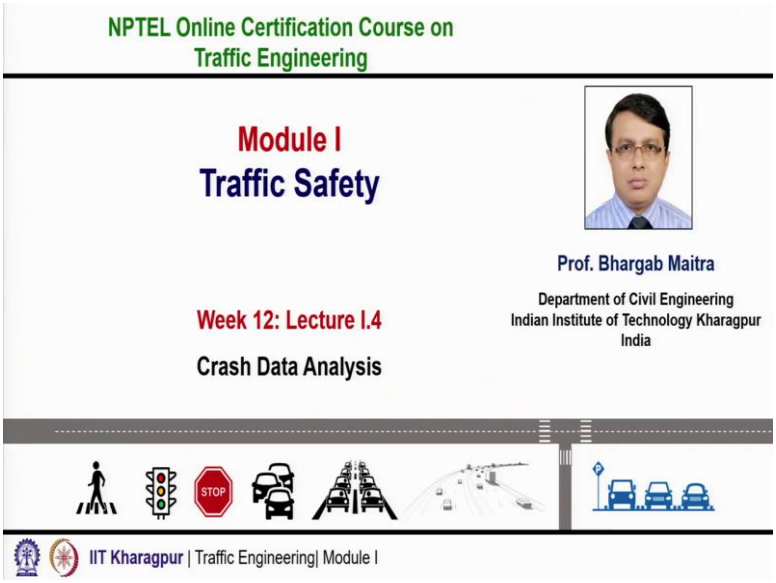
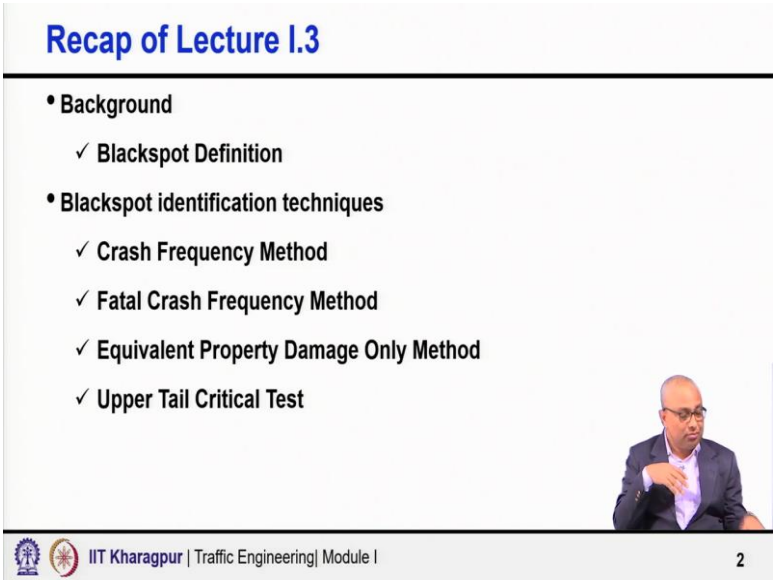
Traffic Engineering
Professor Bhargab Maitra
Department of Civil Engineering
Indian Institute of Technology Kharagpur
Lecture 62
Crash Data Analysis

(Refer Slide Time: 00:16)



Welcome to module I, lecture 4. In this lecture we shall discuss about crash data analysis.

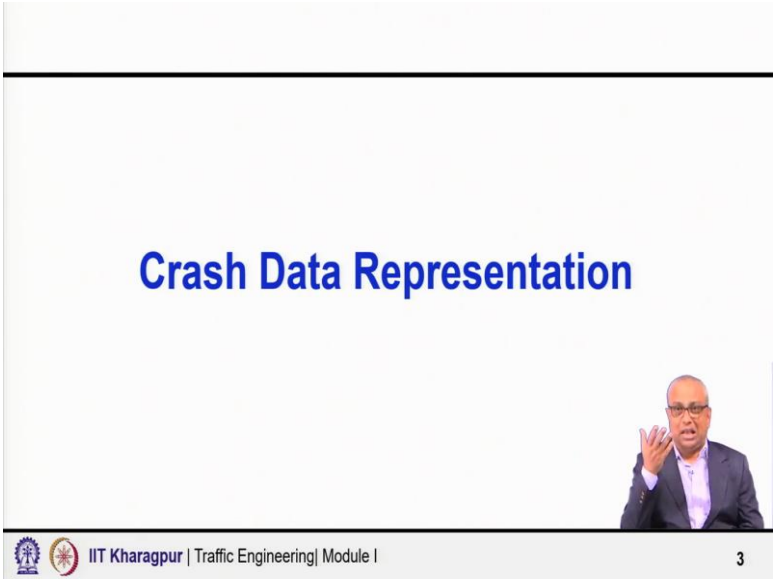
(Refer Slide Time: 00:22)



In lecture 3, we discussed about blackspots. How we can define black blackspot or what is really blackspot and then different techniques for identification of blackspot. For example, crash frequency method, then fatal crash frequency method, equivalent property damage only

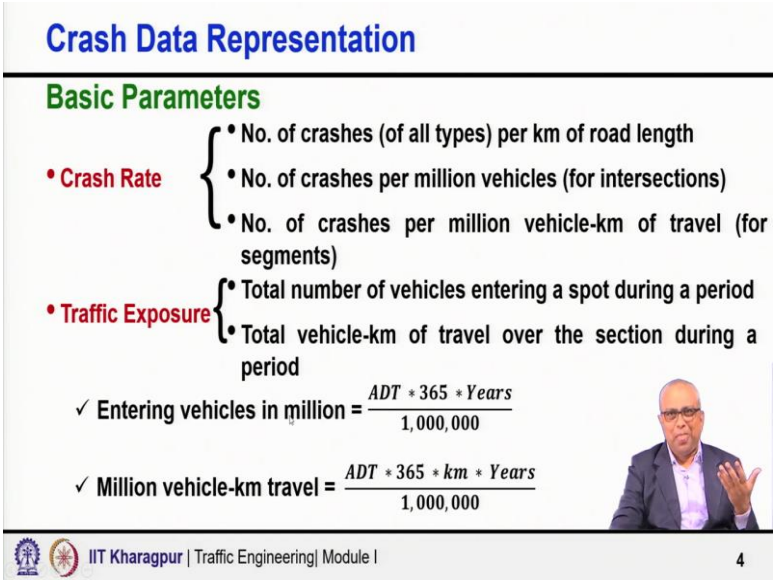
method, upper tail critical test and then with some example data I have also shown you how you can identify blackspot.

(Refer Slide Time: 01:00)



Now, the data is there. So, we want to do the analysis of the crash data.

(Refer Slide Time: 01:09)



So, first the crash data representation. We will represent crash data using a number of meaningful basic parameters. For example, crash rate a very important parameter it will indicate number of crashes of all types per kilometer length of road. So, you have so many hundreds and thousands of kilometers of road and you take different sections and then decide how many number of crashes are happening or have occurred per kilometer length of the road.

Second, it could also be expressed in terms of number of crashes per million vehicle typically for intersections, how many vehicles are using the intersections. So, per million vehicles, how many crashes are happening, that number of crashes per million vehicle kilometer of travel and typically for segments, this is a meaningful expression.

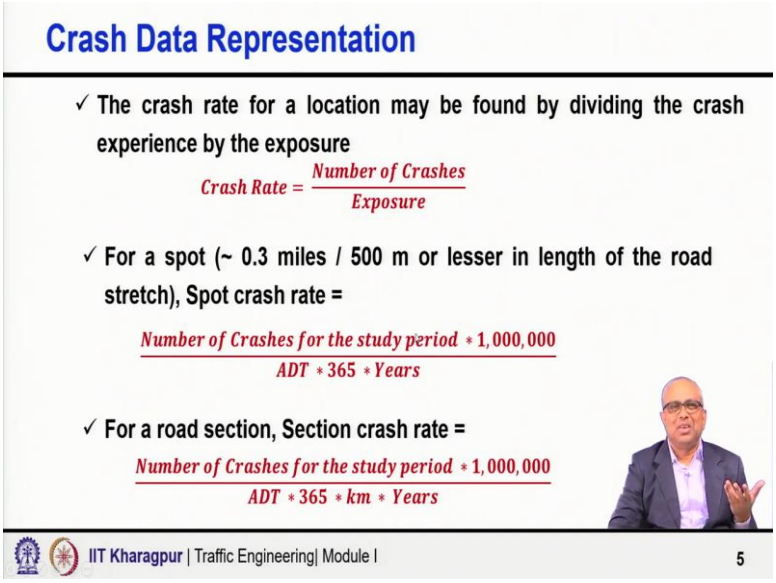
So, crash rate could also be expressed in terms of all these three possibilities, fist number of crash per kilometer length of road or per million vehicles or per million vehicle kilometers traveled. You have to take a suitable quantification as per the context given a context what is more meaningful.

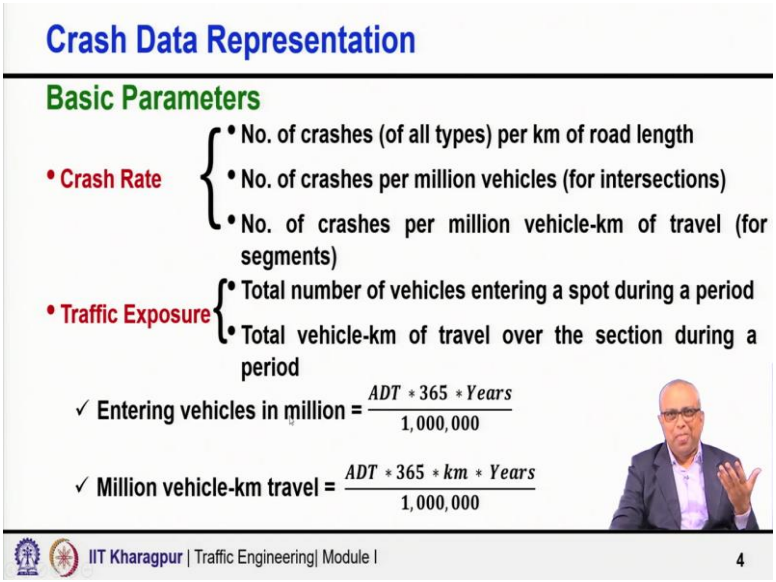
Second, next important basic parameter maybe the traffic exposure. I can express it in terms of total number of vehicles entering a spot during a period that period maybe 1 hour, 1 year, 2 year, 3 year. So, how many total number of vehicles are entering a spot during a given period? Also it could be expressed exposure in terms of total vehicle kilometer traveled over the section during a period, you have taken a section, then during a period considering that overall section, what is the total vehicle kilometer travel that has occurred.

So, if we are expressing traffic exposure in terms of total number of vehicles entering a spot during a period then it can be expressed like this entering vehicles in millions. So, you have ADT Average Daily Traffic or ADT Annual Average Daily Traffic multiplied by 365, that is the yearly traffic multiplied by the number of years we are considering divided by 1 million or 10 lakhs in Indian context. So, I have written it like million so, I have used the commas in that manner.

If I am going for the second one that means total vehicle kilometers traveled over the section during a period. So, I can again express it vehicle kilometer instead of that million vehicle kilometers traveled. So then again ADT or ADT multiplied by 365 into years into kilometer over the section. How much kilometres divided by 1 million that gives you the million vehicle kilometers traveled? So, that shows you how we can quantify the traffic exposure. Now, I will show you once you have quantified the traffic exposure, how you can calculate the crash rate.

(Refer Slide Time: 05:27)



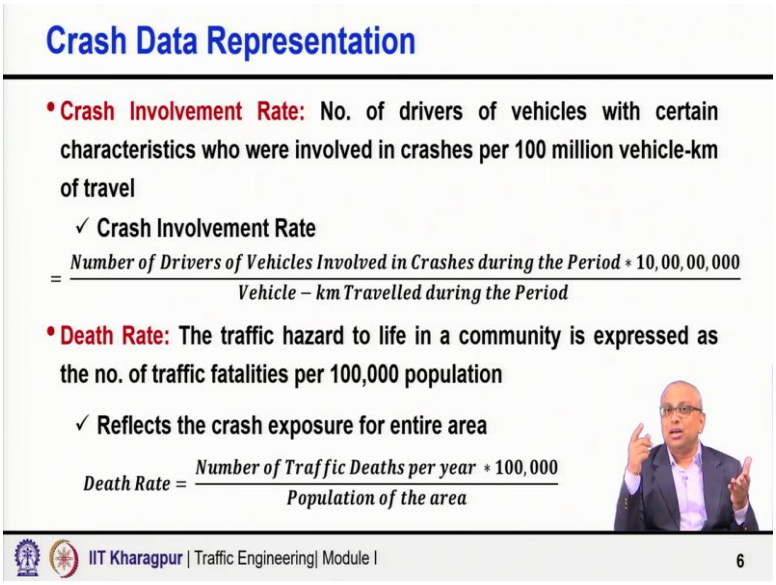


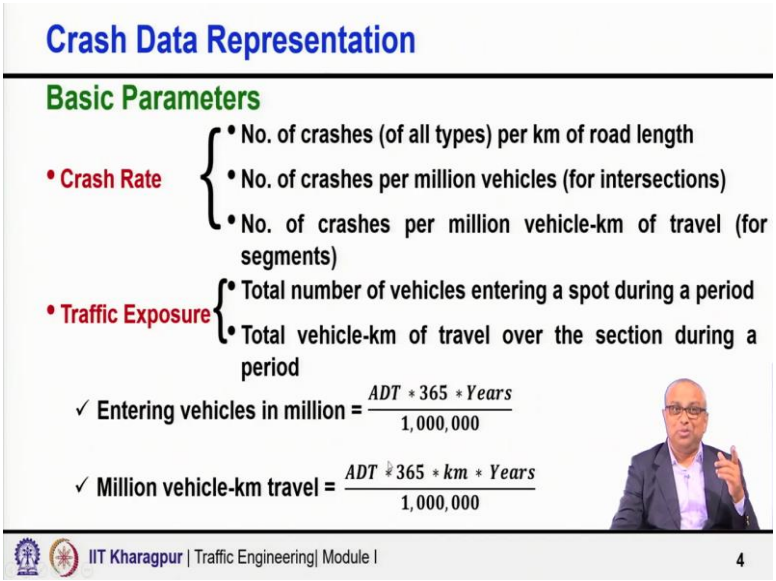
So, the crash rate for location may be found by dividing the crash experience by the exposure number of crashes divided by exposure because what is the rate, it is basically crash rate. So, number of crashes divided by traffic exposure. So, exposure we will use if we want suppose number of crash per kilometer, length of road then one way we will do number of crashes per million we can if we want we will do accordingly.

So, the but all cases basically the crash rate is the number of crashes divided by exposure. So, for a spot or maybe 500 meter or less in length of the road stage, the crash rate can be expressed as number of crashes for the study period, total divided by the exposure. Exposure is what? ADT into 365 into years divided by 1 million. So, 1 million will come here. So, that way you can express the spot crash rate.

For section typically as we said or segments it is million vehicle kilometer. So, how we can express? Number of crashes for the study period divided by ADT into 365 into kilometer into years divided by 1 million. So, here you will get section crash rate and in what vehicle crash number of crashes per million vehicle kilometer of travel. So, this is the expression and the first one rather the middle one is number of crashes per million vehicle that is all.

(Refer Slide Time: 07:43)





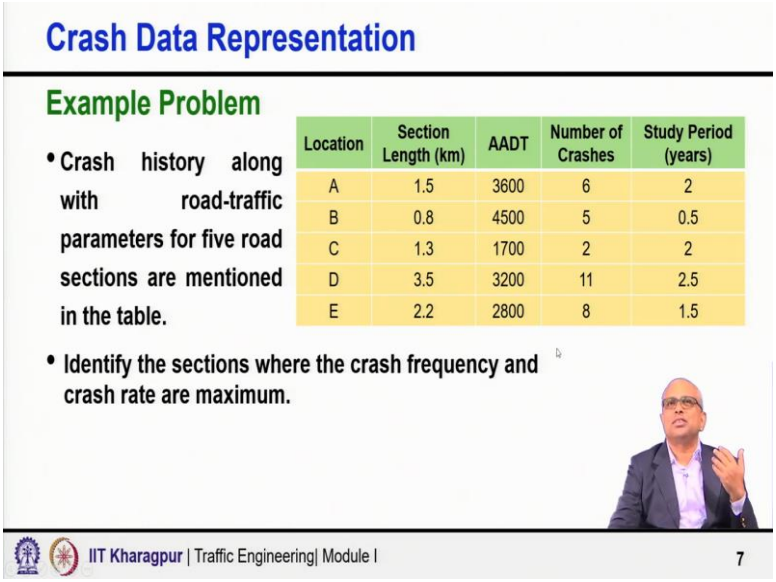
Now, the third meaningful parameter as we have said earlier crash rate, then traffic exposure, then third meaningful parameter is crash involvement rate. What is that? It is that number of drivers of vehicles with certain characteristics please note that we are seeing number of drivers of vehicles with certain characteristics who were involved in crashes per 100 million vehicle kilometer of travel. So, how to express crash involvement rate take that number of drivers of

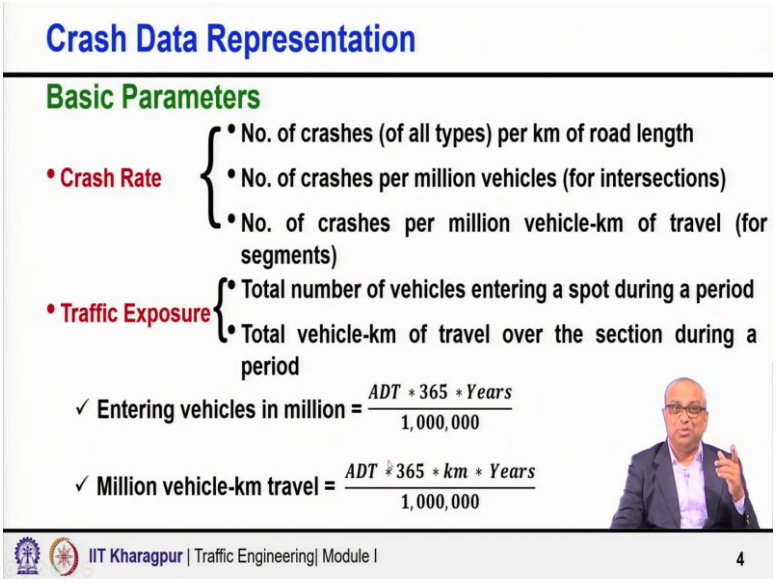
vehicles involved in crashes during the period divided by vehicle kilometer traveled during that period and since we said per 100 million vehicle kilometer, so divided by 100 million. So, that is what is written. So, that is what it is.

Now, the death rate, death rate is the traffic hazard to life in a community and is expressed as number of fatalities per 100,000 population. Per 100,000 population, how many fatalities is happening or has happened? How many fatalities is happening per 100,000 population or how many fatalities has happened per 100,000 population. So, it reflects the crash exposure for the entire area. People are living, so it is very important very meaningful visit.

So, how we can get it, death rate, number of traffic deaths per year into or divided by basically population of the area. But then we are expressing expressing it per 100,000 population. So, it is multiplied by 100,000 that is the death rate.

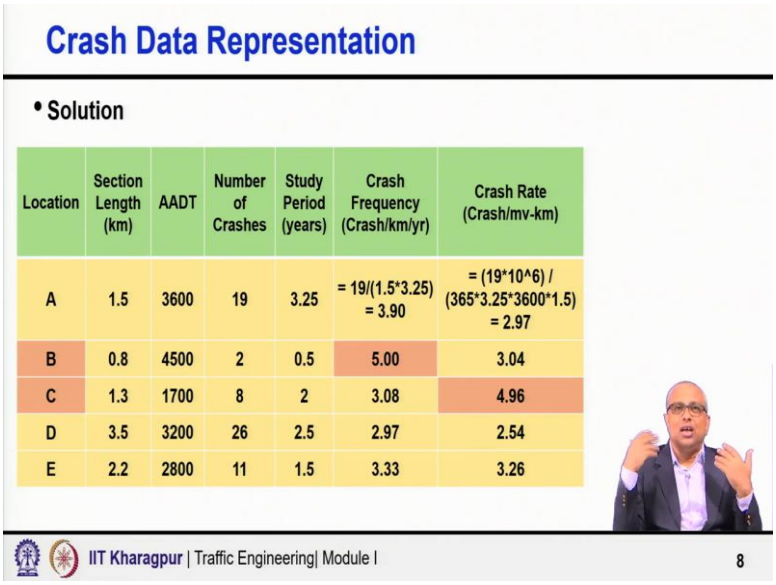
(Refer Slide Time: 10:13)





Let us take an example. Suppose you have five locations for short stage or segment whatever you say. Section links are given different section length, different ADT value, different number of crashes and different study periods. But then, how I compare? How I compare these are the meaningful parameters basic parameters, they can help us to do a meaningful comparison.

(Refer Slide Time: 10:50)



So, let us go ahead. Location is given, section length is given, ADT is given, number of crashes given, study period is also given. So, I am now calculating the crash frequency what is crash frequency number of crash per kilometer per year. So, you take the crash number of crashes 19 divided by per kilometer. So, divided by 1.5 per year 3.25 is the period, so divided by 3.25. So, that is what you get 3.9.

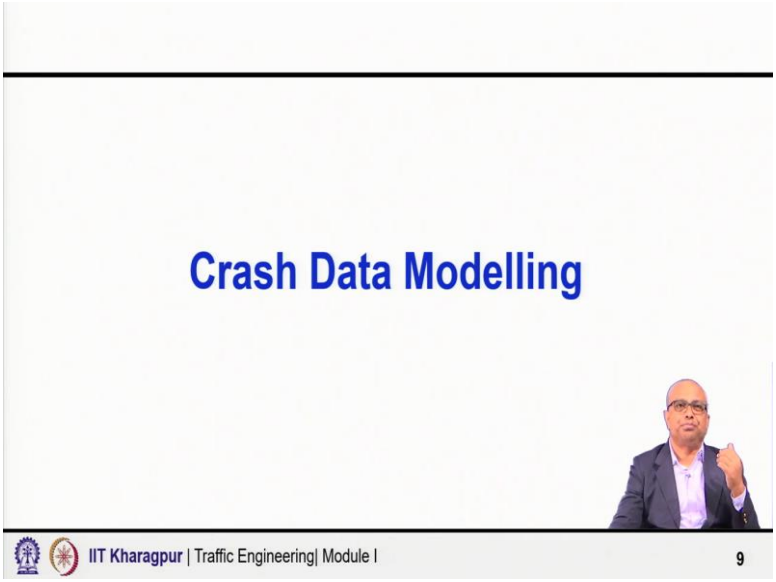
So, now, I have kind of normalized it expressed it in terms of crash per kilometer per year. So, like that the quantities you calculate and you find here say location B that has got the highest value of crash frequency. So, we can identify that segment or the location.

Similarly, I can also calculate the crash rate, crash per million vehicle kilometer traveled. So, 19 crash divided by million vehicle kilometers traveled 365 number of days in the year, 3.25 year multiplied by 3600 ADT into section length is 1.5. So, you calculate it 2.97, like that you calculate for all the sections and you will find here section C has got the highest crash rate 4.96.

So, like that we can identify the sections when sections have different length different ADT, different number of crashes, different study periods, but still we can compare them and different parameters like the crash frequency, while some section may be worst. Crash rate while some other section may be the worst, sometimes it may be the same section telling you that that should be your focus immediate focus.

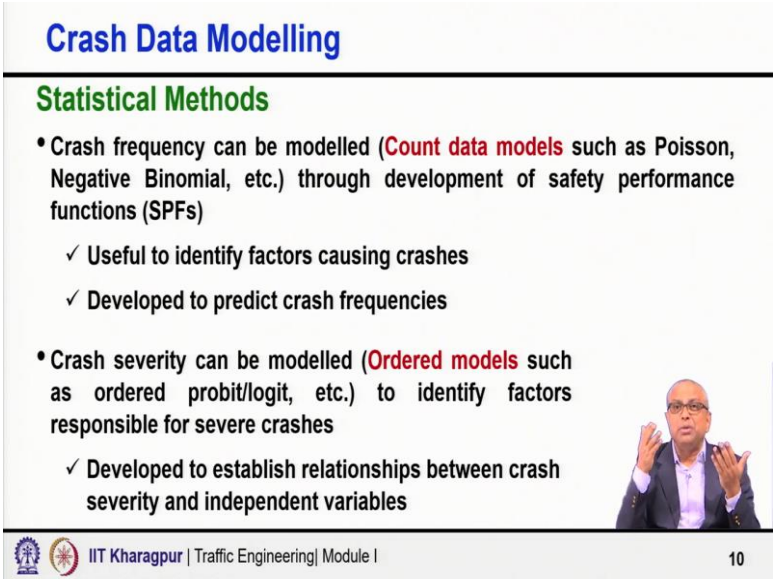
So, it helps us to compare different sections with different numbers of accidents or crashes, different study period length up the road traffic volume is different. So, we can still compare them in a meaningful manner.

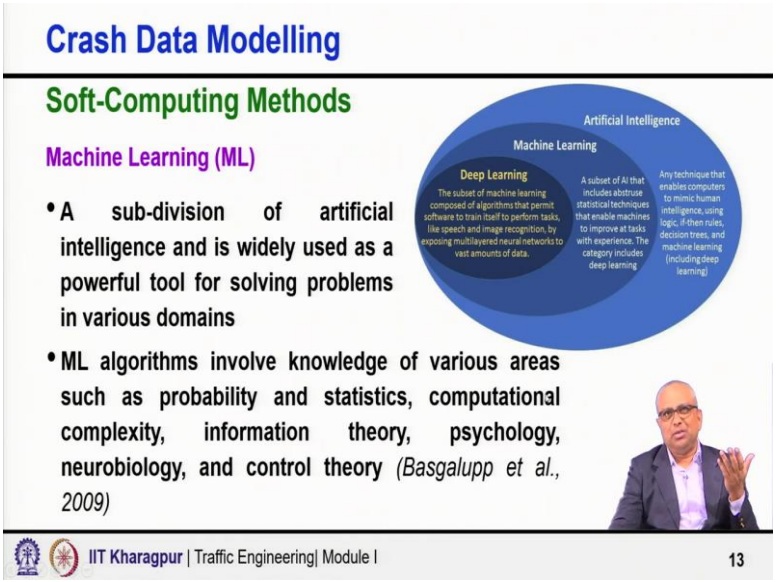
(Refer Slide Time: 13:44)



Now, we go to the next part that is crash data modeling.

(Refer Slide Time: 13:53)





There are different approaches or you shall briefly mentioned to you about two major approaches, statistical methods and then soft computing methods. Obviously, I have some difficulty here because I cannot go into details because any modeling approach if I want to discuss in details that will take a good amount of time. And if you do not have the background, it is very difficult.

But I wanted to tell you that you can actually use very well used statistical methods in different modeling approaches statistical model you can develop, you can also use soft computing techniques, ML based techniques, machine learning based techniques models to do the crash data modeling data meaningfully. So, it is a little bit only some overall understanding or ideas not really going into details because if you know every type of model, it may if I want to explain

you how to do the models, how to develop a model, it will take really a lot of time. So, just a quick overview.

Statistical methods using statistical methods, I can model the crash frequency, you can also say or we also call this model set count data models, because the number we are talking about the crash frequencies, the number of crashes. So, these are count data models. So, I can actually develop count data models.

For example, you may use Poisson model or negative binomial and through development of safety performance functions SPFs want to know more you have to further follow other lectures probably or you have to take some other course, where you can learn more and I am almost towards the end this is the last module of this course. So, I want to touch upon some of the things that how the in which way we should we are proceeding or which way the world is going actually.

So, everything I would not be able to tell you in details, but give you some ideas which way the things are moving. So, here you can model crash frequency using count data model through the development of safety performance function SPFs, now, this is SPFs one way are very useful, because they can help us to identify factors causing crashes, what are the factors that are actually contributing to crashes.

So, number of crash and these factors are related, that is why they are they are in the SPFs safety performance functions, statistically they are contributing. So, if I tell you or give you some SPFs that will tell you that what are the factors that are actually causing crashes. Second, if you have such kind of functions, then they can also be used to predict crash frequencies.

So, I can tell what values of all those factors in a given context means, what is the expected number of crashes I can predict I can forecast. So, one way such kind of safety performance functions which are you can develop town data models it can, they can help us to identify the factors which are causing crisis, but at the other side or on the other hand, they can also help us to do predictions given scenario or given certain conditions how we will able to predict that is the crash frequency. I can also try to model crash severity fatal, non fatal.

In this case, you are only making it two groups and you can easily represent as 01 or it could be fatal, greater injury, minor injury, major injury, minor injury. And or ,or fatal, then major injury, minor injury property damage, you can do more classification you can make it fatal, non fatal to or you can go for more classes.

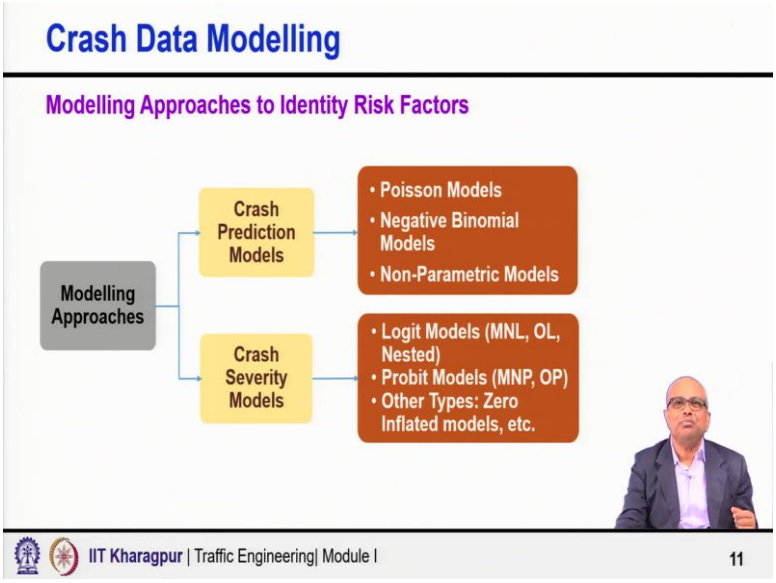
So, I am trying to model the severity using logical explanatory variables and because the I am trying to model the severity which normally is expressed like only using two levels or three levels are four level something like that. So, here what we develop? We develop ordered model because your outputs that 012 like that, ordered model. This ordered models you can develop like ordered probit, ordered logit.

If you have taken my course on urban transportation systems planning there in the context of (20:18) model discrete choice models, I mentioned about logit model, particularly logit binary logit, multinomial logit I mentioned also I indicated what the difference between the logit and the probit is. So, here it is basically ordered logit or ordered probit even other types of model you can also develop.

Now, again if you are really interested to do you know crash data modeling using all these statistical methods probably have to take a course where you can learn more about the how to develop ordered models or that logit to probit or how to develop the count data model if you are trying to model the crash frequency, but here I can only tell you that yes you can model crash frequency you can develop count data models, develop safety performance functions or you can also develop orderd logit or orderd probit to identify factors which are responsible for severe crashes or which are influencing the severity of crash.

Now, this kind of models, ordered models or crash severity models are developed to establish the relationship between the crash severity and the independent variables which are the causal factors and which are statistically contributing or helping you to explain the variation in the crash severity.

(Refer Slide Time: 22:05)



So, if I come to the next slide. Further, again summarize the same thing what I show, so, the statistical modeling or statistical methods when I am applying, I can apply it for crash prediction model number of crash all crashed together. So, are different types of crashes also, the number I can predict crash prediction models or I can use it also for or I can also analyze the data by developing crash severity models.

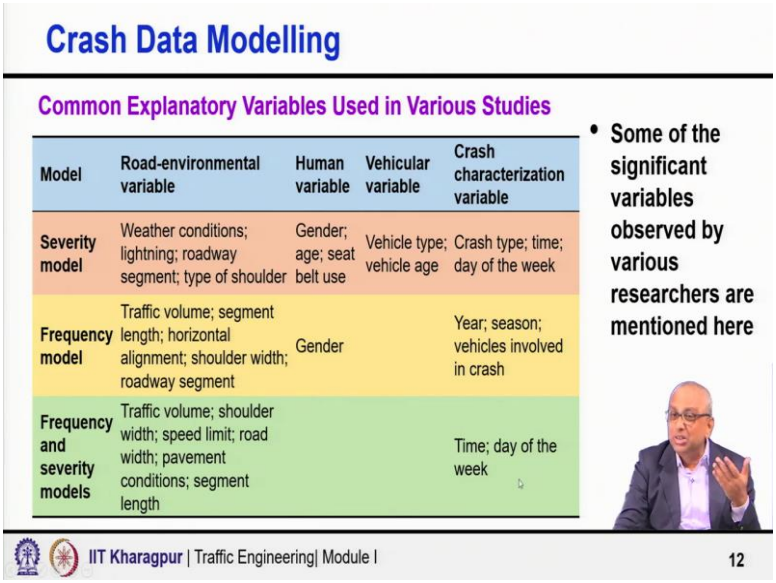
If you are using crash prediction models, these are the some example models, you can use Poisson models, you can use negative binomial models. or several other nonparametric models. The next part also I will go and discuss a little bit, the other way you can use crash severity model. So, there you can use to logit model maybe MNL, ordered logit, nested logit. It depends on how you are formulating what you want to do, what is the type of data what you want to do with that model.

So, depending on that you can go for logit models, you can go for probit model, like MNP or OP order probit and multinomial probit or even go for other types of model for example zero inflated models. There again, I am saying sorry, I have to stop it here I cannot go into detail discussion. But if you are interested, it is only to tell you yes, if you have the crash data, you can try to analyze the crash data and develop models for crash prediction or prediction of the crash severity.

You can do that, people have done it, researchers have done it and several modeling techniques or statistical models can be developed starting from Poisson model negative binomial nonparametric models or if you are doing the crash severity, maybe logit model, probit model,

it depends on how what is your y, what is your x, what is your independent variable, what is your dependent variable, how you are expressing those variables in the database, what you want to do with the model what you are trying to achieve and accordingly you have to select a suitable model specification and then try to calibrate the model for that given database.

(Refer Slide Time: 24:57)



Now, here I am trying to tell you what are the common explanatory variables that have been used in different studies. It is not that everywhere same factors or same variables will come these are the independent variable or the causal variables you can call, not everywhere the same thing will come.

So, I am not saying in your case when you collect data or if you are trying to analyze data, these are the variables that all will become active or these are the only variables which will become active or you will find to be statistically actually significantly contributing to explain you, explain the variation may not be, but it gives you some idea.

So, you know that researchers have used this kind of variable, so, probably in my case, also I can I should include an issue try to see whether they are contributing and if you do not find them significant omit, omit those variables.

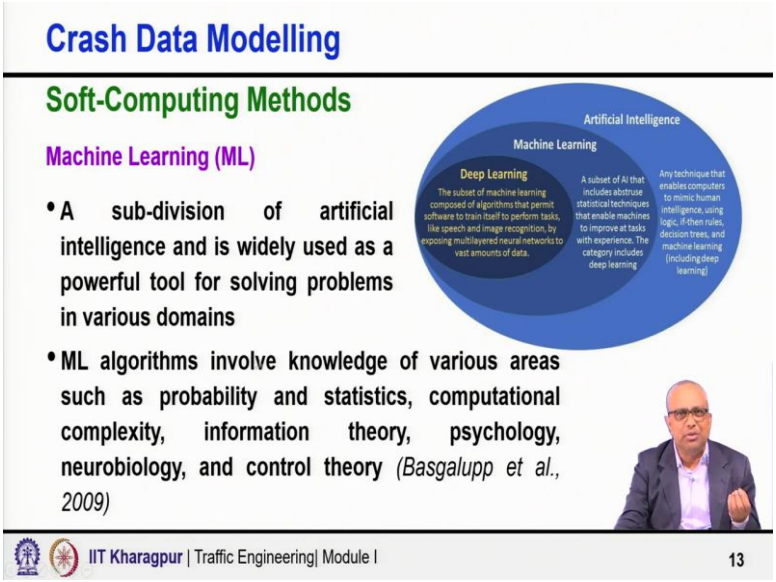
So, people have used for developing severity model they have used different road environment related variables for example, weather condition, lighting, roadway segment and its characteristics type of shoulder that these are the different variables even human variables for example, gender, age, use of seat belt, use of helmet, use of seat belt regular variable such as type of the vehicle age of vehicle and also the crash characterization variable.

For example, what is the type of crash, at what time it happened, what day of the week the crash occurred. So, when you are trying to model the severity, while, while trying to model the severity, researchers have used all such kind of variables which are the road environment related variable human related variable, vehicular related variable and crash characterization related variables.

Similarly, frequency models when they have developed they have used road environment related variables such as traffic volume, segment length, horizontal alignment, shoulder width, roadway segment characteristics human variable like the gender, like crash characterization variable like year season, vehicles involved in crashes.

And sometimes some joint models also researchers have developed frequency and severity models some kind of together or the joint models, there are also they have used very similar kind of variable, where you can see traffic volume, shoulder width, speed limit, roadway, pavement conditions, segment length and also crash characterization variables such as time of the day or day of the week, which day of the week, which time, that kind of variable, but these are all indicative.

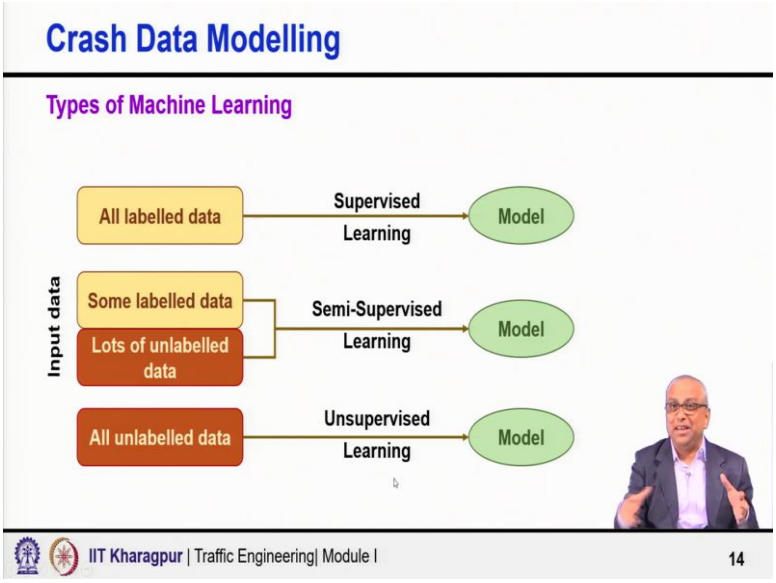
(Refer Slide Time: 28:38)



Now, going to soft computing methods increasingly becoming popular many cases giving very good results very promising way. So, increasingly researchers are using different machine learning techniques and you it is ML techniques, it is subdivision of the AI or artificial intelligence and is widely used as a powerful tool for solving problems in various domains.

And of course, ML algorithm involves knowledge of various areas, I have mentioned some of those.

(Refer Slide Time: 29:23)

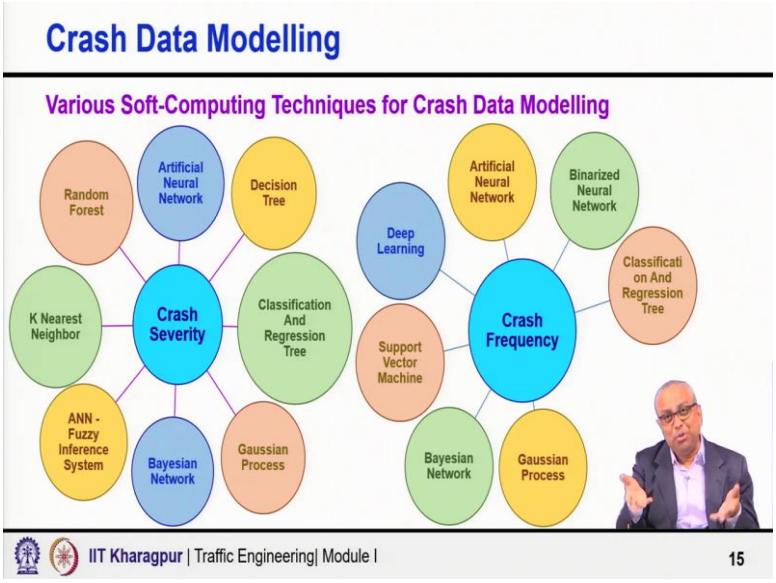


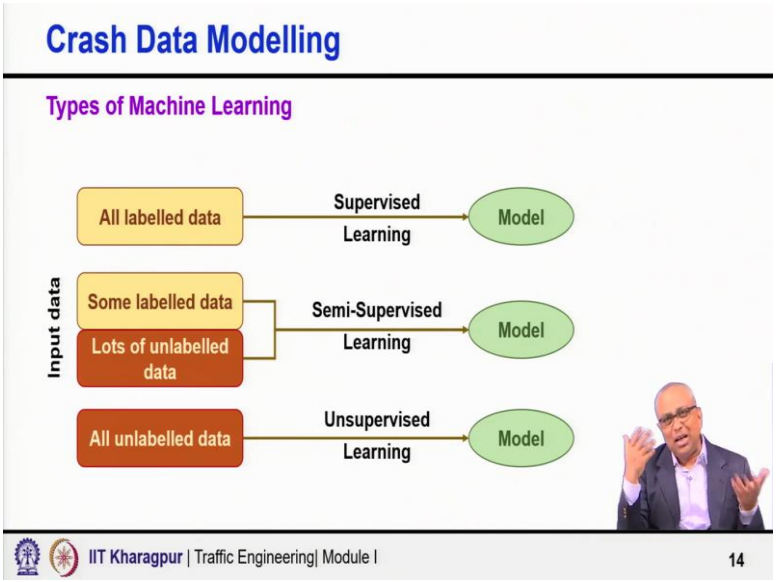
And several techniques have been developed. So, in transportation as a transportation researchers, we do not work on developing new techniques, but rather we try to understand these algorithms and then see given my context, what is the way that I can apply and if I can apply what is the what are the alternative algorithms that I can try, because all algorithms may not work well in all conditions.

So, you have to decide, what kind of algorithms I will use and then within that group, what are the possible things that can be attempted and you try many things and then sometimes a model will give you better performance sometimes some other model will give you in another context some better performance, but overall, very encouraging results are obtained and reported by researchers.

So, depending on the input data, whether you have label data or they have some label data, some unlabel data or whether you have all are label data, depending on those you may use either supervised learning any so, many algorithms out there under supervised learning. So, you may use supervised learning, you may use unsupervised learning again so, many different algorithms that are available or you can use semi supervised learning and develop models.

(Refer Slide Time: 31:01)





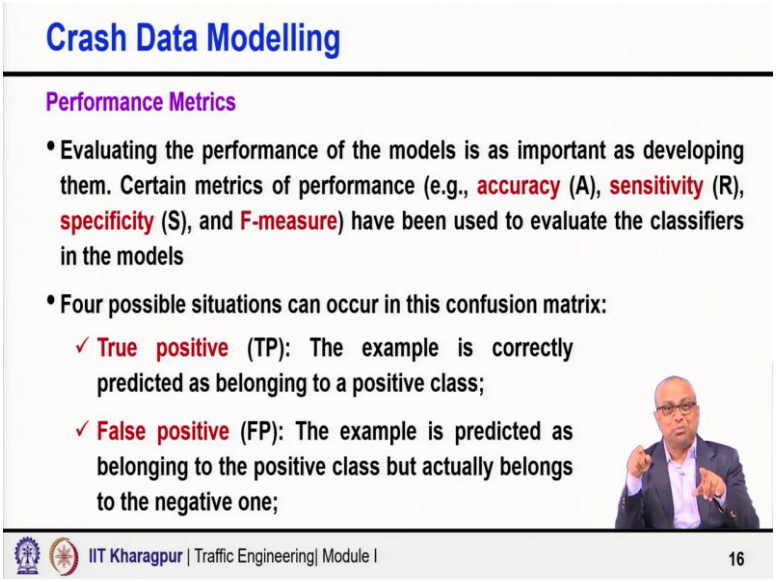
Here I have indicated that several works have been reported in the literature researchers have in various country context various using different databases, they have actually used ML techniques for modeling the crest severity and also the crash frequency and while modeling the crash severity, all such kind of things have been used like starting from random forests to K nearest neighbor to artificial neural network, decision tree, classification and regression tree.

So, number of things they have used while modeling the trying to model the crash severity and many cases very encouraging results are obtained. Similarly, while trying to model the crash frequency, they have again used deep learning artificial neural network, classification and regression tree. Similarly, it will support vector machine, number of things they have tried.

And as I said that it is difficult to tell you a priori which one will exactly what or work better, but this is all you have to see what I am trying to model what are my independent variables, what are my dependent variables and or whatever in the overall input data with its level data or unlabeled data or some label some label so, then decide what kind of learning you want to use supervised learning unsupervised learning or so, then you can think, okay, I have to use supervised learning. So, within supervised learning, what are my possibilities.

And then try different things and compare the models and see they would be in a given context what best is what is working in the best possible manner. Because it is data it has to be data specific, not that you develop maybe in one case random forest gives you very good result, but not every case you will get the always the best result using them for it. So, we have to try alternative things, which maybe tried logically and then see, based on comparison, that what you are accepting and which model finally were accepted. So, often in such work, we try a number of models, statistical model, ML based model, even we compare the performance of different statistical model different level based model and then pick up.

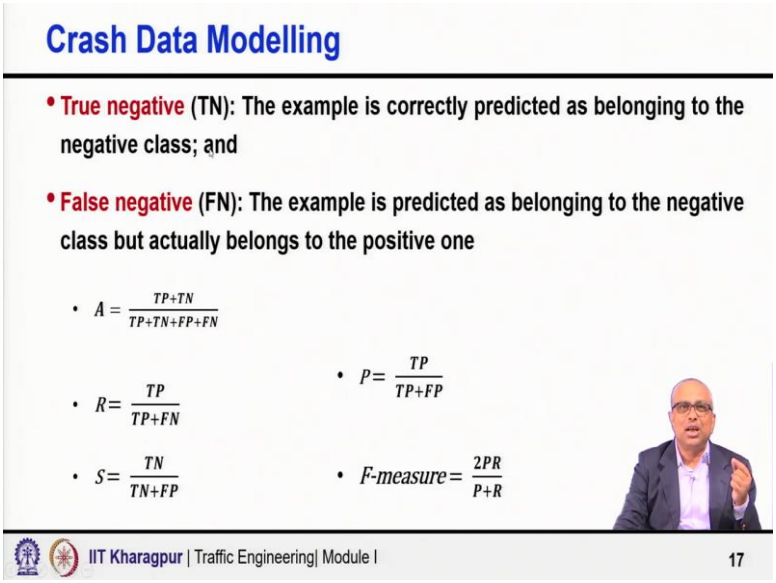
(Refer Slide Time: 33:58)

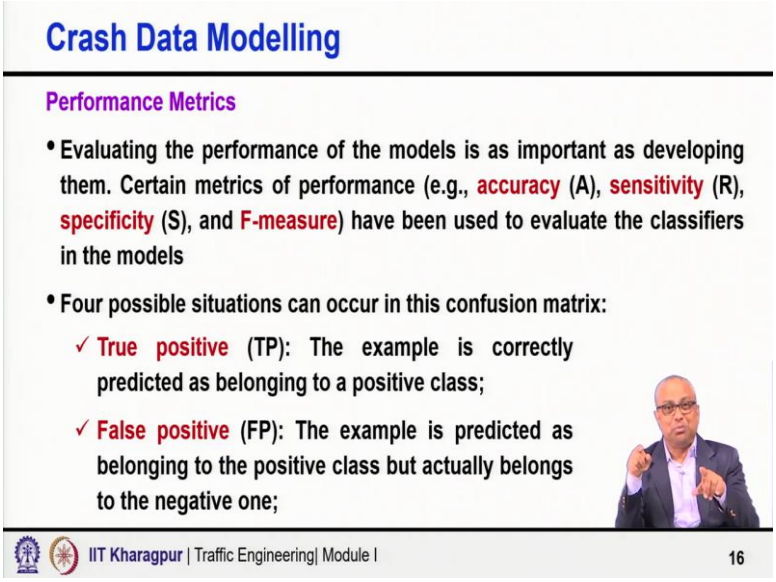


And in this context, the performance metrics is very important. So, evaluating the performance of the model is as important as developing them. Now certain metrics of performance for example, accuracy, sensitivity, specificity, and methods have been used to evaluate the classifiers in the models and there are four possible situations which can occur in this confusion matrix.

One is true positive that means, the example is correctly predicted at belonging as belonging to a positive class. So, example is correctly predicted as belonging to a positive class, that is true positive. False positive means, it is not true positive it is false positive that means, the example is predicted as belonging to positive class it is false positive, it is predicted belonging to positive class, but actually belongs to negative class. So, that is why we are calling it a false positive.

(Refer Slide Time: 35:13)

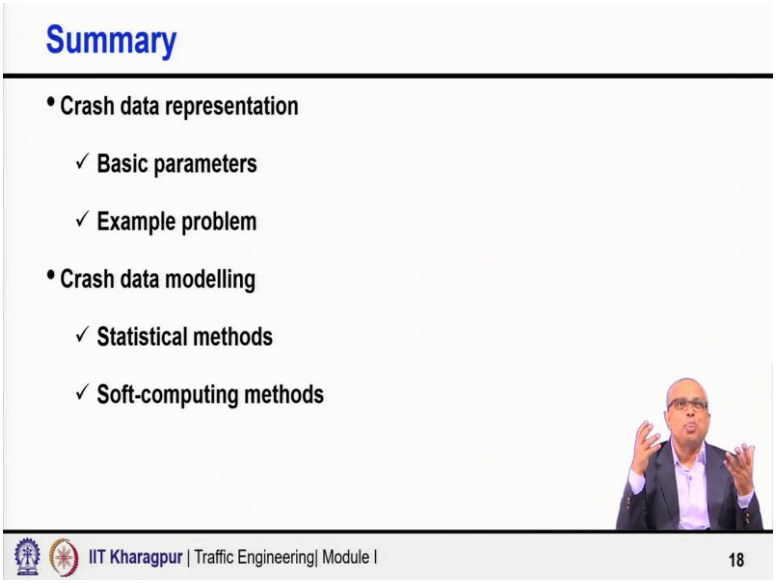




Similarly, it could be true negative that when the example is correctly predicted as belonging to the negative class, it may be false negative that means, the example is predicted as belonging to negative negative class because false negative prediction is negative, but that is false. So, actually it belongs to the positive one. So, the example is predicted as belonging to negative

class, but actually belongs to the positive one. And here I have shown you I have told here an accuracy sensitivity specificity, F-measure, so, A, R, S, F-measure. How you can calculate it using this true positive, false positive, true negative, false negative. So, that is what I have explained.

(Refer Slide Time: 36:09)

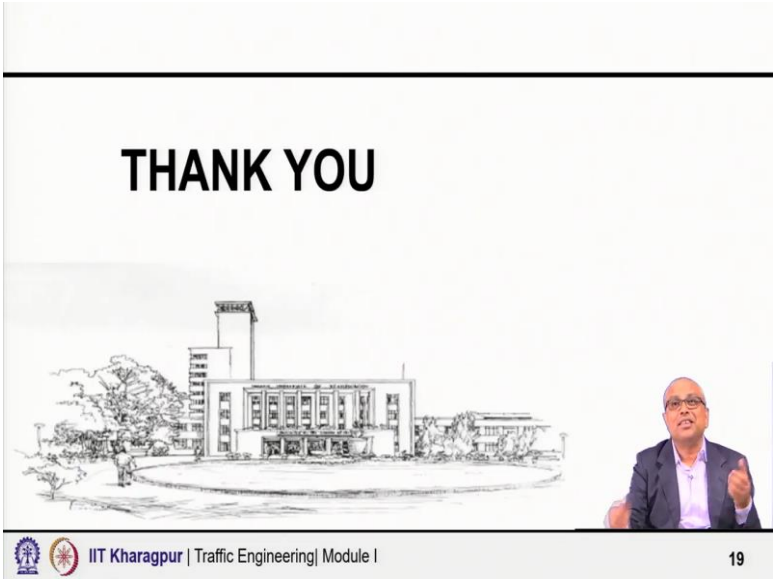


So, with this I would like to summarize that what we discussed here the representation of the crash data use using several important basic parameters, multiple parameters, we discussed basic parameters, took some example problem also to show you one example problem we took to show you different situation, different period volume length, but still how we can meaning compared the meaningfully.

Then talked about the crash data modelling said that you may use statistical methods or you soft-computing methods for doing the crash data modeling. What kind of models you want to develop, what you want to predict, it will predict the severity you may predict the number of crashes.

So, depending on that what are the if you are using statistical methods then what are the approaches what are the alternatives that are available and in soft-computing methods also particularly the ML techniques, I indicated also what all have been attempted by different researchers, while trying to model the crash severity as well as the crash frequency.

(Refer Slide Time: 37:32)



So, with this I close this lecture. Thank you so much.