Introduction to Data Analytics

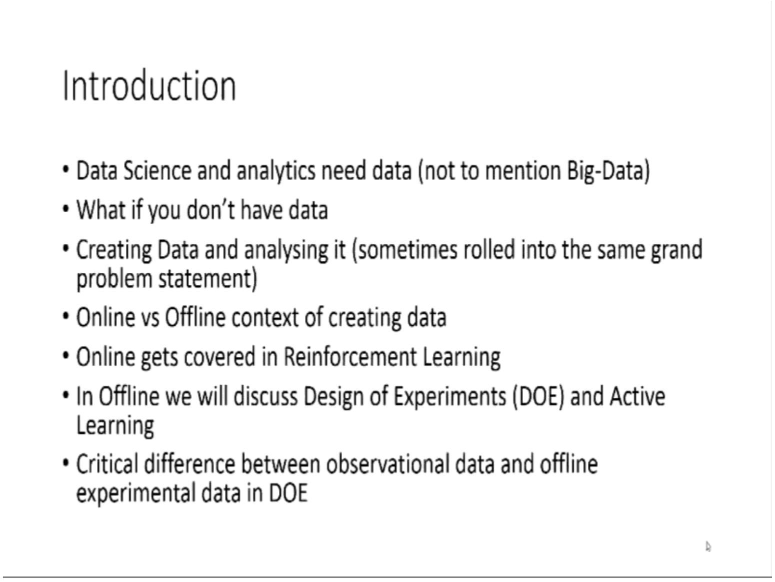
Prof. Nandan Sudarsanam and Prof. B. Ravindran
Department of Management Studies and
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module – 08
Lecture – 43
Introduction to Experimentation and Active Learning – 1

Hello and welcome to our first lecture on Introduction to Experimentation and Active Learning. This is the first lecture of a two part series on these topics and in this lecture we intend to motivate the use of these techniques as well as the concept of reinforcement learning. So, the reinforcement learning is a lecture, a separate lecture that you will have from Professor Ravindran.

And in this lecture we will motivate the need for these topics experimentation, active learning, reinforcement learning, by widely we talking about them in a data analytics course and we will also briefly introduce the topic of experimentation or design of experiments and in the next lecture, we will continue with experimentation and end with active learning.

(Refer Slide Time: 01:04)



Introduction

- Data Science and analytics need data (not to mention Big-Data)
- What if you don't have data
- Creating Data and analysing it (sometimes rolled into the same grand problem statement)
- Online vs Offline context of creating data
- Online gets covered in Reinforcement Learning
- In Offline we will discuss Design of Experiments (DOE) and Active Learning
- Critical difference between observational data and offline experimental data in DOE

So, let us get into the subject, where we take a broader look into data science and analytics. The core idea is that data science and analytics need data and if you were to go with the big buzzwords now, we might even need big data, you know where we can really gain useful insights and this is quite fairly motivated with the easy availability of storage, the easy availability to process data and the internet of things generating a lot of data. It is quite easy to get choose of data and analysis the data and come up with useful insights.

But, in not every situation do you start with a data base full of data and in not every situation, is it easy to create this data, it might either be costly or it might not be, the data that you might have is not the relevant data that you need and in many cases, you just have not started the exercise. So, for all those cases the big question is, basically do you have no scope for data analytics and the answer really is that, there is this whole other set of tools and techniques, the quantitative tools and techniques where which focus really are not just the analysis part of data, but have something to say about what data gives creative, which then goes and gets analyzed.

And that is the focus of these lectures on experimentation active learning and reinforcement learning, which is that data science is not confined or data analytics is not confined to choose of data that are already available, but it is moreover an iterative process, where sometimes the question of creating the data is also intrinsically looked into this grand problem statement.

Now, sometimes we will not even give a second thought to this dichotomy of creating data and analyzing data, because some problems just inherently come with this creation. So, if you take a look at many of the, you know inferential statistics techniques that we saw earlier in this course. Let us take an example where we used a two sample t test, you essentially had to sample n number from, you know class a and class b. So, and you know compare the means, this whole process or sampling was in some sense creating the data.

So, let me give you a concrete example, one of our favorite examples might have then that our 10th standard girls. Is a average height of 10th standard girls higher than the average height of 10th standard boys and there we said, we need to go to take a sample of 10th standard girls and sample of 10th standard boys and we came up with a conclusion that and we did a hypothesis test there.

So, when you do a sample, you are essentially creating the data and in other situations as well, where you would doing some kind of an engineering experiment perhaps and somewhere we did not really use a word experiment excessively, but the idea is that sometimes we never thought about it, but those are fairly the simpler cases. There are a lot more cases, where you need to explicitly think rule the creation of the data before any kind of analysis can take place and that is what we will focus on in this lecture.

Now, when we speak about creating data, a very important factor becomes whether this is an online context or an offline context for creating the data. What do you mean by that? Online and offline do not… Online does not mean that you are in the internet, it means something different here. What it means is, essentially when you are in a online setting; that means, you are experimenting or you are creating data on a system that is currently creating a product or producing, you know involved in performing a service or a task, which is going to the end user and for the purpose of getting this data, you are not just starting a passive observation exercise.

So, you are actually either querying the system or you either playing with the system in order to get the data that you need. Now, that can be a problem sometimes, because you are actively interfering with the system to create the data that you need and this system is right now either producing a product that is going to the end user or this system is a live system, which could influence the experience an end user is having and I am using the word end user in a fairly broad way. If you experimented on a traffic signal, the end user or the motorists need to go through the traffic signal.

If you clear on with a manufacturing process which is producing a product, the end user is the ultimate user of the product that goes and reaches the, you know the customer. If it is process, then again the process itself has some outcomes. It either gives the end user some information or the process itself performs a task for the end user and all of these things could be getting compromised, if you are playing with the live system.

And so, a major area that we would be looking at is the one of the reinforcement learning, where you are looking at an online system. So, you care about learning, where you care about this learning in a very supervised learning frame work, where you have some outputs and you want to understand how these inputs effect the output, that is one side of the story. But, this is second side of the story which is, you do not want to… You are interested in doing as well as you can, even while you are experimenting.

Because, some consumer or end user is experiencing the effects of your experiment and you yourself could be, the experimental could also be the consumer. So, we are going a little abstract out here. So, but the advantage of going this abstract is that you can really envision any scenario and any domain and this kind of a frame work should apply there. So, we will see that primarily, the online setting or I liked to kind of call it the live setting, live experimenting gets covered in reinforcement learning and those lectures will support that.

Now, in going forward now, in this lecture in the next one we are primarily going to talk about an offline setting. What we mean by an offline setting? Just the opposite of the online, which means, that the unit that we are experimenting on is not producing products that are going to go to the end user. So, you either created and you can do this in many ways. So, if you want to conduct an experiment and you want to gather data, you can create a model of your system and go experiment on that model, you can artificially create a lab setting.

Let us say I want to experiment on what fertilizers worked on my fields. I do not have to go pour those fertilizers on my fields, I can actually create a green house and choose some specific plans and try these fertilizers out and I can essentially create this lab setting, which is suppose to mimic the real world. But, it is not always creating models I mean it could be creating models, it could be creating artificial environments, it could be creating computer simulations and then you go experiment on that. But, it could also be the real process except, now if turned off the real process.

So, let me give you an example. Let us say you wanted to do an experiment on a machine and this is a machine that used to manufacturing and you wanted to know, what how to set various boiler plate stuff on this machine. So, you wanted to know what the speed should be, what the turning radius should be, different, different parameters associated with this machine. Now, if the machine is currently manufacturing the product which is going to the end user that is the problem.

But, what if you stopped the entire manufacturing process and you said, we are going to now exclusively commit some resources to experimentation. So, we learn about this system and so you clear on with the machine, clear on with various settings on the machine, make it you know create products and then you measure the products and see how well you did or how badly you did and you learn about the systems, you created the data, you analyze the data, you learnt about the system, but these products that are sacrificed in some sense, they are not going to the end user.

In other words, you do not care how well you do when you are experimenting and that is the core of offline experimentation. A bad experiment is not one that gives you poor results, but a bad experiment is one where you cannot learn about the system, your focus is about creating data to learn about the system and that is not mean there is no cost to experimenting. There is a cost, but you can think of it as a fixed budget that has been sacrificed or you can think of it as there being no cost at all.

However, you want, but it is not that while you are experimenting you are trying to perform as well as possible that would be the online setting. So, now, that we would understood this difference. It is also important to understand the difference between observational data and offline experimental data and I use the word in DOE. So, for the first time you started using the word DOE and here we mean design of experiments, it is a more formal way of talking about experimentation of talking about statistical experimentation.

And the idea here with this difference between observational data and offline experimental data is the fact that with observational data you are not interfering with this system. So, let us go back to the original problem. We said, hey how do you do data analytics when you do not have data. One approach is to say, so I need to start collecting the data and so perhaps I will turn on a few sensors or put some sensors in certain places and I have start collecting the data.

Now, that in itself is not the subject we are talking about here, that is just turning on this switch of collecting data and once you have the data, you analyze the data and that is fair game that is data analytics in some sense. But, that data analytics has coming from observational data. What we focusing on right now is, is there some way for me as the agent that wants to collect the data to actively engage for the system and choose, what data need gets collected and that is what we will be doing both in design of experiments and active learning.

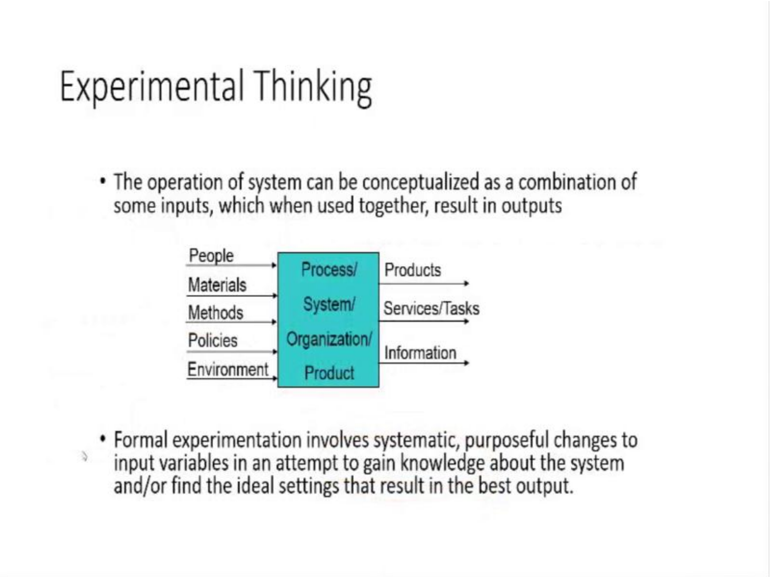
You are in an offline setting, meaning that you are right now committing all your resources to collecting the data and you can collect whatever data you want. But, the point is that you are going to be controlling what data gets collected and as a result, the only way to do that is not to passively observe the data that get generated, but to actively in the case of design of experiments you will actively go and change some of the settings in a system or very specifically go query certain points and that is how the data gets generated.

A typical problem statement and experimentation would actually say, go set the machine to setting a and setting b and setting, setting c and then let us see what the output is. And in active learning is seen a little bit more is that entire data, the input space is available and you get a query a particular point in the input space and then get the answer.

Now, these are just two difference ways of describing it, because they come from slightly difference context of application, but the core problem statement is the same which is that you as the user and experimenter or learner has the choice of generating an output at a pre decided point in the input space and that is has some critical difference over observational data and lot of advantages just to give you some intuition, the biggest advantage is one concerning multi collinearity, if you just observe the data it is possible that two input variable are so highly correlated to the point where there is almost a perfect correlation.

In which case you would never know if the output was increasing or decreasing as a result of variable input variable a or input variable b, because input variable a and b are so highly correlated. So, in the case of design of experiments are active learning you would come to that realization at some point without passively observing data and say o we need to try out an experiment or we need to query a point, where input variable a is high and input variable b is low and vise versa to try and understand which of these two input variables is having an impact on the output.

(Refer Slide Time: 15:17)



Experimental Thinking

- The operation of system can be conceptualized as a combination of some inputs, which when used together, result in outputs

People
Materials
Methods
Policies
Environment

Process/
System/
Organization/
Product

Products
Services/Tasks
Information

- Formal experimentation involves systematic, purposeful changes to input variables in an attempt to gain knowledge about the system and/or find the ideal settings that result in the best output.

So, let us start with focusing our first lessons on experimentation and design of experiment. The core idea design of experiments is that as long as you can conceptualize the operation of a systems has some combination of inputs which when use together results in outputs, you have the scope for this kind of black box creation and a black box I mean essentially and understanding of the way the inputs and the outputs relate to each other.

So, the inputs themselves can be anything you know they can be really broad, the only important thing is you want to able to quantify them in some way. The black box that you are experimenting or can be anything, it can be a process, it can be a system, it can be an organization which is performing certain organizational functions, it can be an actual product. And typically what are the outputs here you are interested in? The result of the black box could be the some products being created and you can measure these products and therefore, measure how good or bad they are.

The output could be some services or tasks that are created by this black box and again you should be able to evaluate the output. And the third is a little bit more abstract which is some information is getting created, again as long as you quantify these outputs and quantify the inputs you have a system were by there is scope for experimentation. So, formal experimentation what is it and how is it different from just observing data and analyzing.

Formal experimentation essentially involves systematic, it is a very important word systematic in purposeful changes that you make two input variables in an attempt to gain knowledge about this system and or find the ideal settings that result in the best output. So, that is sentence is little long wind it, but let us kind of break it down, so the idea is that in experimentation you proactively go and systematically or purposefully change the input variables it would mean that you actually go and say, oh I need to understand about little bit about the systems.

So, I am going to try setting input variable a to a particular value input variable b to another value and there I am going to go look at what output I get and the purpose of this. Now, in some rare cases it could just be that you are not physically changing the input variable, but you are physically choosing the input variable that you want to observe. Because, you do not have the information about how the output is going to look at every point in the input space.

So, you are not making a modification to the system there exists this large enough repository of information, which is very expensive to query. Because, if it is not all you have is a huge data set which requires a supervised learning task. But, for some reason if this is very expensive to query you could also think of experimentation and light of choosing very carefully the input variables that you want to gain knowledge about. But, more often than not the typical context is that you have this system where you go actually change the input variable set them to different values.

So, think of this perhaps this foundry process, where you are trying to create castes and your input variable could easily be the temperature of the molten metal that you pouring in, the pressure that is being applied the kind of material that the cast is made off and your output could be the number of defects that you see in the cast. Now, in an experimental process you will go and systematically purposefully try different temperatures, different pressures, different materials to make the cast, different practices in the cast, how long be you weight before you open out the cast, what kind of a room do you place it, you would go actually physically change these settings to different values and observe what happens to the output, which in this case is the number of defects you see.

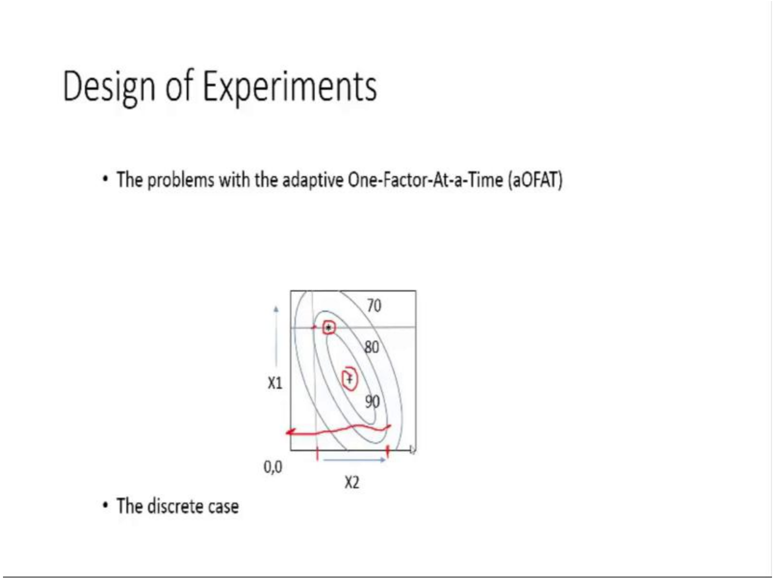
So, the emphasis here is in the systematic and purposeful changes to the input variables.

Now, why do you do it, you could do it for two reasons, you could do it to gain knowledge and you or you could do it to find those settings that you want to set the inputs to get the best output. Now, the gaining knowledge could just be an independent process that could be the final goal, sometimes the approaches is to gain knowledge and once you gain the knowledge.

And what knowledge we are talking about here? We are talking about the knowledge associated with how the input variables affect the output variables. Now, you could then once you gain this knowledge use that information to figure out what are the ideal settings for the inputs. But, you also have a algorithms which say you know what I do not care about the knowledge my goal is to right now find out what those settings should be. So, that I get the best output and that is takes on a different form. So, that is the core idea of experimentation.

So, this point a very natural question arises see you got some two or three or four input variables and you got an output variable. What is the problem? Why do not we just clear on with each input variable, you know one at a time and see what is the best setting for each input variable and do this sequentially, turns out it is not that simple.

(Refer Slide Time: 20:58)



Design of Experiments

- The problems with the adaptive One-Factor-At-a-Time (aOFAT)

- The discrete case

This is simple problem with that and we I am going to illustrate that with this diagram. Let us say that this plot that you has two input variables, input variable one x called x 1 and input variable two called x 2 and the output is nothing but, a hill that is projecting out of this green and this hill is shown with a contour plot. A contour plot is basically is one which connects which is often seen in maps, where you basically draw a line, where the height about sea level usually is the same.

So, here imagine that there is flat surface which is the base the rectangle and then there is hill projecting out of this flat surface, out of this screen and coming towards your face as if you are looking at this screen and these eclipse is that you see on this screen these kind of circular looking things are nothing but, the contour plots. So; that means, everything that is on this line is 70 meters or feet or inches whatever you choose to think of it.

So, this is an example of response surface are basically try to characterize in this picture, how the variables x 1 and x 2 have an influence on the output, why which is the hill coming out of the screen. So, it is an abstract concept, now let us take a look at what happens, when you just play with one variable at a time and we are going to call that adaptive one factor at a time experimentation. The idea behind this algorithm is that what I am going to do is at any given point of time I am going to play with anyone variable.

So, I am going to start with let us say x 1 and I am going try different values of x 1 and I am going make a conclusion at some point by saying at what value of x 1 did I see the best y and I am going go with that I am go on a fix x 1 now to that value and play with x 2. Now, sequentially do that with all the input variables until I come to a conclusion. Now, take this example where I stop playing around with x 1 and I arbitrarily fix some value for x 2 and it turns out that I fix the value right here.

So, I put a star there, so I started with x 2 set to this value and I just started playing with x 1. So, what is that mean when you playing with x 1; that means, you keep changing x 1 and here we are actually just changing x 1 to different values and as you go to through different values you see different heights. For instance, when you set x 1 to this value you seeing a height of 70. Why? Because, this is the contour line of 70 when x 1 is equal to this value you see a height of 80 and so you keep going through this process and until you find the highest point and the highest point is here, because of this point you are touching 80 for instance at this point at x 1 you are not touching 80 your some across the 75.

So, you conclude that this is the best value of x 1 and then you set x 1 to this value. So, x 1 gets set to this value and then you go about changing x 2. So, and basically when you changing x 2 you are staying on this grey line out here and as you keep going through x 2, you find that the point where there is a star is the peak and you conclude that is a highest point. So, you choose to set x 1 and x 2 such that you are at star.

Clearly, what is the problem with this, your problem is that you miss the peak, there it two problems on this, one is that you miss the peak the peak was really out here, if you will take look at this contour map in your understand the contour plot, the plus sign is where the peak is and you erroneously concluded that the star is where the peak is. And the reason for it is fairly simple, the way this hill is drawn it is clear that x 1 and x 2 have some interaction effects, there is x 1 is really good out here, when x 2 is set to this value.

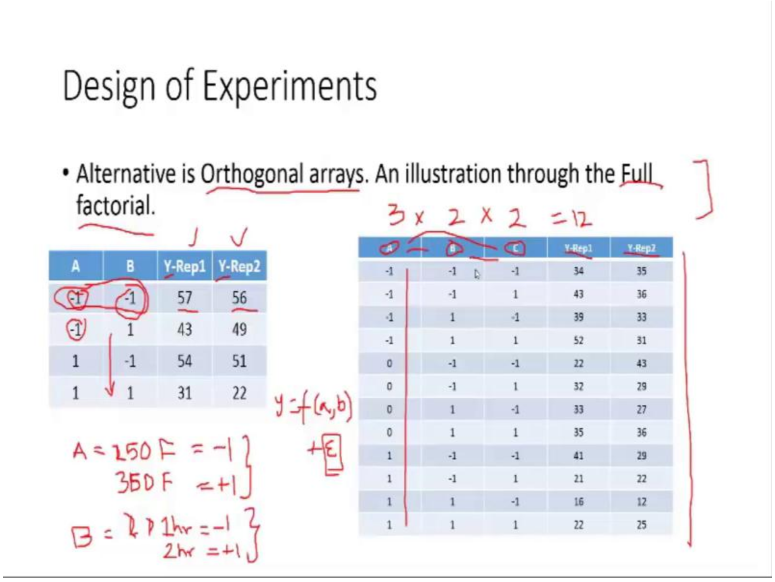
Now, if you said x 2 to another value, let us say x 2 here the highest point of x 1 is probably somewhere here. So, this value of x 1 winds up being highest. So, it really what we mean by an interaction is where you conclude x 1 to get the best output depends on where you set x 2 to. So, this is interactive effect that you can sometimes gets fooled by, there is an another thing that we have not really discussed here, which is that few do an experiment at a given x 1 and x 2 setting are you always going to get the exact same value and the answer is probably not.

Experimentation is typically carried out in sarcastic setting meaning that even if you understand that your output y is some function of your input variables, in this case there are two input variables. So, there is some mathematical function that is associated with input variables x 1 and x 2 and the output variable y, but on top of that if you said x 1 and x 2 to specific values are you going to get the same y. The answer is in a stochastic system you do not, because there is another factor which is just you can call it noise, you can call it irreducible error, you can call that luck whatever it is, it is just concept of just there being uncertainty above and beyond you are and this kind of uncertainty is what we deal with in supervised learning is what we deal with in much of what we have discussed and in this course.

So, that aspect also can throw you off in an approach like this. Now, this is been illustrated to you in a continuous frame work, where you are able to change x 1 continuously and see different, but in more practical settings you do not have infinite experiments. So, you might just set x 2 to a particular value and sample x 1 at someone two or three predetermined values, more often than not in design of experiments you will see that we are only interested in linear effects most of the cases you are interested in linear effects.

So, you would really look at trying out each variables at two different points. Because, where two points you can draw straight line and the two points are typically get coded. So, what is another approach that you can do, that you can employ to overcome this problem?

(Refer Slide Time: 28:15)



Design of Experiments

- Alternative is Orthogonal arrays. An illustration through the Full factorial.

A	B	Y-Rep1	Y-Rep2
-1	-1	57	56
-1	1	43	49
1	-1	54	51
1	1	31	22

3 x 2 x 2 = 12

A	B	C	Y-Rep1	Y-Rep2
-1	-1	-1	34	35
-1	-1	1	43	36
-1	1	-1	39	33
-1	1	1	52	31
0	-1	-1	22	43
0	-1	1	32	29
0	1	-1	33	27
0	1	1	35	36
1	-1	-1	41	29
1	-1	1	21	22
1	1	-1	16	12
1	1	1	22	25

$y = f(a, b) + \boxed{\varepsilon}$

A = 250 F = -1
350 F = +1

B = 1hr = -1
2hr = +1

The other approach you can do to overcome this problem and one of them is defined is called is a broadly called as orthogonal arrays and a specific type of an orthogonal arrays called the full factorial which is what I show here. So, take a system where variable A can take on two states. So, let us go back to this casting problem and let us say you are interested in variable A which is let us say the temperature of the molten metal that is boarded and so the temperature could be something like 250 Fahrenheit and you might be interested in studying the effects at 350 Fahrenheit.

Now, I am not an expert in this I do not know those are reasonable temperatures for molten metal I am guessing it is a little too low actually, but who knows. Now, typically what you do is when you have just these two settings, you kind of lay code them and you called them 250 as minus 1 and 350 as plus 1. Now, you do the same thing with the second variable input variable of interest, now variable b could be something like pressure, where you have lope or let us say the time that you weight before you remove the cast.

So, that could be 1 hours or 2 hours again I do not know those are reasonable numbers, they to illustrate a point is to 1 hour again you call it minus 1 and 2 hour you call it plus 1. So, what you go about doing in a full factorial is you try every combination of every variable with every other variable. So, you try the minus 1 minus 1 setting, you try the minus 1 plus 1 setting. So, and plus 1 and minus 1 setting and you get the picture, you essentially try every possible combination and that is called a complete enumeration of the designed space.

Because, you have discrete data points and you might choose to do some experiments at each of these combinations of points. In this example of shown you two replicates these are called replicates and these are both the same output variable y, but you choose to take two readings, it is actually unfair to call it two readings, because it is not like you do the experiment just once and just use the same measuring device and just take two readings, you actually read do the experiment and the reason you do that is because of the concept we just discussed which is you acknowledge that your y is some function of in this case variables a and b.

But, on top of that the sum error, this error often you sought of is being Gaussian which sum with mean 0 and standard deviation equal to sum value. But, even without going it to the they just some noise and you can see that even though you set 8 to minus 1 and b to minus 1 first time around you got 57, second time around you got 56 and you see different levels of uncertainty. Now, the same concept extended to three variables is what shown here, on the right hand side and you have so a, b, c three variables and I am also showing you a case where you are not just interested in two levels, but you might be interested in three levels.

So, variable a has 3 levels, variable b has 2 levels and variable c has 2 levels. So, complete enumeration of them would be nothing different than three times, two times, two which is equal to 12 and we call the 12 treatment combinations. So, you can have 12 treatment, 12 rows out here and again here I am choosing to take two replicates just to get a better idea of the noise that is there in this system as well.

But, the hope is that to taking on such an approach would enable or strict perform or perform of analysis on the data which helps us understand which helps us not get strict by this Gaussian this noise that is there on the system which can make you conclude the wrong things, if you an art where of it and at the same time also understand that they can be some interactive effects between a and b or a and c or b and c and that is about we will be focusing on.

So, in this part we are focused little bit more and how the experiments are itself designed and this is just one way of designing experiments and this is called the full factorial design and there are other ways of doing the same thing. In the next lecture we will be briefly talk about, how do you analyze the data that you gets from such an experiment and we close to be talking about active learning.

Thank you.