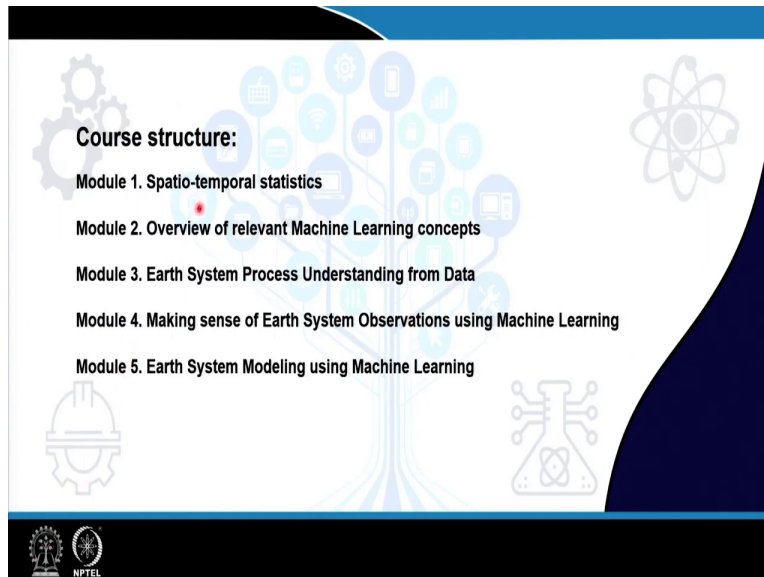


Machine Learning for Earth System Sciences
Prof. Adway Mitra
Department of Computer Science and Engineering
Centre of Excellence in Artificial Intelligence
Indian Institute of Technology, Kharagpur

Module - 01
Spatio - Temporal Statistics
Lecture - 01
Introduction

Hello everyone. I am Adway Mitra, an Assistant Professor at Indian Institute of Technology, Kharagpur. And, today we are going to start with our new course on Machine Learning for Earth System Science.

(Refer Slide Time: 00:45)



So, in this course, we will basically have 5 modules. The first module will be on spatio-temporal statistics. The second model will be overview of relevant machine learning concepts. Third module will be about earth system process understanding from data. In the fourth module, we will try to consider the earth system observations using say the observations we obtain from remote sensing and so on. We will try to interpret them using machine learning and try to get useful information out of them. And in the fifth module, we will consider like earth system

modeling which is something which many like scientists in various domains like hydrology, climate science, have been using in their own ways. We will see how those process based models can be improved using Machine Learning.

So, we will start off with the first module on spatio-temporal statistics. So, this will basically give you the necessary toolbox to understand and do the necessary computations with the kind of data which will be handling in the domain of Earth System Sciences.

(Refer Slide Time: 01:58)

Expected Background:

- Familiarity with probability and statistics (different distributions, variance, covariance, conditional and joint distributions, expectation)
- Familiarity with basic machine learning and data science concepts
- Interest in some domain(s) of earth system sciences
- Ability to read and understand research papers and from online sources

The slide features a background graphic of a tree with various icons (gears, atom, hard hat, circuit) and a small video inset of a speaker in the bottom right corner. The NPTEL logo is visible in the bottom left corner.

The background which is expected for this course is that the students should be familiar with basic probability and statistics like very, especially the concepts like say, various probability distributions like the Gaussian distributions, Gamma distributions, Beta distributions, Dirichlet distributions, and so on.

Then the basic concepts like variance, covariance, conditional and joint distributions, expectations, and so on. Like, you are not expected to be like greatly experts in statistics, but at least a clear understanding of these basic concepts is absolutely necessary. The second expected background is that you should be familiar with the basic machine learning and data science concepts.

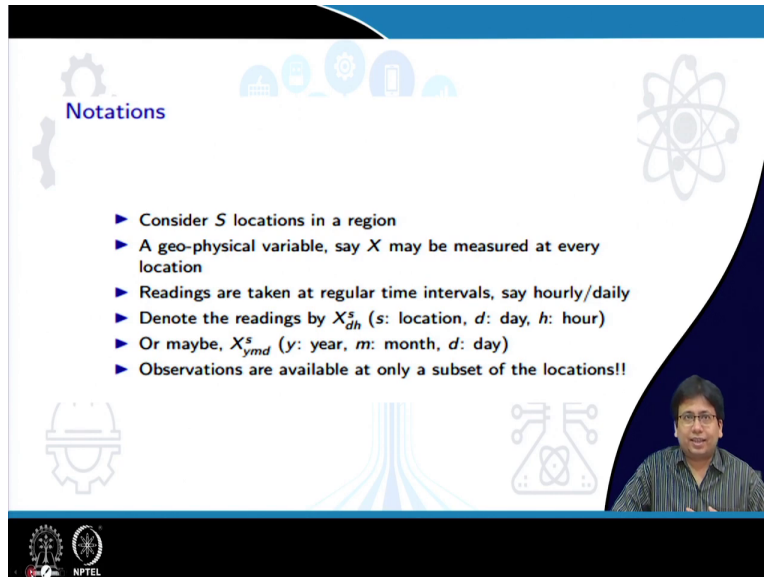
Some of the more advanced concepts, say example those related to deep learning and so on, we will try to like give a brush-up or overview of those in the second module. But I would also expect that either you have already got the enough background knowledge of these topics or you will be required to update yourself as the lectures proceed, otherwise you may not be able to follow like the subsequent modules like say module 3, module 4 and so on.

Apart from that, it is also a necessary to have interest in at least some domains of earth system sciences, be it climate or hydrology or say seismology or like any other thing, even remote sensing will do. But like at least some exposure to this, at least some domains of earth system science is a these are desirable criteria.

And finally, you should be able to read and understand research papers and there are very few standard textbooks and these kinds of things for this course because this is largely a state-of-the-art topic in which most of the reading materials will be from online sources only. So, you should have that ability to search online sources and choose the relevant materials and learn them.

And also, as we move to the later modules that is modules 3, 4 and 5, we will be dealing mostly with research papers, so which you will have to read and understand. So, maybe not, I mean suddenly not word by word details of research papers, but at least the by reading a paper you should be able to understand like what the paper is about, what is what the aim is, what are the methodology used and most importantly what data sources are they have used, because in these kinds of subjects the data is like the most important thing.

(Refer Slide Time: 04:47)



Notations

- ▶ Consider S locations in a region
- ▶ A geo-physical variable, say X may be measured at every location
- ▶ Readings are taken at regular time intervals, say hourly/daily
- ▶ Denote the readings by X_{dh}^s (s : location, d : day, h : hour)
- ▶ Or maybe, X_{ymd}^s (y : year, m : month, d : day)
- ▶ Observations are available at only a subset of the locations!!

The slide features a blue header and footer with various icons. A small video inset in the bottom right corner shows a man with glasses speaking.

So, now in this introductory lecture, I will give you a like a template problem like which might arise in this domain. So, various problems which we will be dealing with like you can say, this is a template for those problems. So, the data which we will be handling in this course will be mostly spatiotemporal data. That means what? That means, the measurement of something which is at a given location and a given time point.

So, let us say that there are capital S locations in a particular region and we are interested to measure a particular geophysical variable which may be something like temperature, rainfall or something like that, let us denote it by X , and that is measured at each of the S , capital S number of locations. So, these capital S locations, they can be either like point or points on the space or they can be grid points.

So, like as you know, like when we are studying any like earth science related matter like the standard is to divide the whole earth surface into grids, like of latitudes and longitudes and so which divides the earth surface into small square or rectangular regions. So, each of those regions can be called as a grid box.

So, either we can have like one measurement at every grid box. So, which will mean that we are taking, like there may be a spatial variations within the grid box also, but we can consider its

average value. Of course, that will make sense only if the grid box is small enough, otherwise, we do not want to take an average over a region which has a high variance.

So, we will like; so, either we will consider that the earth surface or the region in consideration is divided into small grid boxes at each of which we are reading the locations. Or, we can have what is known as in-situ locations that is something like we have placed a thermometer, let us say on one particular building of IIT Kharagpur and that case we are measuring the temperature precisely at that point. So, when we are talking about measurement, special measurements it can be either grid-wise measurements or in-situ measurements.

Now, this geophysical variable X which we are trying to measure let us say we are try taking its readings at regular intervals of time which may be hourly, daily like per minute, per second or is some kind of time frequency.

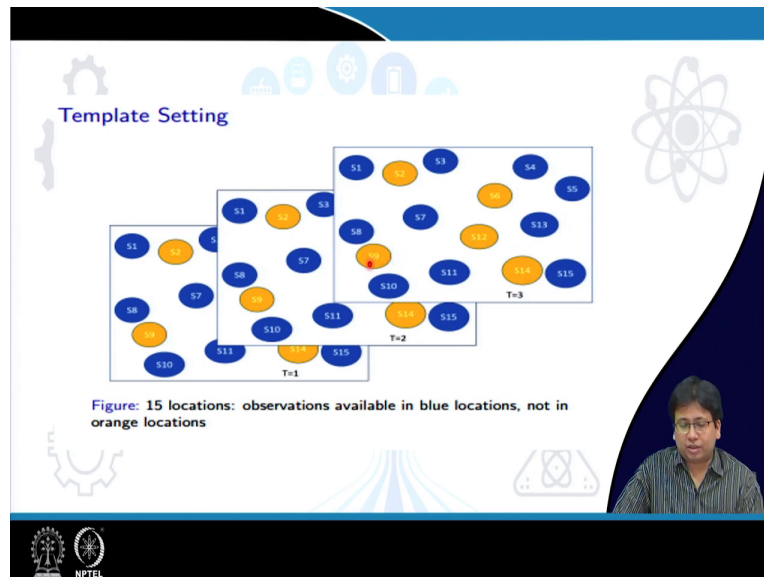
So, let us denote these readings by X_{dh}^s . So, X denotes the variable, the superscript s denotes the location, and the subscripts dh denotes the time point. So, in this case, d stands for day and h stands for hour. This is assuming that we are measuring the that variable at every hour of every day.

Now, it need not be that kind of measurement. We may also be measuring that thing every year, every month, and every day. So, in that case, the like we will index the variables like this as ymd . So, so basically the indexing depends on at what spatial and temporal frequency we are making the observations. So, this like, so this is the basic notation like as for specific applications we will be changing it slightly to suit the needs.

And also, let us assume that these observations or of these geophysical variables, the that is we are able to observe it only at a subset of the locations and at a subset of the time points. That is we do not have the complete observations of this set, but there may be only a certain number of locations where we have let us say the abilities to sense these or measure this geophysical variable.

And further, we may not be able to do it at all time points also within that period. So, there might be misses, they are like it is possible that from time to time that measuring system will be down and we will not be able to measure it. So, such gaps also can be there in the data, ok.

(Refer Slide Time: 09:22)



So, this is the basic setting. So, just to illustrate it. So, let us say that, like let us consider these time slices. So, T equal to 1, T equal to 2, T equal to 3. We are considering 3 time slices here and these bubbles these you can say are the spatial locations which we are talking about. So, like as I said the spatial locations can be either in-situ measurements that is at specific points or they can be over a uniform grid system.

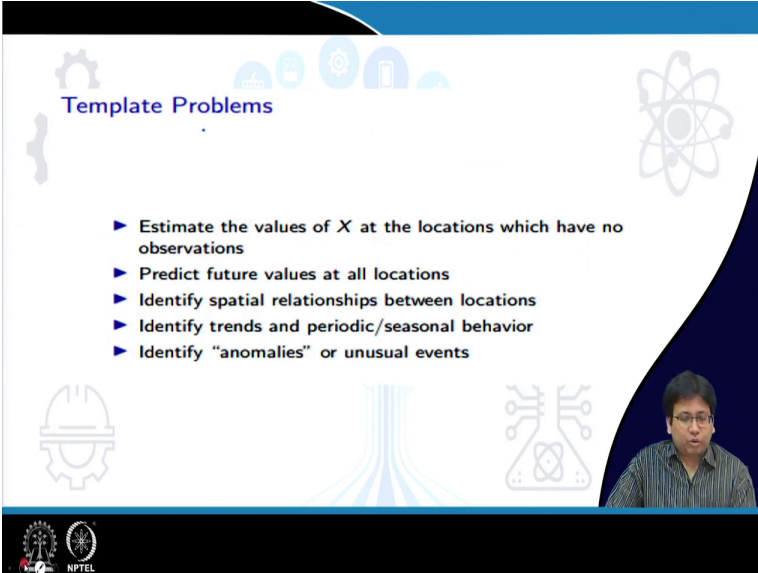
Now, as this suggests, this image suggests here we are considering in-situ measurements because you as you can understand that this is not like observations from a uniform grid. And these bubbles as you can see they are either blue or yellow.

Now, blue means that is we have the observations in those locations and this yellow or orange means that we do not have the observations. And you may also notice that at every point of time T equal to 1, T equal to 2, T equal to 3, we never get any observations at these orange locations like s_2 , s_9 etcetera. We have, that is because we do not have any sensing facilities in these locations.

In the other remaining blue locations, we do have sensing facilities, though in some of the time points even they may be temporarily down or faulty or something. So, some of these blue

observations may be missing or unavailable at certain points of time, but these yellow or orange observations will never be available.

(Refer Slide Time: 10:56)



The slide is titled "Template Problems" in blue text. It features a list of five bullet points, each preceded by a blue right-pointing triangle. The background is white with various light blue icons: a gear, a lightbulb, a person, a door, a network of nodes, a hard hat, a stylized tree, and a circuit board. In the bottom right corner, there is a small video inset showing a man with glasses and a striped shirt speaking. The bottom of the slide has a black bar with the NPTEL logo on the left.

- ▶ Estimate the values of X at the locations which have no observations
- ▶ Predict future values at all locations
- ▶ Identify spatial relationships between locations
- ▶ Identify trends and periodic/seasonal behavior
- ▶ Identify "anomalies" or unusual events

So, now given this setting what are the questions which we may ask or what are the problems which we may be interested in answering. So, one obvious answer is that. So, for like estimate the values of X at the yellow or orange locations, where we have no observations. That is the first and rather obvious problem or question which we may ask.

The second question is like, that is can we predict the future values. So, we have, let us say we have observations at only the blue locations till a particular point of time. Let us say till T equal to 500 we have observations. Then, in the subsequent time steps what the observations are going to be at, both the yellow and the blue locations can we make a forecast of that is like the another kind of problems.

Then, the next set of problems are related to are more diagnostic in nature. So, like can we have the observations at all the blue points, so can we identify some sort of special relationships between them? That is to say can we say for example, say things like whenever the observation in s_1 goes up that or in s_3 also goes up while that in s_{13} goes down. These kinds of relations we shall we be able to identify.

And the second related to that are like can we identify some kind of trends or temporal behavior in the observations at the different locations at the blue or the yellow locations. And finally, identify the anomalies or unusual events. So, there might be certain situations in which some of the observations at some locations have gone very high or and persists to be like that for a few days.

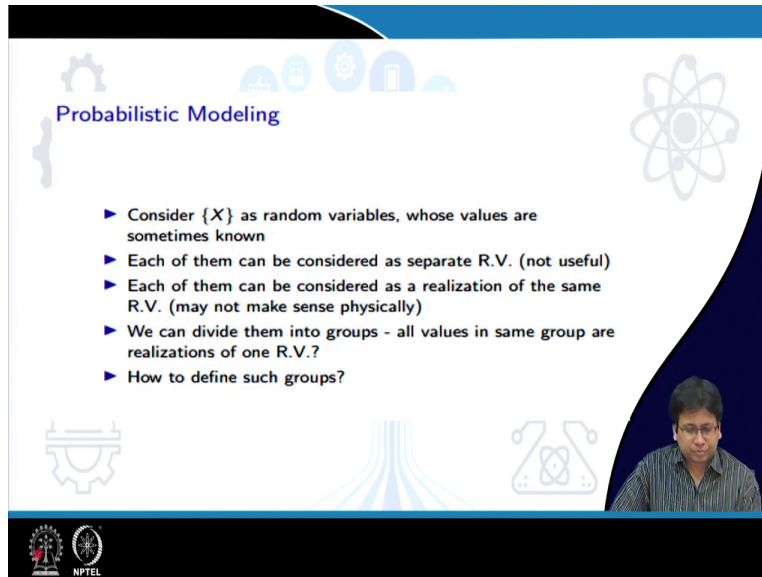
Now, this might happen for two kinds of reasons, either the measurement itself has gone wrong maybe the sensors with which we are using, there is like some error in those sensors etcetera, in which case it is giving faulty readings or it might also happen that some unusual thing is actually happening there, ok.

So, like if it is the first one, then the sensors need fixing. If it is the second one then like it is a matter of concern. So, suppose we are dealing with gas sensors, and we have a network of gas sensors and we are measuring the volumes of gases.

Now, suppose the volumes of the sensors are giving high readings for some time, then in a particular region that might indicate that a gas leak has taken place. Either it can mean that or it can also mean that the sensors have gone wrong.

But if it is like, if it is at only an isolated point then it might suggest that it is a problem with that sensor itself, but if it is like including several sensors in a given like in nearby locations that it is are very unlikely that all of them will be like behaving badly. So, most likely it is some things urgent has happened there which needs our attention. So, that is the anomaly detection problem.

(Refer Slide Time: 14:03)



The slide is titled "Probabilistic Modeling" in blue text. It features a list of five bullet points in blue text, each preceded by a right-pointing triangle. The background is white with a blue header and footer. The header contains several small icons: a gear, a lightbulb, a person, a gear, a person, and a lightbulb. The footer contains the NPTEL logo and a small video inset of a man speaking. The slide is framed by a blue border on the top and right sides.

- ▶ Consider $\{X\}$ as random variables, whose values are sometimes known
- ▶ Each of them can be considered as separate R.V. (not useful)
- ▶ Each of them can be considered as a realization of the same R.V. (may not make sense physically)
- ▶ We can divide them into groups - all values in same group are realizations of one R.V.?
- ▶ How to define such groups?

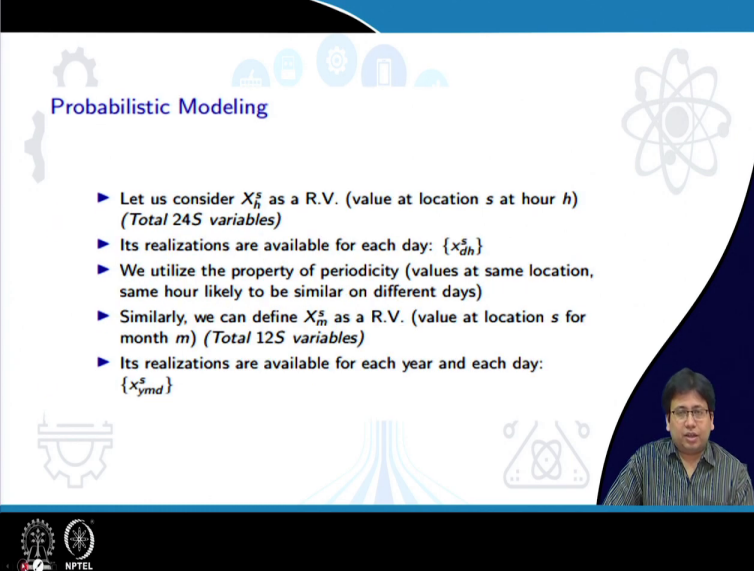
Now, the mathematical tool which we will be dealing with for solving this problem is Probabilistic Modeling. So, this observation X , we will be treating them or the variable X which we are dealing with we will be treating it as kind of a random variable whose values are sometimes known and sometimes unknown.

Say, so for like in for the blue observations, the blue locations where we have the observations or where we can measure the observations we will just say that it is the values of the random variable are known that is we have instances of that random variable. At the orange location, we do not have them.

Now, the these like we can either consider each of them as separate random variables that is the value of X , the measurement of X , at each of these bubbles at each time point; either we can consider all of them as separate random variables which we will see in the next lecture is not very useful or, we can consider each of them as the realizations of the same random variables or, we can divide them into groups and just imagine that all the values in the same group are realizations of one random variable.

And if that is what we are doing then the question will be, how to define such groups?

(Refer Slide Time: 15:30)



Probabilistic Modeling

- ▶ Let us consider X_h^s as a R.V. (value at location s at hour h) (Total 24S variables)
- ▶ Its realizations are available for each day: $\{x_{dh}^s\}$
- ▶ We utilize the property of periodicity (values at same location, same hour likely to be similar on different days)
- ▶ Similarly, we can define X_m^s as a R.V. (value at location s for month m) (Total 12S variables)
- ▶ Its realizations are available for each year and each day: $\{x_{ymd}^s\}$

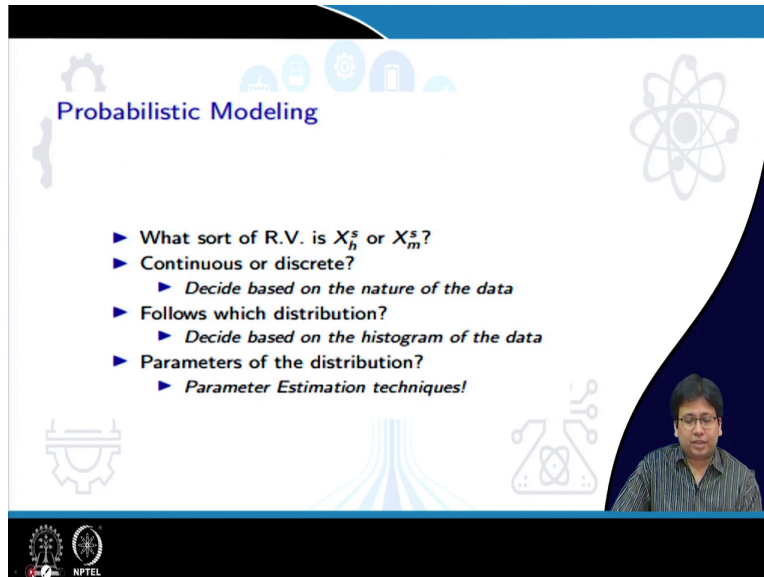
The slide features a blue header and footer with various icons including gears, a lightbulb, and a network diagram. A small video inset in the bottom right corner shows a man speaking.

So, we will try to understand or address some of these questions in the next lecture. And then we will use the idea of probabilistic modeling. So, like you; so, we will use what we already have is the data at certain locations and at certain time points.

So, now you like using that we may try to like estimate, like that is we can actually try to build some kind of a probabilistic model using those random variables, assuming some kind of a relation between the different random variables or trying to put in some other properties or like to capture their spatial and temporal variations and so on.

Like again, we will understand this in the next step.

(Refer Slide Time: 16:15)



The slide is titled "Probabilistic Modeling" in blue text. It features a list of four questions with bullet points, each followed by a sub-point. The questions are: "What sort of R.V. is X_h^s or X_m^s ?", "Continuous or discrete?", "Follows which distribution?", and "Parameters of the distribution?". The sub-points are: "Decide based on the nature of the data", "Decide based on the histogram of the data", and "Parameter Estimation techniques!". The slide has a blue header and footer with various icons. A presenter is visible in the bottom right corner.

- ▶ What sort of R.V. is X_h^s or X_m^s ?
- ▶ Continuous or discrete?
 - ▶ *Decide based on the nature of the data*
- ▶ Follows which distribution?
 - ▶ *Decide based on the histogram of the data*
- ▶ Parameters of the distribution?
 - ▶ *Parameter Estimation techniques!*

So, if we are doing that, then we will there are some design choices which arises. So, what sort of random variables are this going to be? Like so, we know that random variables can be continuous or discrete. Now, that is, of course, we can decide like that based on the nature of the data.

So, if we are like measuring something like say the amount of rainfall or the temperature then obviously, that is a continuous random variable. So, we will be using some probability distribution which is suitable for these continuous random variables. Like say Gaussian distribution, Gamma distribution and so on.

Similarly, if this, if the data is at discrete in that case we will be we should be using some kind of discrete distributions like maybe like multinomial or categorical or Bernoulli, Poisson or things like that. So, that, so these are some of the design choices which we must make understable and those choices which we will understand how to make them based on the nature and various properties of the data.

And then the whenever we are considering any probability distribution, it will also have its own parameters. Like say, the Gaussian distribution has the mean and variance parameters, the

gamma distribution has the shape and scale parameters, the lambda the Poisson distribution has the lambda parameter which is the intensity and so on.

So, how to estimate those parameters? So, the answer is that we will have to estimate the parameters from the data, from the observations by such a parameter estimation techniques such as maximum likelihood estimate and so on, ok. So, maximum likelihood estimate, we may not cover in this class, but like you are expected to read up more about it from various sources.

(Refer Slide Time: 18:14)



And here are some of the references which we will follow during this course. So, we like, so this first one, the Handbook of spatial statistics, this will be important for module 1. In for module 2 and the subsequent module, you can consider this edited book, Deep Learning for Earth System Sciences.

This is available sorry; for Earth Sciences. This is available freely in online, and this is a like it contains like a huge bibliography of related research papers and we will be going over those research papers like for especially for modules 3, 4 and 5 as and when the need arises, ok.

So, with this we come to the end of this introductory lecture. We will continue again on the next lecture where we will be understanding these Spatio-temporal statistics in a bit more details.

Thank you, everyone.