Guest Editorial

Problems in Big Data Analytics in Learning

Krishna Madhavan[a] and Michael C. Richey[b]

[a]*Purdue University,* [b]*The Boeing Company*

Introduction

When the National Academy of Engineering issued its grand challenges – specifically the one on "advancing personalized learning" (National Academy of Engineering, 2008) – it called for the development of new instrumentation, tools, and methodologies to bring learning closer to the learners and their personal choices. Big data plays a critical role in beginning to meet this grand challenge. A tremendous amount of data on students' learning and behavioral experiences is captured in a wide variety of institutional systems. These data range from student demographics and socio-economic backgrounds to data about academic progress and extend down to individual mouse clicks when students are accessing course materials or their time spent viewing a screen of course information. Data captured from automated software-based learning environments in combination with more traditional forms of educational data provide a unique opportunity to understand how learning occurs and to engineer these processes in unprecedented ways.

Data by themselves have only limited value. True transformation of educational ecosystems lies in converting these data into actionable intelligence (meaning insights and knowledge that enable learners and other stakeholders to act). Learning data vary significantly in modality (such as visual, auditory, and tactile) and dimensionality (range of observed characteristics). More important, the contexts from which these data are derived may vary enormously. It is this unique combination of factors that makes big data in learning such an interesting research artifact.

The ability to couple data about learners (actors or agents) and the system structure (courses, schools, university, or industry settings) within which they function is powerful. Many proponents of big data in learning believe that these data, considered as a single unified ecosystem, could eventually uncover the distributed nature of human cognition that is embedded in a thick network of human behaviors. The ability to shed light onto the so-called "ghost in the machine" (Koestler, 1967), where thoughts are embodied in learner actions, has a powerful appeal. In this guest editorial, we discuss the notion of big data and its potential for transforming learning and educational ecosystems.

Defining Big Data

William Gibson coined the term "cyberspace" to describe a

> consensual hallucination experienced daily by billions of legitimate operators, in every nation, by children being taught mathematical concepts. . . . A graphic representation of data abstracted from the banks of every computer in the human system. Unthinkable

Journal of Engineering Education
January 2016, Vol. 105, No. 1, pp. 6–14

DOI 10.1002/jee.20113

> complexity. Lines of light ranged in the non space of the mind, clusters and constellations of data. Like city lights receding. (1984, p. 51)

The notion of "big data" is as old as the Internet itself and is foundational to our technology-driven way of life. Technological leaps in software services such as search engines, online social networks, and online advertising coupled with significant progress in hardware technology – particularly mobile technologies that integrate a wide range of sensors – have pushed humankind into a data generation overdrive. Lohr (2012) pointed out that big data is

> a meme and a marketing term, for sure, but also shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions. There is a lot more data, all the time, growing at 50 percent a year, or more than doubling every two years.

Given the genesis of big data as a technology paradigm, Cuzzocrea, Song, and Davis provided a much more technical and traditional definition:

> "Big Data" refers to enormous amounts of *unstructured data* produced by high-performance applications falling in a wide and heterogeneous family of application scenarios: from scientific computing applications to social networks, from e-government applications to medical information systems, and so forth. (2011, p. 101)

While this definition provides a bit more insight about the nature of the data generated, it lacks several important components that are needed to completely define big data – namely, fit and conformance with traditional database architectures, ease of data mobility, and value derivation.

Dumbill (2012) described big data as

> data that exceed the processing power of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it.

He also presents the ideas of 3Vs – volume, velocity, and variety. Volume simply refers to the large amount of data and is a critical aspect of big data. It is not that data is just growing in volume – but it is also growing at an extremely rapid pace (velocity). For example, consider the velocity and volume with which the metadata associated with the number of queries needed to refresh the Facebook Newsfeeds of every user in the world grows every minute. Variety refers not only to the types of data but also to the lack of structure within the data (unstructured data). Subsequently the big data community added veracity (trustworthiness of the data) and value (ability to transform data into insights) to round out the dimensions of big data – now referred to as the 5Vs.

New Types of Data and Workflows

Beyond the five dimensions of big data just mentioned, the power of big data perhaps lies in its ability to allow humans to envision a new way to generate and understand knowledge. boyd and Crawford pointed out that

> Just as Ford changed the way we made cars – and then transformed work itself – Big Data has emerged a system of knowledge that is already changing the objects of

> knowledge, while also having the power to inform how we understand human networks and community. (2012, p. 665)

The notion of "big" – specifically volume and velocity – is somewhat misleading in the context of data related to educational contexts. Given that many Internet companies involved in search or social media accumulate and process several petabytes of data every day (a fairly conservative estimate), the educational enterprise does not deal with similar volumes or velocity of data. Rather, in the context of education, another more important aspect of big data needs to be considered. Through better instrumentation, work on big data has led to the observation, archiving, and processing of new types of data that were previously unobserved or unobservable. Lohr (2012) highlighted this important point when he stated,

> It is not just more streams of data, but entirely new ones [read: types of data]. For example, there are now countless digital sensors worldwide in industrial equipment, automobiles, electrical meters and shipping crates.

Similarly in the educational context, new data types or streams of data – such as learner behavior data from the learning environment (electronic or physical environments), institutional data about student learning, and data regarding co-curricular activities – are now increasingly collected.

The availability of these new types or streams of data leads to the question, what sorts of research problems can be developed to fully take advantage of data of this scope? In this guest editorial, we focus on research and development efforts that can provide direct value to learners. In the following, we discuss one paradigm-shifting research issue.

Real-Time Analytics

When it comes to big data analytics, the educational research enterprise follows, by and large, a linear model. Figure 1 shows this linear flow of educational research with very little immediately benefiting the learners who generated the data in the first place. Students interact with learning environments that generate rich behavioral or interaction data. In virtual environments such as in MOOCs and learning management systems, much of this data is in the form of clickstream (Carr, 2000; Moe & Fader, 2004). We define *clickstream* as a proxy for user behavior that not only records that a click was made by the user on an electronic object, but also captures the associated metadata, such as the system time when the click was made (for standardization), the object on which the click event occurred, the dwell time on the clicked entity, and a few other navigational details. Insofar as clickstreams embody behaviors, they are also a manifestation of learner intent. For this reason, the idea that these data represent only a stream of user behaviors sells their potential somewhat short. Perhaps a better name for these data would be *cognitive streams* – a term that better represents the true analytic potential of these data to lead to actions that benefit learners.

In addition to clickstreams, data from many different learning contexts are part of the educational big data ecosystem. They include those derived from the institutional records of students – for example, demographics, courses taken, grades in these courses, academic trajectory, co-curricular activities, and ID card swipe events at various campus locations and events. For an exhaustive list of broad educational data types and research studies based on them, see Romero and Ventura (2010, Table 2, p. 603). Nonetheless, it is important to point out here that just because we are now able to store all these data in some database system does not of itself make these data useful for improving learning. The combination of analytic engines, decision support models, and predictive capability applied to the data is

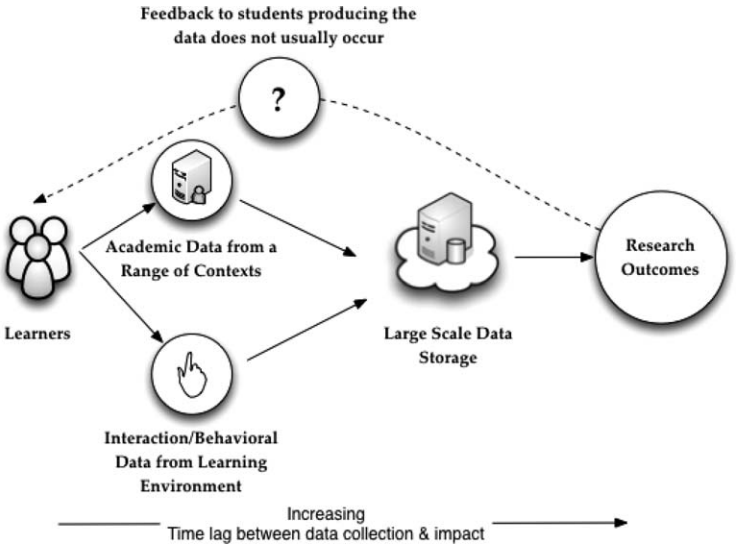

Figure 1 Linear model of big data research in education showing latency between data collection to research impact with questionable flow of insights back to learners who generated the data.

what makes the data valuable. The fundamental challenge here is not one of scale of data, but rather that the data from institutional settings, virtual learning environments, and other experimental contexts have not been reconciled, indexed, or organized very well. So the value of all these data to enable learning is difficult to determine. There are some efforts to bring order to this rather chaotic landscape. But a discussion of those efforts is outside the scope of this editorial.

Focusing back on Figure 1, once the data are collected in data repositories (or databases), they either are never utilized for any meaningful research, or, more often, analyzed and presented as traditional research products, such as peer-reviewed papers and conference proceedings. This figure illustrates the latency and sometimes missing pathway (dotted line) for providing feedback to learners based on their actions. Immediate feedback plays a critical role in learning and big data could enable just that.

Feedback is a fundamental concept in the design, analysis, and control of systems from neurons to societies (Aström & Murray, 2009; Tustin, 1952; Wiener, 1961). An essential mechanism in theories of learning, it has been investigated from many perspectives since the start of the twentieth century. For example, the foundational work of Thorndike (1913) on the law of effect showed that stimuli that followed an organism's response affected learning. Thorndike's work led to behaviorism. Newell and Simon's (1972) work on means-ends analysis using rules provided an alternative to state-based representations to reduce differences between states. This work is a key part of cognitive and information processing models of learning. Vygotsky's (1978) work on the scaffolding and feedback provided by parents and peers in a zone of proximal development has been the basis of many socially based theories of learning. In the context of big data in learning, the problem is that although learning data are continuously captured from and about students, we lack the ability to provide immediate and

sound feedback by way of action recommendations to learners (and indeed other stakeholders such as teachers, administrators, and academic advisors).

There are some efforts to provide learners with more real-time feedback based on the data collected. For example, work on providing real-time feedback with recommender systems – systems that recommend potential selection pathways based on user choices or actions – has a long history in the technology-enhanced learning context (Manouselis, Drachsler, Verbert, & Duval, 2012; Verbert et al., 2012). It is also important to note that the recommender systems used in educational settings usually have very restrictive contexts within which they function. At present, most education research using big data is somehow an antithesis to the overall paradigm that the availability of these data represents. When big data from learning contexts is treated as a stored, static commodity, the results of whose analyses will not immediately improve feedback and action recommendations provided to learners; it then loses a significant amount of value. LaValle, Lesser, Shockley, Hopkins, and Kruschwitz (2013) made the case that analytic sophistication in the use of big data should reduce the time it takes to provide value to end users. The fundamental problem in our field is that, as in the pre-big data era of educational research, it takes a significant amount of time for the research results to change teaching and learning practice. Shortening the time between data collection and its impact on learners is a challenge for big data in learning.

The educational data mining and the learning analytics communities are acutely aware of the need to move to more real-time approaches to dealing with data. While the educational data mining community has focused on relatively pure computational methods for processing educational data, the learning analytics community has emphasized approaches that focus on "human in the loop" methods for analyzing data. Baker and Inventado (2014) and Siemens and Baker (2012) provided a full review of the differences in philosophies and approaches between the educational data mining and learning analytics communities. Significant work has already been undertaken on intelligent tutors, cognitive tutors, and small-scale, domain-specific adaptive systems that do provide real-time feedback to students – but on relatively constrained problem domains. Baker and Yacef (2009) summarized these advances. The research methodologies they describe are means by which big data, machine learning approaches, and traditional educational theory building converge. These are novel and innovative methods. Yet, some significant research challenges are just beginning to emerge.

The transition from analyzing data that are relatively static to providing real-time feedback to learners in a large range of contexts is a significant problem. Researchers are increasingly adept at developing performance and feedback models for utilizing data that emerge from very specific and narrowly defined contexts. These performance and feedback models deal with everything from characterizing performance in courses to analyzing the emotional state of learners (sentiment analyses). Some of these models are predictive and utilize machine-learning techniques to customize learning pathways. However, if we were to think of a student traveling through the academic system with the goal of completing a course of study, the analytic contexts need to span decision making across multiple levels. For example, decision-making systems need to provide students with the ability to make informed decisions at a course or content level (micro-level decisions). At the same time, these systems need to include capabilities to help learners navigate institutional requirements (macro-level decisions). Work on even understanding the types of models that could be used for predictive efforts in spanning analytic boundaries between micro- and macro-level decisions is still very much in its infancy. Once these boundary-spanning models are

better developed, the next stage is to optimize them for the high-speed execution required in real-time analytics.

In order to provide real-time and actionable feedback to learners in the wide range of decision-making scenarios they encounter, we need to discuss streaming data in a bit more depth. The idea of streaming data was inherently present in the definition of "clickstream" we introduced earlier. The use of streaming data (Gaber, Zaslavsky, & Krishnaswamy, 2005; O'Callaghan, Meyerson, Motwani, Mishra, & Guha, 2002) for decision making is by no means new. Imagine an observer standing on a bridge watching a river of data flow by. Once the data have passed the observation point, it loses its immediacy for decision making and becomes part of a static repository. Such data are called one-pass data streams (Garofalakis, Gehrke, & Rastogi, 2002); the concept highlights the brief period of immediate efficacy of these data streams (in the context of learning for learners).

The thinking behind switching from analyzing static data to streaming data is to make learning data immediately useful to the learner who provided the data in the first place. In the above example, if we placed algorithms that execute at extremely high speed and precision at the point where the data crosses the observer on the bridge, and connect these algorithms to robust context-sensitive prediction models, it is possible to provide insights that allow better decision making on the part of the students. This line of research requires us to develop supervised (human in the loop) and unsupervised machine-learning techniques and embed these approaches into easy-to-use visual interfaces. For example, the ability to move to more real-time analytics would fundamentally change how we utilize formative assessments to probe what is "in the black box" (Black & Wiliam, 1998) of the learner's mind. It will also lead to providing better feedback to improve learning with reflective assessments where, for example, clickstream data reveal how students evaluate their own learning and the conditions in which learning occurs. Self-assessment or metacognitive assessment is recursive and can reveal how the learners think about their own thinking (Shute & Becker, 2010; Toth, Suthers, & Lesgold, 2002; White & Frederiksen, 2000). By enabling better reflective assessments, real-time analytics can be especially helpful to a learner developing cognitive routines for self-monitoring and error-detection strategies (Butler & Winne, 1995).

In general, most of the algorithmic approaches in learning analytics today focus on fairly standard web-mining techniques that have been well researched. Romero and Ventura (2007) pointed to some standard techniques, such as clustering, classification, association rule mining, sequential pattern mining (primarily linear sequences), and text mining. Romero and Ventura (2010) also pointed to advancements in the sophistication of the techniques through the use of Bayesian approaches. The noise and sparseness of learning data make building predictive models extremely difficult. However, a new class of algorithms based on variations of techniques known as Kalman filtering (Kalman, 1960) and ensemble Kalman filtering (Evensen, 2003) are starting to emerge to deal with data noise and sparseness. Ensemble techniques use a combination of algorithms in specific sequence to maximize accuracy and convergence capability of a model.

It is not just the algorithmic aspect of big data in learning research that is underdeveloped. Exploring, developing, and testing visual interfaces that shield non-experts in data mining from the complexity of the underlying data models and algorithms are all significant challenges. In discussing visual analytic systems aimed at non-experts in data mining, Madhavan et al. (2014) pointed out that "affordance is innovation" (p. 1831): it is less important for such data analytic systems to showcase their technical complexity than to place the users' decision-making needs at the center of the design process.

Conclusions

Engineering disciplines have been dealing with big data for a long time now. However, in engineering education, little has been written about the basic tenets of big data and their relevance to our field. The latency between collecting data from students and the ability to translate these data into actionable intelligence is a significant barrier. The relationships between research in learning, educational psychology, and theories of instructional design are complex, more like an interacting ecology of ideas and practices than a clear hierarchical organization. We are aware that in this editorial, we have not addressed a number of other critically important problems related to big data in learning – such as ethics of data usage or sharing research data. However, we decided to focus on one crucial research barrier that has profound influence on achieving the dream of real-time feedback to learners by collecting all the data they provide.

References

Aström, K. J., & Murray, R. M. (2009). *Feedback systems: An introduction for scientists and engineers*. Version v2.100b. http://www.cds.caltech.edu/~murray/books/AM08/pdf/am08-complete_22Feb09.pdf

Baker, R. S. J. d., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 61–75). Berlin, Germany: Springer.

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.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7–74.

boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679. http://dx.doi.org/10.1080/1369118X.2012.678878

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*(3), 245–281.

Carr, N. G. (2000). Hypermediation: Commerce as clickstream. *Harvard Business Review*, *78*(1), 46–47.

Cuzzocrea, A., Song, I.-Y., & Davis, K. C. (2011). *Analytics over large-scale multidimensional data: The big data revolution!* Paper presented at the ACM 14th International Workshop on Data Warehousing and OLAP, Glasgow, Scotland, UK.

Dumbill, E. (2012, January 11). What is big data? An introduction to the big data landscape. *O'Reilly*. https://beta.oreilly.com/ideas/what-is-big-data

Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, *53*(4), 343–367. http://dx.doi.org/10.1007/s10236-003-0036-9

Gaber, M. M., Zaslavsky, A., & Krishnaswamy, S. (2005). Mining data streams: A review. *ACM Sigmod Record*, *34*(2), 18–26.

Garofalakis, M., Gehrke, J., & Rastogi, R. (2002). *Querying and mining data streams: You only get one look. A tutorial*. Paper presented at the SIGMOD Conference, Madison, WI.

Gibson, W. (1984). *Neuromancer*. New York, NY: Penguin Patnam.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, *82*(1), 35–45.

Koestler, A. (1967). *The ghost in the machine*. London, UK: Hutchinson & Co.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2013). Big data, analytics and the path from insights to value. *MIT Sloan Management Review*, *21*.

Lohr, S. (2012, February 11). The age of big data. *New York Times.* http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?_r=1&scp=1&sq=Big Data&st=cse

Madhavan, K., Elmqvist, N., Vorvoreanu, M., Chen, X., Wong, Y., Xian, H., . . . Johri, A. (2014). DIA2: Web-based cyberinfrastructure for visual analysis of funding portfolios. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1823–1832. http://dx.doi.org/10.1109/TVCG.2014.2346747

Manouselis, N., Drachsler, H., Verbert, K., & Duval, E. (2012). *Recommender systems for learning*. New York, NY: Springer Science & Business Media.

Moe, W. W., & Fader, P. S. (2004). Capturing evolving visit behavior in clickstream data. *Journal of Interactive Marketing*, *18*(1), 5–19. http://dx.doi.org/10.1002/dir.10074

National Academy of Engineering. (2008). *Grand challenges for engineering in the 21st century*. http://www.engineeringchallenges.org/

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

O'Callaghan, L., Meyerson, A., Motwani, R., Mishra, N., & Guha, S. (2002). *Streaming-data algorithms for high-quality clustering*. Paper presented at the 2013 IEEE 29th International Conference on Data Engineering (ICDE), San Jose, CA. http://dx.doi.org/10.1109/ICDE.2002.994785

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

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics Part C: Applications & Reviews*, *40*(6), 601–618. http://dx.doi.org/10.1109/TSMCC.2010.2053532

Shute, V. J., & Becker, B. J. (2010). Prelude: Assessment for the 21st century. In V. J. Shute & B. J. Becker (Eds.), *Innovative assessment for the 21st Century* (pp. 1–11). Heidelberg, Germany: Springer.

Siemens, G., & Baker, R. S. J. d. (2012). *Learning analytics and educational data mining: Towards communication and collaboration*. Paper presented at the 2nd International Conference on Learning Analytics and Knowledge, Vancouver, BC. http://dx.doi.org/10.1145/2330601.2330661

Thorndike, E. L. (1913). *Educational psychology: Vol. 2. The psychology of learning*. New York, NY: Teachers College, Columbia University.

Toth, E. E., Suthers, D. D., & Lesgold, A. M. (2002). "Mapping to know": The effects of representational guidance and reflective assessment on scientific inquiry. *Science Education*, *86*(2), 264–286.

Tustin, A. (1952). Feedback. *Scientific American*, *187*(3), 48–54.

Verbert, K., Manouselis, N., Ochoa, X., Wolpers, M., Drachsler, H., Bosnic, I., & Duval, E. (2012). Context-aware recommender systems for learning: A survey and future challenges. *IEEE Transactions on Learning Technologies*, *5*(4), 318–335. http://dx.doi.org/10.1109/TLT.2012.11

Vygotsky, L. S. (1978). *Mind in society: The development of higher mental process*. Cambridge, MA: Harvard University Press.

White, B., & Frederiksen, J. (2000). Metacognitive facilitation: An approach to making scientific inquiry accessible to all. In J. Minstrell and E. van Zee (Eds.), *Inquiring into inquiry learning and teaching in science*. (pp. 331–370). Washington, DC: American Association for the Advancement of Science.

Wiener, N. (1961). *Cybernetics: Or, control and communication in the animal and the machine* (2nd ed.). New York, NY: MIT Press.

Authors

Krishna Madhavan is an associate professor of engineering education in the School of Engineering Education at Purdue University. He focuses on the research, design, and development of big data tools for a wide range of academic and nonacademic contexts.

Michael C. Richey, Ph.D., is an associate technical fellow involved in technology and innovation research at The Boeing Company. He is responsible for improving the learning experience for students, incumbent engineers, and technicians. His research interests are sociotechnical systems, learning curves, and engineering education.