JOURNAL OF LEARNING ANALYTICS

(2017). Shape of educational data: Interdisciplinary Perspectives. *Journal of Learning Analytics*, *4*(2), 6–11. http://dx.doi.org/10.18608/jla.2017.42.2

Shape of Educational Data: Interdisciplinary Perspectives

Colleen M. Ganley and Sara A. Hart
Florida State University, Tallahassee, FL, USA
ganley@psy.fsu.edu

ABSTRACT. This paper is a guest editorial for a special section that forms the proceedings of the Shape of Educational Data meeting. This special section features papers that apply methods from multiple fields — including mathematics, computer science, educational psychology, and learning analytics — to describe and predict student learning in online platforms. The special section is organized such that the first set of articles discusses different online learning systems (WEPS, WeBWorK, and inVideo) and data that can be analyzed from these systems. The second set of articles involves descriptions of topological data analyses that can be helpful to researchers in learning analytics and educational psychology to better model student learning in online courses. The third set of articles uses data obtained from online systems to study factors related to student learning. Due to these multiple approaches, we can gain insight into the types of data available, the ways in which we can measure particular constructs related to learning using these data, and the ways we can analyze these data, including statistical approaches and visualizations.

Keywords: Topology, geometry, learning analytics, educational psychology, data visualization

1 INTRODUCTION

This special section came together as the result of the research that Mika Seppälä conducted over a number of years. This work spans multiple fields, including mathematics, computer science, learning analytics, and educational psychology. Mika had a vision of applying mathematics, specifically concepts in topology and geometry (his mathematical content areas of expertise), to understanding how students learn in online courses. He was passionate about improving his own teaching as well as understanding the ways that students work through online resources. His goal was to use this information to improve the materials in his courses and create a recommendation system for how students should use the course. He built an online education platform — World Education Portals (WEPS; Seppälä, 2014) — as a complete virtual learning environment based on Moodle, supporting online and hybrid courses across 15 institutions worldwide. He advocated for the use of learning analytics to study how students interacted with WEPS and similar learning platforms, to better understand and predict student performance. He hoped that by combining learning analytics with geometry and topology he could help increase student learning.

Mika Seppälä believed that topological and geometric methods in data analysis have a transformative power in understanding the shape of data. This emerging mathematical field of topology (Carlsson, 2009, 2012a, 2012b; Edelsbrunner & Harer, 2010) gives us the potential tools that allow us to extract

(2017). Shape of educational data: Interdisciplinary Perspectives. *Journal of Learning Analytics*, *4*(2), 6–11. http://dx.doi.org/10.18608/jla.2017.42.2

essential information from complex data with noise. Hybrid topological numeric methods developed by Carlsson and others (Carlsson, 2009, 2012a, 2012b; Edelsbrunner & Harer, 2010; Lum et al., 2013) have revealed subgroups in data sets that traditional methodologies have failed to find. These methods can be applied to an individual data set or can be used to find relations between data sets. Seppälä thought that education was a particularly attractive application area for topological methods because of the inbuilt structure (Seppälä, 2013) that may allow us to separate noise from the core data, and to get valuable results that will deepen our understanding of learning.

The goal of his work was to bring researchers in the mathematical community, who are experts in topology and geometry, to collaborate with researchers who run learning platforms that collect massive amounts of data on how students learn, as well as educational psychologists and learning analytics experts. Mika Seppälä unexpectedly passed away in early 2015 and we were fortunate to be able to collectively continue some of his work. Many of the authors in this special issue collaborated with Mika and the themes of his work can be seen across the articles.

2 PAPERS IN THIS SECTION

The articles in this section represent the proceedings from the Shape of Educational Data meeting, held 7–8 April 2016 and featured presentations from researchers in mathematics, computer science, learning analytics, and educational psychology. The special issue is organized such that the first set of articles discusses different online learning systems (WEPS, WeBWorK, and inVideo) and data that can be analyzed from these systems. The second set of articles involves descriptions of topological and geometric data analyses that can be helpful to researchers in learning analytics and educational psychology to better model student learning in online courses. The third set of articles uses data obtained from online systems to model factors related to student learning. Though these articles represent varying approaches from multiple fields, they coalesce in that they discuss the tools, data, and analyses involved in the shared goal of better understanding student learning in online systems.

2.1 Online Systems and How They Are Set Up

In the first article, Pauna introduces the WEPS learning platform, developed in collaboration with Mika Seppälä and others. Pauna describes this platform and what types of learning materials are available on it as well as the types of data available for researchers interested in modelling student learning. The platform contains course material including lectures, as well as quizzes and workshops, which are available to students throughout the course. The math assessment systems, STACK and WeBWorK, which can examine higher level skills than traditional multiple choice questions (as discussed in more detail by Gage in his article), are used for the quizzes. Students are also required to grade each other's workshops within the system, which provides additional data on student use of the system. Pauna ends with an example of how data obtained from the course quizzes and workshops can be analyzed along with course performance and discusses the strengths and weaknesses of attempts to identify what

(2017). Shape of educational data: Interdisciplinary Perspectives. *Journal of Learning Analytics*, *4*(2), 6–11. http://dx.doi.org/10.18608/jla.2017.42.2

these metrics actually measure. As more data become available in WEPS courses, these analyses can be refined and retested.

In collaboration with Pauna and Seppälä, Gage integrated the WeBWorK mathematics assessment database into the WEPS portal on Moodle. In his paper, he discusses this integration and the advantages of this approach and outlines the types of data available to instructors and researchers interested in student mathematics learning. These tools are a large improvement compared to what is available for homework, practice, and assessments within LMSs because these new tools are able to assess student responses more holistically. WeBWorK does not require one specific form of an answer, as long as the response of students is the correct answer in some form, the system identifies it as correct. This helps instructors, as they do not need to rely on multiple choice questions or questions that require very specific answers, but can actually look at how students are understanding problems and can see the types of incorrect answers they are providing. It also benefits the students, because they are able to check their answers and know that, no matter what form they are in, the grading of the answer will be accurate. The very large database of math problems also provides data ripe for researchers to examine student thinking and learning in mathematics.

Wang and Kelly discuss a tool uniquely designed for analyzing video information, inVideo. This tool is capable of searching data about both the audio and images from a video file. inVideo can do this through keyword searches, and is capable of doing so with multiple languages. It can also conduct searches using reference images. The authors argue that this system has the potential to increase student engagement because students are able to leave timestamped comments on videos and instructors can parse videos into smaller, more digestible segments with assessments for students embedded between these small segments. The authors report on the results of research conducted with a number of computer science courses that use these tools. Their research suggests that the inclusion of the tools available through inVideo is related to greater student engagement with the course content. This finding highlights the potential of this software for increasing student engagement and learning.

2.2 Mathematics Concepts for Understanding and Modelling Data

In her article introducing concepts from topology, Munch describes how concepts in topological data analysis can be used to identify and visualize underlying patterns in data, including data about student learning. She introduces two topological signatures: the persistence diagram, which models the structure of a set of data as loops and holes to understand how data points are clustered and connected, and the mapper graph, which produces a model of the shape of a dataset and is a very useful tool for producing visualizations of data. Critically, with immense amounts of data in the educational domain available as well as the availability of software that allows researchers to model these complex mathematical models, there is a unique opportunity to use these methods to understand and visualize student learning in online systems.

(2017). Shape of educational data: Interdisciplinary Perspectives. *Journal of Learning Analytics*, *4*(2), 6–11. http://dx.doi.org/10.18608/jla.2017.42.2

Buser and Semmler explain the mathematical approach initially proposed by Mika Seppälä as a strategy to visualize the flow of students through the activities available within the online learning platform WEPS. Using two examples, that of students moving through the decision points of the Finnish school system and that of students moving through the decision points in the engineering program at École Polytechnique Fédérale de Lausanne, the authors show how Riemann surfaces can be used to model the movement of students. These models can account for many different possible movement patterns including bifurcations (where students move down different paths at a decision point) and merges (where students coming from different paths are joined on the same path). These map well onto the examples used in this paper and can also map onto more complex learning systems such as WEPS. These models provide an excellent example of how one can use mathematical models to visualize patterns in data from educational settings.

Caprotti uses the WEPS online system to show how mathematical principles can be applied to data on student learning. Her work uses clickstream data from the WEPS system to develop graphs, specifically a Markov Chain, to examine the ways that students move through the activities provided in a calculus course. Through these graphs, the most common patterns of progress through the course can be seen, and common skipping patterns can be observed. Initial analyses show that students who are more diligent in the course (i.e., skip fewer assignments in a row) tend to get higher grades in the course as well. She also demonstrates how a look-back, look-ahead, and in-section value can be calculated to understand the strategies that students use when moving between activities. Importantly, she finds that students who are more diligent are more likely to look ahead than those who are less diligent. She ends by discussing how information such as that from the current article can be used to potentially inform recommender systems for what students should do next in the course, based on data from successful and unsuccessful students in past courses.

2.3 Analyzing Data Collected from Online Systems

In their article, Ostrow, Wang, and Heffernan discuss the importance of considering more data than the dichotomous “correct” or “incorrect” often used when examining data on student performance. Instead, they forward a partial credit strategy that takes into account the hints that students receive and the number of attempts they need in order to get the correct answer. With this strategy, more information can be taken into account when using just a small number of test items. Using data from two online learning tools — ASSISTments and Cognitive Tutor — they test the efficiency of this partial credit scoring strategy compared to dichotomous scoring in the categorization of students who are higher or lower performing. They find that, for most skills, partial credit scoring provides more efficient categorization of students, but with no increase in Type I error rates. They discuss the implications of these findings for examining the effectiveness of randomized controlled trials in online education platforms.

To start their paper, Gašević, Jovanović, Pardo, and Dawson recommend to the field of learning analytics that it must draw from educational theory, and not simply be in the habit of atheoretical data mining. With this in mind, the authors draw from the literature concerning student approaches to learning,

(2017). Shape of educational data: Interdisciplinary Perspectives. *Journal of Learning Analytics*, *4*(2), 6–11. http://dx.doi.org/10.18608/jla.2017.42.2

specifically focusing on deep versus surface approaches to learning. The authors were interested in examining the association between a student's self-reported learning approach and their study strategies extracted from digital trace data, and then relating these with course grade data from a hybrid, first-year undergraduate engineering course in computer systems. The authors found four study strategies that could be extracted from the online behaviours of students, differentiated by self-reported approaches to learning, which also predicted performance in the class. The authors conclude that the inclusion of self-reported measures, based on educational theory, in addition to the online trace data, is important in understanding student learning within an online course, although they point out limitations of both types of data.

In our study, Hart, Daucourt, and Ganley, we investigate potential factors related to student success in a flipped Calculus II course on WEPS. We draw on psychological and educational research about the attitudinal and cognitive predictors of student performance, and also include predictors that use online log data and draw on the field of learning analytics. Using a dominance analysis to identify the most critical predictors of performance from among multiple affective, cognitive, and online factors, we find that the key predictors seem to be perceptions of the importance of math, approximate number system (ANS) ability, total amount of discussion forum posting, and time grading peer workshop submissions. These results suggest that combining information from multiple sources, including student self-reports of attitude, direct assessment of related cognitive skills, and information from log data, provides the most information about student success. These findings can help to suggest information that could be used to build a recommender system to improve student learning in this course, and this work utilizes a strategy that can be used across any online courses.

National Science Foundation program officer Eamonn Kelly writes a conclusion, drawing on some of the themes of the work in the special section and highlighting some future directions for the field based on his collaborative work with Mika Seppälä.

3 CONCLUSION

We believe that this collection of articles highlights connections between fields whose collaboration is critical to conducting the most informative research on student learning with data that can be obtained from online learning systems. One theme that came up across articles is the importance of understanding how to best use mathematical principles to visualize data about student behaviour and learning from online systems. Another issue is that it is critical to think carefully about a number of the decisions that we make in regard to data and its analysis in online systems. This is especially critical given the wealth of data available through these online systems, which can be overwhelming and difficult to organize, analyze, visualize, and interpret. Several articles also address issues around what the most appropriate data are to extract from the system to give the most information about how students are using the system. Another critical consideration is the identification of what, exactly, is measured with certain pieces of information extracted from these systems (e.g., What does using hints

(2017). Shape of educational data: Interdisciplinary Perspectives. *Journal of Learning Analytics*, *4*(2), 6–11. http://dx.doi.org/10.18608/jla.2017.42.2

imply? What does it mean to be a person who completes quizzes many times?). These questions are critical for the interpretation of results obtained from data in online systems. Together, the interdisciplinary research included in this special section presents interesting research methods and findings, while also providing ideas for future research on student learning that capitalize on new analytic methods currently outside of the learning analytics field. This work is all centred around the shared goal, championed by Mika Seppälä, of understanding and improving student learning in online courses.

ACKNOWLEDGEMENTS

We sincerely thank the National Science Foundation and our program officer, Barry Sloane, for the support of the Shape of Educational Data meeting as well as of some of the research in this special section through NSF grant 1450501. We thank the authors of these articles for their contributions to this special section and also thank the researchers who reviewed articles. This section could not have happened without the work of Mika Seppälä. Mika was an innovative thinker dedicated to bringing together researchers in many fields with the ultimate goal of using all available tools and knowledge to improve student learning. On a personal level, Mika was a great colleague to many and incredibly supportive of other researchers in this area. Thanks also to Cici Safkan-Seppälä and Sara Seppälä for attending the Shape of Educational meeting and for sharing our enthusiasm for the work presented.

REFERENCES

Carlsson, G. (2009). Topology and data. Bulletin (New Series) of the American Mathematical Society, 46(2), 255–308.

Carlsson, G. (2012a). The shape of data. In F. Cucker, T. Krick, A. Pinkus, & A. Szanto (Eds.), Foundations of Computational Mathematics, Budapest 2011. London Mathematical Society.

Carlsson, G. (2012b). Shape of data. Retrieved from http://youtu.be/kctyag2Xi8o

Edelsbrunner, H., & Harer, J. (2010). Computational topology: An introduction. American Mathematical Society.

Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., ... & Carlsson, G. (2013). Extracting insights from the shape of complex data using topology. Scientific Reports, 3, 1236. http://dx.doi.org/10.1038/srep01236

Seppälä, M. (2013). Riemann surfaces, quadratic differentials, and MOOCS. Retrieved from http://youtu.be/qbS_Cum07xg

Seppälä, M. (2014). World Education Portals (WEPs). https://geom.mathstat.helsinki.fi/moodle/ (formerly https://myweps.com)