Developing standards and systems for MOOC data science

Kalyan Veeramachaneni, Franck Dernoncourt
Colin Taylor, Zach Pardos, Una-May O’Reilly

Any Scale learning for All
CSAIL, MIT
Overview

• Background and motivation
  • What did the data looked like?
  • Where are the bottlenecks?

• Data science @ scale
  • Organize- MOOCdb
  • Create multiple views- MOOC En Images
  • Provide APIs- MOOCdb Access

• Why standardize?
What did the data look like?
What did the data look like?

- MOOC
- Server logs JSON lines
- Browser logs JSON lines
What did the data look like?

- SQL Dump
- Production server
- Forums data
- Server logs
  - JSON lines
- Browser logs
  - JSON lines
What did the data look like?

- Course information XML files
- Course website
- SQL Dump Production server
- Forums data
- Server logs JSON lines
- Browser logs JSON lines
What did the data look like?

- Course information: XML files
- Course website
- SQL Dump: Production server
- Forums data
- Server logs: JSON lines
- Browser logs: JSON lines
- Emails
- Surveys
Example Research Question

RQ1: How does amount of time spent on the video correlate to performance on the homework?

What is the time period between two consecutive homeworks?
Identify what are the video urls that correspond to these homeworks?
Example Research Question

RQ1: How does amount of time spent on the video correlate to performance on the homework?

What is the time period between two consecutive homeworks?

Identify what are the video urls that correspond to these homeworks?

Calculate the time spent on these urls by the user?
Example Research Question

RQ1: How does amount of time spent on the video correlate to performance on the homework?

What is the time period between two consecutive homeworks?

Identify what are the video urls that correspond to these homeworks?

Calculate the time spent on these urls by the user?

What answers did the user submit during this period?

What are the correct answers for those problems?
Example Research Question

RQ1: How does amount of time spent on the video correlate to performance on the homework?

What is the time period between two consecutive homeworks?

Identify what are the video urls that correspond to these homeworks?

Calculate the time spent on these urls by the user?

What answers did the user submit during this period?

What are the correct answers for those problems?

How many did the student get right?
Example Research Question

RQ1: How does amount of time spent on the video correlate to performance on the homework?

- What is the time period between two consecutive homeworks?
- Identify what are the video urls that correspond to these homeworks?
- Calculate the time spent on these urls by the user?
- What answers did the user submit during this period?
- What are the correct answers for those problems?
- How many did the student get right?
- Run corr function in MATLAB
Where are the bottlenecks?

A typical process

- Preprocess
- Assemble covariates
- Build models
  - Analyze
Where are the bottlenecks?

What is now?

Preprocess $\rightarrow$ Assemble covariates $\rightarrow$ Build models

What we desire?

Preprocess $\rightarrow$ Assemble covariates $\rightarrow$ Build models

sizes represent the time spent in doing the activity
How about we organize the data using a concise schema that we will always adhere to?

Then all the scripts we write can be reused.
How about we organize the data using a concise schema that we will always adhere to?

Then all the scripts we write can be reused

**Challenge**: Generalizable, loss-less
Steps to Data science @ scale

**Step 1:** Develop a generalizable, loss-less schema for the data

**Step 2:** Define and design APIs to extract multiple views of data

**Step 3:** Design user friendly APIs

**Step 4:** Design a platform where users can contribute and share the scripts
Step 1: Develop a generalizable, loss-less schema

- Students interact with the system in the following ways
  - Observe
  - Submit
  - Collaborate
  - Give feedback
Step 1: Develop a generalizable, loss-less schema

The observing mode

- User
  - id
  - name
  - gender
  - age
  - country
  - ip
  - timezone_offset
  - final_grade

- Events
  - event_id
  - resource_id
  - user_id
  - duration
  - timestamp

- Resources
  - resource_id
  - name
  - url
  - type_id

- Resource type
  - type_id
  - type_name
Step 1: Develop a generalizable, loss-less schema

The observing mode

Is this generalizable?
Step 1: Develop a generalizable, loss-less schema

The observing mode

Is this generalizable? yes
Step 1: Develop a generalizable, loss-less schema

The observing mode

Is this generalizable? yes
Is this loss-less?
Step 1: Develop a generalizable, loss-less schema

The observing mode

Is this generalizable? yes
Is this loss-less? yes
Step 1: Develop a generalizable, loss-less schema

MOOCdb
Step 2: Define and design APIs to extract multiple views of data

MOOC En Images

Student cohorts

by grade
by access patterns
by certificate winners
custom cohort

by week
by deadline
by midterm
by homework
by modules
custom time points

by country
by IP
by city

Space

Time
Step 2: Define and design APIs to extract multiple views of data

MOOC En Images

Average number of submissions per country for students who got a certificate (countries with less than 50 students were removed)
Step 2: Define and design APIs to extract multiple views of data

MOOC En Images

Number of submissions per day

Day
Step 3: Design user friendly APIs

MOOCdb - Access

MATLAB

```matlab
%% Running queries

sql = [' SELECT observed_events.observed_event_duration ' ...
    ' FROM moocdb.observed_events AS observed_events ' ...
    ' WHERE observed_events.observed_event_duration < 200' ...
    ' LIMIT 100000; '];

cursor = exec(connection,sql);
a = fetch(cursor);
data = cell2mat(a.Data);

boxplot(data)
title('Distribution of the duration of observed events')
ylabel('Duration (in seconds)')
print('-dpng','-r300', ['observed_events_duration_boxplot'])
saveas(gcf, 'observed_events_duration_boxplot', 'fig')

plot(sort(data))
title('Distribution of the duration of observed events')
ylabel('Duration (in seconds)')
print('-dpng','-r300', ['observed_events_duration_plot'])
saveas(gcf, 'observed_events_duration_plot', 'fig')
```

Python

```python
import google_charts_wrapper

def main():
    """
    This function plots the number of new registered students every day from 2012-02-13
    """
    sql = ""
    -- Takes 1 second to execute
    SELECT DATEDIFF(users.user_joined_timestamp, '2012-02-13 00:00:01') AS 'Day',
    COUNT(*) AS 'Number of new registered students'
    FROM moocdb.users AS users
    GROUP BY 'Day'
    HAVING 'Day' >= 0
    ORDER BY 'Day' ASC
    
    options = google_charts_wrapper.options()
    options.set_data(google_charts_wrapper.get_data(sql))
    options.set_chart_type("area_chart")
    options.set_chart_title("New registered students every day from February 13, 2012")
    options.set_height(500)
    options.set_width(900)
    options.set_page_title("New registered students every day from February 13, 2012")
    options.set_v_axis("New registered students", titleTextStyle={color: 'blue'})
    options.set_output_file("/output/users_join_date.html")
    print options.get_data()
    google_charts_wrapper.generate_html(options)

if __name__ == "__main__":
    main()
```
Step 4: Design a platform where community can share scripts

- MOOC En Images
- Who talked to who?
- IRT
- BKT
- Mobility

User friendly APIs
- MATLAB
- Python
- R

Query Optimizers

Scientific Community
- Analysts
- Database experts
- Privacy experts

MOOCdb

Scripts Archive
Leading to standardization

MOOC Curators → Database → MOOCdb

Insert logs into database
Transform Database into standard format

MOOCdb

Database

Shared Standardized schema

Scientific Community

Analysts
Database experts
Privacy experts

Public access Platform

Scripts archive
Sql scripts

Crowd

Feedback to MOOC Organizers, sharing of results, predictions descriptive analytics.

MOOCdb + Sql scripts =

MIT
Benefits of standardization

• Reduction in time to analyze (TTA)
• Publicly available scripts to create data views
• Unified representation to DB/Privacy experts
• Crowd sourcing of feature engineering
• Crowd sourcing of data analytics
• Elimination of entry barriers
Thank you!

Look for MOOC En Images, updates on MOOCdb and open release of all the tools.