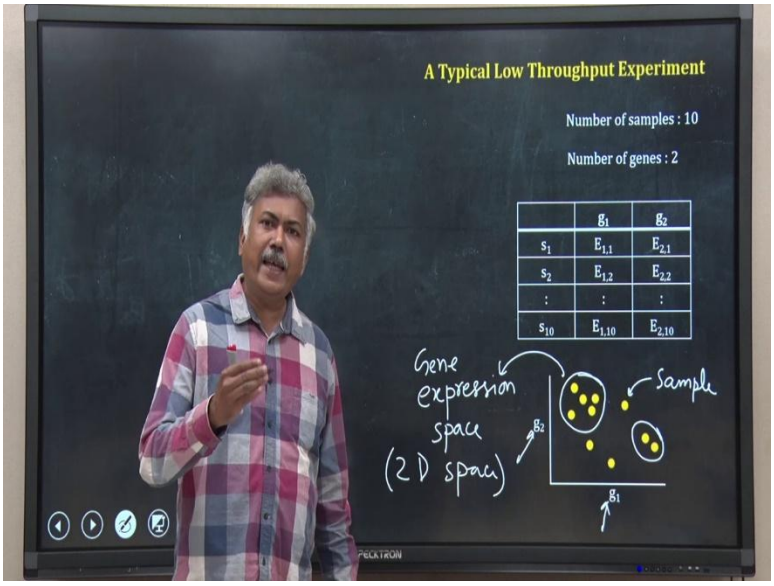
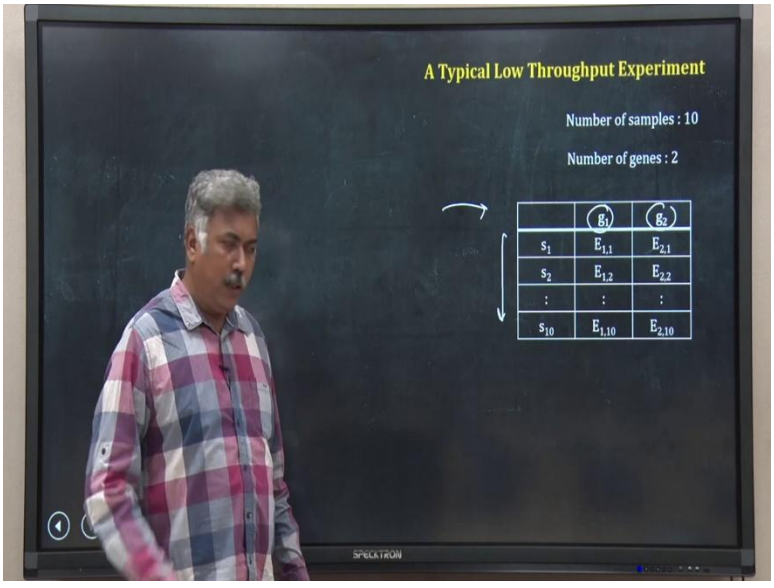
Data Analysis for Biologists
Professor Biplap Bose
Department of Biosciences and Bioengineering
Mehta Family School of Data Science and Artificial Intelligence
Indian Institute of Technology, Guwahati
Lecture 44
Higher – Dimensional Data in Biology

Hello, everyone, welcome back.

(Refer Slide Time: 0:42)





	g_1	g_2
s_1	$E_{1,1}$	$E_{2,1}$
s_2	$E_{1,2}$	$E_{2,2}$
:	:	:
s_{10}	$E_{1,10}$	$E_{2,10}$

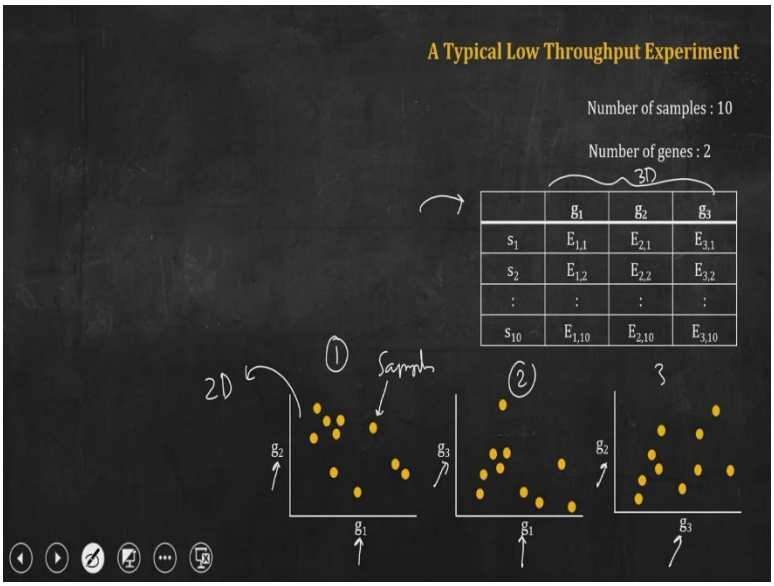
Suppose I am doing a real time PCR experiment. And I have ten samples, maybe ten tumor samples. And I am measuring the level of expression of two genes only. And I am doing real time PCR and maybe this is the table or tabular format for that data. So, this is my table. And this is gene 1 is one gene, gene 2 is another gene, and you have the sample from sample 1 to sample 10.

Now, I want to visualize this, I want to see the gene expression behavior of these 10 genes, 10 samples, and I have measured for two genes. So, what I can do, I can represent these data in a scatterplot, very easy to do, you have already learned how to plot a scatter plot. So, what I will do, I will have gene 1 suppose in the horizontal axis and the expression of gene 2 in the vertical axis. Now, I have 10 yellow dots here, each of this is a sample.

So, this whole space, where I have shown the position of each of these sample, each of these yellow dot, I can call this is Gene expression space. And what is the dimension of the space? I have two genes, the expression level of two genes shown here in two axes, horizontal axis and vertical axis. So that mean, this is a two D space.

So this is how usually you visualize data. And then you may use some statistical test or some other analysis, for example, cluster analysis to find, cluster in these samples. So these are maybe one cluster, this may be one cluster, something like that. So you can perform conventional data analysis on this data. Now, let me add another gene in it, third gene.

(Refer Slide Time: 2:46)



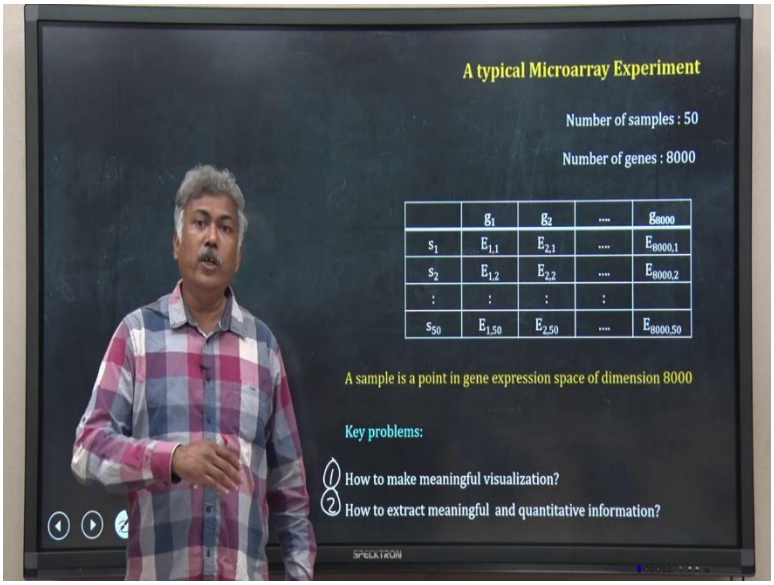
So I am now doing the same experiment as the 10 sample, but other than two genes, I am measuring the expression of 3 genes, g1, g2 and g3. Now, again, I have this tabular data. Now, how should I visualize it, because remember, that I can visualize in three dimension, that is the maximum I can visualize, but usually, we feel more comfortable when we visualize something in two dimensions.

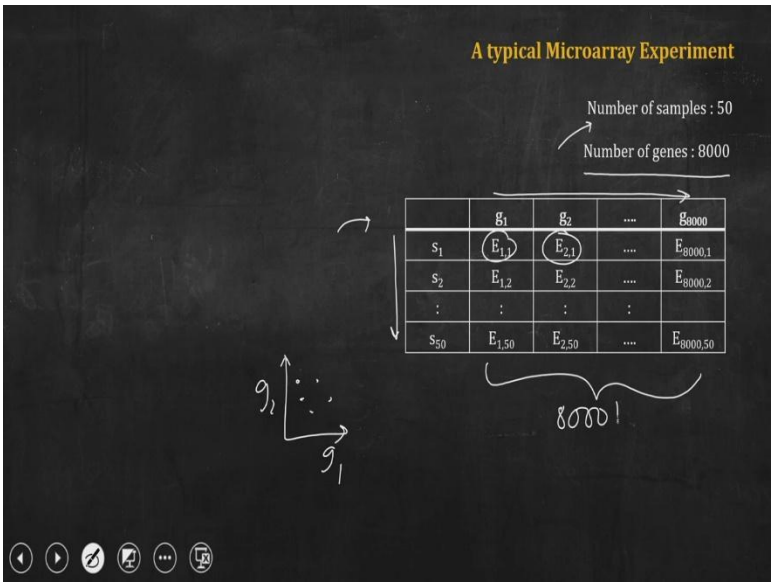
So, you may try something like what I have done here, so I can take any of these two genes. For example, in this first plot, what I have done, I have taken gene 1 and gene 2. So I have a 2D space here, involving two coordinates, one for gene one another from the gene two, and in that I have all these 10 samples. So this is the way you can see the variation of gene one and gene two expression simultaneously in 10 samples.

Whereas, and the next plot could be you take g1 and g3. So this may be your second plot. And this may be your third plot where you take g3 and g2. So what you are doing, you are embedding you are putting your samples in multiple two dimensional space, although you have actually three dimensional data, you have three dimension, because you have three genes that expression you have measured, gene one, gene two and gene three, but it is easier to visualize in two dimension.

So I am breaking down the data and visualizing it in two dimension and in this case also I can actually perform conventional analysis very easily, there is no big deal in this. Now, let me scale up this experiment. I do not want to measure gene expression of 2, 3 or even 4 genes, I want to measure in hundreds and thousands simultaneously. And if I say that immediately it comes to come to mind that maybe you should go for a microarray experiments.

(Refer Slide Time: 4:48)





	g_1	g_2	….	g_{8000}
s_1	$E_{1,1}$	$E_{2,1}$	….	$E_{8000,1}$
s_2	$E_{1,2}$	$E_{2,2}$	….	$E_{8000,2}$
:	:	:	….	:
s_{10}	$E_{1,10}$	$E_{2,10}$	….	$E_{8000,10}$

Let us see a data, here is a hypothetical data. So I have now, suppose 50 samples, maybe 50 tumor samples, and then you are doing a microarray where you can measure the change in gene expression for suppose 8000 genes, I am talking of a lower side of microarray you can go higher also, you can have even ten thousand, twelve thousand, twenty thousand genes in a microarray possibly, but I am saying suppose it has 8000 genes can be detected in this microarray.

So, I have a gene expression microarray, now, that data if you will visualize in a tabular format, will look something like this. So, I have 8000 columns and I have 50 rows, each of these is a expression measure coming from my microarray experiment. Now, how should I visualize it? Let me get back in the earlier case where I had two genes and 10 sample I am creating a gene expression space which is two dimensional.

So, you had g 1 here and g 2 there and each of these points is a sample. So, your sample's positions are embedded in a two dimensional space you can visualize you can do analysis everything. Now, in this case, if I consider each of the samples as a dot in a gene expression space, what is the dimension of my gene expression space now?

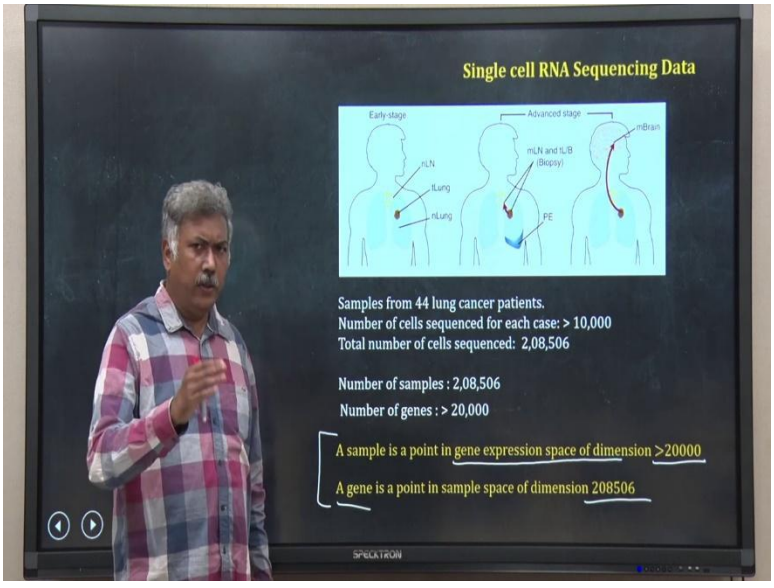
Earlier I had two genes, so, it was two dimensional, then I gave the example where I have three genes, so, I had three dimensional, but in this case I have 8000 genes that means, it is a 8000 dimension space, gene expression space. So, that means, if I somehow can plot it imagine I cannot visualize it, but you can suppose I want to plot it somewhere somehow, that means, in that plot a sample will be a point in a gene expression space of dimension 8000.

Obviously, you cannot visualize it. So, once you have this type of higher dimensional space, where you have that data, you land up with two typical problems, what are those? Obviously, the first problem would be the problem in visualization, how should I visualize. If I have three dimensional space, a space of gene expression, I could have created three plot the way I have shown, one for gene one gene two, one for gene two gene three and other one.

But if I have now 8000 genes, I cannot break them down and create so many two dimensional plot that will be meaningless, nobody will be able to visualize that and see that and make meaning out of it. At the same time, I cannot go beyond three dimensional visualization also. So, what should I do with this such a huge dimensional data set? You cannot visualize it directly.

This is the first problem. And the second problem is, which is not so obvious, but it is a recurring problem is that when you were dimension explodes like these from two to three to 8000, 20,000 then the conventional mathematical methods face problems. So, we cannot use them easily. So, now, I have a data set which is very rich in information, but neither I cannot visualize it, neither I cannot properly analyze them mathematically using conventional tools. Let me move further, give you another example of a huge higher dimensional data.

(Refer Slide Time: 8:30)



This is a example of single cell RNA sequencing, RNA sequencing is becoming almost now a gold standard to study gene expression and you can do that at single cell level. So, what these

people have done in their work is that they have taken samples from 44 lung cancer patients and then from each of these sample each of these patients out of 44 they have taken biopsy samples.

So, they have large number of cell, on an average the number of cells per sample is bigger than 10,000. And as per their report the total number of cells that they sequenced in their experiment by RNA sequence method for all these 44 Lung cancer is two lakh eight thousand five hundred and six. So, you can consider I have two lakh eight thousand five hundred and six samples, each of these cell is a sample and what they did, they did RNA seek to measure the gene expression level in each of these cell for how many genes?

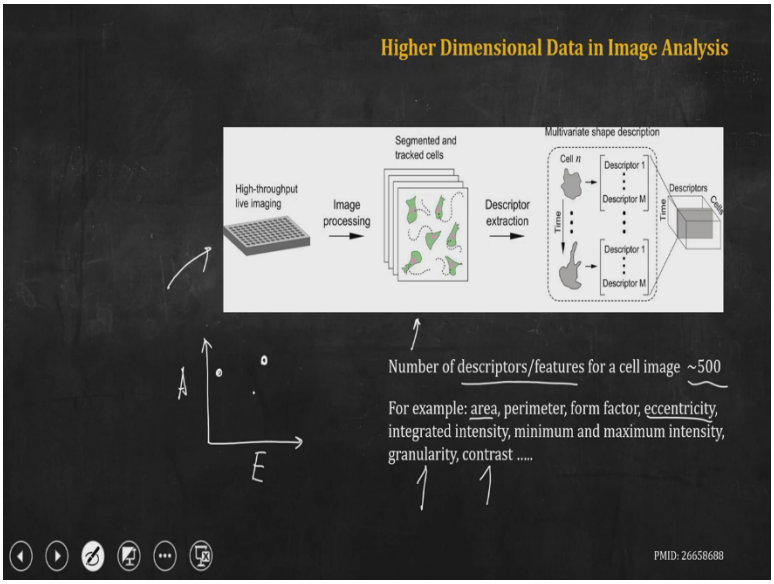
More than 20,000 genes. So, I have more than two lakh samples and what is the dimension of gene expressions space, its more than 20,000. So now, let me look into very carefully what is the dimension of this data? Now, if I consider each of the sample as a point in the data space, then the dimension is greater than 20,000.

So, the each data point is sample is in a gene expression space of dimension greater than 20,000, in two dimensional gene expression experiment I have two axes, two coordinate, gene one and gene two now, you have more than 20,000 coordinates, it is a huge dimensional system. You can invert it also, you can now say that samples are variable.

Each sample is variable and each I want to know the behavior of each gene in the sample space, then what is the dimension. So, the dimension will be then this two lakh eight thousand five hundred and six. I have a space where I have two lakh eight thousand five hundred and six coordinates and 1 dot, 1 point of data in that space, that sample space is one gene whatever way you look this data, its a huge dimensional data set.

And obviously, you cannot visualize this data using conventional plotting visualization techniques. And you will end up in trouble in analyzing this data mathematically also. Let me give you another example of higher dimensional data and that is from something else not from gene expression.

(Refer Slide Time: 11:35)



This thing what I will discuss now is becoming very popular now, because the microscopy techniques has improved. So, what they have done in this work is that they are doing high throughput live cell imaging. So, what they have, they may have suppose 96 wells, in each well you have mono layer of some cells and then you are treating the cells with some drug or some molecule and then you are doing live cell imaging that means, for each well you are taking multiple snapshot in different position repeatedly at different set intervals.

So, you have a large image data sets and that is shown here. So, you have lots of images and during this time period what is happening because of the drug treatment or because of the molecules that you have added the cells are changing their state they may be changing their gene expression, they may be changing their morphology, they may be moving around.

So, you have now this large number of images coming from the single wells and these are time dependent images. So, the next step usually what you will do from these images, you have to use some segmentation algorithm, so, that the machine can identify each of these cells distinctly and then again these individual cells are alive and many of them may be migrating moving.

So, you should have a tracking algorithm to track how one particular cell is moving from one position to another. Now, suppose you have done that the way they have done it here. Now, if my question is something like this, I know during this process during these experimental time, many of these cells they change their morphology and that morphology change has a biological

meaning there is a biological effect. So, I want to track with time how different cells are not only moving around, but at the same time they are changing their morphology.

Now, morphology when you talk of it, seems so easy to identify if you look into one particular image or multiple cell you may point out, you may say see, look at this all circular cell that cell may not be circular, but I as a human being you intuitively will understand what do you mean by that, but when you are doing mathematical analysis using a computer algorithm, you require precise numerical, quantitative data.

So, there are hundreds of different numerical quantity which we call a descriptor or features which you can extract from these images. Those can be used to define the morphology of a cell. I will give you a few example for example, area maybe one descriptor length, maybe one descriptor radius maybe one descriptor, eccentricity maybe another descriptor granularity contrast, intensity, all these things can be a imaging feature or descriptor for a particular cell morphology.

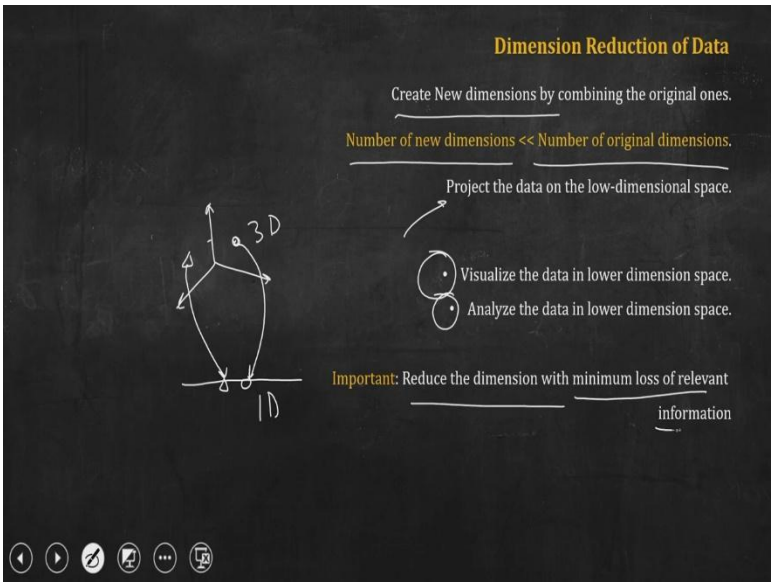
Now, this image analysis algorithm they extract hundreds and thousands of such descriptor, morphological descriptor from each cell. So now, you can imagine I will have a table suppose I have two lakh cells. So, I have two lakh rows and for each cell I have multiple columns, each column has a quantitative value for all these, suppose in this case, almost 500 descriptors or features.

So, all these features define a particular cell. So, what is the dimensionality of this data? Suppose, I want to define a cell just by its eccentricity and area, then I can have a two dimensional space where eccentricity maybe in one axis horizontal one and the area in the vertical axis and one cell will be a single dot here another cell will be single dot here another cells something like that.

So, you may have 10 lakh or 20 lakh dots, all are individual cells in two dimensional space, but in this example, they have almost 500 features. So, the dimensionality of the space this feature space is now more than 500 and in that each cell is embedded each individual cell is embedded. So, again we landed up in a higher dimensional space and we will face the same trouble of visualization and analysis.

So, I gave you three example one from microarray another from the RNA seq and this one is from Image Analysis, all these are high throughput techniques and they generate a huge dimensional data set, very high dimensional data set. So, how should we tackle the problem of visualization and analysis of this higher dimensional data? This is not just unique to biology in many other fields of science and technology, you will find this problem.

(Refer Slide Time: 16:59)



So, the generic solution to handle this problem is to reduce the dimension of data. So, how do you reduce? Suppose, I have 1000 or 500 or 20,000, some value, some very high value of dimensions like that means, the space of the data has so many coordinate axes. So, what you do, your mathematical algorithm what it will do is that it will create new dimensions by combining the real and existing dimension.

That may seems now bit unusual to you. But in separate videos when I will discuss about the method it will be very clear. So what they are doing, what the algorithm, dimension reduction algorithms will do, they will combine the existing dimensions, which are very large in number to create some new dimension. And they will do such a way that the number of new dimension is much lesser than the number of original dimension, that I have to achieve.

And then what I will do, I will project or embed my data from that high dimension to this reduced dimension. As simple as that, if I explain a bit more, suppose I have a three dimensional space. So it is a 3d data. And now I will combine all these three coordinate axis system in such a way, suppose that I become, make a system one dimensional.

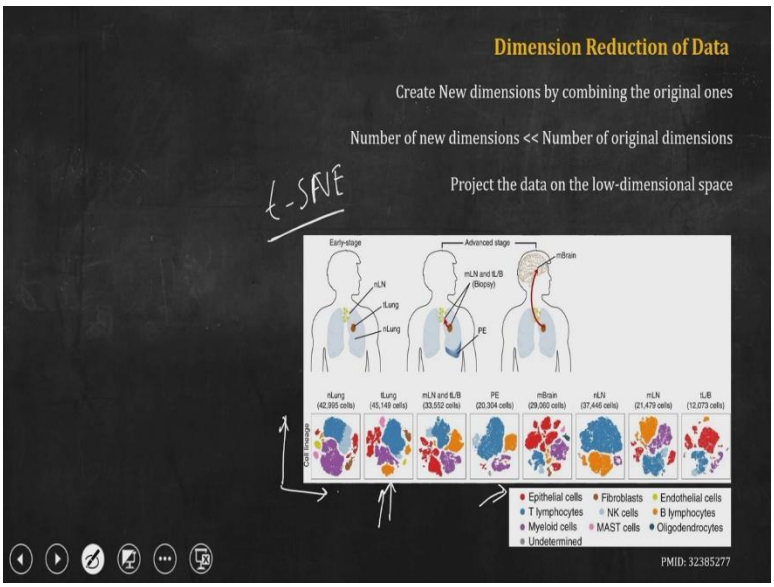
So now, if I have a data point here, I have to project or embed that from three dimension to one dimension, I have another data point here, I have to project or embed that in this new one

dimension. So, what I am doing, I am going from higher dimension to lower dimension where I am projecting or embedding my data set.

Now, once you can do that, once you can do that successfully, then it becomes much easier to visualize the data because now I will not visualize the data in the original higher dimensional space, but rather now I will visualize the data in a lower dimensional space and all the mathematical analysis that I will do, I will not do in the higher dimension space anymore, I will do that in the lower dimension. So, those two problems are now solved.

Now, when you do these type of, use this type of method, which we call dimension reduction method, we have to remember that there may be some loss of information during this reduction of dimension and the algorithm should take care that the reduction in dimension should lead only to minimum loss of relevant information. So, let me go back to that RNA seq experimental work and the image analysis thing to explain that what do I mean in reduction of dimension?

(Refer Slide Time: 19:55)



So what they did, they did, they used a technique in the case called t sne and we will discuss that in another lecture. This is a dimension reduction technique and I have shown a partial data of that analysis. So, what do you have, you have, if you remember we have more than two lakh cells and if I have different samples, so suppose these are all coming from the tumor in lung, the sample

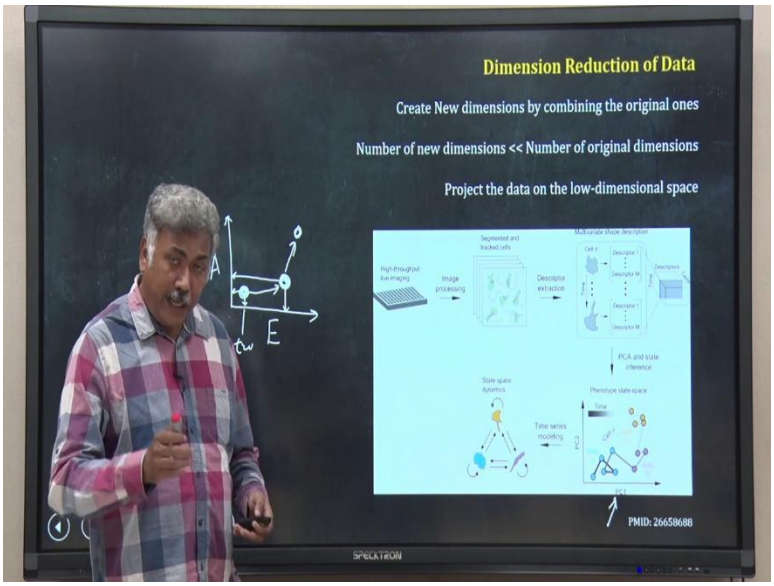
set so, they have 45,000 cells and if you even remember we have done RNA seq for individual cells.

So, in my mind the number of genes that I have sequenced is essentially, the expression measure is more than 20,000. So, originally I have a 20,000 dimensional space which I cannot even imagine, in that then, that space I have dots, each.is one of those 45,000 cells. Now, what they do is they reduce that higher dimensional thing to two dimension using this method called T Sne and they have visualized it here, where I have two dimensions one dimension in this direction and other dimension in this direction and each of the dot which you cannot see them as a distinct dot, because they have 45,000 of them are one cell.

And then what they have done, they have done some other analysis on that, so, that they can now identify what type of cell those are based upon gene expression. So, for example, the colored epithelial cell as red so, on that map data that projected data on this T Sne two dimension is marked by red. So, you can see there are some patch of red cells, which are maybe 10,000 20,000 of them are there they are epithelial cell type.

So, in this way from a 20,000 greater than 20,000 dimensional space, gene expression space, they have reduced the dimension to two dimension and projected or mapped the data on that two dimension. Now, you can easily visualize, now you can perform analysis on this data also. Let me go to the imaging thing.

(Refer Slide Time: 22:13)



In this case, what they are using, they are using another dimension reduction technique called principal component analysis. And we will learn that also in subsequent lectures. So, what you have, if you imagine initially that I have suppose, suppose what I am doing, I have just two descriptor feature eccentricity and area. And if you remember, they have time dependent data, they are doing live cell imaging.

So, suppose in one cell is here it is initially at t equal to zero, t equal to zero, its eccentricity is this value, and area is this value and as time passed, due to the drug treatment or treatment with the molecules they have used, the cell is changing its morphology and suppose in two hour it has moved somewhere in this space. So, at two-hour time the eccentricity this one has increased, but area has little increased. So, now, in this way, I can say then the trajectory of this cell in this two dimensional space is this one, and then maybe at four hour it has moved here like that.

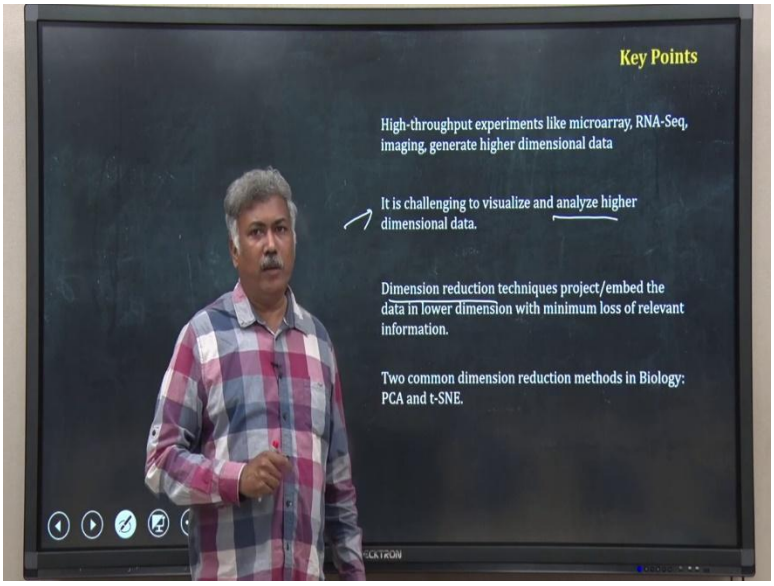
But the problem for me, the dimension of the original data is not two, I have more than 500 descriptor or feature to identify the morphology of the cell. So, I have a five dimensional space and I want to use that data to understand the trajectory, time trajectory of individual cell, how they have changed in each of these samples to 20,000 to or two lakh cells, how they are changing their shape with time.

I cannot do that. So, what they have done, they have used PCA principal component analysis to reduce the dimension to two or three dimension. For example, they have shown here the two dimensional representation of the data, where now the data is in two dimension, again the same way I have that each of these dot is one cell and the trajectory of that cell in this new space is shown.

So, this new space is two dimensional and it has two coordinate and these two coordinate, these two axes has been created by combining the existing 500 dimensions. Now, you can visualize each cell as a dot in the space you can map the time trajectory of each of the cell in the space and what they did. Eventually they create a mathematical model to understand the dynamics or more change in morphology of each of these cell.

So, again in this case, they successfully use dimension reduction technique to reduce that hugely higher dimensional data to a lower dimension two dimension where they can perform visualization as well as a mathematical analysis. So, that brings me to the end of this lecture. So, let me jot down what we have discussed in this lecture.

(Refer Slide Time: 25:23)



In this lecture, I discussed about higher dimensional data in biology and the problems associated with that. As we are moving more and more to automated high throughput experiments in cell and molecular biology, we are generating high throughput higher dimensional data. For example,

you can take microarray RNA seq or live cell imaging, as example, we discussed in the lecture today.

These high dimensional data has two unique challenges. The first one is that they are difficult to visualize, and at the same time, they are difficult to analyze in that higher dimensional space. And that is where the techniques or algorithm for dimension reduction comes. What these techniques will do?

These techniques will reduce the dimension of the data space to very lower dimension such that we do not have much loss of relevant information. And I have given two example, from the publication where in one case they have used T Sne and in other case they have used PCA and in subsequent lecture, I will discuss both of these algorithms. That is all for this lecture. Thank you for learning with me today. See you in those lecture of dimension reduction.