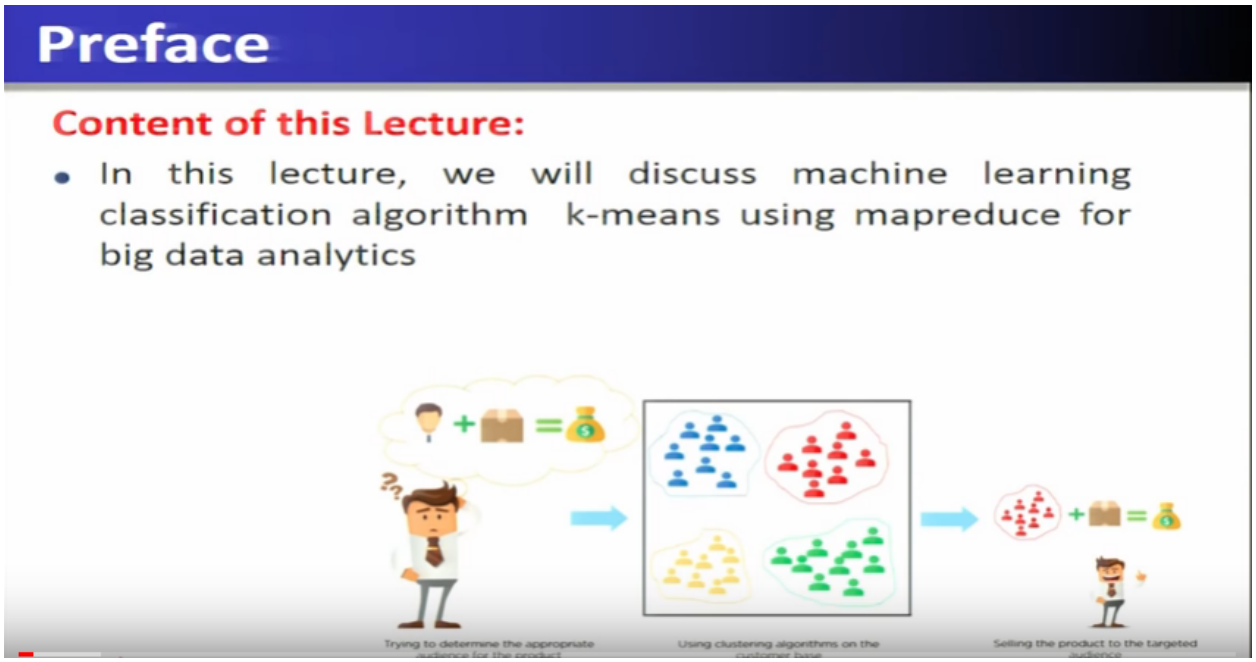
Lecture-25
Machine Learning Algorithm K-means using Map Reduce for Big Data Analytics

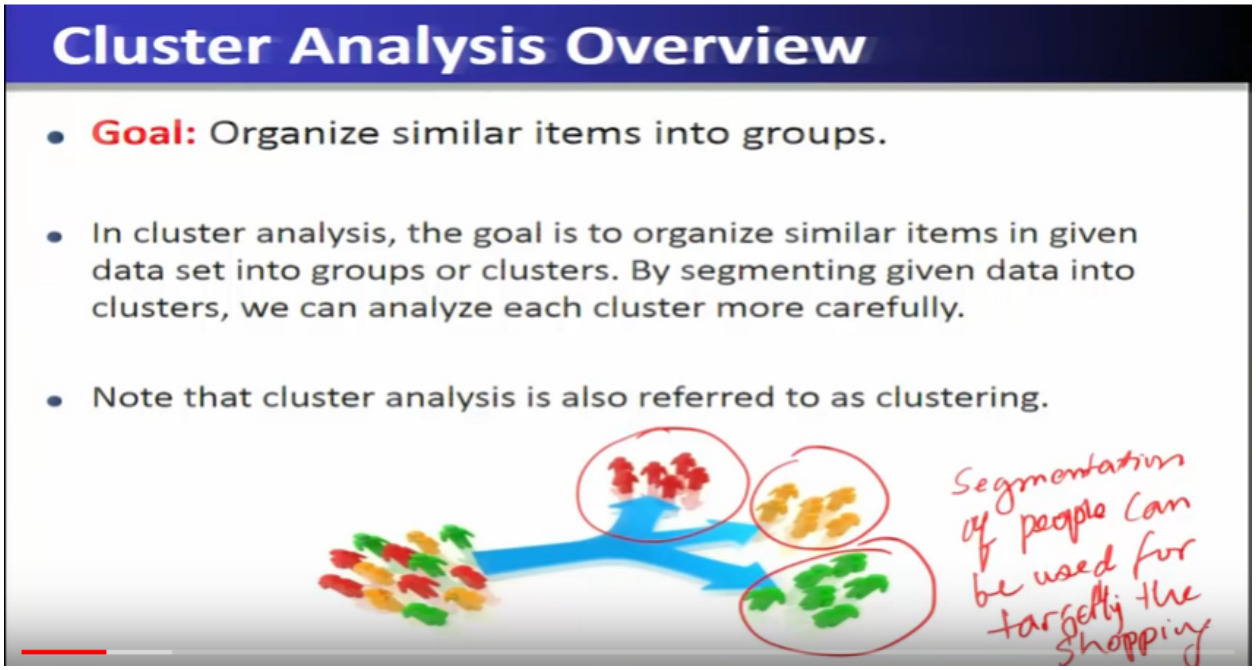
Machine Learning Algorithm, Parallel k-means, using Map Reduce, for big data analytics.

Refer slide time: (0:21)



Preface, content of this lecture: In this lecture, we will discuss machine learning algorithm, k-means using Map Reduce, for big data analytics.

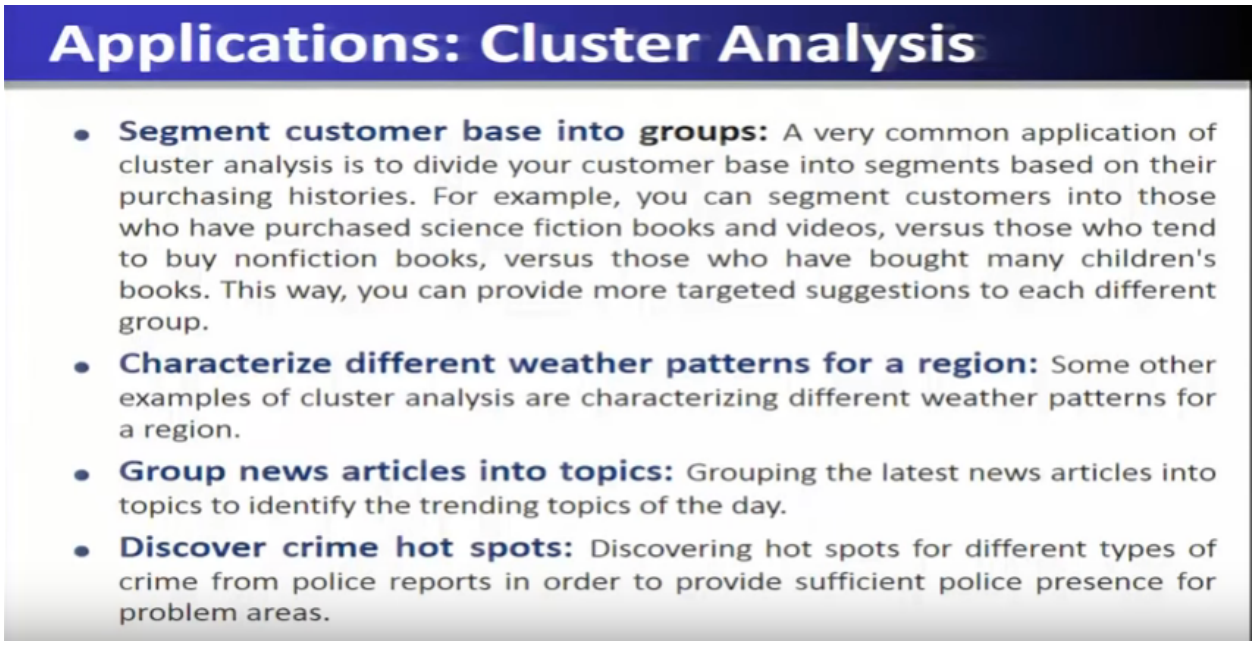
Refer slide time: (0:33)



Before we go, into the deeper of k-means algorithm, let us, understand some of the preliminaries, which are required to understand this k-means, big data analytics algorithm. Now, the term which is called a, 'Cluster Analysis', is a widely known, term in the machine learning framework and the cluster analysis,

goal is to organize, the similar items in a given data set into the groups or they are also called a, ‘Cluster’. By segmenting this way, the given data into the clusters, we can gain the insight into the data set, about the clusters or the groups and this is called the, ‘Segmentation’. And we can analyze each cluster or a group with more careful inside. Now, this cluster analysis is also known as, clustering. For example, here we can see here, there is a group of people. Now, as far as these group of people, can be segmented into different groups, based on their purchasing capacity or are maybe that, different interest groups and soon. So, here we can see that all the interest groups, of a particular interest, they are classified into four different classes. And this can be used; this segmented can be used, for targeting the, the shopping businesses.

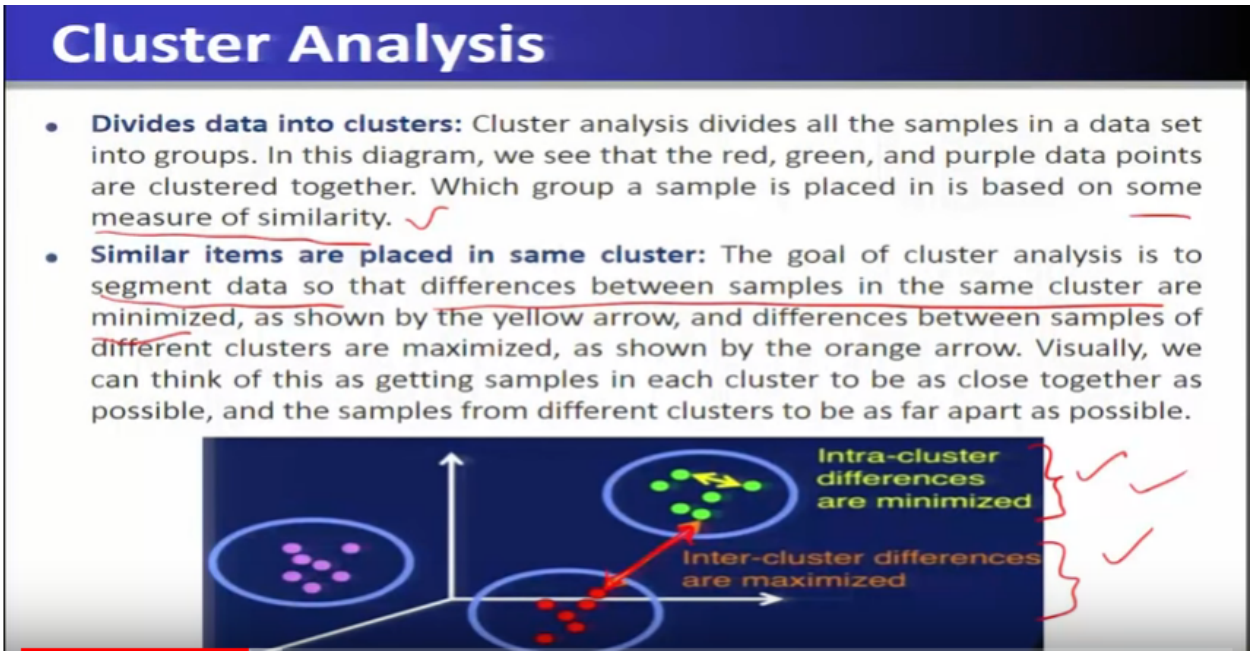
Refer slide time: (2:48)



Let us see, some other applications of this cluster analysis. Here, the first such application, which is well known is called, ‘Segment Cluster Base’, into various groups. So, hereafter identifying these particular customers into different groups, based on their purchasing histories, these segment customers into, who purchased? Which of these items can be useful for different business applications, for targeting advertisements or for targeting different promotions? So that, their businesses can be directly, dealt with the interested, set of people with their, known preferences. Now, similarly the other application is to characterize, different weather patterns, for a region. So, this particular way, the cluster analysis can characterize different, weather patterns of a region and therefore, different disaster management or different applications can be, based on this kind of clustering or segmentation. Similarly as far as, the group news articles, can also be used, so that, they can be identified that means, different news articles, can be grouped according to that trending topics of the day. and also, to discover the crime hotspots ,it is possible that, to identify different type of crimes and they are, located and using cluster analysis,

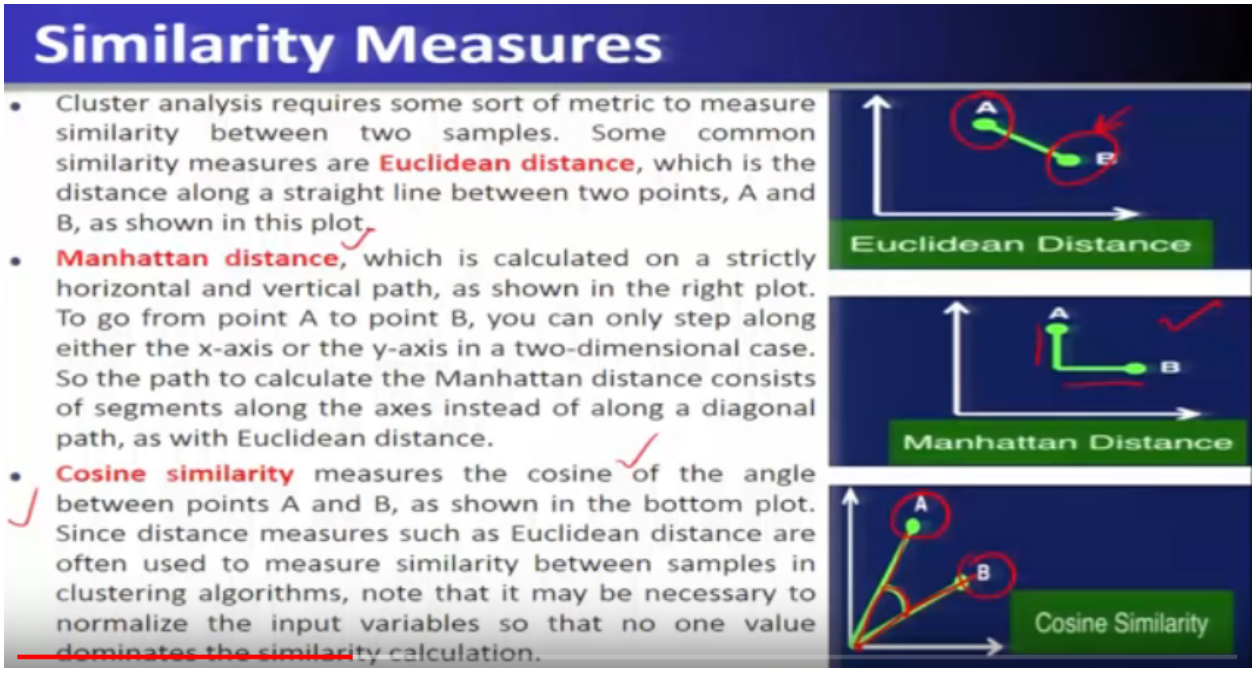
identifies the hottest part so that, the sufficient number of police personal, scan be, can be put on their presence, to solve any problem, related to that spot or for that, crime spot.

Refer slide time: (4:49)



Similarly, we can divide the data into the cluster. So, cluster analysis divides, the samples of in two groups. And in this particular picture, we can see here, there are three different groups which are shown, with their members as, red, green, purple, data points. Now, they are grouped together, based on some measure, which is called a, 'Similarity Measure'. We are going to see, what do you mean by the similarity measures? So, this particular clustering or the cluster analysis is, with reference to some measure of similarity, we are going to see about, this particular term, in more detail in further slides. Now, in this particular cluster analysis, we are going to keep the similar items, together in the same cluster. So, the goal of cluster analysis is, to segment the data. So that, the, so that, the difference is between the, between the samples in the same cluster are minimized. So here, we can see, by an example, that for example, all the green ones, if we see, the differences in terms of measuring, the similarities. So, they are basically having a very little difference, hence they are grouped together in a form of a cluster. So this way, the differences between the samples, of different clusters are maximized, here we can see by this particular arrow. Now, the differences between the elements, across different clusters are maximized. So, there will be a different separation and different clusters, can be easily identifiable. so that means, in a nutshell, we can summarize by saying that, the similar items are placed, are placed in the same cluster .that is their intra cluster, differences are minimum and also, across different clusters, what we can say that, the inter cluster differences are to be maximized. and when these two things, these two conditions are met, then we are done with that cluster and clustering analysis is about, dividing into these clusters, so that similar item can be placed together and the, items which are in different clusters, they are separated with the maximum differences.

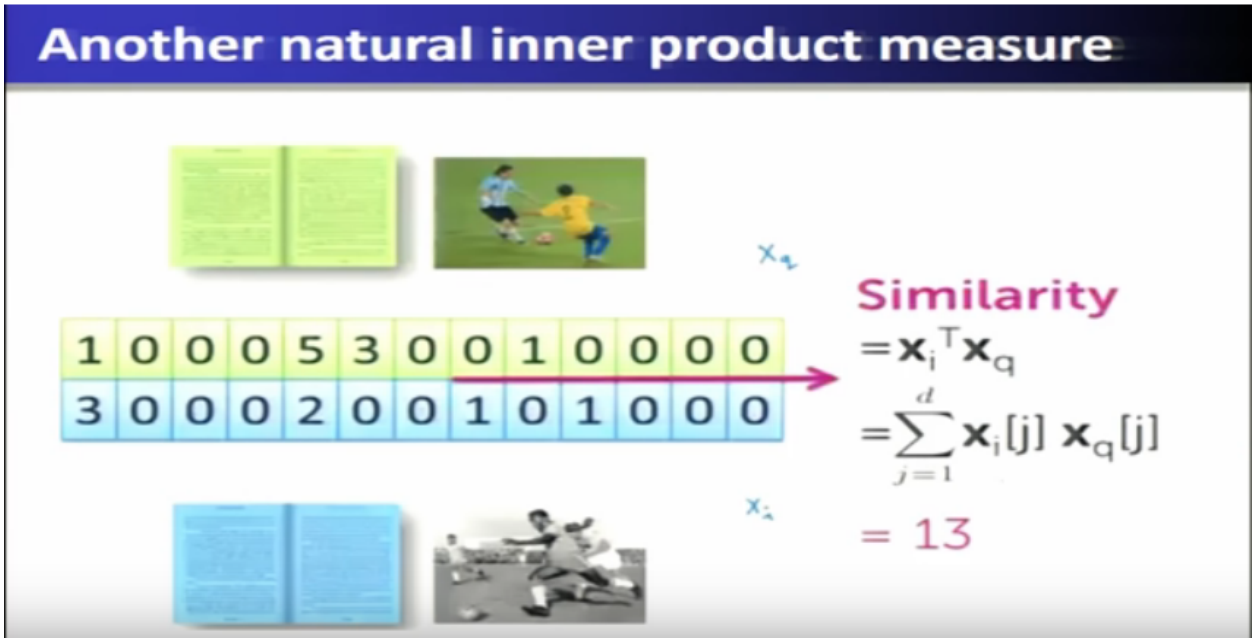
Refer slide time: (7:24)



Now, let us see the similarity measures, which is the, which is the main parameter, by which this particular clustering or cluster analysis or is particularly performed. So this is called, 'Similarity measures'. So here, cluster analysis required, some sort of measures, to some sort of metric, to measure similarity, between two samples. Some common similarity measures are, equilibrium distance, which is the distance along the straight line, between two points, A and B, which are shown over here. So, this is called, 'Equilibrium Distance', here, the measure is about the de, the distance between the any two points, which are shown here, by the straight line. So, there is a equation and equilibrium distance, follows the triangle inequality properties, of mathematics and this kind of measure is, used here, for doing the clustering and in the cluster analysis. Another similarity measure is called the, 'Manhattan Distance', which is calculated, on strictly following the horizontal or on the vertical path, as shown here in this particular diagram. Now, from to go from, A to B it has to follow either the horizontal or the vertical path, it cannot follow the diagonal path, in this two-dimensional space ,that is the case, that it is called, Manhattan distance and using this particular similarity measure, you can also do the clustering ,in certain applications. Finally there is another dissimilarity measure, which is also very popular, besides all these is called, 'Cosine Similarities'. So, in the cosine similarities, there is a measure of cosine, of the angle between two points. For example, the points are A and B and if with the reference, you can plot, you can draw, an angle and the cosine of this angle is, basically called a cosine similarity. which is shown here, in this particular diagram. Now, you can see that, Euclidean distances are used to measure, the similarity between, the samples in the clustering algorithm and it may be necessary, to normalize this input. Now, so that no, no one value will dominate, the similarity calculations and here, as far as, cosine similarity is concerned on, this kind of similarity is used in some of the applications. for example, the documents which are closely related to some topic, their cosine similarity will be less and they are, they are, they are

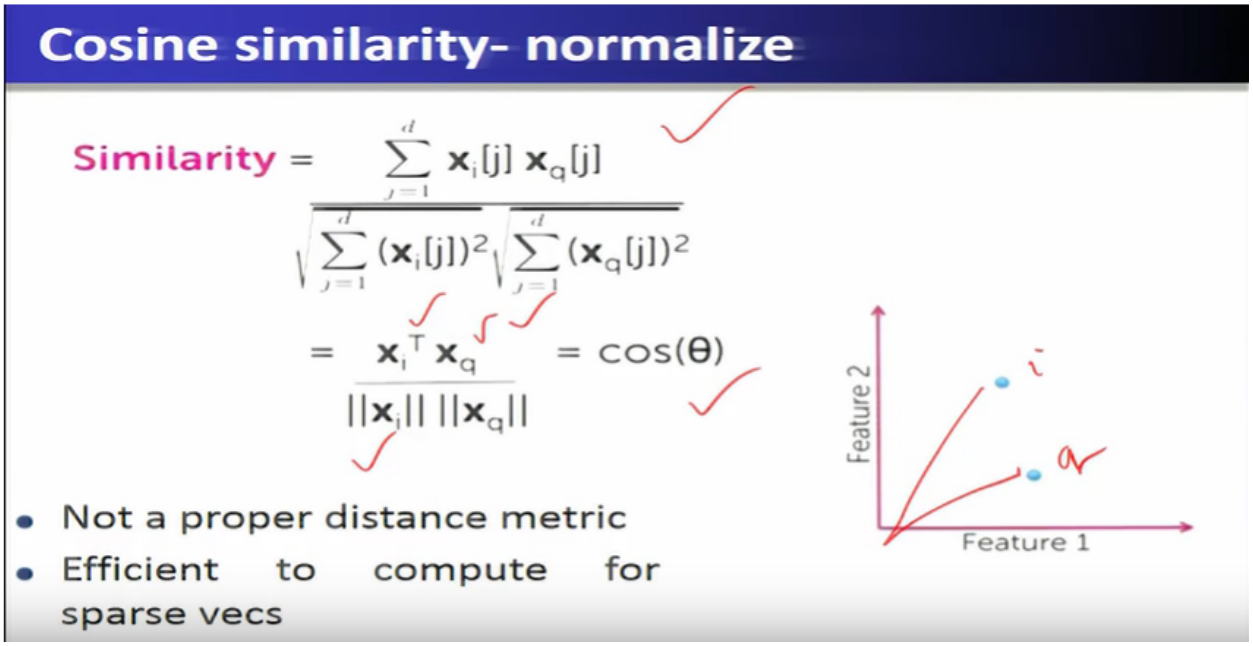
together, will form a cluster or a cluster of topics. So, if a corpus of documents is given, then you can apply, this cosine similarity and this cosine similarity, will segment all that topics, segment the carp us, into topic wise or distribution or classification and this can be used, in news reading promotions.

Refer slide time: (10:47)



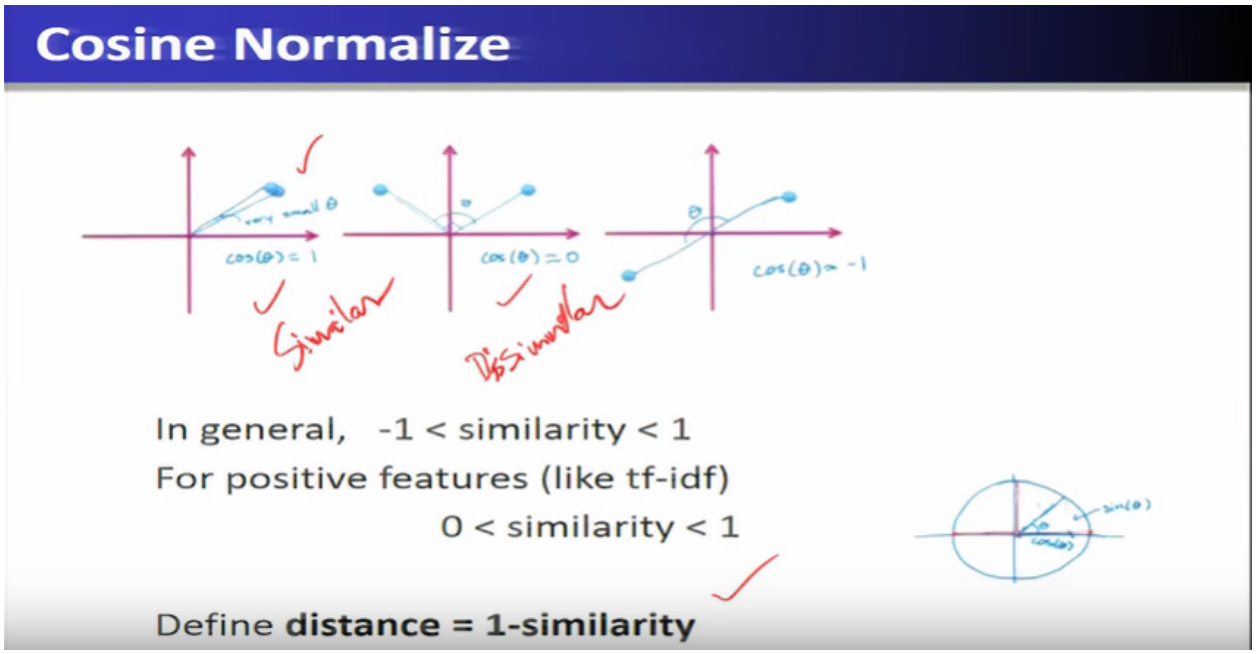
Now, we can see, this particular cosine similarity is, nothing but it's a, it's a dot product, of two vectors.

Refer slide time: (10:57)



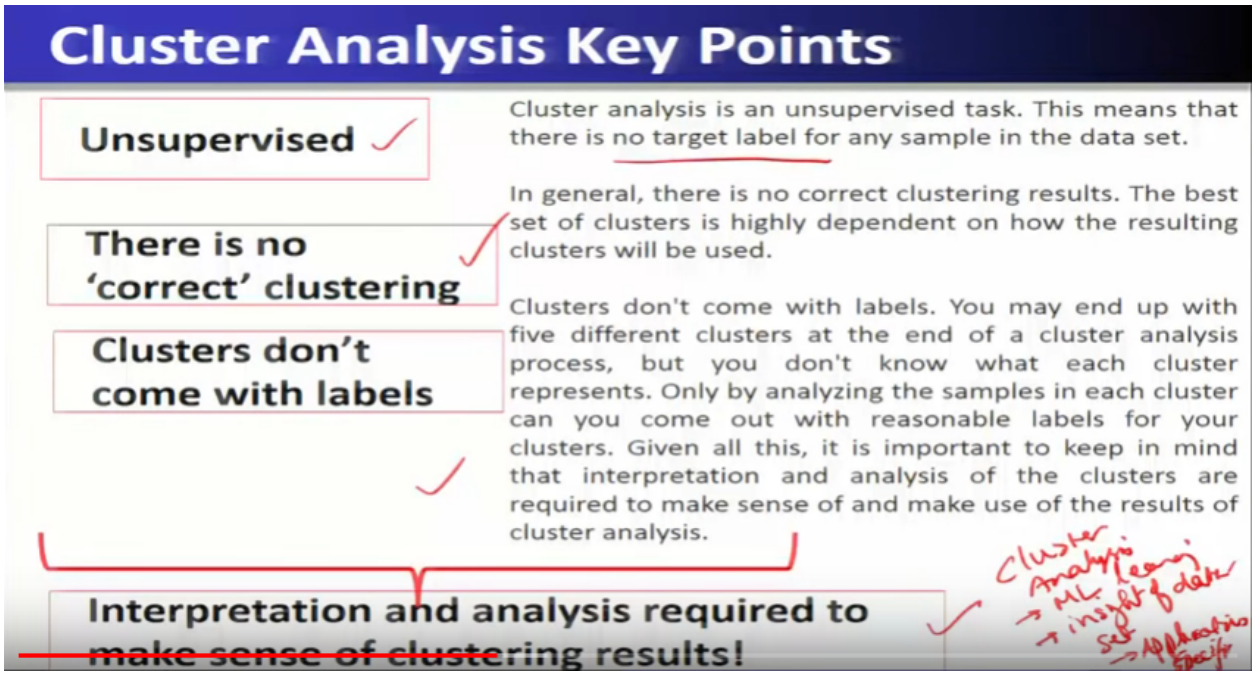
So you can see that, these vectors are nothing but they are the features. And if you take the cosine similarity, which by normalization, it will form this equation that is, the vector X I and the transpose of a dot product with, another vector. Q of it and which is divided by, this particular normalization factor, which is nothing but, all this is, the summation of all points of excise and their square root of it. So, this particular way, this cosine similarity will formulate, the normalized view, normalized way of the measuring, the similarities, among the two points or a to two topics and so on, can be used in various applications. Now, this particular cosine similarity is used, in the document classification, why because, these are very sparse. Hence, this kind of similarity measure, that is called, 'Cosine Similarity Measure', is very useful in such scenarios.

Refer slide time: (12:10)



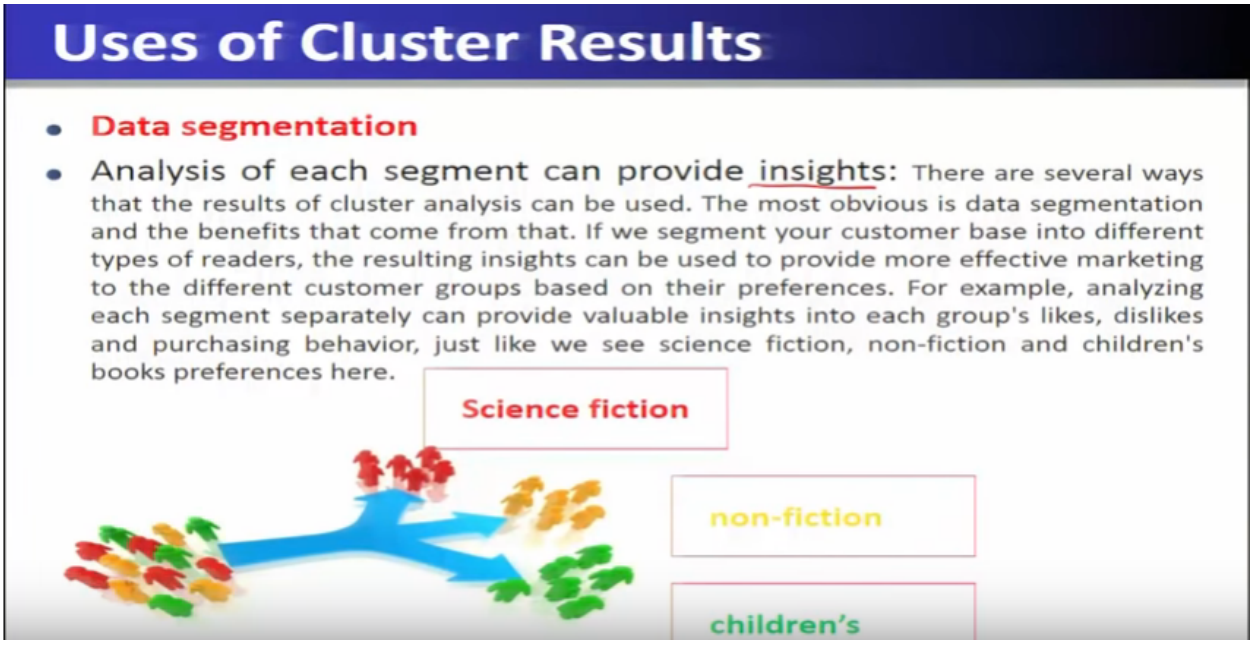
To have more understanding about, cosine similarity. and how this cosine similarity is normalized? We can see here that, if this angle is very small, then this cosine value of that similarity of these two points will become 1. Hence, higher is the value, the I mean, it is closer or more similar, the documents are. And if they are separated apart, then the cosine value of that angle, will become 0 hence, the this similar. This similar, objects or topics are having, the highest, cosine values and otherwise it will become zero. So, the distance is often measured because, one -that similarity hence, the smaller values are taken, as so here in this particular method.

Refer slide time: (13:18)



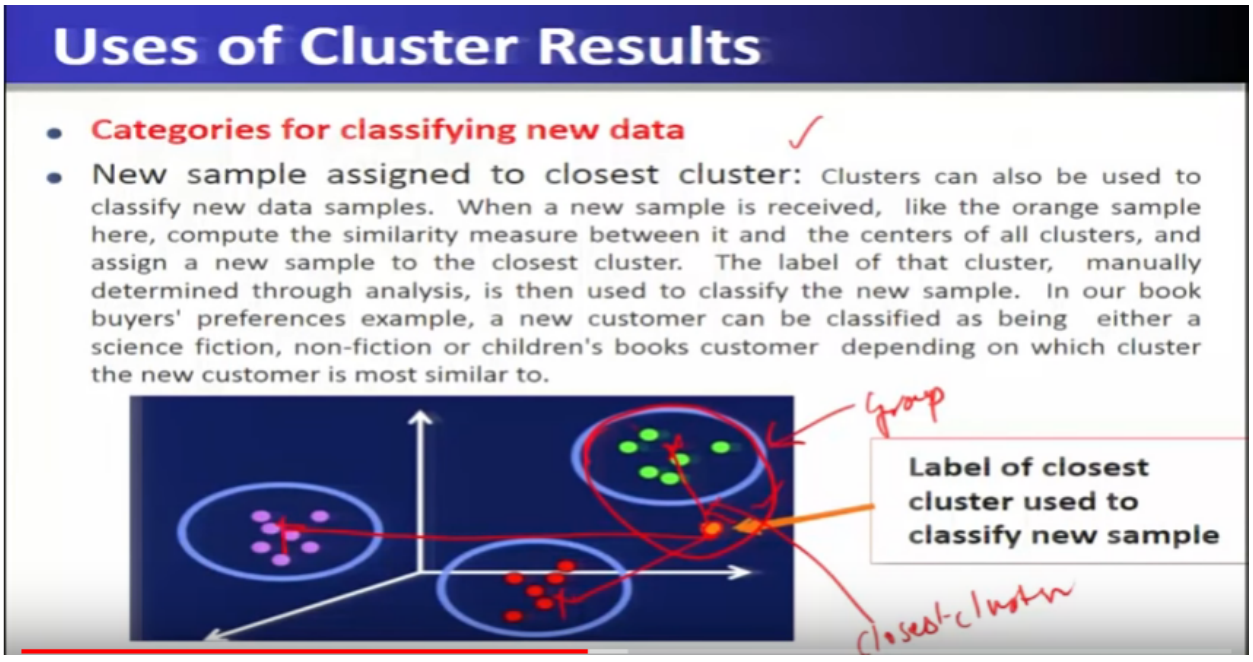
Now, cluster analysis the key points. we can see that ,this particular technique, of clustering or a cluster analysis is, categorized as the unsupervised machine learning, why because, the, the data points or a dataset is not, having any target label, hence, it is called, 'Unsupervised Learning'. Another thing is, the clustering there is no correct way of, the of being doing the clustering. Hence, it requires a different application and their interpretations, based on, this particular technique. Now, another key point about this cluster analysis is that, clusters don't come with the label, that we have seen and using this particular cluster analysis, we don't know, what each cluster will represent, so hence, by analyzing the samples, in each cluster, you can come out, with a reasonable label, for your clusters, given all these it is important to keep in mind, that interpretation and analysis of the clusters are required, to make sense of and make use of, the result of the cluster analysis. So, interpretations and analysis is required to make sense, of these clustering results. So that means, these cluster analysis is, a machine learning technique to gain the insight of the data set. And by doing this interpretation and this analysis, you can use this insight for any application or the business. Hence, the interpretation and the analysis is very easy, application is specific.

Refer slide time: (15:22)



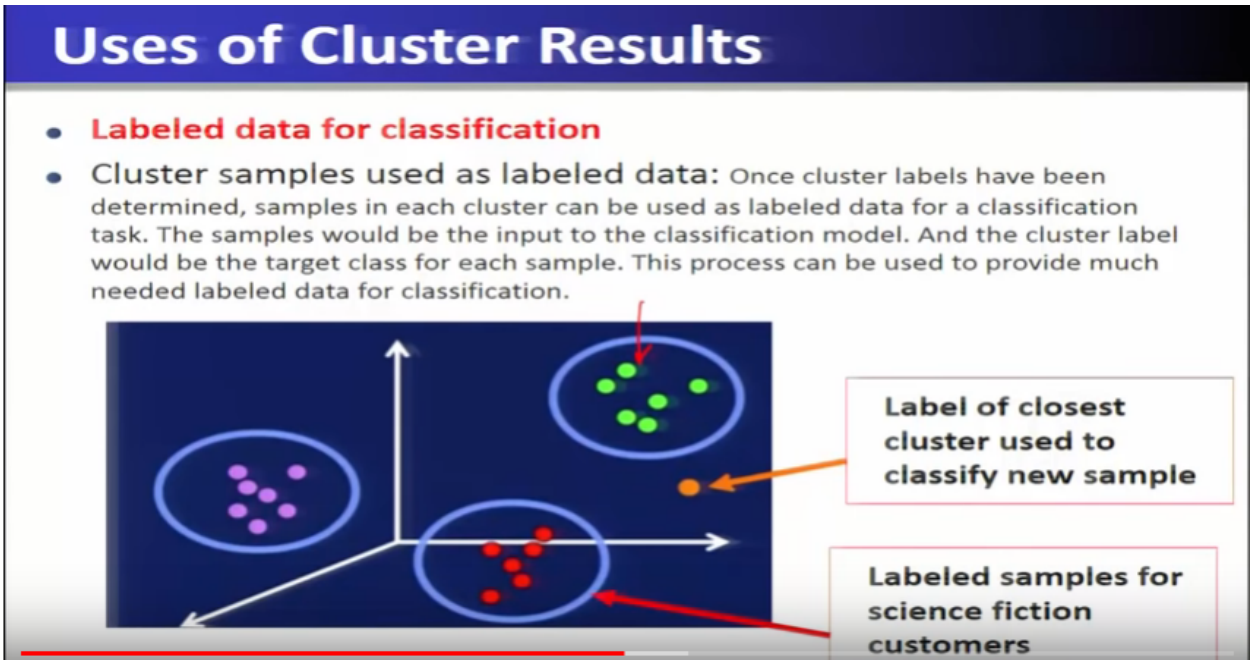
Now, let us see, some of the uses of these cluster analysis results. Now, the first we use is called, 'Data Segmentation', that is the analysis of each segment, can provide an insight, that we have already seen or we have already discussed. So there are several ways that, the results of cluster analysis can be used, one obvious is data segmentation and the benefits that come from that. So, if we segment your customer base, into different type of readers, the resulting insight, can be used to provide, a more effective marketing to different customer groups, based on their preferences. for example, analyzing each segment separately, can provide when you will insight into each groups, likes, dislikes, purchasing behavior, just like we see, a science fiction, non science fiction and children's, book preferences here, in this particular picture.

Refer slide time: (16:25)



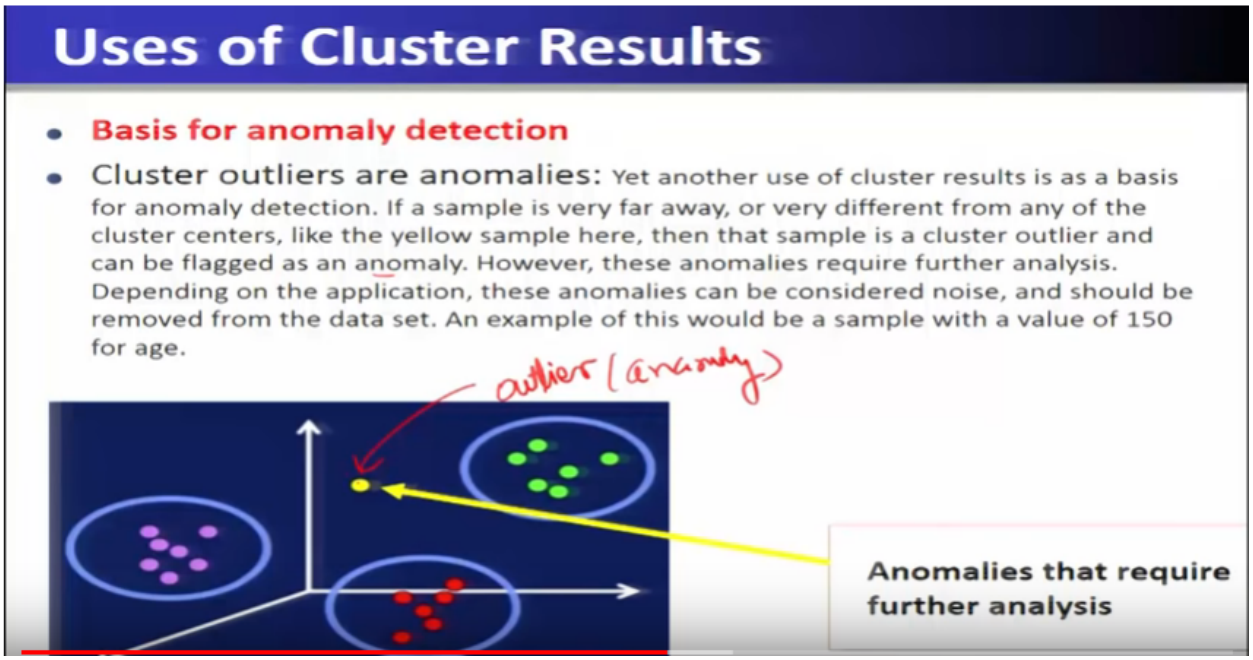
Now, categories of classifying the new data. So, when a new sample assigned to closest: cluster then cluster can be used to, classify the new data sample. So, when a new sample is received, like an Oran sample here, in this example, compute the similarity measure, between it and the Centers of all the clusters. similarities between the, between it and the Centers of our cluster and assign, a new sample to the closest cluster, for example, if this is the closest cluster, in this particular let us assume that, this is the closest cluster .so, this will become a new cluster member, after that. So, the label of that cluster, manually determines through the analysis is then used to classify, the new sample. So, in the book buyers, preference example, a new customer can be classified as being, either the science fiction, non science fiction or a children's, of books customer, depending on which cluster, the new customer is similar to. So, this particular cluster will identify, this particular user to be a part of, this particular cluster group. and hence, this particular classification ,will now know that, this customer is belongs to this kind of group, for example, here the groups which we have identified in the previous example is about, the readers of science fiction, non science fiction or a children's books.

Refer slide time: (18:18)



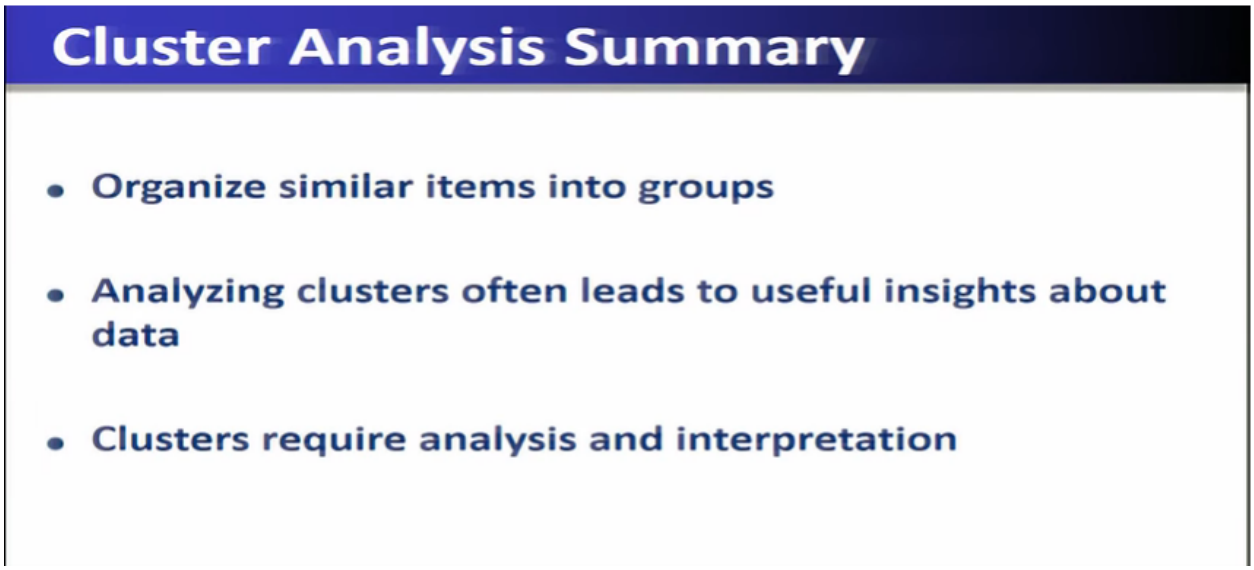
Now, label data for classification. So, cluster samples used as, the label data. Now, once the cluster labels have been determined, samples in each cluster, can be used as, label data as, for a classification task. The sample would be the input, to the classification model and the cluster label would be the target class, for each sample. this process can be used to provide, much needed labeled data for the classification, take this particular example here, the label of the closest cluster is used to classify, the new sample and the label sample, for the science-fiction customers are labeled in this case, so after clustering these particular, different clusters, will now be used to assign the level, for each data points and also when a new data points, comes it will be getting a new label. And this label data can be used for supervised learning.

Refer slide time: (19:27)



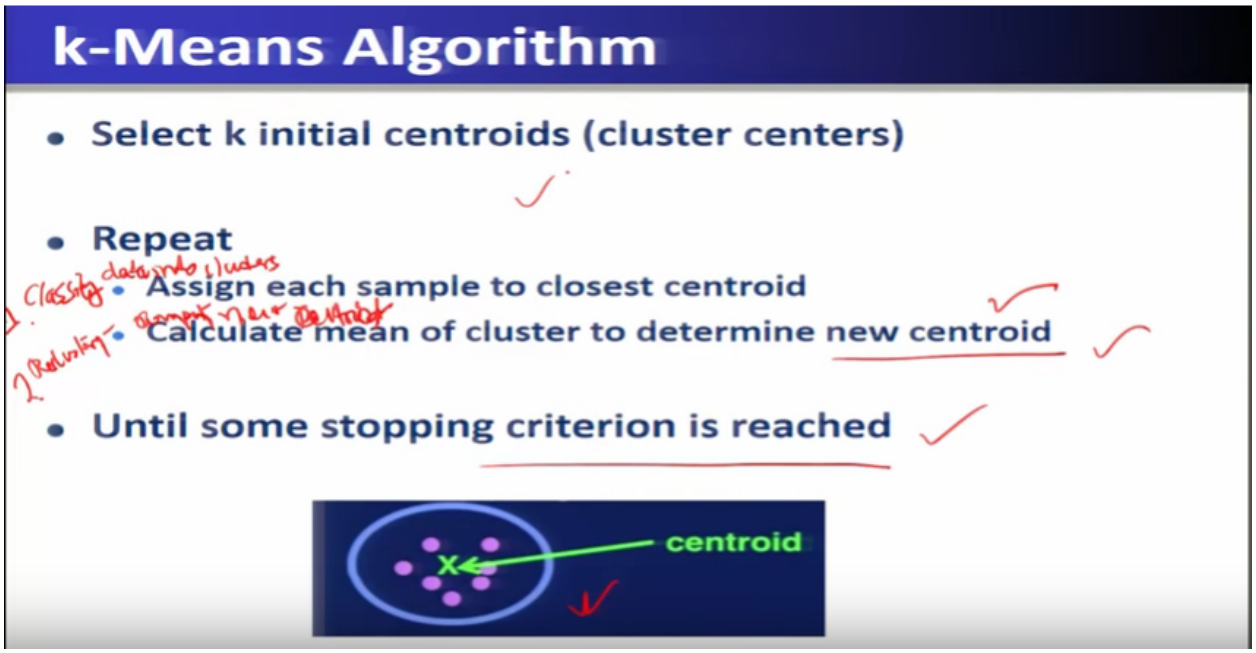
Now, another basis, this cluster can be used, cursor analysis can be used, as a basis for anomaly detection. So cluster outliers are, basically the anomalies. So, yet another use of cluster results is, as a basis for anomaly detection. So, if as ample is very far away or very different from any of the cluster centers, like here shown as, the yellow data, data point. Then the sample is a cluster outlier and can be flagged as the anomaly, however these anomalies required further analysis, depending on the application these anomalies, can be considered noise and should be removed from the data set. an example, of this would be sample with the value 150 for the age, so after doing this analysis and doing the anomaly detection, the quality of data, will be improved and hence, the quality of the predictions, accuracy of the predictions, will further increase in the machine learning context ,of analytics.

Refer slide time: (20:50)



Now, let us summarize, the cluster analysis is that, it will organize the similar items, into the groups and it will analyze, the clusters often leads to a useful insight about the data. And the cluster required the analysis and interpretation.

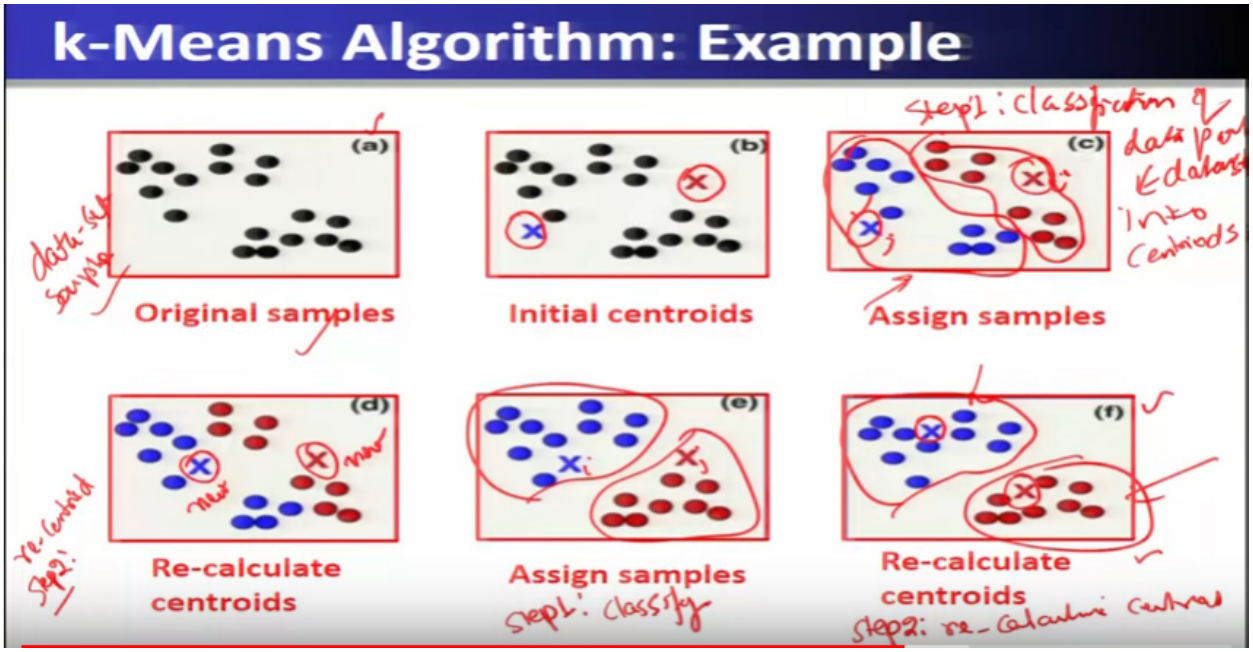
Refer slide time: (21:06)



Now, let us see, the one algorithm that is called k-means, algorithm for doing the clustering. And for performing the cluster analysis, as we have seen in the previous slides, different applications of it. Now, here the k-means, algorithm initially, selects k initial centroids, which are nothing but the cluster centers.

And the next step would be, iterative. iteration of two steps, which says that, assign each sample, data sample to the closest centroid and then, it will then calculate the means of the clusters, to determine the new centroid and this two iterations, will repeat until some stopping criteria is met. So, let us define that, this particular step is called, 'classify the data into the clusters'. and the second step is called,' Re Clustering', which says that, now having defined the clusters instep number one. Now, step number two will identify, by taking the means of all the data points, into a particular cluster and it its mean, will give a new centroid, so it is called, 'Re Clustering', that is computing the new centroids. Now, these two steps, will further repeat and this particular repetition will stop, until the stopping criteria is met, this is that, when we are going to compute the new centroid, there is no new, there is no much differences are happening further, that means, the points are almost saturated and fixed into the clusters, not going to change. Hence, the stopping of that algorithm is required, at that point, that means, all the plus, all the clusters are now being identified and fixed, with the proper centroid hence, this otherwise this particular step repeats. Therefore, the clustering of that data points is, highly dependent upon, the initial positioning or initialization of different centroids.

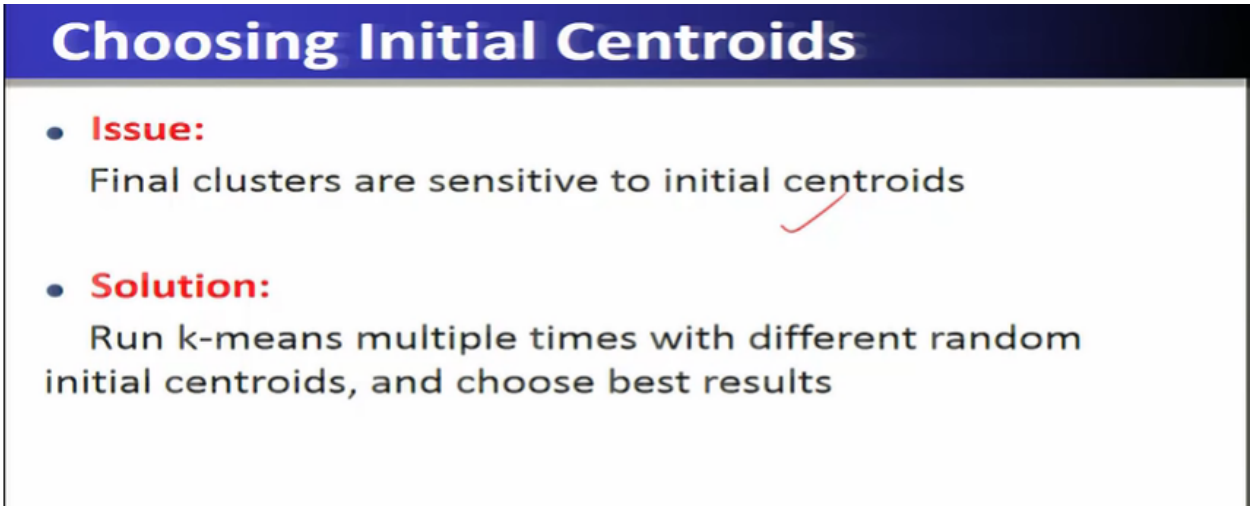
Refer slide time: (23:59)



So, let us see the example, of this k-means, algorithm on a original data set. Let us say, figure 1 or figure a, denotes the original data samples, data set points, this is the original data set sample. Now, let us identify or let us place, two clusters, these are the initial centroids, which are placed into that original data set. now ,then we will perform the second step or we will, we will now, this was the initialization, that is let us identify two centroids ,after identifying the two centroids I step one of the algorithm, which says that, a classification of data points, in the data set, into, into these two centroids. So for example, this particular every data point, will identify, will find out the distance between, these different centroids. Let us say, I and this is J, every data point will compute, these distances, this is called, 'Similarity Measures'. So, the data point which is closer, to a particular centroid will join that, centroid. So here, you can see, in

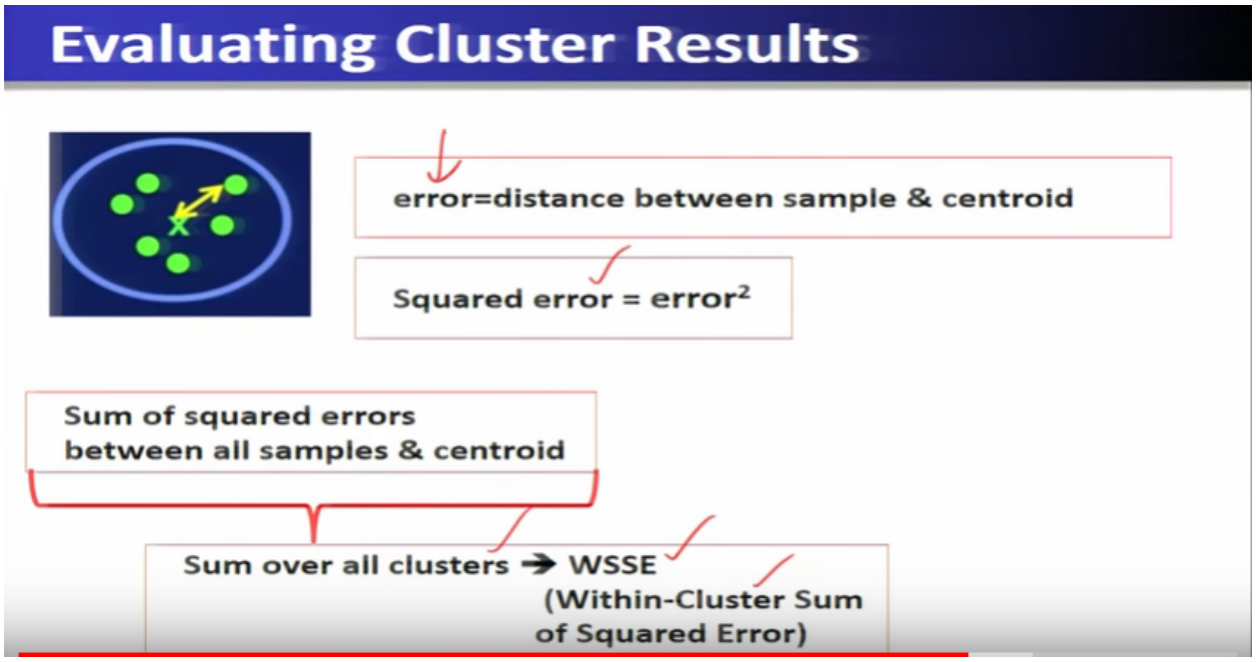
this particular diagram that, that these red dots, shows that, these data points are closer to the to the centroid X I, compare to XJ, where whereas all the blue poor depth dots, they are closer to XJ, compared to X I. hence, the two different groups, that means, the initial data, set samples is divided into two different groups and they are being identified as, the centroid, that is X I and XJ. Now, with this particular center, now we will execute the step number two, which says that, which says that it will recompute, the centroid or a new centroid, will be computed. So, based on these other points mean or an average will be calculated. And it will identify a new centroid. And then, the iteration repeats, for the step number one. and step number one will now, identify the data points which are closer, to this new centroid X I and the data points, which are closer to the new centroid X J. hence, two different clusters or segments is being identified. Now, in the figure f ,we see that, again we have to calculate the new centroids and now ,we can see that, new centroid is not going to change, these membership of these data points, hence the stopping criteria, will be it work to stop this algorithm at this end. Therefore, these two different data points or a cluster of these two points is, being used to identify, the insight into this particular data, which was given as the original data set. So they are basically, quite similar in their similarity measures.

Refer slide time: (28:29)



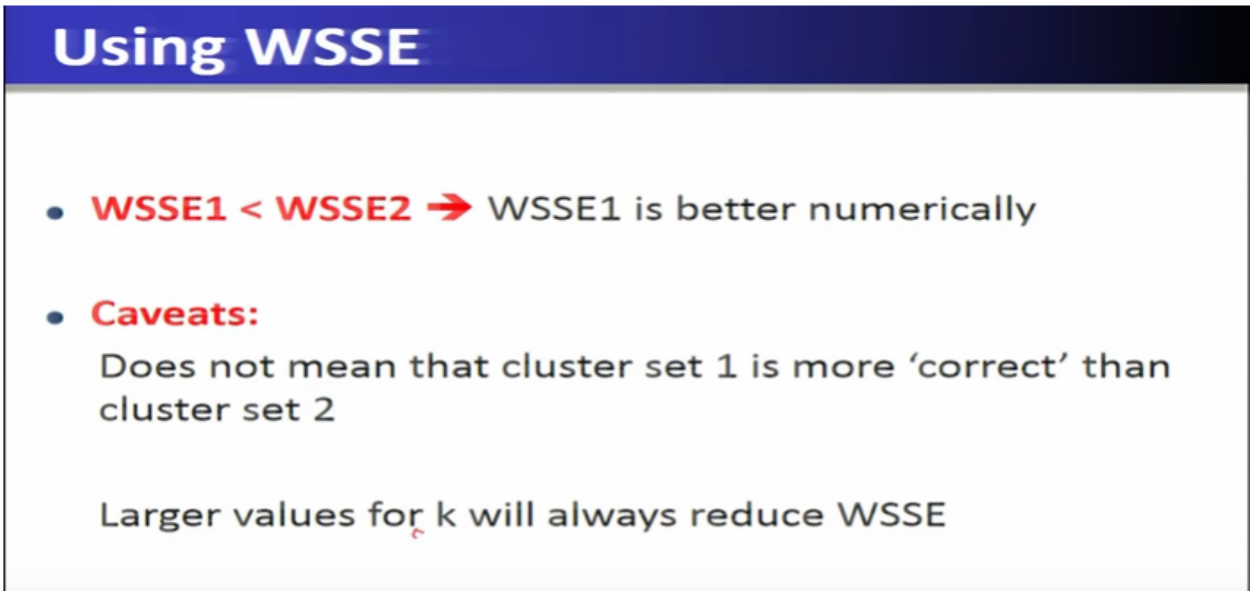
Now, we we have seen that, choosing the initial centroid is going to affect, the final clustering. So, final clustering is very sensitive to the initial centroid, so the issue is, the final clusters are very sensitive to the initial centroids .So what is the solution? to run the k-means, multiple times, with different random initial centroid, we have to choose the best results, hence, we start from that initial centroid after taking the best ones, which gives the best possible results.

Refer slide time: (29:07)



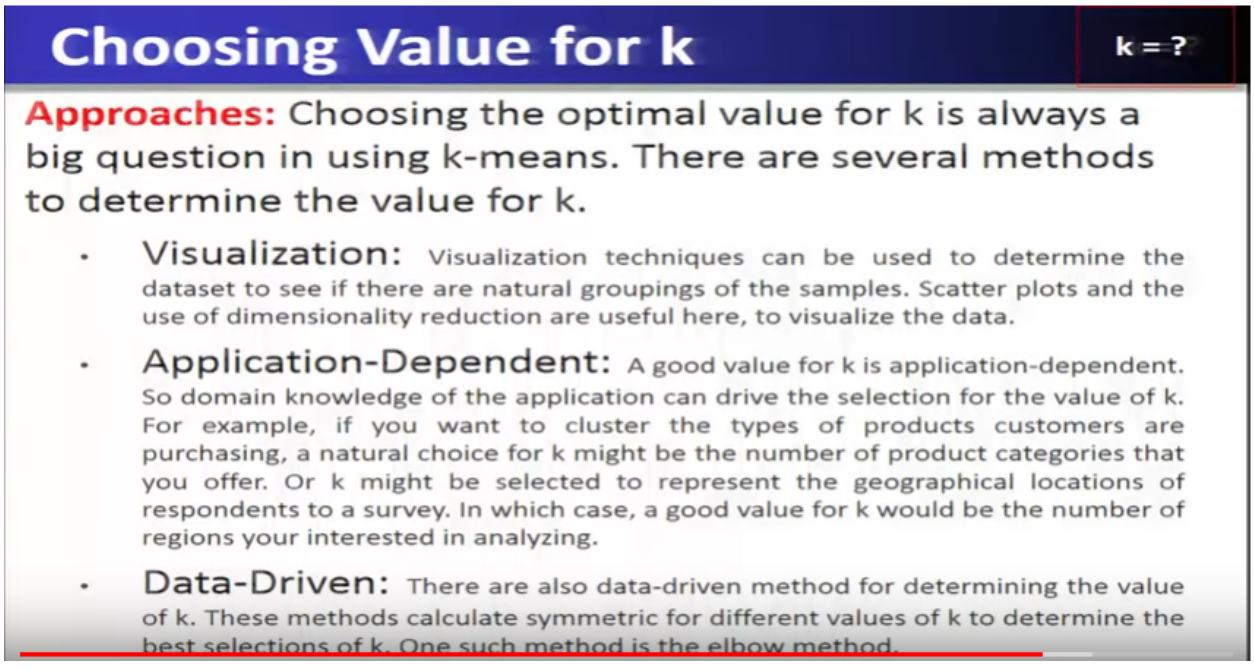
Now, after doing this clustering, using k-means, now we have to do this, we have to evaluate these are not these clusters. And we have to also interpret the results, which are in nothing in the form of clustering. So, for to evaluate that results. Let us, use a term which is called the, 'Error'. So, error is the term the difference between the, the sample and its centroid. So for example, we can see that, we can measure the distance between, the centroid and data points, for all data points and we have to take, the sum of these squared errors. And this is some of squared errors, between all samples and centroid is being calculated, for all over all the clusters, hands, within the cluster sum of square error is, the new term, which we are now saying WSSE. So that means, this particular term, which we have calculated is used to, do this analysis.

Refer slide time: (30:25)



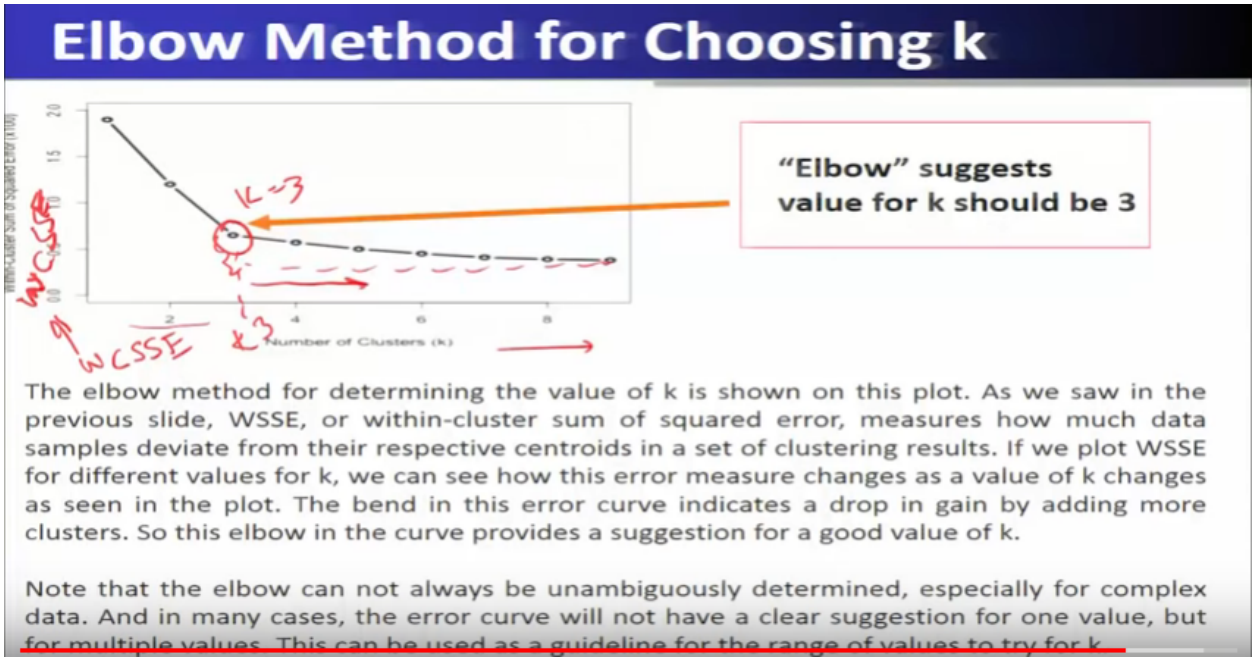
Now, we can say that, a clustering which has WSSE1. And there is another clustering, which is WSSE 2. Now, if we compare these two parameters, these two values of that clustering and if WSSE1 is less than, WSSE 2 then basically, WSSE 1 is a better numerically. And we can see that, this does not mean that cluster 1 is, more correct than cluster 2, larger the values of K, will always reduce this value of WSSE. Now, therefore the value of K, which we are now going to fix, will also has an implication on WSSE.

Refer slide time: (31:10)



So, now we will see, how to choose this value of K and, and, and so on and let us see, the different approaches. So choosing an optimal value for K is always, a big question and using the k-means, algorithm. and there are several methods, to determine the value of K. so, the first method is called, 'Visualization', technique can be used to determine the data set to see if there are natural groupings of the sample, scatter plots and the use of dimensionality, dimensionality reduction can be, useful here to visualize the data. the second approach is called, 'Application Dependent', that is a good value for K is application dependent, so the domain knowledge, of the application can derive the selection, for the value of K. for example, if you want to cluster, the types of products customers are purchasing, a natural choice for K might, be the number of products categories, you offer or K might be selected, to represent the geographic locations, of respondent to survey in which case, a good value of K, would be the number of regions, you're interested in analyzing. Third approach is called, 'Data-Driven. So, there are also data-driven methods, for determining the value of K, these methods calculate symmetric for, different values of K, to determine the best selections of K.

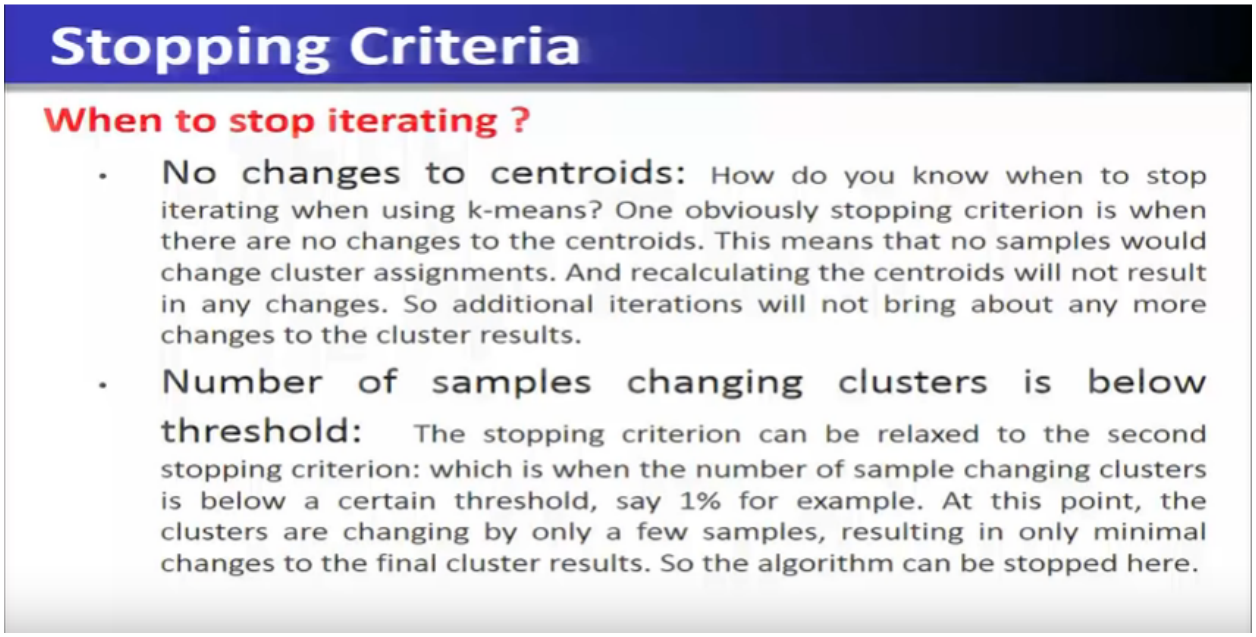
Refer slide time: (32:51)



One such method is called, 'Elbow Method'. which is shown over here, for choosing the value of K. in this elbow method, we can see that this W, within cluster sum of squared errors ,that is w, WCSSE, that we have seen, again we are writing WCSSE , this term is shown on, the vertical axis and on the horizontal side, the number of cluster, that is the value of K. so, as we increase the value of K, we can see that it is coming down, but beyond this particular point, if we increase the value of K, that is the number of clusters ,we see that, this particular reduction in the WSSE is very nominal, it's not very much. So, this is the point, which is called an elbow point. So, elbow point shows that the value of K, should be this elbow point and this is nothing but a k. k is equal to 3, at 3 we see that, the WSSE will not drop much, beyond

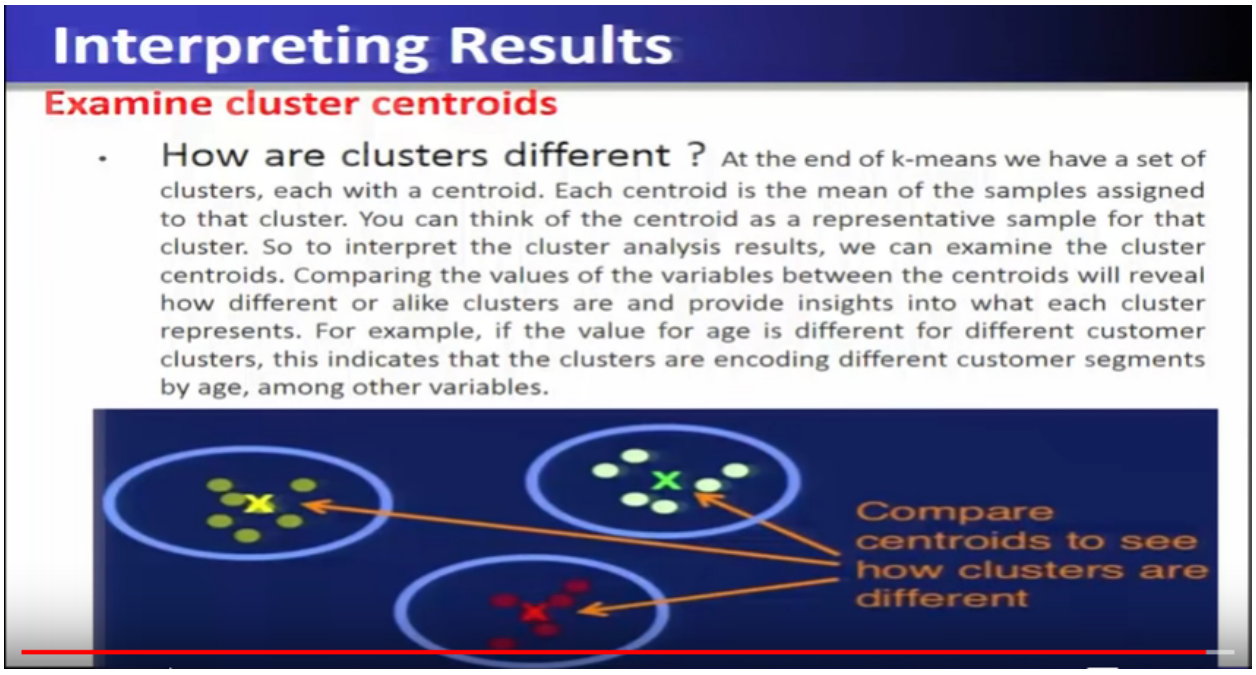
that if we increase the value of K. so even if them, the number of clusters are more beyond 3, it is not going to affect much, on reducing this WSSE hence, the value of K is chosen, based on this elbow method. So, in the elbow method for determining the value of K is shown, on this particular plot, as we see, on the previous slide WSSE or within cluster sum of square or measures, how much data sample deviate from their respective centroid, in a set of clustering results. So, if we plot, W SSE for different values of K, we see that, how this error, measure changes as the value of K changes, as seen in the plot. The bend in this error curve indicates a drop, in the gain by adding more number of clusters. So, this elbow in the curve provides cessation, for a good value of K, note that the elbow, cannot always be unambiguously determined, especially for a complex data, in many cases the error curve, will not have the clear solution for any value, for multiple values. This can be used as a guideline, for trying the value of K.

Refer slide time: (35:37)



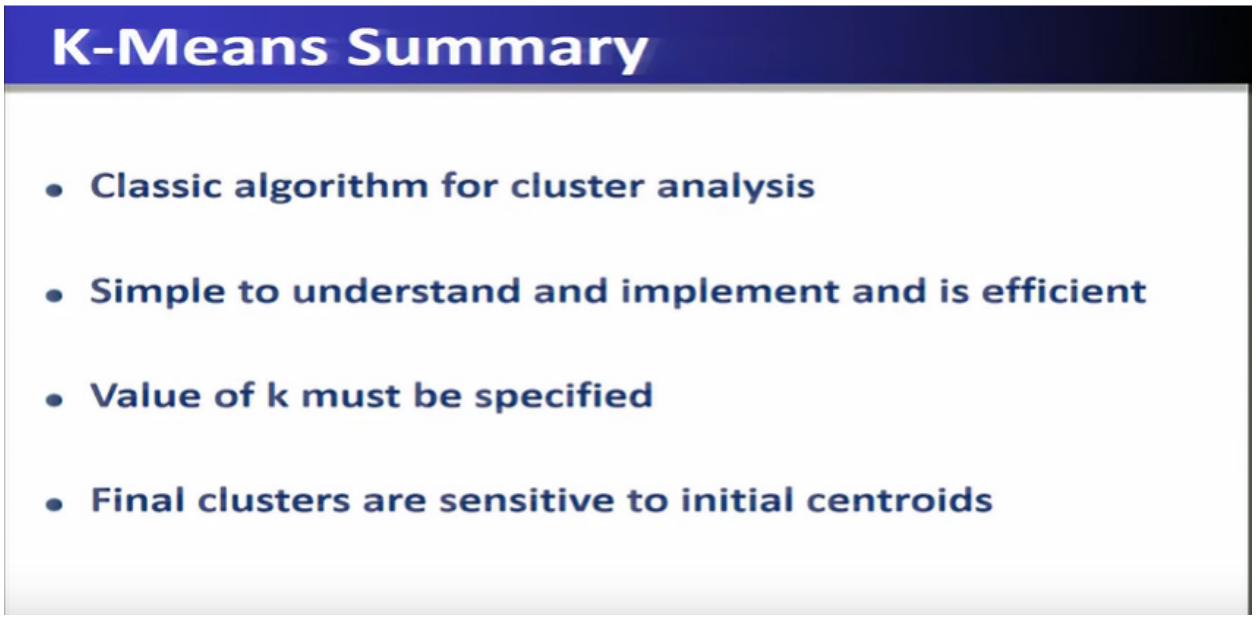
Now, another parameter here in k-means, algorithm is about, stopping criteria. When to stop these iterations of that, that is two iterations, which two steps in an iteration, which we are making, classification and Re computing the new centroid. Now, we will stop iterating, when no changes to the cluster, to the centroid is observed. So, how do you know when, to stop iterating when using the k-means? One obviously stopping criteria is when there is, no changes in the centroids. This means, that no samples would change the cluster assignments and recalculating the centralized, would not result in any change, so additional iterations will not bring out, any more changes in the results. Second way to stop, these iterations in the k-means, clustering is, determined using the number of samples, which are changing, the clusters is below a threshold. So, the stopping criteria can be relaxed, to these second stopping criteria, which is when the number of sample, changing clusters is below a certain threshold, say 1% for example, in at this point, the clusters are changing by only a few samples, resulting only minimal changes to the final cluster result. So, the algorithm can be stopped here.

Refer slide time: (36:57)



Now, finally interpretation of the results is also, going to be very important. Now, examine the cluster centroids, how the clusters are different? So, at the end of k-means, we have a set of clusters and each with a censored. Each centroid is the mean of, the samples assigned to that cluster, you can think of the centroid is the representative, sample for that cluster. So, to interpret the cluster analysis result, we can examine the cluster centroids, comparing the values of variables between the centroid, will reveal how different are alike, the clusters are and provide insight into, what clusters represent? For example, if the value for the age is different for, different, for different customers cluster, this indicates that the clusters are encoding different, customer segments by age and among other variables. So, here we can see here, to compare the centroid to see, how the clusters are different.

Refer slide time: (38:06)



So, let us summarize the k-means, so it's a classical algorithm for doing the cluster analysis or the clustering, it is simple to understand and interpret and it’s efficient also, the value of K here, in k-means, algorithm must be specified. And the final clusters are, very sensitive to the initial centroid.