

Learning Analytics Tools

Professor. Ramkumar Rajendran

Department of Educational Technology

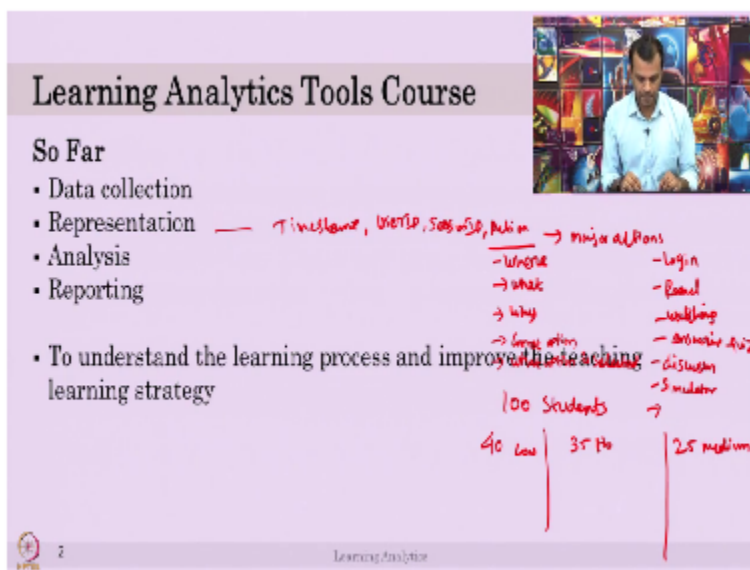
Indian Institute of Technology, Bombay

Lecture No. 59

Revision of Learning Analytics tools course

Welcome back to Learning Analytics Tools Course. This week we will just provide an overview of what we learned in this whole semester, there will not be any new topics in this class but we will revisit what we discussed in an abstract view.

(Refer Slide Time: 0:31)



Learning Analytics Tools Course

So Far

- Data collection
- Representation — *timestamp, UserID, SessionID, Action* → *major actions*
 - *login*
 - *logout*
 - *what*
 - *why*
- Analysis
- Reporting

• To understand the learning process and improve the teaching learning strategy

100 Students

<i>40 low</i>	<i>35 Ho</i>	<i>25 medium</i>
---------------	--------------	------------------

Handwritten notes on the right side of the slide:

- *major actions*
- *login*
- *logout*
- *what*
- *why*
- *login*
- *logout*
- *what*
- *why*
- *login*
- *logout*
- *what*
- *why*
- *login*
- *logout*
- *what*
- *why*

So, in this course so far we talked about

1. data collection (i.e what data to collect and from which environment)
2. how to represent/store the data - I suggested the method that you collect **timestamp** and **user ID**, if you allow the students to do multiple sessions (i.e they have to interact with

the system multiple times) collect the **session ID** and the **action** they are doing (i.e reading, or uploading an assignment), we can also collect the contextual data (i.e where the action is taking place (on some document or simulator), what page they are reading, if they are answering some questions then what is their response etc).

So, I might be interacting with the technology-enhanced learning environment. I might have an interaction like navigating from one window to another window, clicking some buttons. I suggested you capture all the clicks. But do I have to consider everything as action? Not necessarily because if you consider everything as action there will be too many features to come up. So, think of the major actions they do, it is only the major actions that are considered. For example, in a Moodle or in a MOOC system or in a technological learning environment what are the major actions they do?

You might consider login, as the logging into a system is start of the action (it is not necessarily important for us when you do the no analysis in a sequential manner, but you can consider login to know when it is started), the login is associated with the session ID that also tell us whether the session is continued or they logged in multiple times.

And by login ID you can construct features like how many times a login happened in a week, in a day or within hours and the average time they spend on each login time. After logging in what are the main actions they do in the environment? For example, if they use some LMS they might be reading something, watching videos, so reading, watching videos or answering questions, taking a quiz, they might be in the discussion forum, discussions.

In some of the systems we saw they are interacting with the simulator, some systems they are drag and dropping features, so if you have any specific items mark them as actions. Now, for each action say “read” (the student is reading), we can capture - which page they are reading or which PDF or what are the content they are reading, if you chunk the content(reading material) into smaller concepts, it is easy to track that.

Then if they are watching a video, what are they watching the video, if you want to minute data, talk about are they watching in a single 1x speed or 1.5 x speed.

So, watching speed, are they seeking the video from one place to place, you can add all this information as contextual information.

So, the first step, when you do the analysis, is that you plot the frequency number of times actions occur for each student, the time they spent on each of these actions. And if there is some interesting aspect that happens, go look into the contextual information, why is this student looking at it?

A typical example is, for example, you have 100 students, you saw 50 students are low performers(based on a pretest and post test score) and 40 students based on pretest score are high performers and there are some students who are not neither low nor neither high but in between. But taking 50 as cutoff makes no sense, so we can give a median value or some gap for bifurcation.

Now, you have this data, first step you do is after collecting these features, the features you are talking about login, read, watching, plot the graph and find out is there any difference like in the frequency, in the average time they spent, if there is a difference well and good, go and talk about it why the difference whether that difference can impact the students' performance that is hypothesis, test it out with the data correlation or do prediction that is good. But you see that for both students all of them are reading time is same, watching time is also same, kind of same there is no significant difference, we can run a significance test on this data.

Then you might want to go further, if both low scoring students and high scorers are reading for same time on average, how is it that some students are able to get good scores and others are not able to score good? Maybe it lies in what they are reading and how they are reading. Now, we go look into the contextual information and see if there is any difference in the context that comes out, that is how we do the analysis. So, I was trying to explain this in week 2, so you might have had the idea of what is action, what is the context, but I want to give you the picture why it is happening in the learning analytics.

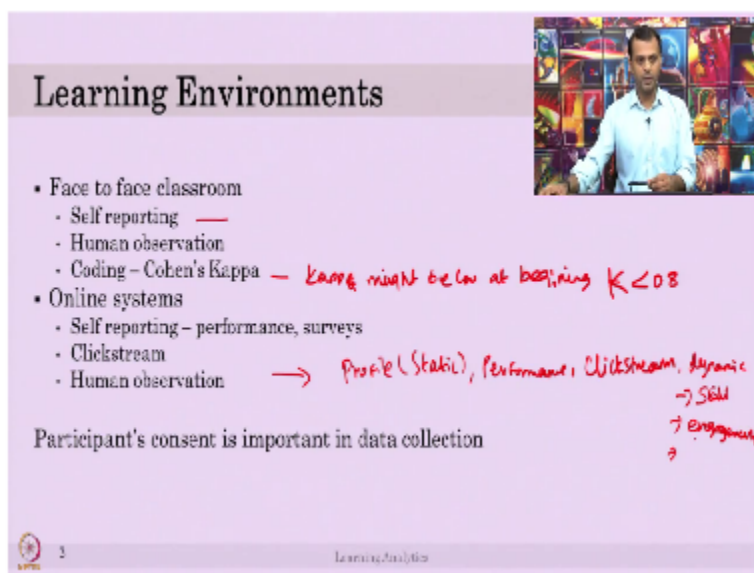
Next is we do analysis, as I was explaining that, then you do reporting, so since there is a lot of data and you need a lot of graphs. Do you want to report everything? So, depending on what you

are reporting and whom you are reporting you have to make a graph and make a plot, so make inference out of it and report it.

So, make sure that you know what are reporting and who are you reporting to, and what is their interest, or what you want to tell them like you do not tell all the story to the everyone in the public, so you might be interested in saying that, hey we know all these other factors are already established in literature, we also found the same, but there is something new we found , which is not accessed anywhere in literature, we want to talk about that, let us talk about that, something new. So, that is a gap you have to identify and report it.

Then after reporting the whole idea is to understand the learning process and improve the teaching learning strategy. So, this is kind of what learning analytics is, a definition I talked about in a week 1 lecture. If you remember what learning analytics is? LA is to collect data, measure, report, analysis of data in the context of students working on it. And in order to improve the students' learning, the primary goal is to improve student's learning. It is not only about how you create a model and predict the student's performance but it is about how you use that knowledge to improve your teaching learning strategies so the students learn better.

(Refer Slide Time: 10:08)



Learning Environments

- Face to face classroom
 - Self reporting —
 - Human observation
 - Coding – Cohen's Kappa — *Kappa might be low at beginning $K < 0.8$*
- Online systems
 - Self reporting – performance, surveys
 - Clickstream
 - Human observation → *Profile (Static), Performance, Clickstream, Dynamic*
 - Skill
 - Engagement
 -

Participant's consent is important in data collection

3 Learning Analytics

So, we talked about different learning environments like face to face, online systems and also LMS kind of things, online systems includes LMS and Tele both. So, in a face to face most of the data collection is by self reporting, this is a key, actually very important, most of the data is collected by self reporting or human observation, there are only things you can do. Self reporting is that you might run a survey, ask students what you are feeling about or how many students are engaged in the class or you ask them some kind of questions, they answer something and based on that you can identify whether they understood or not.

And self-reporting can be done online, also you can use a small piece of paper circulated in the classroom and they will answer some set of questions. Or the second way of directing data is human observation, you are a teacher in, you are an instructor, you are in the class and you are collecting students' performance, student's engagement, student's activities or number of times students interacted with the peer, all these things you are collecting.

Human observation is one of the valid or well accepted observations, the only key point is the researcher should not get biased in labeling each and every action of the students, the participants. So, in order to avoid bias please create an inter rater or inter observable reliability (it is based on the observer or rater), so we use Cohen's Kappa. We talked about Cohen's Kappa also in our earlier weeks. So there we mentioned the Kappa is used to compare performance of two systems or to measure the agreement between two observers. So, look at what is inter rater reliability, how to use Cohen's Kappa if you are doing human observation this is a very key part.

The reason is you should not get bias to labeling the particular variable you are recording, say for example, in the classroom I want to understand the students engagement and you might be the observer standing in the classroom, say there are 5 observers observing a class of 50, each observes observing 50 students, they observing using the Round Robin method, so that is every 2 minutes I will observe student 1, next 2 minutes I will observe student 2, like at the end of twentieth minute you observe the all 10 students and you come back to first student at 20th minute i.e you will be observing the same student again. You are marking in a note saying that student 1 from time 10:00 to 10:02 is engaged or is interacting with the peer etc.

So, you have to first come up with the coding mechanism. What are the things you have to code? There are a lot of things, this is called qualitative research. There are a lot of resources available

in such articles, go and look at the coding existing you might find new coding. Once you have a coding assignment you record the student's engagement. And how do you compare whether you are doing it correctly and that you are not biased. So, what you do is compare your coding with your peers, other person's coding, so that when you compare inter-observable reliability both observers have to observe the same student for a certain period of time.

So, usually when we start this never happens initially this Kappa may not be good or really great. So, what we do is first you will have a set of rubrics (like mark the student as non engaging when a student moves and talks), you have your own rubrics coding mechanism, you discuss with your peer and both understand the coding mechanism and then start observing.

After observation you do the Kappa calculation if the kappa is low, i.e. less than say 0.8, discuss with your peer who is other observer and talk where the mistake is made and why you thought that the particular action is not engagement and you thought it as disengagement, why the particular other observer thought that as engagement.

Discuss and resolve the conflicts, redo the assignment again, observe a new student or same student's for another period of time. Then you check again the Kappa score, make sure the kappa score is more than 0.8.

Sometimes it is not that you will be observing the students in a real live environment, you might be recording the students facial expressions, students actions in the class and you will be looking at the video and recording it because you want more data, if you want to observe the data in a real classroom you may not able to observe all the 50 students data or you do not have too many observers to observe the data, so you will not have real time data.

So, what you might do is, you might keep 2 or 3 cameras, observe, record the students actions and you might look at the student 1 in the camera and mrk down engagement. In those scenarios again talk to your inter observer both watch the same amount of time say 5 minutes or 10 minutes of video and mark down all the rubrics, make sure the Kappa is greater than 0.8, if not again redo the assignment on a new student or a new time frame, not the same time. So, make sure you make the Kappa is more than 0.8. So, even observation can happen in real time also in the video that is the thing I want to discuss here.

The second one is online systems for self reporting, it is not just self reporting of service also the students are answering the questions, the performance all these things can also be considered, other thing is clickstream data, human observation data. In general, in ITS kind of environment what we consider something called a profile, this is kind of a static information, static in the sense it is a profile of a student like age, gender, the year of the study, the prior knowledge, maybe if you are collecting the parents information, all these things kind of profile information when they come to systems kind of static.

And clickstream data is other data you observe all the interactions. And then you might the performance data is another data slack profile, performance, click stream data like that is the data we use by using the clickstream data and the performance by combining these two, you might be coming up with some dynamic data. What I mean is dynamic data: you might be measuring the skill or you might be measuring the engagement, something you might be measuring something for your research that data is dynamic. Why do I say dynamic? At the beginning you might start with no median value or middle value.

Then based on students click stream and performance you will change it whether delta increases or decreases, so a student is performing very well and is interacting with all the artifacts in the system, you might increase the engagement more and more. Some are not interacting with all the systems just simply talking, watching a video, they are not doing any performance, you might reduce the values representing some skills, so this has changed based on the student's performance on the dynamic.

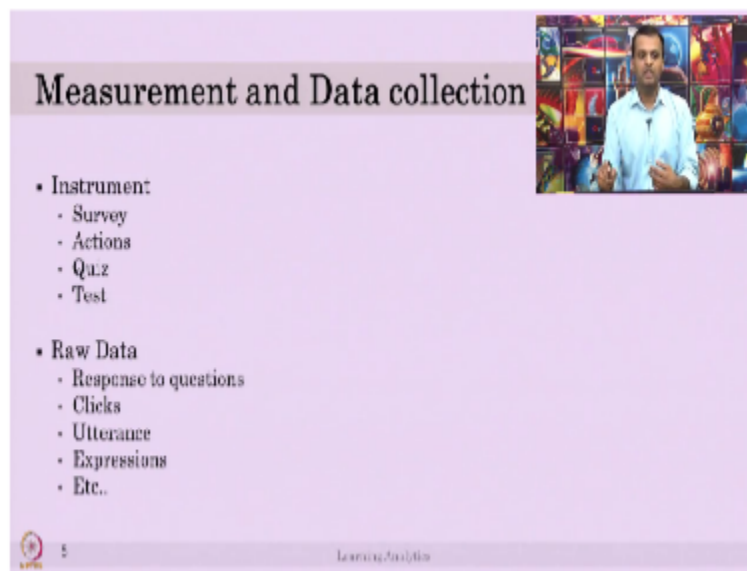
So, in general by using this dynamic data what you actually provide is new feedback, a new content it is part of Intelligent Tutoring systems, so that is the whole idea. So, that is the whole idea of clickstream data, even in online systems you do the human observation that is why I said that if you might capture the videos or something and you might sit down and record the facial expressions and do it.

For example, in online environment you might be capturing students facial expressions or data from eye trackers or something else, if you have facial expressions data, you may not be able to code it directly, you can ask the students to self report about their own emotions watching their

own videos or you can sit, observers can observe the students facial expressions and code their emotions or affective states.

So, in any case make sure that you do the Cohen's Kappa and the Kappa score is more than 0.8, there should be no bias in human observation. If you are not doing that the people do not consider that data is valid that is very very important. So, all this data collection, any environment, any data you collect please get the participants consent and if you are getting a participant consent please make sure with whom you can share the data, who has access to data, in the future what you will do, what are their rights and think about all these things when you do the participants consent data.

(Refer Slide Time: 20:04)



The slide has a light purple background. At the top, the title 'Measurement and Data collection' is written in a bold, black, serif font. Below the title, there are two bulleted lists. The first list is under the heading '• Instrument' and includes 'Survey', 'Actions', 'Quiz', and 'Test'. The second list is under the heading '• Raw Data' and includes 'Response to questions', 'Clicks', 'Utterance', 'Expressions', and 'Etc..'. In the top right corner, there is a small video inset showing a man in a light blue shirt speaking. At the bottom left, there is a small logo with the text 'Lecture 12' and '5'. At the bottom right, the text 'Learning Analytics' is visible.

- Instrument
 - Survey
 - Actions
 - Quiz
 - Test
- Raw Data
 - Response to questions
 - Clicks
 - Utterance
 - Expressions
 - Etc..

So, what we talked about is in environments we might collect data from different instruments like surveys on the clickstream data, quizzes, the test performance all these things we talked about. And these are all raw data and from this we get raw data like response to questions or the clicks they speak some utterances or the facial expressions.

(Refer Slide Time: 20:35)

Activity



How do we extract features from raw data and why?




Learning Analytics

So, now think for a minute how you extract features from this raw data and why we are doing that, that is very important. I can collect all the data, facial expressions they are talking, the utterances and the clicks, every click, performance. Why are we doing that and why how do you extract features from this log data? Please pause for a minute and write down your answer, after writing it down resume to continue.

(Refer Slide Time: 21:02)

Activity: Response



Feature Extraction

- Depends on your research goal
- Domain knowledge
- Frequency and time

→ Peak → number of peak in one session

* * * * * last 5 sessions

4 0 2 11 40 number

→ avg Peak time in 1 session

→ avg Peak before taking Quiz

Learning Analytics

So, the feature extraction depends on like why we are doing it i.e why you have to go and predict something or compare the students with other students. So, what features to extract it depends on the research goal, that is why I mentioned in the first slide the actions. You want to go to the context information level or simply the actions, so you come up with your own number of features by simply using the major actions level or combining actions with the contextual information. And this is where domain knowledge is important.

Not everyone will be extracting a good feature the one who is working on the domain say education domain will be able to extract a good feature, if you are creating your own system and you know the students interaction the system means something so then you might be able to tell from the experience, the experts knowledge you have in that particular system or the particular domain you might say students clicking these 3 buttons or reading for 5 minutes and doing quiz might be look like the student is reading this or something.

How do we know that, it is like, it is your expertise and if I talk to a teacher in a classroom environment, if a student is coming for only 40 percent of time for attendance and he is not submitting 3 assignments at all, will he pass the exam? The teacher will say yeah, he will not pass the exam because I know it based on experience. Domain knowledge is expertise knowledge you might possess in this domain.

So, my focus is to use your expertise knowledge for the creating features. The reason is the domain knowledge will depend on how many years you are working on it, how many systems you created. The one way to start is read the research papers to understand how they extract features, what are the features extracted in a similar environment, look at those features, list down all the features then you might be able to come up with your own set of features or combine some features to create your own features.

And frequency and time is important I said that if there is a major actions, simply create a feature as a frequency of that action over a certain period say over last 1 hour, last few hours, last 2 days, last 3 days, last 5 days over 1 session, over the last 5 sessions, sequence of any major action. For example, if the major action is reading, what I do is a number of times did, number of read in one session, I am not talking about what are they reading, which page they are reading is needed if you go further but I am just saying if you are reading just go about how many times they are

reading. Same, “read action” in the last 5 sessions or the same in 40 minutes so you can extract multiple features from one major action. So, why do we expect multiple features?

You can access multi features and find out which features are redundant by doing a correlation analysis or correlation with a dependent variable or feature extraction methods will help you to do that. You might say it does not make sense to look at the read session for the last 5 sessions, if you have a particular domain knowledge, you might say no this particular scenario I do not think last 5 sessions of reading will not really helpful, you can remove it, that is why I say domain knowledge is applicable.


So, for the same read action I will create multiple features based on the time average, average read time in one session you know it is that and you can talk about average read in last 20 minutes or average read in last 5 sessions or you can be specific that average read before taking a quiz something like that. So, here I am combining two actions so you can come up with a lot of features that is why I am saying that go and look at the papers already which is talked about this and also the domain knowledge will help you to come up with, you can say that if a student is reading more than 3 minutes before taking the quiz then there might be you might be doing better or if the student is reading many times like spending 5 minutes and you did some other actions then quiz this may not be good.

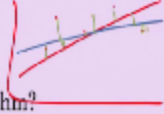
So, you might come up with your own hypothesis, to test it you extract these features from the raw data. So, there are two things, **frequency** and **time** this is a very basics and you can all start with that, how do you improve the set of features, how to come up with the new features and combining multiple major actions or you can combine the major plus contextual information, it is all come up with your domain knowledge.


(Refer Slide Time: 26:58)

Analysis

- Basics of ML algorithms
 - Pattern Mining, Process Mining
 - Clustering Techniques
 - Performance metrics
- How these algorithms are trained?
- What are the hyper parameters in each algorithm?






Learning Analytics

So, and what we did also we talked about basics in machine learning algorithms, we talked about pattern mining, process mining and clustering techniques, performance metrics, some of the regression logistic, we talked about all these things in this course. So, we will revisit pattern mining and process mining, most of you might know the clustering techniques and regression, process mining and pattern mining might be new to many of you, check it again it is very interesting and check for associate mining also for the pattern mining.

And importantly this performance metric might be something new. The reason is not many people cared about what is a metric, what to look for that is very very important, and why this metric is needed is also I wanted to inform you guys, so that how to compare two things. Then how this algorithms are trained, I never talked about it. In a linear regression we talked about how to find the best fitted line.

So, from each point you have to find out the gap, this is the loss function of this and similarly this loss function for this function. I never talked about how it is trained. There is a reason for that and I also never talked about hyper parameters in each algorithm because if I talk about training then I should be involving hyper parameters. At the beginning of the course also in the introduction I clearly mentioned this course is for someone who is new to machine learning and the course does not require any mathematics or anything.

So, I completely avoided calculus in this course. So, in order to avoid calculus I did not talk about how it is trained and what are the hyper parameters in each other. But I request all of you to go and if you are interested to more about that or go and watch videos by professor Andrew Ng each of this, each of these recordings and learn more about it or lot of very good resources are available in internet, so, if you are really interested about machine learning there are very good books available in machine learning also.

The aim or focus of this course is not that, it is not to teach every algorithm with training and training and finding parameters because this course is not for only the one who is already well trained in the mathematics or we know the programming or something like that this is for everyone, so I just kept that part very very low, if you are much motivated please read further.

(Refer Slide Time: 30:18)



So, also in this course we talked about multiple tools iSAT we had a video I hope you would have done the assignment on iSAT, use it the iSAT tool is available just use it and check this code, if it is useful use it and you can contact the developer and he will be happy and you can do that. Also we have given small scripts to run a sequential pattern mining, it is not a tool as a iSAT or the other big tool but we wrote a small script to extract the sequential pattern mining, if you

have find some other tools available online please go and use it there is no harm, it might call us associate mining or something like that.

So, understand what is pattern mining which checks up frequent patterns in a fine grained level and what is a process mining, the process mining absorbs from the whole process. So, process mining we introduced ProM software that is available for free for academic usage, also for commercial that is one software for free. So, we use that software too. Weka is completely free for everyone, use it, Orange is free for academics. So, use these tools do not stop only with these tools, if you are interested go to Tableau go to Rapid Miner, explore more tools, I will talk about that in detail or how you can expand the next steps.



But these 5 tools we covered and make sure you learn everything well and if the videos are not enough go and check YouTube videos for process miner, Weka and Orange, iSAT that is simple, very simple to do, we gave a sample data just upload it you will see everything we just have to click buttons it is very simple. SPM is it is not, it may not be the complete tool you have to run a script, if you are interested such no other softwares which talks about associate mining, we do not find any good software for associate mining in an educational setup that can be shared freely with everyone that is why we have to write our own scripts.


Hope you guys enjoyed interacting with the tools because this course is about tools and this main focus is that you get to know some tools and play on some data.

(Refer Slide Time: 32:32)

Activity

What is the difference between 20th century vs 21st century teachers/learners?




 Learning Analytics 11


So, now I want to ask something different, that is, what is the difference between the twentieth century and the twenty-first century? When I say the twentieth century, imagine the time before 2000 or you can go up to 2003 or 4. What is the difference between twentieth century teachers, learners versus twenty first century teachers, learners? Think about it, write down your answers, after writing it down please assume to continue.

(Refer Slide Time: 33:04)

Activity: Response



- 20th Century Teacher
 - Source of knowledge
 - Teaches everything
 - Always there to answer question.
 - Guides you for advanced topics or studies
- 21st Century Teacher
 - Motivator
 - Guides the student for resources



Learning Analytics

In twentieth century we see teacher has source of knowledge I did not have access to most of the books only the library have limited books, the teacher has all the access to books, on the library books and you might have a notes from somewhere, we might have to read a lot of books and in the library also you have special access to books and the teacher teaches everything with examples, there is a textbook he covers the each and every syllabus and he explains the problem and he goes off and asked us to solve the exercise, solutions and problems if you have doubts the teachers might teach.

And we see the teacher as a guru like he is a known Alexa of knowledge and he has everything and he is the one person to go after for everything to ask for this. And he is always there to answer my questions and he guides in fact not just teaching, the good teacher guides beyond your subject, hey what to do, how to do and a lot of discussions, a lot of things happened, this is twenty first century teachers. If you are a teacher or you are born in the 1980s you might be at, you might have seen the twenty first, twentieth century teachers. But if you are a teacher now, you are continuing that, you are not a twentieth century, you are a twenty-first century teacher.

What is a twenty-first century teacher? You are not a source of knowledge anymore, the whatever data you can access, whatever resource you have, the students have more than that, students have access to lot of more data and lot of more resources, you are not a source of knowledge, do not expect that you have to teach everything and you have to be there and you

know everything that is all gone, student knows more than you in a particular subject and not just students a lot of interest experts and everybody writes their own blog it is all changed after web 2.0 and the videos are coming in YouTube or other video serving platforms.

So, you are a motivator to students like which course to do and the students for resources guides them, look at this particular resources for this particular topic, but that does not work much, they are the rating agencies, you go to Quora, you check out which resource is good, you read about it, but you are kind of pulling together saying that do this particular task, maybe you are guiding them what to do next if they have any issues.

But there are, you are not the only person who is the source of knowledge that is what I am trying to say so you be a motivator and do that. So, I picked up that I am not a twentieth century teacher, I was a twentieth century student, I was asking my teacher for everything, I go to the library, only the library has 4 books I read notes, I read it. But now I am a twenty first century teacher, I did not teach everything about all these videos.

What I trying to do is, motivate you on, make interest on, this is linear regression, this is logistic regression, this is Naïve biase, this can be applied here, this is decision tree I did not talk about anything about decision tree in detail, this is a decision tree, how to do it, there are two important parameters, with that knowledge what is entropy and information gain in decision tree you can go to any resource, now you can pick it up easily.

So, now what is these two and you can watch a new video or read a paper or read a book, you might get more information and more interest all depends on your interest. So, all my videos I always talk about, all these ML algorithms please go to professor Andrew Ng video if you are interested more about that or any videos, I recommend couple of papers in the videos for you to read because that is how this data all this model is applied in a real data, how people have used it.

And for the tools if it is ProM or Weka or Orange or Rapidminer Tableau, go to YouTube there are plenty of videos explaining how to apply each and every algorithm like what data apply which buttons to clicks everything is there, there is no need for us to teach everything which is already existing, we do not want to reinvent the wheel. So, the idea is to motivate you. There is a tool this can be used, go and learn it, that is a whole idea.

And this course is trying to pull out what data to collect, what are the learning resources, what are the different environments and how you can apply it. Pulling these three things, the data collection, the environment and the ML together and how to infer that is what we are trying to do in this course, not to teach every ML or the tools or algorithms. Thank you.