Data Science for Engineers
Prof. Ragunathan Rengasamy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture - 29
Introduction to Data Science

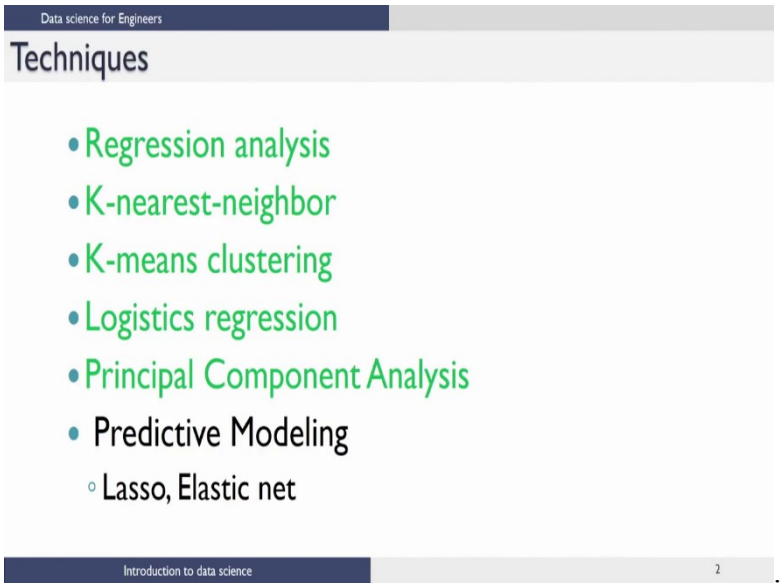
Now, that we have looked at the fundamentals required to understand data science in terms of linear algebra statistics and optimization, we are going to start series of lectures where we introduce data science, describe different techniques that are used in data science and finally, end with one practical industrial example of use of data science. While we introduce the techniques we will also use smaller examples to illustrate how the technique might be used in a data science problem. At the end of this course there will also be a case study for the participants to practice.

So, this is the first lecture on introduction to data science. Before we jump into the techniques I would like to introduce some interesting ways of looking at data science and in a broader context understand what these techniques are doing and how one should think about data science problems. One could teach these techniques as disparate set of methods to solve data science problems; however, the critical thing is in learning how to use these techniques for real problems which is what we call as problem formulation.

What we will do in this course is introduce the participants to a data science problem solving framework, very short lecture on that, to give you a view of how one should think about general data science problems and how you convert a problem you know which is not well defined into something that is manageable using the techniques such you learn in this course.

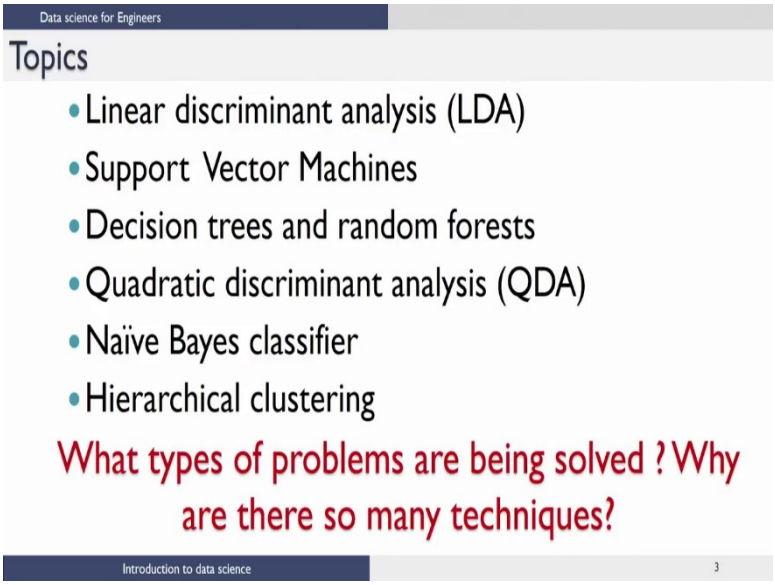
So, let me start with this laundry list of techniques that people usually see when they look at any curriculum for data science or any website which talks about data science or many books that talk about data science. I have just done some colour coding in terms of the techniques that you will see in this course in green.

(Refer Slide Time: 03:05)



And other techniques are out there which we will not be teaching in this course, but which would be a part of more advanced course. So, there are techniques such as Regression analysis, K - nearest - neighbour, K - means clustering, logistics regression, Principle component analysis, all of which you will see in this course then people talk about Predictive modelling under that there are techniques such as Lasso, Elastic net that you can learn.

(Refer Slide Time: 03:38)



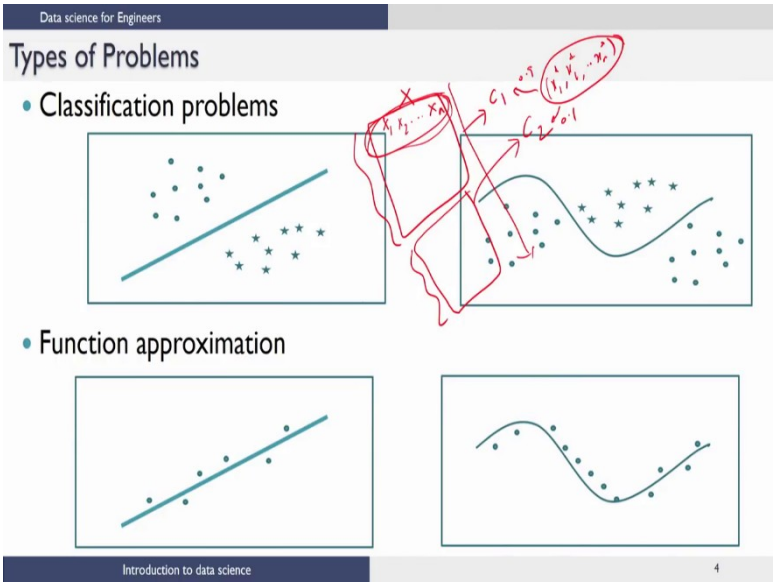
Then there are techniques such as Linear discriminant analysis, Support Vector Machines, Decision trees and random forests,

Quadratic discriminant analysis, Naive Bayes classifier, Hierarchical clustering and many more such as deep networks and so on. So, to get a general idea of data science one might be tempted to ask that if all of these collections of techniques solve data science problems then, one would like to know what types of problems are being solved really and once one understands the types of problems that are being solved then, the next logical question would be - do you need so many techniques for the types of problems that you are trying to solve.

So, this would be typical questions that that one might be interested in answering. What I am going to do is I am going to give you my view of the types of problems that are being solved and why there are so many techniques to solve these types of problems. Since this is a first course on data science for engineers we going to cover major categories of problems that are of most interest to engineers. This is not to say that other categories of problems do not exist or that they are not interesting.

We will keep this viewpoint in the background as we go through the lecture materials of this course. Other than this one could also think of statistics as useful by itself for data science problems. Statistics is also intricately embedded into the data science techniques in terms of the formulation themselves and also in characterizing the properties of the machine learning techniques.

(Refer Slide Time: 05:39)



So, in my mind fundamentally I would say that there are mainly 2 class of problems that we solve in data science. So, I am going to call these as classification problems and function approximation problems.

So, let us look at what classification problems relate to so these are types of problems where you have data which are in general labeled and I will explain what label means and whenever you get a new data you want to assign a label to that data.

So, typical example of this type of problem is called a binary classification problem which is used in many applications I will point out 2 applications for example. In this type of problem what you have is data we will call data x. This data could have many attributes let us say x_1 x_2 all the way up to x_n this is something that we saw in linear algebra and so on and in binary classification problems what you have is you have a group of data which you say can be assigned a label let us say c 1 and I will explain why I use the term c, c refers to the class to which this data belongs and then another block of data with the same attributes may be labeled as c2.

So, now the data science problem is the following if I give you a new data point let us say x_1 star x_2 star all the way up to x_{n*}, the algorithm should be able to classify and say this point is likely to have come from either class 1 or it could have come from class 2. So, assigning a label to this new data in terms of what is the likelihood of this data having come from either class 1 or class 2 is the classification problem. Let us say if you assign the likelihood of this coming from class 1 as 0.9 and from class 2 as 0.1 then one would make the judgment that this data point is likely to belong to class 1.

Now, let us see how this is useful in a real problem. So, I will give you 2 examples one example is something that people talk about all the time nowadays which is called fraud detection. So, let us take one particular case of fraud detection for example, so, whenever we go and use our credit card we buy something and the credit card gets charged. So, let us say there are certain characteristics of every transaction that you record such as the amount the time of the day the transaction is made the place from which the transaction is made the type of product that is bought through the transaction and so on. So, you can think of many such attributes. Let us say those are the attributes that characterize every single transaction.

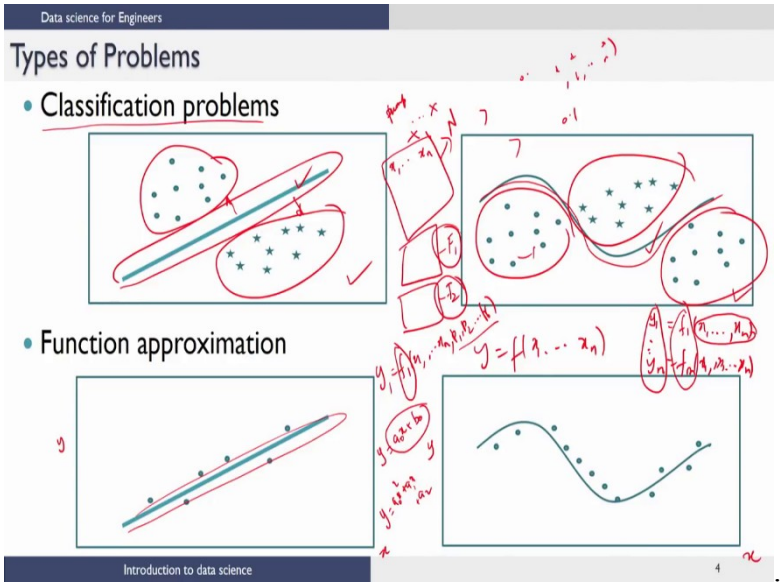
Let us assume that there are many people and they are making transactions and you have transactions listed like this and you find out that of these, these were actually fraudulent transactions these were transactions that was not legal or was not made by the person who owns the credit card and these are transactions which are legal. So, this is something that you label based on exploring each transaction which you think might not be right and actually when you find out that that transaction was not legal then you put it into the basket which is illegal transaction.

Now, if you use a data science algorithm a binary classification algorithm to be able to give the likelihood of a transaction being correct or fraudulent based on this easily calculatable attributes or easily monitorable or measurable attributes. Then, whenever a new transaction takes place you could run it through this classifier and then find the likelihood of this transaction being fraudulent.

And in cases where the transaction has a very high likelihood of being fraudulent, then the company could call that person and then say hey we saw that your credit card was used in such and such a place such and such a time for buying such and such a thing did you actually do this transaction and if you have gone on a vacation to remote place and you made this transaction you tell them I have come to this place on vacation this is the right transaction and so on if not, then you find that transaction is fraudulent and stop the payment. So, this is one example of how a binary classification problem is useful in real life.

Now, when we talked about this data and we talked about the binary classification problem, we talked about just 2 classes x_1 and x_2, but in reality there could be problems where there are multiple classes. One very good engineering example would be fault diagnosis or prediction of failures. Where you might have, let us say a certain equipment a pump or a compressor or distillation column whatever the equipment might be and then the working of that equipment is let us say characterized by several attributes. How much power it draws, how much performance does it give, is there vibration, is there noise, what is the temperature and so on.

(Refer Slide Time: 13:12)



So, now, you could have engineering data x which let us say talks about the characteristic of let us say a pump and the pump is characterized, the operation of the pump is characterized, by let us say several attributes x_1 to x_n. And if you have legacy data or historical data where you have been operating pumps for years and years and then you know that if these variables take values in this block then everything is fine with the pump.

So, I write n for normal and then you could have a block of data and that data might have been the data that is recorded whenever there is a particular type of fault in the pump let me call this fault f one. Then you could have another block of data which could have been seen when there is fault f2 and so on. So, we will just stick to 2 faults f1 and f2. Let us assume these are the only 2 failure modes that are possible. Now you start operating the pump and then at some point you get this data and then you ask the following question. Based on this data would be possible for me to say if the pump is operating normally or is there likely to be failure mode 1 that is the current situation of the pump or is it failure mode 2 that is the current situation of the problem.

So, in this case you see that there are 3 classes n, f1 and f2. So, this is what is called a multi class problem. So, again when a new data comes in we want to label this as either normal f1 or f2 if it is normal, you do not do anything. If it is f1, if it is very severe then you stop the pump and then x it. If it is not very severe you let the maintenance know that this pump is going to come up for maintenance at some time and in the next shutdown of the plant this pump needs to be maintained. So, that is how classification problems are very important in engineering context.

So, we will look at examples of both binary classification and multi class classification as we go through the series of lectures. So, in summary the one type of problem that we are interested in data science is classification and these 2 pictures here show the different types of challenges that we are going to face when we look at classification problems. So, problems where linear equation can be a decision function for us to classify are called linear classification problems or we call these problems are as linear classifiable or linearly classifiable and here we show in 2 dimensions,so, binary classification problem so all of this could be a class 1 and all of these could be class 2 and a line or a plane or a hyper plane could be used to classify this data points.

Now, more complicated problems are where hyper plane or a line might not be enough for us to classify. Here is an example of a classification problem which is non-linear. So, let us assume that this data and this data belongs to class 1 and this data belongs to class 2. However you try to draw a line it would be very difficult ,almost impossible in this case, to classify these 2 classes with just a line in this 2 D picture. However, if your decision boundary are the function that

you are going to use to classify is of this form. So, you see the difference between this and this, this is non-linear, this is linear, then we could easily extend the concepts that we have learnt in terms of the half spaces and so on to do classification for these types of problem using non-linear decision boundaries.

So, you would say if the points are to this side it is 1 class and if the points are to the other side it is another class. One has to do this carefully de ning the equivalent ideas for non-linear decision boundaries equivalent to the linear case very carefully and the minute you move from linear to non-linear then there are a host of other questions that come about. And these questions are really related to what type of non-linear function should one use.

When we talk about linear classifiers the linear functional form is fixed it is very simple. It is only one functional form you have to estimate the parameters of course, but we do not have to really think about what functional form you were going to use. However, if it is a non-linear problem then we really need to choose a particular type of decision function that we need to use and how do you choose this decision functions now the minute you go to the non-linear domain there are infinite number of functional forms that you can choose how do you choose one that works for you it is an interesting and important question that one needs to answer.

So, that is as far as classification problems are concerned, now let us move on to the other type of problem that one solves in data science this is what I would call as function approximation problem. Again, I am showing function approximation problems in 2 dimensional space here. So, I might have an out-put and an input so again in a general case we will have many inputs and many outputs. This is what is called as a case of single input here and a single output. However, you could have many attributes and the output being a function of many attributes.

This is also a function approximation problem or you could also have many outputs which are a function of many attributes. So, this is also possible. So function approximation is the task of nding these functions and whenever we write a function this function is typically parameterized by parameter. So, for example, if you just take let us say one output and then say this is f1, x_1, x_2, x_n these are the attribute values and there will also usually be a set of parameters that you have to use for that function. So, that could be p1 p2 and pr let us say.

So, when I talk about a function approximation problem, then the problem that we are trying to solve is the following. Given several samples of these out-puts and the corresponding attributes that resulted in these outputs. So, this is the data that we are going to talk about and once I have a large amount of this data, how do I find this function

form and once I choose a functional form how do I also identify the parameters that are in the functional form. So, a simple example is if it is a linear functional form then I say $y = a_0 x + b_0$ let us say.

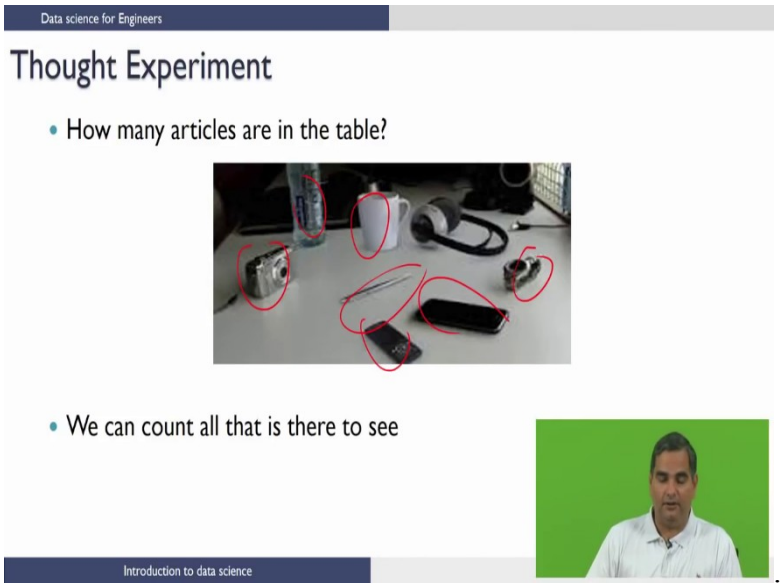
In this case the functional form is linear and the parameters are a_0 and b naught if you assume that it is a quadratic functional form then you could do $a_0 x^2 + a_1 x + a_2$. So, in this case the functional form is quadratic and there are 3 parameters now a_0 a_1 and a_2. So, when you do this function approximation you will have to figure out both the function and the parameters and in classification problems you want to come up let us say in the linear case with a line or a hyper plane where these points are as far away from this as possible. In the function approximation case what you want to do is, you want to find a line or a hyper plane such that these points are clustered around that and this is a linear problem which is what we are going to see in this course as linear regression.

Now the same non-linear version of the pro problem similar to the picture on the top is shown here. Here you want to have a non-linear surface or a curve that goes through these points and these points are clustered around that curve. So, in summary there are really only two types of problems that we predominantly solve from an engineering viewpoint using data sciences, these are classification problems and function approximation problems.

So, if there are only two types of problems that we are really solving then one might ask why are there so many techniques for solving these types of problems and one standard question that comes about whenever someone does data science is if a particular technique is better than another technique and the proponent of one technique will say this is a greatest technique the proponent of other technique will say that is a greatest technique and you know this debate keeps going on and so on.

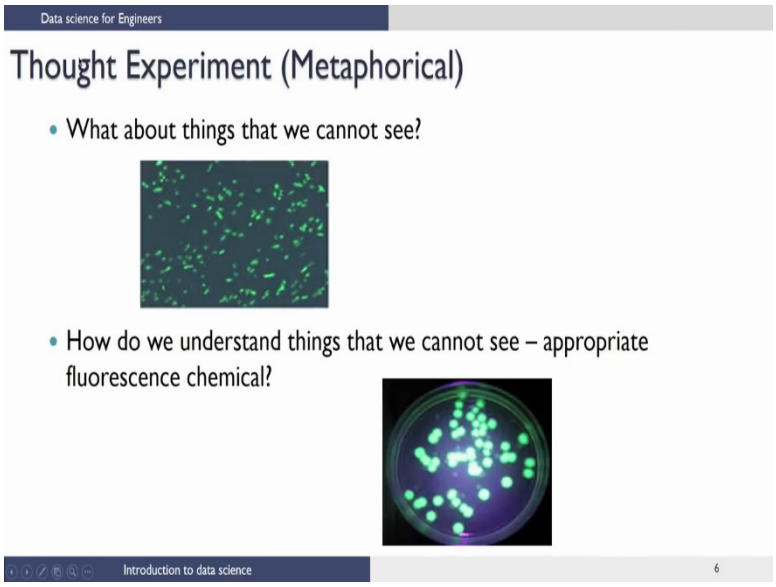
So, I am going to give a slightly different view of why we have so many techniques and you know you can kind of resolve in some sense this question of which technique is better. So to do this let us do a thought experiment.

(Refer Slide Time: 24:44)



So, we have many objects on the table that is shown in the slide, then if I asked you how many articles are there in the table you would quickly say well there is a camera, there is a cup, there are two mobile phones, there is a watch, there is a pen, bottle and so on. So, basically we can kind of count or see whatever there is to see and then enumerate and then say these are the objects or articles in the table. So, in some sense we can count all that is there to see. So, this at this point you will say this is all that is on the table then I asked you the question is that all really that is there on the table.

(Refer Slide Time: 25:41)



And not to take this very literally, but this illustrates the key idea that I want to use when we go back to answering the question as to

why there are so many techniques for data science. So, carrying on I will ask you is that all there is on the table and then if I ask you this question what about things that we cannot really see. So, in the table there might be millions of teaming microorganisms which are not visible to us to the naked eye.

So, if I ask you to enumerate everything that is there on the table you can only enumerate what you can see the things that you cannot see you cannot enumerate. So, let us assume again you have to understand the logic behind what I am trying to explain not to take this too literally. Let us assume that you suspect there could be four different types of microorganisms also that could be on the table, now you cannot see it. So, what you do is just again to do the thought experiment let us say someone came up with some chemical which if you simply spray on you can start seeing these microorganisms.

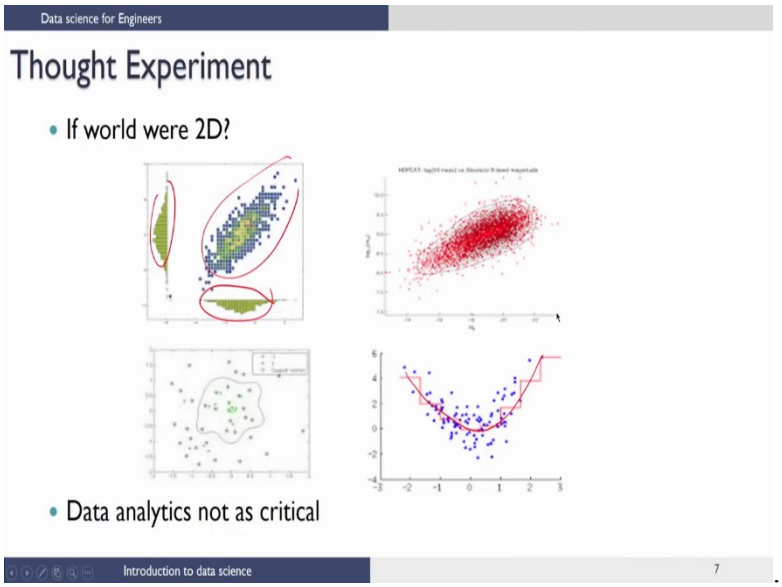
So, let us say there are 4 microorganisms there are four chemicals, now the assumption here is when someone comes up with a chemical like this they have tested it, they have shown that it works for that particular microorganism very theoretically and repeatedly they have shown that it works. So, we cannot re-ally go back and question whether this chemical is good for this microorganism because that has been demonstrated reasonably well.

So, if there are these 4 microorganisms what you would do is. You have to make an assumption as to what exists on the table. So, let us say you make the assumption that microorganism one is what is there on the table. So, you pick up the fluorescent chemical one and then spray it. Now if you see fluorescence and then you would come to the conclusion yes my assumption is right this is the microorganism that is also on the table.

Now, the interesting thing is if it does not for us then the conclusion is not that the chemical is bad because that is been provably shown to work for this particular case, you would only assume that the assumption that you made is not right. So, you have to go back to the next assumption which would be microorganism 2 is there on the table and then you look at the fluorescence chemical 2 and so on.

So, once you do this exercise and let us say when you use chemical 2 and 4 it fluoresced one and 3 did not fluoresce then at the end of that exercise you could go back and then say the articles on the table or the camera and the and the cell phone and so on and also microorganism 2 and 4. Now notice how you have been able to see what you cannot visually see using this assumption validation cycle. So, this is an important thing to understand and I am going to connect this to the techniques and data science in the coming slide.

(Refer Slide Time: 29:34)

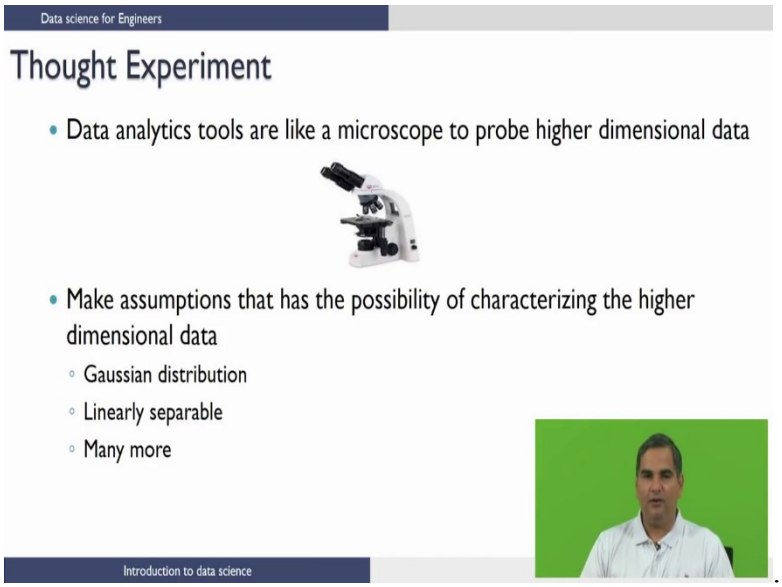


If world were 2 D for example, so all that we need to do could be done with just two attributes and looking at two attributes then whether it is a classification problem or a function approximation problem I can simply visualize it draw whatever I want characterize, and then be done with the problem. So, for example, here I could say this looks like a distribution; it looks like a normal distribution, in the two variables and so on.

If this were a function approximation problem you could really say well a single line will not solve this problem. So, maybe I can use a non-linear function and so on. So, if world were 2 D data analytics is still important because I need to explain the variability and all that; however, it is not as critical as the case where we have many more than two variables or more than two attributes and the situation currently is that for every problem that one tries to solve there is this data deluge.

There are tons of attributes that one could actually measure and monitor and you really want to see how many of these attributes are really going to contribute to the problem that we are trying to solve or how many of these attributes are really important. So, we are going we are going with big data from two dimensions to multiple, multiple dimensions and the question then is how do I understand the organization of the data, in multiple dimensions where I cannot see multiple dimensions I cannot see beyond 3 D.

(Refer Slide Time: 31:28)

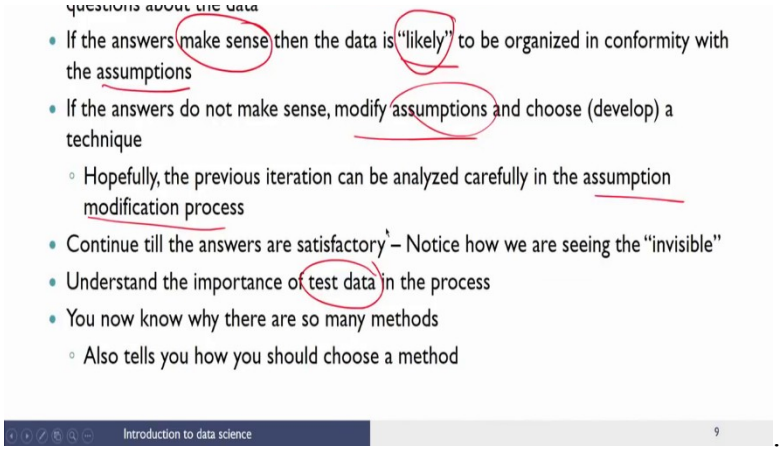


So, how do I do this is a question that we are trying to answer. So, I would think of data analytic tools as microscope to probe higher dimensional data. Data in much more than 2 dimensions that you cannot actually see or visualize and the way we do this is the following. So, we have data and we cannot see it we cannot plot it because there are attributes in the 1000's- 10000's in some cases.

So, what you have to do is much like the microorganism example that I showed you. You have to make some assumptions about the data. To come up with a comprehensive set of assumptions is difficult, but let me explain some of the assumptions that are generally made. You could make the assumption that the data is actually random or data is generated from random process and you could make distribution assumptions such as it is Gaussian distribution and so on. Or you could make assumptions about how the data is organized, such as I think I can use a linear classifier to solve the classification problem in which case we are making the underlying assumption that the problem or the data is structured in such a way that it is linearly separable and there are many more assumptions that you can make it can make combinations of assumptions and so on.

So, you start with multi dimensional data you make these assumptions and then what you do is the following.

(Refer Slide Time: 33:17)



You pick a technique based on the assumptions and this technique should have been proven to solve problems where these assumptions have been made. So, in other words let us say if you make the assumption that the problem is linearly classifiable, then you really want to pick a technique which will work very well, which has been shown to theoretically work very well, for linearly separable problems.

So, this is equivalent to picking the chemical that has been shown to make a certain type of organism fluoresce. So, you choose the technique and then you deploy this technique and if the answer makes sense and we will see what it means when we say make sense mathematically from a data science viewpoint then the data is likely to be organized in conformity with the assumptions that you have made.

So, important the key, it is important to look at the key words that we are using it is "likely" to be organized and assumptions are important. So, likely would mean we will have to do some metric and different people will use different metrics and different levels of satisfaction of that metric to be convinced that what they have is right and wrong. So, that is where subjectivity comes in, but if the answers make sense then we will say the data is likely to be organized in conformity with the assumptions.

If the answers do not make sense, then typically the tendency is to blame the technique it is really not the technique that is a problem, the problem is the assumptions that we have made because we are not able to see this data in multiple dimensions. So, what you should do is you should modify the assumptions and choose our develop if you are a data scientist a technique to solve this problem if these assumptions were true.

Now, hopefully the previous iteration where you actually use some assumptions and saw that the assumptions were violated and that it was not likely that those assumptions are the one that are valid for this

problem. Even though you failed in that attempt you still got something out of it which would help you in modifying the assumption. So, this assumption modification process could be done with more knowledge from failed attempts from before.

Now you continue with this process till the answers are satisfactory and notice in this process how you are seeing the invisible. So, you are able to see data in n dimensions. So, for example, you cannot clearly see hundred variables plot them and then see whether they are linearly separable or not. But if you use a linear classifier and it worked very very well then you know that the data is likely to be organized in such a way that a hyper plane could separate this data into two groups. So, you have started seeing the invisible much like the thought experiment we did with the table case.

Now, this question of likely and makes sense are very important. So, how do I ensure that I test to see whether the results that I have are good enough or not. That is done using test data in many of these data analytic techniques. So, the test data is very important when we do this exercise and as we teach different techniques you will you will see how this is important and will explain this in greater detail.

Now ultimately what I want to point out is the following now we have an answer for why there are so many methods. There are many types of assumptions that you could make and for each of these assumptions based on the assumptions you could come up with techniques which would work very well if those assumptions were true or the data was organized in a way the algorithm assumes it is organized.

So, since there are so many assumptions, there are many many combinations of assumptions you can make. There are many techniques which are ne tuned and developed particularly to solve problems where data is organized according to the assumptions that are used in the technique. So, that is the reason why you have so many techniques. So, in some sense when you look at all of these techniques it is not as important or as interesting to compare these techniques blindly in terms of this is better than the other one and so on.

But it is more important from a data science perspective from a learning and understanding data science perspective to look at each technique in terms of the assumptions that it makes about the problem that is being solved and once you have a mental map of the assumptions that the technique uses to solve a problem and the technique, then you are in a good situation to be able to use a particular technique or a group of techniques for solving a particular problem so this is important to keep in mind.

So, in this lecture the first introductory lecture on data science I wanted to right away address the questions of the type of problems that we solved and in summary most of the problems that you solve in data science can be categorized as either classification problems or function approximation problems that is one take home message from this lecture and the other message is that there are several techniques for solving these data science problems we wanted to know why there are so many techniques.

So, I gave you a slightly different perspective on these techniques in terms of them allowing us to see or visualize or characterize or explained data in multiple dimensions. So, you start seeing data in multiple dimensions which is not possible otherwise. What we will do in the next lecture is to get some of these ideas into a notion of a framework for solving data science problems and I will illustrate that framework using one activity in data science problems which is used in many many problems this is in general the first step in many data science problems which is called data imputation.

So, I will describe a framework for solving data science problems and use this data imputation as an example to explain what that framework is and how does it work. And as part of that process, we will also see how we use this assumption validation cycle within the framework to be able to choose the best technique to solve the problem that you are interested in.

So, I will see you again in the next lecture on the use of a framework for solving data science problems.

Thank you.