# Educational Data Mining and Learning Analytics

*Contributors: Mimi Recker, Andrew Krumm, Mingyu Feng, Shuchi Grover, Ken Koedinger*
Questions, or want to add to this topic or to a new topic? Contact CIRCL.

## Overview

*Educational data mining* (EDM) is the use of multiple analytical techniques to better understand relationships, structure, patterns, and causal pathways within complex datasets. Learning Analytics (LA) is a closely related endeavor, with somewhat more emphasis on simultaneously investigating automatically collected data along with human observation of the teaching and learning context. Overall, cyberlearning emphasizes the integration of learning sciences theories with these techniques in order to improve the design of learning systems and to better understand how people learn within them.

Educational systems are increasingly engineered to capture and store data on users' interactions with a system. These data (e.g., big data, system log data, trace data) can be analyzed using statistical, machine learning, and data mining techniques. The development of computational tools for data analysis, standardization of data logging formats, and increased computation/processing power is enabling learning scientists to investigate research questions using this data (Baker & Siemens, in press).

Research goals which EDM/LA can address include:

1. Predicting students' future learning by creating models that incorporate information such as students' knowledge, behavior, motivation, and attitudes.
2. Discovering or improving models that characterize the subject matter to be learned (e.g. math, science, etc.), identify fruitful pedagogical sequences, and suggest how these sequences might be adapted to students' needs.
3. Studying the effects of varied pedagogical enhancements on student learning.
4. Advancing scientific knowledge about learning and learners through building models of learning processes that incorporate data about students, teachers, understanding of subject matter, pedagogies, and principles from learning sciences.

5. Supporting learning for all students by adapting learning resources to fit the particular needs identified, including adaptations for individual students when warranted.

In addition, researchers are expanding EDM/LA to new frontiers, such as studying learning in constructionist research where the lack of formal structure in learning environments (such as games and maker spaces) make traditional assessments difficult to implement. Another new frontier for EDM/LA is understanding collaboration in formal and informal learning environments.

Large scale use of learning management systems, games, virtual worlds, augmented reality, simulations, and constructionist spaces in learning, as well as the emergence of online open learning materials (such as Khan Academy) and courseware (including MOOCs) has fueled research in EDM/LA. The NSF-funded Pittsburgh Science of Learning Center (PSLC) or 'LearnLab' has spearheaded key research in this field in the past decade. The PSLC Datashop is an important resource serving as a central repository to secure and store research data and provide a set of analysis and reporting tools. Early work in EDM/LA by the PLSC team (Koedinger, Corbett, and others) was conducted in the context of Intelligent Tutoring Systems. The cognitive models of learning they used and developed (drawing on earlier work by John Anderson) have contributed to understanding the design of adaptive, data-rich learning systems, especially in STEM subjects. Other noteworthy efforts include (among others) the development of tools and techniques for mining data and making inferences about non-cognitive aspects of learning (Ryan Baker and colleagues); growing an understanding of conversation analytics (Carolyn Rose's group at CMU); analytics in games (Constance Steinkuehler and Kurt Squire; Taylor Martin and colleagues); LA to serve teacher needs (Mimi Recker et al.); studying collaborative processes and social learning analytics (Dan Suthers; Simon Buckingham Shum; and others); and multi-model learning analytics in constructionist spaces (Paulo Blikstein and colleagues).

## Issues

As practiced in cyberlearning, EDM is often deeply interdisciplinary. Thus in planning EDM efforts, a critical question is how to support multiple disciplinary specialists work together in order to (1) address the most pressing problems of practice, (2) collect useful data both online and through more traditional techniques, (3) analyze data using appropriate techniques to rigorously answer the question at hand, (4) interpret results and elicit feedback from multiple stakeholders to generate appropriate implications for action, and

(5) continue to represent and display data in ways that support valuable uses of the data by researchers and practitioners.

Issues of privacy and ethics of data use are yet to be resolved. Standards that balance the need for data privacy with the need to link student and teacher data across distributed systems need to be established as also mechanisms for informing users about what data are collected in addition to providing users a means to control access, anonymize, and opt-in or out.

An important trend in research, amplified by EDM/LA, is to take products to scale first, and then begin conducting research. (In contrast, a traditional pathway involves slowly scaling research on educational technology over a decade or more.) The Evidence Framework (Barbara/Bakia et al.) provides valuable guidance for research approaches when working in the scale first/then study paradigm.

Overall, EDM draws on traditional statistical techniques and shares further challenges with other analytic uses of research data, such as:

1. Combining needed data from different systems, which can be difficult.
2. Achieving construct validity and interpretability of results.
3. Understanding consequential validity and use of results to drive decisions.
4. Deciding whether use of data to drive high-stakes and/or low-stakes decisions is warranted.
5. Establishing safeguards for privacy and ethics of data use.

Several other key issues have been identified by the Learning Analytics Workgroup report (Pea, 2014). These include foregrounding the needs of learners and challenges of educators, defining success metrics for personalized learning while recognizing that different outcomes of the learning process are relevant for different stakeholders, and creating the necessary infrastructure for supporting research in learning analytics.

## Projects

Examples of NSF Cyberlearning projects that overlap with topics discussed in this primer (see project tag map).

### Analytics/data mining

- Doctoral Consortium for the 2016 Learning Analytics and Knowledge Conference
- CAP: Data Consortium Fellows: A Mentorship Program to Expand the Cyberlearning Data Analytics Community
- CAP: Doctoral Consortium for the 2015 Learning Analytics and Knowledge Conference
- CAP: Advancing Technology and Practice for Learning Reading and Writing Skills in Secondary Science Education
- EAGER: Automatic Classification of Programming Difficulties by Mining Programming Events
- Badge-Based STEM Assessment: Current Terrain and the Road Ahead
- EXP: Transforming High School Science via Remote Online Labs
- EXP: Collaborative Research: Fostering Ecologies of Online Learners through Technology Augmented Human Facilitation
- EXP: Learning Lens: An Evidence-Centered Tool for 21st Century Assessment
- DIP: Collaborative Research: Taking Hands-on Experimentation to the Cloud: Comparing Physical and Virtual Models in Biology on a Massive Scale

More posts: analyticsdata mining

## Readings

Key readings documenting the thinking behind the concept, important milestones in the work, foundational examples to build from, and summaries along the way.

The Learning Analytics WorkGroup Report (Pea, 2014) is a great synthesis of the issues and current state of the field of Educational Data Mining and Learning Analytics.

A new report, Accelerating Science: A Computing Research Agenda by Honavar, Hill, and Yelick (2016) seeks to articulate a research agenda for developing cognitive tools and leveraging big data to augment human intellect and enable new modes of discovery — and may inspire some interesting ways to think about smart and connected communities of learners.

Baker, R., Siemens, G. (2014). Educational data mining and learning analytics. In Sawyer, K. (Ed.) Cambridge Handbook of the Learning Sciences: 2nd Edition.

Baker, R. S. J. D., & 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. (in press). Data Mining for Education. To appear in McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition). Oxford, UK: Elsevier.

Borgman, C. L., Abelson, H., Dirks, L., Johnson, R., Koedinger, K. R., Linn, M. C., … Szalay,A. (2008). Fostering learning in the networked world – The CyberLearning opportunity and challenge: A 21st century agenda for the National Science Foundation (Report of the NSF Task Force on CyberLearning). Arlington VA: NSF.

Computing Research Association. (2005). Cyberinfrastructure for education and learning for the future: A vision and research agenda. Washington, DC: Computing Research Association.

D'Mello, S. K., Picard, R. W., and Graesser, A. C. (2007) Towards an Affect-Sensitive AutoTutor. Special issue on Intelligent Educational Systems – IEEE Intelligent Systems, 22(4), 53-61.

Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A. and Rosé, C. P. (2015). Data mining and education. WIREs Cogn Sci, 6: 333–353.

Pea, R. (2014). The Learning Analytics Workgroup: A Report on Building the Field of Learning Analytics for Personalized Learning at Scale.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33(1), 135-146. doi: 10.1016/j.eswa.2006.04.005.

Romero C.R., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man and Cybernetics, part C: Applications and Reviews, 40(6), 601-618.

Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) (2010) Handbook of Educational Data Mining. Boca Raton, FL: CRC Press.

U.S. Department of Education, Office of Educational Technology (2012). Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief. Washington, D.C., 2012.

Executive Office of President (2014). Big Data: Seizing Opportunities, Preserving Values. The White House, Washington, DC.

## Resources

Organizations:

- International Educational Data Mining Society (publishes Journal of Educational Data Mining and convenes the International Conference on Educational Data Mining)
- Society for Learning Analytics Research, which sponsors LAK conferences
- Learning Analytics Community Exchange (LACE), an EU funded project involving nine partners from across Europe

Online courses and video:

- Big Data in Education MOOC
- Big Data in Education MOOC (archive)

Tutorials:

- LearnLab 2013 (PSLC) Resource Archive
- Introduction to Data Mining for Educational Researchers

## Online Data and Analytic Sharing

LearnSphere integrates existing and new educational data infrastructures–including MOOC data, discourse data, and the LearnLab DataShop–to offer a world class repository of education data. Funded by the NSF Data Structure Building Blocks (DIBBS) program, it is the first education-focused DIBBs project. LearnSphere is led by Ken Koedinger at CMU in collaboration with MIT, Stanford, and U Memphis.

The LearnLab DataShop has 100s data sets on student learning from educational technology and associated analytics.

The KDD Cup 2010 competition has large student data sets available and will evaluate submitted prediction models.

Other potentially relevant repositories the Linguistic Data Consortium at UPenn and the DataBrary.