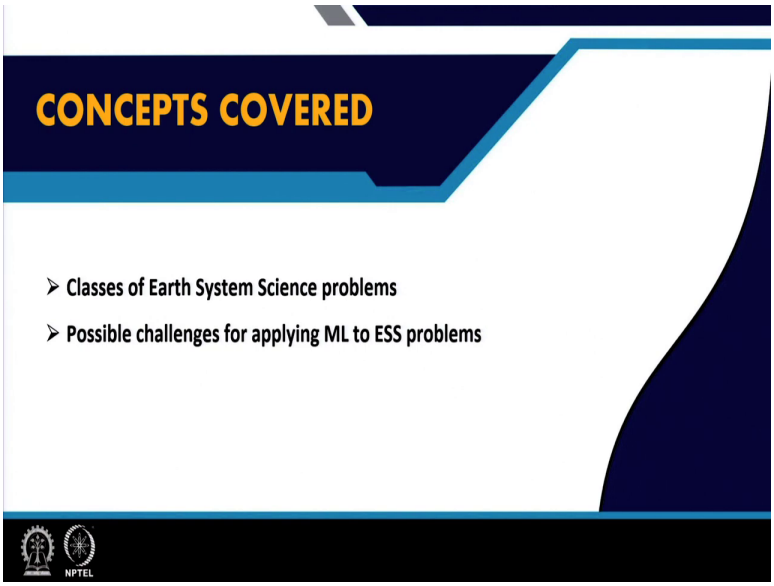
Machine Learning for Earth System Sciences

Prof. Adway Mitra
Department of Computer Science and Engineering
Centre of Excellence in Artificial Intelligence
Indian Institute of Technology, Kharagpur

Module - 02
Machine Learning Review
Lecture - 11
Challenges and Opportunities for ML in ESS

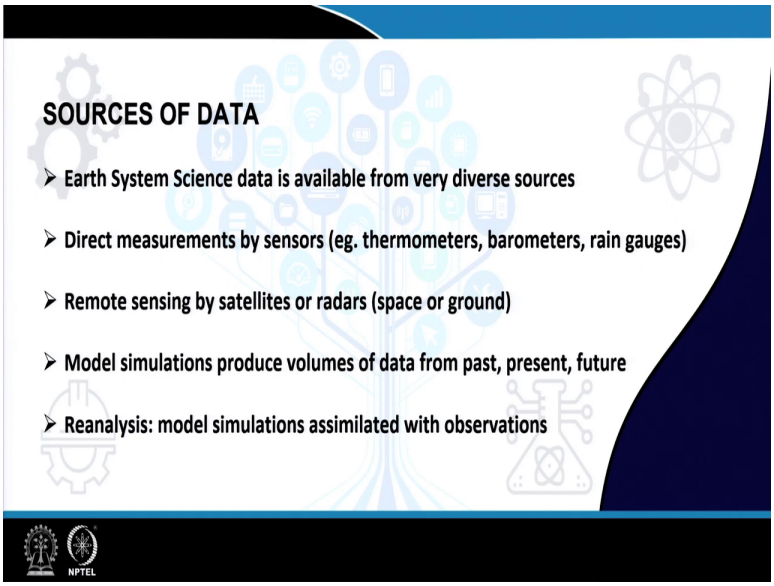
Hello everyone. Welcome to lecture 11 of this course on Machine Learning for Earth System Science. Today we will start module 2 which is a Review of Machine Learning techniques which will be useful for this course. So, the topic of this lecture number 11 are Challenges and Opportunities for Machine Learning in Earth System Science.

(Refer Slide Time: 00:48)



The concepts we are going to cover today are the different classes of earth system science problems and possible challenges for applying the machine learning to earth system science problems.

(Refer Slide Time: 01:00)



So, first of all let us talk about the different sources of data. Now, earth system science data is available from very diverse sources. First of all we have direct measurements by sensors such as thermometers, barometers, rain gauges and so on, which are like distributed all over the earth surface; and, like as the time is progressing we have made more progress in the quality as well as the density of such sensor networks.

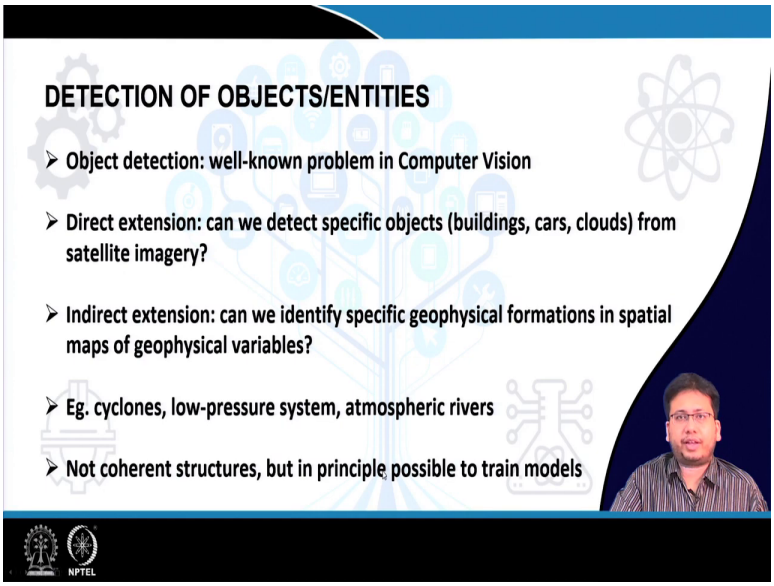
As a result, we are capable of getting very like a very large amount of data at very high spatial and temporal resolutions. And, most of these direct measurements by sensors are in situ, that is the measurements are specifically for the location, for the exact location where they are deployed. Another source of data is from remote sensing by satellites and radars, which may be present either on the space which is usually the case; I mean which is always the case for satellites and often for radars, but there are also ground based radars available.

Apart from these there are model simulations which produce volumes of data from past, present and future. So, these models, like there are specific models for different domains of earth sciences such as climate, hydrology, oceanography, then these various land surface processes, soil and convection and so on. So, like these models they like they are capable of simulating the different variables and their values they take over a period of time.

And, they can be run for like starting from some initial conditions they can be run for arbitrary long into the future, as well as they can be run backwards into the past. And, they generate high or very large volumes of data. In fact, the models are can often provide the values of variables which cannot be measured directly. However, whether the values simulated by these models whether they are accurate not or not, that is a different question.

So, reanalysis data is what we get when we do the assimilation of observations with the model simulations. So, in the last lecture, we had already talked about data assimilation using things like Kalman filter. And, we had seen how the model estimates can be progressively improved at every time step as we keep on receiving more and more data.

(Refer Slide Time: 03:40)



Now, let us talk about various classes of machine learning problems which we may face in the domain of earth system science. So, first of all is the problem of object detection. So, this is a well-known problem in computer vision for which lots of algorithms and models have been proposed for like nearly 20 years now, like earlier they used to build object specific features such as histogram of oriented gradients or haar features etcetera.

And, then use some kind of classifier like support vector machines to like do the object detection. Nowadays of course, for the for about 10 years now, when we are living in the deep learning era

then those features are like detected or computed directly by the neural networks. So, the object detection problem has been studied extensively in computer vision, but can it be extended to the domain of earth system science?

So, analogous problem in this domain will be to detect specific objects such as buildings, car, clouds etcetera; especially from satellite imagery. An indirect extension of this problem is suppose we have spatial maps of some geophysical variables. And, now can we identify specific geophysical formations from those spatial maps.

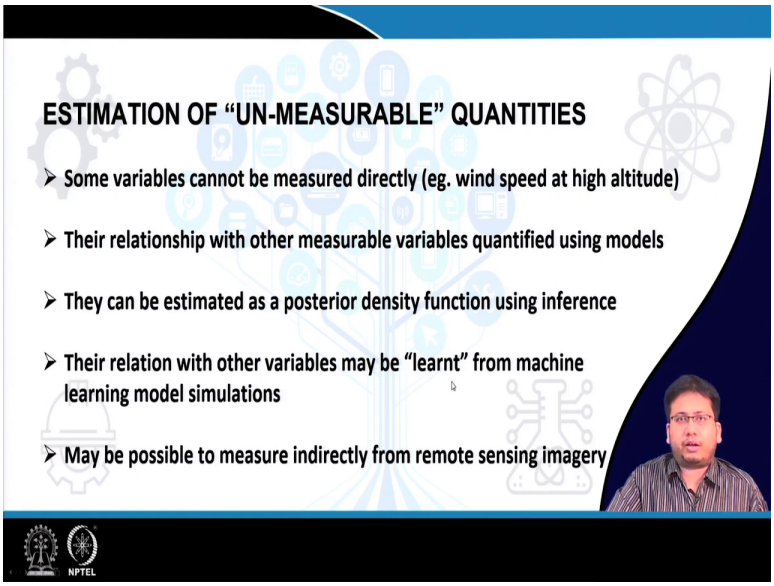
So, each of these maps, they can actually be treated as some kind of images. Even though they are like they such images like by looking at them we cannot make much sense out of it. But, like, an image is basically like a matrix where every element is corresponds to some pixel values. So, the same can be said about the spatial maps of geophysical variables as well.

So, in principle it is possible to like use these object detection models or algorithms to identify specific structures such as cyclones, low pressure systems etcetera from the physical or the from the spatial maps of the corresponding physical variables. Like for example, in case of cyclones we may want to look at the spatial maps of wind speed and pressure and things like that.

Similarly, in there are like there are other specialized structures like low pressure systems, the atmospheric rivers etcetera. Like these are actually spatial structures; for example, the cyclone has a kind of spiralling structure, which many of us have seen in weather forecast images.

So, these structures are not exactly coherent structures like the like say for example, if you want to detect a human face in image in the images taken by an camera. So, a human face is something that has a coherent structure. Now, these cyclones or low pressure systems they may not have a coherent structure like that, but still in principle it is possible to train the models to detect this, if we have sufficient training data.

(Refer Slide Time: 06:41)



Then there are estimations of various unmeasurable quantities. So, there are some variables that cannot be directly measured. For example, wind speed at high altitude. It is difficult to measure because simply we do not have like enough sensors at that altitude. Of course, like there are some in situ measurements can be obtained from satellites or say air balloons or in some cases even aeroplanes. However, like such measurements we can say are sparse and that too not available at like all altitudes, only at specific altitudes we can get.

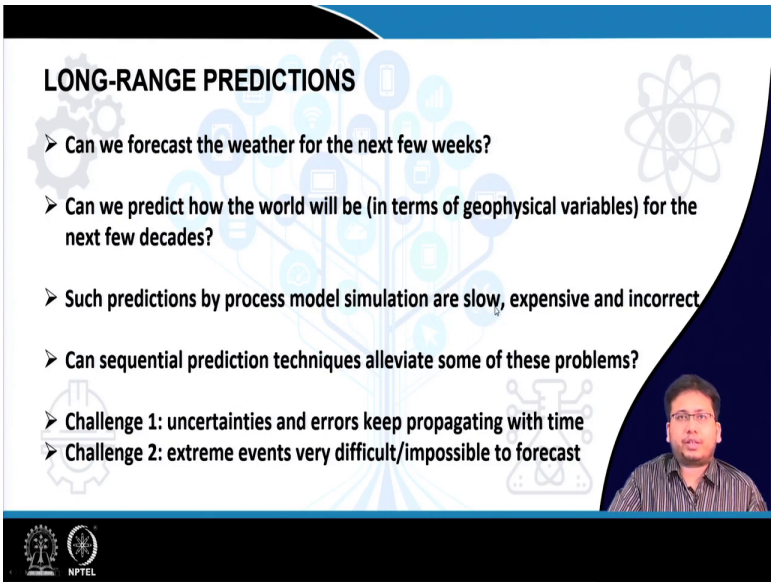
Now, the their relation with the other measurable variables, which we can measure directly, they such relations might be quantified using the various laws of physics and, then using those laws like it may be possible to estimate those unmeasurable variables as some kind of a probability density functions based on the measurable variables, that is, we have to use some kind of a conditional distribution and use the process of probabilistic inference to estimate what those values or those variables may be.

However, the problem is that the relations with these observable or measurable variables is often not very well known. They are based on some laws of physics which may not have been validated very thoroughly. So, the relations with the of the unmeasurable quantities which we are talking about their relations with the other measurable variables may actually be learnt from machine learning using model simulations.

And, then it may be possible to measure these variables indirectly from remote sensing imagery. So, suppose there are like, I want to measure the wind speed at a certain altitude and we have that satellite imagery which is taken from some airborne satellite at a much higher depth, at a much higher altitude. Now, that image by itself it does not tell us anything about the wind speed.

But, it might be possible to come up with some proxy measurements which in turn may help us to give an estimate of the wind speed. So, like there are a class of methods in machine learning, where like these remote sensing images are used to calculate the or they are used as some kind of proxies for various quantities of interest.

(Refer Slide Time: 09:27)



Now, another class of problems in which like machine learning can potentially be useful is that of long-range predictions. Now, can we forecast the weather for the next few weeks? So, this is an example of a long-range prediction problem. Next, can we predict how the world will be in terms of geophysical variables for the next few decades?

So, like both you can say they are long range prediction problems, though the by long the time scales are very different. So, this is a problem like where the time scale is in weeks and in this problem the times scale is in the order of decades. But, like at both scales like it is a like when

we are talking about weather forecasting a few weeks is a long time, when we are talking about climate then a few decades is like a long time.

Now, such predictions by process models like they can be obtained by simulations, that is the process models can be run for a long time. There are weather forecasting models like WRF: Weather Research and Forecast, that is, a numerical model for simulation of weather conditions. Now, that is typically suitable for like a few hours or at most may be a couple of days.

But, like it is in principle possible to run it for a few several weeks also and get some forecast. But the problem is that it will be slow, expensive and then also it will probably be incorrect. Similarly, there are these GCM; so, the General Circulation Models which some people also call as Global Climate Models. They can be run for like long durations of time; say several years or even several decades and we can get some estimation of what the future climate may look like.

Now, of course, weather and climate are different things at different time scales. So, both of these problems are similar in a sense, but they are they require different models. But, both of these problems they have the similar issues that is like in both cases they are slow, they are very expensive to run. They typically need supercomputers and then also the results they produce may often be incorrect.

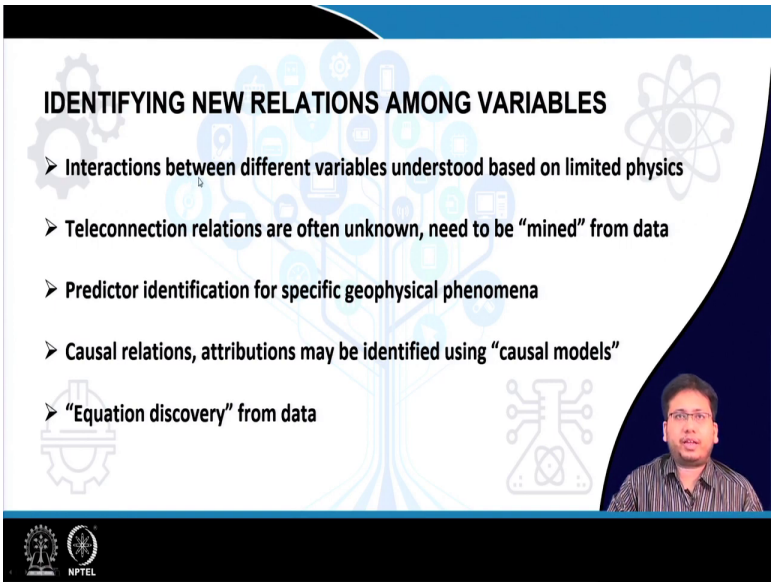
Now, can sequential prediction techniques like we now have machine learning models like LSTM and so on, which are very like we also have things like say reservoir computing which are capable of predicting long like sequential data or sequential time series. Now, can such sequential prediction techniques be used to alleviate some of these problems? But so, theoretically that is possible and it is an interesting task to tackle. However, it comes with two challenges.

The first challenge is that the uncertainty the even using machine learning methods, now any forecast that we make comes with some kind of an uncertainty or there may be some kind of forecasting error. And so, if we have some kind of even a small forecasting error like let us say a few days or a few time steps in advance. Then as we go into more and more time steps, these errors also keep propagating with time. And, what started off with a small error gets amplified greatly over time and soon the like our predictions like go way off the mark, that is one problem.

So, it is necessary to be very accurate right from the on which is often not very like an easy thing to do. The second challenge is that like there are often these extreme events and which are expected to become more and more frequent. So, these extreme events are often very difficult or even impossible to forecast like at a like well ahead of time.

So, if we are unable to forecast these extreme events, the our predictions are likely to be bad. And, like once as I said if we are unable to forecast a particular extreme event then we like immediately we incur some error at that point. And, then like in the future as we simulate more into the future or as we predict more into the future, the error we incurred they will only keep on accumulating and like soon we our predictions may go off track.

(Refer Slide Time: 14:09)



So, these are the two main challenges of long-range predictions. Apart from the next class of problems is related to identifying new relations among variables. So, interactions between different variables, is understood based on limited physics. We often do not know the exact physics of how different variables interact. We have various laws of the nature, but those may also be that is mostly hypothetical which have been which have been validated by on so called small data.

But, like many of these laws that is they may be missing out on important variables which like which are not known to science as of now. So, now the specifically there are teleconnection relations between climatic or geophysical variables in different parts of the world. And, such teleconnection relations are often unknown and they need to be mined from data. Of course, mining them with the relations from data and like gaining insight about how those how that teleconnection actually works; like in the laws of physics, that is a different question altogether.

And, even in that matter like this data dependent data driven machine learning methods may be useful. Then, predictor identification for specific geophysical phenomena is another question of interest. Say, for example, I want to predict the seasonal rainfall, monsoon rainfall over India. So, it is well known that Indian monsoon is like it is impacted by like climatic phenomena in different parts of the world. Some of which are quite far away such as even in the Pacific or the Atlantic Ocean.

So, now exactly which variables in which parts of the world can be possible predictors for the Indian monsoon? Identifying such predictors is like is not a straightforward task and earlier Indian meteorological department used to have some 8 parameter or 16 parameter models, where they used some 8 or 16 predictor variables from different parts of the world. And, use some kind of a regression model to predict the Indian monsoon.

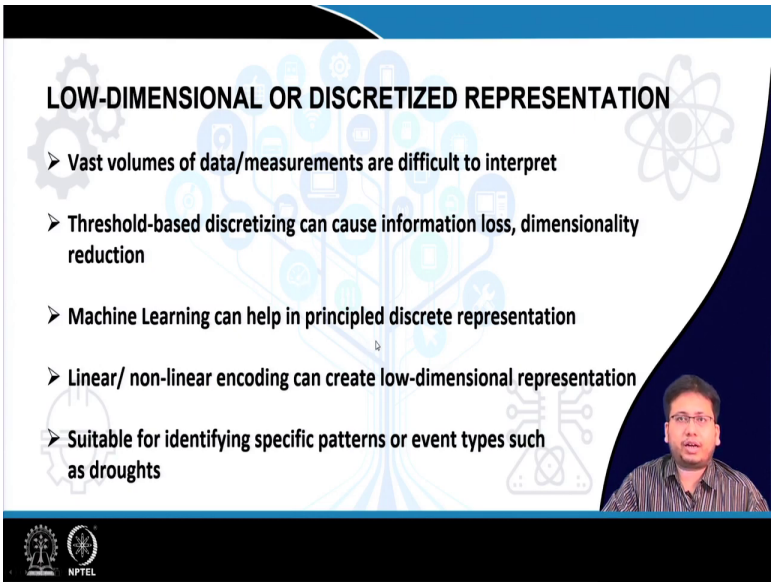
Of course, that practice is now not used anymore. They have moved on to other more sophisticated approaches. But, in like those 8 or 16 predictors which they had obtained may not have been like identified through a very principled way. As a result of which they are often used to be errors. So, it is possible to use machine learning. So, machine learning has a class of methods called feature selection and it is possible to use such methods for identification of predictors.

Now, causal relations or attributions they may be identified using causal models. So, in our lecture on causality, we have already discussed this in some detail like we have also mentioned the problem of attribution. Say suppose one particular event happens; can we identify or attribute that event to a larger cause? So, that is the problem of causal like causal attribution.

So, in general that is not an easy problem to solve. But, then in machine learning there are class of models called causal models which may enable this to be done. And then finally, there is a task of equation discovery from data. So, we already have like various laws and equations in physics and related disciplines. But, then again as I said earlier those are like based on like those are typically hypothesis by like scientists which have been validated on sparse observations or as we can say small data.

Now, using the power of big data can we discover more equations or maybe refine the existing equations? So, then another set of or another class of problems is that of. So, if I want to create a low dimensional or discretized representation of the vast volumes of data we have.

(Refer Slide Time: 18:13)



So, like if we have so much volumes of data or measurements, which I already talked about, say from the dense sensor network from remote sensing imagery from model simulations and so on. Such like if we like right now the biggest problem in this domain is not the unavailability of data, but rather the abundance of data. We have so much data that searching for useful information from it is often like searching for a needle in a haystack.

So, an alternative might be to discretize the data or to somehow reduce its dimensionality; however, if we do that based on some kind of thresholds as is like as has often been done by the

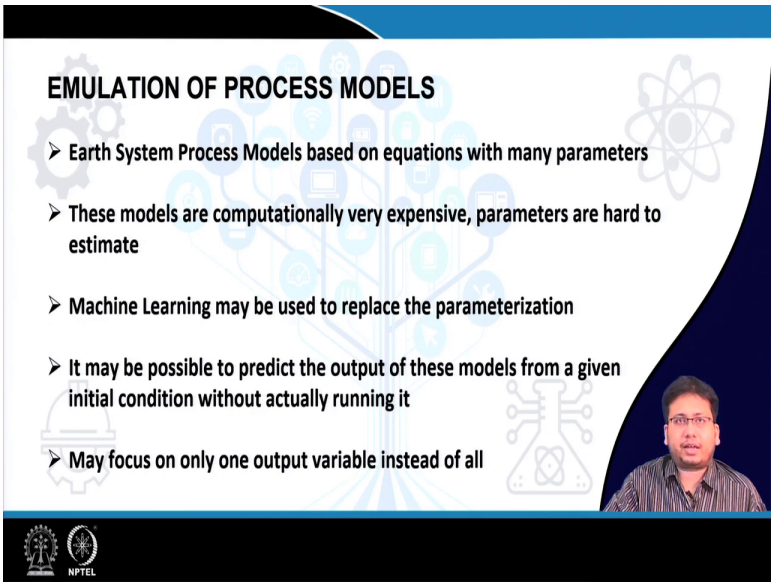
domain scientist, such threshold based discretization can cause information loss. Because, we do not know what threshold to use and it might also happen that like it is important to use different thresholds at different places, which is which may often be difficult.

So, a threshold based methods to discretization or for dimensional reduction that is if we have some high dimensional data and we just prune of some of the dimensions. These are often counterproductive because they may cause loss of important information. So, machine learning can actually help in approaching these problems in a principled way. It can like create a discrete representation or it can do the dimensionality reduction using some kind of linear or non-linear encoding.

And, the thresholds which we talked about like machine learning has the potential to eliminate the need of putting of setting hard thresholds like that. There are soft thresholding and various other sophisticated approaches to thresholding. We can say smart thresholding if you want that can probably alleviate some of these problems. And so, these kinds of tasks are suitable for identifying specific patterns or events in large volumes of data.

Like say for example, droughts or heat waves or like ocean eddies things like that. Things that are that have some kind of complex structures, but it is not so easy to identify them using any kind of thresholds on individual values. So, if the those individual values if they can somehow be reduced to a lower dimension or discretized, then may be identification of these kinds of patterns or events may become simpler.

(Refer Slide Time: 21:08)



Then that there is this crucial domain which is becoming an important issue of in the domain of earth sciences now is the emulation of process models. So, as I said the domain scientists of earth system science, various disciplines of earth system science have developed various models for the processes in their respective disciplines. So, those are typically based on equations, which have lots of parameters.

So, the like as the domain scientists understand the physics and chemistries of those problems, they encode the process through a set of equations which are known as the model equations. And, then those models also have a very large number of parameters. So, the problem is that these equations they may be these equations may be inadequate. Because, as I said earlier they are like they are basically hypothesis of various climate scientists with some small scale validation.

And, besides the parameters which are used, they are often hard to estimate. In fact, the parameters need not be constant; they can be varying over space or time. So, machine learning may be useful for replacing this kind of parameterization. So, typically what happens in process models is that those models work at a particular resolution; spatial and temporal resolution.

But, there are various sub processes which occur at lower resolutions. So, such processes are known as the unresolved processes which the model does not cannot simulate -instead what it does is it just assume some kind of parameters to represent those processes. But that is often not a great idea.

So, it is possible to replace some of those parameters by machine learning models. So, instead of actually just reducing the those unresolved processes to just a parameter value, it is possible to run some machine learning model to predict those or to simulate those processes.

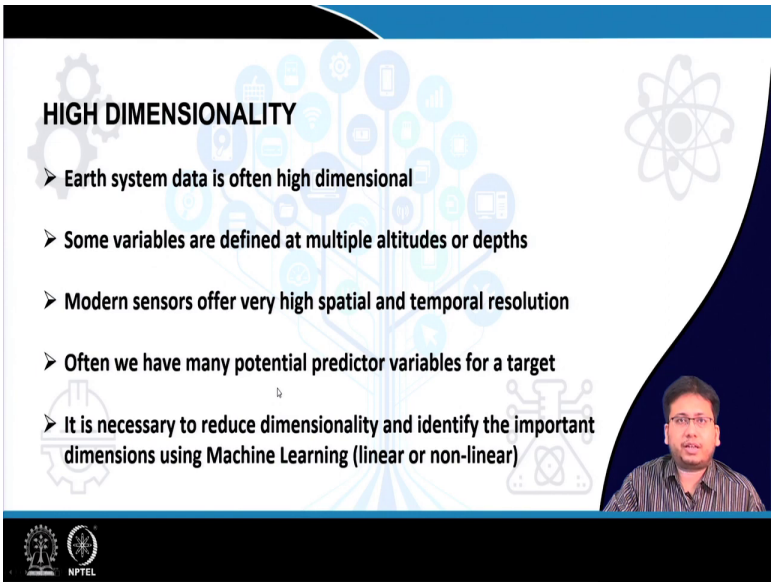
And, whatever values we get as out their outputs, we can then plug those values into the larger process models. So, this is right now a very hot topic of research in machine learning for earth system sciences that is using machine learning to replace the parameterization. Now, it may also be possible to predict the output of these process models starting from any given initial conditions without actually running them.

So, the reason why we may not want to run these models is that like as I have said earlier also, these are slow and expensive and running them often require supercomputers. So, like we may if we do not have so much of computing power or if we want a fast answer, say, suppose I am making a for like a now casting problem, say that is to say forecasting at short notice. So, like say if I want to forecast some event that is that may happen in the next 6 hours.

I do not want process model that will run for say 3 and or 4 hours and then only give me the result, because that way vital time will be lost. So, the alternative is to build the machine learning based or emulators or surrogate models as they say. They like they need not run the entire model, but they may focus only on specific things like specific output variables of interest and just predict what the model may be like may simulate its values to be from a given this initial condition.

So, if the models are used offline to generate a large number of large amount of data, these kinds of machine learning models can actually be predicted or can be actually be trained to predict the models output from any given input. So, this is the process of emulation or building surrogate models of earth system process models.

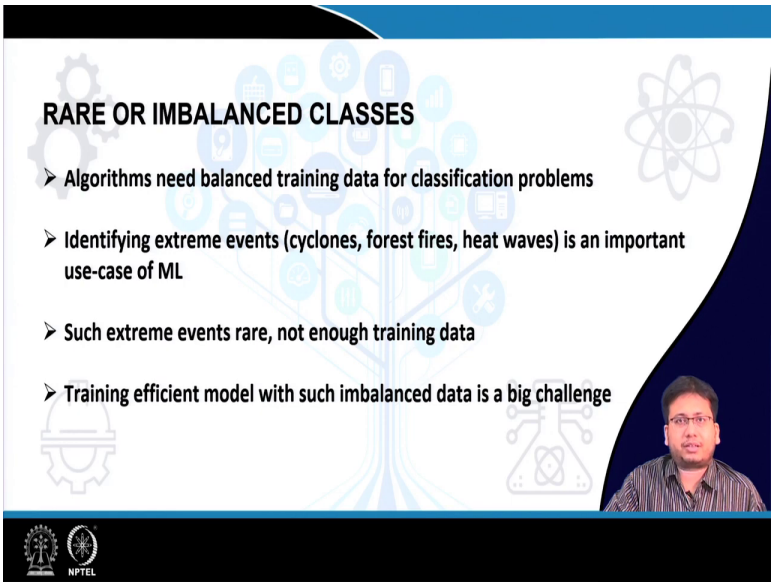
(Refer Slide Time: 25:18)



Now, the coming to the various challenges related to the nature of the data, we come across in the domain of earth system science. First of all these kind of data is often a high dimensional, we have already mentioned. So, it is high dimensional for multiple reasons. Like some variables they may be defined at multiple altitudes or depths. Then there are these modern sensors as I said they produce data at very high spatial or temporal resolutions.

And, like when we are trying to identify the various predictors as I of a particular phenomena as I talked about, we often do not know the which or which variables or variables from which locations to focus on. So, we simply throw in all possible predictors and hope to identify the right predictors. But, that creates a problem of high dimensionality that is the input the inputs which we give is of like very high dimensions. It is necessary to reduce the dimensionality and identify the important dimensions using machine learning techniques, be they linear or non-linear approaches.

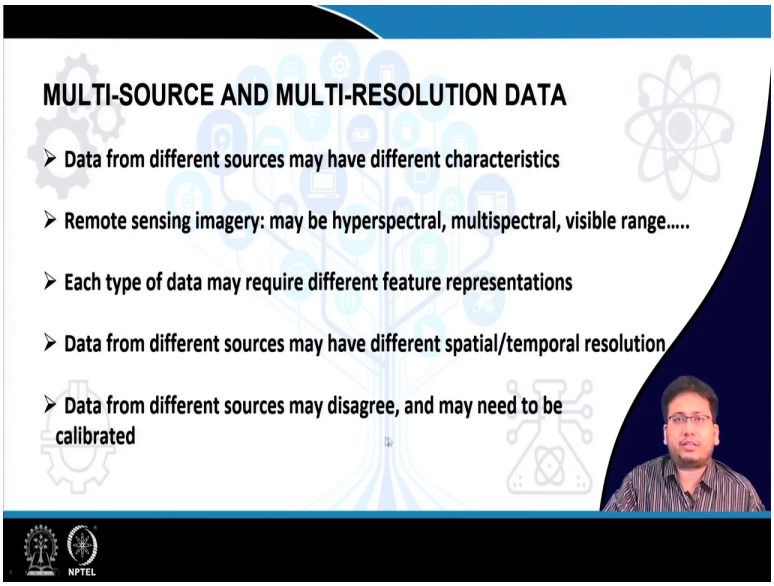
(Refer Slide Time: 26:25)



So, then there are this problem of rare or imbalanced classes. So, we know that classification algorithms they need balanced data from that is to say they need roughly comparable number of examples from the different classes for this kind of a these kinds of classification problems. But, when the problems are related to say identifying of extreme events; say cyclones, forest fires so, heat waves etcetera, like identifying such extreme events from vast volumes of data that is of course, a very important and crucial problem in the domain of earth system science and, climate scientist or earth scientist hope to develop machine learning methodologies for these problems.

But, such approaches or such models are going to be marked by the presence of imbalanced classes. Because, the these extreme events by definition they are very rare; so, we will not get enough training data for them. So, like there are of course, ways to get along get around these problems, but that is another big challenge.

(Refer Slide Time: 27:35)



And, then there is this also this problem of multi-source or multi resolution data. So, as I already said the data comes from different types of sources and each source may have different characteristics. So, even when we are talking about remote sensing imagery, then there are different kinds of remote sensing imagery. There are hyper spectral images, multi-spectral images, visible range images and all that.

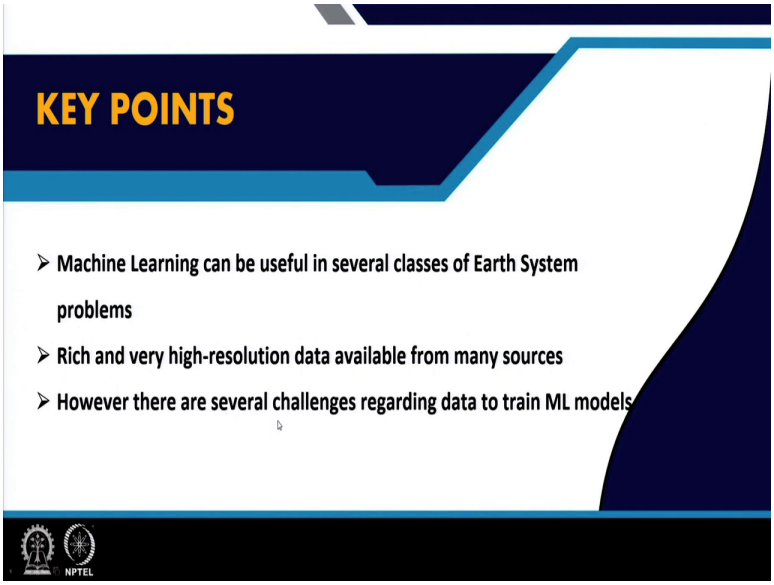
And, then for each kind of data we may need different kinds of feature representation. Like so, like when we are doing any kind of machine learning model, it is always necessary to represent the data using like using suitable features. But, then the features for one kind of data may be may not be suitable for another kind of data. Also the data from different sources, they may have different spatial or temporal resolutions.

The data from different sources they may disagree also, because the measurements are done in using different technology. So, the measurements that of a particular thing which we get; let us say we are measuring rainfall using say radar satellites and rain gauges. It is possible that each of them will give a like different estimate of the same thing. So, it is actually necessary to calibrate that these different sources against each other.

(Refer Slide Time: 28:53)

So, like in recent years say especially say for the past 5 years or so, climates like scientists in different parts of the world especially in US and Europe have been increasingly turning to machine learning methods. And like so, these are some of the important references which discuss about how machine learning may be used for in the various context of earth system sciences.

(Refer Slide Time: 29:20)



So, the key points to take away is that machine learning can be useful in several classes of earth system problems. And, we also have very rich and high-resolution data available from different sources. However, there are several challenges to regarding this data to train the machine learning models. So, in the coming lectures, we will actually discuss the various classes of machine learning models which may be suitable for identifying or for studying these problems. So, we will discuss this in the following lectures.

So, till then goodbye.