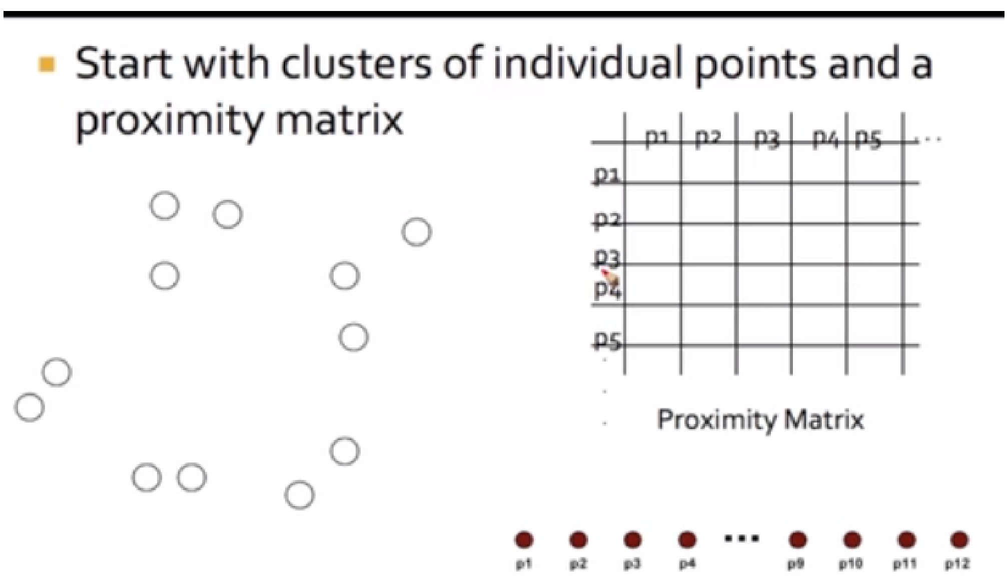


After each step the proximity matrix is updated.

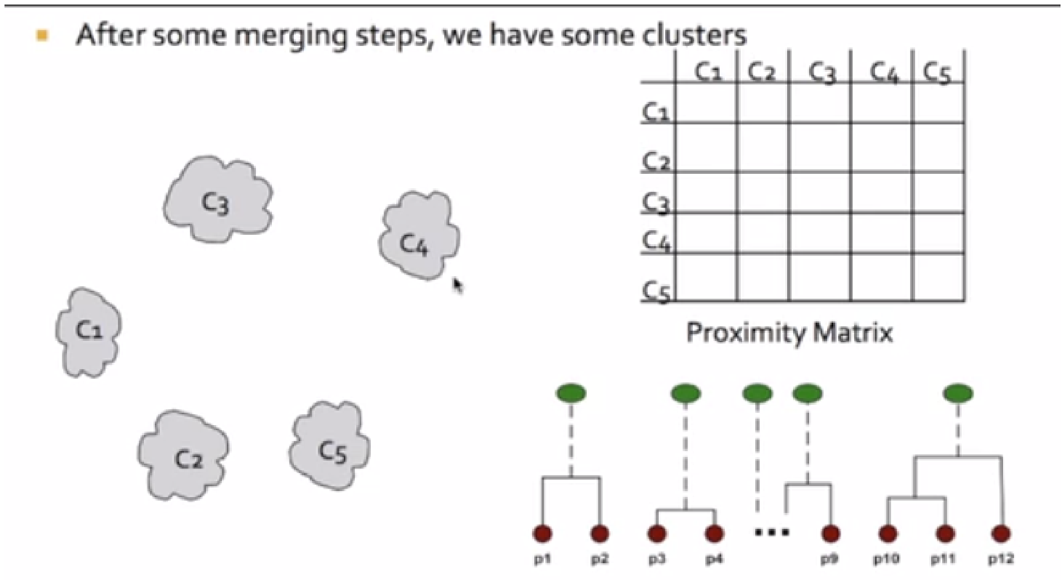


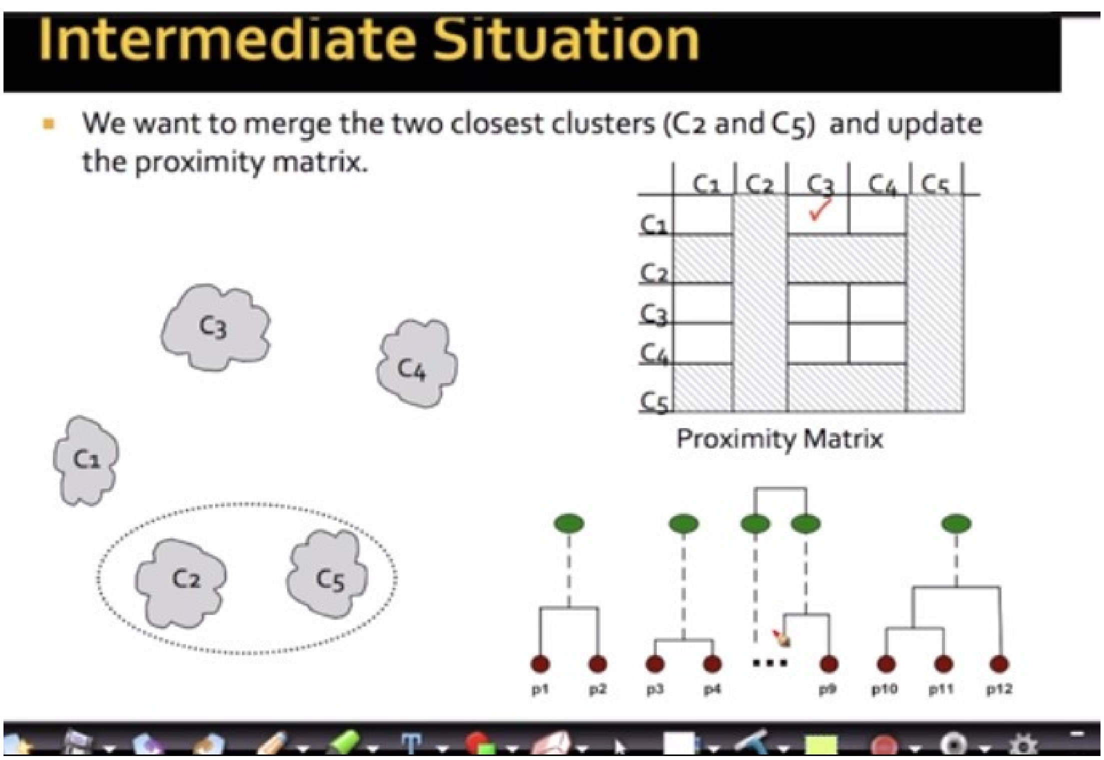
That is the update state. Then we update the rule until the single cluster remains.

How to compute the sim (or) distance between 2 clusters.



In every iteration we are merge the points.

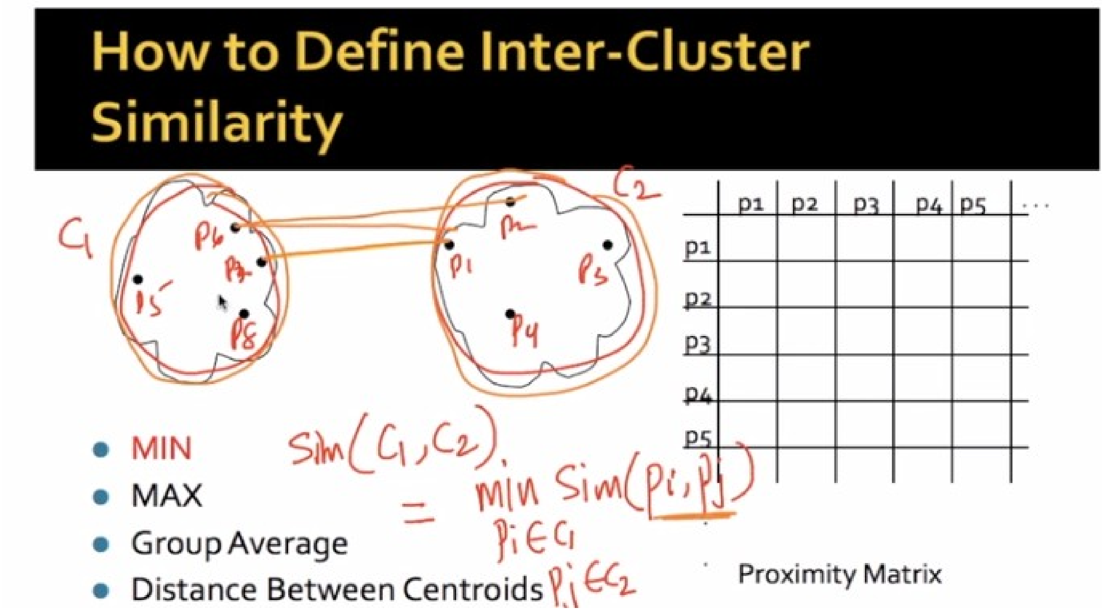




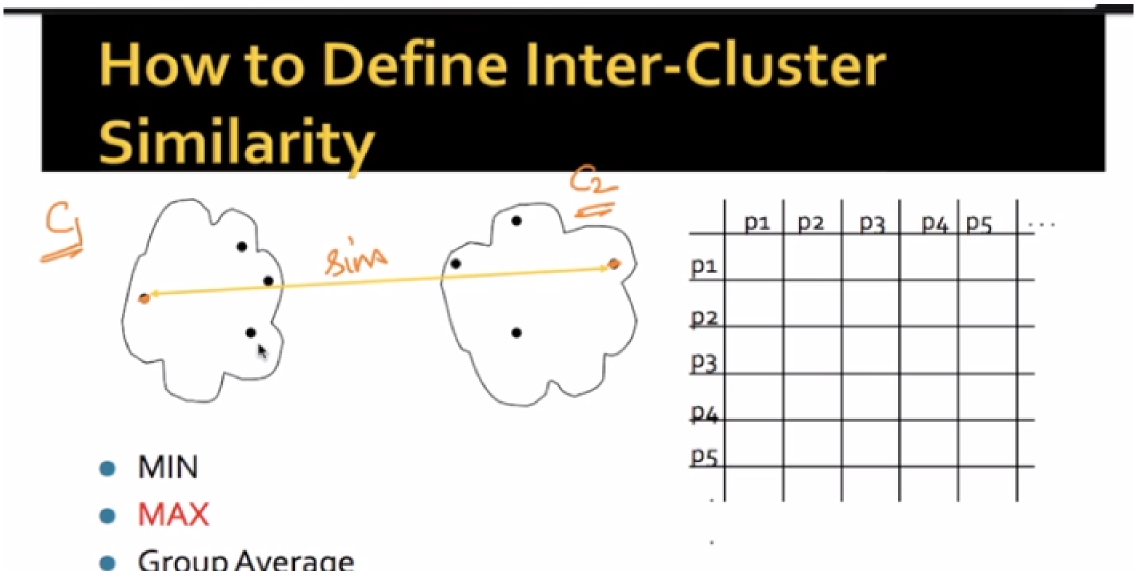
How to define the inter – cluster similarity:

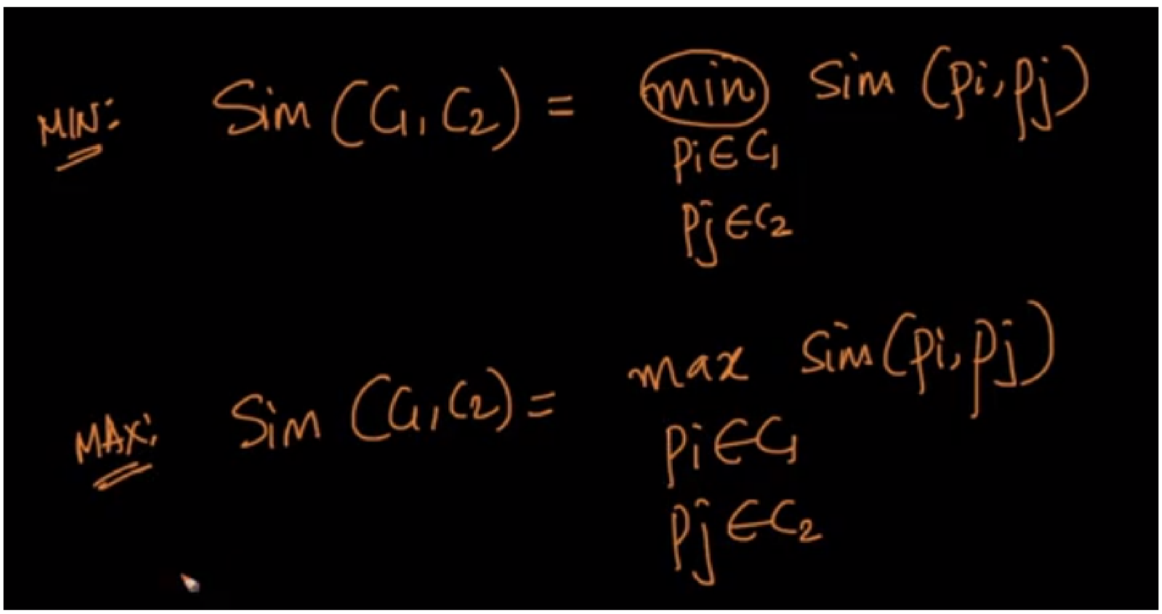
Proximity methods: Advantages and Limitations.
MIN approach:

$$Sim(C_1, C_2) = \min_{\substack{p_i \in C_1 \\ p_j \in C_2}} sim(p_i, p_j)$$

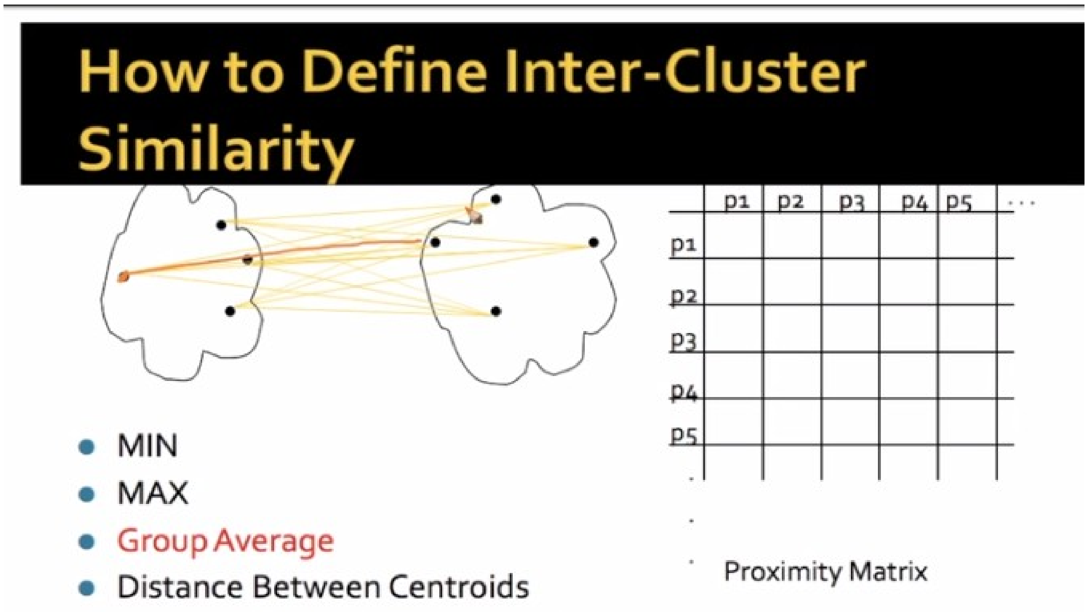


MAX approach:

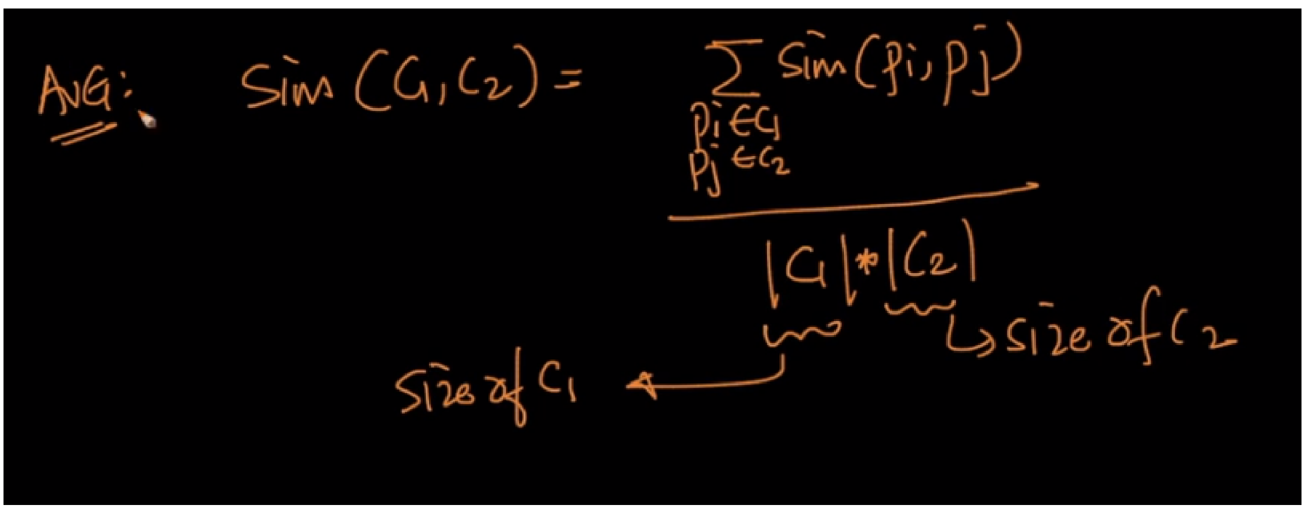




Group average:



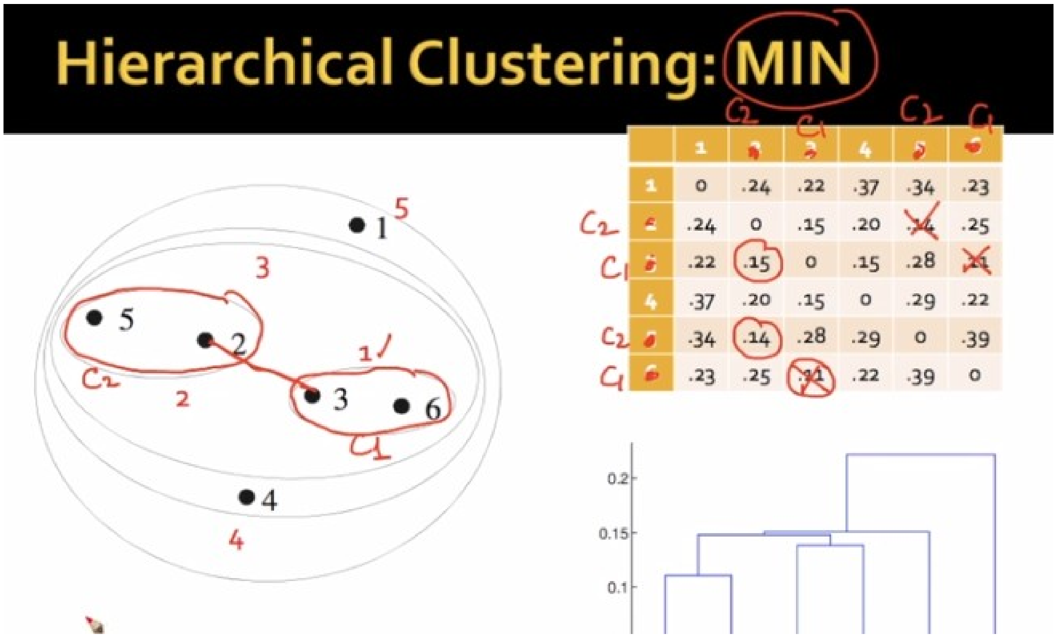
Take the average of the similarity values

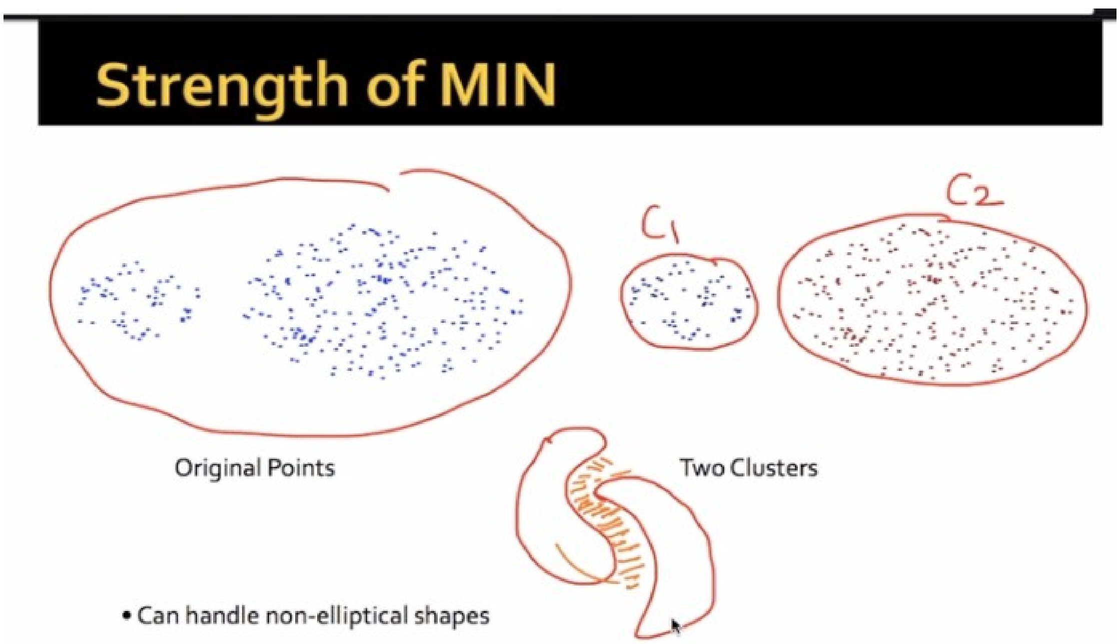


Similarly is the cosine similarity.

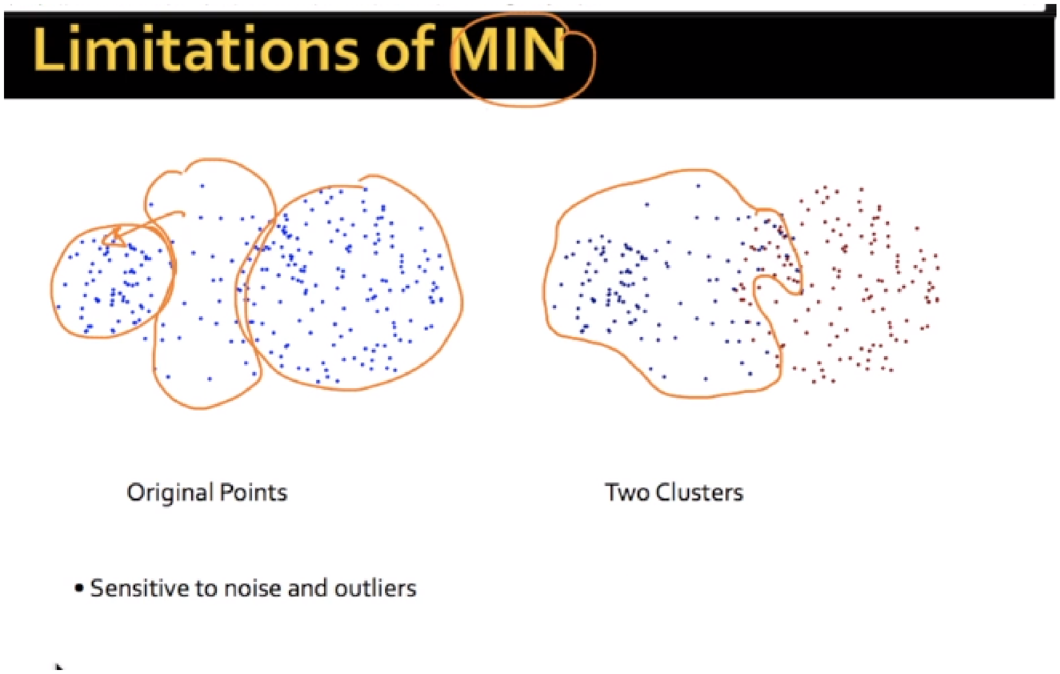
All the methods can be kernelized, except the distance between the centroids.

Hierarchical Clustering: MIN



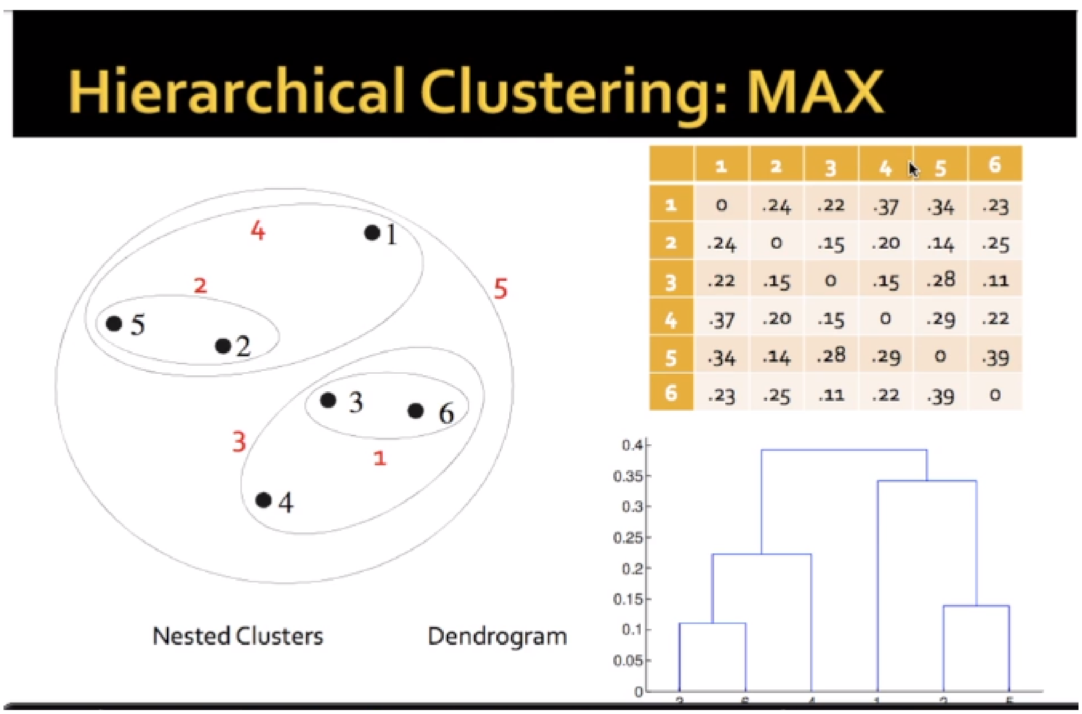


Limitations:



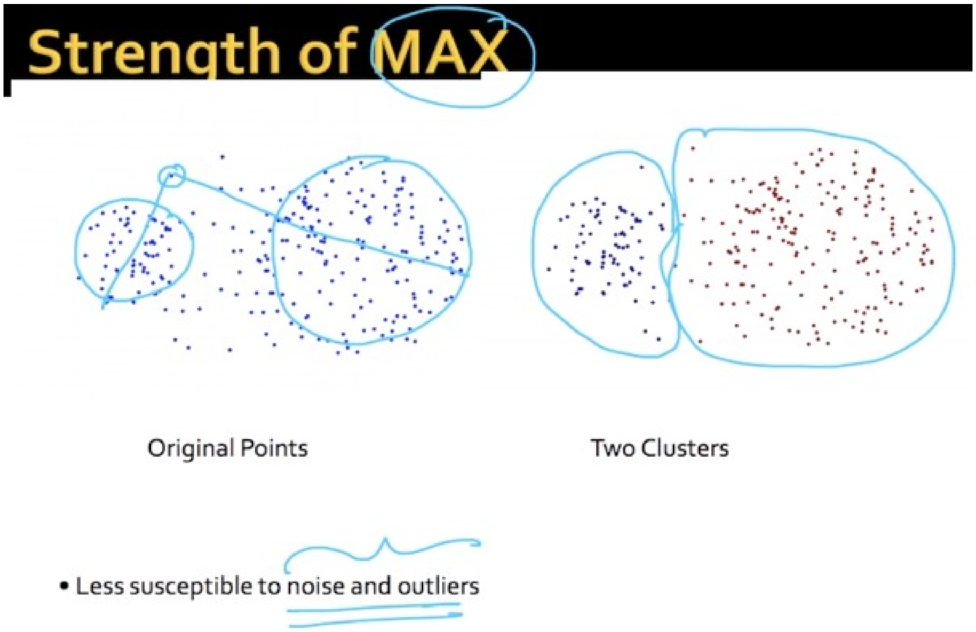
It is extremely sensitive to the noise.

Max:

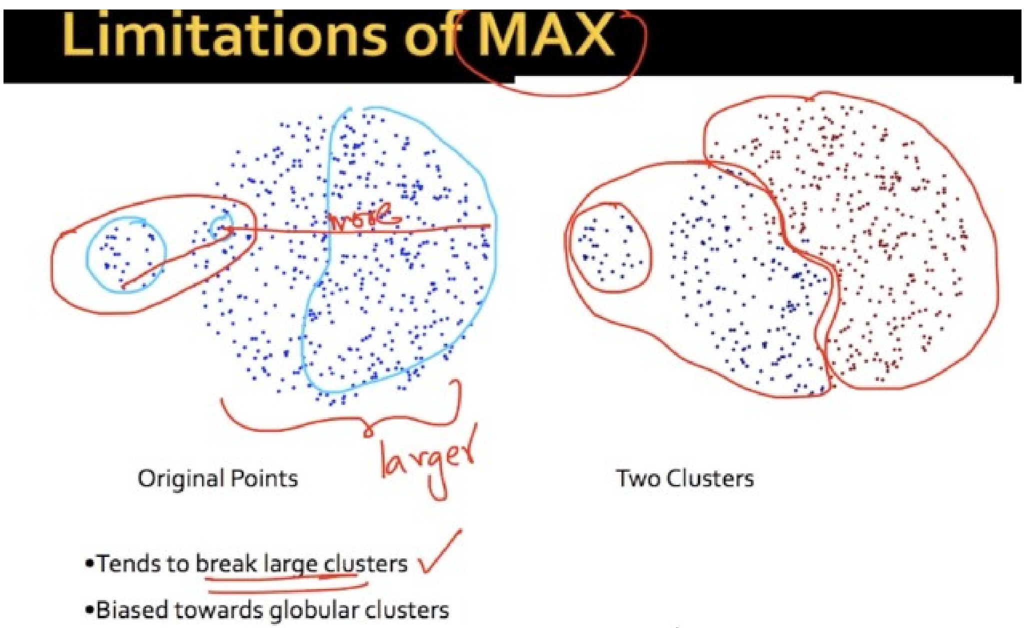


	1	2	3	4	5	6
1	0	.24	.22	.37	.34	.23
2	.24	0	.15	.20	.14	.25
3	.22	.15	0	.15	.28	.11
4	.37	.20	.15	0	.29	.22
5	.34	.14	.28	.29	0	.39
6	.23	.25	.11	.22	.39	0

Strength of MAX:



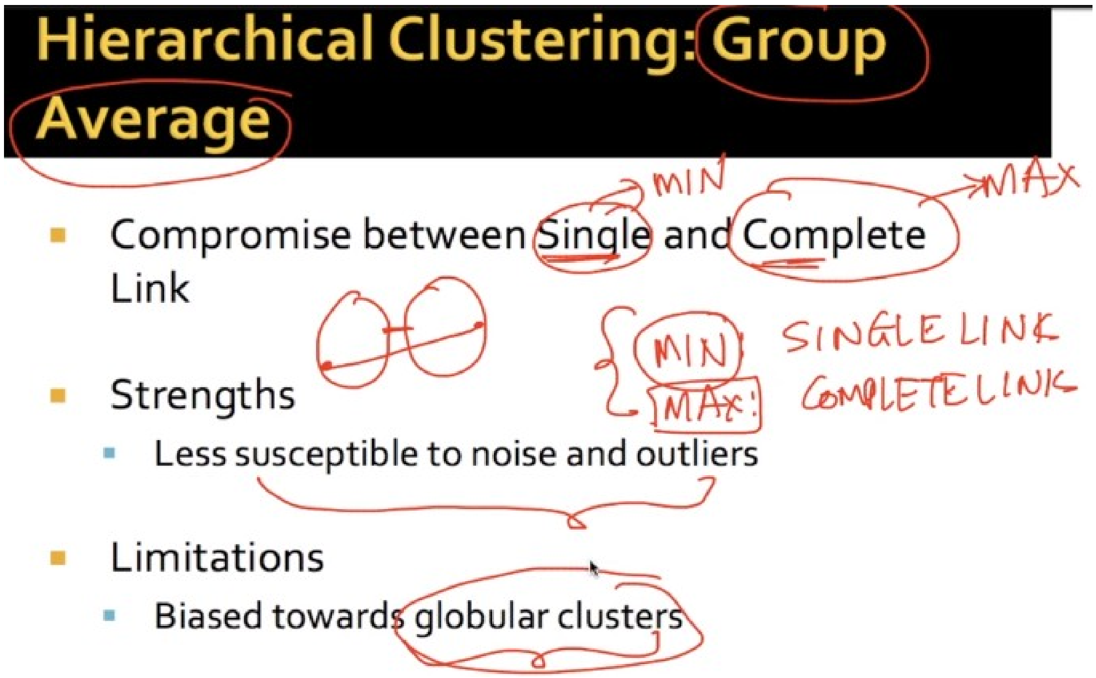
For example, if we have the large and small clusters like below.



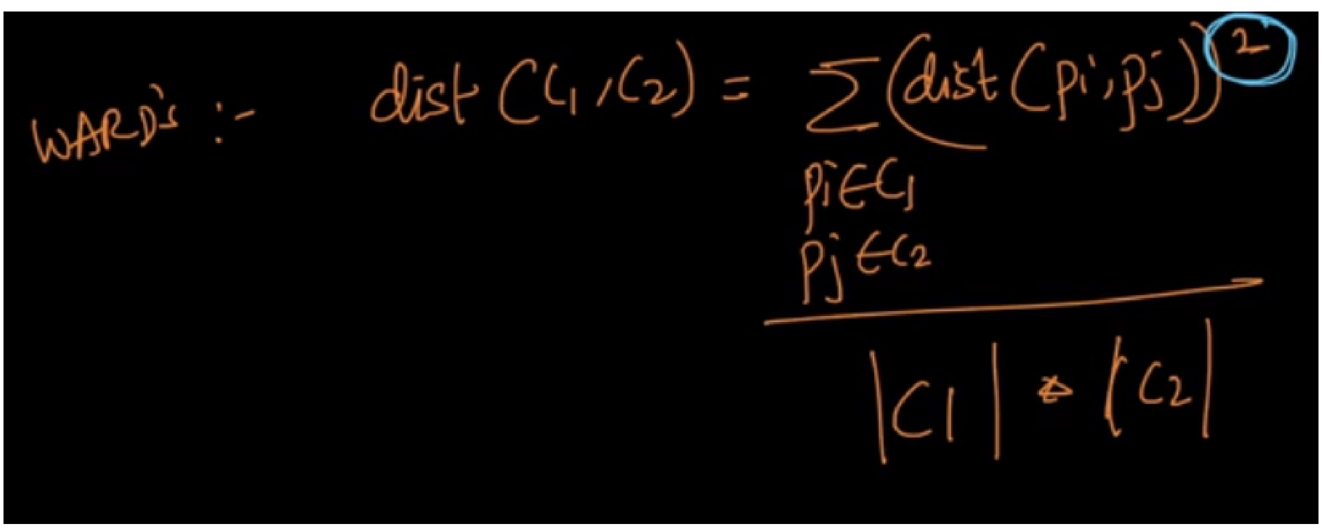
The biased towards globular clusters.

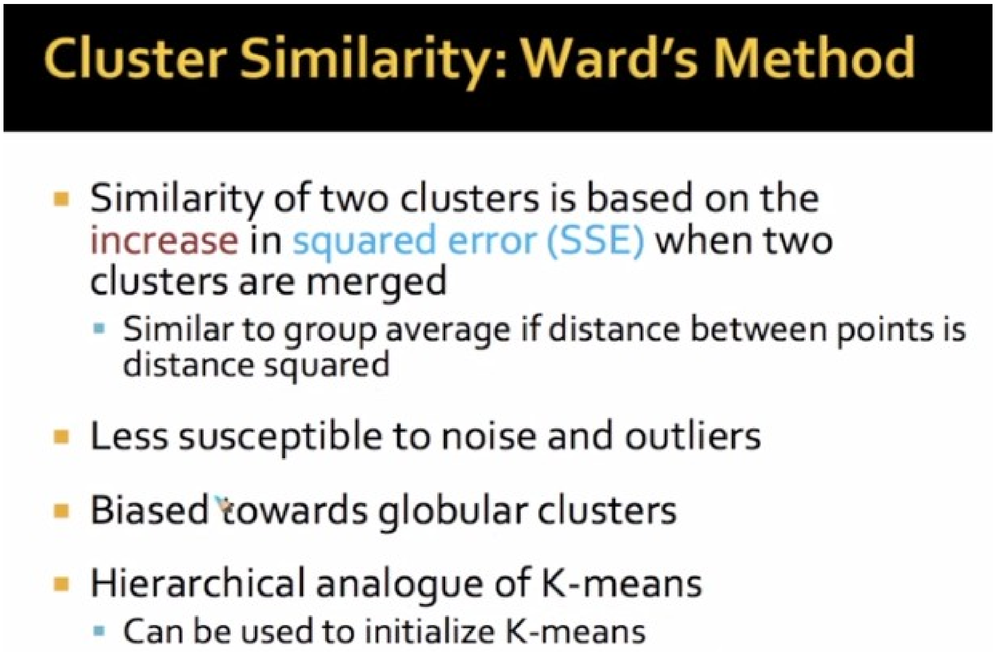
Group average:

It is the average of the single(min) and complete(max) clustering.



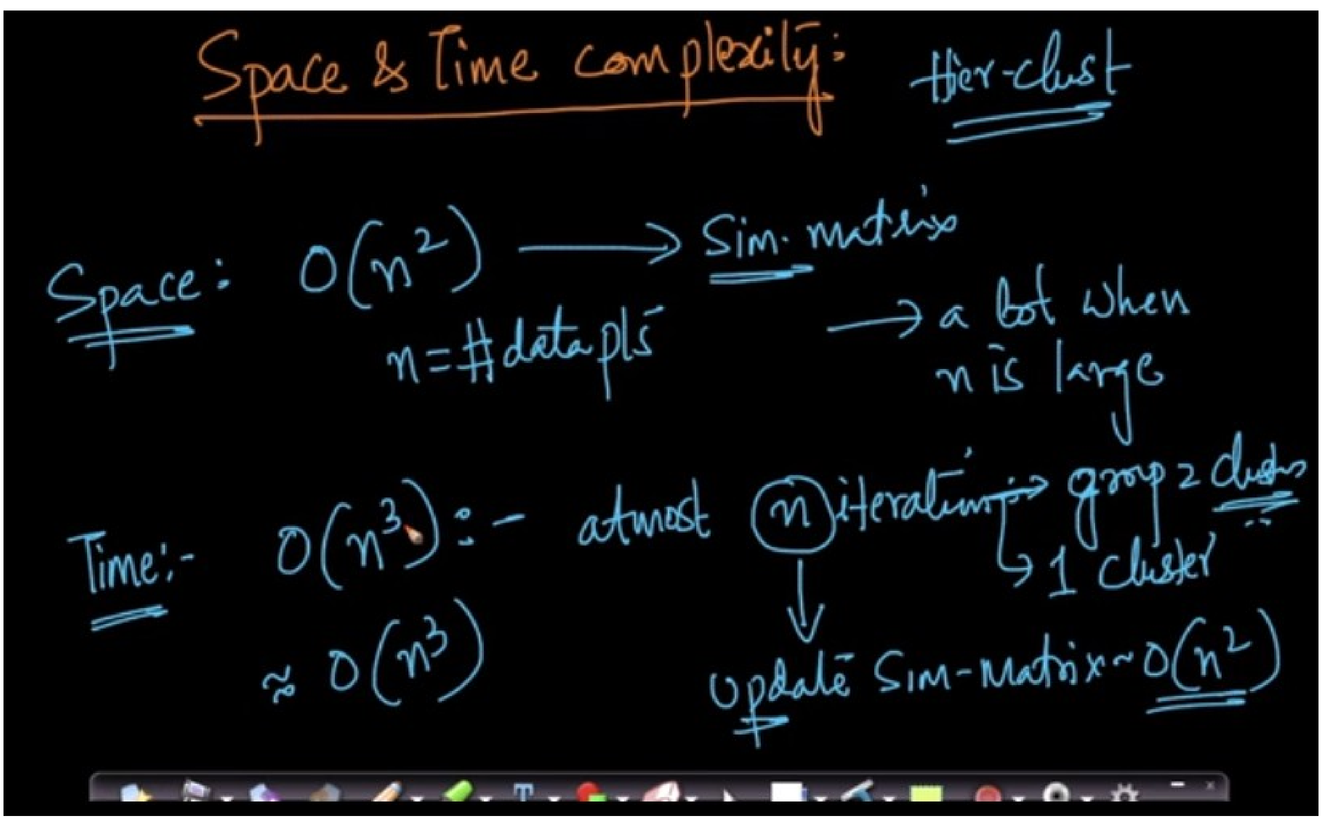
Ward's method:
The distances between the clusters is squared distance.





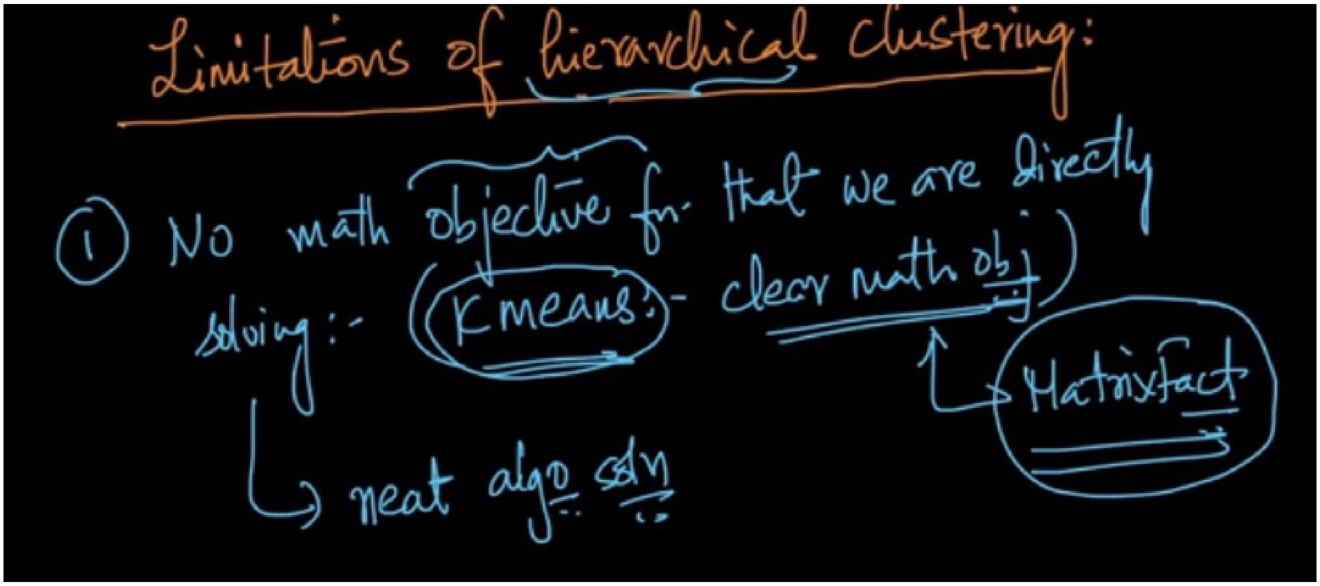
Time and space Complexity:

Sim. Matrix O(n^2):

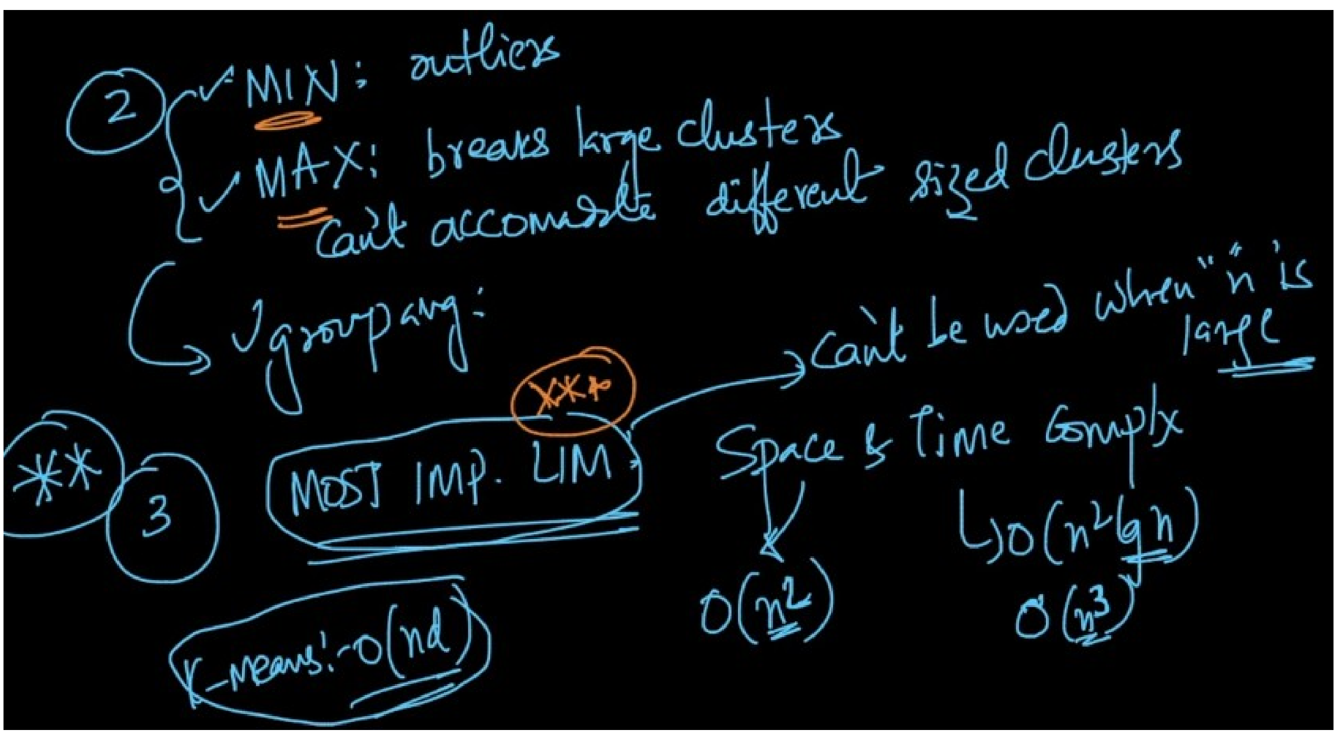


When ‘n’ is large, we need more memory for performing clustering.
We cannot use agglomorative as ‘n’ value is large.
Limitations of Hierarchical Clustering:

It is an algorithmic solution.

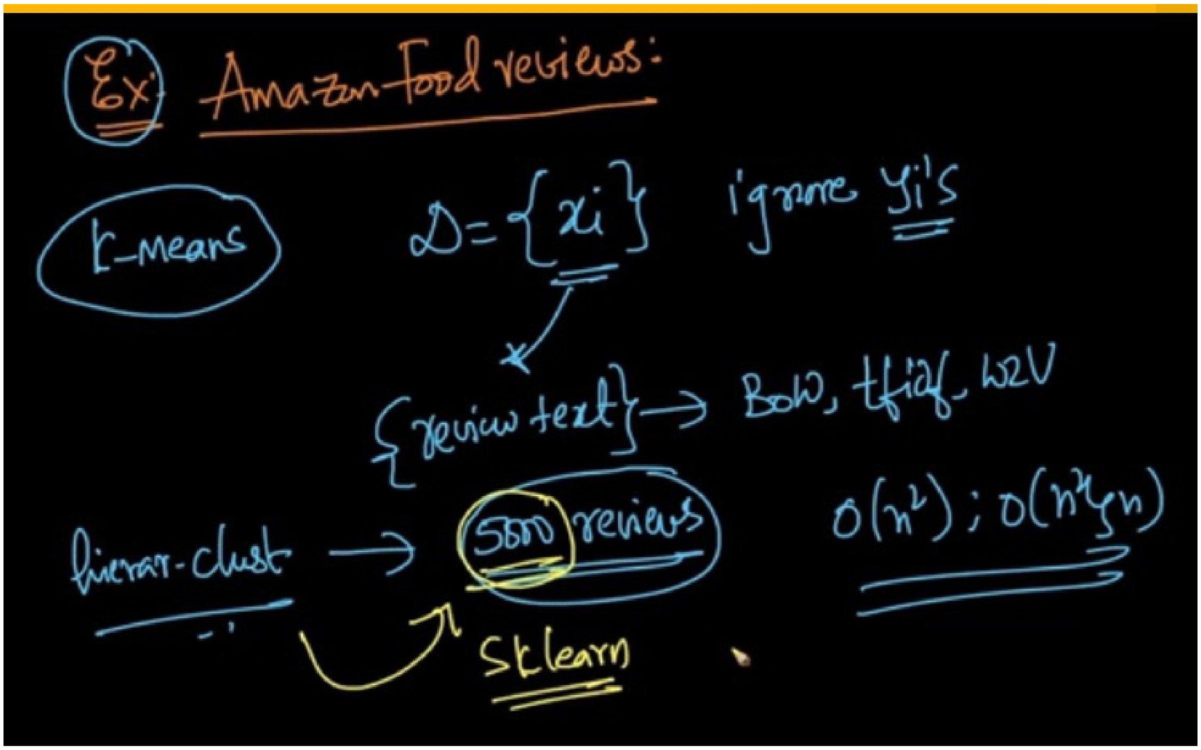


Each of the cluster technique has its own advantage and disadvantage. Its space and time complexity.



Exercise:

Apply on Amazon data set.



Make multiple clusters