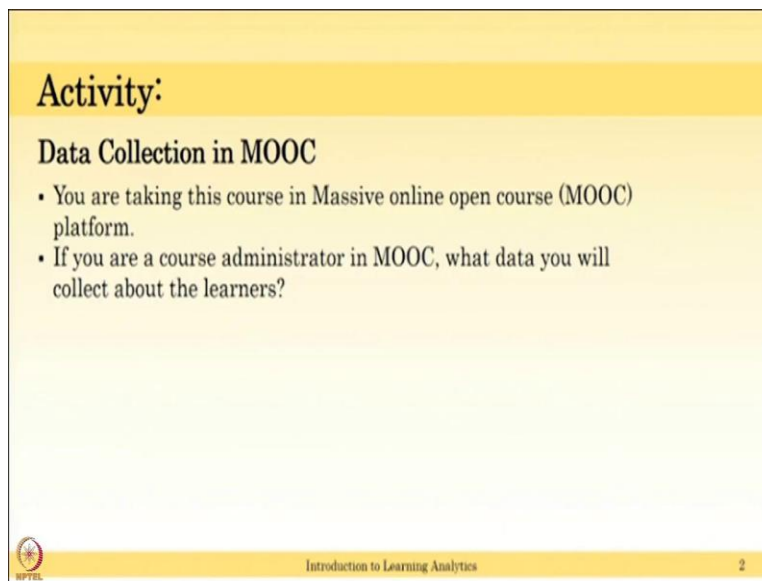


**Introduction to Learning Analytics**  
**Prof. Ramkumar Rajendran**  
**Interdisciplinary Programme in Educational Technology**  
**Indian Institute of Technology, Bombay**

**Lecture – 06**  
**Data Collection: MOOC**


(Refer Slide Time: 00:18)



**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?

 Introduction to Learning Analytics 2

In this learning dialogue, we will discuss about Data Collection in MOOC environment. You are taking this course in massive online open course platform that is MOOC. Assume that you are a course administrator in MOOC and you have access to all the programs, all the data in the MOOC and you can write the scripts which we need to collect data on the MOOC. If you have all the access and your administrative access and you can write code, what data you will collect about the learners in the MOOC? You can pause this video, write your answers, after completing this task, you can resume the video to continue.

(Refer Slide Time: 00:56)

## Activity

### Data Collection in MOOC

- Timestamp of each event/action
- Learner ID, Session ID, IPAddress
- Pages viewed
- Discussion
  - 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



Data collection in MOOC, there is a common parameters that is timestamp of each event or action, learner ID – who is the learner and session ID, the same learner can login to MOOC for multiple times in a day or over the time. So, session ID is important, IP address to know where the user is logging from. So, you can collect this information for all users across MOOC, these are very key information to identify which user which session and where he is from and the timestamp to tell fine grained data such that when the event occurred.

Apart from this you might have thought about you can collect the pages viewed, discussions, navigation, behaviors in the video. I would like to go each one in detail. If it's page viewed, we need to note down what page viewed and what is the time students spent on that particular page, what is a page title and where the student is reading. Depending on the MOOC platform, you can perform various activities in a discussion forum like you can comment, you can delete, reply, up vote, down vote or like the comment or in the thread you can create a thread, you can delete a thread, you can unfollow a thread, you can follow a thread or you can update or just miss it, you just watch the thread you read the thread, but you are not spending any time in it.

You have not created any comment, you have not followed anything, how do I know whether you visited the thread and read or not? If a time stamp tells that you are in the thread for more than several seconds say three seconds or five seconds you might be looking something in the thread or it is possible that you open the thread page and you go out to come back after five minutes that is also possible. So, that should be upper threshold limit also to consider a student is reading a thread or just looking at thread or creating something in thread. Also in this forum, the learner can search something or a particular user. So, these kind of activities should be logged in the MOOC.


Apart from this the student navigates in the MOOC from one page to other page this navigation behavior also can be logged. So, MOOCs contains basically three main things one is there is a lot of resources and videos and discussion forum. So, we need to log all the information as students interacting with the MOOC. In video behaviors you can collect more information such as play, pause, seek, speed change and transcript.

(Refer Slide Time: 03:37)

## 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" }
```

Introduction to Learning Analytics4

Let us look at the raw data of MOOC collected from IIT Bombay X course, you can see the raw data the user name is anonymized and the even source is browser and the event name is seek video. The learner is seeking the video and the timestamp tells when it is happened, also the user id is there, organization which is offering the course IIT Bombay X and course id. When we seek information, the context we need is from time which time to which time the students seek the video, the new time is 0.557, the old time is something else.

So, the students seek the video form 625.9 to 0.557. So, you used the sleek onSlideSeek. So, if particular event or particular action is seek video, we need to know couple of things that is which time the students seek the video, which particular video he seeked in and what is the starting and ending duration in the seek video.

(Refer Slide Time: 04:56)

## 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"}, "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 look at the other example. In this example, the event source is browser, the test book is the event name is actually scrolling; the student is scrolling the reading resources. It has a detail of which page; where it is from again the student id and everything. Apart from this, it has other important information that is, which direction he is scrolling, is he scrolling downward or upward.

We may not need such a rich data, but it is good to log all this data, to understand what student would have done in the particular time, why the student is not performing well in this particular time. So, these two examples, the example shown in the collecting data in the classroom environment also collecting data in MOOC environment to show that, you have to think what kind of data we can collect in a different environment. You collected raw data or you ask the MOOC platform to provide a raw data to you. Most of the MOOC platforms like IIT Bombay X or coursera or edx can provide a log data to you, if you're offering a course in the platform.

(Refer Slide Time: 06:08)

## Data Preprocessing

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



Introduction to Learning Analytics



The raw data should be converted into actions and events. So, the raw data can be saying into, what is the student ID, session ID, what is the action what are the important parameters we need in that action. So, you can develop some scripts maybe in any programming language to convert this log data into a readable data in a CSV format. Each action should contain timestamp, userID, sessionID also what is the action name, is a seek video or is it a scrolling, is it a discussion forum creation, is it a deletion of discussion forum and the context of the action.

For example, in seek video we saw start time, end time, the duration of seek or if it is a question answer response to the answer, what answer is a student enter. If it is a reading page, what is the page name he is reading, speed of the video if he is watching the video. If it is in forum, what is the title of the forum, these kind of information should be captured in the context of the action.

If you want to learn about preprocessing that after collecting data, after writing a script to extract these actions, if you want to apply preprocessing there are lot of videos and lectures about preprocessing in ML and data mining courses. Teaching preprocessing is beyond the scope of this course.