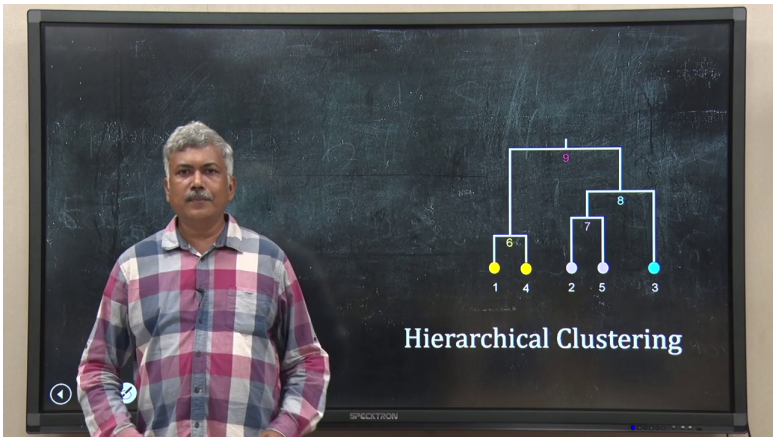
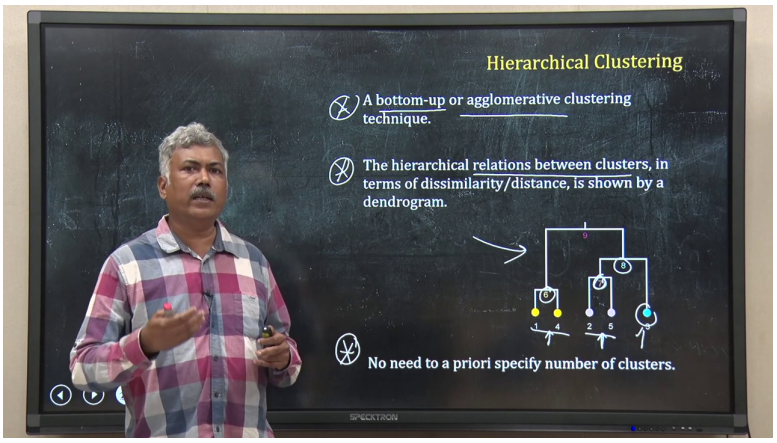
Data Analysis for Biologists
Professor Biplab Bose
Department of Biosciences and Bioengineering
Indian Institute of Technology Guwahati
Lecture 40
Hierarchical Clustering

(Refer Slide Time: 0:33)



Hello, everyone, welcome back. In this lecture, we will learn about hierarchical clustering. Earlier we have learned about k means clustering. Hierarchical clustering is bit different from k means clustering. k means clustering is flat, whereas hierarchical clustering provides me a dendrogram. So, let me see the key features of hierarchical clustering.

(Refer Slide Time: 0:58)



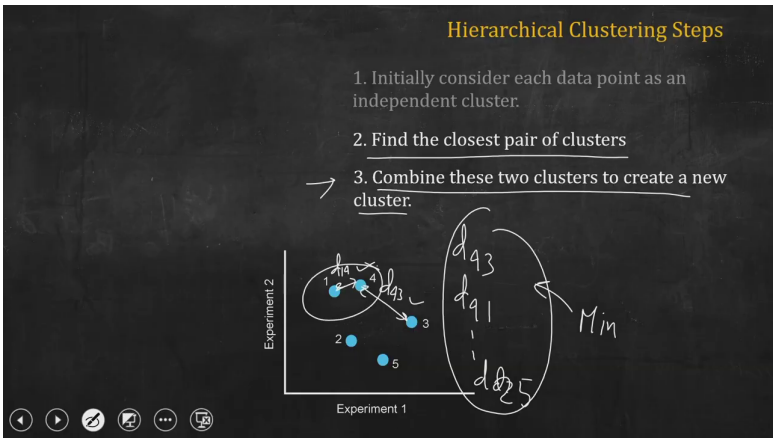
The first one is that hierarchical clustering is agglomerative algorithm or bottom-up algorithm. What do I mean by that? In hierarchical clustering, what we are doing, we initially consider each data point as individual cluster, and then fuse these clusters, these data points to create new cluster, and this is iterated. So, you are taking multiple components, and they are fusing together to create new clusters. So, that is why it is called agglomerative.

Secondly, that is why or the second point is key difference between it and k means clustering. A hierarchical clustering will give me a dendrogram a visual representation, just like this one, that will show the relationship between clusters in terms of their dissimilarity or distances between data points and clusters. For example, in this case, it shows that 1 and 4 form 6, whereas 2 and 5, form the cluster 7, and 3, and 7, this one, and this together form cluster 8.

So, in this way, I get a dendrogram, just like the phylogenetic tree that you may have seen to understand evolution of organism. The third important point for hierarchical clustering is that to perform hierarchical clustering, you do not need to specify the number of clusters in your data. If you remember, in k mean clustering, at the very beginning of running the algorithm, I have to specify the value of k the number of clusters, for example, you may consider k equal to 3, you may consider k equal to 5 or even 10.

Now, the problem in k means clustering is that how do I know how many clusters are there in my data. So, that is what we discussed in our lecture in k means clustering that we perform the same analysis using multiple different values of k and then choose the optimum one. But in hierarchical clustering, we have no such problem, we do not define or specify the number of clusters in the data a proiri before we start the algorithm. So, let us start implementing the algorithm on a data and try to understand how hierarchical clustering works.

(Refer Slide Time: 3:28)



So, I have cooked up a data just like a gene expression experiment, I have two experiment, experiment 1 and experiment 2 and each of these data points is 1 gene maybe this is Gene 3. Usually, there will be hundreds and thousands of data points just for clarity, I have just kept 5 data points and we will build a dendrogram using a hierarchical clustering algorithm on this

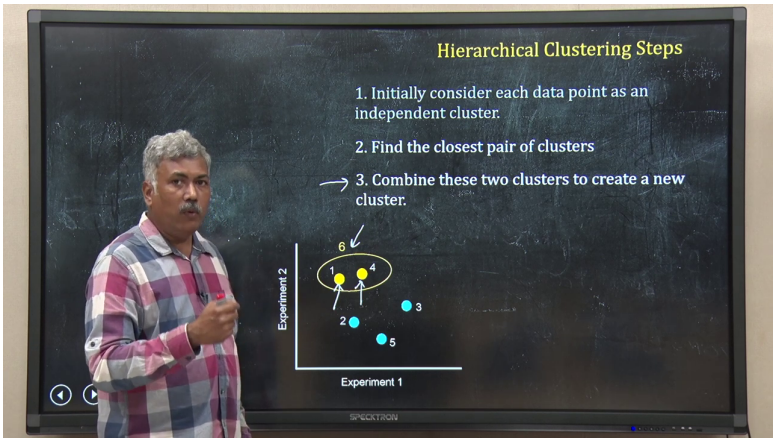
data, the first step as it is agglomerative method is to consider that each data point is an independent cluster.

So, how many data points I have 5 data points, that means the number of clusters will be 5, all these are independent clusters. This is the first step. Next what we do, we find the closest pair of clusters, what do I mean by that? So, I have 5 clusters, in this state actually each cluster is nothing but the original data points. So, I have 5 clusters and we take any pair of any 2 of them and find the distance between them. So, for example, you may take 4 and 3 and calculate the distance there maybe d 4 3.

We have learned many distance measures earlier, Manhattan distance, Euclidean distance, Mahalanobis2 distance, you have to choose which distance measure you want to use depending upon the problem at hand. So, suppose we are using Euclidean distance, which is most commonly used for this type of analysis, so we calculate the distance between 4 and 3, we calculate distance between 4 and 1. And all the pair for example, d 2 and 5. So, all pairwise distance we measure, and then we find out the minimum of those.

So, you find the minimum of this, that will give me the closest pair. So, we have to find the closest pair. So, if you look at this data, visually, it is apparent that 1 and 4 may be the closest one. So, d 1 4, may be the smallest distance. So, now I have to move to the third step, what is the third step? Third step says that combine these 2 clusters to create a new cluster. So, I have chosen the closest pair in this data, 1 and 4 is the closest pair. So, I will fuse these together to create a new cluster. And that is what I have shown in this diagram.

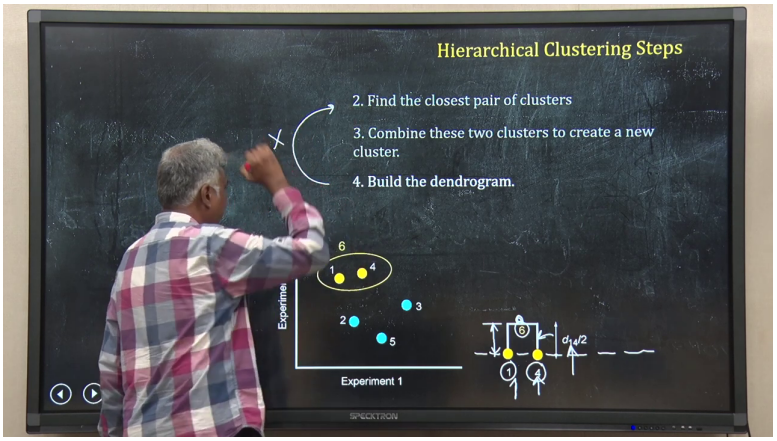
(Refer Slide Time: 6:06)



So, I have fused 1 and 4 as my third step says, to create a new cluster, cluster 6, till now, I have 5 data points, they are all independent clusters so I had 5 clusters. Now, I have created a

new cluster so, this is the sixth cluster. Now, what I have to do, I have to move to the fourth step, where I will draw the dendrogram.

(Refer Slide Time: 6:31)

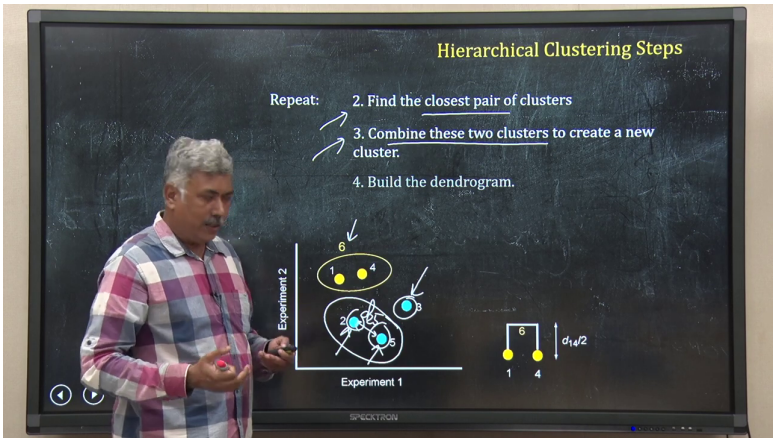


So, to explain the dendrogram it is better if I draw a horizontal line. Remember, in this horizontal line, we will always put the original data points. For example, in this case, my original data point is 1 and 4. So, these yellow dots representing 1 and 4 is on the horizontal line. And then I connect them by 2 arms to represent that they have got fused to create 6. Just like a phylogenetic tree, for example, organism 1 and organism 4 are originating from ancestor 6 something like that.

So, I have 2 branches coming out of cluster 6, and they are joining to 1 and 4. Now, what should be a distance of this one, height of this branch? We use the convention; we scale this height as per the value of the half of the pairwise distance. So, the pairwise distance between 1 and 4 is d 1 4, you half that divided by 2, and you scale that to scale the height of this branch, just like you scale when you draw a map. So, you are scaling it.

Now, what we have to do we have done the first step, we have created the first dendrogram we have to build upon it. So, to build upon it, we have to repeat this process multiple times.

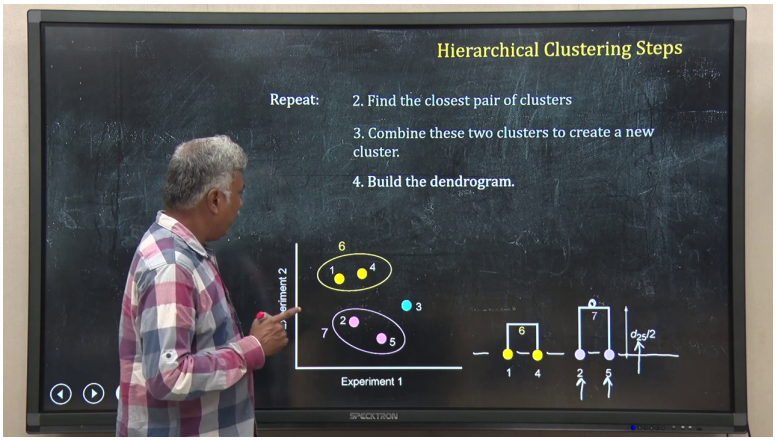
(Refer Slide Time: 8:03)



So, let us repeat, the first repetition is for repetition of 2, again find the closest pair of clusters. Now remember, we have fused 1 and 4. So, 1 and 4 does not do not exist for us anymore, what we are left with we are left with cluster 6, cluster 3, cluster 2 and cluster 5. So, we have to calculate the pairwise distance between these 4 clusters. If you remember 3, 5, 2 are itself clustered at the very beginning. So, they are still independent clusters.

So, we have to take these 4 clusters 6, 3, 2 and 5 and find a pairwise distance between them and find the closest pair the way we have done earlier. So, for example, in this diagram, if you visually observe, you will see possibly this pair the distance d 2 5, maybe the least one, so the closest pair is cluster 2 and cluster 5. So, what I will do, I will move to third step, I will combine these 2 clusters to create a new cluster just like we did earlier. So, now I will combine 2 and 5 to create a new cluster, that is what I have done here.

(Refer Slide Time: 9:24)

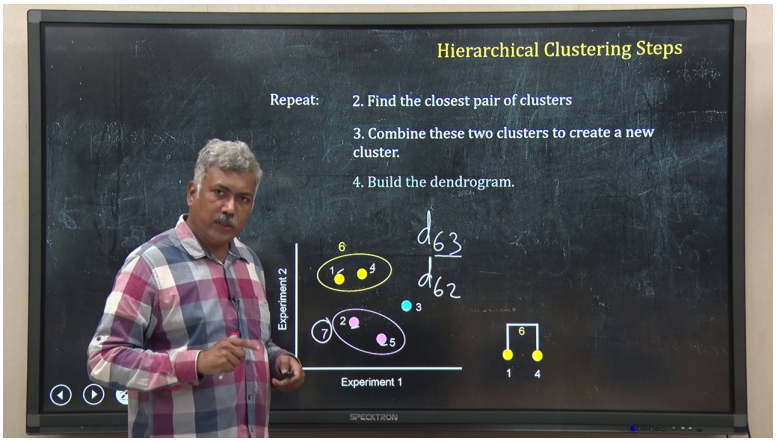


So, I have got a new cluster 7, 2 and 5 now disappears. So, I have got a new cluster by combining cluster 2 and cluster 5 to create cluster 7. Now I will add that to my dendrogram.

So, let us see that again just for clarity, let me draw this horizontal line. And as I said earlier, all my original data points should be on this horizontal line. So, I have all those data points on the horizontal line, and now this is 2 and 5. They have fused combined to give me the seventh cluster. So, this is 7th cluster. So, 7 is connected to 2 and 5 by two branches, what will be the height of these branches? Again, the height will be half of the distance between 2 and 5.

So, d 2 5 divided by 2, that way I have done for 1 and 4 when I was building for 6. So, that is how I have drawn the dendrogram or added the new cluster to the dendrogram. So, remember, I have to keep on repeating this method.

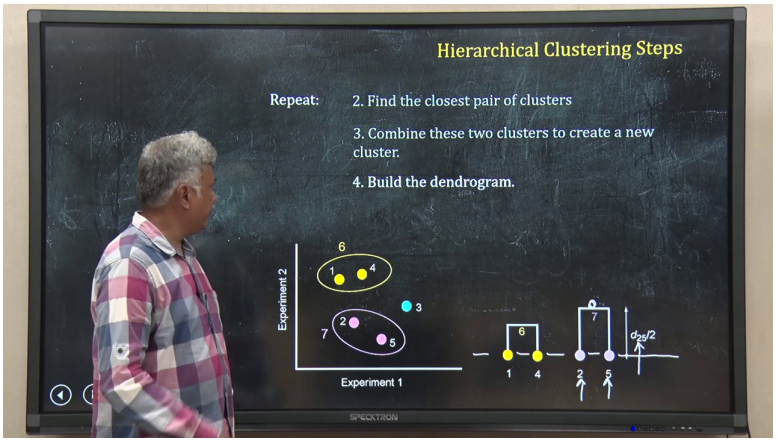
(Refer Slide Time: 10:45)



Now, before I repeat, you may be wondering that earlier, what we have done, we have calculated distances between 1, 4, 3, 2 and 5 pairwise distances. But in this particular state, I have calculated distance between 6 and 3, distance between 6 and 2, and so on. Now, 6 itself is a cluster, it is not a data point. So, cluster 6 has multiple data points. So, this cluster is a set of objects. So, how do I measure distance between a cluster and a data point? I know how to measure distance between 2 data points, I know how to calculate that for you using Euclidean distance measure.

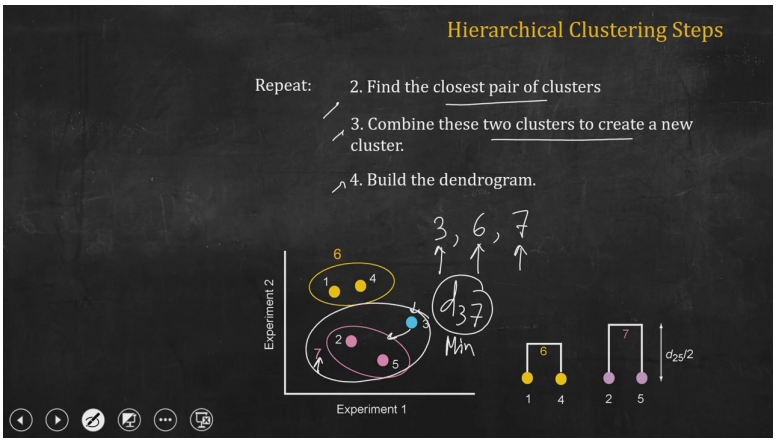
But how should I calculate distance between 2 clusters, or one cluster and one data point? Because the cluster does not have one object, it may have multiple objects. That is a valid question. And I will come back to that how to do this.

(Refer Slide Time: 11:35)



For the time being, you can see that we know that and that is what we are implemented here. For example, we calculated distance between 6 and 3, 6, and 5, and so on. So, now let us repeat the whole algorithm again, and proceed further.

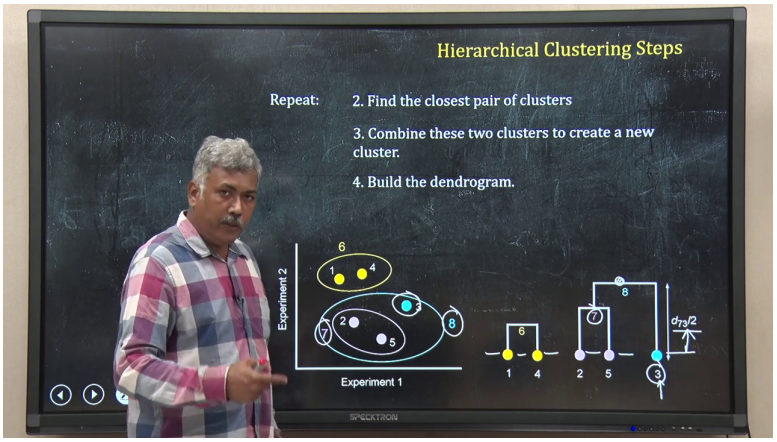
(Refer Slide Time: 11:51)



So, I have to repeat again, so I have to repeat this whole thing for 2, 3, and 4. So, now, if I look into this, I can easily see that 3, and 7. So, I have how many clusters we have, now I have cluster 3, cluster 6, and cluster 7, I have to calculate the pairwise distance between them. Because remember, 2 and 5 does not exist anymore. For me, neither 1, 4 exists for me. So, I have cluster 3, which is the original data point, cluster 6, and cluster 7. And I have to calculate the pairwise distance between these 3 clusters.

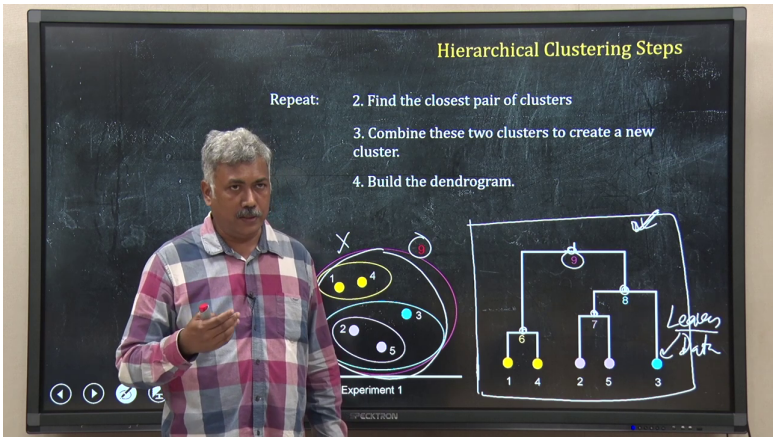
And then I have to combine the closest pair. So, I will combine these 2 closest pair to create a new cluster and add that to the dendrogram. So, if you visually look at it, possibly 3 and 7, d 3 and 7, or d 7 3, whatever you say, is the closest or maybe the minimum one. So, that means I have to fuse these 2, to create a new cluster.

(Refer Slide Time: 13:00)



And that is what I have shown in this slide. So, I have fused cluster 3, and cluster 7, which itself made up of 2 and 5, to create a new cluster, cluster 8. And I have put that clustering on the dendrogram. Just to explain, let me draw the horizontal line, 3 has been placed on this horizontal line, and 3 has fused with 7 to create a new cluster 8. So, from 8, 2 branches, 2 arms go to a 3 and 7. Again, what is the scaled height of these branches two arms, that will be equal to the half of the distance between cluster 7 and cluster 3. So, this iteration is kept on doing, we will keep on doing that.

(Refer Slide Time: 13:49)



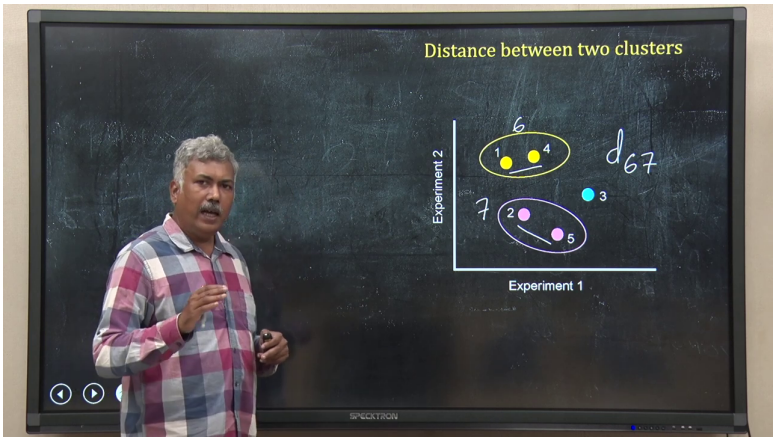
And then eventually, in the next step, I get all data points, and now in this new cluster, cluster 9, and I add that to the dendrogram. So, I am done, because all data points are now are part of this cluster 9, there is no other cluster left, I am left with only one cluster. So, there is nothing more to fuse. So, my algorithm will stop here, what I have got, I have got this dendrogram which shows a hierarchical relationship between my data point and the clusters.

Usually, you will report this one we do not report this one in our reports and articles, because this one I have drawn just to explain these dendrogram you will report. So, what this dendrogram says? With leaves the leaves of this dendrogram these are the leaves. Leaves of the dendrogram are all original data. So, I have 5 leaves, 5 data points.

And the nodes, internal nodes are the clusters and this whole diagram which is scaled with scaled arm, each of these arms represent the distance or dissimilarity between different point data points and clusters. So, looking at this cluster, this diagram, I can easily understand which cluster is close to or dissimilar to which cluster.

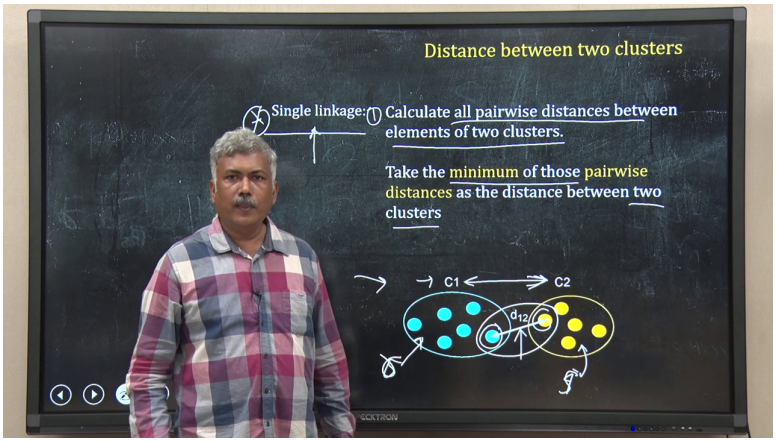
Now, let me move to that unanswered question. In this algorithm I am measuring distances between clusters. In one our previous lecture, we have learned distances between data point, how should I measure now, the distances between clusters? There are many methods, I will discuss few of them.

(Refer Slide Time: 15:41)



Let me explain what we have to do here. So, for in our problem, we had to measure the distance between 6 and 7 and 3. So how do I measure distance between 6 and 7, when 6 and 7 has 2 elements both. So, that is our problem. And I will discuss some of the distance measure for clusters. As a technical term, usually the distance between clusters is usually called the linkage. So, there are many types of linkages are usually used in data analysis, I will discuss 4 of them.

(Refer Slide Time: 16:17)



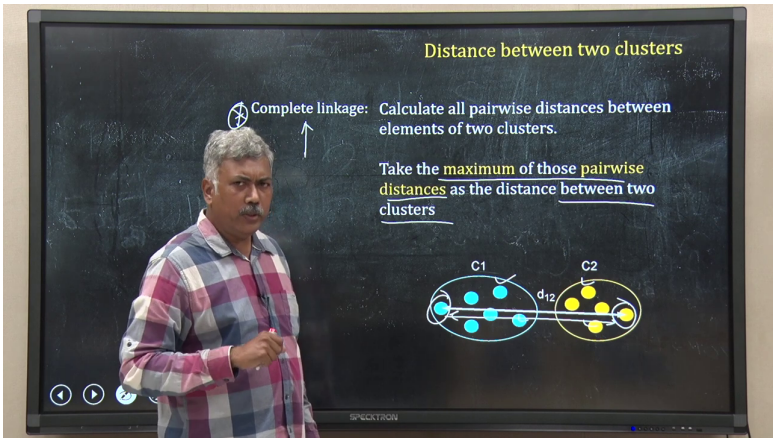
The first one is single linkage. What are you doing in this case, just look at this diagram, I have 2 clusters, cluster 1 and cluster 2, C 1 and C 2. Cluster 1 has 5 objects or 5 elements, whereas this one has 4 objects or 4, I'm sorry, this one has 5, this one has 6. So, now, what I have to do, the first step I have to do is to calculate all pairwise distance between elements of these 2 clusters.

What do I mean by that, in cluster 1, I have 6 objects, whereas in cluster 2 I have 5 objects. So, take one object from cluster 1 and pair it with another object of cluster 2. In this way you create all possible pairs of objects between Cluster 1 and Cluster 2 and calculate the distances. Now, these objects are your original data points.

So, now, you can use those distance measures that we have learned earlier, for example, Euclidean distance or Mahalanobis distance or correlation distance whatever you want to use, you can use that those definition those metrics to calculate the distance between any 2 data point pair that you have picked up from these 2 clusters.

So, you make a list, all the pairs from C 1 and C 2 and their distances and then you pick the minimum distance, minimum of these pairwise distances and that distance you will consider as the distance between these 2 clusters. For example, in this case, I have calculated all possible pairwise distances, and I find that this and this forms a pair and their distance is the minimum distance, so, this distance will be considered as the distance between cluster 1 and cluster 2 and that will be called single linkage. Now, let me move to the second option that you have.

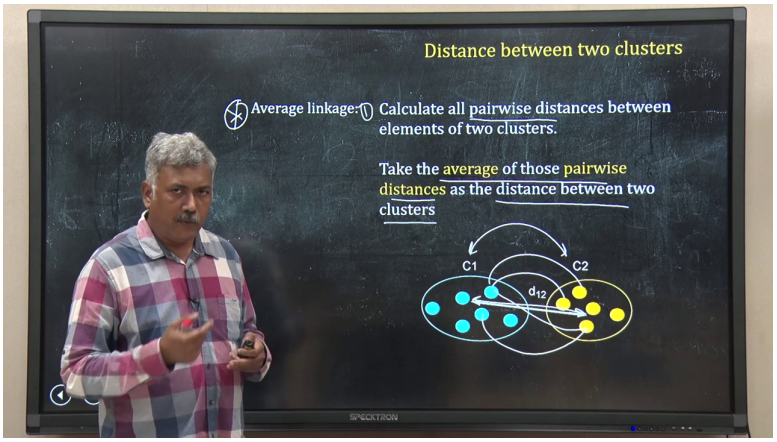
(Refer Slide Time: 18:25)



That is called complete linkage. Now procedure is almost same, you again create all the pairs taking one object from C 1, cluster 1 and another object from cluster 2. So, you make a pairwise combination and then you calculate distance between these pairs. So, you calculate distance between this one and this one, this one this one, so, on. So, you again make a list, in single linkage you have taken the least value, the pair which are close to each other. Now, in case of complete linkage, you choose the maximum of those pairwise distances.

So, you calculate the maximum of those pairwise distances, and that distance will be considered as the distance between my clusters, and that will be called complete linkage. Look at the diagram, in this case, this and this forms a pair. And the distance between these 2 objects is the largest one in all these pairwise distances in this particular example. So, this distance is now the complete linkage or the distance between these 2 clusters, cluster 1 and cluster 2.

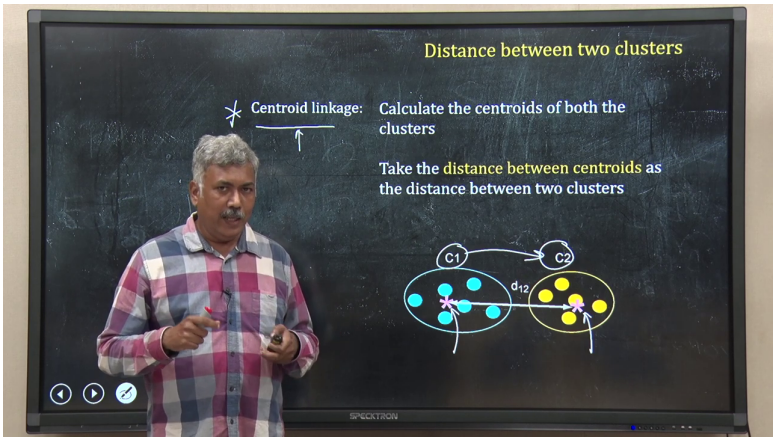
(Refer Slide Time: 19:41)



Let us look into the third option that is called average linkage. And as the name suggests, essentially will do averaging of pairwise distance. So, again you calculate all the pairwise distances between them by taking one object from C 1 and C 2. So, you make all the pairwise combination and calculate the distance between all these objectives from C 1 and C 2, and then you take the average of those pairwise distances, and that distance you will consider as the distance between clusters.

So, in this case what I have to do, I have to calculate distance between these, distance between these, distance between these ones, distance between these ones, all pairwise distances, and then I will make a list table and take the arithmetic mean of that and suppose, that arithmetic mean is this line, arrowhead, double arrow headed line I have shown. So, that distance is the average linkage or the average distance between cluster 1 and cluster 2.

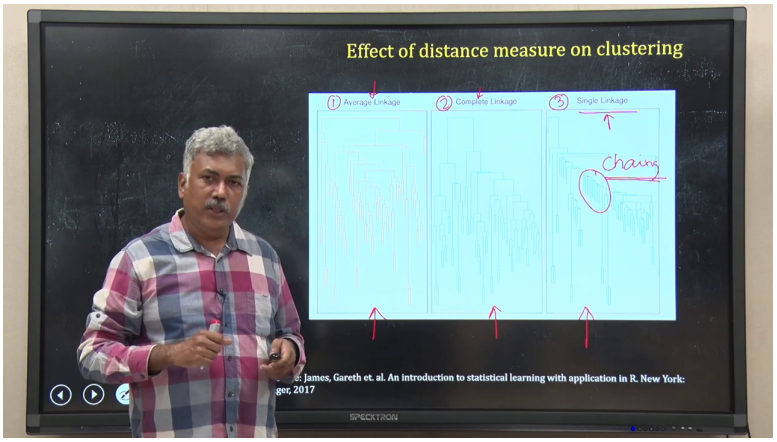
(Refer Slide Time: 20:48)



The last distance measured is bit different from the other 3 that we discussed till now, that is called centroid or centroid linkage, what you are doing here you do not calculate any pairwise distance, what you do you take cluster 1 the data of that and calculate the centroid of that, that means the mean position center of that cluster. So, suppose that is this pink star, you do the same thing for cluster 2.

So, that is that pink star, now you calculate the distance between these 2 centroids and that distance you consider as the distance between cluster 1 and cluster 2 that will be called centroid distance or the centroid linkage. So, you can choose any of these 4 to implement your hierarchical algorithm for clustering. Now, we have to remember that this choice of linkage or choice of distance between cluster will affect the final outcome of our clustering algorithm.

(Refer Slide Time: 21:52)

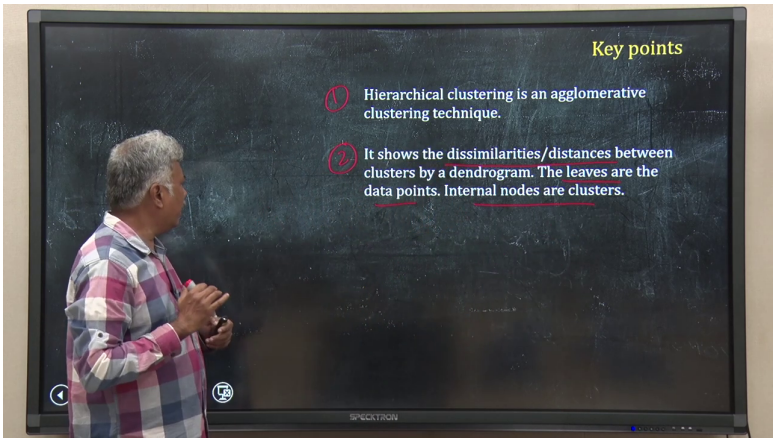


And I have taken the diagram from a book that shows it very nicely. So, let me change the color of my pen. So, this is average linkage they have used on a data set on the same data set they have used the complete linkage and also used the single linkage measure to draw the dendrogram the hierarchical clustering they perform hierarchical clustering, and you can easily see that these 3 diagrams these 3 dendrogram are different. So, that means your choice of linkage that you use in your algorithm will affect the final outcome of your hierarchical clustering.

And if you look at it single linkage has completely unbalanced dendrogram, what we have got here is something called chaining. This happens because, you remember that in single linkage what we are doing, we are calculating the pairwise distances between objects in cluster 1 and cluster 2 and taking the shortest one. Now, that means, if I have quite heterogeneous 2 clusters, which are quite far, but one of the data points are close to each other, then these 2 clusters will come together they will belong to the same cluster.

So, in this way, you get this type of chaining which essentially skews the whole dendrogram. So, usually that is why a single linkage measure is not used mostly in biology, in most of the time biologists use average linkage and sometimes complete linkage because they provide me a balanced dendrogram. That is all for this lecture. To learn about hierarchical clustering, let me jot down what we have learned in this.

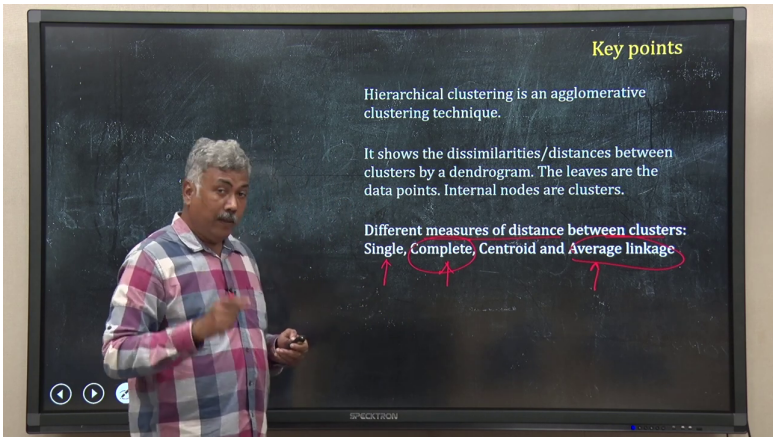
(Refer Slide Time: 23:43)



In this lecture, we have learned about the algorithm the basic steps involved in hierarchical clustering, you have to remember that hierarchical clustering is actually agglomerative method. So, you take all data points that are independent cluster, and then you keep on fusing them to create a bigger cluster.

The second important thing is that the hierarchical clustering gives me a dendrogram a visual representation of data where the dendrogram shows which data point is close to which one which cluster is close to which one. So, it shows a visual representation of similarity and distances between clusters and in this dendrogram the leaves are the original data points and the nodes are the internal clusters.

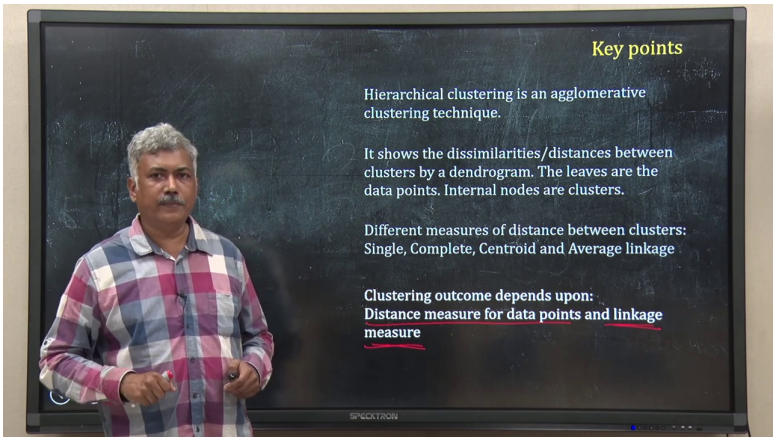
(Refer Slide Time: 24:35)



The third thing that we have learned is that we have learned about different distance measure between cluster this is a bit different from distance measure of data points. So, we have learned about single linkage, complete linkage, average and centroid linkage. And as I

mentioned, most of the time, people use average linkage and complete linkages because they provide us quite a balanced dendrogram.

(Refer Slide Time: 25:01)



But the last not the least the most important thing as I have shown by one data set, that linkage measure that you are using can actually create completely different outcome. So, you have to judiciously choose the linkage measure. At the same time for any clustering algorithm, not just for hierarchical clustering, any clustering algorithm that you use, your distance measure for data points will also affect the final outcome.

So, whenever you are performing hierarchical clustering or any other clustering, you have to be very careful about what distance measure you should use to calculate the similarity or dissimilarity between 2 objects or 2 data points. Euclidean distance is usually the most common choice, but there are many other distance measures we have discussed earlier and you have to judiciously choose them.

And because based on your understanding of the biology and the physical principle that which measure will give physically meaningful meaning of dissimilarity or similarity between 2 objects, and you have to choose that and specifically for hierarchical clustering again, you have to judiciously choose that linkage measure. That is all for this lecture. See you in the next one. Till then happy learning.