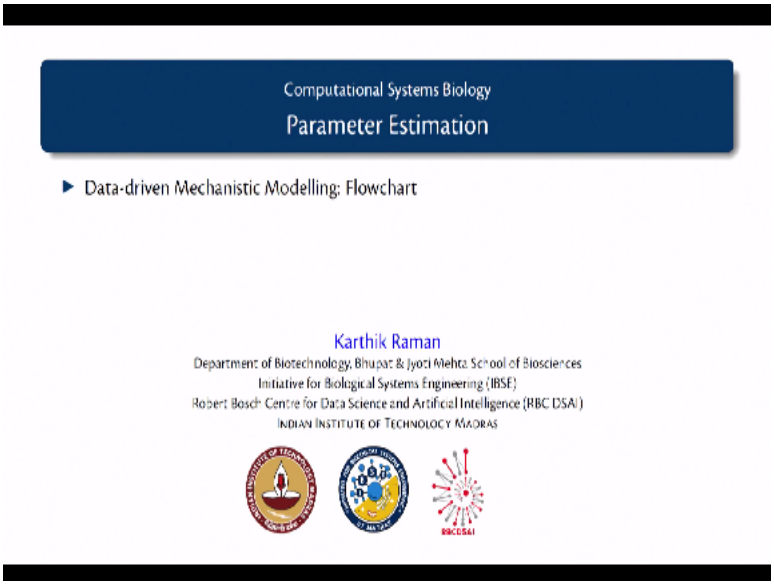
Computational Systems Biology
Karthik Raman
Department of Biotechnology
Indian Institute of Technology - Madras

Lecture – 42
Parameter Estimation

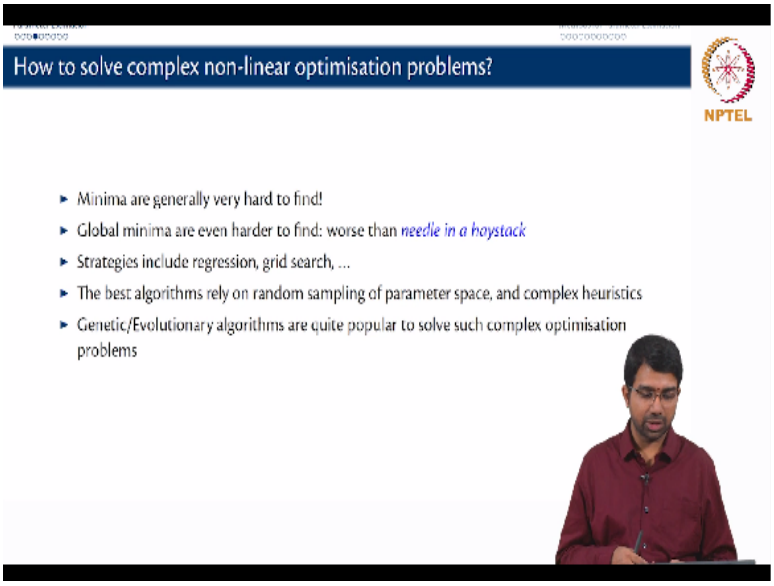
(Refer Slide Time: 00:11)



In this video, we will look at how we carry out data driven mechanistic modelling. So given some data on, you know, time course of various species involved in a biochemical reaction network and so on. How do you build a mechanistic model may be based on mass action? Michaelis-Menten, or Hill and so on.

To accurately predict the time, course of various species involved and, you know, understand what is happening in the biological network. So we will look at the various aspects in terms of how we pay process the data, write till how you diagnose the parameter estimates and the fits and so on.

(Refer Slide Time: 00:46)



How do you solve such complex non-linear optimization problems? The minima are generally very hard to find because it is in a very very complex place. Forget finding global minima is worse than a needle in a hay stack, right and then are many strategies. You can do a grid search, we can do various kind of regression and so on. **"Professor - student conversation starts"** Gradients. Gradients descents are difficult because you do not want to be completing the gradient of that function.

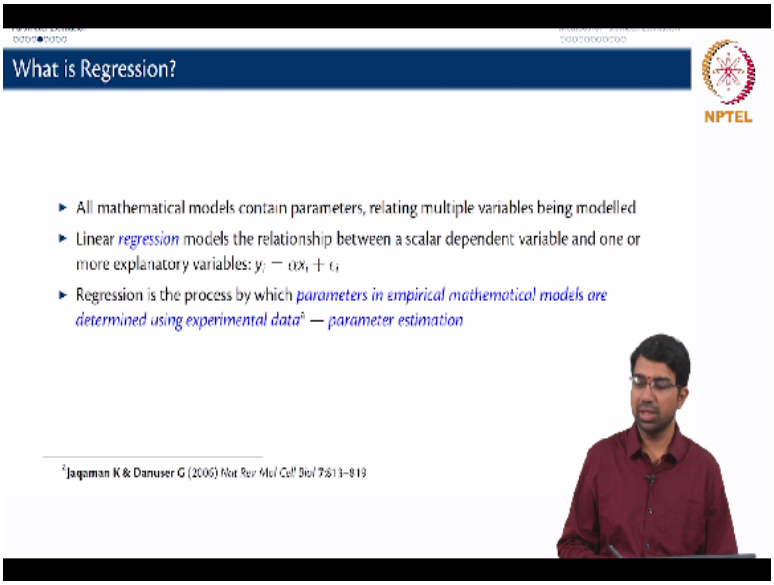
It is very expensive to compute the gradient of that function. **"Professor - student conversation ends."** So gradient descents are again not useful. For the best algorithms rely on random sampling of parameter space and complex heuristics and these are usually grouped under what are known as direct search algorithms. You basically directly search in the space. You do not use any fancy gradient computations, any other logic. You just directly search in the space. Of these, the most popular are genetic and evolutionary algorithms.

You may be study that in one of the coming classes. Grids are just basically, you know, it is how you search for something that you have lost, right. You basically grid the place up and look at every, so it is essentially searching in every place of the parameter space, every section of the parameter space. **"Professor - student conversation starts"** But instead of grid search, sampling will be better thing.

Yes, yes, so you basically divide it into grids and do a sampling. So we try to study some sampling algorithms. So you essentially divide the parameter space into grids and you sample uniformly across the grids. There is something called Latin Hypercube Sampling which does exactly that. (()) (02:20) Yes, again, the noise models are interesting in a way so you can have Gaussian noise, you can have (()) (02:31) noise.

It really depends, right on the system that is being studied. Whether it is able to capture the noise characteristics of the system in question. **"Professor - student conversation ends."**

(Refer Slide Time: 01:47)



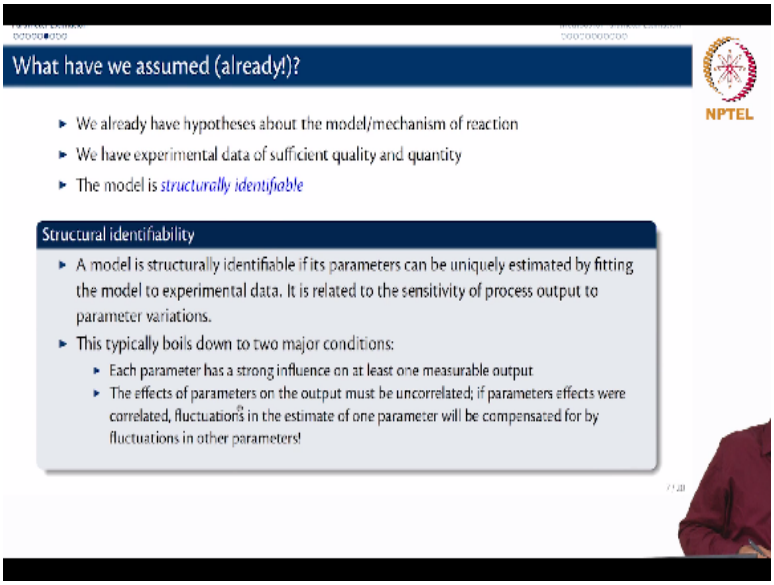
So what is regression of parameter estimation? We know that all mathematical models contain parameters which correct the variables that are being modelled. All of you are very familiar with this classic linear regression, right. So y is alpha a*xi+some epsilon a residual, right and you basically minimize the residuals, right. So you do say sigma yi-alpha xi whole squared and you reduce it and you find the best alpha, right.

You may also have an intercept if you want and so on. So regression is the process by which parameters and these kinds of empirical mathematical models are determined using experimental data, this is what we normally call data driven mechanistic modelling, right that is another common phrase that you will see to describe such models. It is data driven because you are assuming that you have your Michaelis-Menten or whatever, some time course data and you also

know the mechanisms, right.

You do not just make a blackbox model. Just say some alpha x1+beta x2 that kind of thing, right. You know mechanism, you at least propose mass action equations for that or Hill equations or Michaelis-Menten equations for those exact mechanisms that you have, you assume exist and then you build a model and now you have to fit all the parameters in that model back to your, to see if you can reproduce, regenerate your data.

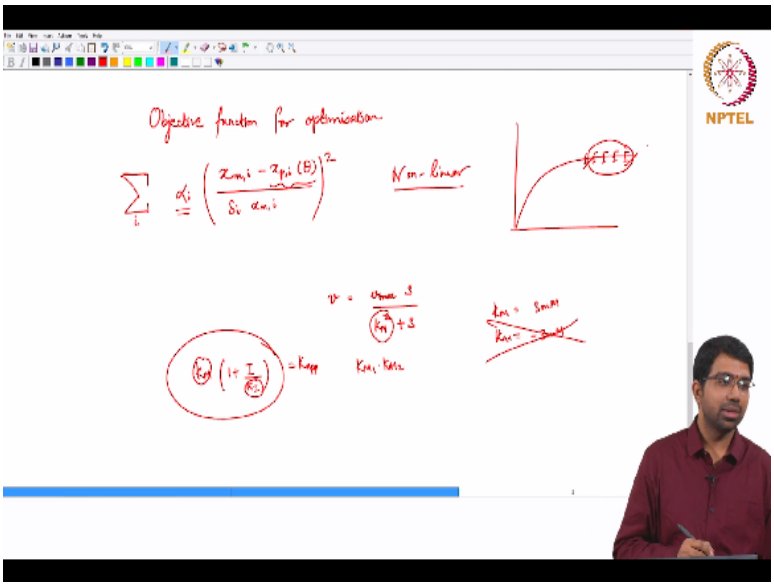
(Refer Slide Time: 04:03)



What have we assumed already, right? First thing is, so these are all surprises for most other engineers, right. If you look at chemical engineer or electrical engineer, this is a bit of a surprise for them because they usually do not assume a model ahead of time, right or a mechanism. So in our case, you have A giving B giving C giving D. I am assuming all those are enzymatic reactions.

I am already starting off with vm km and all of that for each of those reactions. So we start with 4 reactions, 8 parameters and some known mechanisms and very tricky in biology. We assume that we have experimental data of sufficient quality, quantity, reliability, precision, whatever and interestingly we also have assumed that the model is structurally identifiable. What does that mean? There are courses that you can do on identifiability and so on but briefly for our purpose, we will say a model is identifiable if its parameters can be uniquely estimated.

(Refer Slide Time: 05:08)



You already have a problem here; I can give you a very simplistic example. Suppose you have v is v max*s/km squared+s. There is no way you can identify unique parameter set here because km=3 millimolar, might give you the same behaviour as km=-3 millimolar. You probably know that this does not make sense, so you can eliminate it but you have a problem here, right. So this is a very explicit case.

Usually it will be buried so hard that you will never be able to find why your model is not identifiable? What are your issues? So that is another major problem in biology. It is also more than a problem you can pick up it as a feature of biological data because there is robustness in biology, okay. It is not so tightly fixated at one point. At anywhere you move, the system breaks, right.

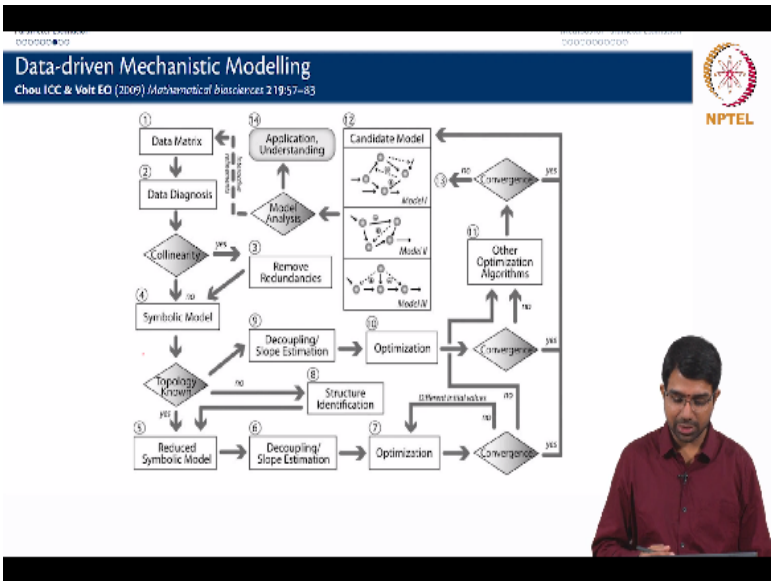
The system has some latitude. It has some flexibility to move around in a particular domain recreating its behaviour. This becomes a very interesting point. We will come back to it. So a model is structurally identifiable if its parameters can be uniquely estimated by fitting the model to experimental data and this is somewhat related to the sensitivity of the output to the parameter variations.

Typically, it means 2 things. Every parameter at least affects say one concentration and the

affects of parameter should also be uncorrelated. If parameter effects were correlated, you can always have high x low y, I mean high k1 low k2 will give you the same behaviour as low k1 high k2 and you can never, so it is like saying, you know, other way to look at it is, if this was km1*km2.

"Professor - student conversation starts" It is only the product... May be the product only matters, right. So those issues will be there. **"Professor - student conversation ends."** In fact, you will see that very commonly. So there will be km*1+i/ki, right for your reactions involving inhibition, right. So maybe this is actually called k apparent. This k apparent is might be the most important component rather than km or ki. So these kinds of issues can potentially be there.

(Refer Slide Time: 07:30)



So this is a very nice flow chart for mathematical modelling, right and it will be nice if everybody followed this flow chart in its entirety. But usually we tend to skimp on 1 or 2 of the steps at least. So let us say you start with a data metrics. First thing you need to do is data pre-processing. Very often, overlooked. Remove colinear points, points that tell you the same thing. For example, one very bad measurement you can imagine is, and so many measurements.

These are not informative. What are these points telling you? Nothing. You might as well have had just this point and this point, right. So these are things that you may want to get rid of. **"Professor - student conversation starts"** Sir but having them is not going to make. Well you,

you fakely assume that you have more data then you have to estimate your parameters, right. You, those are not data points as such. I could, it is as good as having 2 points.

Instead you assume that you have 8 points or 10 points and then you build a model. But... Do you know the topology, yes. Last year's (()) (08:47) there are ways to check. You can systematically perform some diagnoses on these kinds of models. **"Professor - student conversation ends."**

So then do you have a, you have a model, right and do you know the topology, then you try to reduce the model, maybe we will just build this to building a model and then you try to decouple, see if there is anything that can be decoupled. Slope estimation is another method to speed up parameter estimation but basically let us just assume points 4 5 6 as the normal modeling exercise that we do, right and then optimization.

Are you able to fetch the data by sticking your parameter set into the model? Does your parameter set help you regenerate the original data? Convergence grid, then you have a candidate model. If no, you have to fiddle around, right. May be there are issues. This is the place where it gets tricky. When will you know that your model is not good versus the parameter estimation algorithm is no good?

The very open question. There is no clear answer to this, right. So you may have multiple, so the way we go about it usually is, we will start with a parameter space. Let us for simplicity sake stick to Michaelis-Menten which means that I will search on the 2 dimensional grid. In a 2-dimensional space, I will explore to see is that any point that gives me a good fit to my data. If nothing, I do not know, right.

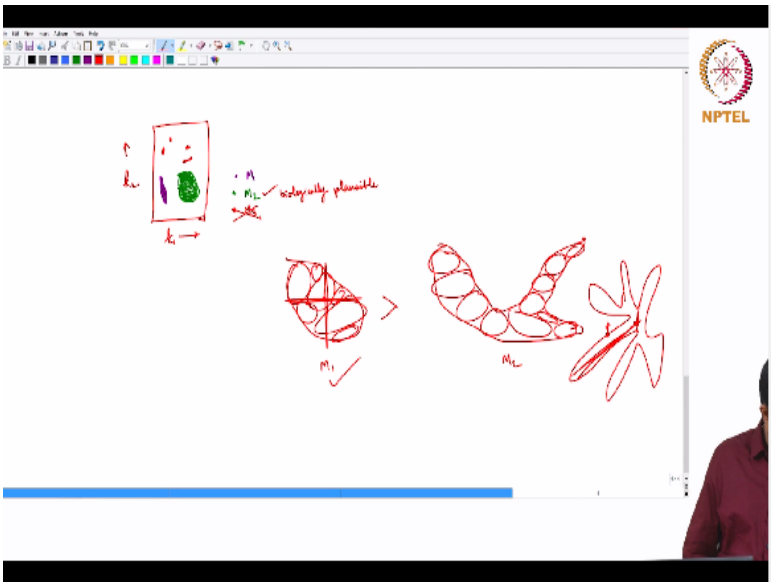
There are 2 issues, either you did not search hard enough. May be your domain of search was not good enough. You were looking in the wrong place or there exists no such point. May be in 2 dimensions, it is easy to say that but in higher dimensions, it becomes very very tricky to say that. So when do you say that, so when do you throw your hands up and say hey, I think my model is wrong.

I need to revisit the reactions I have considered. I need to revisit the topology of my model itself. Do you understand what I mean by topology? The connectivity between different species, okay. This A giving B, B giving C, C giving D. Does B gives C or B gives C+D or may be it does not make so much sense in a metabolic system but in a signalling system, this makes a lot of sense. Does something get phosphorylated once or twice or thrice and so on.

Is there a feedback loop or not because one behaviour will happen if there is the feedback loop? Another behaviour will happen if there is no feedback loop. So how do you handle this? And then hopefully the final part where we analyze the model and develop our understanding. This is the goal of modelling, right, final goal of modelling and the other very important thing is, this is the cycle, right.

You then try to see if you can refine the model. May be the model is off by a little. Can we tweak something in the model or generate new data, perform the one experiment, the one most informative experiment that you needed, right and try to improve on this. This is again a very tricky part. So if you see the topology of these 3 models vary. How do you pick the best? Given a model, given a topology, how do you say is this model good or not?

(Refer Slide Time: 12:24)



This is again an open question but one way to go about it is, you can look at the, let us just

consider, this is k1, this is k2 and now let me say this is for model 1, this is for model 2 and this is for model 3. Which do you think is the better model?

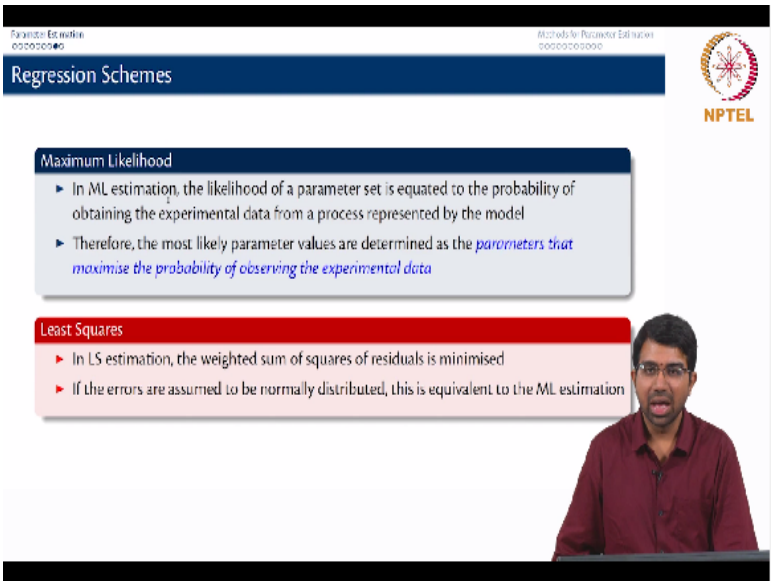
"Professor - student conversation starts" M3 is worst. M3 is worst, okay. Fair enough. The model is... We would normally say that M2 is more biologically plausible. (()) (13:45) because this is somewhat more robust. It admits more changes in parameters without dying, right without basically failing to reproduce the data, right. **"Professor - student conversation ends."**

So this can be a measure of but this is again a very simplified 2 parameter system. How would you evaluate it in many parameters. Here it is very easy. It is just area. In higher dimensions, you are going to talk about a very complex volume. So there are many ways. There is in fact some people who fitted ellipse size in that space and so on. I will share some of those papers. So you basically, so in a 2 dimensional, let us say, let us say this is model 1 and this is model 2, right.

How do you quantify it, right? So you may want to fit some shapes inside this and estimate the area. So now if you, if this area is greater than this area, then you say this is the better model and vice versa. That could be one way to look at it but you can come up with many more ways, exactly. So if you had a model like this, because what would this mean, what does this mean practically?

There are several points where the system is unstable. You move in this direction, you are off. The system does not exist. Whereas here, it is very nice. There is a point which is superstable. You have this much leeway to move, right. Whereas here, there are some directions where you can move nicely, some directions where you cannot move. There are practical issues and in higher dimensions, this becomes really really tricky but these are somethings that one would want to think about.

(Refer Slide Time: 15:50)

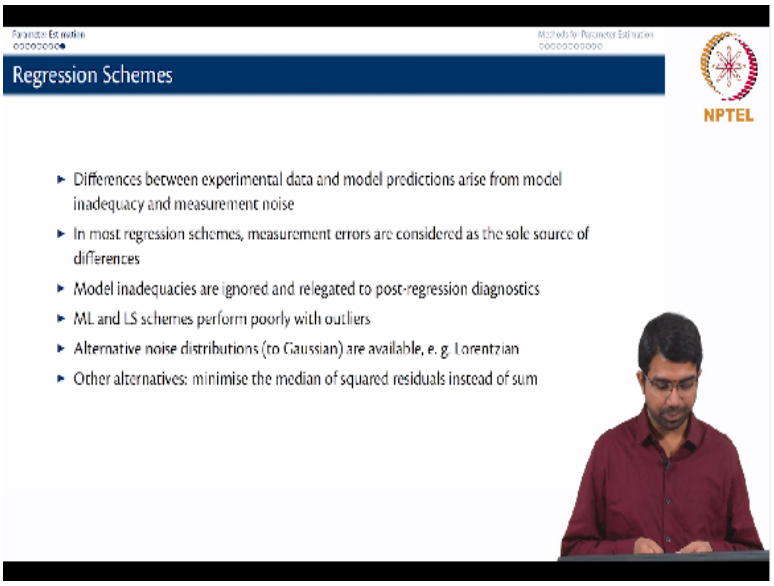


So we were looking at regression schemes. So there are 2 types of regression schemes. One will be maximum likelihood and of course you all know least squares. So in maximum likelihood estimation, how we estimate is we try to maximize the likelihood of a parameter set. So you want to find that parameter set that has the maximum likelihood. So the likelihood of a parameter set is equated to what is the probability of getting the experimental data if these were the parameter set for the model, right.

So what is the probability of observing your data set given your model and given a parameter set, right. So you then find the most likely parameter values as the parameter that maximize the probability of observing the data. Residentially a probability value for, what are the probability of observing the data given your parameter set and you want to basically maximize it but then the parameter set which maximizes it is your best parameter set essentially, right.

In least square estimation, we basically weight the sum square of residuals and if the errors are assumed to be normally distributed, this is basically equivalent to maximum likelihood estimation. If you do the math, it comes out as equivalent to maximum likelihood estimation, right. I will not go into the math because you know you can, it is easy to look this up, this is like in classic textbooks and so on.

(Refer Slide Time: 17:31)

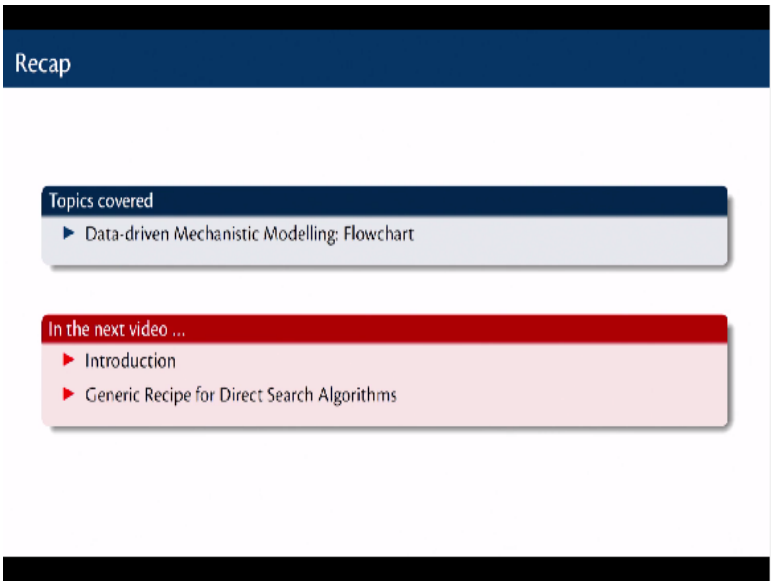


So one thing that we will look for is, we assume that the in most of the errors that we have in prediction, stem from basically the measurement error itself, right. So there are 2 potential sources. So one is model inadequacy. The other one is measurement noise. You basically assume that model is adequate, right and we will relegate that to a little later. We first assume that the model is good enough and then we look at what are your residuals, how bad is your, is the fate and so on.

So in most regression schemes, the measurement errors are considered as the sole source of the difference between your data and your model predictions and model inadequacies are kind of ignored and then relegate it to post-regression diagnostics, right. Once you perform the parameter estimation regression, you then do some diagnostics, right. One issue with both ML and least square schemes is that they both perform poorly with respect to outliers.

So one outlier can hurt it. So there are other alternatives possible. You may want to, you know, minimize the median of the residuals instead of the sum of the residuals but anyway you have weighted least square formulations and there are other noise distributions also that you can use. So all these things become very subjective, right and they very well depend upon the modeling task in hand and how you pick them and so on.

(Refer Slide Time: 19:13)



So I hope in this video, you got a good overview of how we go about mechanistic modelling given data and what are all the possible steps involved and how would you try and discriminate between good models and so on. In the next video, we will start looking at an introduction to stochastic search algorithms or direct search algorithms and particularly discuss what is the generic recipe underlying practically all direct search algorithms.