# Data Mining
# Prof. Pabitra Mitra
# Department of Computer Science and Engineering
# Indian Institute of Technology, Kharagpur

## Lecture - 35
## Clustering – IV

[noise]

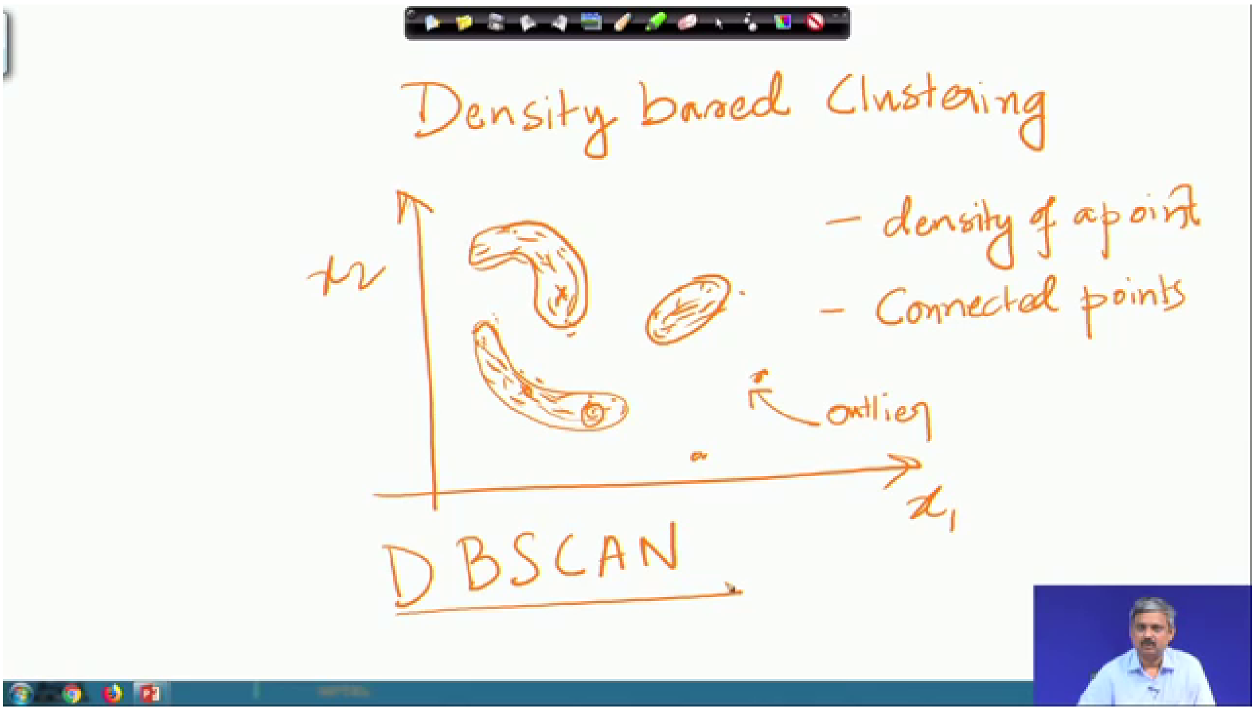
We had earlier talked about two class of clustering algorithms; one was the [noise] hierarchical algorithm where we successively either agglomerate merge together a set of points till we get the desired number of clusters or divisive; where you split the entire dataset into groups till we get the desired number of cluster. And depending on how we measure the distance between two groups of point, we had either the average linkage or the single linkage or the complete linkage clustering algorithm.

We will also found out that these algorithms can find out any arbitrary cluster shapes, non convex or any other shape. But they are sensitive to noise and they take a lot of time when run on the large data sets. The other class of clustering algorithms we described was the partitional, where instead of hierarchically merging and splitting points; we in a single go divide points into [noise] desired number of groups and then update them; so, it is iterative algorithm.

So, in the k means we start with random ah initial seed points and ah all the points which are closer to a particular seed are put together. And then the seed points are updated by the mean value of all these points that are grouped together. This process continues till the seed points do not [noise] change over iterations. So, this is the k means algorithm ah.

This algorithm a limitation is it can produce only spherical clusters and even though the advantage is it is very fast in few iterations it converge. And it is since ah ah it is robust to noise; today we ah talk about a third [noise] category of clustering algorithms known as the density based clustering; the idea is like this [noise].

(Refer Slide Time: 02:36)

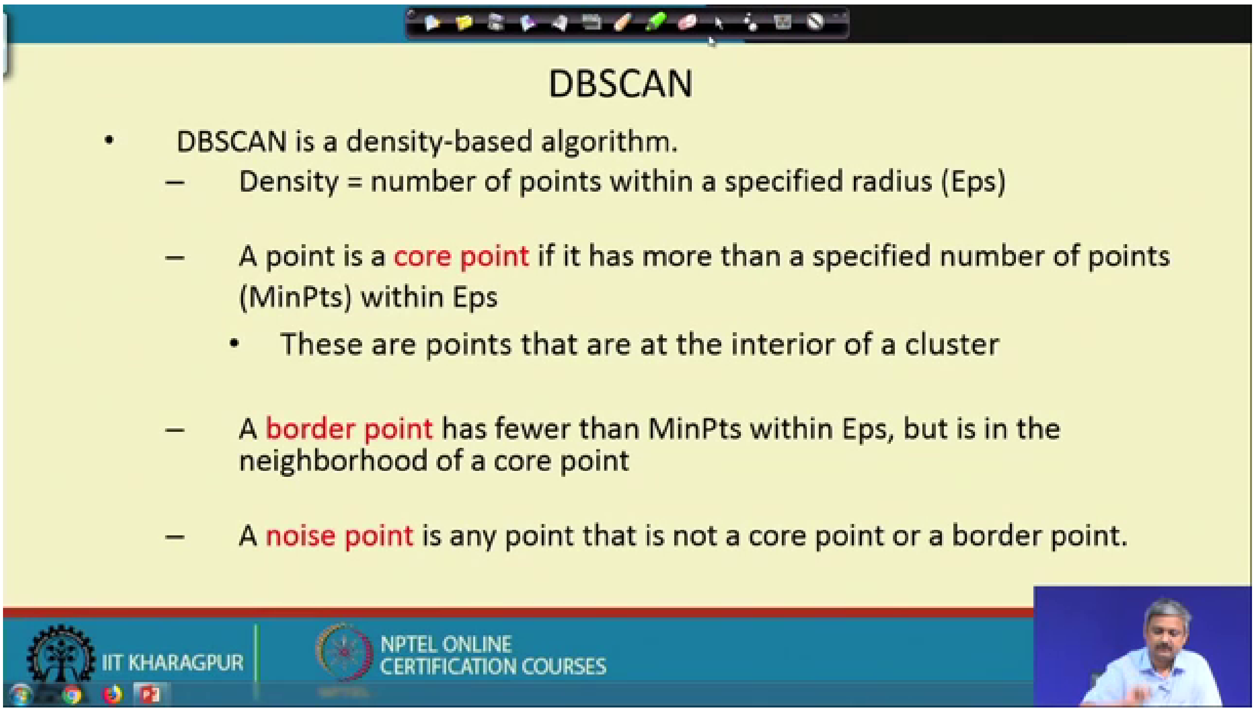


So, we [noise] we will discuss [noise] clustering [noise] the idea is that [noise] if we again I am [noise] drawing points in 2 dimension [noise] it can be in any dimension; if you have say [noise] clusters [noise] like this [noise] and like this [noise] then you see one of the way you can define a cluster is a set of points who have lot of other points in their neighborhood and all such points connected together would form a cluster.

So, we look at two things we study [noise] density [noise] of a point [noise]; that means, if there are other points in the neighbourhood. For example, this is it [noise] all these points are dense, but maybe if we have [noise] one or two separate point they are not dense because there is no other point in the neighborhood. So, we talk about density of a point and we talk about [noise] connected [noise] points [noise] points. So, in [noise] sense it can [noise] dub the air connected [noise] I will I will soon define ah what this connected actually means ok [noise].

So, if we define a cluster to be like that and in points [noise] which are not part of a cluster as outliers; [noise] you can as well have a clustering algorithm [noise]. So, the ah this kind of algorithms are of course, called density based clustering algorithms. So, this idea is used in one clustering algorithm called D B SCAN [noise] Density Based Spatial Clustering [vocalized-noise] Acronym for all that. So, ah let me explain what it ah means [noise] I will come back to this.

(Refer Slide Time: 05:00)



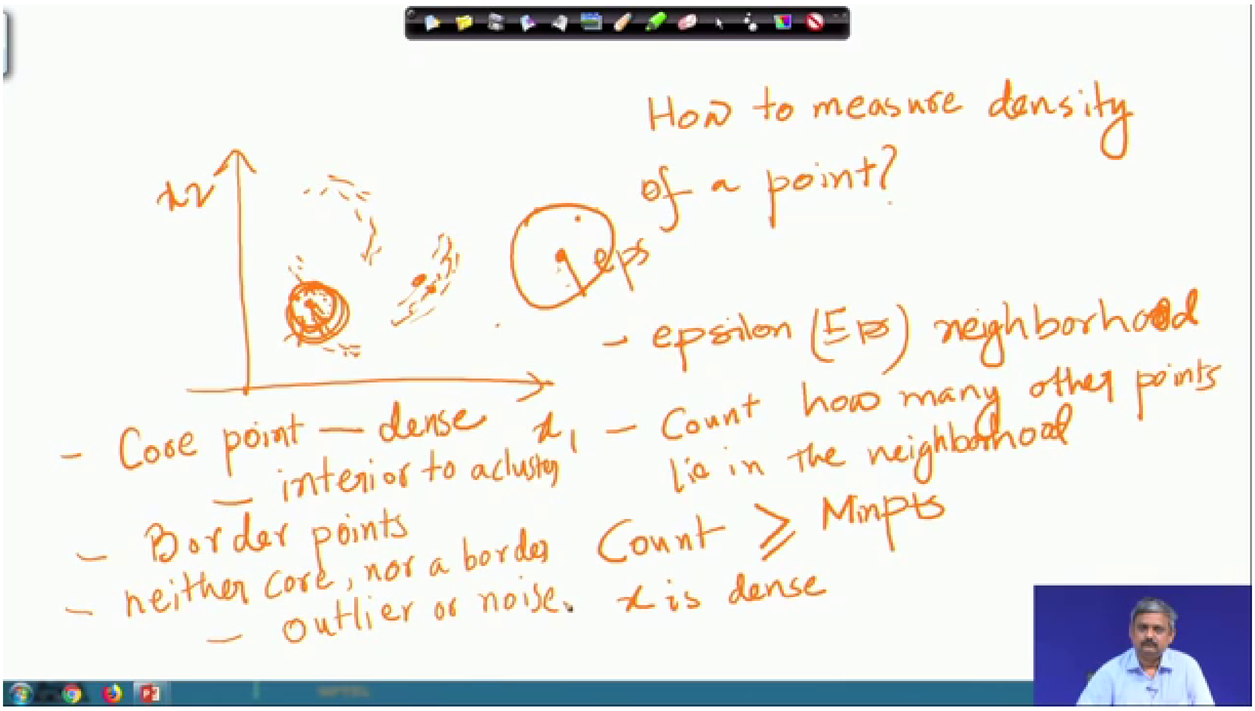
## DBSCAN

- DBSCAN is a density-based algorithm.
  - Density = number of points within a specified radius (Eps)
  - A point is a core point if it has more than a specified number of points (MinPts) within Eps
    - These are points that are at the interior of a cluster
  - A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point
  - A noise point is any point that is not a core point or a border point.

So, how you define density [noise]? You define density as points within a [noise] specified ah sorry as [vocalized-noise] ah as [noise] as points [noise] [vocalized-noise] as points within a specified radius ah of the given point that is the sensitive. A point will be called a core point if it has more than [noise] some main pts number of points within this epsilon radius.

So, these are kind of points interior to the cluster and then you have border points which have which are not dense, but they are connected [noise] to other dense points. And the points which are neither core neither border; we call them as noise or outliers let me explain you again ah for this concept [noise].

(Refer Slide Time: 06:19)



How to measure density of a point?

- epsilon (EPS) neighborhood
- Count how many other points lie in the neighborhood

Count ≥ MinPts
x is dense

- Core point — dense
  — interior to a cluster
- Border points
- neither core, nor a border
  — outlier or noise.

So, [noise]; we draw the [noise] figure again [noise] [vocalized-noise] I draw the figure again [noise] ok; so, suppose these are my clusters and they are some outliers; so, [noise] how to measure [noise] density [noise] of a point [noise] ok?

So, think how dense is a point [noise] ah. So, how dense it is ah the comes from the intuition say you say some some ah some public place a railway station or something; it is very densely crowded or we say some park is not very densely crowded. So, how do you do? Basically you look at a person and see a small area around a region around that point. And there is if there are lot of other people around lying also standing in that area you call it a dense spot.

So, the same thing we do here what we do suppose I am wanting to find the density of this point; what I will do? I will draw an circle centred at that point and with radius epsilon [noise] EPS [noise] you can call it [noise] epsilon ok [vocalized-noise] ah. Usually you will take epsilon to be small and we draw a circle of round ah of radius epsilon.

So, in 2 dimension it is a circle [noise] in higher dimension 3 dimension it is [vocalized-noise] sphere in general d dimension it will be a ball. So, to say the terminology is called a ball ah usually this algorithm will work if the dimension is relatively low maybe less than 3 or less than 7 ok. In higher dimension it is not a very good method; so, that is why

D B SCAN is only applied to it is original only applied to geospatial G I S data sets mainly where you have less number of dimension ok anyway. So, ah we have this epsilon neighborhood [noise] a circle centred [noise] on a point [noise] and radius epsilon [noise]. And we count how many other points [noise] lie in that neighborhood [noise]; how many other points lie in that neighborhood. So [noise] other points [noise] lie [noise] in the [noise] neighborhood [noise] and if that number is big high relatively high.

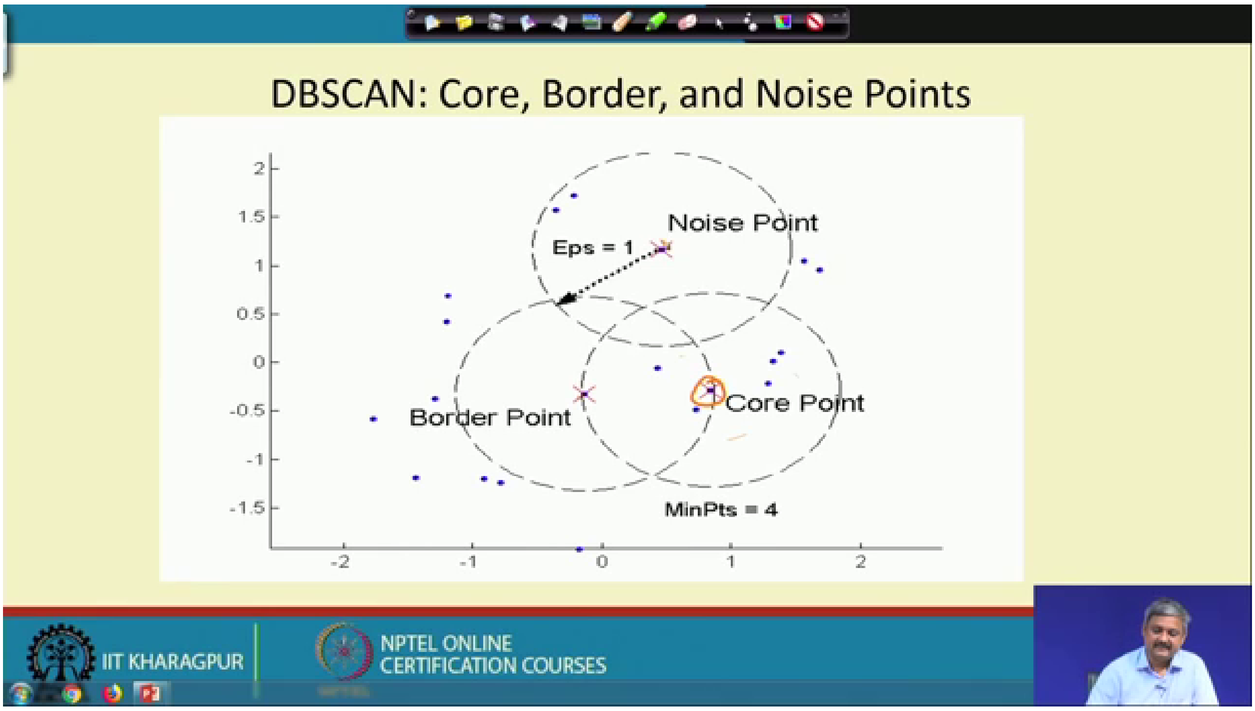
So, how do you do? You check whether this count is greater than [noise] equal to some [noise] threshold that we [noise] said called [noise] min pts [noise] maybe 5 or 7 or 9 or something ah it is the it is a count integer. So, you see how many other points lie in that neighborhood and if that count is [noise] greater than min pts [noise]; that means, we have sufficiently good number of other points lying in that neighbourhood, then we call this point [noise] x to be [noise] dense [noise] x is dense ok. In the D B SCAN terminology, a point will be called a core point [noise] if it is dense [noise]; so, all dense points are core points.

So, they are [noise] usually points which are [noise] inside a cluster ok. So, they are [noise] interior [noise] to a [noise] cluster, but there may be points which are themselves not dense; themselves not dense, but they are in a in [vocalized-noise] inside a ball inside a sphere of another dense point.

So, it may happen [noise] that the [noise] there is a dense point, there is another point which is itself not dense, but it is within the sphere [noise] of another dense point lies in a close proximity of another dense point ah note that the radius of this circle is EPS [noise]. Such points [noise] are called [noise] border points [noise] ok and a point which is [noise] neither a core nor a [noise] border [noise] is called a outlier [noise] or [noise] noise [vocalized-noise] ok it is called a outlier or noise.

So, this is clear let me explain you [noise] this thing [noise].

(Refer Slide Time: 12:34)

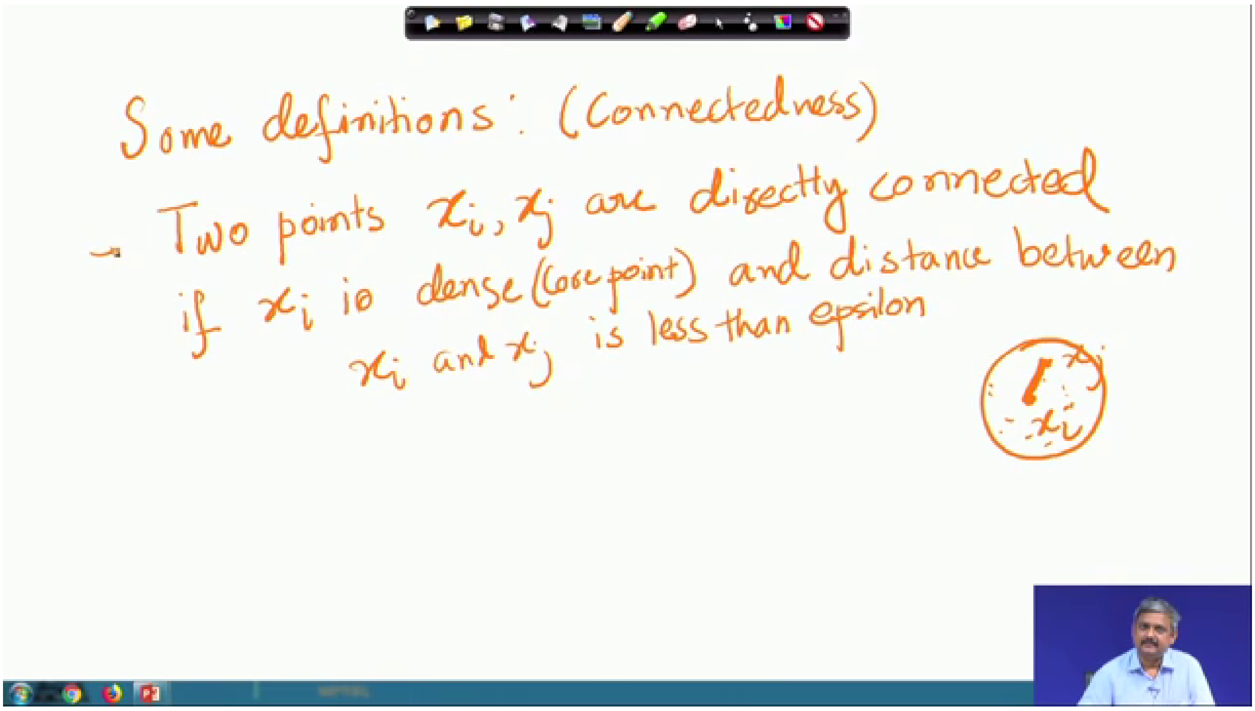


DBSCAN: Core, Border, and Noise Points

So, hm here in this case this blue. So, ah let me put the definition once [noise] first [noise] this is the definition core border noise ok. So [noise] if we take these blue points and say it EPS to be 1 and main pts to be 4; so this is a [noise] core point, this is a core point [noise] this is a core point because [noise] if you see there are [noise] more than four points in the neighborhood this point is not a core ok is not a core, but it is within epsilon of the core [noise].

So, it is a border point and this point is neither core nor border; nor it is itself dense nor it is within epsilon of a dense either noise point ok. So, this is the idea; so, what I will do is that based on this idea [noise] I will define [noise] some terms [noise] [vocalized-noise] [noise].

(Refer Slide Time: 14:15)



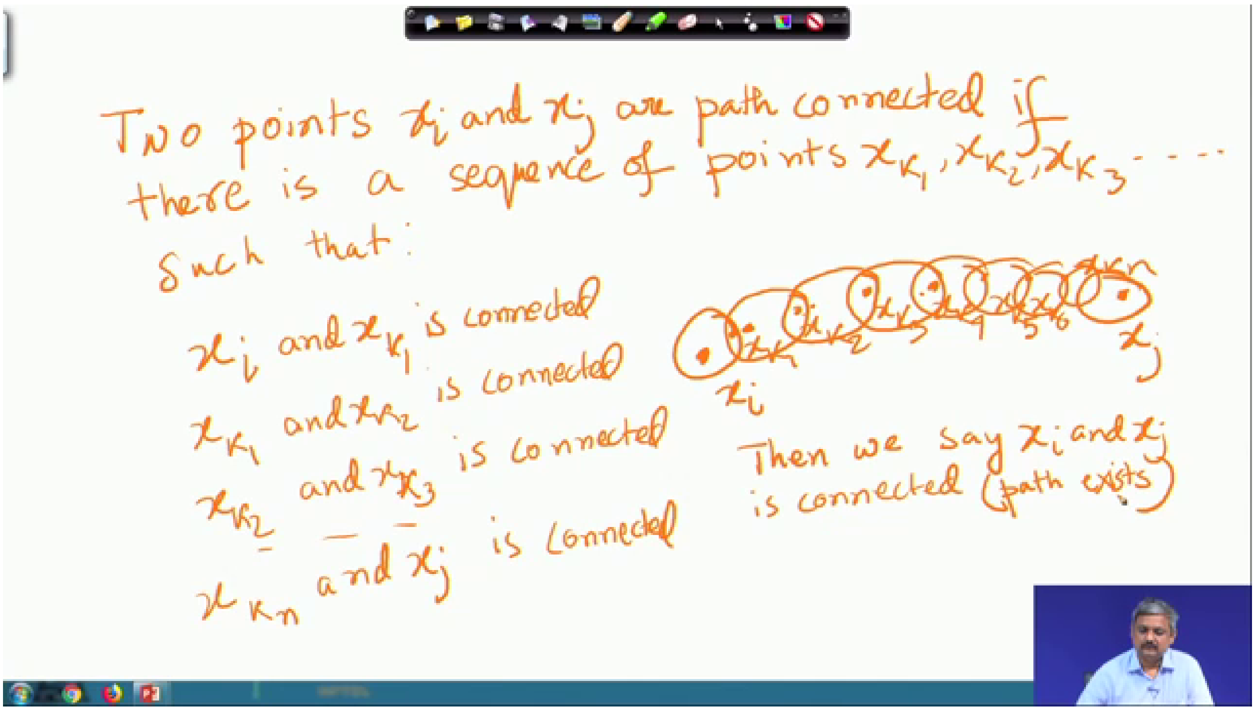
Some definitions: (Connectedness)

→ Two points $x_i, x_j$ are directly connected if $x_i$ is dense (core point) and distance between $x_i$ and $x_j$ is less than epsilon

So, these definitions I say density is one concept core ah then ah the; another concept is connectedness [noise] [vocalized-noise] [noise]. So, x x i is dense there are more than min pts point in a neighborhood of x i of radius epsilon [noise] [vocalized-noise] ok; so, [noise] this is direct connection [noise] ok [noise] clear.

So, x i is connected to x i if x i is dense an x i is within epsilon of x x i [noise].

(Refer Slide Time: 16:33)



Two points $x_i$ and $x_j$ are path connected if there is a sequence of points $x_{k_1}, x_{k_2}, x_{k_3}$ - - -

Such that:

$x_i$ and $x_{k_1}$ is connected

$x_{k_1}$ and $x_{k_2}$ is connected

$x_{k_2}$ and $x_{k_3}$ is connected

— —

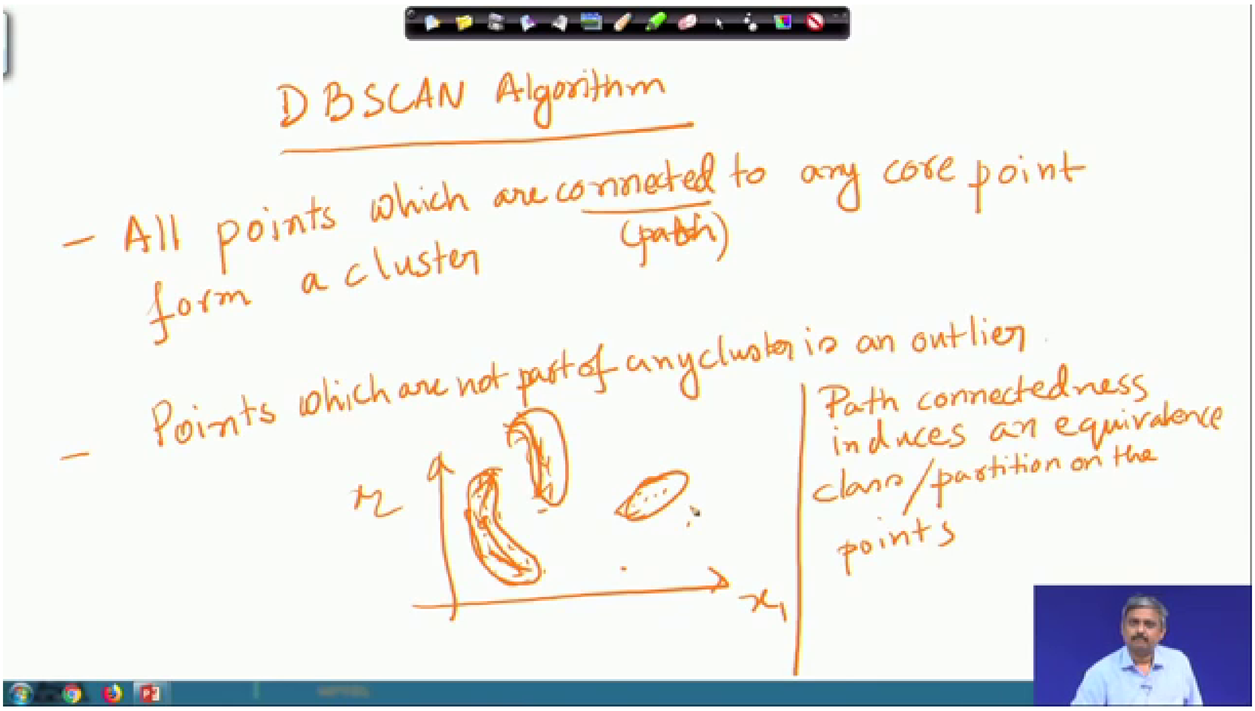
$x_{k_n}$ and $x_j$ is connected

Then we say $x_i$ and $x_j$ is connected (path exists)

[noise] [noise] [vocalized-noise] What it means? [noise] So, you can find the series of points x 1 to x n ok x 1 to x n. So, that x i [noise] this is connected this and this is connected, [noise] this and this is connected, [noise] this and this is connected, [noise] this and this is connected directly [noise] till we end in x a x j ok.

So, ah that is the criteria connected means this is dense [noise] and this is an epsilon this is dense [noise] this is an epsilon this is dense [noise] this is an epsilon ok [noise] connected right alright. So, this is clear; so this is how we have connected first directly connected then path connected [noise]; now let us see this how this helps in clustering [noise] ok.

(Refer Slide Time: 20:41)



DBSCAN Algorithm

- All points which are connected (path) to any core point form a cluster

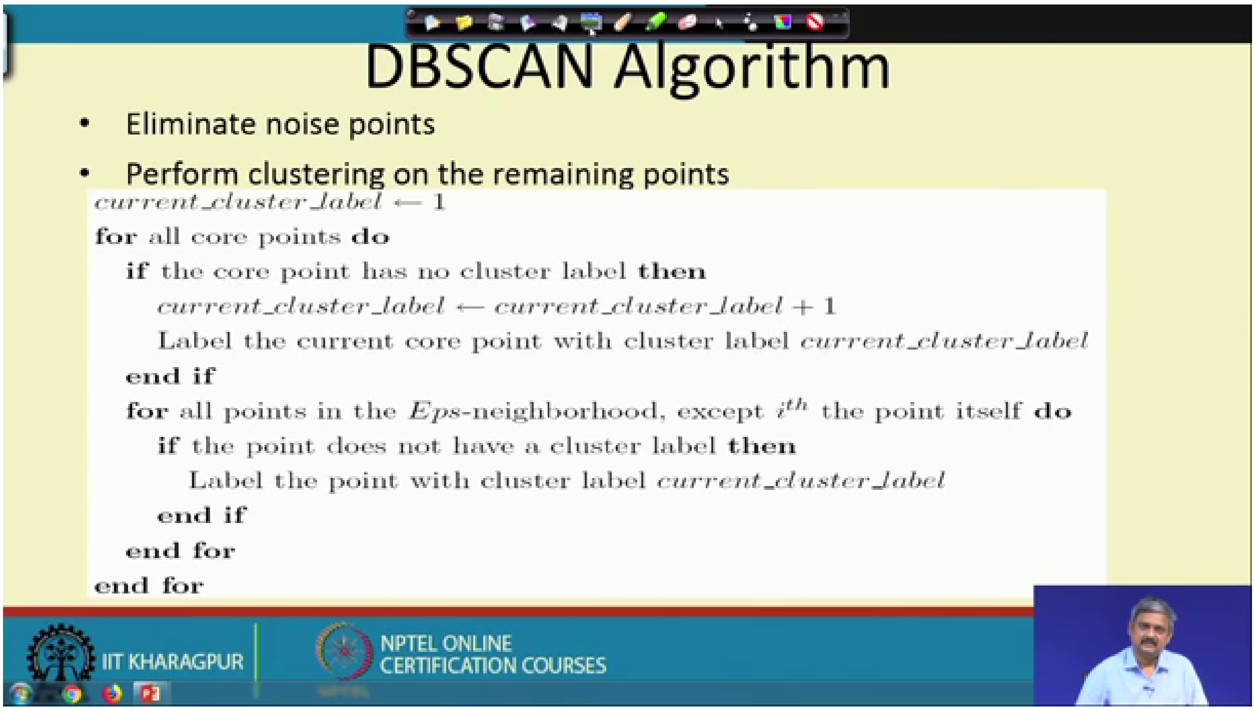
- Points which are not part of any cluster is an outlier.

-

Path connectedness induces an equivalence class/partition on the points

Let me let me first define [noise] I am in path connected I will explain soon [noise] [noise]; that means, [noise]. You see all the [noise] points here there is a dense point [noise] and they are connected to each other [noise] that exist path similarly [noise] all are connected to each other [noise] all are connected [noise] outliers are not connected ok. So, [noise] on the points [noise] alright; so, these are one equivalence class, [noise] one equivalence class, [noise] one equivalence class all right these are single ok.

(Refer Slide Time: 23:31)



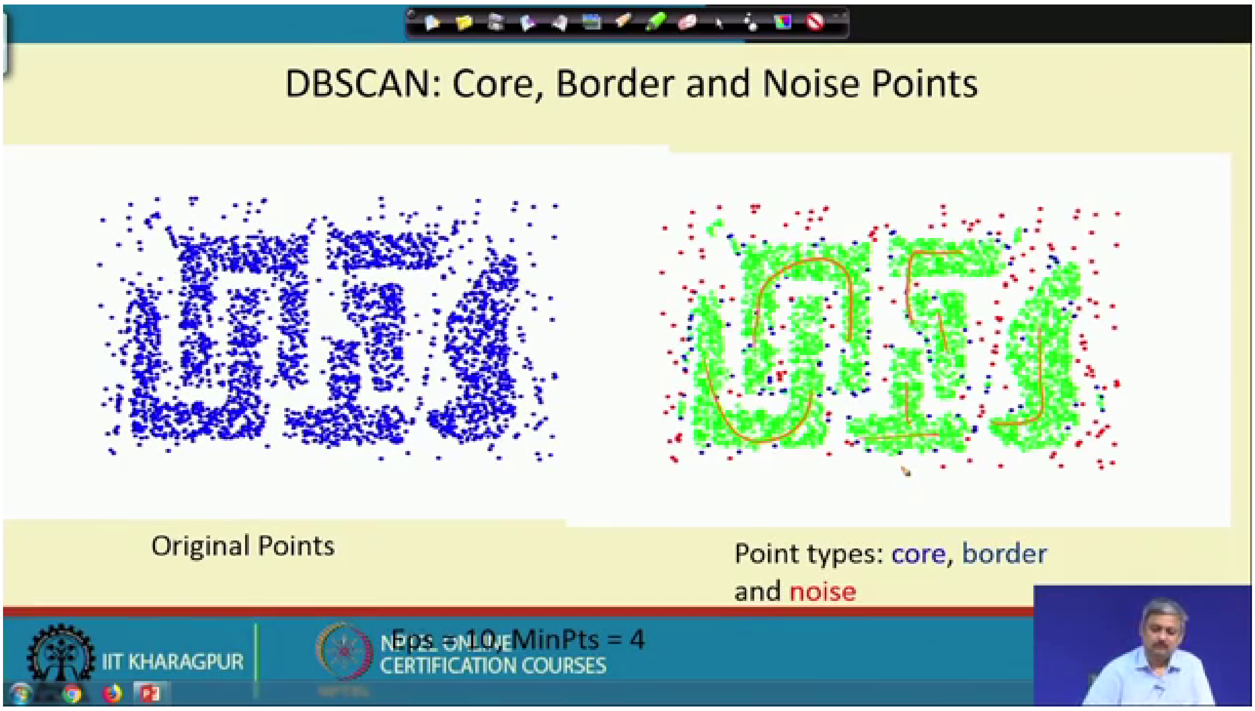
# DBSCAN Algorithm

- Eliminate noise points
- Perform clustering on the remaining points

```
current_cluster_label ← 1
for all core points do
    if the core point has no cluster label then
        current_cluster_label ← current_cluster_label + 1
        Label the current core point with cluster label current_cluster_label
    end if
    for all points in the Eps-neighborhood, except i^th the point itself do
        if the point does not have a cluster label then
            Label the point with cluster label current_cluster_label
        end if
    end for
end for
```

So, this is written down [noise] as an algorithm [noise] [vocalized-noise] which is like this [vocalized-noise] the steps [noise] [vocalized-noise] you can go through for each of the core point a cluster add or add them up; reachable then remove them ah cluster the rest remove them cluster the list till you left with only the noise ok.

(Refer Slide Time: 24:10)



# DBSCAN: Core, Border and Noise Points

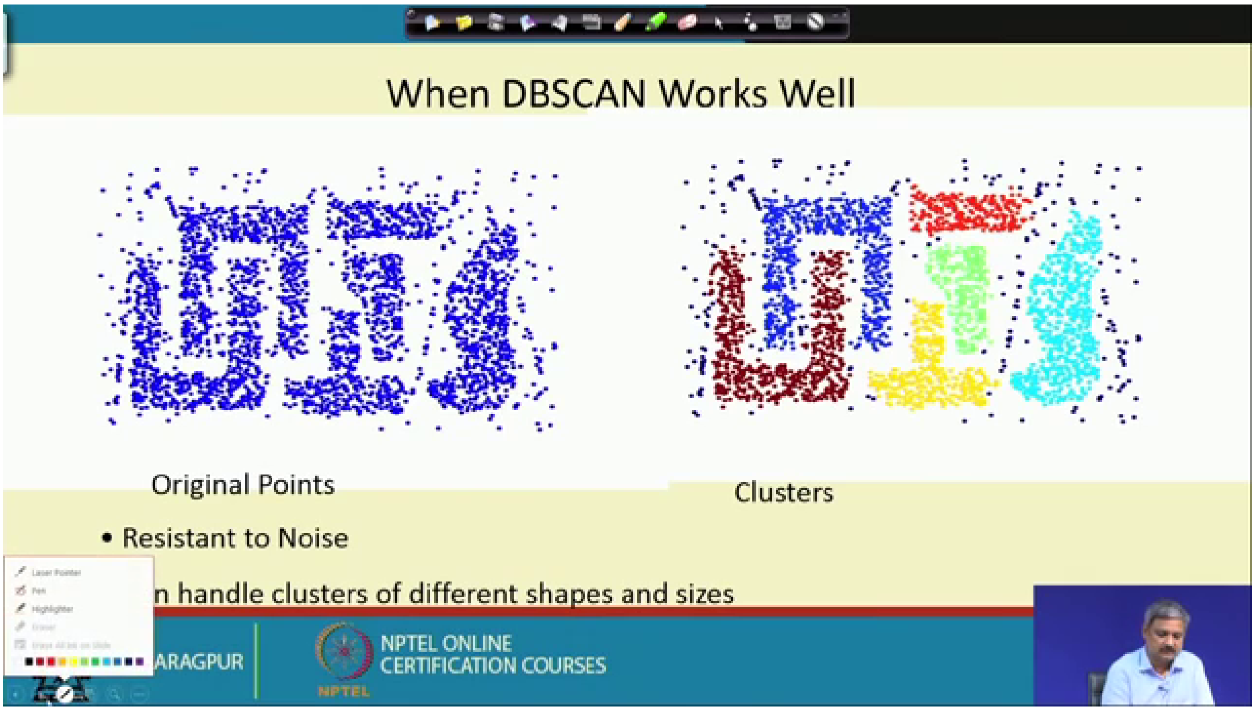
Original Points

Point types: core, border and noise

Eps = 10, MinPts = 4

So, [noise] example of that how we get this ok core border and noise ah these are the noise, these are the [noise] clusters sorry [noise] clusters [noise] these are the noise ok [noise].

(Refer Slide Time: 24:46)



When DBSCAN Works Well

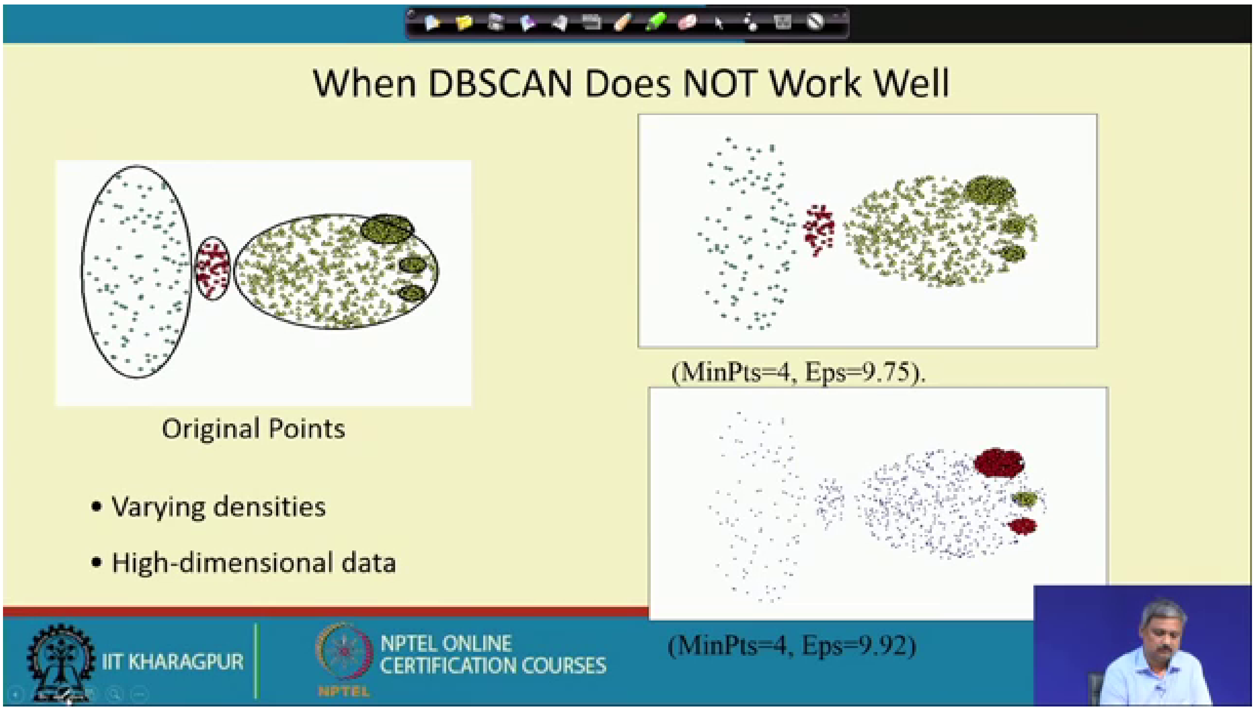
Original Points

Clusters

- Resistant to Noise
- n handle clusters of different shapes and sizes

So, you can see you can get arbitrary separate cluster [noise], but this will work only when you have [vocalized-noise] ok. So, you can get arbitrary say it is resistance to noise also because noise are separated.

(Refer Slide Time: 25:08)



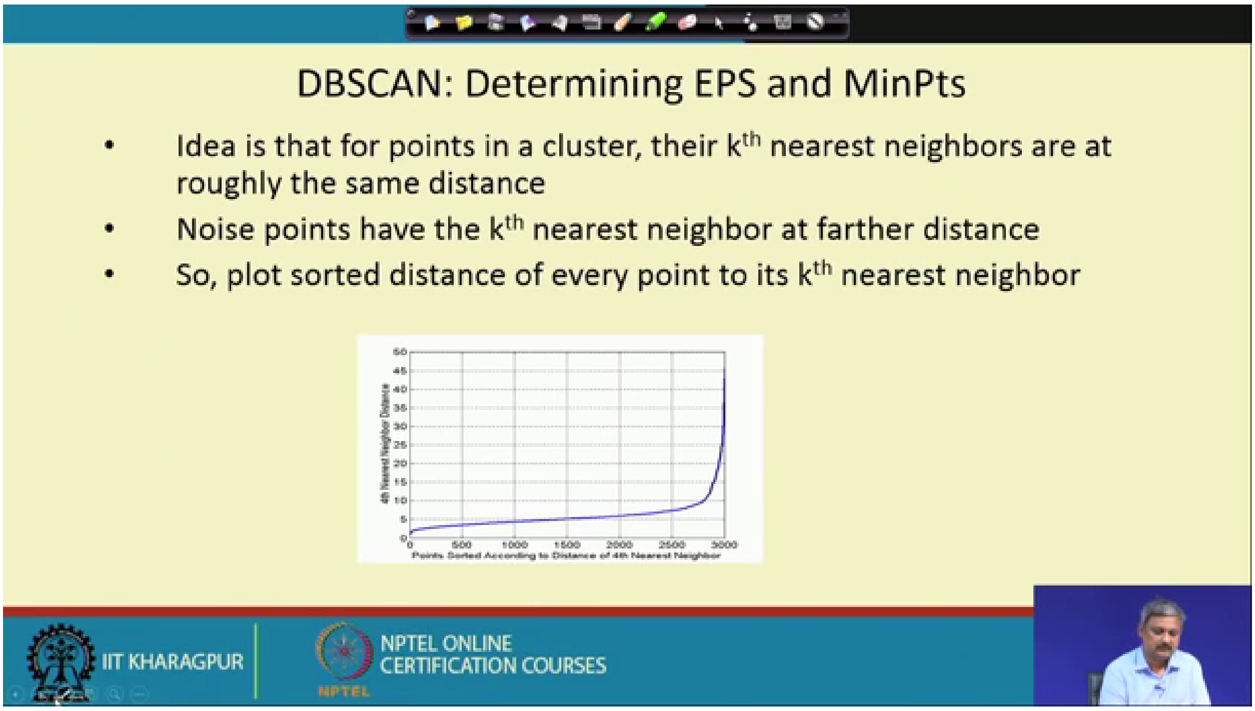
## When DBSCAN Does NOT Work Well

- Varying densities
- High-dimensional data

But this works only when you have low dimensional points; high dimension it does not work ok. So, [noise] also it works when the density is widely varying it does not work some parts are very dense some part are you cannot set them in pts properly alright.

So, in these two cases it does not work.

(Refer Slide Time: 25:35)



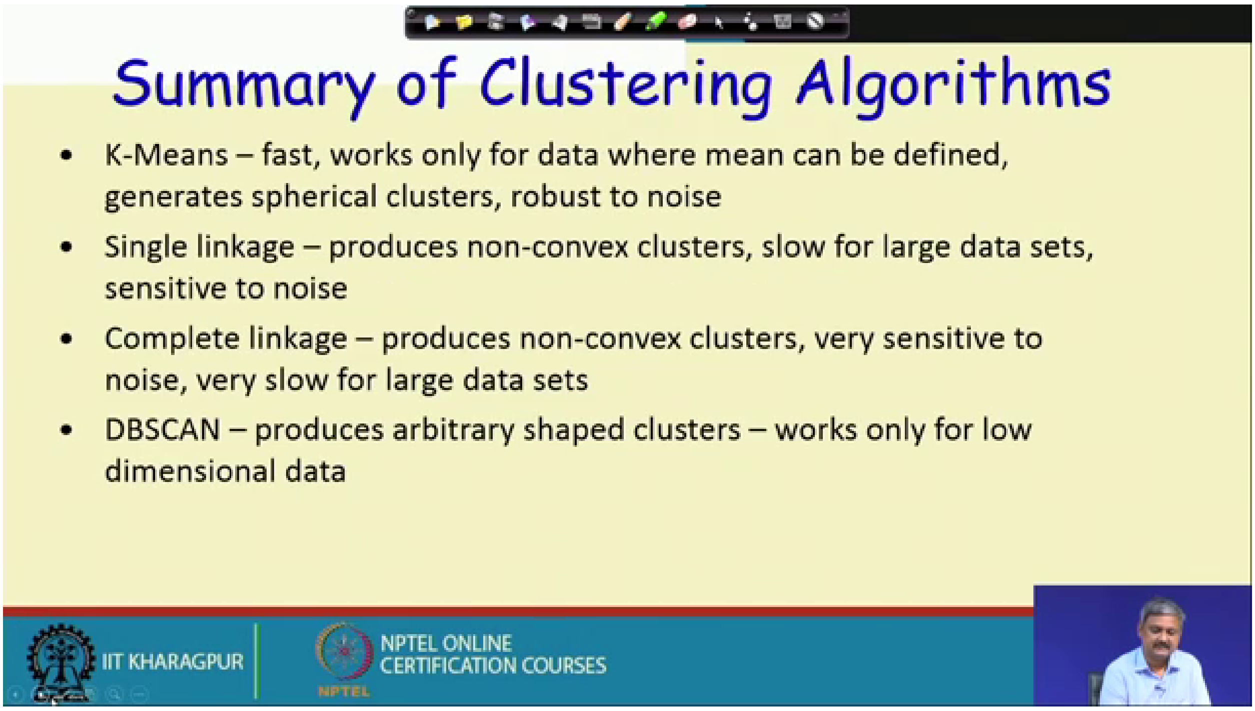
## DBSCAN: Determining EPS and MinPts

- Idea is that for points in a cluster, their k$^{th}$ nearest neighbors are at roughly the same distance
- Noise points have the k$^{th}$ nearest neighbor at farther distance
- So, plot sorted distance of every point to its k$^{th}$ nearest neighbor

4th Nearest Neighbor Distance
0
5
10
15
20
25
30
35
40
45
50
0
500
1000
1500
2000
2500
3000
Points Sorted According to Distance of 4th Nearest Neighbor
IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
NPTEL

So there are different ways of ah estimating a good min pts EPS that I am not going through.

(Refer Slide Time: 25:40)



## Summary of Clustering Algorithms

- K-Means – fast, works only for data where mean can be defined, generates spherical clusters, robust to noise
- Single linkage – produces non-convex clusters, slow for large data sets, sensitive to noise
- Complete linkage – produces non-convex clusters, very sensitive to noise, very slow for large data sets
- DBSCAN – produces arbitrary shaped clusters – works only for low dimensional data

So, this is the summary of the DB SCAN. So, that is the thing it works well for arbitrary cluster noise does not work when there is high dimensional data varying density. It is very much used in spatiotemporal data analysis and roll dimensional data and and ah where you want arbitrary separate cluster ok. So, with this I stop my there are many modifications to db scan ah which ah people have tried different ways, but this is the basic idea.

So, with this I stop this lecture; in the next class I will elaborate on how to evaluate clustering algorithms.

Thank you [noise].