40

Educational Data Mining and Learning Analytics

Potentials and Possibilities for Online Education

Ryan S. Baker & Paul Salvador Inventado

Editor's Note

The following was reprinted from Emergence and Innovation in Digital Learning [https://edtechbooks.org/-uG], an open textbook edited by George Veletsianos.

Baker, S., & Inventado, P. S. (2016). Educational data mining and learning analytics: Potentials and possibilities for online education. In G. Veletsianos (Ed.), *Emergence and Innovation in Digital Learning* (83–98). doi:10.15215/aupress/9781771991490.01

Over the last decades, online and distance education has become an increasingly prominent part of the higher educational landscape (Allen & Seaman, 2008; O'Neill et al., 2004; Patel & Patel, 2005). Many learners turn to distance education because it works better for their schedule, and makes them feel more comfortable than traditional face-to-face courses (O'Malley & McCraw, 1999). However, working with distance education presents challenges for both learners and instructors that are not present in contexts where teachers can work directly with their students. As learning is mediated through technology, learners have fewer opportunities to communicate to instructors about areas in which they are struggling. Though discussion forums provide an opportunity that many students use, and in fact some students are more comfortable seeking help online than in person (Kitsantas &

Chow, 2007), discussion forums depend upon learners themselves realizing that they are facing a challenge, and recognizing the need to seek help. Further, many students do not participate in forums unless given explicit prompts or requirements (Dennen, 2005). Unfortunately, the challenges of help-seeking are general: many learners, regardless of setting, do not successfully recognize the need to seek help, and fail to seek help in situations where it could be extremely useful (Aleven et al., 2003). Without the opportunity to interact with learners in a face-to-face setting, it is therefore harder for instructors as well to recognize negative affect or disengagement among students.

Beyond a student not participating in discussion forums, ceasing to complete assignments is a clear sign of disengagement (Kizilcec, Piech, & Schneider, 2013), but information on these disengaged behaviors is not always available to instructors, and more subtle forms of negative affect (such as boredom) are difficult for an unaided distance instructor to identify and diagnose. As such, a distance educator has additional challenges compared to a local instructor in identifying which students are at-risk, in order to provide individual attention and support. This is not to say that face-to-face instructors always take action when a student is visibly disengaged, but they have additional opportunities to recognize problems.

In this chapter, we discuss educational data mining and learning analytics (Baker & Siemens, 2014) as a set of emerging practices that may assist distance education instructors in gaining a rich understanding of their students. The educational data mining (EDM) and learning analytics (LA)

communities are concerned with exploring the increasing amounts of data now becoming available on learners, toward providing better information to instructors and better support to learners. Through the use of automated discovery methods, leavened with a workable understanding of educational theory, EDM/LA practitioners are able to generate models that identify at-risk students so as to help instructors to offer better learner support. In the interest of provoking thought and discussion, we focus on a few key examples of the potentials of analytics, rather than exhaustively reviewing the increasing literature on analytics and data mining for distance education.

Data Now Available in Distance Education

One key enabling trend for the use of analytics and data mining in distance education is that distance education increasingly provides high-quality data in large quantities (Goldstein & Katz, 2005). In fact, distance education has always involved interactions that could be traced, but increasingly data from online and distance education is being stored by distance education providers in formats designed to be usable. For example, The Open University (UK), an entirely online university with around 250,000 students, collects large amounts of electronic data including student activity data, course information, course feedback and aggregated completion rates, and demographic data (Clow, 2014). The university's Data Wranglers project leverages this data by having a team of analytics experts analyze and create reports about student learning, which are used to improve course delivery. The University of Phoenix, a for-profit online university, collects

data on marketing, student applications, student contact information, technology support issue tracking, course grades, assignment grades, discussion forums, and content usage (Sharkey, 2011). These disparate data sources are integrated to support analyses that can predict student persistence in academic programs (Ming & Ming, 2012), and can facilitate interventions that improve student outcomes.

Massive Open Online Courses (MOOCs), another emerging distance education practice, also generate large quantities of data that can be utilized for these purposes. There have been dozens of papers exploiting MOOC data to answer research questions in education in the brief time since large-scale MOOCs became internationally popular (see, for instance, Champaign et al., 2014; Kim et al., 2014; Kizilcec et al., 2013). The second-largest MOOC platform, edX, now makes large amounts of MOOC data available to any researcher in the world. In addition, formats have emerged for MOOC data that are designed to facilitate research (Veeramachaneni, Dernoncourt, Taylor, Pardos, & O'Reilly, 2013).

Increasingly, traditional universities are collecting the same types of data. For example, Purdue University collects and integrates educational data from various systems including content management systems (CMS), student information systems (SIS), audience response systems, library systems and streaming media service systems (Arnold, 2010). This institution uses this data in their Course Signals project, discussed below.

One of the key steps to making data useful for analysis is to pre-process it (Romero, Romero, & Ventura, 2013). Pre-processing

can include data cleaning (such as removing data stemming from logging errors, or mapping meaningless identifiers to meaningful labels), integrating data sources (typically taking the form of mapping identifiers—which could be at the student level, the class level, the assignment level or other levels—between data sets of tables), and feature engineering (distilling appropriate data to make a prediction). Typically, the process of engineering and distilling appropriate features that can be used to represent key aspects of the data is one of the most time-consuming and difficult steps in learning analytics. The process of going from the initial features logged by an online learning system (such as correctness and time, or the textual content of a post) to more semantic features (history of correctness on a specific skill; how fast an action is compared to typical time taken by other students on the same problem step; emotion expressed and context in a discussion of a specific discussion forum post) involves considerable theoretical understanding of the educational domain. This understanding is sometimes encoded in schemes for formatting and storing data, such as the MOOC data format proposed by Veeramachaneni et al. (2013) or the Pittsburgh Science of Learning Center DataShop format (Koedinger, Baker, Cunningham, Skogsholm, Leber, & Stamper, 2010).

Methods for Educational Data Mining and Learning Analytics

In tandem with the development of these increasingly large data sets, a wider selection of methods to distill meaning have emerged; these are referred to as educational data mining or learning analytics. As Baker and Siemens (2014) note, the

educational data mining and learning analytics communities address many of the same research questions, using similar methods. The core differences between the communities are in terms of emphasis: whether human analysis or automated analysis is central, whether phenomena are considered as systems or in terms of specific constructs and their interrelationships, and whether automated interventions or empowering instructors is the goal. However, for the purposes of this article, educational data mining and learning analytics can be treated as interchangeable, as the methods relevant to distance education are seen in both communities. Some of the differences emerge in the section on uses to benefit learners, with the approaches around providing instructors with feedback being more closely linked to the learning analytics community, whereas approaches to providing feedback and interventions directly to students are more closely linked to practice in educational data mining.

In this section, we review the framework proposed by Baker and Siemens (2014); other frameworks for understanding the types of EDM/LA method also exist (e.g., Baker & Yacef, 2009; Scheuer & McLaren, 2012; Romero & Ventura, 2007; Ferguson, 2012). The differences between these frameworks are a matter of emphasis and categorization. For example, parameter tuning is categorized as a method in Scheuer and McLaren (2012); it is typically seen as a step in the prediction modeling or knowledge engineering process in other frameworks. Still, mostly the same methods are present in all frameworks. Baker and Siemens (in press) divide the world of EDM/LA methods into prediction modeling, structure discovery, relationship mining, distillation of data for human judgment, and discovery with models. In this chapter, we will provide definitions and examples for

prediction, structure discovery, and relationship mining, focusing on methods of particular usefulness for distance education.

Prediction

Prediction modeling occurs when a researcher or practitioner develops a model, which can infer (or predict) a single aspect of the data, from some combination of other variables within the data. This is typically done either to infer a construct that is latent (such as emotion), or to predict future outcomes. In these cases, good data on the predicted variable is collected for a smaller data set, and then a model is created with the goal of predicting that variable in a larger data set, or a future data set. The goal is to predict the construct in future situations when data on it is unavailable. For example, a prediction model may be developed to predict whether a student is likely to drop or fail a course (e.g., Arnold, 2010; Ming & Ming, 2012). The prediction model may be developed from 2013 data, and then utilized to make predictions early in the semester in 2014, 2015, and beyond. Similarly, the model may be developed using data from four introductory courses, and then rolled out to make predictions within a university's full suite of introductory courses.

Prediction modeling has been utilized for an ever-increasing set of problems within the domain of education, from inferring students' knowledge of a certain topic (Corbett & Anderson, 1995), to inferring a student's emotional state (D'Mello, Craig, Witherspoon, McDaniel, & Graesser. 2008). It is also used to make longer-term predictions, for instance predicting whether a student will attend college from their learning and emotion in

middle school (San Pedro, Baker, & Gobert, 2013).

One key consideration when using prediction models is distilling the appropriate data to make a prediction (sometimes referred to as feature engineering). Sao Pedro et al. (2012) have argued that integrating theoretical understanding into the data mining process leads to better models than a purely bottom-up data-driven approach. Paquette, de Carvalho, Baker, and Ocumpaugh (2014) correspondingly find that integrating theory into data mining performs better than either approach alone. While choosing an appropriate algorithm is also an important challenge (see discussion in Baker, 2014), switching algorithms often involves a minimal change within a data mining tool, whereas distilling the correct features can be a substantial challenge.

Another key consideration is making sure that data is validated appropriately for its eventual use. Validating models on a range of content (Baker, Corbett, Roll, & Koedinger, 2008) and on a representative sample of eventual students (Ocumpaugh, Baker, Gowda, Heffernan & Heffernan, 2014) is important to ensuring that models will be valid in the contexts where they are applied. In the context of distance education, these issues can merge: the population of students taking one course through a distance institution may be quite different than the population taking a different course, even at the same institution. Some prediction models have been validated to function accurately across higher education institutions, which is a powerful demonstration of generality (Jayaprakash, Moody, Lauría, Regan, & Baron, 2014).

As with other areas of education, prediction modeling

increasingly plays an important role in distance education. Arguably, it is the most prominent type of analytics within higher education in general, and distance education specifically. For example, Ming and Ming (2012) studied whether students' final grades could be predicted from their interactions on the University of Phoenix class discussion forums. They found that discussion of more specialized topics was predictive of higher course grades. Another example is seen in Kovacic's (2010) work studying student dropout in the Open Polytechnic of New Zealand. This work predicted student dropout from demographic factors, finding that students of specific demographic groups were at much higher risk of failure than other students.

Related work can also be seen within the Purdue Signals Project (Arnold, 2010), which mined content management system, student information system, and gradebook data to predict which students were likely to drop out of a course and provide instructors with near real-time updates regarding student performance and effort (Arnold & Pistilli, 2012; Campbell, DeBlois, & Oblinger, 2007). These predictions were used to suggest interventions to instructors. Instructors who used those interventions, reminding students of the steps needed for success, and recommending face-to-face meetings, found that their students engaged in more help-seeking, and had better course outcomes and significantly improved retention rates (Arnold, 2010).

Structure Discovery

A second core category of learning LA/EDM is structure discovery. Structure discovery algorithms attempt to find

structure in the data without an a priori idea of what should be found: a very different goal than in prediction. In prediction, there is a specific variable that the researcher or practitioner attempts to infer or predict; by contrast, there are no specific variables of interest in structure discovery. Instead, the researcher attempts to determine what structure emerges naturally from the data. Common approaches to structure discovery in LA/EDM include clustering, factor analysis, network analysis, and domain structure discovery.

While domain structure discovery is quite prominent in research on intelligent tutoring systems, the type of structure discovery most often seen in online learning contexts is a specific type of network analysis called Social Network Analysis (SNA) (Knoke & Yang, 2008). In SNA, data is used to discover the relationships and interactions among individuals, as well as the patterns that emerge from those relationships and interactions. Frequently, in learning analytics, SNA is paired with additional analytics approaches to better understand the patterns observed through network analytics; for example, SNA might be coupled with discourse analysis (Buckingham, Shum, & Ferguson, 2012).

SNA has been used for a number of applications in education. For example, Kay, Maisonneuve, Yacef, and Reimann (2006) used SNA to understand the differences between effective and ineffective project groups, through visual analysis of the strength of group connections. Although this project took place in the context of a face-to-face university class, the data analyzed was from online collaboration tools that could have been used at a distance. SNA has also been used to study how students' communication behaviors in discussion forums

change over time (Haythornthwaite, 2001), and to study how students' positions in a social network relate to their perception of being part of a learning community (Dawson, 2008), a key concern for distance education. Patterns of interaction and connectivity in learning communities are correlated to academic success as well as learner sense of engagement in a course (Macfadyen & Dawson, 2010; Suthers & Rosen, 2011).

Relationship Mining

Relationship mining methods find unexpected relationships or patterns in a large set of variables. There are many forms of relationship mining, but Baker and Siemens (2014) identify four in particular as being common in EDM: correlation mining, association rule mining, sequential pattern mining, and causal data mining. In this section, we will mention potential applications of the first three.

Association rule mining finds if-then rules that predict that if one variable value is found, another variable is likely to have a characteristic value. Association rule mining has found a wide range of applications in educational data mining, as well as in data mining and e-commerce more broadly. For example, Ben-Naim, Bain, and Marcus (2009) used association rule mining to find what patterns of performance were characteristic of successful students, and used their findings as the basis of an engine that made recommendations to students. Garcia, Romero, Ventura, and De Castro (2009) used association rule mining on data from exercises, course forum participation, and grades in an online course, in order to gather data related to effectiveness to provide to course developers. A closely related method to association rule mining is sequential pattern mining.

The goal of sequential pattern mining is to find patterns that manifest over time. Like association rule mining, if-then rules are found, but the if-then rules involve associations between past events (if) and future events (then). For example, Perera, Kay, Koprinska, Yacef, and Zaiane (2009) used sequential pattern mining on data from learners' behaviors in an online collaboration environment, toward understanding the behaviors that characterized successful and unsuccessful collaborative groups. One could also imagine conducting sequential pattern mining to find patterns in course-taking over time within a program that are associated with more successful and less successful student outcomes (Garcia et al., 2009). Sequential patterns can also be found through other methods, such as hidden Markov models; an example of that in distance education is seen in Coffrin, Corrin, de Barba, and Kennedy (2014), a study that looks at patterns of how students shift between activities in a MOOC.

Finally, correlation mining is the area of data mining that attempts to find simple linear relationships between pairs of variables in a data set. Typically, in correlation mining, approaches such as post-hoc statistical corrections are used to set a threshold on which patterns are accepted; dimensionality reduction methods are also sometimes used to first group variables before trying to correlate them to other variables. Correlation mining methods may be useful in situations where there are a range of variables describing distance education and a range of student outcomes, and the goal is to figure out an overall pattern of which variables correspond to many successful outcomes rather than just a single one.

Uses to Benefit Learners

As the examples above indicate, there are several potential uses for data mining and analytics in distance education. These methods can be used to learn a great deal about online and distance students, their learning processes, and what factors influence their outcomes. In our view, the primary uses can be categorized in terms of automated feedback and adaptation.

Automated feedback to students about their learning and performance has a rich history within online education. Many distance education courses today offer immediate correctness feedback on pop-up quizzes or other problem-solving exercises (see Janicki & Liegle, 2001; Jiang et al., 2014), as well as indicators of course progress. Research suggests that providing distance education students with visualizations of their progress toward completing competencies can lead to better outcomes (Grann & Bushway, 2014). Work in recent decades in intelligent tutoring systems and other artificially intelligent technologies shows that there is the potential to provide even more comprehensive feedback to learners. In early work in this area, Cognitive Tutors for mathematics showed students "skill bars," giving indicators to students of their progress based on models of student knowledge (Koedinger, Anderson, Hadley, & Mark, 1997). Skill bars have since been extended to communicate hypotheses of what misconceptions the students may have (Bull, Quigley, & Mabbott, 2006). Other systems give students indicators of their performance across a semester's worth of subjects, helping them to identify what materials need further study prior to a final exam (Kay & Lum, 2005). Some systems provide learners with feedback on engagement as well

as learning, reducing the frequency of disengaged behaviors (Walonoski & Heffernan, 2006). These intelligent forms of feedback are still relatively uncommon within distance education, but have the potential to increase in usage over time.

Similarly, feedback to instructors and other university personnel has a rich history in learning analytics. The Purdue Signals Project (discussed above) is a successful example of how instructors can be empowered with information concerning which students are at risk of unsuccessful outcomes, and why each student is at risk. Systems such as ASSISTments provide more fine-grained reports that communicate to instructors which skills are generally difficult for students (Feng & Heffernan, 2007), influencing ongoing instructional strategies. In the context of distance education, Mazza and Dimitrova (2004) have created visualizations for instructors that represent student knowledge of a range of skills and participation in discussion forums. Another example is TrAVis, which visualizes for instructors the different online behaviors each student has engaged in (May, George, & Prévôt, 2011). These systems can be integrated with tools to support instructors, such as systems that propose types of emails to send to learners (see Arnold, 2010).

Finally, automated intervention is a type of support that can be created based on educational data mining, where the system itself automatically adapts to the individual differences among learners. This is most common in intelligent tutoring systems, where there are systems that automatically adapt to a range of individual differences. Examples include problem selection in Cognitive Tutors (Koedinger et al., 1997), where exercises are

selected for students based on what material they have not yet mastered; pedagogical agents that offer students support for meta-cognitive reasoning (Biswas, Leelawong, Belynne, Viswanath, Schwartz, & Davis. 2004), engagement (Arroyo, Ferguson, Johns, Dragon, Meheranian, Fisher, Barto, Mahadevan, & Woolf, 2007), and collaboration (Dyke, Leelawong, Belynne, Viswanath, Schwartz, & Davis, 2013); and memory optimization, which attempts to return to material at the moment when the student is at risk of forgetting it (Pavlik & Anderson, 2008). Intelligent tutoring systems have been used at scale more often for K-12 education than for higher education, but there are examples of their use in the latter realm (Mitrovic & Ohlsson, 1999; Corbett et al., 2010). The use of intelligent tutor methodologies in distance education can be expected to increase in the coming years, given the acquisition of Carnegie Learning, a leading developer of intelligent tutoring systems, by the primarily distance education for-profit university, the University of Phoenix.

Limitations and Issues to Consider

Educational data mining and learning analytics have been successful in several areas, but there are several issues to consider when applying learning analytics. A key issue, in the authors' opinion, is model validity. As discussed above, it is important that models be validated (tested for reliability) based on genuine outcome data, and that models be validated using data relevant to their eventual use, involving similar systems and populations. The invalid generalization of models creates the risk of inaccurate predictions or responses.

In general, it is important to consider both the benefits of a

correctly applied intervention and the costs of an incorrectly applied one. Interventions with relatively low risk (sometimes called “fail-soft interventions”) are preferable when model accuracy is imperfect. No model is perfect, however; expecting educational at-risk models to be more reliable than standards for first-line medical diagnostics may not be entirely realistic.

Another important consideration is privacy. It is essential to balance the need for high-quality longitudinal data (that enables analysis of the long-term impacts of a student behavior or an intervention) with the necessity to protect student privacy and follow relevant legislation. There is not currently a simple solution to the need to protect student privacy; simply discarding all identifying information protects privacy, but at the cost of potentially ignoring long-term negative effects from an intervention, or ignoring potential long-term benefits.

Conclusion

Data mining and analytics have potential in distance education. In general, as with many areas of education, distance education will be enhanced by the increasing amounts of data now becoming available. There is potential to enhance the quality of course materials, identify at-risk students, and provide better support both to learners and instructors. By doing so, it may be possible to create learning experiences that create a level of individual personalization better than what is seen in traditional in-person courses, instead emulating the level of personalization characteristic of one-on-one tutoring experiences.

Application Exercises

- Name five ways educational data mining and learner analytics could help you design an online learning course.
- As taught in this chapter, “Research suggests that providing distance education students with visualizations of the progress toward completing competencies can lead to better outcomes.” Why do you think this is the case?

References

Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. M. (2003). Help seeking and help design in interactive learning environments. Review of Educational Research 73(2), 277–320.

Allen, I. E., & Seaman, J. (2008). Staying the course: Online education in the United States, 2008. Needham, MA: Sloan Consortium.

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher-level cognition and its relation to visual attention. Human-Computer Interaction 12(4), 439–62.

Arnold, K. E. (2010). Signals: Applying academic analytics. Educause Quarterly, 33(1). Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to

increase student success. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, LAK 2012 (pp. 267–70), New York: ACM.

Arroyo, I., Ferguson, K., Johns, J., Dragon, T., Meheranian, H., Fisher, D., Barto, A., Mahadevan, S., & Woolf, B. P. (2007, June). Repairing disengagement with non-invasive interventions. In Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work (pp. 195–202). IOS Press.

Baker, R. S. (2014). Big data and education. New York: Teachers College, Columbia University

Baker, R., & Siemens, G. (2014). Educational data mining and learning analytics. In K. Sawyer (Ed.) Cambridge handbook of the learning sciences: 2nd Edition (pp. 253 – 274). New York, NY: Cambridge University Press.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining 1(1), 3–17.

Baker, R. S. J. D., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. User Modeling and User-Adapted Interaction 18(3), 287–314.

Ben-Naim, D., Bain, M., & Marcus, N. (2009). A user-driven and data-driven approach for supporting teachers in reflection and adaptation of adaptive tutorials. In the Proceedings of Educational Data Mining 2009 (pp. 21–30).

Biswas, G., Leelawong, K., Belynne, K., Viswanath, K., Schwartz, D., & Davis, J. (2004). Developing learning by teaching environments that support self-regulated learning. In Intelligent tutoring systems, 3220: Lecture notes in computer science, 730–40. Maceió, Brazil: Springer.

Buckingham Shum, S., & Ferguson, R., (2012). Social learning analytics. Educational Technology and Society 15(3), 3–26.

Bull, S., Quigley, S. & Mabbott, A. (2006). Computer-Based formative assessment to promote reflection and learner autonomy, engineering education. Journal of the Higher Education Academy Subject Centre 1(1), 8–18.

Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. Educause Review 42(4), 40.

Champaign, J., Colvin, K. F., Liu, A., Fredericks, C., Seaton, D., & Pritchard, D. E. (2014). Correlating skill and improvement in 2 MOOCs with a student's time on tasks. In Proceedings of the First ACM Conference on Learning @ Scale Conference (pp. 11–20). ACM.

Clow, D. (2014). Data wranglers: Human interpreters to help close the feedback loop. In Proceedings of the Fourth International Conference on Learning Analytics and Knowledge, LAK 2014 (pp. 49–53). New York: ACM.

Coffrin, C., Corrin, L., de Barba, P., & Kennedy, G. (2014). Visualizing patterns of student engagement and performance in MOOCs. In Proceedings of the Fourth International Conference

on Learning Analytics and Knowledge (pp. 83–92). New York: ACM.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4, 253–78. Corbett, A., Kauffman, L., Maclaren, B., Wagner, A., & Jones, E. (2010). A cognitive tutor for genetics problem solving: Learning gains and student modeling. Journal of Educational Computing Research 42(2), 219–39.

d'Aquin, M., & Jay, N. (2013). Interpreting data mining results with linked data for learning analytics: Motivation, case study and directions. In Proceedings of the Third International Conference on Learning Analytics and Knowledge, LAK 2013 (pp. 155–64). New York: ACM.

d'Aquin, M. (2012). Putting linked data to use in a large higher-education organisation. In Proceedings of the Interacting with Linked Data (ILD) Workshop at Extended Semantic Web Conference (ESWC).

Dawson, S. (2008). A study of the relationship between student social networks and sense of community. Educational Technology and Society 11(3), 224–38.

Dennen, V. P. (2005). From message posting to learning dialogues: Factors affecting learner participation in asynchronous discussion. Distance Education 26(1), 127–48.

D'Mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B., & Graesser, A. (2008). Automatic detection of learner's affect

from conversational cues. User Modeling and User Adapted Interaction 18, 45–80.

Dyke, G., Howley, I., Adamson, D., Kumar, R., & Rosé, C. P. (2013). Towards academically productive talk supported by conversational agents. In Productive multimodality in the analysis of group interactions (pp. 459–76). New York: Springer US.

Feng, M., & Heffernan, N. T. (2007). Towards live informing and automatic analyzing of student learning: Reporting in ASSISTment system. Journal of Interactive Learning Research, 18(2), 207–30.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning 4(5), 304–17.

García, E., Romero, C., Ventura, S., & De Castro, C. (2009). An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering. User Modeling and User-Adapted Interaction 19(1–2), 99–132.

Goldstein, P. J., & Katz, R. N. (2005). Academic analytics: The uses of management information and technology in higher education. Educause. Retrieved from https://net.educause.edu/ir/library/pdf/ers0508/rs/ers0508w.pdf

Grann, J., & Bushway, D. (2014). Competency map: Visualizing student learning to promote student success. In Proceedings of

the Fourth International Conference on Learning Analytics And Knowledge (pp. 168–72). ACM.

Haythornthwaite, C. (2001). Exploring multiplexity: Social network structures in a computer- supported distance learning class. The Information Society: An International Journal 17(3), 211–26.

Janicki, T., & Liegle, J. O. (2001). Development and evaluation of a framework for creating web-based learning modules: a pedagogical and systems perspective. Journal of Asynchronous Learning Networks 5(1), 58–84.

Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. Journal of Learning Analytics 1(1), 6–47.

Jiang, S., Warschauer, M., Williams, A. E., O'Dowd, D., & Schenke, K. (2014). Predicting MOOC Performance with Week 1 Behavior. In Proceedings of the 7th International Conference on Educational Data Mining (pp. 273–75).

Kay, J., & Lum, A. (2005). Exploiting readily available web data for scrutable student models, In Proceedings of the 12th International Conference on Artificial Intelligence in Education (pp. 338–45), Amsterdam, Netherlands: IOS Press.

Kay, J., Maisonneuve, N., Yacef, K., & Reimann, P. (2006) The big five and visualisations of team work activity. In Proceedings of the International Conference on Intelligent Tutoring Systems (pp. 197–206).

Kim, J., Guo, P. J., Seaton, D. T., Mitros, P., Gajos, K. Z., & Miller, R. C. (2014, March). Understanding in-video dropouts and interaction peaks in online lecture videos. In Proceedings of the First ACM Conference on Learning@ Scale Conference (pp. 31–40). ACM.

Kitsantas, A., & Chow, A. (2007). College students' perceived threat and preference for seeking help in traditional, distributed, and distance learning environments. Computers and Education 48(3), 383–95.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In Proceedings of the Third International Conference on Learning Analytics and Knowledge (pp. 170–79). ACM.

Knoke, D., & Yang, S. (eds.). (2008). Social network analysis (vol. 154), 2nd Ed. Thousand Oaks, CA: Sage.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. International Journal of Artificial Intelligence in Education 8, 30–43.

Koedinger, K. R., Baker, R.S. J. D., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. Baker, R. S. J. D. (eds.), Handbook of educational data mining. Boca Raton, FL: CRC Press (pp. 43–56).

Kovacic, Z. (2010). Early prediction of student success: Mining students' enrollment data. In Proceedings of Informing Science and IT Education Conference (InSITE) 2010 (pp. 647–65).

Lam, W. (2004). Encouraging online participation. Journal of Information Systems Education, 15(4), 345–48.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. Computers and Education, 54(2), 588–99.

May, M., George, S., & Prévôt, P. (2011). TrAVis to enhance online tutoring and learning activities: Real-time visualization of students tracking data. Interactive Technology and Smart Education 8(1), 52–69.

Mazza, R., & Dimitrova, V. (2004, May). 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 and Posters (pp. 154–61). ACM.

Ming, N. C., & Ming, V. L. (2012). Predicting student outcomes from unstructured data. In Proceedings of the 2nd International Workshop on Personalization Approaches in Learning Environments (pp. 11–16).

Mitrovic, A., & Ohlsson, S. (1999). Evaluation of a constraint-based tutor for a database. International Journal of Artificial Intelligence in Education, 10, 238–56.

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014) Population validity for educational data

mining models: A case study in affect detection. British Journal of Educational Technology, 45(3), 487–501.

O'Malley, J., & McCraw, H. (1999). Students' perceptions of distance learning, online learning and the traditional classroom. Online Journal of Distance Learning Administration, 2(4).

O'Neill, K., Singh, G., & O'Donoghue, J. (2004). Implementing elearning programmes for higher education: A review of the literature. Journal of Information Technology Education Research, 3(1), 313–23.

Palazuelos, C., García-Saiz, D., & Zorrilla, M. (2013). Social network analysis and data mining: An application to the e-learning context. In J.-S. Pan, S.-M. Chen, & N.-T. Nguyen (eds.). Computational collective intelligence. technologies and applications (pp. 651–60). Berlin and Heidelberg: Springer.

Paquette, L., de Carvalho, A. M. J. A., Baker, R. S., & Ocumpaugh, J. (2014). Reengineering the feature distillation Process: A case study in the detection of gaming the system. In Proceedings of the 7th International Conference on Educational Data Mining (pp. 284–87).

Patel, C., & Patel, T. (2005). Exploring a joint model of conventional and online learning systems. E-Service Journal, 4(2), 27–46.

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. Journal of Experimental Psychology Applied, 14(2), 101.

Perera, D., Kay, J., Koprinska, I., Yacef, K., & Zaiane, O. R.

(2009). Clustering and sequential pattern mining of online collaborative learning data. IEEE Transactions on Knowledge and Data Engineering, 21(6), 759–72.

Rabbany, R., Takaffoli, M., & Zaïane, O. R. (2011). Analyzing participation of students in online courses using social network analysis techniques. In Proceedings of Educational Data Mining (pp. 21–30).

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to Expert Systems with Applications, 33(1), 135–46.

Romero, C., Romero, J. R., & Ventura, S. (2013). A survey on pre-processing educational data. In Educational data mining (pp. 29–64). Berlin: Springer International Publishing.

San Pedro, M. O. Z., Baker, R. S. J. D., Bowers, A. J., & Heffernan, N.T. (2013). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In Proceedings of the 6th International Conference on Educational Data Mining (pp. 177–84).

Sao Pedro, M., Baker, R. S. J. D., & Gobert, J. (2012). Improving construct validity yields better models of systematic inquiry, even with less information. In Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UMAP 2012), (pp. 249–60).

Scheuer, O., & McLaren, B. M. (2012). Educational data mining. In Encyclopedia of the Sciences of Learning (pp. 1075–79). New York: Springer US.

Sharkey, M. (2011). Academic analytics landscape at the University of Phoenix. In Proceedings of the First International Conference on Learning Analytics and Knowledge, LAK 2011 (pp. 122–26). ACM.

Suthers, D., & Rosen, D. (2011). A unified framework for multilevel analysis of distributed learning. In Proceedings of the 1st International Conference on Learning Analytics and Knowledge (pp. 64–74).

Veeramachaneni, K., Dernoncourt, F., Taylor, C., Pardos, Z., & O'Reilly, U. M. (2013). Developing data standards for MOOC data science. In AIED 2013 Workshops Proceedings (p. 17). Berlin: Springer.

Walonoski, J. A., & Heffernan, N. T. (2006). Prevention of off-task gaming behavior in intelligent tutoring systems. In Intelligent tutoring systems (pp. 722–24). Berlin: Springer.

 Please complete this short survey to provide feedback on this chapter: http://bit.ly/LAandED

Suggested Citation

Baker, R. S. & Inventado, P. S. (2018). Educational Data Mining and Learning Analytics: Potentials and Possibilities for Online Education. In R. E. West, *Foundations of Learning and Instructional Design Technology: The Past, Present, and Future of Learning and Instructional Design Technology*. EdTech Books. Retrieved from https://edtechbooks.org/lidtfoundations/educational_data_mining_and_learning_analytics

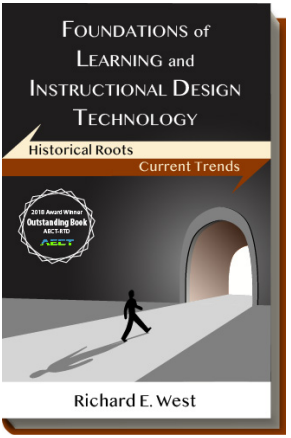
Chapter Copyright Notice

Ryan S. Baker

Dr. Ryan S. Baker is an associate professor in the graduate school of education at the University of Pennsylvania. Dr. Baker develops data mining methods for monitoring the interaction between students and educational software. He is currently serving as the director of the Penn Center for Learning Analytics. He was also a founding president of the International Educational Data Mining Society and is currently a member of the board of directors of that society. He was previously an associate professor at Columbia University in the Department of Human Development. Dr. Baker received his PhD from Carnegie Mellon University.

Paul Salvador Inventado

Dr. Paul Salvador Inventado is a post-doctoral researcher at the School of Design at Carnegie Mellon University, where he is using data mining to help math tutors become more effective. He was an assistant professor with the Department of Software Technology at De La Salle University in the Philippines. Dr. Inventado received his PhD from Osaka University in Japan and was a recipient of the Monbukagakusho Scholarship.



West, R. E. (2018). *Foundations of Learning and Instructional Design Technology: The Past, Present, and Future of Learning and Instructional Design Technology* (1st ed.). EdTech Books. Retrieved from https://edtechbooks.org/lidtfoundations