Surveying the MOOC Data Set Universe

James J. Lohse
College of Computing
Georgia Institute of Technology
Atlanta, GA, USA
jiml@gatech.edu

Christine A. McManus
College of Computing
Georgia Institute of Technology
Atlanta, GA, USA
christine.mcmanus@gatech.edu

David A. Joyner
College of Computing
Georgia Institute of Technology
Atlanta, GA, USA
david.joyner@gatech.edu

***Abstract*— This paper is a survey of the availability of open data sets generated from Massively Open Online Courses (MOOCs). This log data allows researchers to analyze and predict student performance. Often, the goal of the analysis is to focus on at-risk students who are not likely to finish a course. There is a growing gap between the average researcher (who does not have access to proprietary data) and the ready availability of data sets for analysis. Most research papers studying and predicting student performance in MOOCs are done on proprietary data sets that are not anonymized (de-identified) or released for general study. There are no standardized tools that provide a gateway to access usable data sets; instead, the researcher must navigate a maze of sites with different data structures and varying data access policies. To our knowledge, no open data sets are being produced, and have not been since 2016. The authors survey the history of MOOC data sharing, identify the few available open data sets, and discuss a path forward to increase the reproducibility of MOOC research.**

Keywords—MOOC, weblog, analysis, edx2bigquery, Google BigQuery, anonymized data set, de-identification, MOOCdb, Moodle, Educational Data Mining, Learning Analytics, Learning at Scale, Limeade

I. Introduction

Massively Open Online Courses (MOOCs) have not met their promise. After peaking in 2015, enrollment has dropped. The goals of reaching and educating citizens in less developed nations have not been met. Completion rates are as low as 7% [1]. Many methods have been used to predict the at-risk status of students and encourage them to complete the course. Further effort has been spent to attract MOOC students so they return for the second year of classes. An underlying impediment to prediction of at-risk students is access to open, de-identified data sets. Unless a researcher has access to their institution or company's data set, de-identification is the primary barrier to access. The basic scientific concept of producing reproducible research is missing from most studies. This paper identifies a knowledge/content gap in navigating through the options for MOOC-related data. The authors found that many attempts to standardize and share MOOC data have met dead-ends and been terminated.

II. A Brief History of Moocs

To understand why much of this paper focuses on efforts in the first half of this decade (and not so much the second half), it's worth understanding the history of MOOCs. The first known use of the term "massively open online course" was in 2008 when Siemens and Downes ran a course online, "Connectivism and Connective Knowledge". The course attracted over 2,000 students [2]. The first genuinely massive MOOC was in 2012 when Stanford professor Sebastian Thrun posted videos of his Artificial Intelligence course on the Internet. One hundred sixty thousand people from 190 countries signed up to view the courses [3].

Within months, Thrun founded Udacity, the platform where GA Tech offers the OMS-CS degree, which is the first Large Internet-Mediated Accredited Degree (Limeade) program [4]. Roughly in the same timeframe that Thrun founded Udacity, Harvard and MIT created the edX platform, eventually attracting sixty universities to provide content. Stanford spun off another MOOC platform contribution – Professors Andrew Ng and Daphne Koller launched Coursera, gaining over one hundred contributing institutions. Soon after this, Coursera worked with a group at MIT to create MOOCdb[1], which will be presented in-depth in a later section.

MOOCs were heralded as the solution to democratizing post-secondary education. They were supposed to level the playing field by attracting students worldwide, such as those in less developed countries and from at-risk minority populations. Unfortunately, MOOC enrollment peaked in the 2015-2016 school year and primarily attracted students from developed nations [1]. MOOCs suffer from the "twin 7%" measurements – only 7% of those who enroll in a MOOC finish it, and only 7% return for online classes during a second year. While it is not reasonable to expect every student to finish, the 7% figure can undoubtedly be improved. This potential improvement is the focus of virtually all MOOC-related EDM and LA research & analysis efforts.

III. Problem: A Lack of Anonymized Mooc Data

First, to understand the terminology, we define "open data set" and "proprietary data set." An open data set is an export of weblog data from a MOOC that provides a schema and can be downloaded by anyone. The "schema" can be a document summarizing the tables and fields in the exported data. To allow this, as discussed below, data must be de-identified and released to the general public. A proprietary data set is one which still contains personally identifiable information (PII) or requires special permission to access. Examples of proprietary data sets can be found by searching the term "MOOC" at the Harvard Dataverse website [5]. The search results turn up two data sets where the summaries contain language such as, "Data are currently only accessible to qualified reviewers." If there is a gatekeeper, and the data is not open for any outsider to conduct reproducible research, it is considered proprietary in this context.

Second, the term MOOC must be defined more narrowly. Limeades, for example, are not MOOCs. They are degree programs that count for credit and have teaching assistants to provide grading, feedback, and forum moderation. The key distinguishing factor between a MOOC and a Limeade course is how the course scales. Regardless of how many people sign up for a course, if an online course can scale without the addition of more human resources, it is appropriately called a MOOC. If a course, such as those in

[1] There are various capitalizations of MOOCdb, we have settled on this, used in the first MOOCdb paper [7]

978-1-7281-2549-7/19/$31.00 ©2019 IEEE 23-25 October 2019, Milwaukee, WI, USA
2019 IEEE Learning With MOOCS (LWMOOCS)

Limeades, requires the addition of more teaching assistants to grow the course membership, it is not massive [4]. The other distinguishing characteristic of a MOOC is that it is open to anyone with an email address, regardless of admission to a degree program.

A. Ethical Challenges in Sharing of Student Data

More and more research is being published using data sets for studying performance in MOOCs. The research papers this author has reviewed show that companies and universities are almost exclusively using their proprietary data sets.

One problem is that of de-identifying the data. When the data includes essay-style answers to tests and quizzes, the bar of preventing re-identification is even higher. Anonymization is just one factor keeping MOOC data sets from being more publicly available. A paper in the Journal of Learning Analytics [6] illustrates the main challenges in data set openness:

- Transparency: making students aware of data collection and potential sharing with third parties
- Anonymization of student data, especially essays
- Ownership of the data
- Accuracy of data and resulting analysis
- Security of student data per privacy laws

FERPA is the United States Family Educational Rights and Privacy Act law of 1974. FERPA governs the privacy protections of student data. It only applies to schools that receive funding from the US Department of Education. For private entities (Moodle, Coursera, Udacity) to release open data sets, there are the terms of service with their users, as well as evolving laws such as the EU General Data Protection Regulation (GDPR). It is quite easy to understand, given an evolving legal framework such as GDPR (it is challenging to understand the jurisdiction) that companies and universities that run MOOCS are hesitant to release any open data sets. If a student in the EU signs up for a MOOC offered by a US-based company, it appears that the constraints of GDPR must be followed. Any releases of open data sets must clear legal hurdles that represent significant barriers to data sharing.

B. Lack of Standard Data Formats – Call to Action

Another problem with the analysis of open data sets is the lack of a standardized data schema across MOOCs. A 2013 paper issues a call to action to create data extraction tools that would bypass the data-sharing problem [7]:

> Our contention is that the MOOC data mining community - from all branches of educational research, should act immediately to engage in consensus driven discussions toward a means of standardizing data schema and building technology enablers for collaborating on data science via sharing scripts, [and] results in a practical, directly comparable and reproducible way. It is important to take initial steps now.

One barrier to data sharing that Veeramachaneni et al. identify is the amount of time needed to clean the data. The paper estimates that data cleaning takes 70% of the time for analysis, much more than for model building. The paper notes that even different MOOC courses within the same organization suffer from inconsistent, ad-hoc approaches to the structure of data.

To solve this, those researchers created an approach called MOOCdb. The goal was not so much as to make data sharing possible to create data extraction tools to work with a standardized, community-sourced database schema. MOOC developers would be expected to create their courses in such a way that the log data could be exported to match the MOOCdb data standards.

The MOOCdb project was run at MIT with Kalyan Veeramachaneni as the most consistent author through a series of papers expanding the concept [7][8][9][10][11]. This project still exists today as the MOOC Learner Project, maintained by the MIT Alfa group. The project website with Github links is at http://mooclearnerproject.csail.mit.edu/.

IV. MOOC STUDENT PREDICTION USES CLOSED DATA

It is a bit ironic that MOOCs, a genre of courses embraced by academics, universities, and once-sizzling startups for their openness, have not provided the same openness in the underlying tracking data.

An intriguing paper, *GritNet: Student Performance Prediction with Deep Learning* [12], was published by Coursera and uses proprietary data. In another paper from 2018, *Predicting Academic Performance: a Systematic Literature Review* [13], the authors write, "we found almost no data sets that have been published for wider use." A reference in their paper, *Educational Data Mining and Learning Analytics in Programming: Literature Review and Case Studies* [14], similarly noted the scarcity of open data sets for use by researchers outside the institution or company that created them:

> Additionally, we found that open data sets are exceedingly rare in the literature we reviewed. Only four studies (5%) were based on open data sets. Three of these studies were using the Blackbox dataset, while the last provided open tools and detailed data within the paper.

MOOCs have come a long way since Sebastian Thrun left Stanford to found Udacity, but the ability of researchers to reproduce results (without explicit permission) has not.

V. THE SEARCH FOR ANONYMIZED DATA SETS

While our initial hopes about the availability of anonymized MOOC data were mostly misplaced, there are some exceptions and variations on the theme. In this section, the false starts are documented:

A. Pittsburgh Science of Learning Center (PSLC)

From the book *Learning Analytics,* a chapter titled *Educational Data Mining and Learning Analytics* [15], the authors Baker and Siemens note:

> considerable quantities of data are now available to scientific researchers through

> public archives like the Pittsburgh Science of Learning Center DataShop

Unfortunately, very little (arguably none) of this is MOOC data. It is largely drawn from more traditional e-learning tools such as digital assistants, intelligent tutoring, and other data sources that do not capture the entire interaction of students with a learning system. MOOC weblog activity is more robust than most other sources as it represents a student's entire interaction with a course. At PSLC, a significant amount of the student's interaction with the course is outside the data sets being presented.

There is a GA Tech Intro to Psychology "MOOC" with data posted on PSLC. The problem with this data is that it only represents 288 users for the 2018 data. There is not a clearly defined cutoff between a MOOC and a "BOOC" (Big, not Massive), but one source claims a BOOC is 500 students or less [16]. However, as defined for this paper, we distinguish "big" from "massive" in this way:

> … a 'massive' course is one in which the human resource allocation is constant regardless of enrollment, while a 'large' course is one in which the human resource allocation scales linearly with enrollment and where there is no cap to this growth (e.g., lecture hall size) [4].

The PSLC/GA Tech Psychology data for 2013 looks more promising, starting with 5,615 students. Unfortunately, it appears the corresponding data set "Students That Finished" only represents 281 students. Studying a MOOC where only 5% (281/5615) of students finished is probably not the best data set for understanding desired MOOC student behavior. Thus, PSLC does not represent a good source of open MOOC data sets, for the reasons cited and for one more reason: one must use a "Request Access" button and be logged in to the site, so PSLC still acts as a gatekeeper, even if only in a minimal fashion. This does not meet the definition of open data set in this context.

B. edX2bigquery – Google BigQuery

Another variation on the MOOC theme is the advent of blended MOOCs. A blended MOOC uses a MOOC platform to provide an online class environment that counts for credit at a traditional university. In more recent iterations, the credit has been for certificates or Nanodegrees on platforms such as Coursera and Udacity. A 2016 paper titled *Google BigQuery for Education: Framework for Parsing and Analyzing edX MOOC Data* by Lopez et al. talks about moving the Person-Course data set pipeline from MongoDB to Google BigQuery [17].

Initially, HarvardX and MITx were using in-house tools and data storage to process "nightly" MOOC data exports on an in-house cluster using MongoDB. The massive growth of data required processing times such that:

> After roughly two years of MOOC data, processing times soon exceeded the limits of reasonable wait time for aggregating data, with some aggregate data taking longer than 48 hours to generate.

HarvardX and MITx migrated their data processing pipeline to use a set of python tools, edx2bigquery [2]. Unfortunately, as noted before, this did not lead to a current, open data set. There is a HarvardX/MITx edX-based Person-Course data set synced nightly to Google BigQuery [17]. Sadly, it is not currently in the Google Cloud Public Data Marketplace. No MOOC data appears there, nor is there a specific category for education. It will be seen in a later section that some 2013-2014 Person-Course data sets are available for analysis, and a paper was published using those data sets.

C. Stanford Datastage

Another once-promising project appears to be on hold. The Stanford Datastage website initially looks encouraging, but there is a line that says, "If you are interested in studying the data, please fill out our request form." Unfortunately, that request form page states, "As of 1 September 2018, our data sharing activity will be suspended." There is another page listing Coursera data sets[3] that states, "Coursera courses can unfortunately currently only be shared with researchers within Stanford University".

The interesting part about Datastage is that Stanford was one of the original participants in the MOOCdb project. As of this writing, the only open data set this author can find from Stanford is a MOOCPosts dataset [4]. The data is described as a "dataset [that] contains 29,604 anonymized learner forum posts from eleven Stanford University public online classes." This is a proper open data set but does not represent all the activity in a MOOC, just the forum activity.

D. LearnSphere

LearnSphere is a website run by Carnegie-Mellon University, Stanford, MIT and the University of Memphis to centralize access to and analytics of various data sources. LearnSphere is a front end to MOOCdb, DataStage, PSLC Datashop, and two other projects called DiscourseDB and Tigris. There is a graphical front-end to manage a data analysis workflow. LearnSphere also uses PSLC as a data source for non-MOOC data sets.

Unfortunately, this does not represent a source of open data sets. There are workflows, comprised of scripts written by researchers, that can be run on proprietary data sets in LearnSphere. Access to those data sets is strictly limited to qualified researchers, and those researchers must agree not to de-identify those data sets. If a researcher has access to a Spark-based Coursera data dump, they can load it on LearnSphere – this does not satisfy the need for reproducible research. Additionally, the references on the website to MOOCdb and DataStage appear to be out of date, given that MOOCdb has a new name (more on this soon) and DataStage is no longer sharing data through its site.

E. DiscourseDB

According to the website[5], DiscourseDB is, "an NSF funded data infrustructure (sic) project designed to bridge

[2] https://github.com/mitodl/edx2bigquery
[3] datastage.stanford.edu/handleCourseNameRequest.php?platform=coursera
[4] https://datastage.stanford.edu/StanfordMoocPosts/
[5] http://discoursedb.github.io/

data sources from multiple platforms for hosting … learning experiences." The learning experiences involve "collaborative and discussion-based learning," similar to MOOCdb in that they intend to provide a standard data model focusing on, "Chat, Threaded Discussions, Blogs, Twitter, Wikis, and Text messaging". The only open data sets available here are related to an open-source project, OpenFL, and the Crito dialogues by Plato. In the long run, this may be an essential source for analysis of free text in MOOC data; thus it is a project worth watching. At the same time, it does not solve the MOOC open data set problem today.

F. Tigris

Also linked from the LearnSphere website is the Tigris project, which is the "workflow authoring tool" from the LearnSphere front end. This is the GUI layer between LearnSphere and MOOCdb.

G. Kaggle Person-Course Data 2016

Kaggle has a project[6] titled *Online Courses from Harvard and MIT* that points to an expanded version of the previous release of Person-Course data noted above. It is based on a research paper by the same name [18]. Although this initially looks like a full expanded data set was released for 290 courses, the only data published was a summary appendix.[7]

VI. Deep Learning in MOOC Research

Because initiatives such as MOOCdb and the edx2bigquery data pipeline have disappointed in terms of producing open data sets, the most exciting research comes out of the for-profit companies themselves. For example, a paper titled *Gritnet: Student Performance Prediction with Deep Learning* features three authors who all give their address as the Udacity corporate headquarters in Mountain View, CA [12]. The paper's authors sum up the problem nicely:

"Reliable early-stage predictions of a student's future performance could be critical to facilitate timely educational interventions during a course."

They go on to state that few researchers have used deep learning neural networks to improve the prediction of at-risk students earlier than other methods. The authors "recast the student performance prediction problem as a sequential event prediction problem" and have outperformed existing approaches. They identify the traditional method of analysis that typically uses linear regression. They accurately state that MOOCs present a management problem of a scale beyond the abilities of even the most experienced teachers. In an on-campus classroom setting, teachers have more ability to predict student performance. The sheer number of students in a MOOC makes this impossible, both for reasons of distance and learning at scale.

The deep learning system they developed is named GritNet. The authors claim GritNet, "does not need any feature engineering (it can learn from raw input) and … can operate on any student event data associated with a timestamp (even when highly imbalanced)."

[6] https://www.kaggle.com/edx/course-study
[7] http://year4.odl.mit.edu/appendix.html

VII. MOOCdb

MOOCdb was introduced in a 2013 research paper published by MIT, Coursera, and Stanford [7]. The authors had an idea to facilitate data sharing and publication that would allow many more researchers to have access to MOOC log data analysis methods. Like the edx2bigquery approach, their core idea was to create a standard schema across their organizations. The limitation was that MOOCdb would not provide data or computational power [19], remaining dependent on data providers to keep MOOCdb alive.

Researchers wanting to do analysis would perform the same queries on disparate databases, selecting features they wanted for their proposed models. The MOOCdb format would ensure that each disparate data set would return results in a consistent manner. While some old Stanford web pages show course data available via MOOCdb, Stanford stopped taking public data requests in 2018[8].

In 2013-2014 the MoocDB concept was extended beyond a standardized schema to create MoocViz and MoocViz 2.0 [8][9]. A website was developed to allow researchers to create, upload and share scripts used to analyze the underlying MOOCdb data. While this did not move MOOCdb toward being a repository for open data sets, over time it has evolved into an important application of Human-Computer Interface (HCI) methods to the analysis of MOOC data. The current interface allows for the HCI method of direct manipulation to be used on scripts and datasets to chain them together into an analysis pipeline.

Starting with closed data sets, a data extraction script is applied. This results in an extracted data set. Data aggregation scripts are applied to the extracted data set, thus creating public, aggregated data that has been stripped of PII. Visualization scripts are then applied to the aggregated data, resulting in graphical displays summarizing the original closed data set.

The closed data sets are not shared, protecting the PII in them. Researchers are free to share the aggregation scripts, and a graphical user interface is provided to chain these steps together. This system is not open to all, in that many of the scripts are not shared publicly and there is no access to other's data sets. However, a researcher who possesses, for example, exported Coursera data can upload the data to the system and run analysis using the few publicly shared scripts – or write and share their own. This system was given the name MoocViz and has been integrated into the LearnSphere system previously discussed.

VIII. Available Open Data Sets

A. Moodle Research

One open data set we found was on Moodle.net, where a *Learn Moodle* MOOC is offered roughly twice a year (https://learn.moodle.net/). In 2016, the people at research.moodle.net created an open data set and made it available. A fundamental limitation is that it only covers one course, so any research based on this dataset cannot see the longitudinal retention rate of MOOC students over a more extended period. There are also discrepancies between the

[8] https://iriss.stanford.edu/carol/research

documentation and the actual data, as well as data cleaning issues such as duplicated unique keys.

Initially, 6,119 learners enrolled in the course. Less than half of these completed a survey answering whether their data could be used for research. Of those, eighty percent gave their permission. Consistent with MOOC trends, only one-quarter of enrollees finished the first activity (a "hello" forum post). In the end, only twelve percent of students (735) earned the completer badge [20]. In nominal terms, this is a low number. Having additional user data to analyze would make the analysis significantly better, but we used this data set in our MOOC.

B. edX HarvardX/MITx Person-Course on Dataverse

A MIT News article discusses the public release of an open data set from HarvardX/MITx of sixteen edX-based courses [9] [21]. Named *HarvardX-MITx Person-Course dataset AY2013*, this open data set shows that in the early days of MOOCs, there was a recognition that reproducible research is a crucial element of applying the scientific method to MOOC research. The only restrictions of use of the data set are that researchers do not try to re-identify the data, and they do not redistribute the data.

A paper titled *Acting the Same Differently: A Cross-Course Comparison of User Behavior in MOOCs* [22], does a great job of providing reproducible research on this data. The paper uses this HarvardX/MITx open data set and identifies which courses were used for analysis:

> Our study utilizes … courses, including 6.002x (Fall 2012 and Spring 2013): Circuits and Electronics, 2.01x (Spring 2013): Elements of Structures, 3.091x (Spring 2013): Introduction to Solid State Chemistry

These courses are all part of the sixteen-course data set. The producers of this data set note that the raw data from the courses include PII and the following methods were used to de-identify the data set: "de-identification, removing personally identifiable information using best practices and expert determination methods, including aggregation, anonymization via random identifiers, and blurring, among other techniques".

They note the challenge in de-identifying forum data when the original forum posts may be public and still available online. They observe that if a student were to compile their statistics, such as the number of forum posts and grades, and publish them on Facebook, it would be possible to re-identify the student in the Person-Course data set.

C. Canvas Network Data on Dataverse

This open data set [10], released under the Creative Commons Attribution 4.0 International License, has the same structure as the HarvardX/MITx Person-Course data set, but was not produced in affiliation with HarvardX/MITx. It represents "de-identified data from Canvas Network open courses running January 2014 - September 2015". It contains 325,000 records from learner activity in 238 Canvas open courses. Many of the courses on https://www.canvas.net/ are related to administering and analyzing Canvas courses.

Per the Person-Course data model, each record represents one student enrollment in one course. If a student took three courses, there would be three records. The records contain Canvas administrative data, user survey responses, and user-generated activity data such as the number of minutes of video watched.

IX. "MOOCs Are Dead" -- Move Away From Open

One reason private MOOC providers such as edX and Udacity have not been more active in producing open data sets is their self-proclaimed move away from openness. In 2017, Clarissa Shen, a Udacity vice-president, said "MOOCs are dead … MOOCs are a failed product, at least for the goals we had set for ourselves," in an interview with the Indian media [23]. In 2018, Anant Agarwal, CEO of edX, said, "MOOCs are dead," in an IBL News article [24]. In that article, a "member of one participating university," was quoted saying, "MOOC is a philosophy of education; it has never been a business model. We signed up on edX following this principle."

There is a move away from the underlying definition of open in the term Massively Open Online Courses. Genuine openness means that someone who has an email address can sign up and take a course. While this remains possible on Coursera, edX, and Udacity, the model of charging for a verified completion certificate is gaining strength. At the same time, more and more universities are outsourcing their Masters degrees to these platforms.

X. Discussion

There are a few places to get open MOOC data sets, yet there is no one standardized interface to understand them all. More open, publicly accessible data sets are needed. The landscape for what is available is fractured without one single source of information to lead researchers to what they need quickly. This paper tries to remedy that situation.

If another effort to share anonymized MOOC data is underway, we are unaware of it. To open MOOC weblog research to a much wider group of interested researchers, it would be helpful to develop a standard data format and method of anonymizing data. While this goal has been attempted and abandoned in the past, the problem is not insurmountable from a technical standpoint. What is needed is the collective commitment of private and public MOOC providers to work together in creating a common standard for data sharing. Also, the lawyers have to approve, which is likely to be the sticking point. The first step would be a change in perception of the value of sharing open data sets.

Also, from the standpoint of reproducible research, it would be desirable for more open data sets to be available. While some researchers may continue to perform analysis on proprietary data, if there were open data sets that had a standard schema, then the researcher's methods could be reproduced by other, differently biased researchers. A fundamental element of the scientific method is to do research that can be replicated by others. The primary concern is to allow research to be verified by parties that do not have the same self-interest or bias in the results.

[9] https://doi.org/10.7910/DVN/26147
[10] https://doi.org/10.7910/DVN/1XORAL

If the commercial interests such as Udacity and Coursera could see a benefit in terms of higher retention and completion rates, this incentive might be sufficient to obtain their participation in an open data sharing initiative. It is harder to pinpoint the tangible benefit to educational institutions – perhaps they would benefit from having more research opportunities, and this could lead to a pipeline of research talent from schools to companies when the students graduate.

Another area where open data set standards could be improves is the method of analysis. In this area, we have seen MoocDB / LearnSphere has made progress in allowing researchers to write and share scripts that can be run on any data set that adheres to a common data format. By sharing the actual method of analysis, another barrier to reproducible research would fall. This would work like an open source project such as Python Pandas where contributors work together to build an analysis "script" library. LearnSphere is on the right track in this respect, though many of the scripts can be run by anyone but their methods are not transparent.

There are differing opinions on who works under stronger competitive pressures. To the extent that Limeades are hosting their programs on platforms that also offer proprietary certificates and Nanodegrees, it seems to be in everyone's interest to increase MOOC completion rates. Proper research will identify incentives in each environment that will entice students to finish what they start, or at least will allow researchers to discern between students who start with a plan to finish vs. those who sign up solely to see the content.

REFERENCES

[1] J. Reich and J. A. Ruipérez-Valiente, "The MOOC pivot," *Science (80-.).*, vol. 363, no. 6423, pp. 130–131, 2019.

[2] J. G. S. Goldie, "Connectivism: A knowledge learning theory for the digital age?," *Med. Teach.*, vol. 38, no. 10, pp. 1064–1069, 2016.

[3] R. D. Peterson, "MOOC fizzles," *Acad. Quest.*, vol. 27, no. 3, pp. 316–319, 2014.

[4] D. Joyner, "Squeezing the limeade: policies and workflows for scalable online degrees," in *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, 2018, p. 53.

[5] "Harvard Dataverse." [Online]. Available: https://dataverse.harvard.edu/. [Accessed: 08-Jun-2019].

[6] M. Khalil and M. Ebner, "De-identification in learning analytics," *J. Learn. Anal.*, vol. 3, no. 1, pp. 129–138, 2016.

[7] K. Veeramachaneni, F. Dernoncourt, C. Taylor, Z. Pardos, and U.-M. O'Reilly, "Moocdb: Developing data standards for mooc data science," in *AIED 2013 workshops proceedings volume*, 2013, vol. 17.

[8] F. Dernoncourt *et al.*, "MoocViz: A large scale, open access, collaborative, data analytics platform for MOOCs," in *NIPS workshop on data-driven education, Lake Tahoe, Nevada. Retrieved from http://groups. csail. mit. edu/EVO-DesignOpt/groupWebSite/uploads/Site/MoocViz. pdf*, 2013.

[9] P. Thompson and K. Veeramachaneni, "MOOCviz 2.0: A collaborative MOOC analytics visualization platform." 2013.

[10] S. Boyer and K. Veeramachaneni, "Transfer learning for predictive models in massive open online courses," in *International conference on artificial intelligence in education*, 2015, pp. 54–63.

[11] S. Boyer, B. U. Gelman, B. Schreck, and K. Veeramachaneni, "Data science foundry for MOOCs," in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2015, pp. 1–10.

[12] B.-H. Kim, E. Vizitei, and V. Ganapathi, "GritNet: Student Performance Prediction with Deep Learning," *arXiv Prepr. arXiv1804.07405*, 2018.

[13] A. Hellas *et al.*, "Predicting academic performance: a systematic literature review," in *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, 2018, pp. 175–199.

[14] P. Ihantola *et al.*, "Educational data mining and learning analytics in programming: Literature review and case studies," in *Proceedings of the 2015 ITiCSE on Working Group Reports*, 2015, pp. 41–63.

[15] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," in *Learning analytics*, Springer, 2014, pp. 61–75.

[16] D. Hickey, "On MOOCs, BOOCs, and DOCCs: Innovation in open courses." 2013.

[17] G. Lopez, D. T. Seaton, A. Ang, D. Tingley, and I. Chuang, "Google BigQuery for Education: Framework for Parsing and Analyzing edX MOOC Data," in *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, 2017, pp. 181–184.

[18] I. Chuang and A. Ho, "HarvardX and MITx: Four years of open online courses--fall 2012-summer 2016," *Available SSRN 2889436*, 2016.

[19] J. Gardner, C. Brooks, J. M. L. Andres, and R. Baker, "Replicating MOOC predictive models at scale.," in *L@ S*, 2018, p. 1.

[20] Research.moodle.net, "About the 'Learn Moodle' anonymized database," 2016. [Online]. Available: https://research.moodle.net/158/3/anonymiseddatasetreadme.pdf. [Accessed: 06-Oct-2019].

[21] News Office, "MIT and Harvard release de-identified learning data from open online courses | MIT News," *MIT News*, 2014. [Online]. Available: http://news.mit.edu/2014/mit-and-harvard-release-de-identified-learning-data-open-online-courses. [Accessed: 05-Aug-2019].

[22] B. Gelman, M. Revelle, C. Domeniconi, A. Johri, and K. Veeramachaneni, "Acting the Same Differently: A Cross-Course Comparison of User Behavior in MOOCs.," *Int. Educ. Data Min. Soc.*, 2016.

[23] Jeffrey R. Young, "Udacity Official Declares MOOCs 'Dead' (Though the Company Still Offers Them) | EdSurge News," *EdSurge*, 2017. [Online]. Available: https://www.edsurge.com/news/2017-10-12-udacity-official-declares-moocs-dead-though-the-company-still-offers-them. [Accessed: 10-Aug-2019].

[24] "MOOCs Are Dead, Welcome MOOC-Based Degrees | IBL News," *IBL News*, 2018. [Online]. Available: https://iblnews.org/moocs-are-dead-welcome-mooc-based-degrees/. [Accessed: 10-Aug-2019].