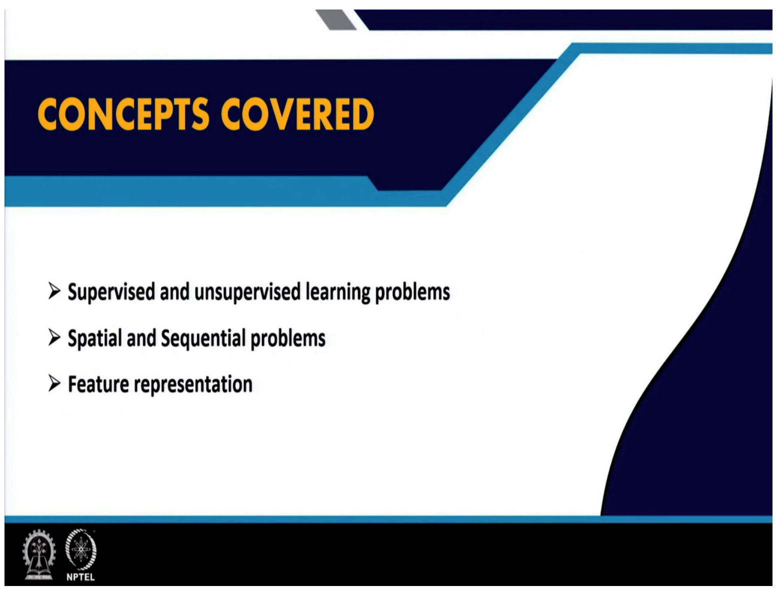
Machine Learning for Earth System Sciences

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 - 12
Types of Machine Learning Problems in ESS

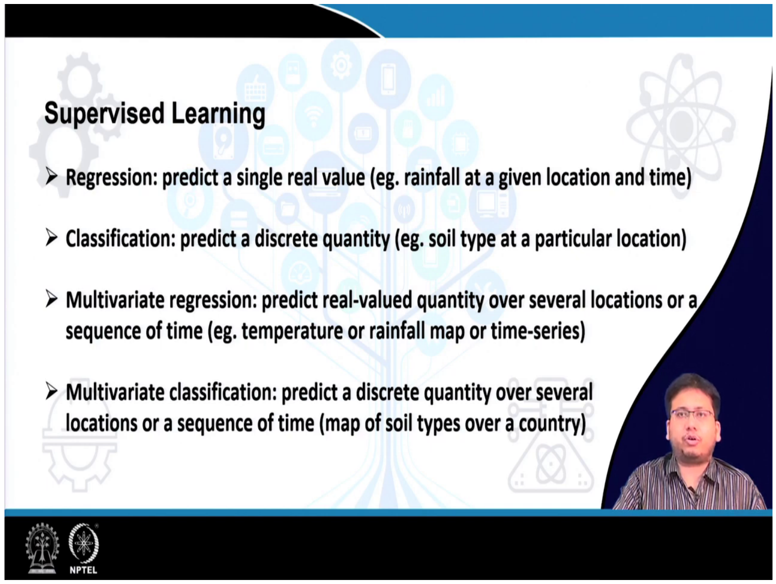
Hello everyone, welcome to Lecture 12 of this new course on Machine Learning for Earth System Science. So, right now we are in Module 2 which is on review of the Machine Learning Techniques which will be suitable for this course. So, today we are talk going to talk about the different Types of Machine Learning Problems in Earth System Sciences.

(Refer Slide Time: 00:47)



The concepts which we will cover today are supervised and unsupervised learning problems, the spatial and sequential problems and feature representations.

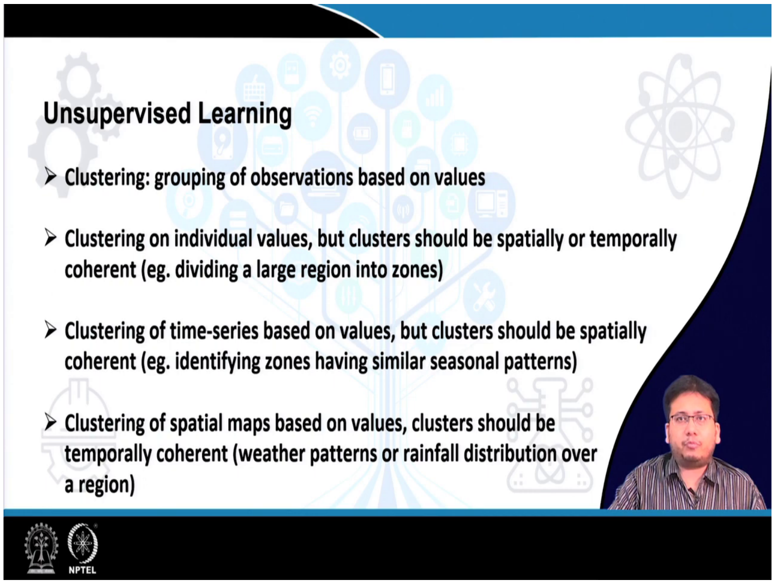
(Refer Slide Time: 00:57)



So, coming to Supervised Learning problems like the basic concept I hope you all know what is supervised learning? So, we know that there are 2 categories of supervised learning problems one is Regression, where we try to predict a single real value; for example, the rainfall at a given location and a time. And another category is the Classification where we try to predict a discrete quantity, for example the soil type at a particular location.

Then there are also multivariate regression problems where we may want to predict real values valued quantities over several locations or for a sequence of time. For example, either a temperature or rainfall map with a spatial map or a time series. Similarly there is also multivariate classification problems, where we try to predict a discrete quantity over several locations or a sequence of time. Such as for example, an entire map of soil types over a country or like over a period of time.

(Refer Slide Time: 01:57)



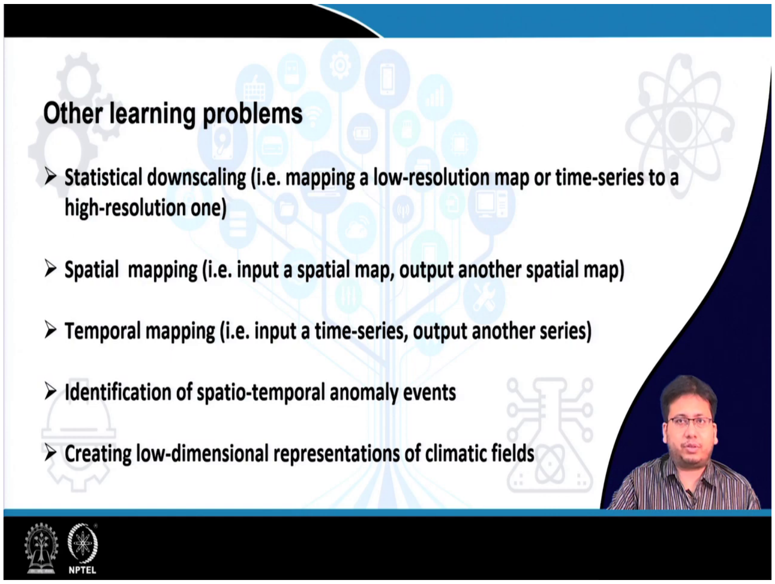
Then there are unsupervised learning problems, say for example Clustering. So, clustering we know is a task of grouping of observations based on their values. Now clusterings can be done they on the individual values of the observations, but like often in the domain of earth sciences, it comes with an additional constrain that the clusters which we get they those clusters should be spatially or temporally coherent.

For example we may want to divide a large region into zones, but those zones like they should be like contiguous zones or like they should have some kind of a meaningful shape, otherwise it may not make great sense. Then additionally we may want to do the clustering of time series based on values, but then those clusters again may be should be spatially coherent.

Say for example, if you want to identify the zones having similar seasonal patterns or I may also be interested to do the clustering of spatial maps based on values, but those clusters should be temporally coherent. For example, I may want to do a weather patterns or rainfall distribution over a region.

So, these are all clustering problems, but they often come with additional constraints over the way in which normal clustering problem is defined.

(Refer Slide Time: 03:21)



Now, there are other learning problems also. Say for example, statistical downscaling where we are interested to learn a mapping from a low resolution map to a high resolution map or say from a low resolution time series to a high resolution time series.

Then so there is a spatial mapping problem, where the input is a spatial map and the output is another spatial map or there is a temporal mapping problem where the input is a time series and the output is another time series. So, this statistical downscaling problem which I mentioned is like is a good example of both of these.

Then there is the task of identification of Spatio temporal anomaly events and finally, the task of creating low dimensional representation of say climatic or other earth other fields in earth sciences.

(Refer Slide Time: 04:33)

So, like in our previous lecture we already talked about the that why it is often necessary to create a discrete or low dimensional representation of the data, like that is for like to make it more comprehensive and also to identify specific or structures in the data which might be difficult to identify in the like in the raw data.

An important category of machine learning problems is that of predictor selection. So, we know that in machine learning like we need to represent the quantities of interest using features or predictors. So, like we that is why we have the concept of feature vectors.

Now for example, if I want to predict the monsoon rainfall over India we may need the like a set of predictors the set of values of different variables which may be used in as inputs to a machine learning algorithm to predict the quantity of interest.

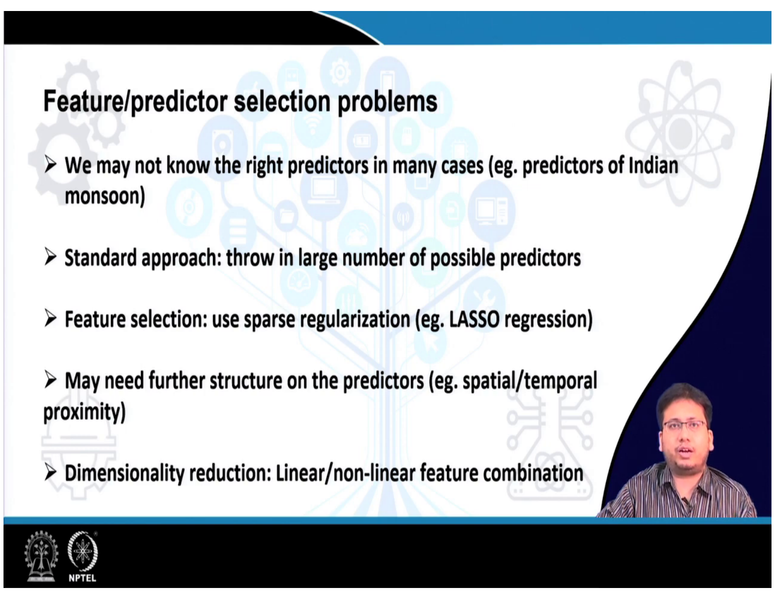
Now like the those predictors may be say for example, sea surface temperature and wind speed from different places of the world. Now these predictors they may include a variety of covariant variables at differ measured at different locations or time, but not all of these predictors or covariates are going to be like suitable or are going to be equally useful in making the prediction. Typically only a few of them will be useful the others will be useless.

So, identifying which of the predictors are actually useful and which are not. So, that is the task of feature selection or of machine learning for which there is a plethora of methods. Then there is like in case of unsupervised problems we have this feature representation, like say for example I want to represent the let us say that task is to do the clustering of these spatial patterns or temporal patterns or whatever. So, in that case typically the input we need may they may include the values of different variables at several places and times.

So, like now which places and which times or which variable these are all questions which we need to answer to get a proper representation of the data. I mean whenever we are solving some kind of an unsupervised learning problem any kind of clustering problem. So, that depends entirely on the features like how we represent the different data points.

So, like in this case the there like we may have to represent the data using not only one variable of interest, but may be also various other covariates which may have some influence on the value on the variable that we are actually interested in. We may need to consider it at different locations at different time points, so that way it becomes a like it becomes a bit of a challenge.

(Refer Slide Time: 07:27)



Now, in machine learning there is the set of approaches which are known as the Feature Selection approaches. So, the idea is that so we have a very large number of features, let say a very high dimensional input vector the like which are to be used as input to some kind of a prediction algorithm.

Now, the prediction algorithm may identify which of the features are useful and which are not. So, there is a concept of sparse regularization, where some like say for example in case of regression problem each of the features are provided with some kind of a coefficient which the these magnitudes of these coefficients they indicate how the relative importance of that feature for that particular problem.

Now, like I may impose the constraint of sparsity on these coefficients which mean that most of these coefficients should be 0 and only those features which are really important the most crucial features they will have the non-zero coefficients. So, there is this class of methods you know like, for example the LASSO regression and there are other variants of it also which do this kind of sparse estimation of the coefficients.

So, additionally we may also need some further structures on the predictors may be beyond sparsity. For example, they should be spatially or temporally proximal to the thing to the quantity that is being measured.

Say for example, if I want to predict the rainfall over India like I we know that from common sense that the most important predictors should be from nearby regions. Of course, there are teleconnections as we had mentioned say; for example, some quantities in Pacific Ocean or Atlantic Ocean they are also important for predicting the Indian monsoon.

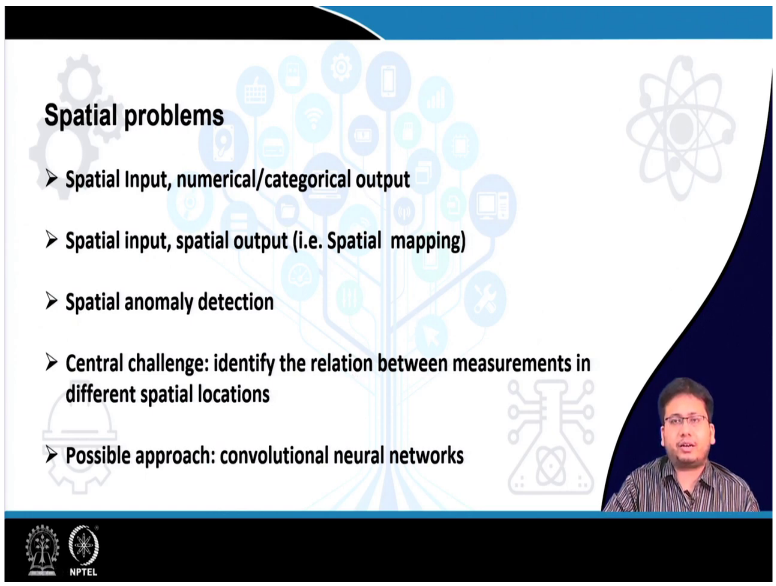
But even those teleconnections the teleconnections do not happens magically, the teleconnection happens by due to some or a chain reactions of events. And so and whatever influence of those remote phenomena are impacting the Indian monsoon that is the target quantity, like it will the impact will definitely involve other variables at more at more proximal locations.

So, to identify the predictors of a particular event we may want to give more importance to local predictors. So, that is we that is we may want some kind of spatial or temporal proximity as some kind of a structure on the predictors that are actually selected in the in this feature selection problem. So, that requires us to build more sophisticated regularization function beyond sparsity.

So, there are like elastic net regularizers and so on which actually try to impose these kinds of sparsity constants and they have been used in earth system sciences.

Now, similarly when we talk about the concept problem of dimensionality reduction what we want is some kind of linear or non-linear combination of the different features. So, what should the ideal combination of features be? That is another problem which machine learning may be interested in.

(Refer Slide Time: 11:01)



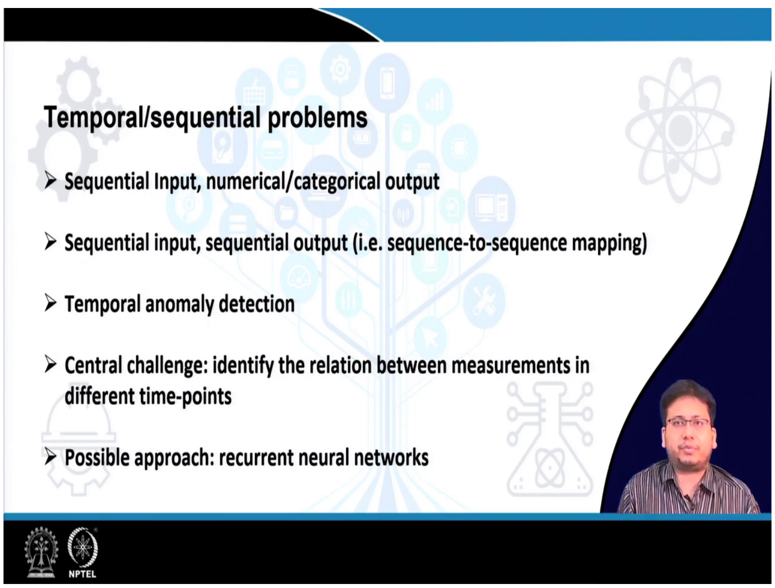
So, now there are spatial, so since the data is spatio temporal in earth system sciences there are Spatial problems and Temporal problems. In case of spatial problems the input is like is some kind of a spatial distribution of a quantity or of several quantities and the output may be just a numerical or a single numerical or categorical scalar value or it may be that the output is also a some kind of a spatial map. So, that we can call as a spatial mapping problem or a like for example, something like an image to image translation.

So, an image is a spatial map and the output is also another image which is also a spatial map. So, that is an example of a spatial mapping problem. Then there are spatial anomaly detection problems. So, suppose we I have a spatial map of a particular variable out of which I want to identify some anomaly events which are like either individual values which are different from their neighbors or some specific events like say droughts and things like that. The central challenge here is to identify the relation between measurements in different spatial locations.

So, like we know the that there are spatial influences when we are talking about one particular that is a measurement of a variable at a particular location, it cannot be viewed in isolation. We have to consider like other variables also and how they impact or how they influence that particular the variable at that particular location that we are interested in.

So, one approach of doing this is the convolutional neural network. So, convolutional neural network which we will discuss in more detail in the next lecture, they like they are actually capable of like quantifying or like or representing the interaction between different spatial locations.

(Refer Slide Time: 12:55)

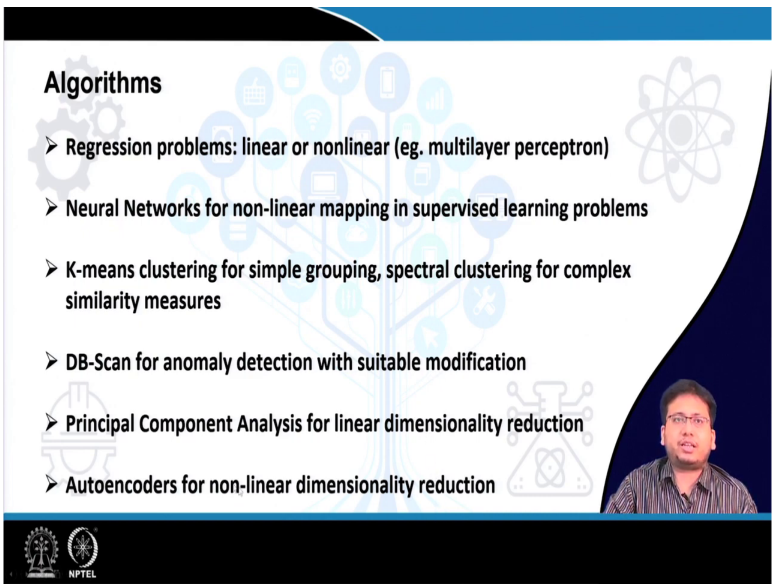


Next is the other is the Temporal or Sequential problem, where the input is something like a sequence or a something like a time series and the output might be let us say some kind of a numerical or categorical value real that is a real number.

Or it can also be the output can also be sequence, that is a task of sequence to sequence mapping. So, an apart from that there is a concept of temporal anomaly detection problem. So, suppose I have a time series of a variable I may want to identify individual time points or segments of the time series in which we see some kind of unusual value being taken. So, that is the temporal anomaly detection problem.

So, just like the spatial case in this case also the central challenge is to identify the relations between the measurements at different time points and one possible approach to do that is the recurrent neural networks, which actually try to like store the previous observations like in some kind of a hidden structure and try to predict the previous I mean the future values based on those previous values using some kind of complex mapping.

(Refer Slide Time: 14:09)



Now, coming to the various algorithms of machine learning which are which might be suitable for the problems which we discussed. Say in case first of all there are the for regression problems we can use various linear or non-linear regression problem. So, for linear we already know there are there is the simple linear regression, then there are the various regularized regressions in which impose some kind of sparsity some kind of structure like sparsity etcetera on the different coefficients which I already mentioned.

Apart from that there are non-linear regressions methods like say multilayer perceptron and things like that. Then neural networks are very effective in this domain especially in the non-linear mapping purposes in the supervised learning. So, we talked about several these spatial or temporal problems in which the input may be a spatial map or a time series. So, and the and as I mentioned the central challenge is to represent the interactions between let say different location in the map or different time points in the time series.

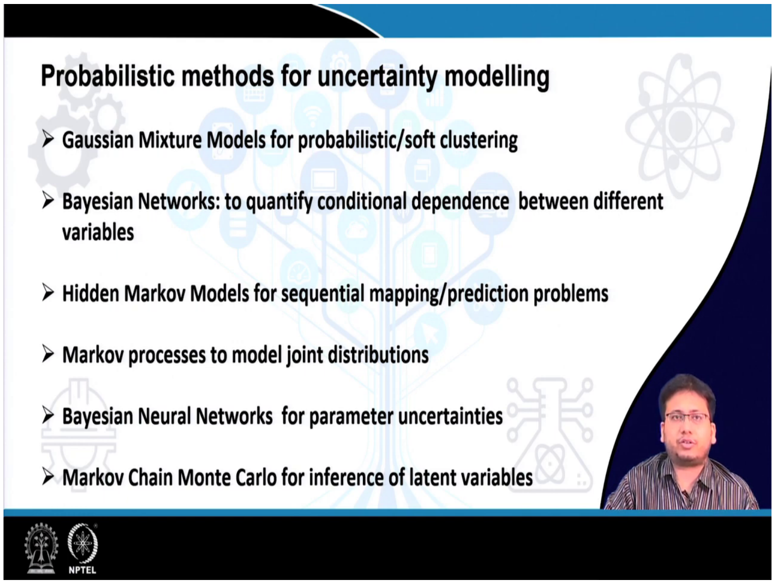
So, these kinds of interactions especially the non-linear interactions they are captured well or efficiently through neural networks. Now in case of unsupervised problems the k means clustering can be used for simple grouping, if you want more but k means clustering as we know it considers the Euclidean distances. But like say suppose between different data points which need to be clustered, the distance cannot be simply Euclidean it has to be something more sophisticated than that it may be something like customized by the user. So, or customized for that particular problem.

So, in that case we may use something like a spectral clustering which allows the like allows the user to define the similarity matrix between the different data points. Then for the anomaly detection problems it is possible to use the DB-Scan algorithm for clustering though which we already discussed, but we also identified that it has certain drawbacks.

Now, it might be possible to modify the DB-Scan algorithms to cover those drawbacks in certain specific settings of some specific problems. Now for we also talked about the dimensionality reduction problem, for that problem the principal component analysis can be useful. But that is of course a linear dimensionality reduction in the sense that the new dimensions which we get are actually linear combinations of the original dimensions.

But it if we by any chance or for any particular application, it might make sense to go for some non-linear combination of the input of the original dimensions. So, in that case we may go for auto encoders which is a non-linear dimensionality reduction approach.

(Refer Slide Time: 17:09)



Then probabilistic models are also very suitable in this domain especially for uncertainty modeling. So, like if you are doing a clustering problem it may often be difficult to define like a hard clusters like that, I may often want to define soft clusters or if you can say probabilistic cluster. Where like we can define some kind of a like we can attach some kind of a probability distribution about which point belongs to which cluster.

So, Gaussian mixture model is something like it is a it is a an elegant method for soft clustering like that or probabilistic clustering. Now we can there are also these Bayesian networks which quantify the conditional dependence between the different variables through like conditional probability distributions and so on. And these Bayesian networks are another powerful tool for solving many problems especially the inference problem or inverse problems.

That is let us say like we know the or we there are some variables which we can observe and some variables which we cannot observe and we know the relation between them. Now using the observed variables can we somehow estimate the values of the unobserved variables.

So, for these kinds of methods these Bayesian networks are often very useful especially when there is some kind of conditional independent structure involved. That is we know that each like measurable variable is depends only on a small number of the unmeasurable variables and vice versa. So, that like what I mean by that is that there are like it is not that all variables depend on each other there is some kind of conditional independence and this in conditional independence independences are utilized in a very elegant and effective way in Bayesian networks and they are useful for like probabilistic inferences also.

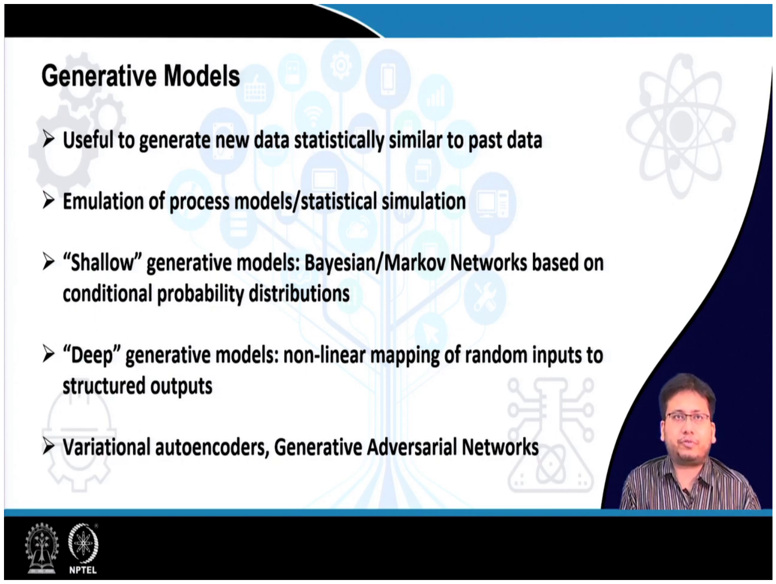
Then there is the set of problem the like in the sequential or mapping problems or the prediction problems, where like we have a sequence as an input and we may want to predict the values of that sequence in future times or we may want to do a clustering of different sequences and things like that. So, hidden Markov model is the good sequence modeling approach and which is often useful for solving these problems.

And when we have various joint distributions between a large number of variables we often use various Markov processes, like which again define some kind of conditional independence. Especially the idea of Markov is that like if we have certain if we have a target variable we say that given the neighboring variables or condition on the neighboring variables, the non neighboring variables are independent of that target variable.

So, that this Markov processes they can be either in the temporal domain or in the spatial domain or both and they are very useful for modelling in joint distributions of these random variables. And finally, we also have nowadays Bayesian neural network, so a neural network of course, has millions of parameters which are basically the weights of the different neuron to neuron collections. But instead of like putting fixed values on these different edge weights we may actually want to represent them with some kind of a distribution.

So, as to allow for some kind of uncertainty in the predictions which are made by the neural network so that is the idea of Bayesian neural network and so whenever we have these all of the above approaches they involve some kind of inference problem or some kind of that is there are some latent variables or parameters. Which are expressed as random variables and we and based on observations we want to estimate their values?

(Refer Slide Time: 21:30)



So, that is often this kind of probabilistic inference is often done using a class of methods known as Markov chain Monte Carlo which is basically like sampling a large number of values.

Then there are the class of models called the generative models, where the task is to generate new data which is statistically similar to the past data, but may but not exactly similar. So say for example, let us say I have given you lots of examples of lots of human faces and I want to draw new human faces which may not be exactly identical to the any of the person faces which we have come across earlier.

But so this is kind of a like we can say a face generation problem, which is it is different from the face recognition or face detection problems, but it is like a it is like we want to draw new human faces. So, like these kinds of generative models they can be useful in the domain of earth system sciences also.

Say for example, when we are talking about these process models which do the simulations of these geophysical processes and they basically give you artificial data about the future are starting from some kind of initial conditions. It is basically the same as what we talked about that is it is creating some kind of distribution of the variables of interest in the future, which is like which and it is which can be expected that those simulated values they will be statistically similar to past historical or observed values, but not identical to them.

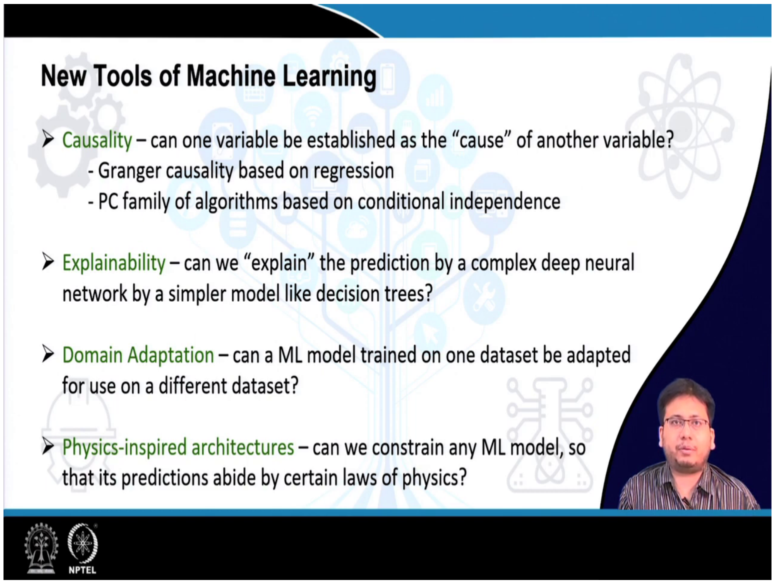
So, this is often achieved with either through statistical simulations or by emulations of these process models, that is we can by statistical simulation I mean that instead of the process the models they are often based on large number of model equations but statistical simulation is like is more using probability distributions and things like that. So, statistical simulation is one way in which machine learning can be used for these simulations, alternatively as we discussed in the previous lecture also machine learning models can be used as surrogates to emulate the equation based process models.

Now, in either cases the task of course is to generate data which is statistically equivalent to past or observed data. So, one idea is shallow generative models in which we build some kind of a Bayesian or Markov networks based on conditional probability distributions. The other is the deep generative models where we involve some kind of non-linear mapping of random inputs to structured outputs.

So, random inputs like because like the inputs should be when we are doing the simulations either the inputs should be random or like they may be specified by the user. In either case some kind of randomness has to be there to take into account the uncertainties of the simulations.

So, deep generative models they actually like have a highly complex non-linear structure to propagate the input uncertainties like at multiple levels and finally reaching the output variables. So, variational autoencoders and generative adversarial networks, these are examples of generative models that are suitable for like such simulation processes.

(Refer Slide Time: 25:12)



Now, in recent years various new tools have come in machine learning the tools of Causality, Explainability and so on; which are all very suitable and appropriate for the domain of earth science problems. Say we have already talked about causality in the previous some of the previous lectures. So, like can we can one variable be established as a cause of another variable.

So, we have already discussed the Granger causality approach and then there is also the concept of Pearl causality. So, there are family of algorithms called the PC family of algorithms named after their founders like. So, these algorithms are all based on the concept of conditional independence between different variables. So, that is about causality.

Then there is another new like highly popular emerging area called the explainable machine learning. So, suppose we have the predictions by some complex deep neural network. Now can we explain those predictions by simpler models like decision trees, because these complex neural networks are often like they are often like black boxes. That is they do some very highly complicated non-linear mapping of the input to produce the output.

Those outputs may be like the their predictions may be very accurate also. But often it is very difficult for domain scientist to understand what exactly the model did, why at all it was able to do the correct prediction.

So, for such situations we need explainable AI, where like we try to for individual prediction that is an individual mapping of the input to the output we can actually build something like a simply simple or interpretable model like say decision tree. So, decision tree is very unlike neural network which is often like a black box, decision tree is something which is very understandable because we see a sequence of features on the basis of which we are taking the final decisions.

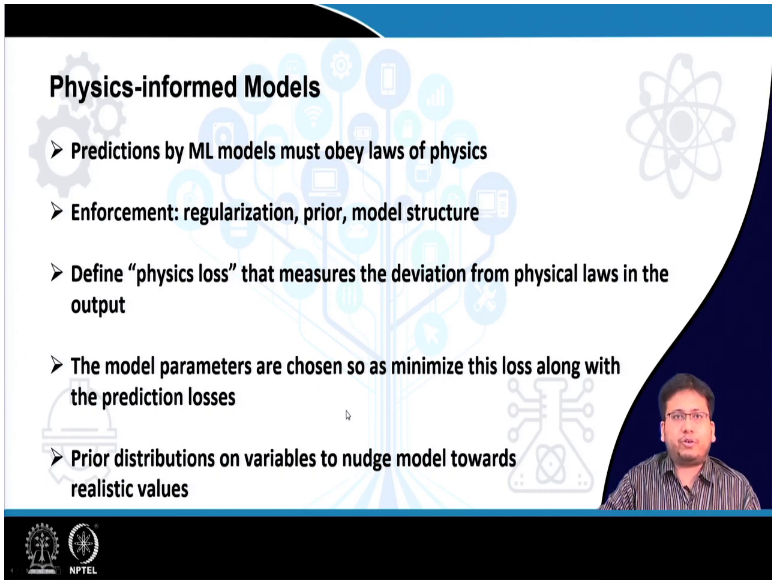
So, like if we are to find the explanation of a particular prediction decision trees are often useful for building this kind of explainable models. Then there is a concept of domain adaptation. So, suppose we train the machine learning models on one data set like then can that same model be adapted for use to a different data set.

Now, why do I want to do that why do not I train it separately there are 2 reasons, one is that machine learning models as I said the recent or a state of the earth AI models they have millions of parameters and they need a huge volume of training data. And now it is not feasible to keep training these models repeatedly for every single problems.

So, often we take a pre-trained network and just maybe fine tune it for one for the problem at hand. So, if I already have trained my model on one data set, I need not train it right from the beginning on another data set I may simply fine tune it. Alternatively it might also happen that like we know that some parts of the world have a very good sensor network say the more developed parts of the world, while if we consider say places like Africa then such sensing facilities or observation facilities are not good.

So, but if I want to use some model which has been trained on say Europe and I want to use it in over Africa, then it is a like we do not know if the if it will work properly because the conditions might be different. Now since I do not have enough observations from Africa I cannot train an Africa specific model from the scratch. So, what we I do is domain adaptation, I take the Europe trained model and do whatever little data is available from Africa I do the some kind of fine tuning on it and try to like use it further. And then finally, there are the Physics informed architectures.

(Refer Slide Time: 29:33)

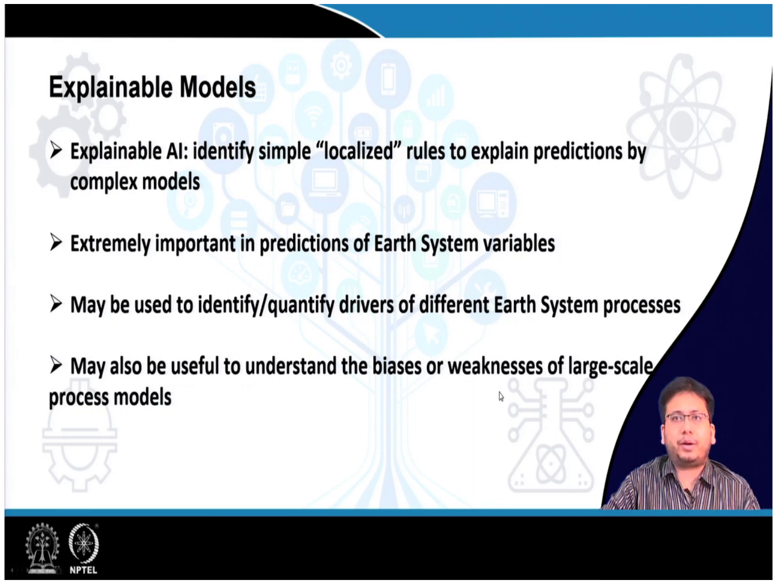


So, physics informed models like we know that the predictions like by the machine learning models they must obey the laws of physics, that is let say conservation of mass, conservation of energy and various other domain specific laws but machine learning model per se cannot ensure that. So, it has to be enforced through regularizations or priors or even through the model structures.

So, typically what people do is they define some kind of a physics loss that measures the deviation from the physical laws in the output. And just like we have the usual empirical loss which measures how different the predictions are from the like from the true values in case, on the training data and try to tune the architect the models parameters, so as to minimize that empirical loss. So, along with the empirical loss we can add the physics loss also.

So, as that the learning algorithm it can modify the parameters, so as to minimize this kind of physics loss. And then of course, in case of probabilistic models we can have prior distributions on different variables to nudge the model towards realistic values.

(Refer Slide Time: 30:44)

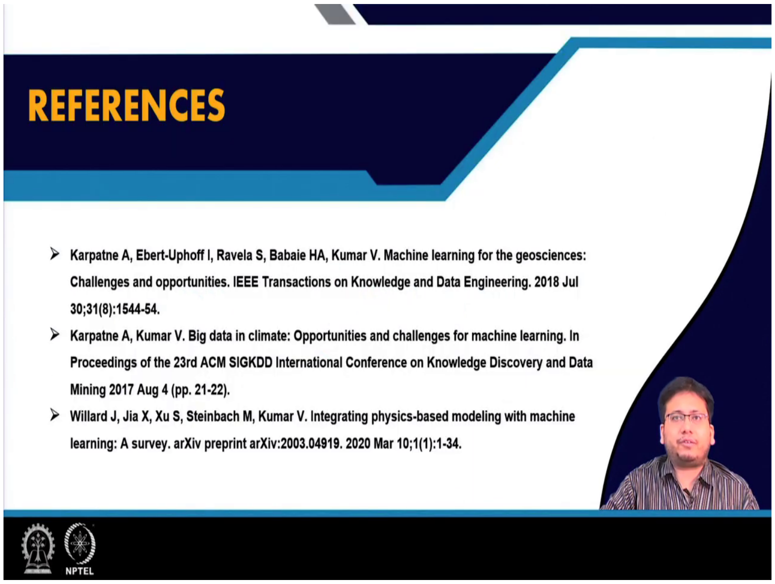


And then there is of course the concept of Explainable Models, which we have already talked about. These are also extremely important for the predictions of earth system variables.

And in fact, these explainable models or the concepts of explainable AI may it actually be used to identify or quantify the drivers of different earth system processes and we may also it may be also be useful to understand the biases or weaknesses of large scale process models.

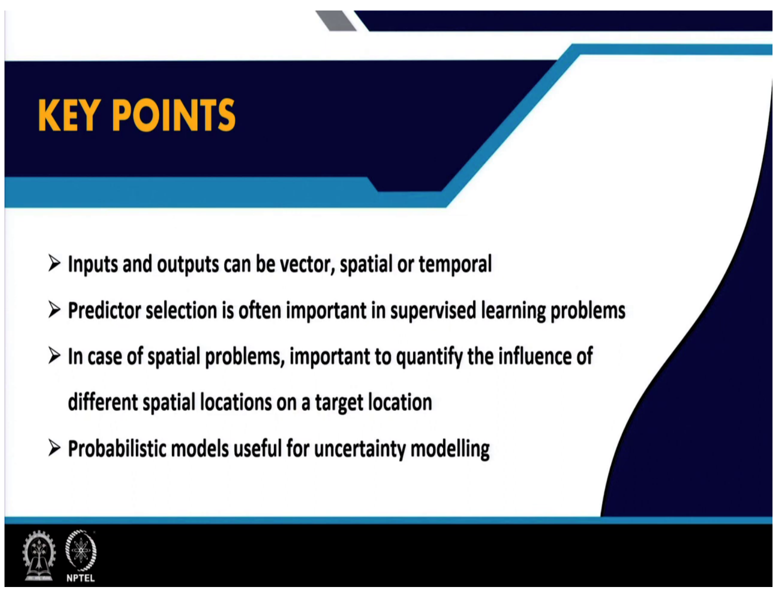
See suppose I learn suppose our process model is run and it is found that it is creating some errors or biases of a particular variable. Now why did that bias happened, what is the source of that bias? So, for that I may want to like build some kind of an explainable model which shows why it predicted what it did? And then we can identify the source of it is bias.

(Refer Slide Time: 31:37)



So, the these are the various references as we talked about in the previous lecture also which talk about the how machine learning and deep learning can be used in the earth system sciences.

(Refer Slide Time: 31:48)



So, the key points to take away from this lecture is that the inputs and outputs of these machine learning problems from earth system sciences, this can be either vector or this can be like I mean they can be either some individual vector or scalar quantities. So, they can be spatial or temporal. The prediction problems is often important in supervised learning problems.

Now, in case of spatial problems it is important to quantify the influence of different spatial locations on the target locations which is achieved by things like CNNS. In case of temporal problems also we have like equivalent problems in the temporal domain and when it comes to uncertainty modeling, so probabilistic models are very suitable for that.

So, with that we come to the end of this lecture. In the following lectures we will discuss a bit more about specific types of these machine learning models like this convolutional and recurrent models, which we talked about like briefly today. So, we will meet you again in the next lecture. So, till then goodbye.