Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology - Roorkee

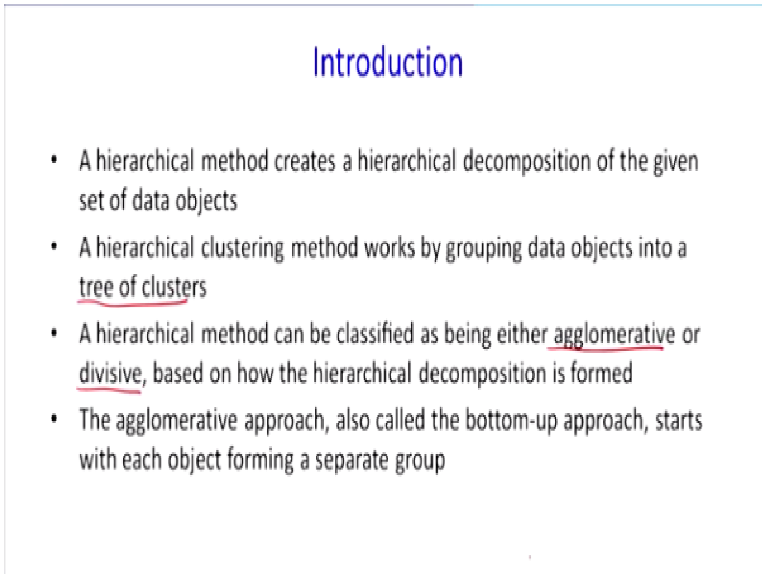
Lecture – 55
Hierarchical Method of Clustering - I

In our previous lecture, I have explained about K - means algorithm that is one type of clustering technique.

(Refer Slide Time: 00:32)

There is another type of technique is hierarchical method of clustering, let us see what is hierarchical method of clustering in this lecture and we will compare that partitioning method versus hierarchical clustering methods in this lecture. So, the agenda for this lecture is introduction to hierarchical clustering, then comparison of partitioning versus hierarchical clustering methods.

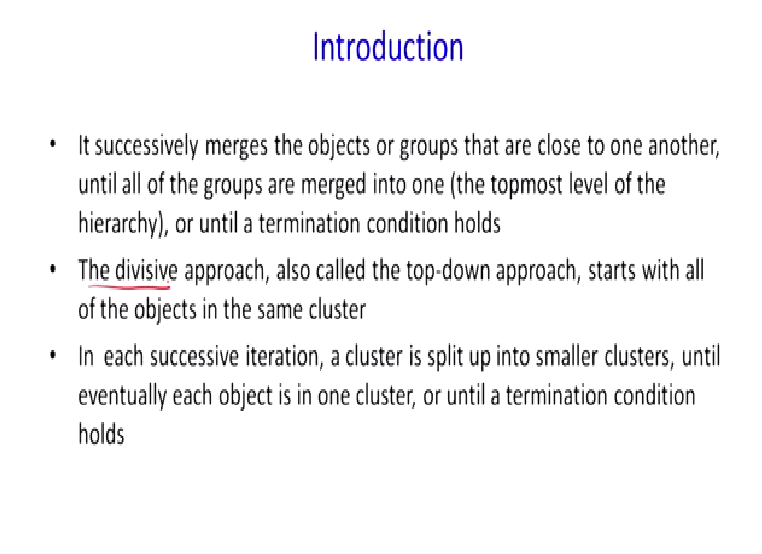
(Refer Slide Time: 00:53)



A hierarchical method creates a hierarchical decomposition of the given set of data objects, a hierarchical clustering method works by grouping data objects into tree of clusters. Here the tree of clusters, I will explain what is this tree of clusters in next slides. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed.

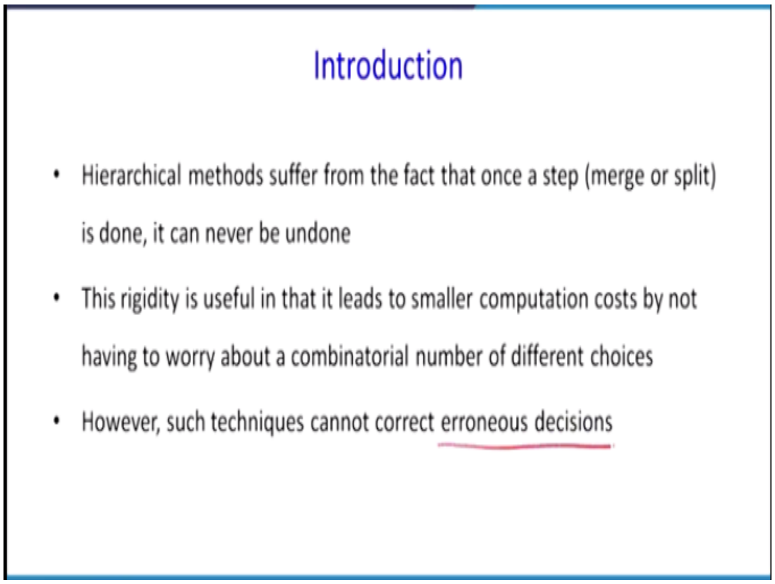
There are 2 way we can say in hierarchical method, one is agglomerative, the second one is divisive, the agglomerative approach also called bottom up approach, start with each object forming a separate group.

(Refer Slide Time: 01:42)



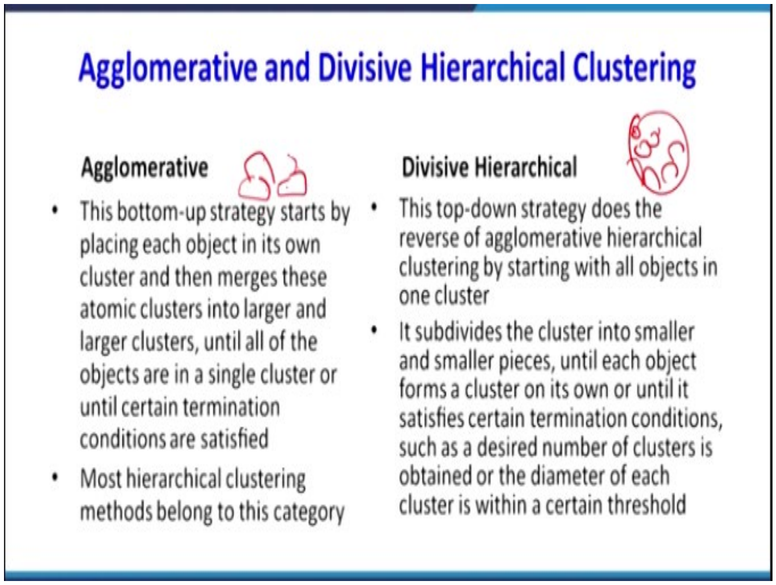
It successively merges the objects or groups that are close to one another, until all of the groups are merged into the top most level of the hierarchy or until a termination condition holds, on the other hand, the another classification in the hierarchical method is in divisive approach also called top down approach starts with of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster.

(Refer Slide Time: 02:27)



Or until a termination condition hold, hierarchical methods suffer from the fact that once a step is done, it can never be undone, so the problem with the hierarchical cluster is once a step is done, you cannot go back and correct the mistake. This rigidity is useful in that it leads to smaller computation cost by not having to worry about combinatorial number of different choices, however, such techniques cannot correct erroneous decisions that is only drawback of this hierarchical methods.

(Refer Slide Time: 03:04)

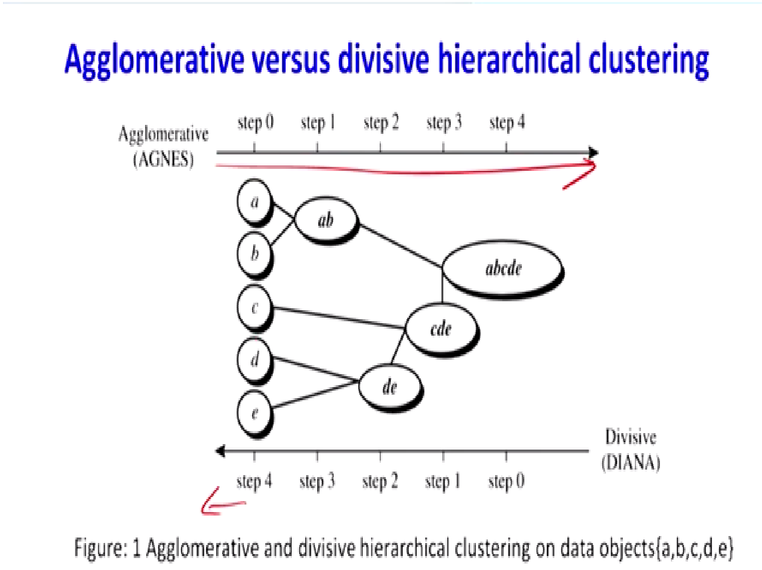


Let us compare agglomerative method versus divisive method, both are hierarchical method; let us see how this it is differ from each other. The agglomerative method this is bottom up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters until all of the objects are in a single cluster or until certain termination condition is satisfied.

Most hierarchical clustering methods belong to this category, on the other hand, the divisive method is a top down strategy does the reverse of agglomerative hierarchical clustering by starting with all object in one cluster. So, in divisive methods what we are doing, so you start from a bigger cluster, then you make smaller one, like it is a cutting a big cake into small pieces. On the other hand, the agglomerative is each object is separate clusters.

Then from that you can form all possible types of clusters that is kind of a tree, it is up to you to decide where you need to have the termination condition. In the divisive method, the divisive method sub divides the cluster into smaller and smaller pieces until each object forms a cluster on its own or until it satisfy certain termination condition such as a desired number of cluster is obtained or the diameter of each cluster is within the certain threshold.

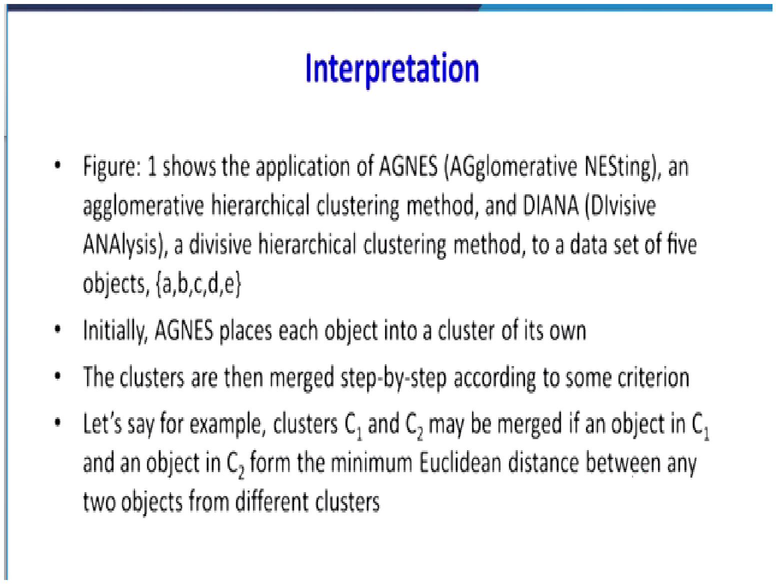
(Refer Slide Time: 04:47)



This picture explains the difference between agglomerative and divisive method, you see that the arrow for agglomerative method is going on this side, it says that there are a, b, c, d, there are 5 objects, we start with each objects are in a separate cluster, then this a and b in step 0, all are clusters; 1, 2, 3, 4, 5 cluster is there, each cluster having only 1 unit. In step 1, see a and b are clubbed that is ab.

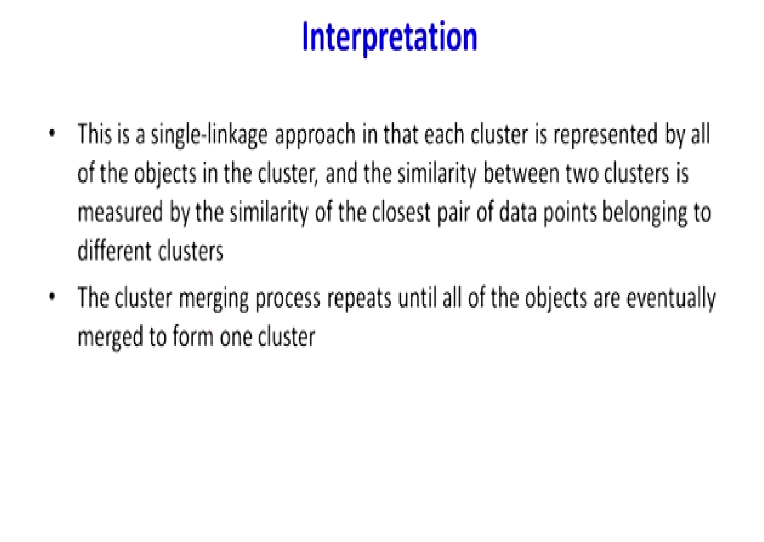
In step 2, d and e are clubbed, in step 3 this c, d, e are clubbed, in step 4 all these a, b, c, d is clubbed, so this is going in from left to right that is agglomerative method whereas in the divisive method, you start from; you see that look at this arrow it is going this side, start from the a, b, c, d that is a big by considering all the elements. A step 0; look at the step 0, only one cluster, in step 1, the c, d, e is 1, in step 2 de is another cluster. In step 3, from a, b, c, d again the ab has come out, in step 4 all individual elements are separate clusters that is a basic difference between agglomerative versus divisive hierarchical clustering.

(Refer Slide Time: 06:20)



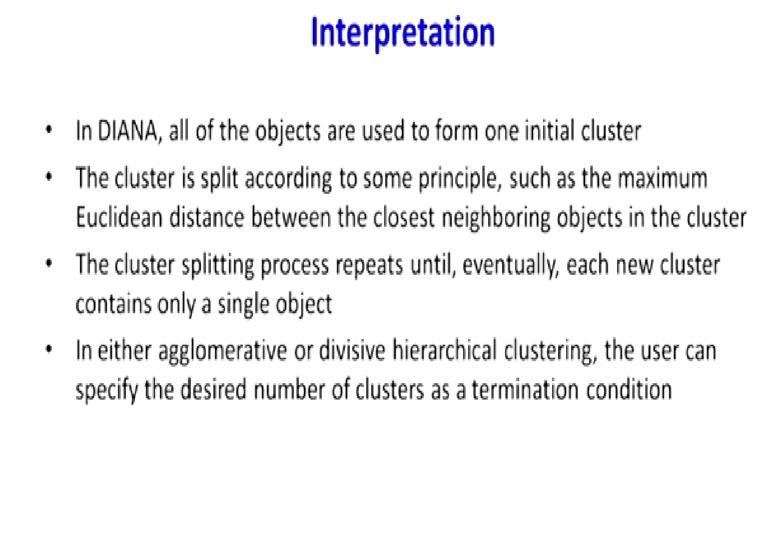
What is the interpretation of the previous slides? Figure 1 shows that application of aggromerative AGNES; agglomerative nesting, an agglomerative hierarchical clustering methods and DIANA divisive analysis, a divisive hierarchical clustering method to a data set of 5 objects; a, b, c, d, e. Initially, agglomerative method places each object into a cluster of its own, there is only one item. The clusters at then merged step by step according to some criterion, let us say for example, cluster C1 and C2 may be merged if an object C1 and the object C2 from the minimum Euclidean between any 2 objects from different clusters.

(Refer Slide Time: 07:13)



This is a single linkage approach, in that each cluster is represented by all of the objects in the cluster and the similarity between 2 cluster is measured by the similarity of the closest pair of the data points belonging to different clusters. The cluster merging process repeats until all of the objects are eventually merged into form 1 cluster. So, what is happening here it is start from the step 0, it goes up to step 4, you see in step 0, there are 1, 2, 3, 4, 5 clusters in step 0.

(Refer Slide Time: 07:56)



But in step 1, all are merged into only 1 cluster that is a, b, c, d, e. In DIANA that is in the divisive method, all of the objects are used to form 1 initial cluster, this cluster split according to some principle such as maximum Euclidean distance between the closest neighbouring objects in the cluster. The cluster splitting process repeats until eventually each new cluster contains only 1 object, only a single object.

In either agglomerative or divisive hierarchical clustering, the user can specify the desired number of clusters as a termination condition. So, here the explanation of divisive method is start from here, in step 0 only 1 cluster is there. In step 4, there are 5 clusters is there, that is a difference.

(Refer Slide Time: 08:43)

Dendrogram

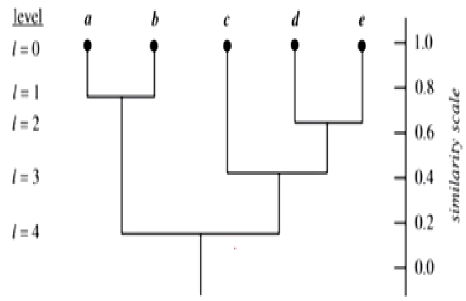


Figure 2: Dendrogram representation for hierarchical clustering of data objects{a,b,c,d,e}

In hierarchical clustering, another important terminology which you have to understand is dendrogram. What is a dendrogram? It is a kind of a tree kind of structure and it says there are different levels, level 0, 1, 2, 3, 4 on left hand side and right hand side there is a similarity scale. At level 0, there are 1, 2, 3, 4, 5 clusters a, b, c, d, e all are forming its own cluster. In level 1, ab is forming 1 cluster.

In level 2, the position of c is compared, in the position of c we are finding the distance between c and between these cluster a and b and cluster d and e. If it is closer to d and e, then c, d, e form an another cluster that is a level 3. Level 4, it is only 1 cluster, all are 5 elements are present in there, so this kind of picture is called dendrogram.

(Refer Slide Time: 09:45)

Dendrogram

- A tree structure called a dendrogram is commonly used to represent the process of hierarchical clustering
- It shows how objects are grouped together step by step
- Figure: 2 shows a dendrogram for the five objects presented in Figure:1 , where l =0 shows the five objects as singleton clusters at level 0
- At l =1, objects a and b are grouped together to form the first cluster, and they stay together at all subsequent levels

Dendrogram, a tree structure called a dendrogram is commonly used to represent the process of hierarchical clustering; it shows how objects are grouped together step by step. Figure 2 shows a dendrogram for the 5 objects presented in the figure 1, where l equal to 0, at level 0 shows the 5 objects are singleton clusters, there is only 1 element in the clusters at level 0. At level 1, object a and b are grouped together to form the first cluster and they stay together at all subsequent levels, this hierarchical structure can be understood with the help of this dendrogram.

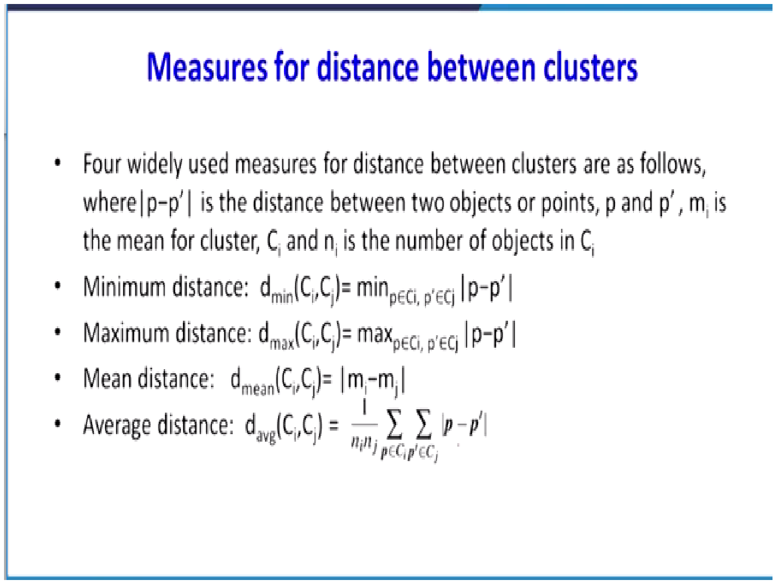
(Refer Slide Time: 10:27)

Dendrogram

- We can also use a vertical axis to show the similarity scale between clusters
- For example, when the similarity of two groups of objects, {a,b} and {c,d,e}, is roughly 0.16, they are merged together to form a single cluster

We can also use a vertical axis to show the similarity scale between the clusters actually, it is given on the right hand side of the picture. For example, when the similarity of 2 groups of object a and b and c, d, e is roughly 0.16, so they merged together to form a single cluster.

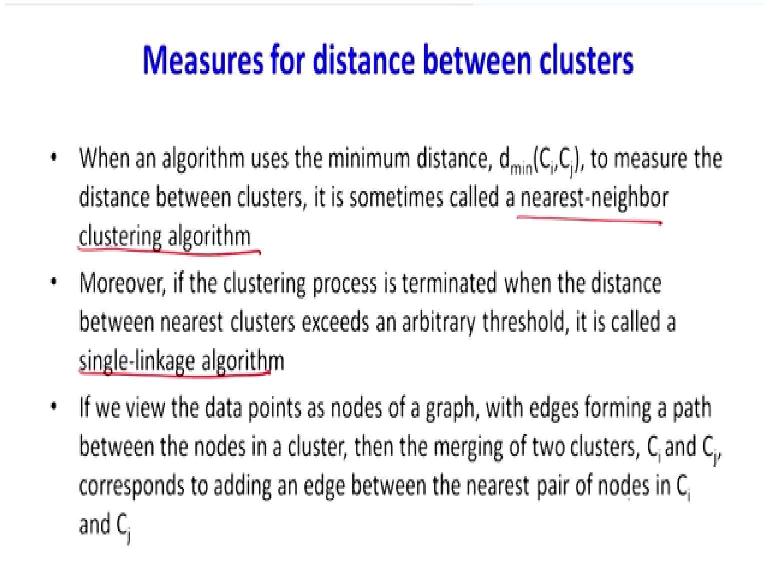
(Refer Slide Time: 10:49)



Now, let us go to another important idea of measures of distance between clusters, there are 4 widely used measures for distance between clusters are as follows, where modulus of p – p dash is the distance between 2 objects or points that is a p and p dash, where mi is the mean of the cluster, Ci and ni is the number of objects in Ci. The first measure is minimum distance, the minimum distance between cluster Ci and Cj equal to, so minimum of modulus of p – p dash.

The maximum distance is d max between cluster i and j equal to maximum of modulus of p – p dash, the mean distance; there is a another measure is d mean is the modulus of mean of the 2 clusters. The average distance is 1 divided by (ni – nj), here n represents number of object in cluster 1, then sigma of p – p dash modulus.

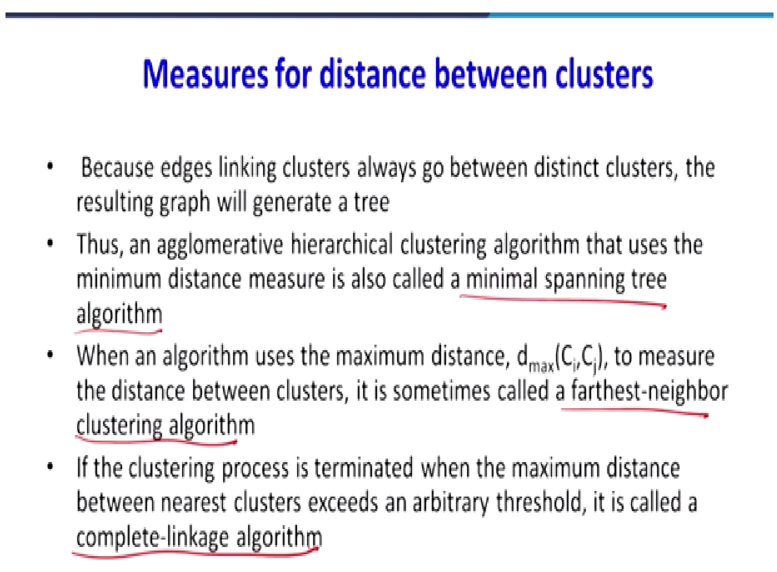
(Refer Slide Time: 12:08)



When an algorithm uses the minimum distance that is a d minimum Ci between Cj, that is a distance between Ci and Cj. To measure the distance between clusters, it is sometime called nearest neighbour clustering algorithm, I will show you in picture in coming slides. Moreover, if the clustering process is terminated, when the distance between the nearest clusters exceeds the arbitrary threshold, it is called single linkage algorithm.

If we view the data points as nodes of graph with edges forming a path between the nodes in a cluster, then the merging of 2 clusters; Ci and Cj corresponds to adding an edge between the nearest pair of nodes Ci and Cj.

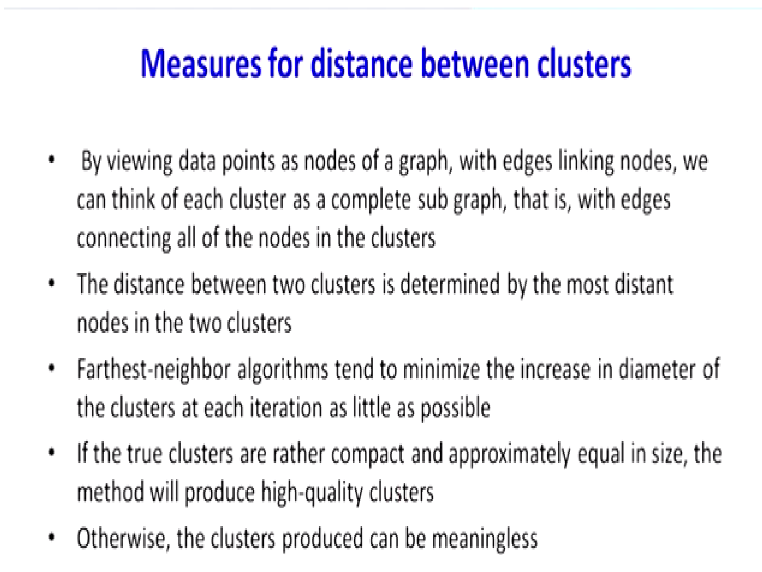
(Refer Slide Time: 12:55)



Because edges linking clusters always go between distinct clusters, the resulting graph will generate a tree, thus an agglomerative hierarchical clustering algorithm that uses the minimum distance measure is also called minimal spanning tree algorithm, even in your subject operation research also, there is a topic network problems in that you might have studied minimal spanning tree algorithm.

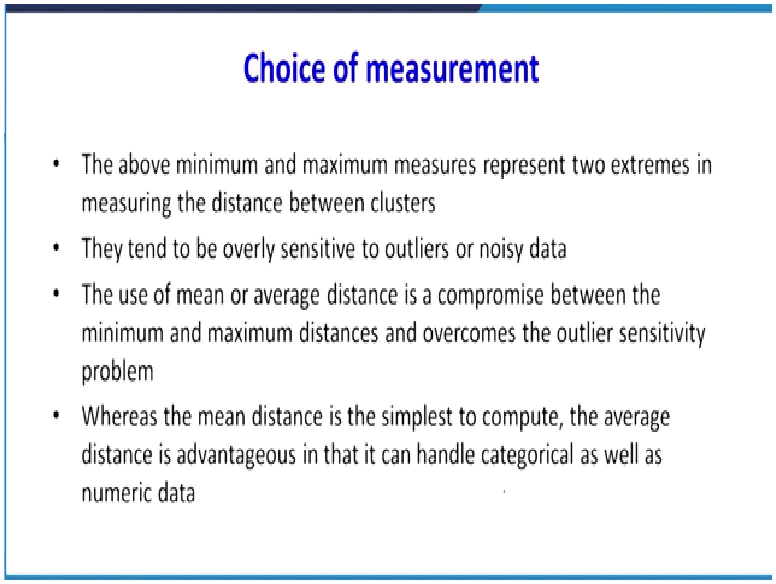
When an algorithm uses the maximum distance between cluster i and j, to measure the distance between clusters, it is sometime called farthest neighbour clustering algorithm. If the clustering process is terminated, when the maximum distance between nearest cluster exceeds an arbitrary threshold, it is called complete linkage algorithm.

(Refer Slide Time: 13:57)



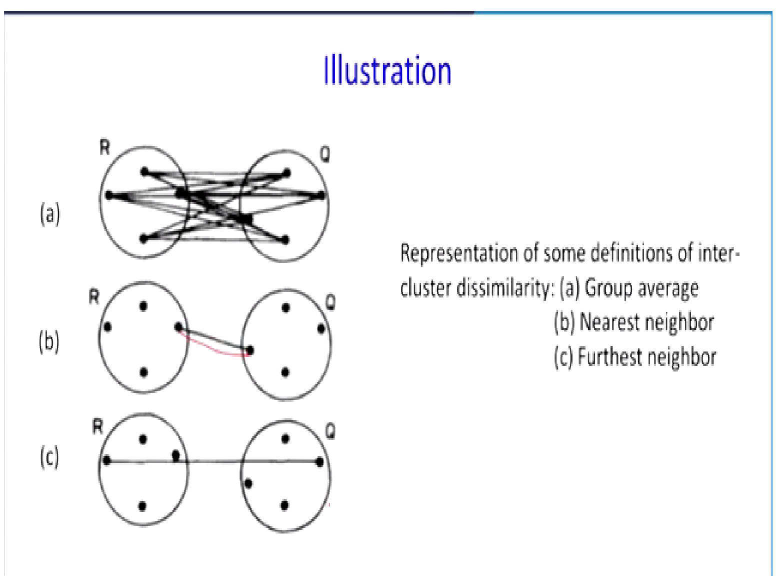
By viewing data points as nodes of graph with edges linking nodes, we can think of each cluster as a complete sub graph that is with edges connecting all of the nodes in the cluster. The distance between 2 cluster is determined by the most distant nodes in the 2 clusters, farthest neighbour algorithm tend to minimise the increase in diameter of the clusters at each iteration as little as possible. If the true clusters are rather compact and approximately equal in size, the method will produce high quality clusters, otherwise the clusters produced can be meaningless.

(Refer Slide Time: 14:40)



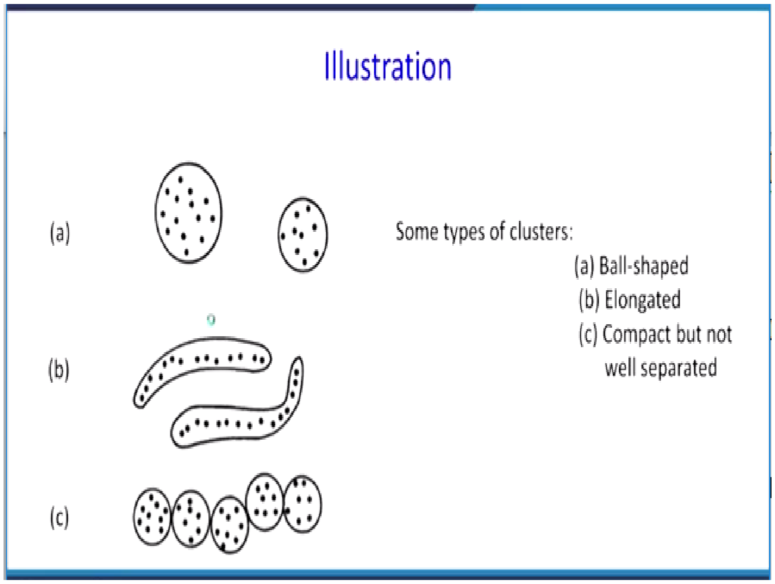
Let us go for choice of measurement; the above minimum and maximum measure represents 2 extremes in measuring the distance between clusters, they tend to be overly sensitive to outliers or noisy data. The use of mean or average distance is a compromise between minimum and maximum distances and overcome the outlier sensitivity problems, whereas the mean distance is the simplest to compute, the average distance is advantageous in that it can handle categorical as well as numeric data.

(Refer Slide Time: 15:24)



This picture shows the distance measures, see the first one represents group average, so you see that all the points are connected with all other points in that cluster. This R is one cluster; Q is one cluster you see that we have finding the beverage that is a group average. This is representation of some definition of inter cluster dissimilarity. The second one is the nearest neighbour. See, this is a nearest neighbour, the third one is the farthest neighbour that I have explained in my previous slides, this is a different type of distance measures.

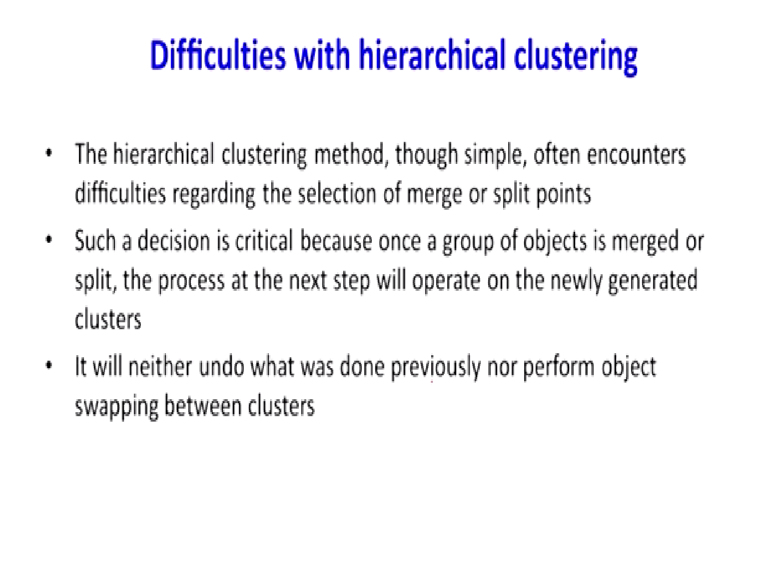
(Refer Slide Time: 16:08)



So, this picture shows some type of clusters, see the cluster here is the ball shaped one, the second one is elongated one the last one is compact but not well separated. So, what will happen; if we follow the group average, your; the final cluster may be this shape, what is that; the ball shaped one. If we follow this distance measures that is a nearest neighbour, the final cluster may look like this one that is elongated.

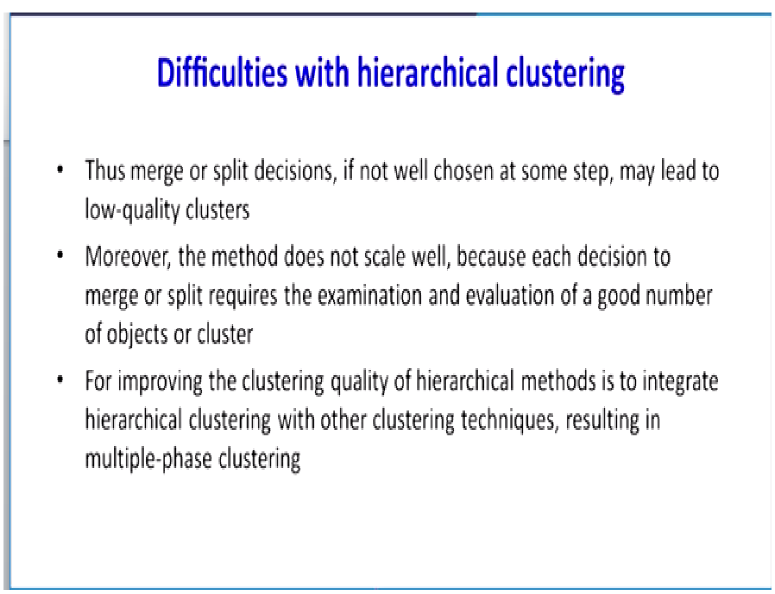
You see that any time, it can form with this point any time we can go to that other cluster, in case if you follow the farthest neighbour distance measures, your final cluster may be in this format that is a compact but not well separated, that is why choosing the correct distance is more important, based on the your distance measures, your shape of final cluster also will vary.

(Refer Slide Time: 17:15)



The difficulties with the hierarchical clustering; the hierarchical clustering method, though simple often encounters difficulties regarding the selection of merge or split points, such a decision is critical because once a group of object is merged or split, the process at the next step will operate on the newly generated clusters. As I told you, this also one of the drawback, once the cluster is formed, you cannot, any mistake has happened that cannot be rectified if you follow hierarchical clustering methods. See it will neither undo what was done previously nor perform objects swapping between clusters; these are the some of the disadvantages of hierarchical clustering.

(Refer Slide Time: 18:02)



Thus, merge or split decisions if not well chosen at some step may lead to low quality clusters, moreover, the method does not scale well because each decision to merge or split requires the examination and the evaluation of good number of objects or clusters. For improving the cluster quality of hierarchical method is to integrate hierarchical clustering with other clustering techniques resulting in multiple phase clustering. So, what we can do; if you want to improve the quality of hierarchical clustering, it can be clubbed with other clustering algorithms, so that you can improve the quality of our clustering.

(Refer Slide Time: 18:49)

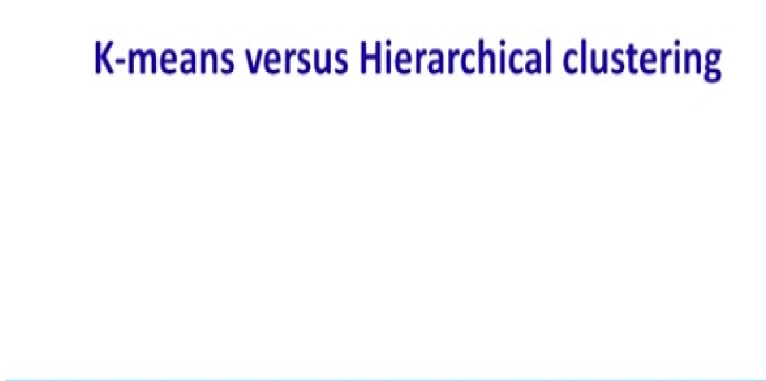
Partitioning Vs. Hierarchical

Method	General Characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape – Distance-based – May use mean or medoid (etc.) to represent cluster center – Effective for small- to medium-size data sets
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels) – Cannot correct erroneous merges or splits – May incorporate other techniques like microclustering or consider object "linkages"

Now, let us compare the partitioning clustering algorithm versus hierarchical clustering algorithm, first we will see what is this partitioning methods, what are the general characteristics that is a K means algorithm is a partitioning method. Find mutually exclusive cluster of spherical shape, this partitioning method is a distance based, may use mean or medoid to represent cluster center, effective for small to medium sized dataset.

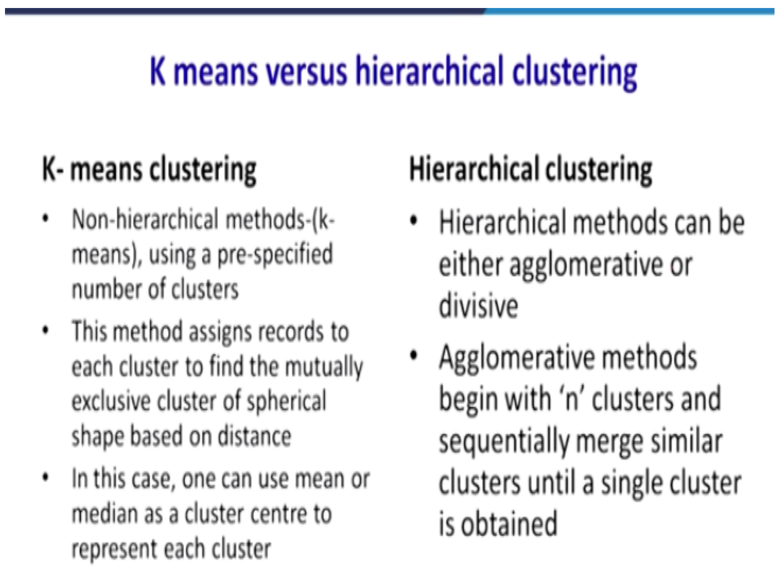
Hierarchical methods; clustering is a hierarchical decomposition at multiple levels, cannot correct erroneous merges or splits, may incorporate other techniques like micro clustering, consumer object linkages.

(Refer Slide Time: 19:39)



Let us compare K means versus hierarchical clustering.

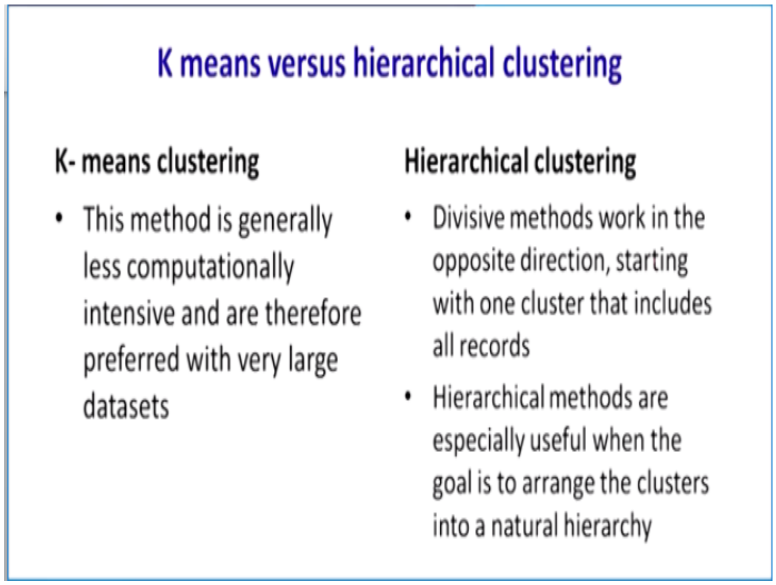
(Refer Slide Time: 19:43)



K means clustering; it is a non-hierarchical method because there will be K means using a pre specified number of clusters. So, when we doing K means clustering in advance we know, how many cluster we are going to have. This method assigns records to each cluster to find the mutually exclusive cluster of spherical shape based on distance. In K mean clustering one can use mean or median as a cluster centre to represent each cluster.

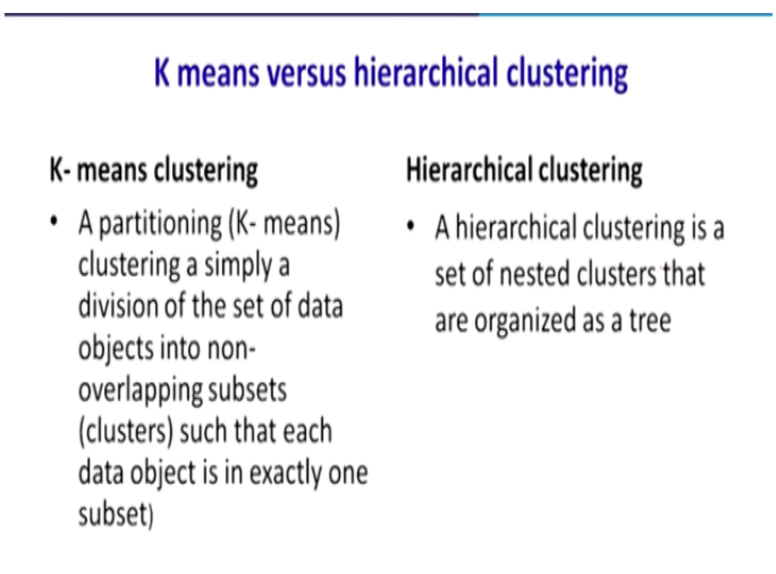
For hierarchical clustering, this method can be either agglomerative or divisive. Agglomerative method begins with n clusters and sequentially merge similar clusters until a single cluster is obtained.

(Refer Slide Time: 20:32)



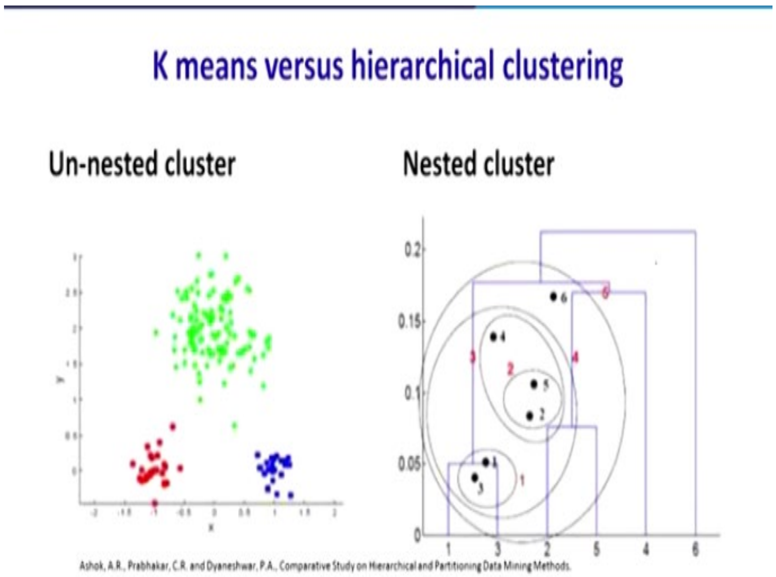
K means clustering methods is generally less computationally intensive and are therefore preferred with very large data set. In hierarchical clustering, divisive methods work in the opposite direction starting with one cluster that includes all the records. Hierarchical methods are especially useful when the goal is to arrange the cluster into a natural hierarchy.

(Refer Slide Time: 21:01)



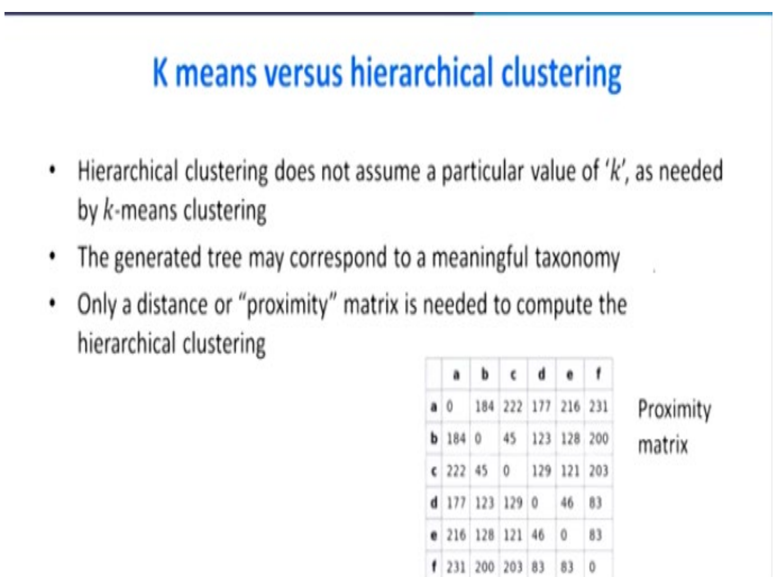
A partitioning that means a K means clustering simply a division of the set of data objects into non overlapping subsets clusters such that each data object is in exactly one subset; a hierarchical clustering is a set of nested clusters that are organised as a tree.

(Refer Slide Time: 21:23)



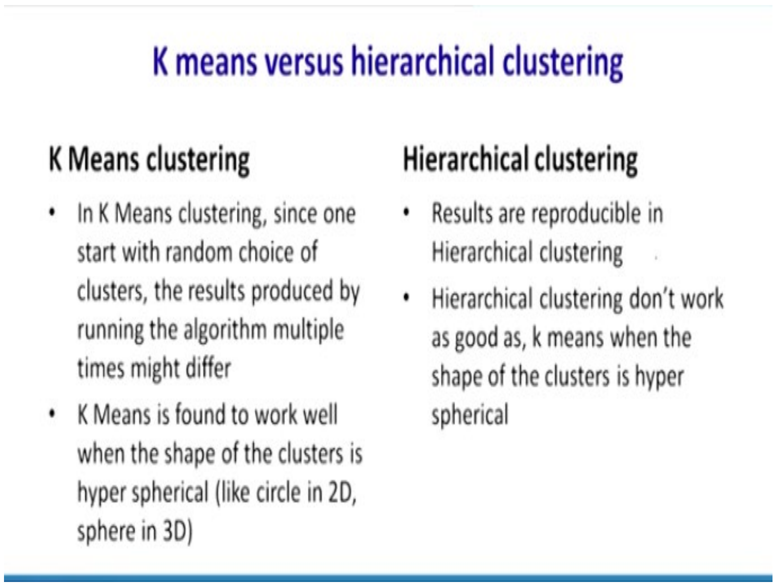
When you look at this picture, the picture shows in the left hand side is un-nested clusters, we can say it is a K means clusters. In the right hand side, the name is called nested cluster, this is nothing but your agglomerative or hierarchical clustering.

(Refer Slide Time: 21:41)



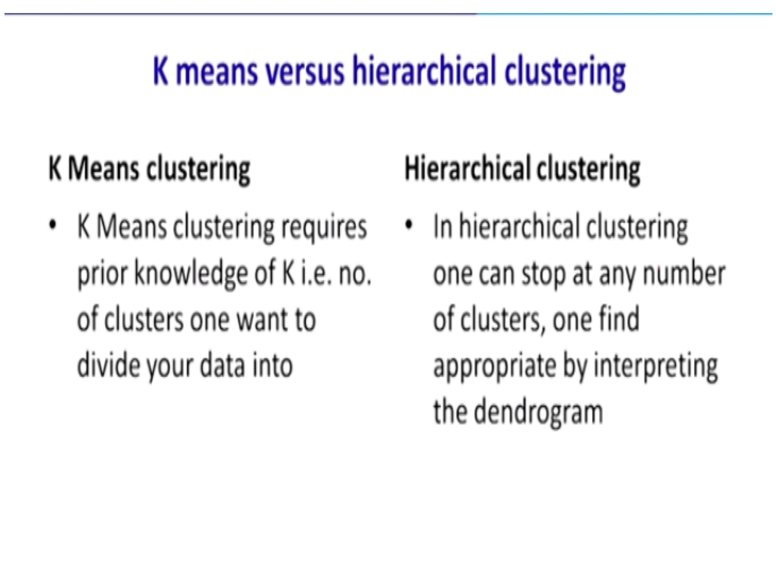
Hierarchical clustering does not assume a particular value of K as needed by K means clustering, the generated tree may correspond to a meaningful taxonomy, only a distance or proximity matrix is needed to compute the hierarchical clustering. This is an example of proximity matrix, see the between a and a, the distance is 0, proximity 0, between b and a, the distance is 184.

(Refer Slide Time: 22:12)



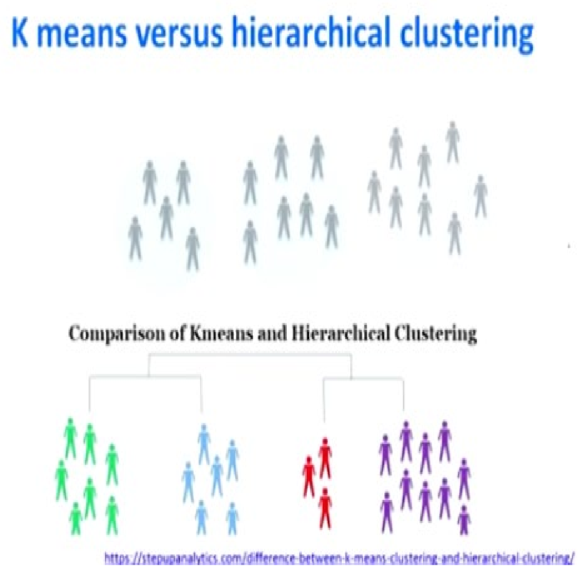
In K means clustering, since one start with random choice of clusters, the result produced by running the algorithm multiple times might differ. K means is found to work well, when the shape of the cluster hyper spherical like circle in 2 dimension, sphere in 3 dimension. In hierarchical clustering, results are reproducible, hierarchical clustering do not work as good as K means, when the shape of the cluster is hyper spherical.

(Refer Slide Time: 22:51)



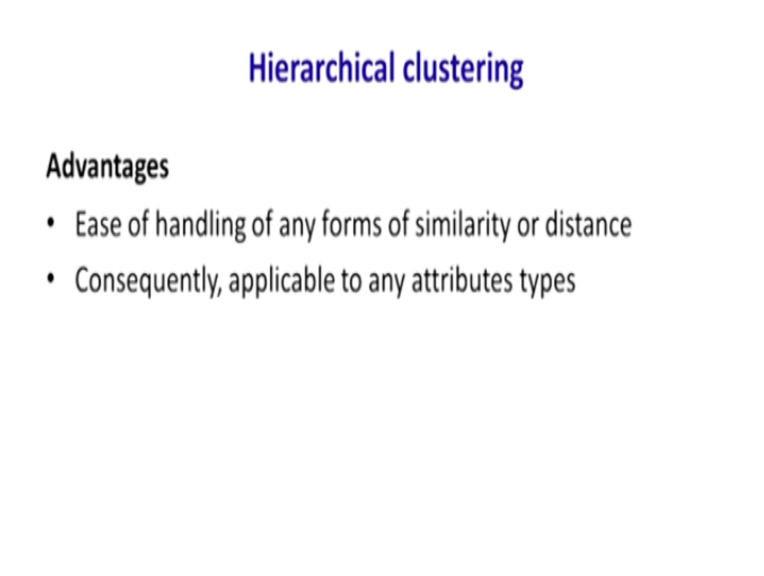
So, the K means clustering suitable for hyper spherical clustering, K means clustering requires prior knowledge of K that is number of clusters one want to divide your data into. In hierarchical clustering, one can stop at any number of clusters, one find appropriate by integrating the dendrogram.

(Refer Slide Time: 23:04)



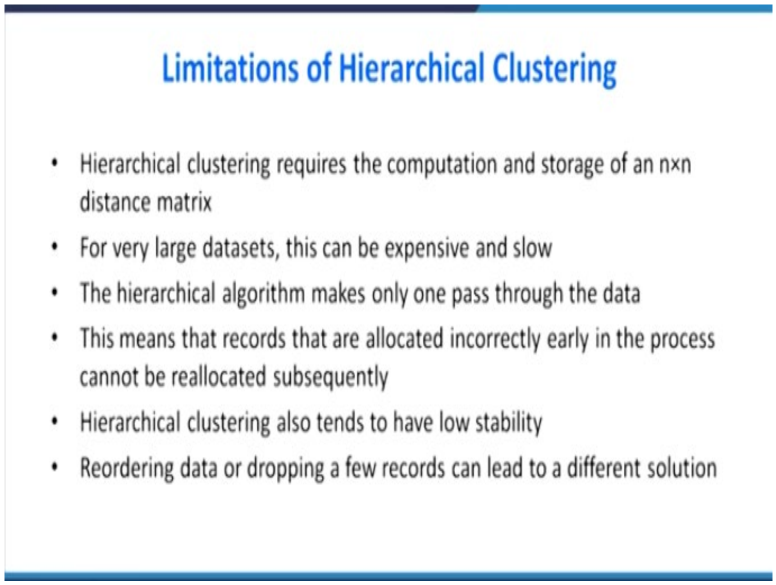
There are 2 pictures, the top one is example of K means clustering where K equal to 3, the bottom one is hierarchical clustering, you see that there is a hierarchy is there, so this is an example of.

(Refer Slide Time: 23:18)



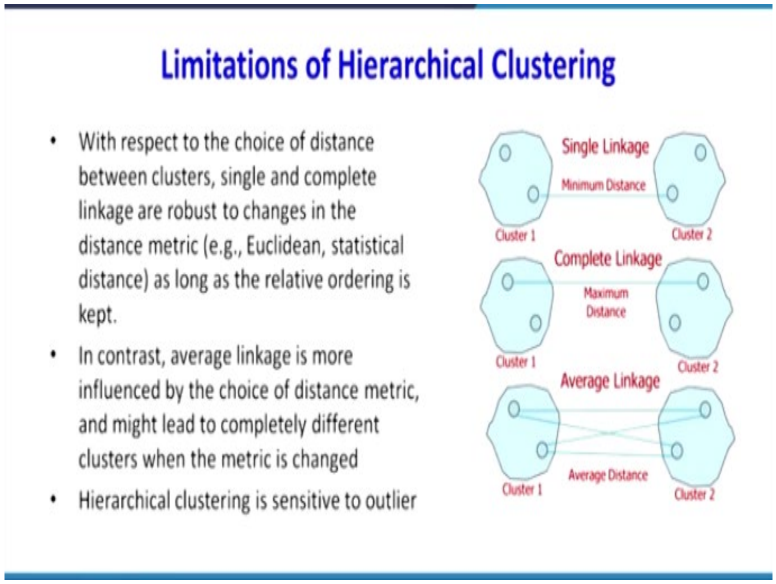
Advantage of hierarchical clustering; ease of handling of any form of similarity or distance, consequently applicable to any attribute types, here attribute is the variable types, it may be interval, it may be ratio, it may be binary or categorical.

(Refer Slide Time: 23:38)



Some of the limitations of hierarchical clustering; hierarchical clustering requires the computation and storage of n cross n distance matrix, here n is number of objects. For very large dataset this can be expensive and slow, the hierarchical algorithm makes only one pass through the data, this means that the records that are allocated incorrectly early in the process cannot be reallocated subsequently. Hierarchical clustering also tends to have low stability, reordering data or dropping a few records can lead to different solution.

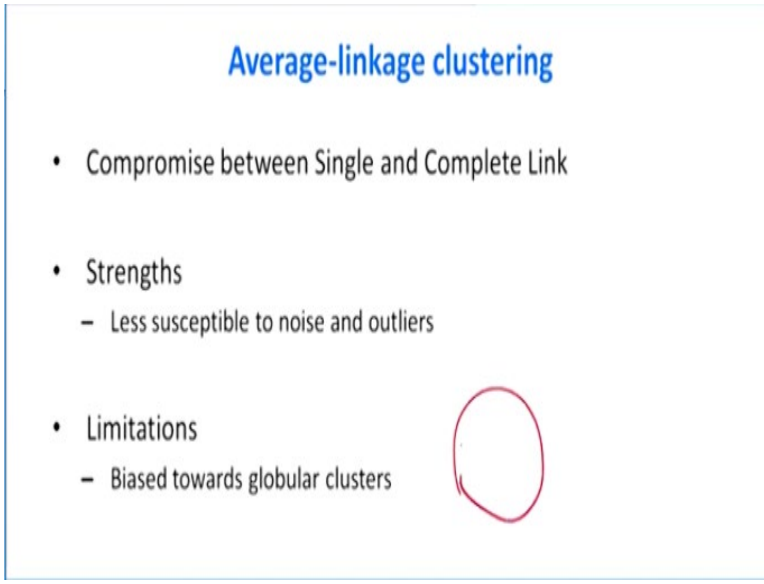
(Refer Slide Time: 24:18)



Limitation of hierarchical clustering; with respect to the choice of distance between clusters, single and complete linkages are robust to changes in the distance metric as long as the relative order is kept. So, what is the example of single linkage when you look at this, there is a cluster 1, cluster 2, so the minimum distance is called single linkage and the distance between the farthest points that is called complete linkage.

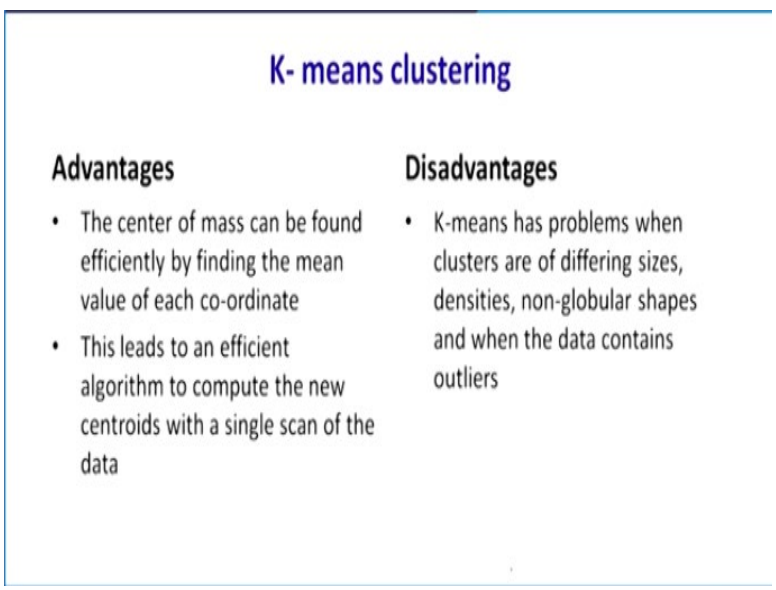
So, if you use these distance measure, then the cluster what you got is very robust, in contrast average linkage is more influenced by choice of distance metrics and might lead to completely different clusters when the metric is changed, hierarchical clustering is sensitive to outlier. If any extreme dataset is there that may provide different kind of clusters.

(Refer Slide Time: 25:14)



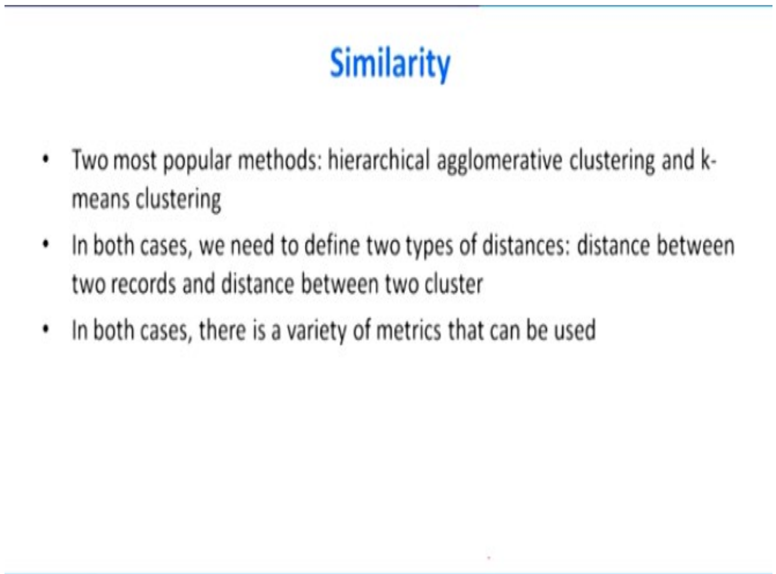
Then, when we will go for average linkage clustering, what is an example of average linkage clustering? This one, you see that, all the distance are connected then we found average. It is a compromise between single and complete link. The strength of average linkage clustering is less susceptible to noise and outliers, the limitations are biased towards globular clusters. So, when you use average linkage clustering, many times, the cluster may be like a spherical shape.

(Refer Slide Time: 25:58)



Now, let us see the advantage of K means clustering, previously we have seen advantage of hierarchical clustering, the advantage of K means clustering is the centre of mass can be found efficiently by finding the mean value of each coordinate, this leads to an efficient algorithm to compute the new centroids with a single scan of data. The disadvantages are K means has problem when the cluster of different sizes, densities, non- globular shapes and when the data contains outliers.

(Refer Slide Time: 26:34)



What is a similarity between hierarchical clustering and K means clustering; 2 most popular method is hierarchical agglomerative clustering and K means clustering, in both cases we need to define 2 types of distance, distance between 2 records and distance between 2 cluster, in both cases, there is a variety of metrics that can be used. In this lecture, I have explained introduction to hierarchical clustering.

Then, I have compared the difference between K means clustering techniques and hierarchical clustering techniques and also I have explained the advantages and disadvantages. In the next lecture, we are going to take one numerical example, with the help of numerical example; I am going to explain how to do a hierarchical clustering, thank you very much.