Learning Analytics and Educational Data Mining: Towards Communication and Collaboration

George Siemens
Technology Enhanced Knowledge Research Institute
Athabasca University
gsiemens@athabascau.ca

Ryan S J.d. Baker
Department of Social Science and Policy Studies
Worcester Polytechnic Institute
rsbaker@wpi.edu

ABSTRACT
Growing interest in data and analytics in education, teaching, and learning raises the priority for increased, high-quality research into the models, methods, technologies, and impact of analytics. Two research communities – Educational Data Mining (EDM) and Learning Analytics and Knowledge (LAK) have developed separately to address this need. This paper argues for increased and formal communication and collaboration between these communities in order to share research, methods, and tools for data mining and analysis in the service of developing both LAK and EDM fields.

Categories and Subject Descriptors
H.2.8 [**Database Applications**]: Data Mining

General Terms
Algorithms, Human Factors, Measurements.

Keywords
Educational data mining, learning analytics and knowledge, collaboration

1. INTRODUCTION
In education, the emergence of "big data" through new extensive educational media, combined with advances in computation [1] holds promise for improving learning processes in formal education, and beyond as well. Increasingly, very large data sets are available from students' interactions with educational software and online learning - among other sources - with public data repositories supporting researchers in obtaining this data [2].

Two distinct research communities, Educational Data Mining (EDM) and Learning Analytics and Knowledge (LAK), have developed in response.

The first workshop on Educational Data Mining was held in 2005, in Pittsburgh, Pennsylvania. This was followed by annual workshops and, in 2008, the 1st International Conference on Educational Data Mining, held in Montreal, Quebec. Annual conferences on EDM were joined by the Journal of Educational Data Mining, which published its first issue in 2009, with Kalina Yacef as Editor. The first Handbook of Educational Data Mining was published in 2010 [7]. In the summer of 2011, the International Educational Data Mining Society (IEDMS) (http://www.educationaldatamining.org/) was formed to "promote scientific research in the interdisciplinary field of educational data mining", organizing the conferences and journal, and the free open-access publication of conference and journal articles. The EDM community brings together an inter-disciplinary community of computer scientists, learning scientists, psychometricians, and researchers from other traditions. A first review of research in EDM was presented by Romero & Ventura [3], followed by a theoretical model proposed by Baker & Yacef [4]. A very comprehensive review of EDM research can be found in [6].

The Learning Analytics and Knowledge conference series was initiated in early summer, 2010, with the development of global steering and program committees (https://tekri.athabascau.ca/analytics/node/5). The conference explicitly emphasized its role as bridging the computer science and sociology/psychology of learning in declaring that the "technical, pedagogical, and social domains must be brought into dialogue with each other to ensure that interventions and organizational systems serve the needs of all stakeholders." The first conference, held in Banff, Canada attracted over 100 participants, with proceedings published in ACM [5], validating interest in inter-disciplinary approaches to analytics in learning. In summer of 2011, the Society for Learning Analytics (SoLAR -- http://www.solaresearch.org/) was formed to provide oversight for the conference, develop and advance a research agenda in learning analytics, as well as advocate for, and educate in the use of, analytics in learning.

With growing research interest in learning analytics and educational data mining, as well as the rapid development of software and analytics methods, it is important for researchers and educators to recognize the unique attributes of each community. While LAK and EDM share many attributes and have similar goals and interests, they have distinct technological, ideological, and methodological orientations. As schools, university, and corporate learning and curriculum organizations begin to adopt data mining and analytics, both LAK and EDM can benefit from building off work occurring in the other community. This paper details the overlap between these different communities and discusses the benefits of increased communication and collaboration.

2. SIMILARITIES BETWEEN COMMUNITIES
The EDM and LAK communities are defined in relatively similar ways. The International Educational Data Mining Society defines EDM as follows: "Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in."

The Society for Learning Analytics Research defines Learning Analytics as: "… the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs."

EDM and LAK both reflect the emergence of data-intensive approaches to education. In sectors such as government, health care, and industry, data mining and analytics have become increasingly prominent for gaining insight into organizational activities. Drawing value from data in order to guide planning, interventions, and decision-making is an important and fundamental shift in how education systems function. LAK and EDM share the goals of improving education by improving assessment, how problems in education are understood, and how interventions are planned and selected. Extensive use by administrators, educators, and learners of the data produced during the educational process raises the need for research-based models and strategies. Both communities have the goal of improving the quality of analysis of large-scale educational data, to support both basic research and practice in education.

3. KEY DISTINCTIONS BETWEEN COMMUNITIES

The similarities between EDM and LAK suggest numerous areas of research overlap. Additionally, the organizational deployment of EDM and LAK requires similar data and researcher skill-sets. However, these two communities have different roots and some distinctions are important to note. Table 1 shows some of the key differences between the communities. It is important to note that these distinctions are meant to represent broad trends in the two communities; many EDM researchers conduct research that could be placed on the LAK side of each of these distinctions, and many LAK researchers conduct research that could be placed on the EDM side of these distinctions. By identifying these distinctions, we hope to identify places where the two communities can learn from each other, rather than defining the communities in an exclusive fashion. Certainly, communities that grow organically as these two communities have done will not have rigid edges between what work appears in the two communities.

One key distinction is found in the type of discovery that is prioritized. In both communities, research can be found that uses automated discovery and research can be found that leverages human judgment through visualization and other methods. However, EDM has a considerably greater focus on automated discovery, and LAK has a considerably greater focus on leveraging human judgment. Even in research which combines these two directions, this preference can be seen; EDM research which leverages human judgment in many cases does so to provide labels for classification, while LAK research which uses automated discovery often does so in the service of informing humans who make final decisions.

This difference is associated with another difference between the two communities: the type of adaptation and personalization typically supported by the two communities. In line with the greater focus on automated discovery in EDM, EDM models are more often used as the basis of automated adaptation, conducted by a computer system such as an intelligent tutoring system. By contrast, LAK models are more often designed to inform and empower instructors and learners.

A third difference, and an important one, is the distinction between holistic and reductionistic frameworks. It is much more typical in EDM research to see research which reduces phenomena to components and analyzing individual components and relationships between them. The "discovery with models" paradigm for EDM research discussed in [4] is a clear example of this paradigm. By contrast, LAK researchers typically place a stronger emphasis on attempting to understand systems as wholes, in their full complexity. The debate between reductionist and holistic paradigms has often paralyzed discussion between education researchers from different "camps"; encouraging discussion between EDM and LAK researchers is a key way to prevent this common split from reducing what EDM and LAK researchers can learn from one another.

Two other differences are in the most common origins and methods of researchers in these two communities. Researchers' origins tend to drive the preferred approaches discussed above, and these preferred approaches in turn drive preferred methods. Greater detail on these issues is given in Table 1.

Table 1: A brief comparison of the two fields

	LAK	EDM
Discovery	Leveraging human judgement is key; automated discovery is a tool to accomplish this goal	Automated discovery is key; leveraging human judgment is a tool to accomplish this goal
Reduction & Holism	Stronger emphasis on understanding systems as wholes, in their full complexity	Stronger emphasis on reducing to components and analyzing individual components and relationships between them
Origins	LAK has stronger origins in semantic web, "intelligent curriculum," outcome prediction, and systemic interventions	EDM has strong origins in educational software and student modeling, with a siginficiant community in predicting course outcomes
Adaption & Personalization	Greater focus on informing and empowering instructors and learners	Greater focus on automated adaption (e.g. by the computer with no human in the loop)
Techniques & Methods	Social network analysis, sentiment analysis, influence analytics, discourse analysis, learner success prediction, concept analysis, sensemaking models	Classification, clustering, Bayesian modeling, relationship mining, discovery with models, visualization

4. CALL FOR COMMUNICATION AND COLLABORATION: EDM and LAK

There is a positive value to having different communities engaged in how to exploit "big data" to improve education. In particular, different standards and values for "good research" and "important research" exist in each community, allowing creativity and advancement that might not otherwise occur in a single, monolithic research culture. For example, EDM researchers have placed greater focus on issues of model generalizability (e.g. multi-level cross-validation, replication across data sets). By contrast, LAK researchers have placed greater focus on addressing needs of multiple stakeholders with information drawn from data. Each of these issues are important for the long-term success of both fields, a key opportunity for the two communities to learn from one another.

Friendly competition between the two communities will keep both communities vigorous, and is generally beneficial for science. This type of competition has occurred in the past, such as in the split between the International Conference on the Learning Sciences and the International Conference on Artificial Intelligence in Education in 1992. Research networks are increasingly global, as reflected by the multi-national executive committees of IEDMS/EDM and SoLAR/LAK, but reflect different nations to a significant degree. Hence, the existence of both communities broadens the number of researchers working and collaborating in the broader area of data-driven discovery in education. At the same time, it is very important to keep competition healthy. Healthy competition requires that both communities disseminate their research to each other through their respective conferences and journals to ensure awareness of important ideas and advances occurring in the other community. The two communities must communicate, in order to bring the greatest possible benefits to educational practice and the science of learning.

5. CONCLUSION

Given the overlaps in research interests, goals, and approaches between the EDM and LAK communities, the authors of this paper recommend that the executive committees of SoLAR and IEDMS formalize approaches for dissemination of research and enacting cross-community ties. A formal relationship will allow each community to continue developing their specialized and distinct research methods and tools, while simultaneously increasing opportunities for collaborative research and sharing of research findings between the communities.

This alliance would also strengthen our opportunities to influence non-academic research and practice. A particular concern now facing both EDM and LAK is the rapid development of analytics and data mining tools by commercial organizations that do not build off of either community's expertise, algorithms, and research results. To give one example, there is increasing consensus in the EDM community that cross-validation needs to be conducted at multiple levels (in particular the student level, but also the classroom and lesson/unit levels). However, there is not direct support for this goal in many of the data mining/analytics tools now emerging. To the extent that EDM and LAK can jointly articulate quality standards for research in this area, it may be possible to more effectively communicate these standards to the wider community of tool-developers and analytics practitioners, as well as the broader research community. As such, both communities would be facilitated in communicating their vision for data-driven science and practice in the field of education.

Both the LAK and EDM communities anticipate that the impact of data and analytics within education will be transformative at primary, secondary, and post-secondary levels. An open, transparent research environment is vital to driving forward this important work. As connected, but distinct, research disciplines, EDM and LAK can provide a strong voice and force for excellence in research in this area, guiding policy makers, administrators, educators, and curriculum developers, towards the deployment of best practices in the upcoming era of data-driven education.

6. ACKNOWLEDGMENTS

Our thanks to Jaclyn Ocumpaugh, and the anonymous reviewers for their valuable input and assistance on this paper.

7. REFERENCES

[1] Mayer, M. (2009) *Innovation at Google: The physics of data [PARC forum]* (11 August, 2009: 3:59 mark). Available from < http://www.slideshare.net/PARCInc/innovation-at-google-the-physics-of-data>

[2] 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 Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.)*Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press, pp.43-56.

[3] Romero, C., Ventura, S. (2007). Educational Data Mining: A Survey from 1995 to 2005. *Expert Systems with Applications*, *33*, 125-146.

[4] 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.

[5] Gasevic, D., Conole, G., Siemens, G., Long, P. (Eds). (2011) LAK11: International Conference on Learning Analytics and Knowledge, Banff, Canada, 27 February - 1 March 2011.

[6] Romero, C., Ventura, S. (2010) Educational Data Mining: A Review of the State-of-the-Art. IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews. *40* (6), 601-618.

[7] Romero, C., Ventura, S., Pechenizky, M., Baker, R. (2010) *Handbook of Educational Data Mining*. 2010. Editorial Chapman and Hall/CRC Press, Taylor & Francis Group. Data Mining and Knowledge Discovery Series.