

## **Learning Analytics Tools**

**Professor Ramkumar Rajendran**

**Department of Educational Technology**

**Indian Institute of Technology Bombay**

### **Lecture No 2.1**

#### **Data Collection from Different learning environment**

Welcome to learning analytics tools course. In this lecture, we will talk about data collection in different learning environments. You might have seen, in this course, we are offering in a learners Centric MOOCs model. So, what is learners centric MOOCs model? So, the video lectures which you are watching now, it is called learning dialogues. That is the only place where we interact with you so that you can understand the concept and each learning dialogue will have a reflections part.

So, we will pause the video in between, we ask you to think about it so that you can think about the answers and write down in a paper. Then we discuss the generic answers, the users come up with. So, when we ask you to pause the pause, please pause the video, think about it, that will help you to learn better.

(Refer Slide Time: 1:05)

## Learner Centric MOOCs

- In the learner centric MOOCs (LCM) model
  - Learning dialogues (video lectures)
  - Learn by doing
  - Learn from answering focused questions in the forum - LxI



And there are “learning by doing activities” that is we will give some problem or some questions which you have to solve by applying the knowledge you learned in learning dialogues. These are however usually not graded for your assignment or something, but we highly recommend you to do that because it will help you to understand and apply the knowledge you learned in a different environment or different data set.

The most important thing is we learn not only by attending lectures, instead we learn by interacting with your peers, with friends in the class. To make that same environment possible, we want you to interact with your peers in the forum. To help you to start a discussion in the forum, we will have focused questions. So please respond to others’ comments and answer your peer’s questions, like them, comment them. Also, you can answer the questions. So, we recommend you to go to new discussion forums so that you will interact and learn from the other learners. That is called LxI. It is called the Learner Experience Interaction.

(Refer Slide Time: 2:29)

## Learner Centric MOOCs

- In the learner centric MOOCs (LCM) model
  - Learning dialogues (video lectures)
  - Learn by doing
  - Learn from answering focused questions in the forum



Also, we have Learning Extension Trajectories. This is a very important part. For example, we are in the twenty-first century, here we do not need to be like teachers in twentieth-century teachers. For example, in twentieth-century teachers, the teacher has the source of knowledge, the teacher has the access to some books in the library so that teacher has all the knowledge.

The students who sit in the class learn only from the teachers, whatever teachers say is true. They cannot go and verify because they do not have access to all the knowledge, book or anything. However, in twenty-first century, the students have more access to knowledge, access to more materials compared to teachers or usually, the teacher also have similar access, so they can go and check the internet, they can watch videos, they can do a lot of things.

So, there is no need for the teacher to teach everything in the course. And in fact, the teacher need not to teach. The teacher has to just guide the students, motivate them to learn a particular topic. They just have to motivate them so that the students can go and watch videos from other lecture. Also, there are very good lectures, very good videos on YouTube, or in a MOOC which explain the same concept in a very beautiful way which students appreciate and student understand.

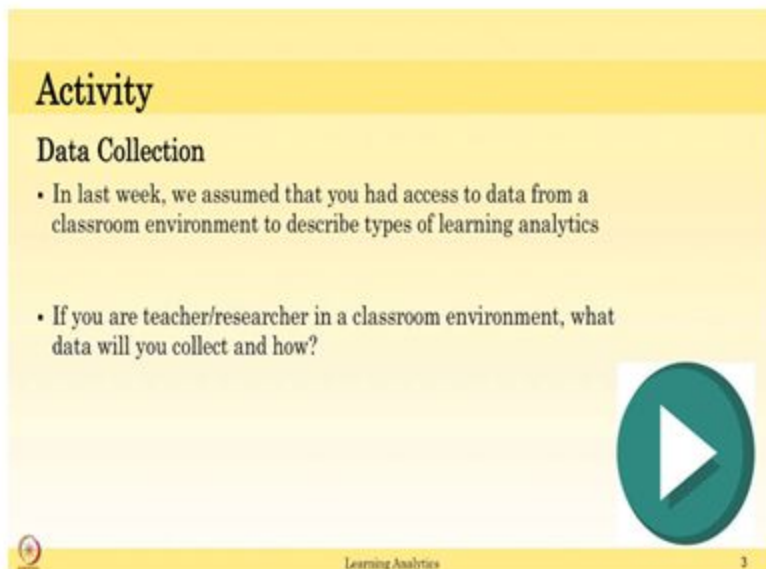
Which means why we are doing this course? There are lot of courses on data analytics. This course learning analytics is to teach you about how to collect data in the learning environments

and use the tools for the data you collect from the learning environment. So, the other courses exist online may not talk about how to collect data or the domain is not education.

So, in order to apply data analytics in education domain, we are teaching this course. So, I said that there are a lot of other videos which means we will give the basic motivation and the videos which are needed to understand what is learning analytics. However, for the advanced users to understand more on the topic, we recommend you to go and watch other videos. This is called as “LxT”. We will provide the resources/links to those videos.

And also we will have an assessment question based on that which is not going to be graded. But I recommend the interested users to go and watch the other LxTs to understand the concept better and learn more about the topic we discuss in this lecture. So, this is about LCM’s, so, learners centric MOOCs. So, this course, we will be offering in the LCM mode. So, when you see LxI, LxT or LBD, please participate actively. Just watching videos and going for an exam will not help you to learn the course better.

(Refer Slide Time: 5:32)



**Activity**

**Data Collection**

- In last week, we assumed that you had access to data from a classroom environment to describe types of learning analytics
- If you are teacher/researcher in a classroom environment, what data will you collect and how?

Learning Analytics 3

So, let us start with the activity. Last week, we assumed that you had access to data from the classroom environment, about performance and attendance. We used that data to describe the

types of learning analytics. Let us consider you are the researcher or you are the teacher. So, you are going to collect data from the classroom environment. What data will you collect and how do you collect? Think about it. Pause this video, write down your answers and resume to continue.

Please consider what data and how you collect data? Suppose if you want to collect about students marks in the mid-sem, you will collect by conducting an exam. Similarly, list down all the data which you can collect in the classroom environment and write down how do you collect it. So, you might have answered performance because that is the most useful data or it is meaningful to understand the student's knowledge. So, let us see, performance is the most common data, everybody thinks about it. How do you collect the performance?

(Refer Slide Time: 6:24)

**Activity**

**Data Collection**

- Performance
  - Mid Term – Question and response to each question
  - Semester Exam – Score
  - Scores in Sub-topics
  - Course Project/Assignments
  - Presentations – Different scales like research, communication, lit survey
  - Open Book Assessments
  - Quiz

Learning Analytics 4

So, we can collect performance in the student's mid-sem exam or midterm exam where we have questions and responses to each question, if you can classify the question to a particular topic, that is your mapping question number 1, 2 to concept one, question number 3, 4 to concept two, you can have a richer data and understand which concept the students understood. Similarly, we will have a semester exam that is end-sem exam course.

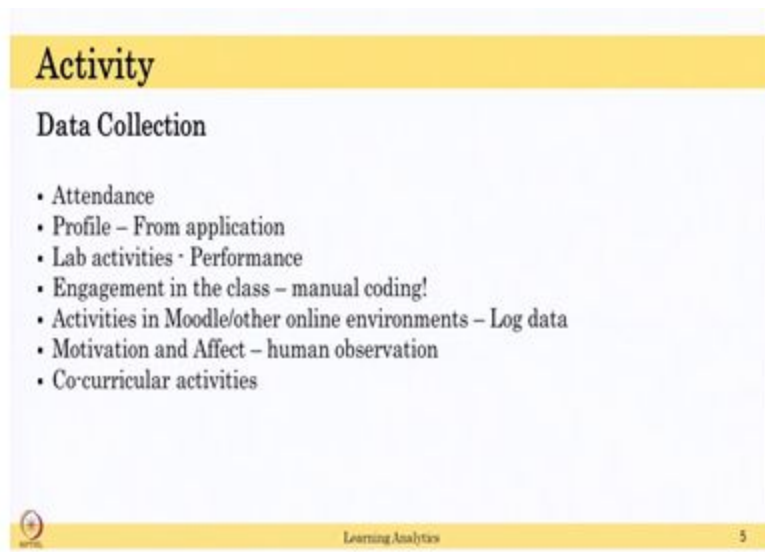
For example, you break down your whole course or whole chapter into multiple subtopics and you can conduct tests on each subtopic to understand students knowledge. Based on that you can

consider redoing or re-teaching something like that. Also you might have given the projects to the students, from the projects, the performance can be collected.

Or you might ask them to present some video or present some topic in the class. And you might create a rubrics to assess the student's communication skill, whether they did a proper literature review, they learnt some research skill, they are able to identify the gap in the literature. You can score on these dimensions and you can use this as performance data.

Or you can give open-book assessment like ask them to use books, solve the complex problem and you can assess their knowledge on how they apply concepts learned in the class in a timed manner. Or you can conduct quizzes, surprise quizzes something like that. These all ways you can collect performance data. So, that is not only data we can collect for classrooms.

(Refer Slide Time: 7:53)



The slide features a yellow header with the word "Activity" in a bold, black serif font. Below the header, the text "Data Collection" is written in a bold, black sans-serif font. A bulleted list follows, containing seven items: "Attendance", "Profile - From application", "Lab activities - Performance", "Engagement in the class - manual coding!", "Activities in Moodle/other online environments - Log data", "Motivation and Affect - human observation", and "Co-curricular activities". At the bottom of the slide, there is a yellow footer bar containing a small circular logo on the left, the text "Learning Analytics" in the center, and the number "5" on the right.

**Activity**

**Data Collection**

- Attendance
- Profile - From application
- Lab activities - Performance
- Engagement in the class - manual coding!
- Activities in Moodle/other online environments - Log data
- Motivation and Affect - human observation
- Co-curricular activities

Learning Analytics 5

We can collect student's attendance. That is simple, you can collect the student's attendance by marking their attendance. And you might have the students profile information and background information such as students which year they are in? Which department or are they from which kind of school or the family background. All the information you might collect it from the admin department. Or whatever data you can collect, you can collect those data from the students.

Also, students might have a corresponding lab activity. That data also can be useful to predict the student's performance. So, you can collect data from the student's lab activities. Also, you can collect students engagement in the class by observing the students engagement by coding it in a sheet or manually coding it. If you want to use some web camera and record the student's engagement in the class and post-class you can code them manually or use some software to code it.

Also, it is like the student's activities in the Moodle or other online environments like LMS or library systems. This we can collect from the log data of the system. From the log data, you can see what are the activities, how many times they log into modal or how many times they download or access the particular course material, something like that.

You can also collect students motivation and affect using human observation, i.e. learners centric emotions such as boredom, confusion, frustration by using human observers in the class. Which you can do in a real class or in a live environment or you can record the student's facial expressions using a web camera and you can sit down and code then after the class.

The camera is a bit tricky because if you have a large classroom, you may not able to capture all the student's facial expressions, so, it is better to use human observation in a live classroom. Also, there are co-curricular activities, which students might be participating in some events related to the course or they might be taking an extra course in MOOCs or something like that.

You can use this data also to understand the students learning process, also to improve your teaching processes. So, we said that we can collect a lot of data. The question is why we have to collect this data from the classroom? So, it is good that we have access to a lot of data, we can create a nice database of all this information of all the students in the class. But why we have to do? So, the main purpose is that, in the first week of this course, we talked about descriptive analytics, diagnostics and predictive analytics. So, you have to understand why we have to collect this data and how this data can be used to predict something so that you can improve your teaching-learning process.

(Refer Slide Time: 10:42)

## What to Predict!

- Why we have to collect these data from classroom?
- How it can be used to improve the teaching-learning process
  - Do you want to predict the students' performance in the final exam using their behaviour in class?



So, do you want to predict the student's performance in the final exam using their behaviours in the class? The behaviours which we discussed in the last slide. Or do you want to predict which student will do better in the mid-sem exam or do you want to predict which student will do well in the quizzes or you want to understand which student is struggling in the class on which topic so that you can teach him better or teach her better. So, this kind of research goal is on up. So, you have to set the research questions. Then you collect data in order to find that. So, let us move on to the other type of learning environment- MOOCs.

(Refer Slide Time: 11:26)



## MOOC

- Massive open online courses
- Students can access the course content from anywhere
- NPTEL – Swayam



So, MOOC is massive, open online courses. This course on Swayam or NPTEL which you are learning is actually a MOOC. So, here students can access the course content from anywhere. So, MOOC is course on Swayam or NPTEL kind of platform. So, now you know what is MOOC. So, consider you are the course administrator in MOOC or you have access to the MOOC software and you know how to collect data, you have a team of people to collect data, whatever you want.

(Refer Slide Time: 11:59)

### Activity:

#### Data Collection in MOOC

- You are taking this course in Massive online open course (MOOC) platform.
- If you are a course administrator in MOOC, what data you will collect about the learners?



If you are The MOOC administrator, course administrators, what data you want to collect from MOOC from the students participating in your course. That is, what kind of data you want to collect from the student's interaction with the MOOC in your course. So, please pause this video, write down the answers, also write down how do you collect this data? Not just what data, like how do you collect this data from the log file of the MOOC. And after writing it down, resume the video to continue.

(Refer Slide Time: 12:29)

The slide is titled "Activity" and "Data Collection in MOOC". It lists several data points to be collected:

- Timestamp of each event/action
- Learner ID, Session ID, IPAddress
- Pages viewed
- Discussion forum
  - Comment – delete, reply, upvote
  - Thread – create, unfollow, delete, reply, update, visit
  - Forum search, follow a user
- Navigation
- Behaviours in Video – play, pause, seek, speed change, transcript

At the bottom of the slide, there is a blue rounded rectangle containing the text: "Collect the learner's interaction with the system – Clickstream data". The slide footer includes a logo on the left, the text "Learning Analytics" in the center, and the number "9" on the right.

The basic and very important data in the learning environment or be it a classroom or MOOC or any other type is the timestamp. In a classroom, it is not possible to record a timestamp in a very accurate level, but atleast the date, time, the class section is good. But in an online environments like MOOC or Tele, we should record the timestamp of each action or each activity students do in the MOOC.

Other than that, you also collect students learner ID, session ID, IP address. Learner ID is each student will have a unique ID and session ID is that, in the same MOOC, a student might be logging it multiple times because the course run for eight weeks or twelve weeks, so the student has to log in multiple times in every week. So that we have to know the session ID. And also the IP address is useful to know the location where the student is accessing the data. That might be useful to do some adaptiveness or provide some feedback to the students.

Let us consider you want to understand the student's page view behaviour like what are the pages has students viewed in this course. Suppose you have a MOOC which has a lot of content in a PDF and also we have a lot of videos and you have a discussion forum. What is the PDF content the student is reading or which pages or which menu she is spending more on it. So, we need to understand what are the pages viewed.

So, how much time he spends on each page will be obtained from the time stamp data you collected. Suppose you collected a data saying that the student was on page 1 from time 10 AM to 10.2 AM, then you know that students spent 2 minutes on page 1, something like that. So, we need a timestamp data and what page they viewed in order to generate this data.

In discussion forum very useful data is present like has a student commented, has he deleted a comment, replied to some comment or he supported the comment or he created a comment, has he started a thread. Delete, unfollow, reply, update a lot of activity is possible within the thread. Also in forum search if the student is following some user or replying to the same user multiple times.

So, this kind of information can be obtained from the forum data, also the navigation information. For example, a student is navigating from one page to another page or the student will be watching the videos immediately going to answer the assignment questions. Or after assignment questions are going back to watch a video in a particular space, which minutes he is watching. So, is he watching videos to answer the questions all this information is possible to capture in the MOOC. That is called navigation.

Also in a video behaviour like are they playing or when they are pausing the video are they skipping the video from one particular place to another place in the video. Or they are changing the speed watching the video in 1.5 x or watching the video in 0.75 x or they are looking at the transcript, all these kind of information also can be captured from the behaviours in video watching. So, all this information can be captured in MOOC. So, simply the idea is that please collect all the learner's interaction with the system that is called clickstream data.

Wherever students are clicking buttons using the mouse or your keypad clicking buttons, typing all the data, you just capture it. Then you come up with the features which can be used from this

data so, that you can predict the student's performance or predict which students are going to dropout, based on this data one can predict whatever research question is. So, simply collect the learner's interaction data using clickstream data capture. This is the one type of data format which you use for Edx course but that raw data has a different format.

(Refer Slide Time: 16:41)



**Data from MOOC**

**Raw Data**

- {"username": "XXX", "event\_source": "browser", "name": "seek\_video", "accept\_language": "en-US,en;q=0.9", "time": "2018-05-15T11:27:13.618189+00:00", ... "context": {"user\_id": "9583xx", "org\_id": "IITBombayX", "course\_id": "course-v1:IITBombayX+...", "ip": "xx.xx.64.13", "event": {"code": "wvF9OwAdCxA", "new\_time": "557", "old\_time": "625.9213540286103", "duration": "832.68", "type": "onSlideSeek", "id": "f5238968f3814cd19ec97ea710a37e8a"}, "event\_type": "seek\_video"}

Learning Analytics 10

Look at this raw data, this raw data says there is user name (we hid the user name), browser, the action name you know it is called seek video you know I told you what is seek video. Seek video is moving video from one particular time to another time. So, at some point of time he was watching the video at 1 minute now he seeks the video to the third minute. So, the time he was watching and the new time also should be recorded, what is the old-time and new time and even type of seek video. So, this information can be captured from this log data. So, this is of general format for log data (it's one type of format and is most usually used format).

(Refer Slide Time: 17:14)

## MOOC Raw data example

```
• {"username": "XXX", "event_source": "browser", "name":  
  "textbook.pdf.page.scrolled", "accept_language": "en-us", "time": "2018-05-  
15T12:14:13.955573+00:00", "page": "https://courses.edx.org/...", "host":  
  "courses.edx.org", "...Introduction_to_Software_Engineering_IIT_Bombay.pdf",  
  "context": {"user_id": 4244xxx, "org_id": "IITBombayX", "course_id": "course-  
v1:IITBombayX+CS101.1x+1T2018", "path": "/event/1", "ip": "xxx.yyy.164.3",  
  "event": {"chapter": "Preamble_IIT_Bombay.pdf", "direction": "up",  
  "page": 3, "name": "textbook.pdf.page.scrolled"}, "event_type":  
  "textbook.pdf.page.scrolled"}
```



Let us see this data, can you take a minute pause the video and identify, try to identify what is this log data means what action this student is doing? So, here also student name is hidden, the action. even name of textbook pdf page scrolled. So, the student is scrolling the pdf page. Which page there is a pdf called “Preamble IIT Bombayx” pdf and the direction is upward scrolling in the mouse you can scroll upwards. So, actually, he is watching and reading the page.

Based on the the OS use, Mac OS or Windows OS, you can say whether the student is watching, the reading the page, going to next page, going back to the previous page. So, this information can be captured from this log data. So, this is another type of clickstream data we are capturing like it is not a clickstream as every action a student does like scrolling also on the page is captured.

So we call it as a trace data so there are two type of data, clickstream and trace data. In general, all the platforms which allow MOOC will help you to collect this data. Unfortunately, NPTEL will not record all this “user information” because the number of users in NPTEL is really huge and we do not have space to keep all the data in the server. So, we might be coming of the new projects to collect all the data. But if you see the courses offered in Edx course there they might record all this data.

So, I was giving you an example that you can collect all the information from this information I want you to create a log features in one specific format. The format is, so, you collect raw data.

So, you should convert the raw data into actions or events by writing some scripts like a Python script, R script.

(Refer Slide Time: 19:17)

**Data Preprocessing**

- Raw Data should be converted into actions/events - Scripts

What is actions or events?

- Similar to the log data we listed in classroom environment we need to identify the features from these raw data. For example
  - Number of pages viewed in X minutes
  - Average time on a page
  - Read - long, short? < 2
  - Number of comment in forum < 5 sec - Read Short

Domain expertise is important to construct these features 5 sec - 1 min Read

Tool: <https://www.featuretools.com/> > 1 min - Read long

> 5 min

Learning Analytics 12

So, the raw data you saw in the previous slides should be used to convert this data into particular actions or events. So, what is that actions or events? Similar to the log data we listed in the classroom environment, we also need to identify the features from these raw data. For example, you want to know the number of pages viewed in the last 10 minutes. How do you do that? You have to write a script from the log data, that data you captured to identify how many pages viewed?

$$Pages_{viewed\ in\ last\ 10\ min} = Pages_{viewed\ in\ time\ x} - Pages_{viewed\ in\ time\ (x-10)}$$

all the page views should be counted, that number should be listed. That is the feature.

Why do you want to know the number of pages? We do not know. That is domain expertise is required to understand which features will be useful to predict the student's knowledge or students performance. For example, the average time the students spend on a page number 3 so, you have to capture whenever a student is reading the page number 3 in all the sessions, and average time has to be computed from that.

Or if a student is reading a page, you might classify them as a read long or read short. Why? For example, I opened a page number 1. I spend only 1 second. Do you consider that as a read? May not be. So, you might expect a student has to spend at least certain times say, 5 seconds to read at least 1 line in that particular page. So, you can come up with the threshold to classify the read long and short.

For example, if the student is watching the page for 5 seconds, you can classify it as a “read short”. If the student is watching the page, from 5 seconds in this particular page to say, 1 minute, you can consider it as a “read”. The threshold is based on your knowledge on what is the content, whether the content has a lot of pictures and mathematical equations, it might take more than 10 minutes also.

So, this is based on your knowledge you applied here. Then you might say it is read long, a student is reading this particular page for long time and if a student on the page spends more than 5 minutes, for example, you can ignore that content. A student might have probably opened the particular page and he might have left, he might have moved to another tab, he is doing some other activity, watching some other videos. So, you should ignore it.

Also if the student is spending less than say 2 seconds or something like that, you can ignore it. So, if you have timestamp information and you know what data student is watching, it will help you to capture these kinds of information- “read short”, “read long”. Why this data is useful? You can say, there are some students who are not reading at all, they might be attempting the quiz or assignment questions, they are not able to solve it.

Though you know the reason because they did not read. Or some students will be reading a lot of time. They may be reading, reading, reading, they are not taking any assessment, we also can send a message saying that, Hey, why cannot you go and take a quiz. Or some students doing good, they read long and they take the assessment, they have high probability of passing the exam.

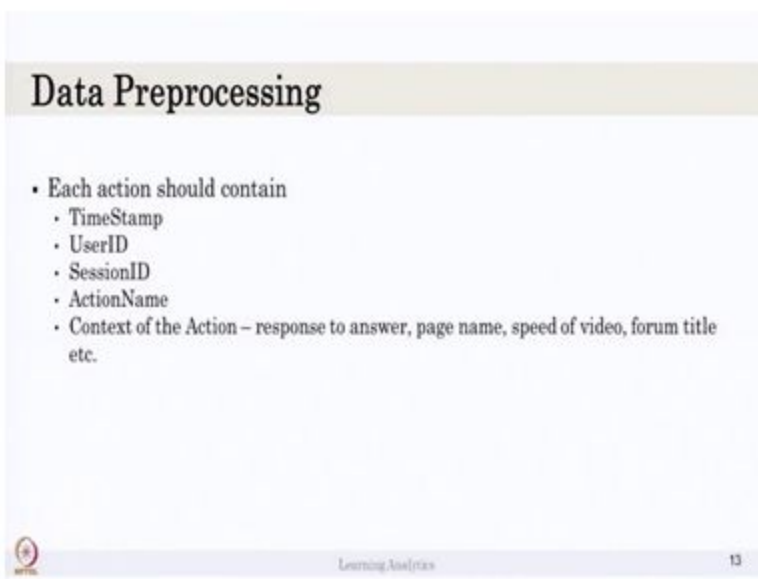
So, you know what is the student’s behaviour from the log data. So, also you can come up with a number of comments in a forums (discussion), which you can capture. So, there is a lot of data you can come up with. So, what is that, how do you come up with these kinds of features? That

is through Domain expertise. So, the feature construction is not just, I can capture all the raw file, I can use that raw file to predict something, no.

Instead, it is also about applying your domain knowledge or your expertise. That is where you need the domain expertise in education or domain expert in teaching experiences. Apply that knowledge to come up with the list of features. To get these features from the raw file, you might need a knowledge on writing scripts like a Python or R. That is what I said about you might need a small bit of programming knowledge. But for this course, we will give you all the features extracted from the raw file.

So, there is a tool which is used very heavily in, in industries for feature, is called feature tool. Please check that tool. This course is not focused on training that tool because that tool is not important to us because we extract the features based on our domain expertise. This feature tool will help you to construct more features if you have the knowledge of domain. So, if you do not have any knowledge on the domain, this feature tool also will not help you to create features. So, please check this tool called feature tool. This is used heavily in industries nowadays.

(Refer Slide Time: 24:12)



The slide is titled "Data Preprocessing" and contains a bulleted list of requirements for each action. The list includes: TimeStamp, UserID, SessionID, ActionName, and Context of the Action (with examples like response to answer, page name, speed of video, forum title, etc.). The slide also features a small logo in the bottom left corner and the text "Learning Analytics" and "13" in the bottom right corner.

- Each action should contain
  - TimeStamp
  - UserID
  - SessionID
  - ActionName
  - Context of the Action – response to answer, page name, speed of video, forum title etc.



So, I mentioned that there should be a specific format in which data should be stored. The format is a timestamp, user ID, session ID, action name. What is the action name? Action name may be reading or watching videos, taking quiz something like that, these are the action names. In each action, you might have a context.

The context may be in a reading page, what page he is reading, what page number he is in, what type of video he is watching, is he seeking the video or is he playing the video in a particular time, speed. All this information can be captured in “context of the actions”. So, the action name will come from your domain. For example, if I use MOOC as a domain, I know that there are 4 major actions in MOOC.

That is, video watching behaviour that is, play, pause, seek some kind of behaviour in the video watching. Interaction in the forum, commenting or creating a thread, other actions. Also in the reading behaviour, they might be reading some PDF or something like that. Also, they might be navigating some menu, or going from one tab to other. So, these are the 4 major actions I might have, 4 or 5 actions.

So, you have to come up with the actions and combine the time to create long or short actions or the actions can be repeated multiple times. You might get a “mult” kind of suffix to it. Then you can have the context of the action, where it is done? Like is he reading page number 3, is he reading, which video he is watching, what is the speed. That kind of information can be used to provide meaningful data collection and that can be used to predict something. So, what you want to predict, that is your question. You want to predict the student’s performance in the classroom, students performance in a particular course or who will dropout in next week, something like that.

(Refer Slide Time: 26:00)

## Data Preprocessing

- Each action should contain
  - TimeStamp
  - UserID
  - SessionID
  - ActionName
  - Context of the Action – response to answer, page name, speed of video, forum title etc.
- Learn about pre-processing in ML or Data Mining courses



So, also you have to learn about pre-processing in other courses like Machine Learning or Data Mining courses. This is your LxT where you can go and watch external course content. However, I recommend for an educational video, we do not need to do much, instead, you need to understand simple things like -if you have missing values, how to replace the missing values.

Some suggest missing values can be replaced with 0 or missing values can be replaced with the mean of other values but it depends on the missing data and also your domain knowledge. So, apply logically what should be replaced and first try to understand why the value is missing.

Also, I recommend you to normalize all the data to 0 to 1. For example, the performance score is measured is a scale of 0 to 100. But the number of upwards is measured in the scale of may be 0, 10, 2, 3. How do you compare these two scales in a single comparison? The Machine Learning algorithm might work if you do normalize it to 0 to 1. Some suggest doing standardization. So, think about it, normalization or standardization. Then you apply them based on your requirement.

(Refer Slide Time: 27:02)

## Summary

- Data collection in
  - Classroom environment
  - MOOC



So, in this video, we talked about data collection in a classroom environment also in a MOOC environment. I also talked about how to extract features from the log file. We will discuss that in detail, what are the features, how to extract features. I will show examples in the next video. Thank you.