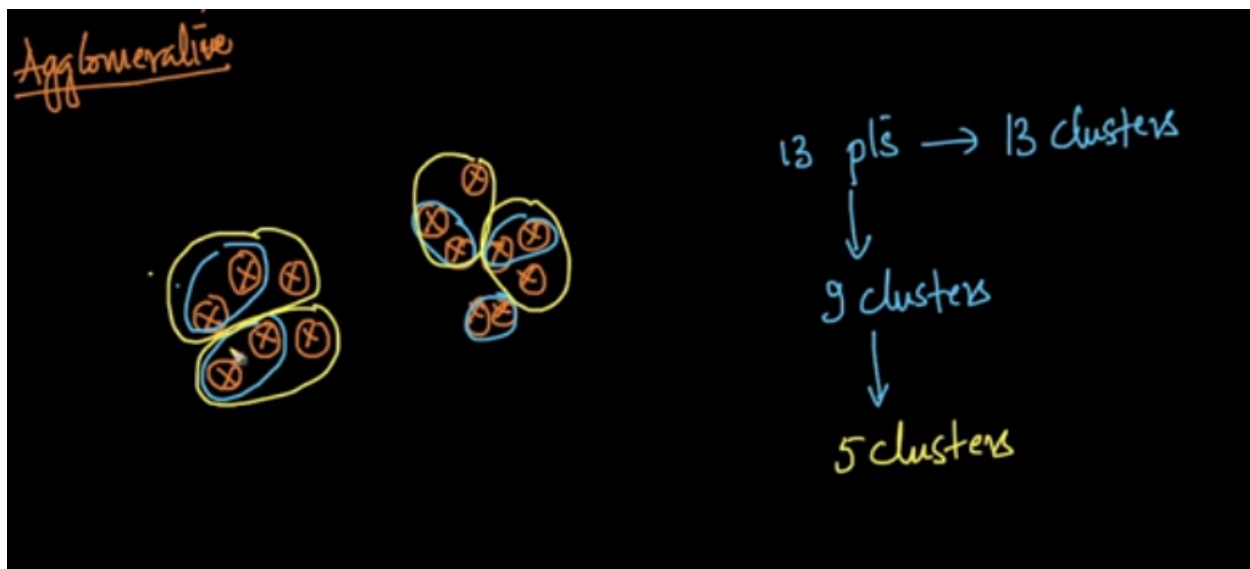
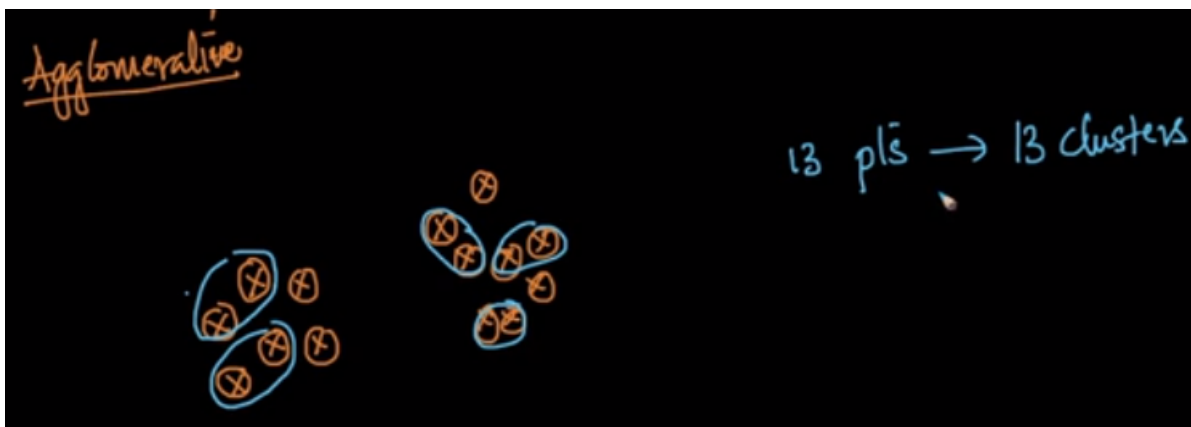


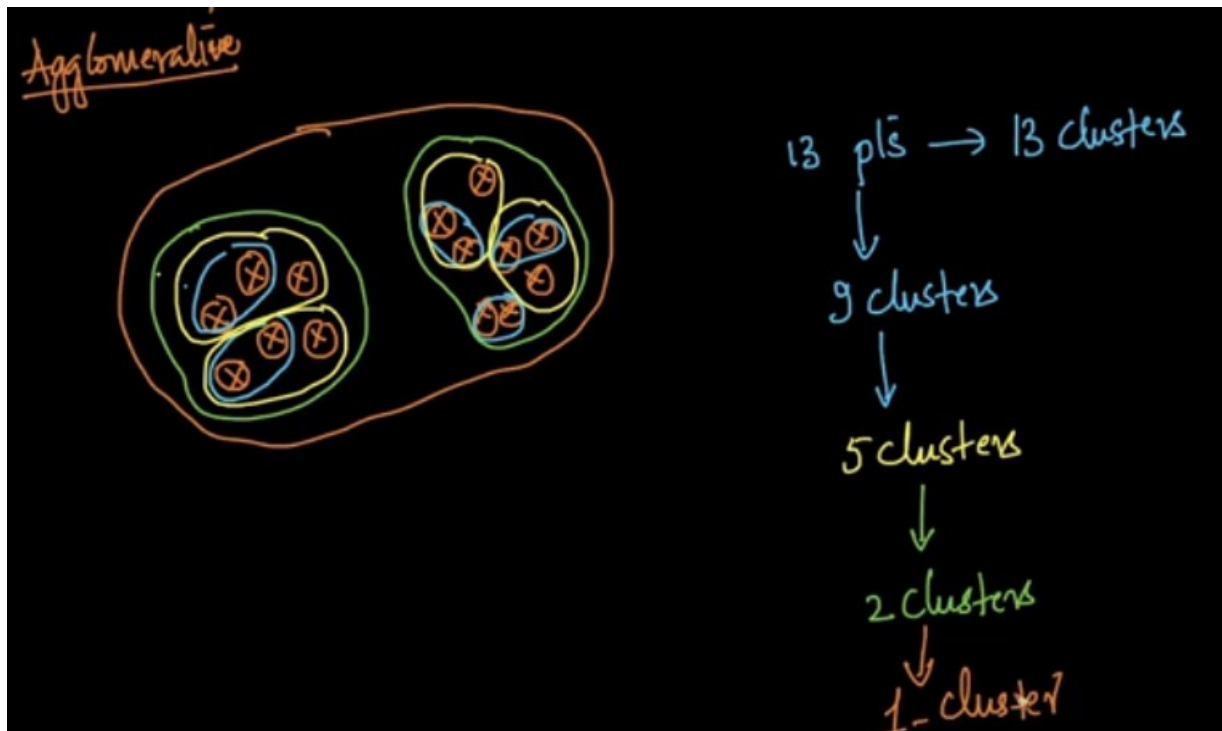
Agglomerative and divisive, Dendograms:

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...

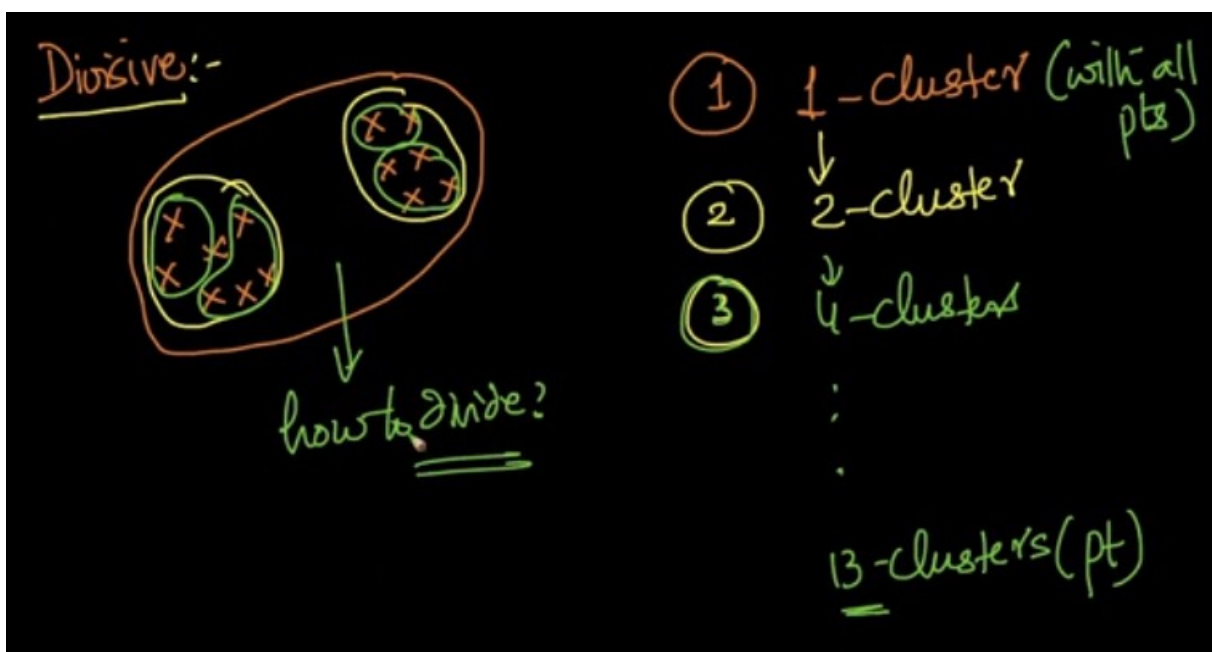


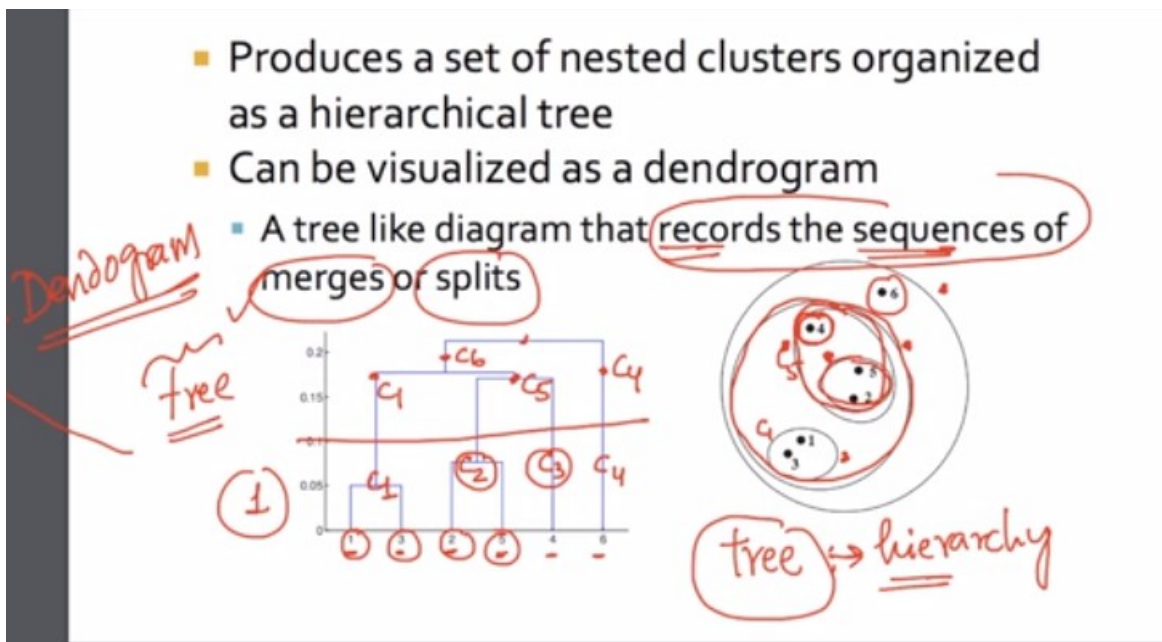
In agglomerative clustering we group with some similarity (or) distance.

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.

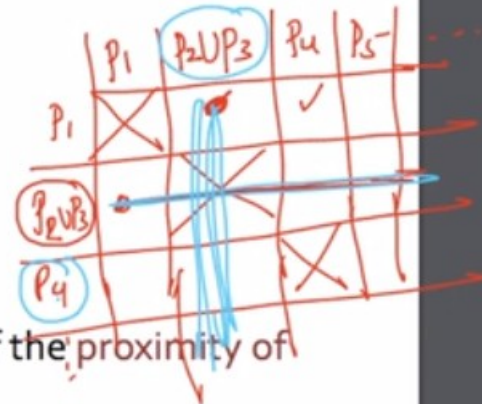
- Basic algorithm is straightforward
  1. Compute the proximity matrix
  2. Let each data point be a cluster
  3. Repeat
    4. Merge the two closest clusters
    5. Update the proximity matrix
  6. Until only a single cluster remains
- Key operation is the computation of the proximity of two clusters
  - Different approaches to defining the distance between clusters distinguish the different algorithms

	$p_1$	$p_2$	$p_3$	$p_4$	$p_5$
$p_1$	⊗	□	□	□	□
$p_2$	□	□	□	□	□
$p_3$	□	□	□	□	□
$p_4$	□	□	□	□	□
$p_5$	□	□	□	□	□

After each step the proximity matrix is updated.

- Basic algorithm is straightforward

1. Compute the proximity matrix
2. Let each data point be a cluster
3. **Repeat**
4.     Merge the two closest clusters
5.     Update the proximity matrix
6. **Until** only a single cluster remains



- Key operation is the computation of the proximity of two clusters

- Different approaches to defining the distance between clusters distinguish the different algorithms

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.

- Start with clusters of individual points and a proximity matrix



	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
...						

Proximity Matrix



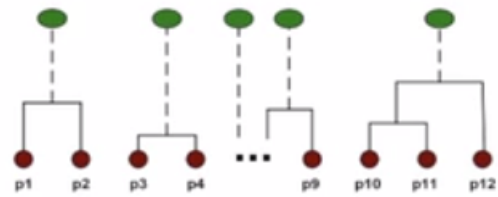
In every iteration we are merge the points.

- After some merging steps, we have some clusters



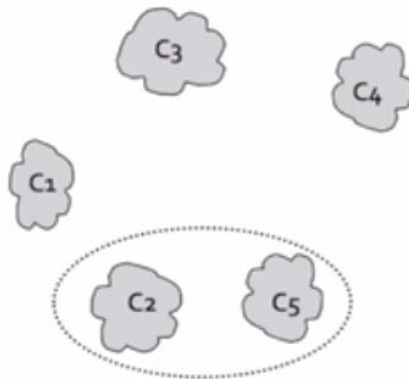
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix



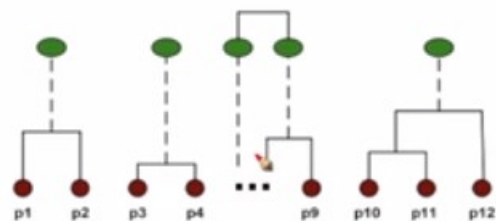
## Intermediate Situation

- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.



	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix

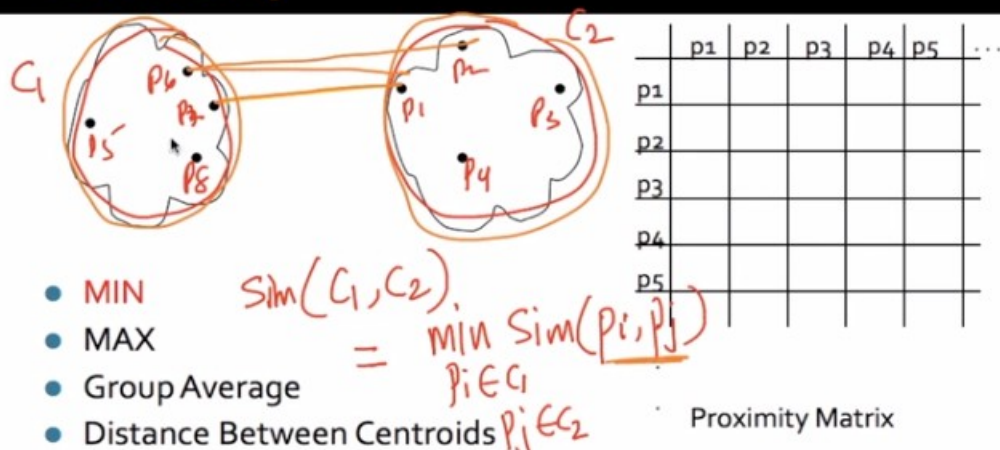


How to define the inter – cluster similarity:

Proximity methods: Advantages and Limitations.  
 MIN approach:

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

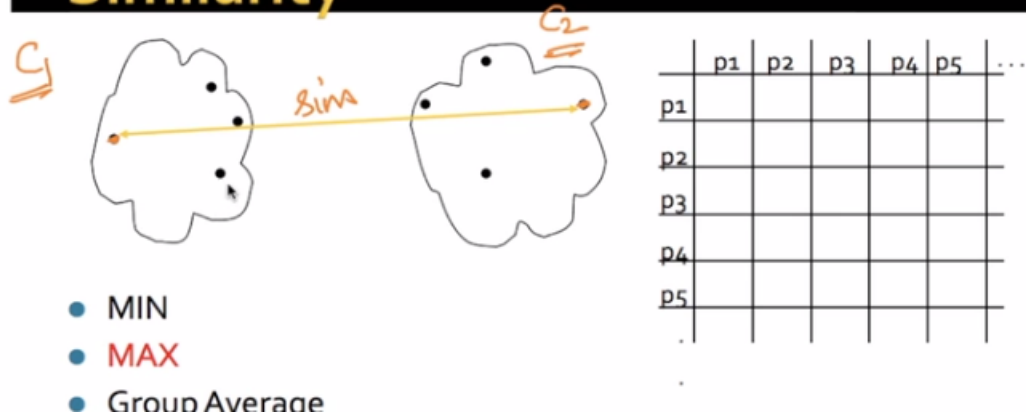
## How to Define Inter-Cluster Similarity





MAX approach:

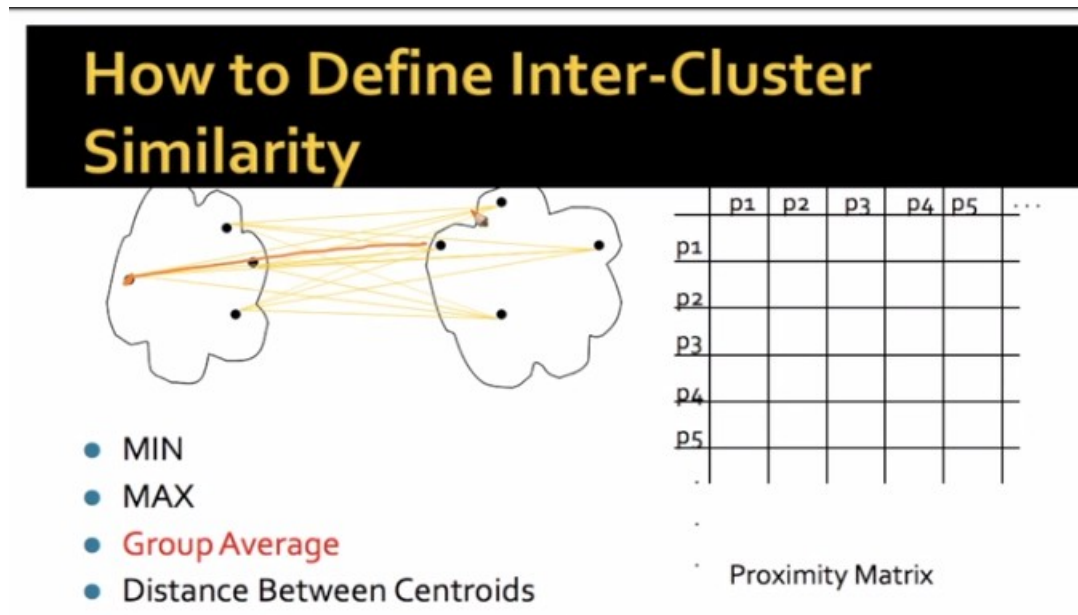
## How to Define Inter-Cluster Similarity



MIN:  $\text{Sim}(G_1, G_2) = \min_{\substack{p_i \in G_1 \\ p_j \in G_2}} \text{Sim}(p_i, p_j)$

MAX:  $\text{Sim}(G_1, G_2) = \max_{\substack{p_i \in G_1 \\ p_j \in G_2}} \text{Sim}(p_i, p_j)$

Group average:



Take the average of the similarity values

ANG:  $\text{Sim}(C_1, C_2) = \frac{\sum_{\substack{p_i \in C_1 \\ p_j \in C_2}} \text{Sim}(p_i, p_j)}{|C_1| * |C_2|}$

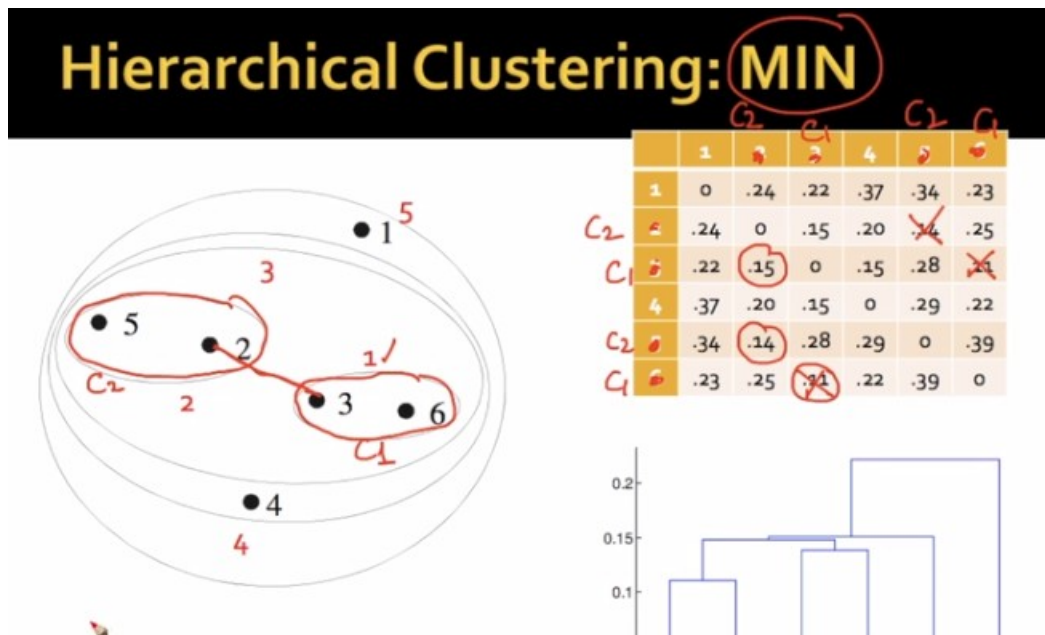
size of  $C_1$   $\swarrow$   $\nwarrow$  size of  $C_2$

Similarly is the cosine similarity.

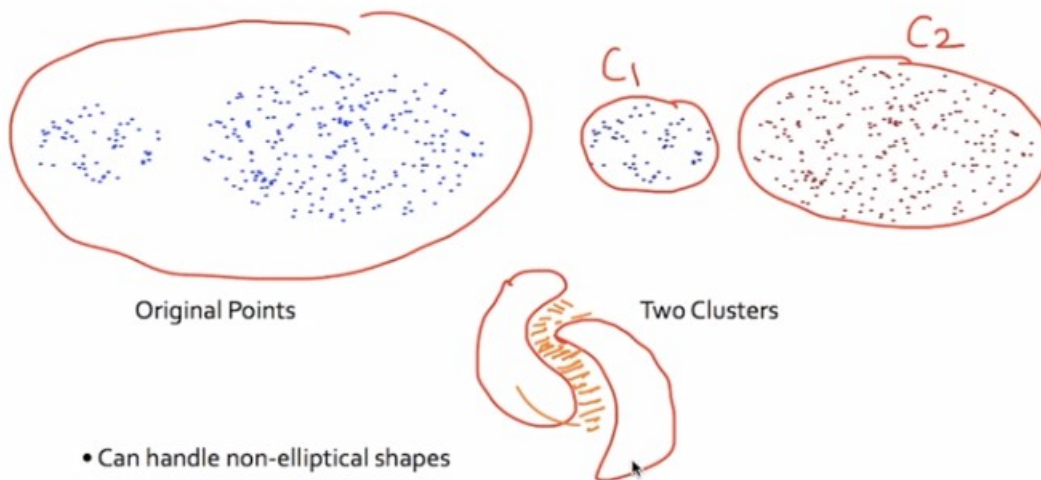


All the methods can be kernelized, except the distance between the centroids.

Hierarchical Clustering: MIN

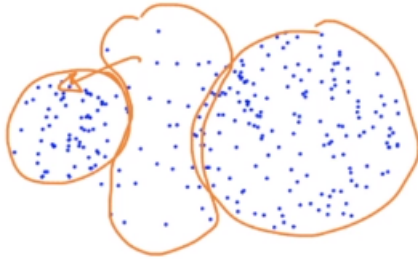


## Strength of MIN



Limitations:

## Limitations of MIN



Original Points



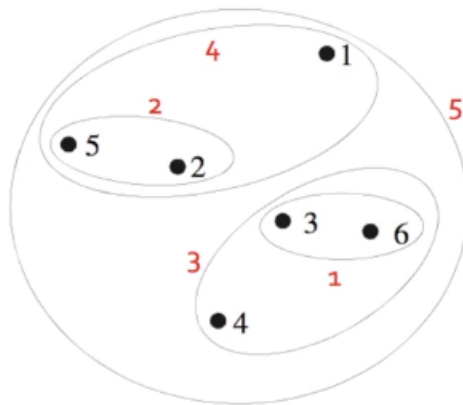
Two Clusters

- Sensitive to noise and outliers

It is extremely sensitive to the noise.

Max:

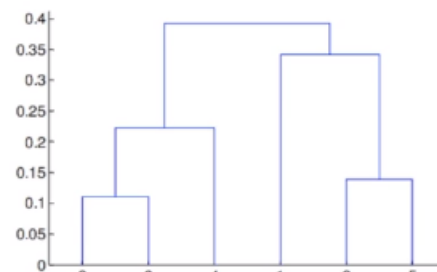
## Hierarchical Clustering: MAX



Nested Clusters

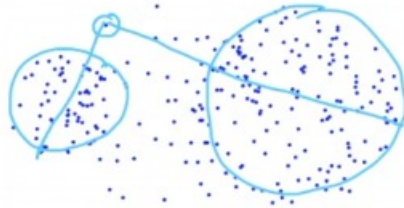
Dendrogram

	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

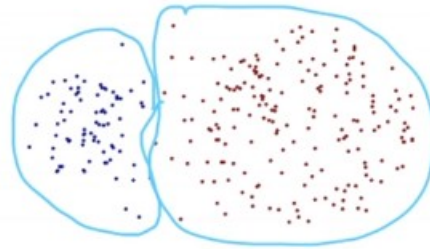


Strength of MAX:

## Strength of MAX



Original Points

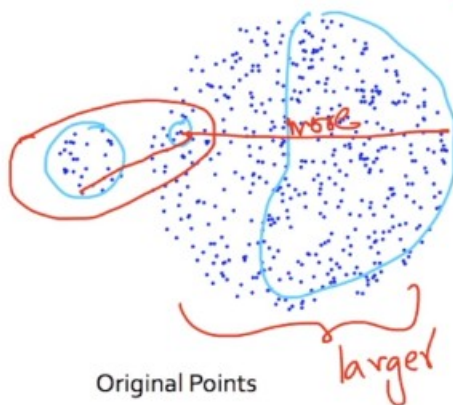


Two Clusters

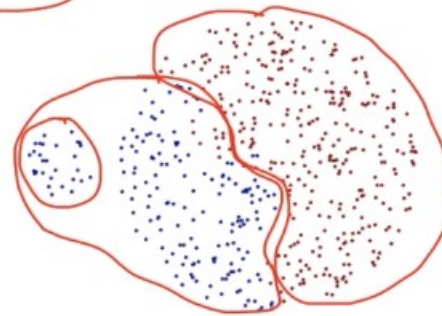
- Less susceptible to noise and outliers

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

## Limitations of MAX



Original Points



Two Clusters

- Tends to break large clusters ✓
- Biased towards globular clusters

The biased towards globular clusters.

Group average:

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

## Hierarchical Clustering: Group Average

- Compromise between Single and Complete Link
- Strengths
  - Less susceptible to noise and outliers
- Limitations
  - Biased towards globular clusters

MIN MAX

SINGLE LINK COMPLETE LINK

Ward's method:

The distances between the clusters is squared distance.

Ward's :- 
$$\text{dist}(C_1, C_2) = \frac{\sum_{\substack{p_i \in C_1 \\ p_j \in C_2}} (\text{dist}(p_i, p_j))^2}{|C_1| * |C_2|}$$

## Cluster Similarity: Ward's Method

- Similarity of two clusters is based on the **increase** in **squared error (SSE)** when two clusters are merged
  - Similar to group average if distance between points is distance squared
- Less susceptible to noise and outliers
- Biased towards globular clusters
- Hierarchical analogue of K-means
  - Can be used to initialize K-means

Time and space Complexity:

Sim. Matrix  $O(n^2)$ :

Space & Time complexity: hier-clust

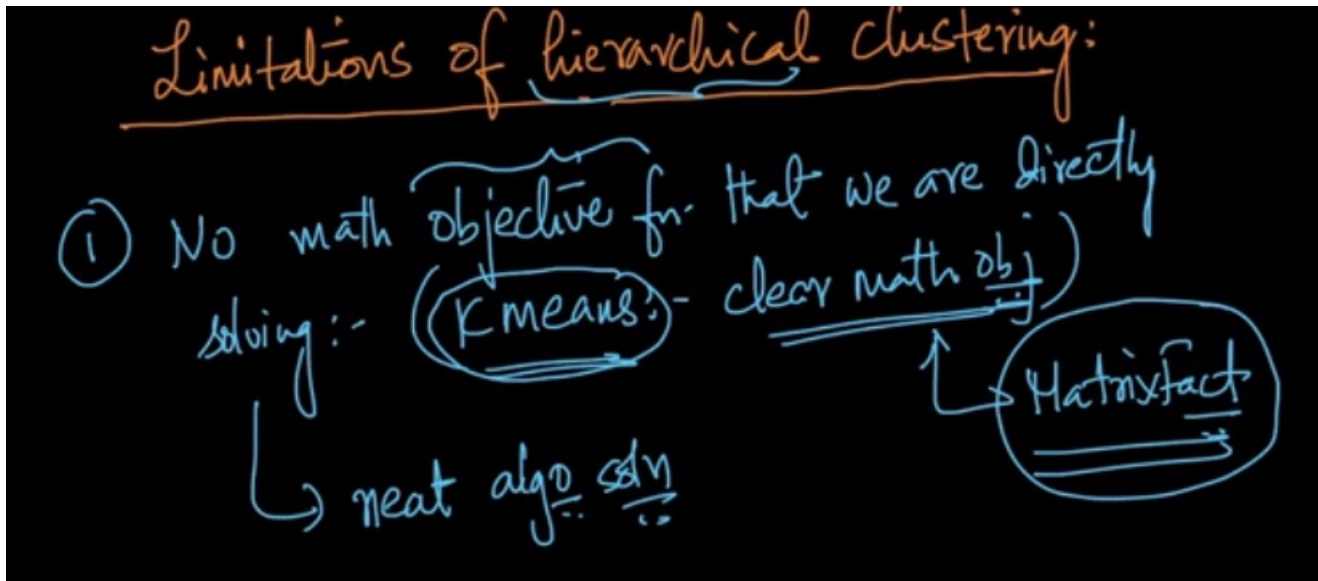
Space:  $O(n^2)$   $\rightarrow$  Sim. matrix  
 $n = \# \text{ data pts}$   $\rightarrow$  a bit when  $n$  is large

Time:  $O(n^3)$  :- atmost  $(n)$  iterations  $\rightarrow$  group 2 clusters  $\rightarrow$  1 cluster  
 $\approx O(n^3)$   
 $\downarrow$   
update Sim-matrix  $\sim O(n^2)$

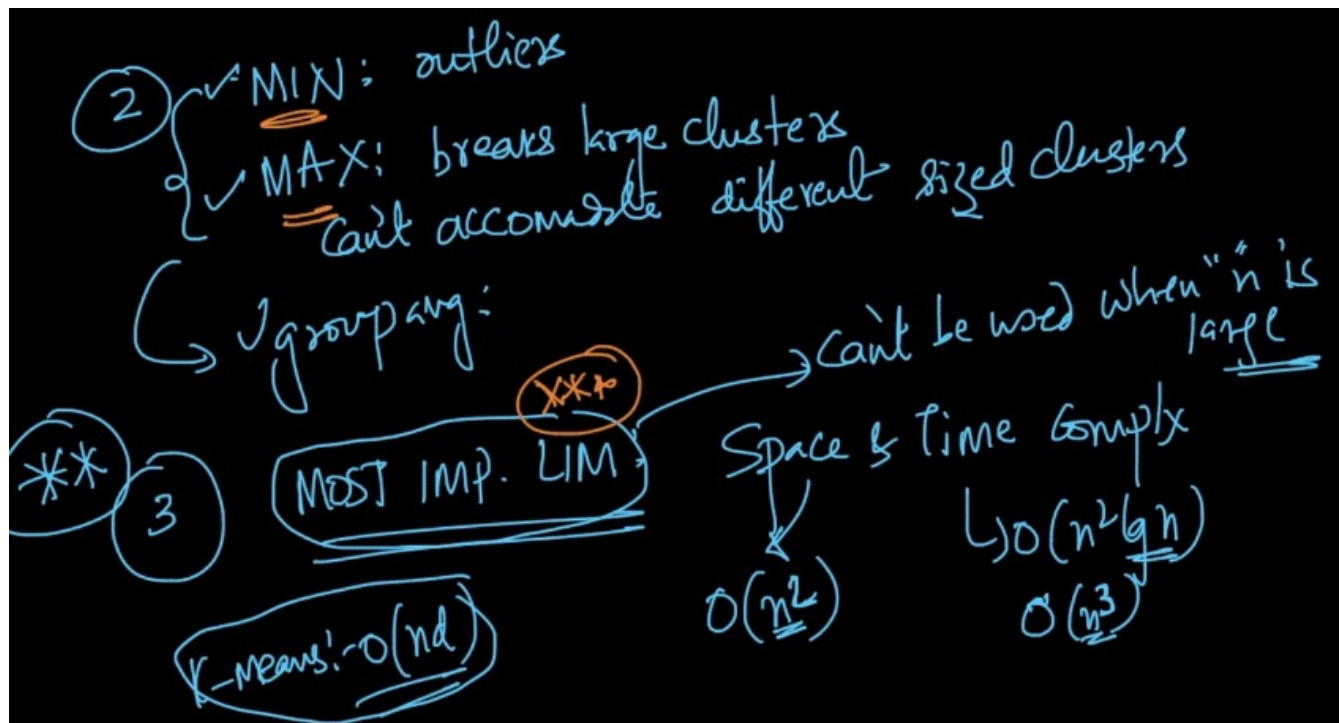


When 'n' is large, we need more memory for performing clustering.  
We cannot use agglomerative as 'n' value is large.  
Limitations of Hierarchical Clustering:

It is an algorithmic solution.



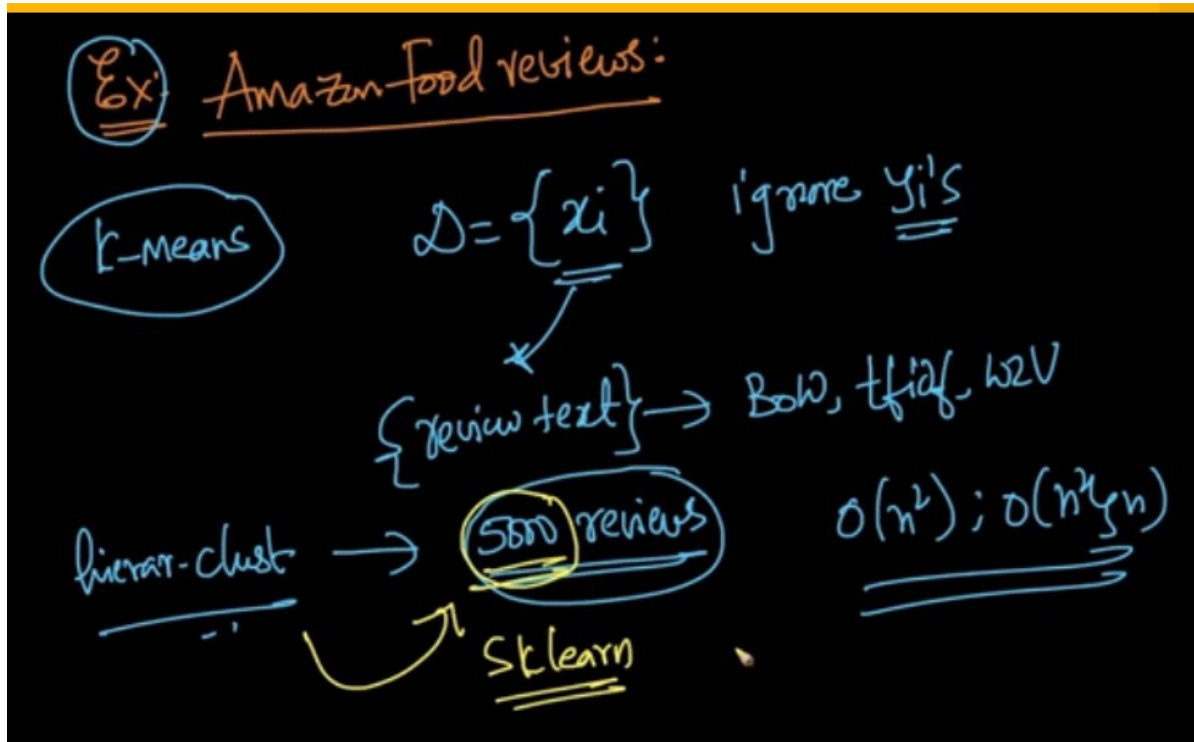
Each of the cluster technique has its own advantage and disadvantage. Its space and time complexity.





Exercise:

Apply on Amazon data set.



Make multiple clusters

