

Preprint

Accepted for publication in *Journal of the Association for Information Science and Technology (JASIST)*

MOOC Visual Analytics: Empowering Students, Teachers, Researchers, and Platform Developers of Massively Open Online Courses

Scott R. Emmons
CNS, SOIC, Indiana University
1320 W 10th Street
Bloomington, IN 47405, USA
Phone: 1.812.855.3259
Fax: 812.855.6166
sremmons@indiana.edu

Robert P. Light
CNS, SOIC, Indiana University
1320 W 10th Street
Bloomington, IN 47405, USA
Phone: 1.812.856.3465
Fax: 812.855.6166
lightr@indiana.edu

Katy Börner
CNS, SOIC & IUNI, Indiana University
1320 W 10th Street
Bloomington, IN 47405, USA
Phone: 1.812.855.3256
Fax: 812.855.6166
katy@indiana.edu

ABSTRACT

Along with significant opportunities, MOOCs provide major challenges to students (keeping track of course materials and effectively interacting with teachers and fellow students), teachers (managing thousands of students and supporting their learning progress), researchers (understanding how students interact with materials and each other), and MOOC platform developers (supporting effective course design and delivery in a scalable way). This paper demonstrates the use of data analysis and visualization as a means to empower students, teachers, researchers, and platform developers by making large volumes of data easy to understand. First, we introduce the insight needs of different stakeholder groups. Second, we compare the wide variety of data provided by major MOOC platforms. Third, we present a novel framework that distinguishes visualizations by the type of questions they answer. We then review the state of the art MOOC visual analytics using a tabulation of stakeholder needs versus visual analytics workflow types. Finally, we present new data analysis and visualization workflows for statistical, geospatial, and topical insights. The workflows have been optimized and validated in the Information Visualization MOOC (IVMOOC) annually taught at Indiana University since 2013. All workflows, sample data, and visualizations are provided at <http://cns.iu.edu/2016-MOOCVis.html>.

INTRODUCTION

Sites like Class Central (<https://www.class-central.com>) and MOOC List (<https://www.mooc-list.com>) help students find relevant courses across platforms. On March 22, 2015, more than 2000 courses by more than 50 different providers were listed. The top-five Learning Management System (LMS) providers from Class Central are

- Coursera, a social entrepreneurship company founded by computer science professors Andrew Ng and Daphne Koller from Stanford University, offering 989 courses.

- EdX, a not-for-profit enterprise with MIT and Harvard universities as founding partners, offering 473 courses.
- Canvas.net, an online course network developed and supported by Instructure, an education technology company that partners with educators, institutions, and technologists, offering 242 courses.
- Miríada X, (129 courses) a platform widely used to teach Spanish MOOCs, offering 129 courses.
- Udacity, a for-profit educational organization founded by Sebastian Thrun, David Stavens, and Mike Sokolsky, offering 74 courses.
- Google Course Builder (GCB), the open source education platform by Google, is covered as 'Independent' in Class Central (but missing in MOOC list) and used to teach 104 courses (<https://code.google.com/p/course-builder/wiki/ListOfCourses>).

As the number and quality of MOOC courses increases, the number of MOOC students increases as well. In September 2015, Coursera reported that it alone has 15M students registered for its offering of 1,000 courses in 35 languages by its 120 partner institutions; 2.5M students completed courses (Coursera, 2015). David Malan's CS50x, an introductory computer science course offered by Harvard and edX, attracted 150,000 student enrollments in its 2013 offering (Malan, 2013). However, course completion rates are less than 10%, indicating a need to explore novel means, including visual analytics, to help people manage course materials, understand requirements, and understand their very own learning progress to ultimately increase completion rates. Major reasons for incompleteness are inability to commit time, poor prior knowledge, lecture fatigue, poor course design, clunky communication tools, and bad peer review (Colman, 2013). Students that pay a fee—even a minimal fee such as \$50 for a Signature Track program—reach completion rates of 70% (Kolowich, 2013).

The remainder of this paper discusses MOOC data analyses and visualizations that aim to help MOOC students, teachers, researchers, and platform developers understand and improve learning dynamics, trajectories, and progress at the individual and aggregated levels. The subsequent section reviews the insight needs and tasks of these four user groups. The *Methods* section discusses data types and formats that different MOOC platforms support, and it presents data analysis and visualization workflows that address the needs of different user groups. The *MOOC Visual Analytics Workflows* section showcases re-runnable workflows and discusses key insights gained from MOOC data analyses. The paper concludes with an outlook to future challenges and opportunities.

RELATED WORK

There is value in the richness of real-world classroom interactions. When aiming to teach or take a MOOC, students and teachers can quickly feel like they are driving blindfolded in heavy traffic on a German autobahn that has no speed limit—they simply have no means to tell who is driving next to them, how fast they are travelling, and when to expect a major collision. In response, some teachers decide to use MOOCs as a platform for high bandwidth delivery of lecture videos and low bandwidth means for multiple-choice assignments that are graded automatically. However, a growing number of users—students and teachers but also learning researchers and platform developers—are embracing MOOCs as a means to “teach the world,” i.e., to improve learning outcomes for millions of students. The challenges encountered are numerous. Some are quantitative in number: scaling up to 100,000 students per class is non-trivial. Because most online course delivery platforms are not designed for high-volume traffic, few make it easy to effectively participate in or manage 1000 discussion threads actively used by 100,000 students. Other challenges are qualitative in nature: teaching students from 100+ countries with vastly different expertise and cultural backgrounds requires language and time zone support but also sensitivity to cultural expectations and foreknowledge (Karen, 2015). While some challenges are shared by students, teachers, researchers, and platform developers, others are specific to one stakeholder group. Key insight needs by

the different stakeholders are discussed here in non-exhaustive lists. Exemplary visualizations are shown in Table 1, which tabulates the four different user groups (columns A-D) versus five types of analysis and visualization discussed in the section *MOOC Visual Analytics* (rows 1-5). References to original works are given in the lower right of each table cell. Workflows for visualizations with a red triangle in the upper-right hand corner of the cells are discussed in the section *MOOC Visual Analytics Workflows* and available at <http://cns.iu.edu/2016-MOOCVis.html>. Table 2 provides captions with context for each cell of Table 1.

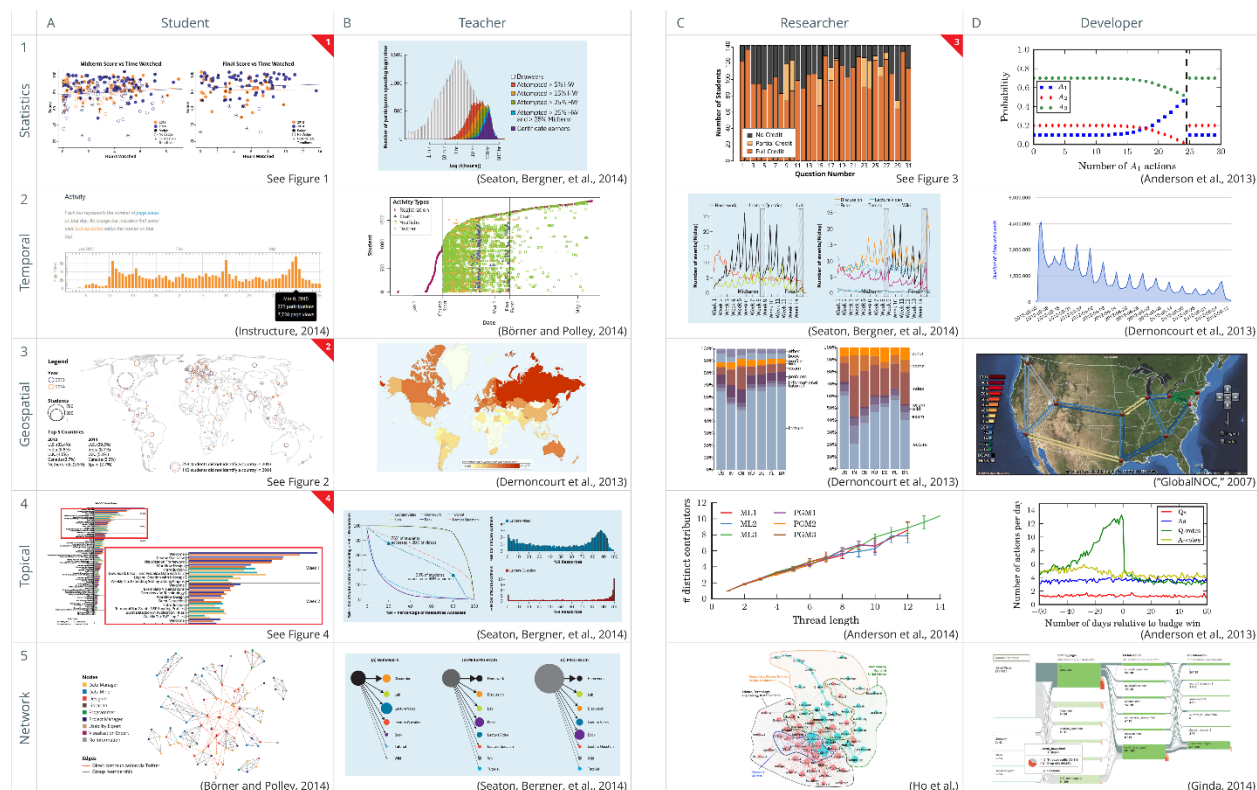


Table 1. Analysis types vs. User Needs. Full-size version at <http://cns.iu.edu/2016-MOOCVis.html>.

Cell	Caption	Citation
A1	Scores vs. time invested watching course videos for students who took the 2013 (blue) and 2014 (orange) IVMOOC midterm (left) and final exam (right) and got at least 50% correct. Filled circles indicate students that earned a badge in the IVMOOC while unfilled circles indicate other students that took the exam. The orange and blue lines indicate the trend lines for the respective years.	See Figure 1
A2	A bar graph from Canvas, showing students' daily use of online course materials. The height of each bar encodes the number of page views	(Instructure, 2014)

	on a given day. Selecting a bar additionally details the date and number of participants on that date.	
A3	Proportional symbol map of the world showing the location of IVMOOC students from 2013 (blue) and 2014 (orange). Circles are area size coded by the number of students per country. Not all students reported their country and missing values are given in the lower part of the map. The top-five countries per year are listed in the lower left.	See Figure 2
A4	A bar graph illustrating the number of views for each video of the IVMOOC. Both the 2013 (blue) and 2014 (orange) offerings of the class are shown. Each video is categorized as either “Theory” or “Hands-On.”	See Figure 4
A5	A network graph of student collaboration during the final assignment of the IVMOOC. The project entailed working on real-world client projects. The nodes of the graph represent the students who completed the projects in groups, had designated roles, and communicated with each other via Twitter.	(Börner & Polley, 2014)
B1	A histogram of the number of students who spent a given amount of time in MIT’s 2012 offering of 6.002x. Time is displayed on a horizontal log axis. The bars of the histogram are grouped and colored by how much overall progress students made on the course, measured by homework completion, taking the midterm, and earning a certificate.	(Seaton, Bergner, et al., 2014)
B2	Activity over time for each student with a unique login in the 2013 IVMOOC. Tracked student actions include registering for the course (purple square), taking an examination (blue triangle), watching a YouTube video (green square), and using the course’s hashtag on Twitter (orange diamond). Students are sorted vertically by registration date.	(Börner & Polley, 2014)
B3	The percentage of students earning a certificate on a country-by-country basis for Stanford’s Coursera offering of Cryptography I in 2013. Darker coloring indicates a higher percentage of certificate earners.	(Dernoncourt et al., 2013)
B4	The percentage of students earning certificates (%N) accessing more than a given percent (%R) of each of the resources in MIT’s 2012 offering of 6.002x. The line graph plots usage curves for which the density of users equals the opposite of the curve’s derivative. The blue histogram illustrates lecture video access, and the red histogram illustrates lecture question access.	(Seaton, Bergner, et al., 2014)
B5	A directed network of student movement while working on the homework (a), the midterm (b), and the final (c) to other course components in MIT’s 2012 offering of 6.002x. The thickness of the edges encodes number of student movements, and the size of the nodes encodes time spent on a course component.	(Seaton, Bergner, et al., 2014)
C1	A stacked bar graph showing student scores per question on the midterm of the Information Visualization MOOC’s 2014 offering. A	See Figure 3

	total of 142 students took the 31-question test, receiving either full credit, partial credit, or no credit for each question.	
C2	Line graphs of student activity, filtered to include certificate earners only, in MIT's 2012 offering of 6.002x. Each point represents the number of times a given resource was accessed divided by the number of people active on the day of access. Students took the midterm and final exams in the blocks of time enclosed by the grey, striped rectangles.	(Seaton, Bergner, et al., 2014)
C3	Relative resource usage by country for MIT's 2012 offering of 6.002x (left) and Stanford's 2013 Coursera offering of Cryptography I (right). The country of each student is inferred from the student's log-in IP address.	(Dernoncourt et al., 2013)
C4	The average number of distinct contributors for a given thread length on the discussion forums of three successive Coursera offerings of Machine Learning (ML) and Probabilistic Graphical Models (PGM). In general, each new comment on a thread is by a new contributor, reflective of the question-and-answer behavior of the forums.	(Anderson et al., 2014)
C5	An enrollment network of HarvardX courses (blue nodes) and MITx courses (red nodes). Directed edges between nodes indicate that a student who completed the course of the source node subsequently enrolled in the course of the destination node. The edges are filtered only to include those with over sixty subsequent enrollments. Node size encodes a sum of in- and out-degree.	(Ho et al., 2015)
D1	The probability that users of Stack Overflow, an online question-and-answer site, take one of three actions. After completing the A_1 action on the website 25 times, users earn a badge. A_2 represents all other actions on the website. A_3 is the "life-action" of offline activity. As users near earning a badge, they increase usage of the website as a whole and take more of the badge-encouraged action.	(Anderson et al., 2013)
D2	Student activity measured by the number of observed events per day for MIT's 2012 offering of 6.002x. This graph is among a set of interactive visualizations developed as part of MoocViz, an open access analytics platform.	(Dernoncourt et al., 2013)
D3	The Abilene nationwide advanced network supports the Internet2 by providing an effective interconnect among the regional networking aggregation points, or gigaPoPs, pioneered by Internet2 universities. The GlobalNOC Real Time Atlas shows live traffic for Abilene with high line utilization in red.	("GlobalNOC," 2007)
D4	The number of actions per day taken by users of Stack Overflow, an online question-and-answer site, relative to the day that they earned the "Electorate" badge. Among four possible actions—questions (Q), answers (A), question votes (Q-votes), and answer votes (A-votes)—we see that users increase their Q-voting activity as they near the Electorate badge, which is awarded for Q-votes.	(Anderson et al., 2013)

D5	Sankey graph by Google Analytics showing the flow of traffic on http://cns.iu.edu from July 25 to August 24, 2014. In 895 recorded sessions, most visitors came from the United States (Country / Territory). From the 314 users who visited the home page (Starting pages), visitors most likely went to the current team page (1st Interaction). Red flows indicate drop-offs—visitors who idle or leave the site.	(Ginda, 2014)
----	--	---------------

Table 2. Captions for Table 1. Supporting information for each of the cells in Table 1, a full-size version of which can be found at <http://cns.iu.edu/2016-MOOCVis.html>.

Students taking MOOCs need to be extremely organized and disciplined. While some MOOCs provide class "meet ups" of various kinds, most MOOCs have no hand-holding or encouragement via weekly in-class teacher encounters (Kizilcec & Halawa, 2015). Students have different use behavior and learning needs in MOOC environments based on their demographics and learning styles (Guo & Reinecke, 2014; Liegle & Janicki, 2006). As a result, there has been a recognized need for research in personalized learning environments, or PLE's, a term which encapsulates MOOCs, and for research in the challenges students face to use MOOC platforms effectively (McLoughlin, 2013).

Visual analytics tools can help students keep track of

- Key learning goals and the most efficient study strategies, e.g., how to best benefit from lecture videos, e-textbooks, notes, forums, and the internet.
- How they are performing (e.g., are major milestones reached and good grades accumulated) and how their progress compares to other students (e.g., leading or lagging on exams, see cell **A1** of Table 1, enlarged in Figure 1).
- Who else is taking the course and who might be a good study partner or teammate, e.g., based on expertise, performance, or time zone (see geographic distribution of students in cell **A3** of Table 1, enlarged in Figure 2, and for collaboration patterns see cell **A5** of Table 1).

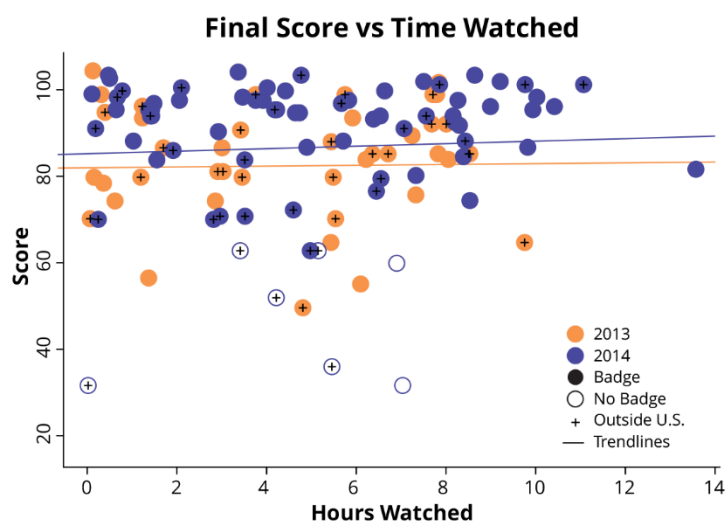
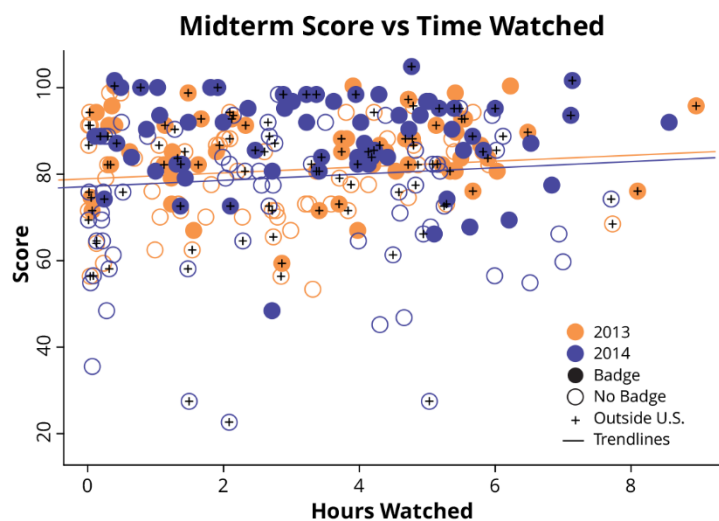


Figure 1. Exam Score vs Time Watched. Scores vs. time invested watching course videos for students who took the 2013 (blue) and 2014 (orange) IVMOOC midterm (top) and final exam (bottom) and got at least 50% correct. Filled circles indicate students that earned a badge in the IVMOOC while unfilled circles indicate other students that took the exam. The orange and blue lines indicate the trend lines for the respective years. (Full-size version at <http://cns.iu.edu/2016-MOOCVis.html>)

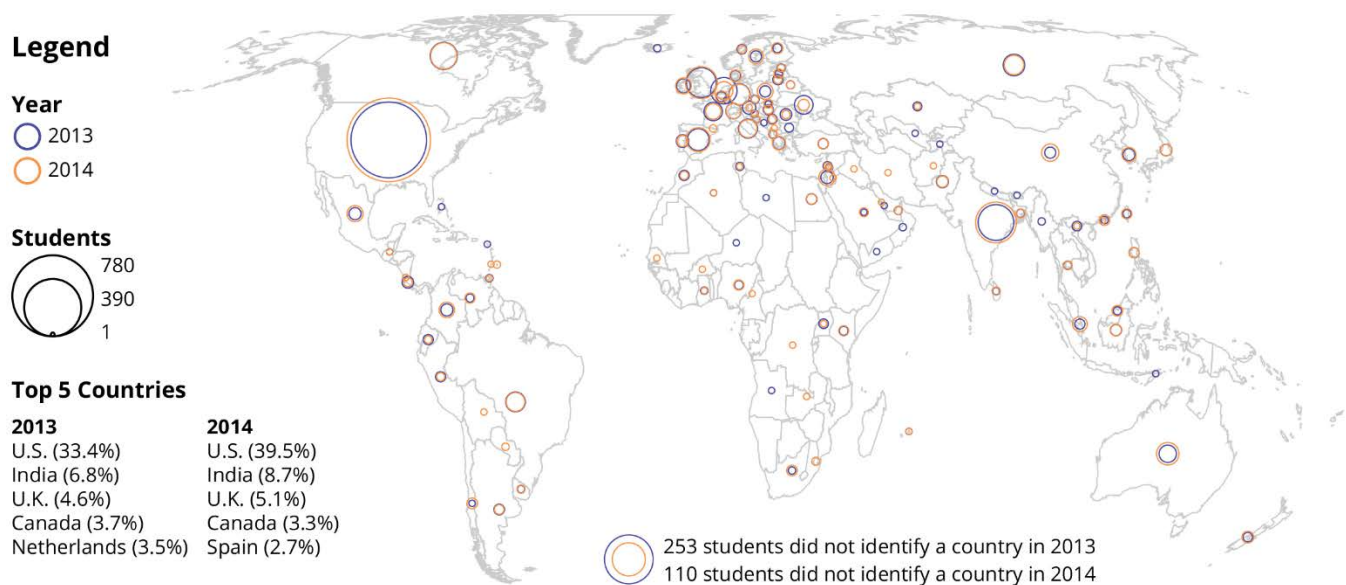


Figure 2. Location of IVMOOC Students. Proportional symbol map of the world showing the location of IVMOOC students from 2013 (blue) and 2014 (orange). Circles are area size coded by the number of students per country. Not all students reported their country and missing values are given in the lower part of the map. The top-five countries per year are listed in the lower left. (Full-size version at <http://cns.iu.edu/2016-MOOCVis.html>)

Teachers (a term which also includes course staff and others helping with teaching a course) of MOOCs need effective means such as visual analytics to keep track of and guide the activities, progress, and problems encountered by thousands of students (Mazza & Dimitrova, 2004). They need to understand the effectiveness of materials, exercises, and exams with respect to learning goals in order to continuously improve course schedules, activities, and grading rubrics, see also course monitoring goals discussed in (Stephens-Martinez, Hearst, & Fox, 2014). Note that this insight is important while the course is running but also to evaluate past courses and help prepare future courses.

Important metrics for teachers in a MOOC context are:

- Students' demographics (e.g., number, background, level of expertise, age, gender, language); motivational factors such as degree or career goals or intended usage of the newly acquired

knowledge; and learning styles (e.g., individual vs. team or textual vs. visual learner) (Kizilcec & Halawa, 2015).

- Student activity and learning progress indicators that make it possible to provide extra support for students that fall behind or to offer more advanced materials to students that master materials quickly (Martinez-Maldonado, Clayphan, Yacef, & Kay, 2014; Taylor, Veeramachaneni, & O'Reilly, 2014; Whitehill, Williams, Lopez, Coleman, & Reich, 2015).
- Bursts of activity, such as those caused by problems with learning materials or inappropriate student behavior, which teachers must counteract and resolve quickly. Activity bursts may also be caused by external events creating unique “teachable moments” that contextualize a particular topic or idea.
- Student performance across exercises, exams, and projects including the analysis of who did what in a team project or how active a student was in the online discussions (Instructure, 2014), see cell **B2** of Table 1 and Figure 3. “Open Learner Models” can be used to analyze group collaboration and design interfaces that enhance student learning (Clayphan, Martinez-Maldonado, & Kay, 2013; Guerra, Hosseini, Somyurek, & Brusilovsky, 2016).
- Student feedback collected via online surveys to reveal strengths and weaknesses of course materials or teaching methods and to reveal additional topics students would like the course to have covered.

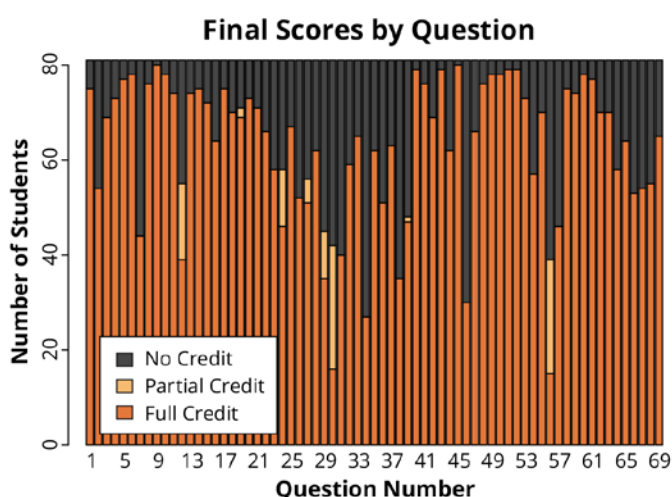
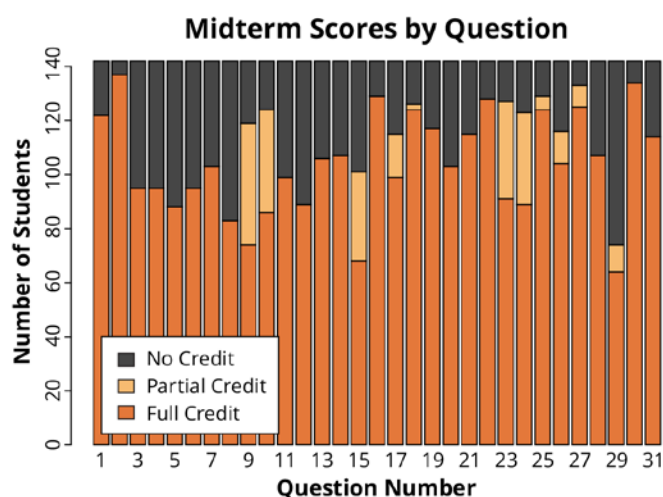


Figure 3. Exam Scores by Question. Student scores per question for midterm (left) and final exam (right) for IVMOOC 2014. (Full-size version at <http://cns.iu.edu/2016-MOOCVis.html>)

Researchers who study human learning and are keen to understand what teaching and learning methods work well in a MOOC environment now have massive amounts of detailed data with which to work. As all student interactions—with learning materials, teachers, and other students—are recorded in a MOOC, human learning can be studied at a level of detail never before possible. Many MOOC teachers double as learning researchers as they are interested to make their own MOOC course work for different types of students.

Visual analytics tools can help researchers study

- Whether factors such as gender, age, education level, disciplinary background, country of origin, and language influence study strategies and learning outcomes. Related work shows that students have fundamental differences in how they interact with course material (Anderson, Huttenlocher, Kleinberg, & Leskovec, 2014; Ho et al., 2015), how they navigate through MOOCs (Guo & Reinecke, 2014; Seaton, Nesterko, et al., 2014), and ultimately how they perform (Kizilcec & Halawa, 2015).
- The popularity and temporal dynamics of student access to content. For example, Seaton, Bergner, et al. (2014) found spikes in textbook usage before examinations (see cell **C2** of Table 1), indicating that students use these resources as references.
- The effectiveness of different student activities and tests or media in teaching; see also cells **B4** and **B5** of Table 1 (Seaton, Bergner, et al., 2014).
- Whether students' study strategies—such as the amount of time spent interacting with course content—correlate with grades, see also Figure 1.
- The importance of motivation and personal organization for completing a MOOC course.

Platform developers need to design systems that support effective course design, efficient teaching, and secure but scalable course delivery. They need to support times of high traffic and resource consumption and schedule maintenances during low activity times.

Visual analytics tools can help developers monitor

- Aggregated user activity patterns to optimize system setup, to detect broken links, or to identify irregularities such as hackers or bots, see cells **D1** to **D5** of Table 1 and associated references.
- Usage of course materials to improve widely-used functionality and omit irrelevant features, see cell **C3** of Table 1 (Demoncourt et al., 2013).

METHODS

This section first discusses the different data types and formats that major MOOC platforms support and then compares existing and novel data analysis and visualization workflows that address the needs of different user groups.

MOOC DATA

Before selecting one of the more than 50 existing MOOC platforms, setting up the course, and opening registration for the first students in a class, it is important to identify what data is needed to monitor student activity relevant for reaching clearly defined learning objectives. Different insight needs (see previous section) can only be satisfied if specific types of data can be recorded by the MOOC platform, obtained via custom surveys, or accessed via existing databases and services (e.g., university student records or LinkedIn data). For example, in order to perform learning outcomes assessments, one must know student knowledge and skills before and after taking the course; to examine gender differences for essay exams vs. multiple-choice exams, one must know the gender of each student.

Each MOOC LMS platform supports the collection of a wide variety of data. Canvas supports a “Course Stream” that lists recent announcements, conversation messages, assignment notifications, and

discussions, and it provides “Course Analytics” that show activity such as page views and student actions over time, assignments submitted on time or late or missing, and grades as a box-and-whisker plot per assignment/exam. GCB provides guidance on how to collect data on three basic categories: "assessments", which cover data from homework and tests tracked within GCB; "reach and engagement", which cover both the location of students enrolled in the course and the activity of each student acquired via Google Analytics; and "happiness", which are student responses to surveys about how satisfied they are with specific aspects of the course (<https://code.google.com/p/course-builder/wiki/MeasureEfficacy>).

Here, we distinguish four general types of data: demographic, performance, activity, and feedback data. Each type is explained subsequently.

Demographic Data: General student demographics, including age, gender, language, education level, and location. Demographic data is commonly acquired during the registration process, and additional demographic data can be acquired via feedback surveys that are discussed ahead.

Performance Data: Student performance based on graded assessments. This is generally collected from homework, quizzes, and examinations, but it also includes results from pre-course surveys designed to examine student knowledge before they take the course.

Activity Data: How students are using class resources, such as the time and date of watching videos, reading material, turning in homework, taking quizzes, or using the discussion forum. Most platforms break down usage by content and media type (i.e., page views, assignment views, textbook views, video views). Following students’ paths through the content via inbound and outbound links (see cell **D5** in Table 1) is important for understanding learning trajectories.

Feedback Data: Student input and feedback. Feedback data allows course providers to learn more about student learning goals and motivation, their intended use of course, and the content students hope to learn.

Feedback data also contains information about what students liked or disliked in terms of course content, structure, grading, or teacher interaction.

MOOC platforms differ widely in how they allow access to data. The ideal platform depends on what student, teacher, learning researcher, or platform developer needs are to be supported as different data analyses and visualizations require rather different data inputs. However, all four data types can be acquired in general. For example, although Canvas does not directly provide the demographic data of gender, age, location, and level of education, it can be acquired using feedback surveys.

Regardless of the adopted platform, API access is usually superior to manual data export or dashboard data access for real-time visual analytics. Access to Google Analytics data is typically restricted to those who run a web site and likely not visible to students. Some data, such as performance data for all students, is only accessible by teachers, and privacy concerns require anonymization or aggregation before it can be shared with other users.

MOOC VISUAL ANALYTICS

A visualization which aims to answer all possible questions for different stakeholders is likely too complex to be understood by many users (Börner, Maltese, Balliet, & Heimlich, 2015). Instead, most visualizations aim to answer “When?”, “Where?”, “What?”, and “With Whom?” questions using temporal, geospatial, topical, and network approaches, respectively. Börner’s (2015) visualization framework is used here as a guide to review existing work and to identify appropriate workflows given a very large combinatorial space of different MOOC datasets and hundreds of different data analysis and visualization algorithms.

Specifically, Table 1 provides a visual overview of exemplary visualizations, whose captions are in Table 2, that is organized by the four different stakeholder groups introduced in the section *Related Work* (columns A-D) versus five types of analysis and visualization (rows 1-5). Subsequently, we review existing work on MOOC visual analytics as well as workflows used to visualize data from the Information

Visualization MOOC (<http://ivmooc.cns.iu.edu>) taught each spring at Indiana University and detailed in the section *MOOC Visual Analytics Workflows*.

Statistics Analysis and Visualization

Line graphs, correlation graphs, and box-and-whisker plots are all examples of how statistical data can be rendered visually. Shown in Table 1, cell **A1** is a graph that shows the return on time investment for students in terms of class score and badges achieved. For details, see Figure 1 and explanatory text. **B1** plots the frequency of hours spent on a course by each student. Color indicates the percentage of attempted assessments (none in gray, >5% in red, ... >25%—and >25% on the midterm—in blue, certificate earners in purple) (Seaton, Bergner, et al., 2014). **C1** depicts statistical data relevant for learning researchers that shows what exam questions potentially need revisions. For details, see Figure 3 and explanatory text. **D1** features data relevant for platform developers that shows how the probability of certain student activities changes after certain actions or exams are completed (Anderson, Huttenlocher, Kleinberg, & Leskovec, 2013).

Temporal Analysis and Visualization

Temporal analyses and visualizations tell when students are active over the span of a course. Data might be examined at different levels of aggregation: by minute, hour, day, week, or semester, by course modules, or before and after a midterm or final. Table 1, cell **A2** depicts a Canvas dashboard showing the number of page views by day for the 2015 Information Visualization MOOC (Instructure, 2014). The visualization reveals light periodic activity over the entire course and activity spikes before examination deadlines. **B2** depicts student activity such as registration (purple), video watching on YouTube (green), tweeting (orange), and taking exams (blue triangles) over time (Börner & Polley, 2014). It shows a continually increasing number of students and activity decreases after major exams—as students realize they may not get good grades and choose to withdraw. **C2** shows an alternative depiction of student activity by Seaton, Bergner, et al. (2014) a temporal line graph of activity magnitude vs. study time to the number of unique users for each day of the week. A periodic spike in users can be seen on Sunday, the

day graded homework and labs were due. **D2** shows the number of observed events per day in MIT's class 6.002x (Dernoncourt et al., 2013) containing periodic spikes indicating weekly activity distribution.

Geospatial Analysis and Visualization

Geospatial data might be examined at different levels of aggregation: by address, city, country, or IP address. **A3** shows a proportional symbol map of students registered for the IVMOOC. Students can use this map to understand the student composition in the course and to form local study groups. **B3** depicts a choropleth map colored by the ratio of certificate winners to the number of registrants by country for the Stanford Cryptography I Coursera course. Red indicates the highest value, 17.24%, for Russia (Dernoncourt et al., 2013). **C3** shows differences in resource usage by country—United States (US), India (IN), China (CN), Russia (RU), Germany (DE), Poland (PL), and Brazil (BR)—for two MOOCs hosted on different platforms (Dernoncourt et al., 2013). The visualization on the left is from MIT's 6.002x offered on edX, and the one on the right is from Stanford's Cryptography I offered on Coursera. There is a pronounced difference in resource usage both by country and by MOOC platform. **D3** shows an internet traffic map allowing developers to determine what portion of their usage is being sent to parts of the world with less stable internet connectivity.

Topical Analysis and Visualization

Topical analysis provides an answer to the question of “what” is going on in a course. **A4** shows how often class videos were watched in the Information Visualization MOOC. For details, see Figure 4 and explanatory text. **B4** Seaton, Bergner, et al. (2014) investigated “What course materials are students using?” with line graphs by categorizing student activity by the type of course material that was used. They found that, among certificate earners, labs and homework assignments were universally used, lecture videos were utilized by roughly half of certificate earners, and the textbook was used little with the important exception of the time of midterm and final exams. **C4** Anderson et al. (2014) sought to understand whether course forums exhibit the dynamics of a discussion board or of a question and answer site. Visually, they used line graphs to study thread length vs number of unique posters and to study course

grade vs thread position. They found that thread length grows linearly with the number of unique posters and that there is a significant spike in course grade from the first to second poster, showing that course forums exhibit the traits of a question and answer site. **D4** Anderson et al. (2013) investigated the efficacy of badges, such as milestones for number of posts, to motivate the behavior of online users. They found that badges were able to influence a significant spike in badge-defined activity for users as they neared completion of a badge.

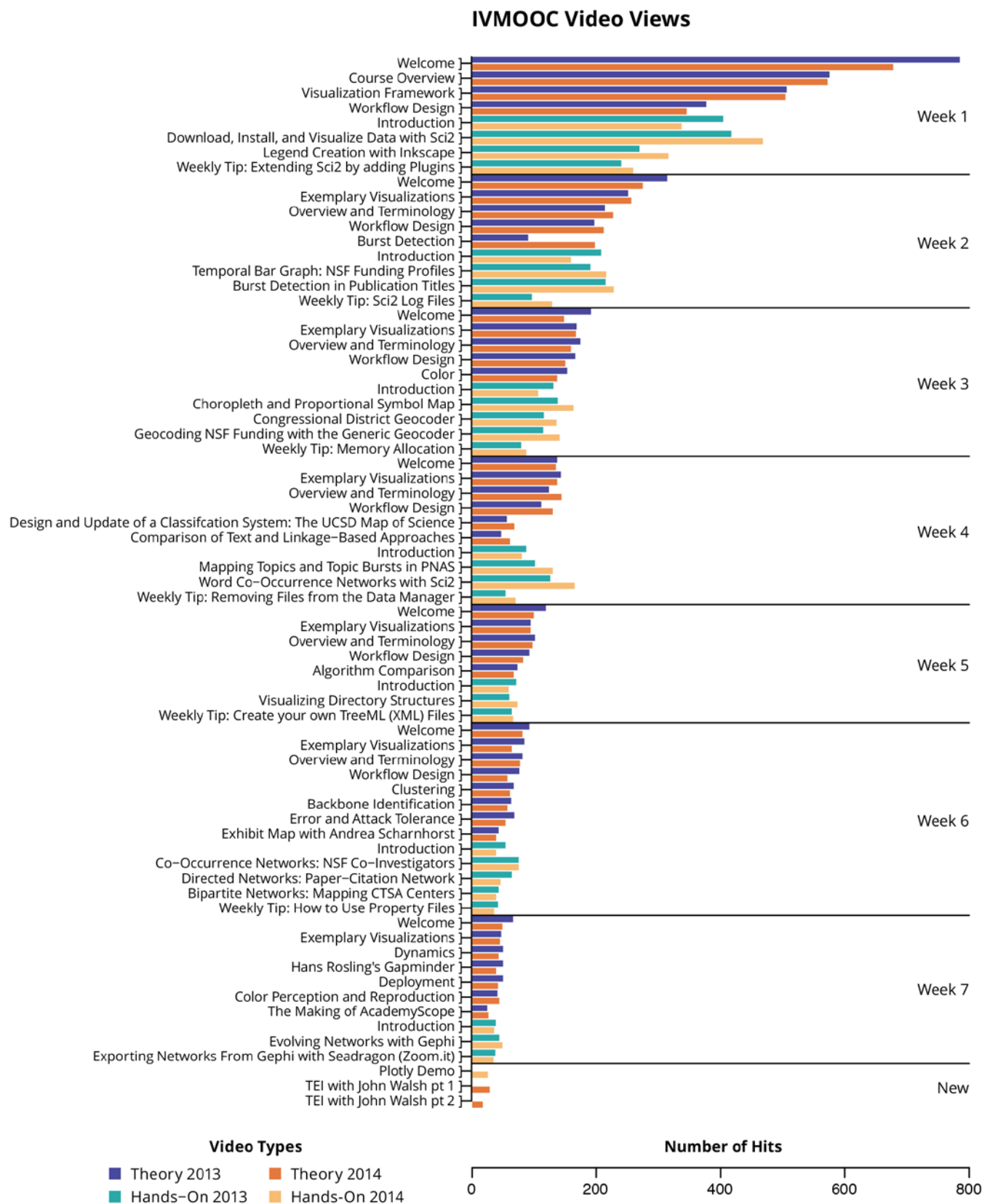


Figure 4. IVMOOC Video Views. IVMOOC video views in 2013 (blue) and 2014 (orange). (Full-size version at <http://cns.iu.edu/2016-MOOCVis.html>)

Network Science Analysis and Visualization

Network analysis examines interactions in a course. These interactions can be between students, between materials on a course website, and between the different courses of a MOOC platform. **A5** shows a network of student interaction to complete the final assignment of the Information Visualization MOOC, which entailed working on real-world client projects (Börner & Polley, 2014). Nodes are colored based on areas of expertise, and the modular structure of teams emerges. Students and teachers can analyze such a network to guide the formation and growth of future teams. **B5** illustrates the transitions students made to other learning components while they worked on homework (left diagram), the midterm (middle diagram), and the final exam (right) (Seaton, Bergner, et al., 2014). The graphs reveal that the textbook was used as a reference for the exams significantly more than for the homework, knowledge which teachers can use in editing resources and which students can use to study effectively. **C5** Ho et al. (2015) created a network of how HarvardX and MITx courses interrelate by examining what courses are taken together or in sequence by 304,414 students. Four curricular content areas can be seen: *Science, Technology, Engineering & Mathematics* in lower left; *Humanities, History, Religion, Design & Education* in yellow on top; *Government, Health & Social Science* in green on right; and *Computer Science* in blue in the middle. Learning researchers can use this information to study how student learning compares and contrasts across the different content areas. **D5** reflects movement through a website acquired via Google Analytics, valuable information for a developer that can show how students are processing information and reveal possible broken links as students route around them.

MOOC VISUAL ANALYTICS WORKFLOWS

This section presents four MOOC data visualizations that are relevant for students, teachers, researchers, and platform developers. Exemplarily, we use data from the Information Visualization MOOC (IVMOOC) run in 2013 and 2014 on GCB. However, the data analysis and visualization workflows run using the R statistics package (R Core Team, 2013) and the Sci2 Tool (Börner, 2011) are generic. All code

and scripts are open source and available at <http://cns.iu.edu/2016-MOOCVis> so that they can be used to visualize other MOOC data. While student activity data cannot be shared easily, applying the same workflows to datasets collected from different MOOCs will support meta-studies to compare results across platforms and course topics.

Empowering Students

As discussed in the section *Related Work*, students need to identify effective study strategies, understand their own performance in the course, and find study partners and project teammates.

The first visualization, Figure 1, uses performance data to run a statistical analysis of scores vs. time invested watching course videos. Each circle denotes a student and non-US students are highlighted by a plus sign within the circle. The graph, which has nearly horizontal trend lines, shows little correlation between time invested watching course videos and examination grade. In 2014, for every hour of video watched there is a 0.7 score increase ($P=0.28$) for the midterm and a 0.3 increase ($P=0.63$) for the final exam. The p-values for these results are not significant, and this data is not presented as any sort of conclusion, but rather shows how visualization can provide a wellspring of questions for future exploration in ways to empower students. Are these trends supported by larger datasets? Are they explained by other factors such as starting level of education or subject familiarity? One possible explanation for this is that time spent watching videos does not capture total time studying because other resources—such as course slides, homework, and an accompanying textbook—are also provided. These results may also indicate that some students watch lots of material passively without learning it or that the material in the lecture videos is not adequate to prepare students for the examination.

During the 2014 IVMOOC, 64 students scored 80% or more of all possible points on exams. Of these, six had no recorded video views on YouTube and another seven only showed an hour or less footage watched. These students may have been downloading videos or slides directly (a zip file comprising all lecture slides and videos is available for those with low bandwidth), but there is no way to link those direct

downloads to specific accounts. Four of the 32 students with similar scores in 2013 recorded less than an hour of video watching. More videos were made available in 2014 and the final score achieved increased. A second visualization addresses another challenge that MOOC students face: identifying study partners or team members across different continents, cultures, and time zones. This insight need is temporal and geospatial: Where are other students from? At what times can I connect with them on the course forum? Activity data can be used to answer these questions. Exemplarily, Figure 2 shows a proportional symbol map of geolocations of the 1,901 registered students in the 2013 offering (blue) and 1,972 students registered students in the 2014 offering (orange). Circles are area size coded by the number of students per country. The circle in the lower left denotes the number of students per year that did not provide address information. In the IVMOOC, student profiles can be searched to help students to connect with each other, including the possibility to form local, in-person study groups. Note that this visualization is also relevant for teachers interested to understand student origin, native languages, time zones, or the geospatial distribution of the ratio of certificate earners to the total number of registrants, see **B3** in Table 1.

Empowering Teachers

As discussed in the section *Related Work*, teachers need effective means to understand the demographics, activity, performance, and feedback of their students in order to adjust and optimize learning materials and mentoring as needed.

For example, teachers must understand how students are interacting with the course materials (e.g., videos, homework, self-tests, and exams) and how this activity correlates with final scores (e.g., see Figure 1) in order to guide them. The insight need is both temporal and topical: What materials are students accessing? When are they accessing them? A stacked line graph of activity data, see **C2** in Table 1, addresses both the temporal and topical aspects of this need as it graphs the magnitude of student registrations and access over time for individual course materials. Topically, teachers can see the popularity of individual videos,

assignments, and readings and are empowered to modify the content and emphasis of class material. Temporally, teachers can see how course materials are used (continuously each week or in bursts right before exams?) and are empowered to send timely emails and forum posts to encourage students to remain engaged and on track (i.e., send reminders for homework/exam to those students that have not yet completed it).

In addition, detailed micro-level data of student performance in exams can be used to communicate how well students understand the material. The question is dual-natured: How strong is performance on each topic? Which students are thriving, and which are struggling? Teachers can use this information to identify poor questions, strengthen course content in areas of difficulty, and reach out to individual students who are struggling. Question breakdown graphs such as those in Figure 3 help teachers understand performance over particular assessment areas.

Teachers also need to understand where around the globe students are accessing the course. Does performance vary by region, possibly caused by language barriers? Are project collaborations difficult due to time zone differences? Figure 2 provides a first step towards answering this question, but the complementary geospatial visualizations in Row 3 of Table 1 provide additional insight.

Empowering Researchers

As discussed in the section *Related Work*, researchers are interested to understand the impact of student demographics, activities, study strategies, or motivation in addition to the type and quality of learning materials on learning outcomes.

A motivating question is if one can predict student dropout based on behavior, which would provide both insight into learning dynamics and the ability for instructor intervention to reduce attrition rates (Whitehill et al., 2015). This question can be answered with activity data that MOOC platforms provide to a fine level of granularity. In particular, researchers have used clickstream data from the first week of class to

predict student dropout to a cross-validation score of 0.81 ± 0.01 , a promising step in this area (Coleman, Seaton, & Chuang, 2015).

Performance data can be used to understand how well students were able to answer any of the many questions asked on examinations in a statistical and topical way. For example, Figure 3 shows student scores by question for the 142 students taking the 2014 IVMOOC midterm (left) and the 82 students taking the final (right). Answers that received no credit (black), partial credit (light orange), and full credit (deep orange) are shown. Note that this visualization is relevant for and can be shown to students and teachers. Questions on which performance was poor can indicate to teachers that additional materials or exercises are needed and indicate to students areas on which to focus when preparing for future exams. Furthermore, comparison to the class as a whole enables both students and teachers to put performance in perspective. If a majority of the class answered a question incorrectly, it marks a potentially difficult question, under-covered subject, or poorly-worded question. If the class performs better than average in a certain area, it either indicates unusually easy questions or that this is an area of strength for the course that is well-presented and well-understood.

Researchers might also like to understand the global impact of online education. Geospatial maps—like the one in Figure 2 and the ones in Table 1, Row 3—can help answer questions such as: What is the reach of a MOOC? How does performance vary by region? How much do the language barriers or foundational education levels in an area matter?

Empowering Platform Developers

As discussed in the section *Related Work*, platform developers are interested to understand what functionality and resources (e.g., storage, compute power, or bandwidth) courses are using and what factors might indicate existing problems or cause problems in the future.

Activity data can be used to explore how course resources are being accessed. This can include monitoring for unwanted behaviors such as spamming or acts of academic dishonesty. Using geospatial tracking of where access is being made, a profile can be formed to describe what percent of users might have only limited internet accessibility. This can inform decisions on what forms of media are ideal with which to offer content and whether low bandwidth alternatives need to be made available. Temporal tracking access over time can not only help point out unexpected bursts in activity but help course managers understand the temporal usage pattern of their users. This can be helpful in planning updates and scheduling events like exams.

Activity data can also be used to explore how users move through the course using a network analysis approach. Many courses, even asynchronous ones, have a delineated path in which content is expected to be consumed, and seeing how well students conform to this path can provide valuable insights to developers. When steps on that path are either missed or approached from unexpected directions, it can reveal problems. A module could be malfunctioning, or it could simply be difficult to access in the intended way. If many students are returning to the main page to access a section, it may indicate that the “next” link that was intended to lead there from the prior section is not working.

DISCUSSION AND OUTLOOK

This paper reviewed existing and presented novel analyses and visualizations of MOOC data for students, teachers, researchers, and platform developers. Hands-on workflows for generic statistical, geospatial, and topical analysis have been made available online in support of effective visualization of MOOC data by others. This section discusses challenges and opportunities for different stakeholders that will impact their insight needs and the development of even more effective visual analytics tools and services.

In a world of nearly-free, any-time, any-place learning, students can become masters of ‘independent majors’, selecting courses and acquiring credits from different universities and other providers. Return on investment (objective or subjective) of student investment of time, money, and energy in terms of

knowledge, skills, credits, and connections (professional and friendship) gained will impact students' decision making. Students need effective means to understand and optimize MOOC activity and performance across the diverse MOOCs they might be taking, which will likely be across different MOOC platforms.

In the near future, we expect to see teachers employ visual analytics to explore diverse methods in an attempt to improve retention rates and to scale-up learning outcomes by personalizing and customizing course content to individual learning styles and needs. Detailed assessments using micro-data of student-teacher activities may empower them to continually improve learning outcomes in a systematic manner. They will work closely with learning researchers and aim to solicit both formal and informal student feedback to optimize teaching and learning.

Learning researchers will have unprecedented opportunities to study learning globally, to study learning across platforms, fields of science, languages, and cultures, and to study learning across learning and teaching styles. Standardization efforts are needed in support of meta-studies, i.e., analyzing data collected from different platforms using standardized data structures, to increase our collective understanding of how different learning and teaching styles combined with technology help or hurt learning.

MOOC platform developers will be expected to mastermind cross-platform open standard databases such as MOOCDB (Veeramachaneni et al., 2014) and visual analytics frameworks such as MOOCViz (Dernoncourt et al., 2013) that are under development and that aim to support plugins for major statistics (e.g., MATLAB, R) and visualization tools (e.g., Google charts and d3js). Ultimately, students will select courses based on content, instructors, and teaching style—not based on platform—making it highly desirable to standardize learning experiences across platforms or to create porous systems that combine the best features of different platforms and visual analytics frameworks.

While current MOOC platforms frequently provide dashboards providing easy access to some data, the dashboards do not cater to the needs of all stakeholders and generally focus primarily on the teacher. We envision the following integration of visual analytics functionality into LMS’:

- Students with LMS course access when they login. In the IVMOOC, students had such access in the format of the “Weekly Scoreboard.”
- Teachers with access to the relevant statistics for the courses they are teaching by means of a “teacher” login in the LMS.
- Learning researchers, who often analyze data across multiple years or course sections, with special access to the appropriate data. This access could come via the LMS or via separate means.
- Platform developers with real-time access to data for a particular course in addition to data from a large number of diverse courses. This access could come via the LMS or via separate means.

We discussed with different LMS platform providers how the visualization analytics workflows presented in this paper might be added as core system functionality. Until integration is completed, anyone is welcome to use, adapt, and extend the code provided at <http://cns.iu.edu/2016-MOOCVis.html>.

Performing this research, it became clear that formal evaluations are needed to determine the utility and effectiveness of visualizations for different stakeholders. Currently, visualizations seem to serve two main purposes: to communicate the results of MOOC data analyses and to increase the number of features offered by MOOC platforms. However, to convert MOOC data into truly actionable knowledge for different stakeholders, qualitative and quantitative evaluations are needed to optimize visualizations for their intended usage.

It is our hope that the framework of general stakeholder types and their needs, major data analysis and visualization types, and well-documented data mining and visualization workflows will speed up the adoption, evaluation, and standardization of MOOC visual analytics.

ACKNOWLEDGMENTS

We would like to thank Samuel T. Mills and Sara Bouchard for their help in designing some of the figures. We thank all of the IVMOOC students for their feedback, comments, enthusiasm, and support. Comments by the anonymous reviewers helped improve the paper considerably. All R code and Sci2 Tool workflows are available and documented at <http://cns.iu.edu/2016-MOOCVis>. This work is supported in part by a MOOC Visual Analytics Tools award by Google Inc. and an Indiana University Learning Analytics Fellows Award.

REFERENCES

- Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2013). Steering User Behavior with Badges. In *Proceedings of the 22Nd International Conference on World Wide Web* (pp. 95–106). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. Retrieved from <http://dl.acm.org/citation.cfm?id=2488388.2488398>
- Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2014). Engaging with Massive Online Courses. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 687–698). New York, NY, USA: ACM. <http://doi.org/10.1145/2566486.2568042>
- Börner, K. (2011). Plug-and-play macroscopes. *Communications of the ACM*, 54(3), 60. <http://doi.org/10.1145/1897852.1897871>
- Börner, K. (2015). *Atlas of knowledge: Anyone can map*. Cambridge, Massachusetts: The MIT Press.
- Börner, K., Maltese, A., Balliet, R. N., & Heimlich, J. (2015). Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. *Information Visualization*, 1–16. <http://doi.org/10.1177/1473871615594652>
- Börner, K., & Polley, D. E. (2014). *Visual Insights: A Practical Guide to Making Sense of Data*. Cambridge, MA: The MIT Press.

- Clayphan, A., Martinez-Maldonado, R., & Kay, J. (2013). Open Learner Models to Support Reflection on Brainstorming at Interactive Tabletops. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education* (pp. 683–686). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-39112-5_88
- Coleman, C. A., Seaton, D. T., & Chuang, I. (2015). Probabilistic Use Cases: Discovering Behavioral Patterns for Predicting Certification. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale* (pp. 141–148). New York, NY, USA: ACM. <http://doi.org/10.1145/2724660.2724662>
- Colman, D. (2013, April 5). MOOC Interrupted: Top 10 Reasons Our Readers Didn't Finish a Massive Open Online Course. Retrieved from http://www.openculture.com/2013/04/10_reasons_you_didnt_complete_a_mooc.html
- Coursera. (2015). *Impact Revealed: Learner Outcomes in Open Online Courses*. Retrieved from https://d396qusza40orc.cloudfront.net/learninghubs/LOS_final%209-21-.pdf
- Dernoncourt, F., Do, C., Halawa, S., O'Reilly, U.-M., Taylor, C., Veeramachaneni, K., & Wu, S. (2013). MoocViz: A Large Scale, Open Access, Collaborative, Data Analytics Platform for MOOCs. Presented at the NIPS Workshop on Data-Driven Education, Lake Tahoe, Nevada, USA.
- Guerra, J., Hosseini, R., Somyurek, S., & Brusilovsky, P. (2016). An Intelligent Interface for Learning Content: Combining an Open Learner Model and Social Comparison to Support Self-Regulated Learning and Engagement (pp. 152–163). ACM Press. <http://doi.org/10.1145/2856767.2856784>
- Guo, P. J., & Reinecke, K. (2014). Demographic Differences in How Students Navigate Through MOOCs. In *Proceedings of the First ACM Conference on Learning @ Scale Conference* (pp. 21–30). New York, NY, USA: ACM. <http://doi.org/10.1145/2556325.2566247>
- Ho, A. D., Chuang, I., Reich, J., Coleman, C. A., Whitehill, J., Northcutt, C. G., ... Petersen, R. (2015). *HarvardX and MITx: Two Years of Open Online Courses Fall 2012-Summer 2014* (SSRN

Scholarly Paper No. ID 2586847). Rochester, NY: Social Science Research Network. Retrieved from <http://papers.ssrn.com/abstract=2586847>

Instructure. (2014, November 8). Canvas Guides. Retrieved from <http://guides.instructure.com/>

Karen, H. (2015). The Single Canon: MOOCs and Academic Colonization. In C. J. Bonk, M. M. Lee, T. C. Reeves, & T. H. Reynolds (Eds.), *MOOCs and open education around the world*. New York: Routledge.

Kizilcec, R. F., & Halawa, S. (2015). Attrition and Achievement Gaps in Online Learning. In *Proceedings of the Second ACM Conference on Learning at Scale*. Vancouver, Canada.

Kolowich, S. (2013, April 8). Coursera Takes a Nuanced View of MOOC Dropout Rates. Retrieved from <http://chronicle.com/blogs/wiredcampus/coursera-takes-a-nuanced-view-of-mooc-dropout-rates/43341>

Liegle, J. O., & Janicki, T. N. (2006). The effect of learning styles on the navigation needs of Web-based learners. *Computers in Human Behavior*, 22(5), 885–898. <http://doi.org/10.1016/j.chb.2004.03.024>

Malan, D. J. (2013, May 3). Data, data, data (from CS50x). Retrieved from <http://harvardx.harvard.edu/blog/data-data-data-cs50x>

Martinez-Maldonado, R., Clayphan, A., Yacef, K., & Kay, J. (2014). Towards Providing Notifications to Enhance Teacher's Awareness in the Classroom. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Intelligent Tutoring Systems* (pp. 510–515). Springer International Publishing. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-07221-0_64

Mazza, R., & Dimitrova, V. (2004). Visualising Student Tracking Data to Support Instructors in Web-based Distance Education. In *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters* (pp. 154–161). New York, NY, USA: ACM. <http://doi.org/10.1145/1013367.1013393>

- McLoughlin, C. E. (2013). The pedagogy of personalised learning: exemplars, MOOCS and related learning theories (Vol. 2013, pp. 266–270). Presented at the EdMedia: World Conference on Educational Media and Technology. Retrieved from <http://www.editlib.org/p/111968/>
- R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>
- Seaton, D. T., Bergner, Y., Chuang, I., Mitros, P., & Pritchard, D. E. (2014). Who does what in a massive open online course? *Communications of the ACM*, 57(4), 58–65. <http://doi.org/10.1145/2500876>
- Seaton, D. T., Nesterko, S. O., Mullaney, T., Reich, J., Ho, A., & Chuang, I. (2014). Characterizing Video Use in the Catalogue of MITx MOOCs. In *the Second MOOC European Stakeholders Summit*. Lausanne, Switzerland.
- Stephens-Martinez, K., Hearst, M. A., & Fox, A. (2014). Monitoring MOOCs: Which Information Sources Do Instructors Value? In *Proceedings of the First ACM Conference on Learning @ Scale Conference* (pp. 79–88). New York, NY, USA: ACM. <http://doi.org/10.1145/2556325.2566246>
- Taylor, C., Veeramachaneni, K., & O'Reilly, U.-M. (2014). Likely to stop? Predicting Stopout in Massive Open Online Courses. *arXiv:1408.3382 [cs]*. Retrieved from <http://arxiv.org/abs/1408.3382>
- Veeramachaneni, K., Halawa, S., Démoncourt, F., O'Reilly, U.-M., Taylor, C., & Do, C. (2014). MOOCdb: Developing Standards and Systems to Support MOOC Data Science. *arXiv:1406.2015 [cs]*. Retrieved from <http://arxiv.org/abs/1406.2015>
- Whitehill, J., Williams, J. J., Lopez, G., Coleman, C. A., & Reich, J. (2015). *Beyond Prediction: First Steps Toward Automatic Intervention in MOOC Student Stopout* (SSRN Scholarly Paper No. ID 2611750). Rochester, NY: Social Science Research Network. Retrieved from <http://papers.ssrn.com/abstract=2611750>