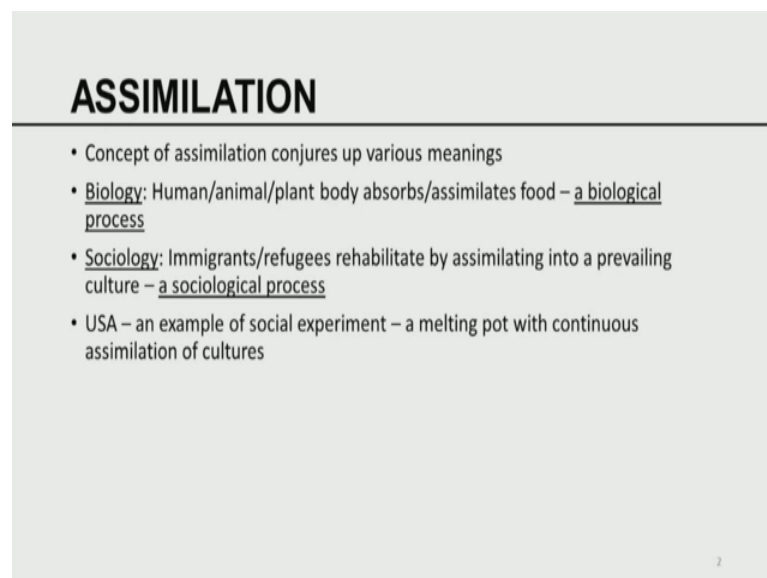


**Dynamic Data Assimilation**  
**Prof. S Lakshmivarahan**  
**School of Computer Science**  
**Indian Institute of Technology, Madras**

**Lecture - 01**  
**An Overview**

Welcome to this course on data assimilation. We are going to provide a very broad background about the theoretical aspects of data assimilation. Our first lecture will be on providing an introduction and an overview. I would like to start with a question, what is assimilation. The concept of assimilation in general conduce of various meanings in different disciplines.

(Refer Slide Time: 00:39)



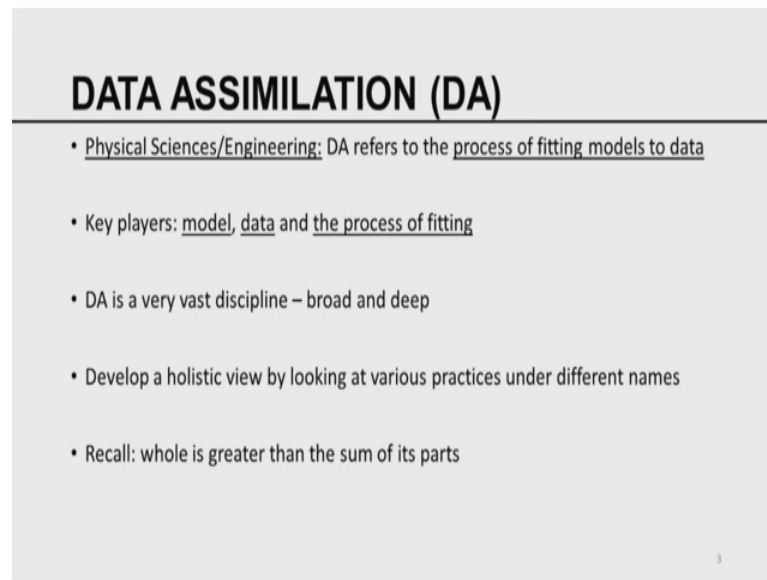
**ASSIMILATION**

- Concept of assimilation conjures up various meanings
- Biology: Human/animal/plant body absorbs/assimilates food – a biological process
- Sociology: Immigrants/refugees rehabilitate by assimilating into a prevailing culture – a sociological process
- USA – an example of social experiment – a melting pot with continuous assimilation of cultures

2

In biology for example, the human animal plant body assimilates or absorbs food. So, here the word assimilation refers to the process of absorption of food by living organism be it human animal or plant. So, here assimilation refers to a biological process. Assimilation also happens within the context of sociology. Immigrants, refugees from one country, they try to rehabilitate in another country by assimilating into the prevailing culture. In this context, assimilation refers to the sociological process. As an example, united state is considered to be an example of social experiment. It is a melting pot of salts; where there is continuous assimilation of all cultures from all around the world.

(Refer Slide Time: 01:45)



## DATA ASSIMILATION (DA)

- Physical Sciences/Engineering; DA refers to the process of fitting models to data
- Key players: model, data and the process of fitting
- DA is a very vast discipline – broad and deep
- Develop a holistic view by looking at various practices under different names
- Recall: whole is greater than the sum of its parts

3

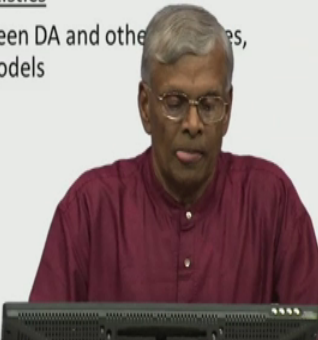
However, within the context of physical sciences and engineering, assimilation refers to the process of fitting models to data. So, in this process the key players are model, data and the process of fitting data to the model. So, this involves 3 different disciplines. One is science of model building, second one is ability to absorb nature and the observations give rise to data, data is available model is available, but we need to bring them together.

This bringing together that is the assimilation process that process is also called fitting models to data. Looking from this prospective data assimilation is a very broad and a vast discipline. It is very broad, it is very deep. We would like to develop a very holistic view by looking at various practices that go with of a names, but the underlying process is always related to data assimilation. We call the whole saying the whole is greater than the some other parts. So, the process of fitting models to data the process of various types of assimilation, these are all various parts that come to we called the science of data assimilation, the whole is much bigger than many of it is parts.

(Refer Slide Time: 03:24)

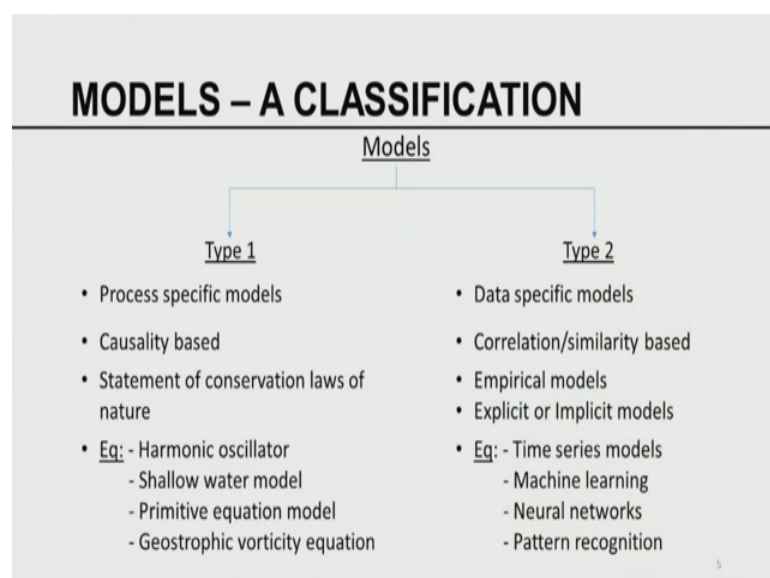
## DA AS CURVE FITTING/REGRESSION

- The notion of fitting in DA is the same as that is used in
  - a) Deterministic curve fitting/interpolation in Numerical analysis
  - b) Statistical regression analysis in Statistics
- To further understand the relation between DA and other practices, introduce a high level classification of models



First, I would like to talk about data assimilation as a curve fitting or a regression process. The notion of fitting in data assimilation is the same as that is used in deterministic curve fitting or interpolation that is often used in numerical analysis. It is also intimately related to statistical regression analysis. To further understand the relation between data assimilation and other practices, we need to be able to introduce different classifications of models. Please recall, models, data and the process of fitting are the 3 components. Now we are going to look into properties of various kinds of models that occur in most of all sciences and engineering.

(Refer Slide Time: 04:17)



Models in general can be classified into 2 types. Type 1 is called process specific model, type 2 is called data specific models. Let us first concentrate on type 1 models. Here we understand the underlying scientific process it is apply. These are developed based on causality. If there is a force there is a reaction to that. That is always a cause and effect relation. In sciences many of these causal relations are nicely captured in the form of conservation laws.

Here are some examples of models sub type 1. The harmonic oscillator that describes the motion of a pendulum in a friction free environment. The shallow water model that can be used to describe the motional waves in the ocean. The primitive equation model that is the basis for many types of whether forecasting. Geotropic vorticity equation in particular one of the equations which is very generic and has applications in particular it is it is the basis for hurricane predictions.

So, we can see models comes from various shapes and forms, each one of them are developed based on fundamental scientific principles. Causality based process based these are essentially statements of conservation laws. As opposed to these the second type of model they are called data specific models. Here there is no underlying causation, there is no cause or effect, there is well established. All that we have is the availability of a bunch of data. These data are observations of mother nature. Given a bunch of observations of mother natures behaviour at various times, it is up to as to be able to mine the underlying information, and that mining is done with respect to finding the correlation between the data.

This is also called correlation based or similarity based. These models are essentially empirical as opposed to based on physical principles. These models can be either an explicit model or an implicit model. For example, I have a time series. The time series could be value of IBM stock over each day. The maximum temperature in Bangalore each day. They unemployment rate in the state of Karnataka every month. This is an example of a time series, we would like to be able to build models for using these time series data to be predict how the unemployment will be in a year from now, what will be the maximum temperature in Bangalore in middle of January. These are called explicit models, these are empirical models these are essentially based on correlation.

Machine learning provides again lots of opportunities for model building based on data. Neural network is an example of a specific kind of machine learning process; where we try to build a specific neural network to be able to undertake a specific task of being able to classify. Neural network models are essentially implicit models. You given output you try to tune the model until you get the required output.

You probably may not able to explain why and how except that if you did the right thing you can make it work. Again, these are data specific for example, the time series model that is used for IBM stock price may not be the same title that can be used to predict tomorrows maximum temperature in Bangalore or unemployment percentage so on and so forth. So, each one of this data set has to be looked at separately, for each one of this sad data set we have to uncover the underlying serial correlation. Here I would like to make a statement about they definite need for correlation.

If a collection of data are not correlated means one does not influence the other if one does not influence the other, we cannot be we may not be able to predict anything at all. So, ultimate aim in model building is to be able to predict. Once we have model I can pull the model solution forward to be able to make predictions. So, correlation is very fundamental attribute of data specific models. And it is using this correlation the model is able to extrapolate make predictions. Pattern recognition is another example of explicit models that are again data specific.

(Refer Slide Time: 09:44)

## DA AND ESTIMATION THEORY

- Development of Type 1 models and DA for these models are separate processes
- Structure of the model in terms of spatio-temporal evolution/relation among the state variables are well specified
- Yet, these models have several unknowns: Initial conditions (IC), Boundary Conditions (BC) and parameters
- Goal of DA is to estimate these unknowns from the knowledge of the available data that is supposed to contain information about the unknowns
- DA is closely related to classical Estimation Theory

Now, another dimension to data assimilation I would like to bring to forth. Data is the development of type 1 models and data assimilation models of are these models are separate processes. For example, if physicists to may try to develop a primitive equation model, an observation specialist may try to get data from satellite or radar. A data assimilation person then comes into play to be able to understand the data set and to be able to bring them together. So, model development and data assimilation process are 2 separate processes, it is the data assimilation person is the one who sits in middle who tries to talks to the both models and data.

The structure of the model in terms of spatio temporal evolution, relation among the state variables are specified by the details of the model. Yet, many of these models may have several are unknowns. For example, in a primitive equation model, we may have in initial condition we may have a boundary condition. Depending on the type of process is involve different kinds of parameters may also became a part of this equation.

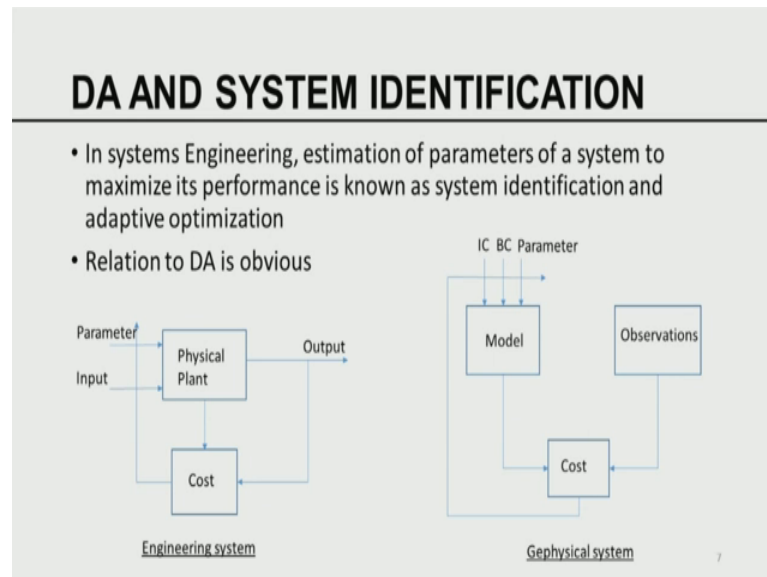
I know some class of primitive equation model helps you to predict whether in a particular situation, but to be able to initialize the model to (Refer Time: 11:08) the model forward I would need initial conditions I would need boundary conditions I would need parameters. So, models are specified modulo initial conditions boundary conditions and parameters, to be able to run the model forward we need the values of these. And how do we get to know the values of this parameters, that is where we use the data. So, in here the goal of data assimilation is to be able to estimate, the unknown from the knowledge of available data that is supposed to contain information about the unknowns.

I would like to explain this little bit further. For example, a (Refer Time: 11:48) a causy geostrophic watercity equation is often used to be able to predict the movement of a hurricane, but that is a very generic kind of equation. If I want to be able to use that model to be able to predict the movement of a hurricane, I need to know where it is today and what is the, what are the coordinates of that what are the pressure differences what are the various other attributes.

So, somebody has to fly planes into these hurricanes they collect data they bring the data to the office so that tells you observations about the phenomena I understand the science the physics behind the hurricane we would like to be able to bring these 2 together in order to be able to estimate the unknowns initial condition boundary conditions and

parameters. Therefore, you can readily see data assimilation is closely related to the classical estimation problem. This is fitting the observation of hurricane to the hurricane forecast models because developed earlier. This aspect is related to estimation theory.

(Refer Slide Time: 12:59)



There is another point of view one can take. Data assimilation has also lot in common with systems identification as used in systems engineering. In systems engineering estimation parameters of a system to maximize the performance; is known as system identification or adaptive identification.

So, let us look at the picture here. This one is an engineering system. That is a physical plant. The physical plant may be a chemical plant, it could be an aircraft it could be a ship. Or it could be any kind of engineering device. It has input it has parameters which are knobs that meant to content to change the behaviour of the system. For example, if this where to be chemical plant, the input could be a raw materials, the parameters could be the presence of catalyst. It could be temperature, it could be a pressure, it could be a concentration.

So, these all are various parameters that one can control. So, the output of the plant depends on the physical properties of the engineering system along with parameters and input would like to be able to understand the dependence of output of the input. So, we try to express that relation in the form of a cost function. We would like to be able to

maximize or minimize. If it is cost is always minimization, if it is profit it is generally return as a maximization is an optimization problem.

So, we would like to be able to change the input of the parameters. In a feedback loop, such that I tried to maximize the functioning of this plant this often occurs in many of the engineering system. You can readily see in here data assimilation is done online. A physical plant is operating, it provides an output that is the function of the input, I do not know what is the maximum output possible I am going to learn the maximum output by sequentially changing the input and the parameters in the loop based on a pre-specified criterion called the cost function. This often happens in the all branches of engineering.

Now, let us come to physically occurring systems such as meteorology, such as geophysical sciences. There are observations of geophysical sciences we often do the observations geophysical come from satellites, radars, balloons, ships collect information aeroplane collect information. And these days we have ground base stations all around over the ocean we have lots of (Refer Time: 15:48).

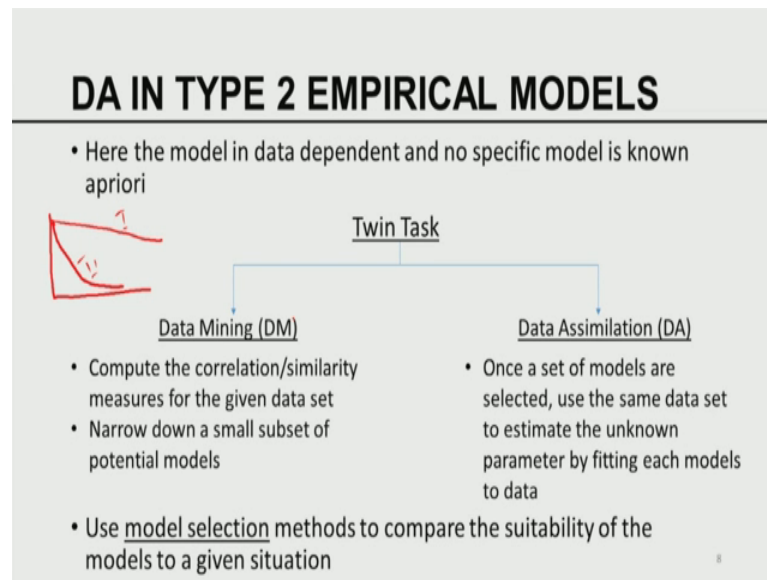
So, observation comes in various shapes and forms, the observation contains secrets about the functioning of the nature. The modellers believes the mother nature behaves in the particular way the encapsulate the behaviour the model based on their understanding in the form of a process-based models that we talk about. The process base models in general can be a dynamic model. If it with the dynamic model can be based on ordinary differential equations or partial differential equations. These models have initial conditions, boundary conditions and parameters.

You only know that model describes the overall observations in some specific ways, but we would like to be able to fit the model to the data so that a specific model can be used to accommodate the specific set of the observation to be able to generate predictions about a particular hurricane a particular tornado, a particular natural occurring event. So, this is how we use the data which are contained the observation and the model fitting model to data.

So, this is an alternate view this is the view that often occurs within the geophysical domain, oceanography, hydrology, atmospheric sciences, volcanology and so on.



(Refer Slide Time: 17:22)



Now, I would like to talk about how data assimilation takes place in empirical models. In empirical model, there is nothing we do not have physics. All that we have is the bunch of data. So, I have to use this data in 2 ways. So, we have a twin task. One is to be able to build the model, and then once we have a general understanding of a class of models that we can use, then we have to do data assimilation on these model using the data. So, let us look at the data mining part the data mining part is the process by which I would like to be able to understand the structure that underlie the creation of data. So, what is that we do? We compute the correlations. We compute the similarity measures in a given data set. Based on the correlation we can narrow down a potential class of models just small subset.

As an example, in the case of time series we may say looking at the correlation structure, is the correlation persist for a long time; that means, it has a long memory. If the correlation comes down to low values very fast, if it decreases for example, your correlation could be if a correlation structure could be like this. Here the correlation does not decrease too fast. So, this these kinds of process supposed to have long memory, another key would be correlation comes down to 0 very quickly.

So, the class of models for this this is one kind of model. This is second kind of model. There are guidelines to choose certain kinds of models for correlations of type 1 for correlations of type 2. Even here one has to contain with the orders of the model, but by

looking at it we cannot say this model work better than that model work better. So, we have to be as we have to be flexible. Try to be able to have at least couple of different ways of looking at it. So, once we have narrow down the choice of models, it is like having a barotropic watercity equation, it is having a primitive equation.

So, we have to come to the same level as in the process-based models by looking at the data to do data mining process, we will talk a lot about data mining a little later in this lecture series in this introductory chapter. Now once the model is available, again in the case of time series models the models are dynamic, given any dynamic model, we will have initial conditions we have parameters to be able to pull the model forward, we need to be able to know the values of these parameters again.

So, once the set of models are selected we can use the same data set to estimate the unknown parameters by fitting each of the models of the data. So now, we can readily see independent of the types of models. The process of data assimilation acquires the presence of a model. The presence of an associated set of data, and then the process of fitting that story must come pretty clear by now. So, the you so, here there is an added difficulty in the case of empirical thing. In the case of process-based models for hurricane we know what class of models to use. For certain times of atmospheric motions, we know what class of models to use. For certain times of atmospheric motions, we know what seven types of models to use. For oceanography they have pretty good understanding what types of models to use.

So, a given process has associated with it the already understood class of models. In the case of empirical model building models of types to there is no such clear-cut association. There is no clear-cut way to be able to decide this model will work better than that. So, we have to always contend with a class of models. So, within this context the notion of model selection becomes a very important and a fundamental task.

There are several methods one can look into to be able to develop, and compare the quality of each of these models to a particular given situation. And we are ongoing to (Refer Time: 21:49) into the model selection process those of us. Who are working in the empirical model building has to be cognizant of the requirement of being able to use very good criteria to be able to select good models for a particular given kind of a data set.

(Refer Slide Time: 22:06)

## DM AND DA – TIME SERIES ANALYSIS – EXPLICIT MODEL

- DM step: Given a time series, first compute and plot the correlogram,  $\rho(k)$  vs.  $k$  where  $\rho(k)$  is the Correlation between data that are  $k$ -steps apart for  $k \geq 0$
- Compare this correlogram visually with those in an album of correlograms for various known classes of ARIMA models and select the ones that are “close”
- DA step: Estimate the parameters of the different models using the same data set
- Model selection: use well known measures to compare the suitability of models for a given situation

$t_k$  DATA  
k  
ARMA(1,1)

So now we have seen 2 aspects of data mining and data simulation. I am going to use time series analysis again to be able to reinforce some of the basic thought process we have already discussed. This is an example of explicit model. Here data mining step, what is a data mining step here? Given a time series first compute and part the correlograms, and what is the correlograms is the one that we already saw. I am going to give another example of a correlogram. This is the time index  $k$ , this is row of  $k$ . So, what does  $k$  refers to?  $K$  refers to the separation between 2 data set, time  $t$  and time  $t$  plus  $k$ . How are the data and time  $t$  and time  $t$  plus  $k$  are associated with each other? For example, if I have a maximum temperature tomorrow or today.

How is today's maximum temperature related to yesterday's maximum temperature? Day before yesterdays, 10 days ago. You can readily see today's maximum temperature may be related to yesterdays, but to a lesser degree day before yesterday to a lesser degree to a week almost no link to a month from today. So, this correlation of maximum temperature one can think of it like this. If you think of again unemployment percentage unemployment percentage is not vary day by day they view repeat of month. So, today's unemployment is very much related to yesterday's unemployment. So, what is the correlation between unemployment. So, in time separated by one day one month 6 months one year.

Where the basic units of time could be one day. So,  $k$  refers to the number of days that separates 2-time epochs between which I am interested in correlation. So, this is what is called a correlogram, plot of row of  $k$  versus  $k$ , row of  $k$  is a correlation between the data are case steps apart. Then what is that we do. So, this is data dependent. So, this a created a particular data. So, this is some sort of a summary of what this data tells us. Then how do you utilize this?

Mathematicians have helped to creates several different types of models for time series analysis. These models are called ARIMA. AR for auto regressive. I for integrated. MA for moving. So, it is a 3 different families of models is called AR model, integrated model, moving average model. ARIMA models; these ARIMA models are a very broad class of potential model that one can built. Mathematicians have helped us to be able to build all these models out of time, and they have analysed the underlying correlation properties of each of these models and have catalogued.

In fact, they are they have developed albums and albums of properties of correlations of various types of models for example. And AR model could be an first order model, second order model, an MA model could be first order model second order model. In AR MA first order AR second order MA. So, in general one can have AR MA  $p$   $q$ . So,  $p$  refers to AR type  $q$  refers to MA type by changing,  $p$  is an integer  $q$  is an integer. By changing  $p$  and  $q$  I can get a whole family of models. If each of these models are created, I can mathematically compute what their correlation should be, and I can plot this correlations and create an album. This album is the fundamental basis for almost all of time series analysis. How do I use this album? You have already cranked out the correlation for the specific data set.

Then you visually compare the given picture the photo with the album that you already have. You narrow down which of the pictures in the album, click symbol, the one that you have that depicts the particular data set we have. And that help you to narrow down. May be it looks like AR 2 and 3. May be it is AR MA 1 1. You do not simply say this is it you look at the things that are closed, you narrow down the model 2 3 4. Each order will have different kinds of parameterization.

So, use the same data to be able to estimate the various parameters of the each of the models. On once you fit different models to the same data, then you can compute what is

called the error of the forecast, the error in the model. Use that error to further select much more finally, the appropriate model that could be used in this particular case. So, the data assimilation step now. So, the that is so, the first one is analysis of data. Second one is comparison of the given data with the album.

Once a particular data has been narrowed down, then you come back to the old step; called the data assimilation step, estimate the parameters of the different models using the same data set. So, here I would like to emphasize the difference in the process-based models, models are created from by the scientist's observations are created by measurement scientists. They do not talk to each other, but our job is to bring them together. So, the model building and data assimilation part are separate. But in the case of empirical models, while you are given is only the data.

Based on the data you have to develop a model, once you develop a model use the same data to be able to assimilate the data into the model. And again, model selection becomes an important situation in these cases. Here comparing the album, I would like to bring an analogy here, all of us happen to go to the doctor once a while. In order to understand what happens with you. Health doctors order different kinds of observations on you blood tests. It just x ray, may be an MRI or may be other chemical analysis of your blood. Once a chemical analysis of the blood x ray MRI are made available, a doctor hangs them all together and he looks at various he or she looks at various signals that is of the observation provide us.

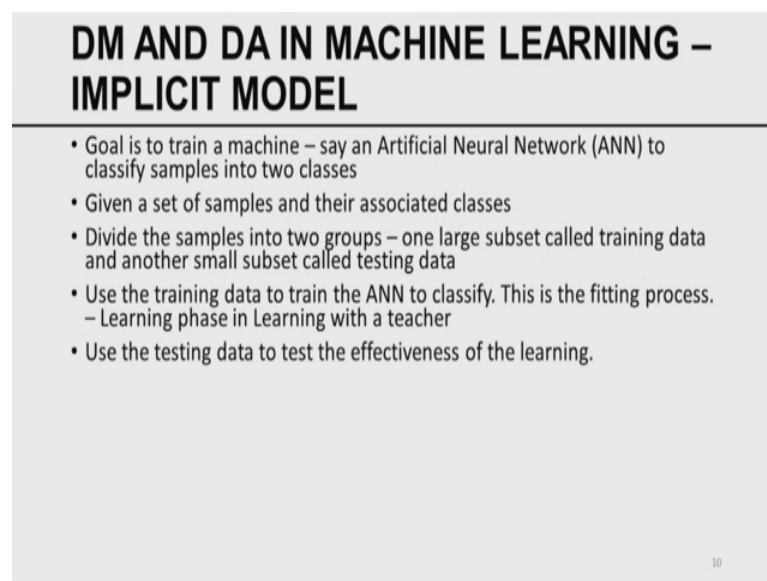
Then they go to the anatomy book. The anatomy books describes what an ideal body should be, they compare the present situation with an ideal situation. Oh this looks right, this looks right, oh this looks not that good. They try to narrow down a particular possibility. And then do further steps do further steps do further steps. Once they understand then they go for the treatment.

So, to label you have this disease to label, you have that disease to label you have this problem. That is a prediction problem, to be able to predict I need, to analyze the model and how do they use the model they use the mm models or to them the anatomy textbook. So, there is a data, there is an album, they compare the data with the album to be able to generate to be able to make a forecast. Once the forecast is done with respect to medical diagnostic forecast then leads to recover treatment and then recovery. In the

case of hurricane forecast, we make a forecast, then we tell the public there is going to be a 20-inch rain tomorrow. Move people up, up to the to the higher grounds and so on so forth.

So, the aim of data assimilation in general is to be able to fit models to data. So, that we can generate useful forecast. And the forecast can be used in many different ways for public consumption. So, that is the overall underlying philosophy that will be pursuing in this course.

(Refer Slide Time: 30:33)



**DM AND DA IN MACHINE LEARNING – IMPLICIT MODEL**

- Goal is to train a machine – say an Artificial Neural Network (ANN) to classify samples into two classes
- Given a set of samples and their associated classes
- Divide the samples into two groups – one large subset called training data and another small subset called testing data
- Use the training data to train the ANN to classify. This is the fitting process.
  - Learning phase in Learning with a teacher
- Use the testing data to test the effectiveness of the learning.

10

So, data mining and data assimilation DM for data mining DA for data assimilation. They are also used in machine learning in this case the models are implicit. We are not going to do much a machine learning in this class. We are not going to do much of times series in this class. That is why I would like to spend a little bit of time to start with to be able to see the other relations to data assimilation.

So, goal is to train a machine, say an artificial neural network, to classify samples into 2 classes. Given a for example, medical diagnosis for example, blood analysis. A blood chemistry gives the blood chemistry the whole idea is that can you feed this blood chemistry information into neural network. You change the neural network. So, it is a well to be able to predict. Yes, you have hepatitis a. No, you have hepatitis b, no you have hepatitis c.

50 years ago, we did not know that there was a disease called hepatitis. Then they know there is this is called hepatitis. Then they found out hepatitis is not one kind it's multiple kind a and b. Then they found out it is not a b; that is, also a c. Do you have you discovered all aspects of hepatitis the answer is no we never know that to be c d e f g h. So, once you know, that these are the characteristics the blood that corresponds to a b c, we would like to be able to automate the decision process.

So, neural network is essentially a classificatory mechanism into, which you feed this chemical information, it prints out the particular kind of disease that corresponds to that that is a classification for a typical classification problem. Another classification problem is in a post office I would like to be able to develop a machine that can read the addresses, to be able to sort them, human can sort, but I would like to automate the sorting process by machines.

So, that is a pattern recognition process, but I write in it a very different from yours, but the machine cannot handle all the ways of handwriting. So, we are going to say, hey if you want to be able to use the machine to classify you have to type in a particular form. Once you have typed in a particular format, a machine can be taught to be able to classify addresses in different groups that is artificial network. Making this network do these processes is called learning phase. So, given a set of samples and their associated classes. So, what is that they mean these corresponds to type a hepatitis these corresponds to type b hepatitis these corresponds to type 3 c hepatitis.

So, there must be an expert who already knows a particular analysis also the classification he has labelled them. So, once I have a sample of their associated classes, then what do I do you divide the samples into 2 groups. One large subset called the training data, another a small subset called the testing data. So, we use the larger training data to be able to train the machine; that is called learning with the teacher. They always learn with the teacher, there are 2 kinds of learning. learning with a teacher, learning without a teacher in a e.

In case of learning with a teacher you learn by recognizing that you have made a mistake. In the case of learning with a teacher, I already know for this input this must be the output. I know for specified. So, you can train using this data the known classification and artificial neural network to classify. So, this is essentially a fitting

process. In the learning this is called the learning with the teacher during this phase, I make sure on this data set this machine behaves in the best possible way.

Then I would like to understand, whether the machine will be able to do things that it has not already seen. So, if I use the same data that I use to teach, and then certify that cheating. So, I would like to be able to test it with the set of data that was not used in the training phase, and that is why we have talked about 2 subsets a smaller subset. Now you feed the data set that the machine has not seen earlier. If on this data set, the machine does very well, then you have succeeded in making the machine learn.

Though this is from a higher angle, you can look at it as a data assimilation process, we are fitting the behaviour of the network to suit the classifications of a given data set. So, you can see machine learning and data simulation have a very strong interconnection.

(Refer Slide Time: 35:18)

**DM AND DA IN ANN – IMPLICIT MODEL**

- DM phase: Choice of the structure of the ANN – number of stages, number of neurons in a given stage, number of input, number of output
- DA phase: Fit the chosen structure to the training data by minimizing the error in classification
- Testing phase: This is the prediction phase – to see how the ANN handles new situations

11

Data mining and data simulation in artificial neural network is an implicit model again. This is a quick summary of what I already described. The data mining phase, here what is the data mining phase the choice of the structure of the neural network. I simply talked about neural network how do I define the neural network it has a number of stages, number of neurons in a given stage, the total number of inputs, the total number of outputs.

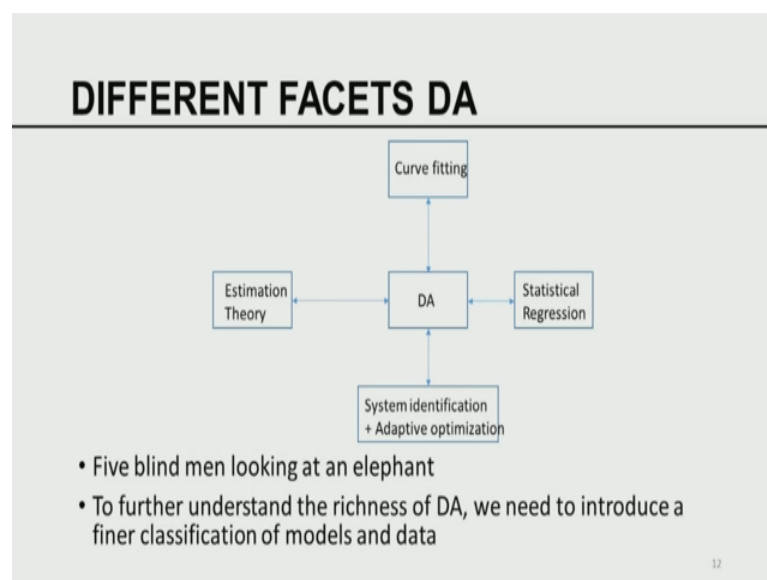


These essentially describe the topology of the underlying artificial neural network. This essentially comes from experience, this is the data mining phase, the mining relates to a lot of experience that one has should I have 2 layers should I have one layer should I have 5 layers should I have one output. Should I have one output or 2 outputs. So, these are all the decisions that engineer has to make in the design of the networks.

Once the network is designed the structure of the network is designed, we are the same level as in the ARMA case, I have picked the model. We are on the same level as I had picked a hurricane, I have picked the hurricane model. So, the model is already fixed in here. The data assimilation phase is to fit the chosen structure. To the training data by minimizing the error in classification. The testing phase is the prediction phase to see whether the artificial neural network handles the new situations, that is all.

So, you can see the data assimilation is intimately associated with learning. My own background my PhD work was in a essentially in machine learning, that is why I could jump into data assimilation part rather easily data assimilation is the basis for any and any kind of learning mechanisms or learning devices.

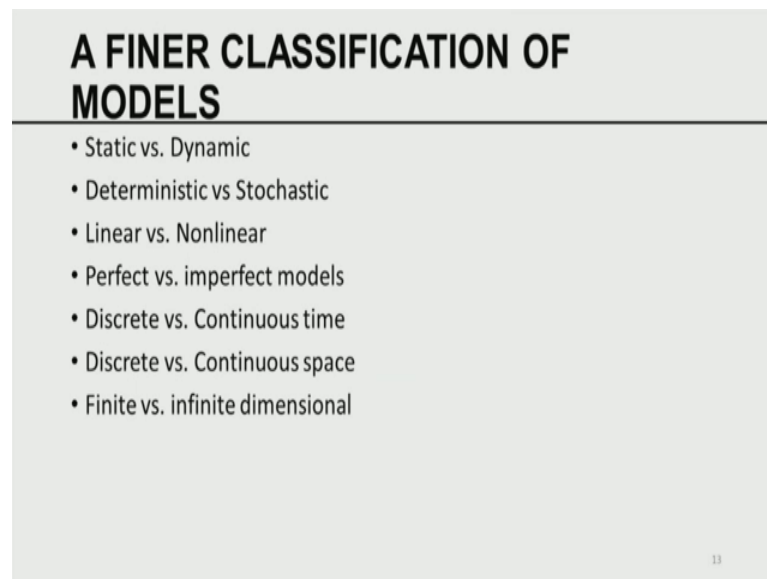
(Refer Slide Time: 37:03)



So, different facts about DA all in 1 place now. DA can be looked at curve fitting. DA can be looked at estimation theory. DA can be looked at statistical regression. DA can also be looked at system identification adaptive optimization. So, looking at DA is like 5 blind men looking at an elephant. A person and car fitting looks at, curve fitting only. A

person size (Refer Time: 37:32) statistical regression only. Our job is to be able to bring the big elephant. This what I was trying to tell in the very first slide. The sum is greater than the parts, the sum the data assimilation. To further understand the richness of data assimilation, we need to introduce further classifications of both models and data. And that is what we are now going to look at.

(Refer Slide Time: 38:01)



Until now we classified models only at a broad angle namely, process-based models or empirical models, are data specific models. That is the classification at the highest topmost level. Now I need to bring down finer nuances into the classification. So, models in general can be classified along different dimensions. A model can be static or dynamic. A model can be deterministic or stochastic. Your model can be linear or it can be non-linear. A model can be perfect or it can be imperfect.

Almost all the models are imperfect, I do not think, there is any model that is perfect. Except few I am not saying there is none no nonexistence. For example, the model that governs the motion of the planet earth around the sun. It is very nearly perfect why? How do we know that we are able to predict the loner solar eclipse over the next 100 years? I am sure enough when we saved there will be a loner eclipses you (Refer Time: 39:08) it happens.

But there are very few systems where we can make such accurate predictions. So, perfect models are few imperfect models are large. A model can operate in discrete time or in

continuous time. A model can work in discrete space and in continuous space. A model can be finite dimensional or model can be infinite dimensional. For example, a person working in static regression, he may be interested only in static deterministic linear non-linear case. A person in system identification may be interested in dynamic deterministic models linear non-linear.

A person working in time series analysis would be interested dynamic stochastic linear non-linear imperfect models that operate in discrete time. Most of the models in geophysical sciences they work in continuous time and continuous space. If a model is given by partial differential equation that is infinite dimensional in nature, if models are given by ODE there are finite dimensional.

So, you can see data simulation looks at the entire domain of modeling, with all its nuances with all its abilities to classify models along these many different dimensions. And in our course, we are going to take such a global view. So, what does it mean in this course, you can do data simulation in several different ways, you can handle any kind of different models. So, we are looking for the ultimate generality that underlies the notion of what data simulation what models and what is the holistic view of data simulation. That is your purpose.

(Refer Slide Time: 41:01)

**EXAMPLES: DETERMINISTIC, DYNAMIC, CONTINUOUS TIME, INFINITE DIMENSIONAL MODELS**

---

- a) ODE:  $x(t) \in \mathbb{R}^n$  called state,  $t \geq 0$  time,  $\alpha \in \mathbb{R}^p$  - parameter  
Linear:  $\dot{x}(t) = Ax(t)$ ,  $A \in \mathbb{R}^{n \times n}$ ,  $x_0$  - I.C  
NonLinear:  $\dot{x}(t) = f(x, \alpha)$ ,  $f: \mathbb{R}^n \times \mathbb{R}^p \rightarrow \mathbb{R}^n$ ,  $x_0$  - I.C,  $\alpha$  - parameters
- b) PDE:  $u = u(t, x)$ ,  $\varphi = \varphi(x, y)$   
Linear:  $\frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} = -f(x, y)$  - Poisson's equation with suitable B.C  
NonLinear:  $\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0$  - Burgers equation with suitable initial and boundary condition
- c) Integral equations  $f(x) = \int_0^T K(x, t)m(t)dt$  -  $f$  and  $K$  are given and find  $m$

Examples; deterministic dynamic continuous time infinite dimensional models ODE  $x$   $t$ . So, I am now going to introduce some notations.  $X$  is a state of a system.  $T$  is the

continuous time.  $\mathbb{R}^n$  is the set of all vectors of size  $n$ . When I say  $x, t$  belongs to  $\mathbb{R}^n \times \mathbb{R}$  consists of  $n$  components. Each one of them are functions of time.  $x$  is considered to be the state of the system.

For example, in a meteorological setup, what is the current temperature pressure in downtown Bangalore? Say, at midnight, midday 8 o'clock. So, what is the state? How do I describe the state of an atmospheric system? Temperature, pressure, humidity, wind speed, all these constitute different components of the vector, and that is called the state of the system. The model also can have a set of parameters  $\alpha$ . There could be a number of them. We will use  $n$  to be the size of the vector which represents the state.  $p$  to be the size of the vector that represents the parameter. We will use  $\alpha$  the Greek letter for the parameter. The  $x$  as our state.

So, what is your linear model  $\dot{x} = Ax$ . So, this is a differential equation  $\dot{x} = Ax$  by  $dt$  the rate of change of the system.  $A$  is a matrix.  $x$  is a vector.  $\dot{x}$  is a vector. Matrices real matrix of size  $n$  by  $n$ , it needs an initial condition  $x(0)$ . So, that is the ODE. Now PDE  $u_t$  is the time variable  $x$  is the space variable in the case of ODE. There is no space unit time  $\phi$  is a function of  $x$  and  $y$ . That is essentially space no time. So,  $\phi$  is a function of space, but no time 2 dimensional.  $U$  is a 2 dimensional, but one is time another space a linear partial differential equation could be a Poisson's equation.

Double derivative of  $\phi$  with respect to  $x$  the second derivative of  $\phi$  with respect to  $y$  is equal to minus  $\phi$ ,  $\phi_{xx} + \phi_{yy} = -\phi$ . It can be it has to be solved under certain kind of boundary conditions. A standard non-linear model is  $u_t + u \frac{\partial u}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2}$ , the time derivative  $x$  plus  $u$  times the space derivative  $x$  must be 0. That is called the Burgers equation with a suitable initial or boundary condition, (Refer Time: 43:44) I think there is a spelling error it is got burgers, b u r g e r. Then there is an integral equation  $f(t) = f(0) + \int_0^t k(t,s)f(s)ds$ . Here  $f$  and  $f$  and  $k$  are given our job is to find the  $m$ . So, given  $k$ ,  $k$  is called the kernel  $f$  is called in the forcing, my job is to be able to find  $m$ . So, you can see models of deterministic dynamic continuous time infinite dimensional can occur either as an ODE or an PDE or an integral equation.

(Refer Slide Time: 44:23)

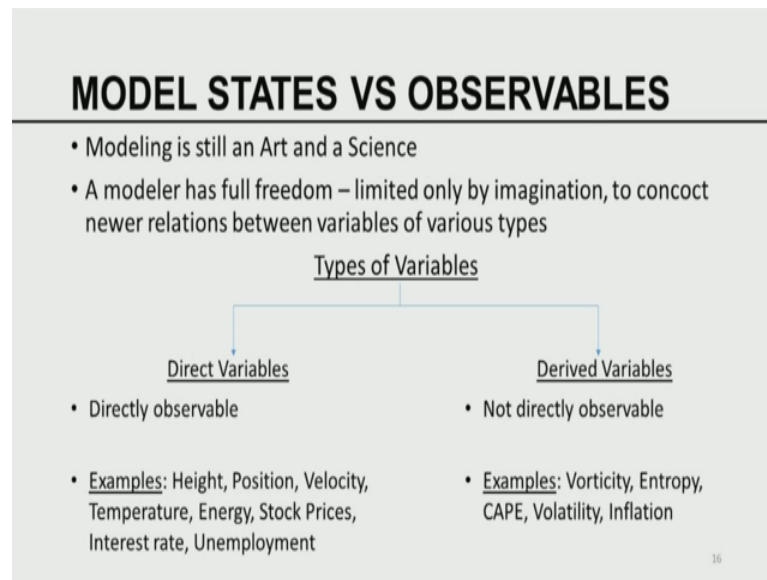
### EXAMPLES: DETERMINISTIC, DYNAMIC, DISCRETE TIME, FINITE DIMENSIONAL MODELS

- Replacing the derivative with a suitable discrete approximation and integrals by sum, we get a variety of discrete models
- Linear:  $\begin{cases} Ax = b, A \in R^{n \times n}, x \in R^n, b \in R^n \\ x_{k+1} = Mx_k, M \in R^{n \times n}, x_0 - I.C \end{cases}$
- Nonlinear:  $x_{k+1} = M(x_k, \alpha), M: R^n \times R^p \rightarrow R^n, x_0 - I.C, \alpha - \text{parameters}$

15

Examples are deterministic dynamic discrete time infinite dimensional models. Replacing the time derivative with suitable discrete time approximation, I can express differential equation into a difference equation. By replacing an integral by a sum in the standard way that we do in numerical integration, we can get a variety of discrete time models. The integral equation gives rise to a linear equation  $Ax = b$ .  $A$  is known,  $b$  is known, my job is to find  $x$ . That is a static model. The dynamic model is  $x_{k+1} = Mx_k$ . There is a linear discrete time model,  $x_0$  is the initial condition. A non-linear model  $x_{k+1} = M(x_k, \alpha)$  as a parameter. In this case I have initial condition I have parameters. So, are now giving life to a representation for good different kinds of model that we talked about are in general.

(Refer Slide Time: 45:26)



Now, I would like to bring the distinction between model state versus observables are related to quantities that I can observe directly for example, vorticity, but I cannot measure vorticity directly has to be measured indirectly volatility the price changes that is called volatility. Volatility is a quantity that affects us, but I cannot measure it directly. So, there are observables, but there are states not every state as is observable. For example, if the observable is pressure I can directly measure pressure if the observable I am sorry if the state is a vorticity I cannot measure vorticity directly I have to measure indirectly.

So, I would like to be able to relate the states of the model and what is being observed and relate them. So, I would like to be able to bring the distinction between the 2 modeling in spite of all the efforts that have gone on for centuries is still an art as well as a science. A modeler has full freedom, limited only by his her imagination. To concoct newer relation between variables of various types. So, the novelty in modeling relates the ability limited only by the imagination, to be able to describe various relation that are underlying between quantities that describe the system.

So, here we can discuss the variables into 2 types direct variables derived variables direct variables are directly observable pressure temperature derived variables the model needs a direct variable, but it is not directly observable examples height, position, velocity, temperature. Energy stock prices interest rate unemployment these are all direct

variables, which are directly observable vorticity entropy cape volatility inflation some of you may come from meteorological geophysical sciences, but my interest comes from applied mathematics to me. Analysis of time series models in the financial setting is no different from analysis of time series models to analyze climate data underlying mathematics are said, that is why I am trying to give you a broader explanation of variables.

So, vorticity entropy cape, this is the a convective available potentiality which is very important in severe storms. Volatility is very fundamental to many things we do in life. For example, the current delusion in in madras, in Chennai is because of a low (Refer Time: 48:23) they that came and sat in on the top of madras did not move at all and dumped. It is a very rare event. So, if you look at the frequency of these events, if you look at the intensity of these events, this event were a high volatility. It is the effect was way too much same thing inflation.

Inflation was very high when I went to United States in the late 70's we had 14 or 15 percent interest rate. Today is close to 0. So, inflation in those days was very high, inflation that is virtually none. Inflation affects everybody, but inflation is something we cannot be measured it has been inferred. So, inflation is a derived variable. Volatilities is a derived variable, entropy is a derived variable. But my model may need these derived variables to be able to make analysis.

(Refer Slide Time: 49:23)

## RELATION BETWEEN STATE VARIABLES AND OBSERVABLES

<u>State variables</u>	<u>Observables</u>
Dynamics of sea surface temperature in equatorial pacific	Radiated thermal energy measured by the satellite
Vorticity dynamics	Prevailing wind field
Total water content in a cloud	Radar reflectivity
Speed of a car by cruise control	Voltage generated which is proportional to speed

So, here I am going to talk about the relation between state variables and observables. Dynamics of sea surface temperature in equatorial Pacific is a variable. This derived variable is of fundamental interest in predicting El Niño. As we all know we are in a very severe grip we are under the very severe grip with a linear very strong. They say the high temp they the average temperature is more than 3 to 4 very close to 4 degrees. And it is affecting weather in different parts of the world. Vorticity dynamics is another state variable for vorticity equation.

The total water content in a cloud system that is very much you needed in cloud physics. Speed of a car in a cruise control today we are talking about driverless car Google is developing driverless car. So, I need to be able to measure the speed of a car very accurately to be able to feed back to the control element. All avionics a flight starts from Bangalore airport at 2 AM goes to London in 8 hours. You cannot see anything, everything is automated. So, the autopilot has to be able to adjust the speed based on various measurements.

So, a plane has tons and tons of observables relating to the situation, where it flies the models of the dynamics is already programmed to the autopilot. So, the autopilot model varies the various control devices based on the observations. Is that a headwind? Is that a tailwind? So, on and so forth. So, to be able to control I need to understand the state variables, it could be sea surface temperature it could be vorticity dynamics it could be total water content, it could be speed of a car in a cruise, control observables.

How do we measure the equatorial Pacific temperature thousand miles west of Hawaii? Nobody can go it is very difficult to develop a network of buoy systems. So, we have to measure the temperature essentially from satellites. Satellite measure only the thermal energy that was radiated in the infrared domain. So, we have to estimate the temperature based on the thermal energy received by the satellite and invert using the very well-known radiation physics laws, Stefan's law, Rayleigh Planck's law and so on.

What is the dynamics would use the observables are the prevailing wind? The  $u$   $v$   $w$  component from which you have to essentially compute the vorticity. The total water content in a cloud, there is no way to directly measure it is measured through the reflectivity of the radar. The raindrops are hanging in the cloud once you send the beam the radar beam it gets reflected by the water droplets.



So, the intensity of the returned reflected beam is displayed in various colors, and we can estimate the amount of water that inherent in the cloud by looking at the reflected energy. So, that is the obser reflectivity is observable, but is related to the state variable called total water content. Voltage generated which is proportional to the speed, and that is what cruise control uses. You sit the speed for the 70, if this because the road condition with the speed comes to 68 the accelerator knows that there is a difference it pushes the car forward.

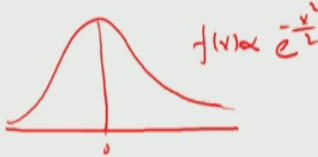
If the road is smooth if you have set 70, and the current setting it may go 72 it shuts off a little bit to be able to control. So, I need to be able to measure the speed, to be able to be able to do good cruise control. With the interest in driverless cars this notion of model's observation data simulation becomes very fundamental to replace a driver by a computer. Because we need to know the dynamics of the model, we need to know a different road conditions. So, a car is fitted with lots of observing devices measuring device.

You are not bored computer has to process all the devices almost instantaneously to be able to steer the car in the particular lane, and they have to be limited they have a (Refer Time: 53:59) success in doing this. It is a long way to go, but intellectually stimulating project where modeling observations bringing the model almost online to be able to use and to be able to steer. The decision there is to be able to steer. The decision meteorologist to be able to predict. And so, you can see the underlying aspects of data simulation again coming along from different directions.

(Refer Slide Time: 54:27)

## OBSERVATION NOISE

- Observation of physical variables – temperature pressure, wind, etc are subject to measurement error/noise
- Following Gauss (1777 – 1855), observation noise is modeled as white Gaussian noise with known covariance structure
- Economic Variables – stock prices, interest rate, foreign exchange rate are intrinsically random but are observable with no error



18

Now, I liked to little bit we talked a lot about a model I have not talked much about the observations. So, I am now going to talk a little bit about observation because is one of the 3 key players in the system.

Observations of physical variables temperature, pressure, wind are subjective measurement errors. These measurement errors are noise, are called noise. For example, what I may read as 2.5, you may say it is 2.3, one might say is 2.6. So, observational errors when humans measure certain things even instruments. If you look at any instrument they will say the following.

So, if you buy voltmeter, they will say it can read from 0 to 100, the accuracy is 10 percent. What does it mean? If it is say 60, it is plus or minus 10 percent of 60. And where the error comes in? The engineering aspect of the of the design of the instrument itself. So, that is one way the error can come in. The other way the error can come in is how do I calibrate. The meter shows 60, is it really 60? We have to calibrate a meter again standards. So, it could be design errors within or design differ there is no design error, there is a design differences, there are manufacturing differences, there is calibration differences.

Then the human being able to read these, all these together lumped into one phase called measurement noise or the measurement error. Following goes one of the very well-known mathematicians of all time who lived between 1777 and 1855. He was the first

one to be able to analyze the properties of observation noise. In fact, in his time, the only kinds of data that that was available was measurements about very many celestial objects made by humans using very simple telescopes. That is all that they had, they had nothing else.

So, he was to be able to make sense out of these observations, and these observations were not consistent. These observation contained a lot of errors. So, by analyzing a given set of observation, he modelled the statistical properties of the observation errors, and developed the so-called bell curve. So, what did he show he showed the following. If this is 0, the errors from a bell-shaped curve. On an average error is 0, but on a given circumstance the probability of an error could be either positive or negative it took a bell-shaped curve.

This is what is called the gaussian curve, and we all know this curve is given by  $f$  of  $x$  is proportional to  $e$  to the power of minus  $x$  square by 2. There is a purpur constant of proportionality that comes into play. So, this curve is a bell-shaped curve. What is one of the most of fundamental contributions of gas is to be able to establish that observation noise essentially followed this bell-shaped curve, and we use this to this date. And that is a very enduring aspect of gaussses discovering. So, this kind of noise, it is also called the white noise. What does it mean? I measure the temperature today, I measure the temperature tomorrow, I measure the temperature day after tomorrow by the same instrument.

The error today is not correlate with the arrow tomorrow, the error is not correlate with a with the error day after tomorrow. So, there is a there is no correlation when sequence are not correlated there is a particular way to characterize this. There is called white gaussian noise, gaussian refers to the bell-shaped curve. Noise is the error, white means they are uncorrelated. White noise means errors are uncorrelated. These noise are also have an known correlation structure. For example, if you buy a voltmeter they say a 10 percent accuracy, 15 percent accuracy, 20 percent accuracy. You can measure you calibrate and you can compute the error.

So, observation noise has to be associated with the error properties, the error properties have to be related to the known covariance structures of the errors. In economics stock prices, the interest rate, foreign exchange rate, they are all intrinsically random. Pressure,

temperature, there is a variability from night and day, that comes because of the phases of the day and night, the phases of the moon, summer, winter, autumn, spring. So, these are all the variations induced by season.

Variations is induced by the day and night. December 31st, the maximum temperature in downtown Bangalore, this year, next year or the past 100 years. If you if you part them there is a natural variability there. So, things could have natural variability. Things could have underlying stochastic properties. So, when we talk about random, the randomness can come from either from the noise or from intrinsically random variations. We will stop here at this moment, we will continue these topics in our next lecture.

Thank you.