

Learning Analytics: The Emergence of a Discipline

American Behavioral Scientist
57(10) 1380–1400

© 2013 SAGE Publications
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0002764213498851
abs.sagepub.com



George Siemens¹

Abstract

Recently, learning analytics (LA) has drawn the attention of academics, researchers, and administrators. This interest is motivated by the need to better understand teaching, learning, “intelligent content,” and personalization and adaptation. While still in the early stages of research and implementation, several organizations (Society for Learning Analytics Research and the International Educational Data Mining Society) have formed to foster a research community around the role of data analytics in education. This article considers the research fields that have contributed technologies and methodologies to the development of learning analytics, analytics models, the importance of increasing analytics capabilities in organizations, and models for deploying analytics in educational settings. The challenges facing LA as a field are also reviewed, particularly regarding the need to increase the scope of data capture so that the complexity of the learning process can be more accurately reflected in analysis. Privacy and data ownership will become increasingly important for all participants in analytics projects. The current legal system is immature in relation to privacy and ethics concerns in analytics. The article concludes by arguing that LA has sufficiently developed, through conferences, journals, summer institutes, and research labs, to be considered an emerging research field.

Keywords

learning analytics, models, distributed learning

The slightest move in the virtual landscape has to be paid for in lines of code.

—Bruno Latour

¹Athabasca University, Alberta, Canada

Corresponding Author:

George Siemens, Technology Enhanced Knowledge Research Institute, Athabasca University, 121 Pointe Marsan, Beaumont, Alberta, Canada T4X 0A2.

Email: gsiemens@athabascau.ca

When P. W. Anderson stated in 1972 that “more is different,” he argued that the quantity of an entity influences how researchers engage with it. As the quantity of data has increased, the attention of researchers, academics, and businesses has turned to new methods to understand and make sense of that data. In some sectors, the relatively recent emergence of big data and analytics is now viewed as having the potential to transform economies and increase organizational productivity (Manyika et al., 2011, p. 13) and increase competitiveness (Kiron, Shockley, Kruschwitz, Finch, Haydock, 2011). Unfortunately, education systems—primary, secondary, and postsecondary—have made limited use of the available data to improve teaching, learning, and learner success. Despite the field of education lagging behind other sectors, there has been a recent explosion of interest in analytics as a solution for many current challenges, such as retention and learner support.

Science is concerned with discovering or recognizing the nature of the universe, particularly in terms of how entities are connected or related to each other. New discoveries are added to “the network of theory” (Kuhn, 1996, p. 7), and when discoveries and innovations “align and come together” (Morin, 2008, p. 51), scientific paradigms, models, and programs result. This new knowledge leads to revisions and questions about our prior understanding and beliefs and reflection on the connections within “the network of theory.” For example, connections that have been now demonstrated to be false (earth-centered universe, personality caused by the four humors) have been replaced by new connections between entities that can be validated. Over time, these connections may be further modified and revised and linked into the new nodes and areas of knowledge. The improvements in the process of discoveries can be considered as “more important than any single discovery” (Nielsen, 2012, p. 3).

Analytics is another approach, or cognitive aid, that can be applied to assist scientists, researchers, and academics to make sense of the connective structures that underpin their field of knowledge. The methods of science and the questions investigated have rapidly changed as large data sets have become available. Early attempts to manage knowledge through classification systems (such as early attempts by Yahoo to organize the web into categories) have now been replaced by the big data and algorithmically driven approach of Google. The emphasis on large quantities of data for discovery has important implications for education. Through the use of mobile devices, learning management systems (LMS), and social media, a greater portion of the learning process generates digital trails. A student who logs into an LMS leaves thousands of data points, including navigation patterns, pauses, reading habits, and writing habits. These data points may be ambiguous and require additional exploration in order to understand what an extended pause of reading means (perhaps the student is distracted or engaged in other tasks, or perhaps the student is grappling with a challenging concept in the text), but for researchers, learning sciences, and education in general, data trails offer an opportunity to explore learning from new and multiple angles. As stated by Latour (2008), the “slightest move in the virtual landscape [is] paid for in lines of code.”

The view that data and analytics offer a new mode of thinking and a new model of discovery is at least partially rooted in the artificial intelligence and machine learning fields. Halevy, Norvig, and Pereira (2009) argue for the “unreasonable effectiveness of

data” (p. 8), stating that machine learning and analytics can help computers to tackle even the most challenging knowledge tasks, such as understanding human language. Hey, Tansley, and Tolle (2009) are more bold in their assertions, arguing that data analytics represent the emergence of a new approach to science.

This article reviews the historical developments of learning analytics as a field, tools and techniques used by practitioners and researchers, and challenges with broadening the scope of data capture, modeling knowledge domains, and building organizational capacity to use analytics.

Defining Learning Analytics and Tracing Historical Roots

As the field of learning analytics (LA) is further refined and established, an authoritative definition will emerge. At present, the vast majority of LA literature has begun to adopt the following definition offered in the 1st International Conference on Learning Analytics:

Learning analytics is the measurement, collection, analysis, and reporting of data about learners and their contexts, for the purposes of understanding and optimizing learning and the environments in which it occurs.¹

Other definitions are less involved and draw language from business intelligence:

Analytics is the process of developing actionable insights through problem definition and the application of statistical models and analysis against existing and/or simulated future data. (Cooper, 2012b)

Where LA is more concerned with sensemaking and action, educational data mining (EDM) is more focused toward developing methods for “exploring the unique types of data that come from educational settings”.² Although the techniques used are similar in both fields, EDM has a more specific focus on reductionist analysis (Siemens & Baker, 2012). As LA draws from and extends EDM methodologies (Bienkowski, Feng, & Means, 2012, p. 14), it is a reasonable expectation that the future development of analytic techniques and tools from both communities will overlap.

Analytics in education can also be viewed as existing in various levels, ranging from individual classroom, department, university, region, state/province, and international. Buckingham Shum (2012) groups these organizational levels as micro-, meso-, and macroanalytics layers. Each level affords access to a differing set of data (quantity and diversity) and contexts. As such, different questions and analytic lenses can be applied that provide detailed and nuanced insight into the specific organizational layer. For example, classroom analytics might include social network analysis and natural language processing concerned with assessing individual engagement levels, whereas department-level analytics might be more concerned with risk detection and intervention and support services, and institution-level analytics might be concerned with improving operating efficiency of the university or comparing performance with other peer universities. In essence, as the organizational scale changes, so too do the tools

and techniques used for analyzing learner data alongside the types of organizational challenges that analytics can potentially address.

Historical Contributions to LA

LA, as a field, has multiple disciplinary roots. While the fields of artificial intelligence (AI), statistical analysis, machine learning, and business intelligence offer an additional narrative, the focus here is on the historical roots of analytics in relation to human interaction and the education system. AI is foundational in analytics and will become more so as Bayesian models and neural networks increase in prominence in LA. However, it is likely in the near future that the use of AI and machine learning in education will first involve an extended period of experimentation rather than critical or core adoption (Cooper, 2012a, p. 9).

The following is a brief summary of the diversity of fields and research activities within education that have contributed to the development of learning analytics:

- *Citation analysis*: Garfield (1955) was an early pioneer in analytics in science by emphasizing how developments in science can be better understood by tracking the associations (citations) between articles. Through tracking citations, scientists can observe how research is disseminated and validated. PageRank, a key algorithm in Google's early search engine, adopted Garfield's model of analyzing and weighting links on the web in order to gain "an approximation to 'importance'" of particular resources (Page, Brin, Motwani, & Winograd, 1999). Educationally, citation or link analysis is important for mapping knowledge domains (detailed below in Knowledge Domain Modeling).
- *Social network analysis* is prominent in sociology, dating back to the work by Granovetter (1973) and Milgram (1967). Wellman (1999), active in social network research since the early 1970s, transitioned into analysis of networks in digital settings. Haythornthwaite (2002) has more recently explored the impact of media type on the development of social ties.
- *User modeling* is concerned with modeling users in their interaction with computing systems. User modeling contributed to a shift in computing where users were treated "as individuals with distinct personalities, goals, and so forth" (Rich, 1979, p. 329), rather than treating all users the same. User modeling has become important in research in human-computer interactions as it helps researchers to design better systems (Fischer, 2001, p. 70) by understanding how users interact with software. As detailed later, recognizing unique traits, goals, and motivations of individuals remains an important activity in learning analytics.
- *Education/cognitive modeling* has been applied to tracing how learners develop knowledge. Cognitive models have historically attempted to develop systems that possess a "computational model capable of solving the problems that are given to students in the ways students are expected to solve the problems" (J. R. Anderson, Corbett, Koedinger, & Pelletier, 1995, p. 168). Cognitive modeling has contributed to the rise in popularity of intelligent or cognitive tutors. Once

cognitive processes can be modeled, software (tutors) can be developed to support learners in the learning process.

- *Tutors*: Computers have been used in education for decades as learning tools. In 1989, Burns argued for the adoption and development of intelligent tutor systems that ultimately would pass three levels of “intelligence”: domain knowledge, learner knowledge evaluation, and pedagogical intervention. These three levels continue to be relevant for researchers and educators.
- *Knowledge discovery in databases* (KDD) has been a research interest since at least the early 1990s. As with analytics today, KDD was “concerned with the development of methods and techniques for making sense of data” (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, p. 37). The EDM community has been heavily influenced by the vision of early KDD.
- *Adaptive hypermedia* builds on user modeling by increasing personalization of content and interaction. “Adaptive hypermedia systems build a model of the goals, preferences and knowledge of each user, in order to adapt to the needs of that user” (Brusilovsky, 2001, p. 87). As will be presented later in this article, personalization and adaptation of learning content is an important future direction the learning sciences.
- *E-learning*: The growth of online learning, particularly in higher education (T. Anderson, 2008; Andrews & Haythornthwaite, 2007; Haythornthwaite & Andrews, 2011), has contributed to the advancement of LA as student data can be captured and made available for analysis. When learners use an LMS, social media, or similar online tools, their clicks, navigation patterns, time on task, social networks, information flow, and concept development through discussions can be tracked. The rapid development of massive open online courses offers additional data for researchers to evaluate teaching and learning in online environments (Chronicle of Higher Education, 2012).

By early 2000, analytics and data-driven approaches to decision making were gaining attention in the academy. While intelligent tutors, user modeling, and adaptive hypermedia emphasized research challenges in learning, *academic analytics* involved the adoption of business intelligence (BI) to the academic sector (Goldstein, 2005). While sometimes referred to as LA, the BI roots of academic analytics are more concerned with improving organizational processes, such as personnel management or resource allocation, and improving efficiency within the university. Academic analytics is also more concerned with organizational operation and “describes[s] the intersection of technology, information, management culture, and the application of information to manage the academic enterprise (Goldstein, 2005, p. 2).

LA Tools, Techniques, and Applications

The fields that have contributed to the development of LA as a discipline, reviewed in the previous section, have also contributed a range of technologies and techniques that are now being used by LA researchers and practitioners.

Tools

Learning analytics tools can be broadly grouped into two categories: commercial and research.

Commercial. Commercial tools are the most developed, with companies such as SAS and IBM investing heavily in adapting their analytics tools for the education market. The use of SPSS, Stata, and NVivo for LA and modeling is an extension of the research activities that students and academics have previously conducted with these tools. Statistical software packages are as central in analytics as they are in quantitative research.

A recent, and alternate, wave of commercial offerings has evolved from education market vendors, such as Ellucian and Desire2Learn. Student information systems, curriculum management software, and learning management systems are already widely used in the education sector. Adding analytics layers to existing systems provides a rapid way to add value for education administrators, managers, and teachers. Several prominent analytics tools already rely on data captured in an LMS. For example, Purdue University's Signals (Arnold, 2010) and University of Maryland–Baltimore County's "Check My Activity" (Fritz, 2010) both rely on data generated in Blackboard. Recommender systems, such as Degree Compass (Denley, 2012), similarly draw on data captured in existing information technology systems in universities. Web analytics tools, such as Google Analytics and Adobe's Digital Marketing Suite (formerly Omniture), are also used for LA.

As the above notes, the current focus on analytics in education has motivated existing commercial vendors to either modify or extend the range of features within established products. However, the growth in this field has also prompted the emergence of a new suite of commercial analytics tools and infrastructure, notably, Tableau Software and Infoclimps. These tools are designed specifically to remove the complexity surrounding many analytic tasks, such as data importing, cleaning, and visualization. These products reflect that analytics is no longer a discrete area of specialization but now attracts individuals with a vast array of skills, expertise, and backgrounds. For example, IBM's Many Eyes allows users to upload text and data files and perform basic analytics without the need for specialized programming or visualization skills. As ease of use, affordability, and accessibility of tools improve, there will be a corresponding increase in the level of adoption across the education community.

Research/open. Research and open analytics tools are not as developed as commercial offerings and typically do not target systems-level adoption. Tools such as R and Weka are focused on individual analytics tasks, not currently designed to be used as part of an integrated institutional support system for analytics.³ Similarly, tools such as SNAPP (Dawson, Bakharia, & Heathcote, 2010), a browser plug-in for social network analysis of discussion forum interactions, or Netlytic (Gruzd, 2010), a cloud-based text and social networks analyzer, are easily accessible for individual researchers but do not have systems-level integration and support.

Techniques and Applications

LA has two overlapping components: techniques and applications. Techniques involve the specific algorithms and models for conducting analytics. Applications involve the ways in which techniques are used to impact and improve teaching and learning. For example, an algorithm that provides recommendations of additional course content for learners can be classified as a technique. A technique, such as prediction of learner risk for dropout, can then lead to an application, such as personalization of learning content to reflect learners' comfort with the subject area. The distinction between technique and an application is not absolute but instead reflects the focus of researchers. A statistician may be more interested in creating probability models to identify student performance (technique), whereas a sociologist may be more interested in evaluating how social networks form based on technologies used in a course (application). Both, however, are important in advancing LA as a field.

Baker and Yacef (2009) address the technique dimension of LA/EDM in listing five primary areas of analysis:

- Prediction
- Clustering
- Relationship mining
- Distillation of data for human judgment
- Discovery with models

Bienkowski, Feng, and Means (2012) offer five areas of LA/EDM application:

- Modeling user knowledge, behavior, and experience
- Creating profiles of users
- Modeling knowledge domains
- Trend analysis
- Personalization and adaptation

Baker and Yacef's model details various types of data mining activity that the researcher conducts, whereas Bienkowski et al.'s model is focused on application. The distinctions between these two models are revealing as they indicate the difficulty of definitions and taxonomies of analytics. The lack of maturity about techniques and analytics models reflects the youth of LA as a discipline.

Techniques, especially those prominent in EDM, are technical and increasingly reflect machine learning and AI techniques. Through statistical analysis, neural networks, and so on, new data-based discoveries are made and insight is gained into learner behavior. This can be viewed as basic research where discovery occurs through models and algorithms. These discoveries then serve to lead into application (see also Herskovitz, Baker, Gobert, Wixon, & Pedro, 2013).

Application areas of LA involve user modeling, knowledge domain modeling, analysis of trends and patterns, and personalization and adaptation. Application areas

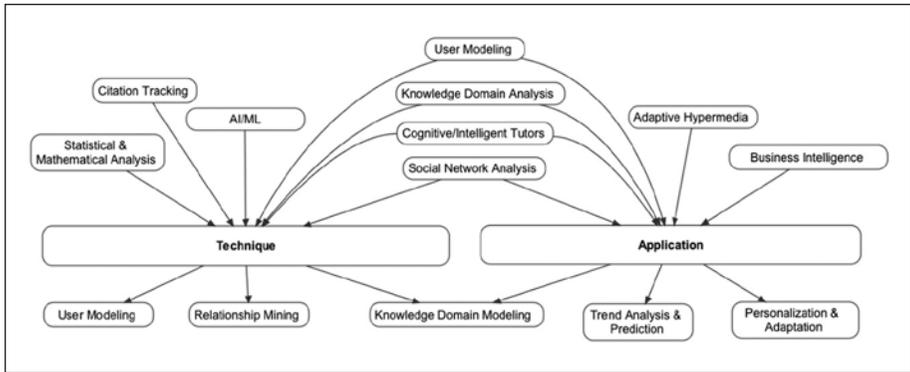


Figure 1. Historical influences in development of learning analytics.

influence the development of curriculum (such as ontologies that can be automatically evaluated against learner-produced work), social network analysis, and discourse analysis. The multidisciplinary roots of LA and the current techniques and applications are detailed in Figure 1.

Prominent analytics techniques are presented in Table 1. It is important to note that analytics models and approaches continue to borrow heavily from the traditional fields, as presented in Figure 1. More recent analytics models that target learning are being developed by LA researchers, such as those tracking behavior, persistence, achievement (Macfadyen & Dawson, 2010; Morris, Finnegan, & Wu, 2005), attention metadata (Wolpers, Jehad, Verbert, & Duval, 2007), participatory and peer learning (Clow & Makriyannis, 2011), and social LA (Buckingham Shum & Ferguson, 2012).

Scope of Data Capture

Analytics requires data sources that reflect the complexity of the learning process. The development of student models predicting success or identifying at-risk learners, intervention strategies, and adaptive learning requires an analytics system to generate learner models or profiles. Simply put, “quality” data are required. Ideally, data that are captured as learners are engaged in authentic learning (where collection is unobtrusive), as contrasted with contrived learning tasks, will provide researchers with greater insight into the social and pedagogical dimensions of learner performance. To date, LA has relied heavily on two sources: student information systems (SIS; in generating learner profiles) and learning management systems (in tracking learner behavior and using it for prediction).

The expansion of data beyond SIS and LMS into a broad range of sources, including the physical interactions that currently do not leave data trails, is important in increasing the quality and depth of analysis. One approach to increase data capture is through “sensor-based modeling of human communication networks” (Choudhury & Pentland, 2003). Sensor-based modeling involves wearable computing devices that

Table 1. Learning Analytics (LA) Techniques and Applications.

LA Approach	Examples
Techniques	
Modeling	Attention metadata Learner modeling Behavior modeling User profile development
Relationship mining	Discourse analysis Sentiment analysis A/B testing Neural networks
Knowledge domain modeling	Natural language processing Ontology development Assessment (matching user knowledge with knowledge domain)
Applications	
Trend analysis and prediction	Early warning, risk identification Measuring impact of interventions Changes in learner behavior, course discussions, identification of error propagation
Personalization/adaptive learning	Recommendations: Content and social connections Adaptive content provision to learners Attention metadata
Structural analysis	Social network analysis Latent semantic analysis Information flow analysis

capture social connections and conversations. Other approaches include “passive acquisition” of “physical activity data” through “pedometers, heart rate monitors, accelerometers, and distance trackers” (Lee & Thomas, 2011, p. 867). With the prominence of mobile devices and emergence of wearable computing, such as Google Glass,⁴ and the “quantified self” movement,⁵ the scope and quantity of data available for analytics will continue to increase.

The role of active data collection and human interaction with, and evaluation of, data before visualization and presentation are reflected in the process of making Google Maps. In addition to collecting images through Google Street View cars, data and images are evaluated and updated by people in order to create maps that are current and accurate: “The sheer amount of human effort that goes into Google’s maps is just mind-boggling” (Madrigal, 2012). Many current LA models rely on data automatically collected. However, these accessible data points, particularly in relation to the learning context, are often incomplete and mere static snapshots in time. To be effective, holistic, and transferable, future analytics projects must afford the capacity to include additional data through observation and human manipulation of the existing data sets.

Open online courses may provide additional data sets for researchers. With the development of new models of learning (Downes, 2005), the adoption of active learning models will influence the types of data available for analysis. Lecture hall data are limited to a few variables: who attended, seating patterns, student response system data, and observational data recorded by faculty or teaching assistants. By contrast, when learners watch a video lecture, data sources are richer, including frequency of access, playback, pauses, and so on. When videos are used as part of an interactive learning system, such as edX,⁶ additional data can be captured about student errors or returns to videos for review. Anant Agarwal has stated that the edX platform is a “particle accelerator for collecting data on learners and helping researchers to understand the learning process.”⁷

A single data source or analytics method is insufficient when considering learning as a holistic and social process. Multiple analytic approaches provide more information to educators and students than single data sources. In fields of network analysis, researchers are using multiple methods to evaluate activity within a network, including detailing different node types, direction of interaction, and human and computer nodes (Contractor, Monge, & Leonardi, 2011; Kim & Lee, 2012; Suthers & Rosen, 2011). These same techniques, drawing on multiple entities and sources of data and user interactions, must be adopted and evolved in LA to further advance the field.

Knowledge Domain Modeling

In addition to improving the scope of data capture and developing advanced analytics tools, personalizing the learning process for individual students is important for the future of LA. Knowledge domains have a structure that can be traced and visualized (Chen & Paul, 2001). Hendler and Berners Lee (2010) state that “the problems that our society faces today are such that only the concerted effort of groups of people, operating with a joint power much greater than that of the individual can hope to provide solutions” (p. 157). In extending this call for community-centric and multidisciplinary approaches to tackle complex problems, the authors argue for the need for “data structures and computational techniques” (Hendler & Berners Lee, 2010, p. 158) to enable human-computer interactions that provide a new level of intelligence and problem solving. A subject area can be mapped and defined. Google’s Knowledge Graph is an example of articulating and tracing the connectedness of knowledge.⁸ Similarly, Börner, Chen, and Boyack’s (2003) work in visualizing knowledge domains details connectedness in the sciences (see also Börner, 2011).

Once knowledge domains have been articulated or mapped, learner data, profile information, and curricular data can be brought together and analyzed to determine learner knowledge in relation to the knowledge structure of a discipline. Data trails and profiles, in relation to curriculum in a course, can be analyzed and used as a basis for prediction, intervention, personalization, and adaptation. Adaptation is not exclusively technological—sensemaking and wayfinding through social systems have demonstrated their value over the last several years through recommender systems, small network clusters, and so on. Adaptation and personalization are multifaceted and

consist of more than just recommending content, incorporating technology, socialization, and pedagogy.

Curriculum in schools and higher education is generally preplanned. Designers create course content, interaction, and support resources well before any learner arrives in a course (online or on campus). Through the use of analytics, educational institutions can restructure learning design processes. When learning designers have access to information about learner success following a tutorial or the impact of explanatory text on student performance during assessment, they can incorporate that feedback into future design of learning content.

Learning content provided to learners can be personalized—a real-time rendering of learning resources and social suggestions based on the profile of a learner, including conceptual understanding of a subject and previous experience. For example, an integrated learning system could track a learner's physical and online interactions, analyze skills and competencies, and then compare learner knowledge with the mapping of knowledge in a discipline. Based on evaluation of a learner's knowledge, an LMS or learning system could provide personalized content and learning activities.

Organizational Capacity

Organizations face a concern about capacity in initiating analytics projects. Individuals who demonstrate the full range of skills and attributes needed to make sense of numerous data sets are rare, resulting in numerous predictions of significant skills shortages (Manyika et al., 2011). For example, an analytics project will require accessing, cleaning, integrating, analyzing, and visualizing data—before any attempts at sensemaking. As such, analytics scientists require programming skills, statistical knowledge, and familiarity with the data and the domain represented in that data in order to be able to ask relevant questions. They will need to be familiar with a variety of data tools and analytics models. They will also need access to server logs and databases. It is unlikely, especially given the rapid development of big data and analytics, that one individual will have a complete set of these skills.

Additionally, effective analytics practices require organizational support. If analytics is to have an impact on how a university supports its learners, great inter- and intra-institutional collaborations are required. Consider the example of analytics that supports identifying learners who are at risk of dropping a course or a program. Data sources might include an SIS, an LMS, and a student success system (i.e., tracking points of contact that the university has with a student, similar to a customer relationship management system from business). Once data access and sharing across departments and courses have been confirmed and a model of weighting important variables has been developed, both automated and human support systems are necessary to intervene and assist learners in a manner that is both timely and targeted in addressing their specific support requirements. Intervention and organizational support require cross-departmental collaboration so that the appropriate help for learners can be identified.

Macfadyen and Dawson (2012) argue that even if analytics is seen to have value and is championed by senior administration, the success of an analytics initiative depends

on faculty support and navigating “the realities of university culture” (p. 160). The insights gained through analytics require broad support organizationally. Prior to launching a project, organizations will benefit from taking stock of their capacity for analytics and willingness to have analytics have an impact on existing processes. In this context, Greller and Drachsler (2012, p. 43) outline six dimensions that must be considered “to ensure appropriate exploitation of LA in an educationally beneficial way”:

- Stakeholders: Those who are interested in or impacted by analytics
- Objectives: Goal or intent of analytics
- Data: Data sets and sources
- Instruments: Tools and technologies
- External limitations: Ethical, legal, managerial/organizational
- Internal limitations: Acceptance of analytics and skill level or competencies to perform analytics within an organization

A restrictive element to the analytics process is the need for multiple areas of expertise in analytics projects. As an illustration, even simple analytics activities will generally require access to server logs or databases. Once these data have been accessed, they need to be cleaned and rendered into a suitable format for analysis. If complex statistical analysis is required, many educators will need additional help with statistical analysis. On even the most basic analytics projects, multiple departments and skill sets are required. Some universities and schools have overcome these data challenges by developing integrated systems for analytics that hide the complex technical processes from end users.

The effective process and operation of learning analytics require institutional change that does not just address the technical challenges linked to data mining, data models, server load, and computation but also addresses the social complexities of application, sensemaking, privacy, and ethics alongside the development of a shared organizational culture framed in analytics.

LA Model

The use of data for improving learning is common in universities. Much of this activity currently happens at a small scale in individual classrooms, where educators use data collected manually or through analysis of server logs to provide individual educators with feedback on which exam questions cause learner confusion or which learning activities or lectures need greater clarity as measured by learner performance on exams or tests. This type of “bottom-up approach” to data use, while helpful for the faculty member and students, fails to take advantage of systems approaches to analytics. The LA model (LAM) detailed below introduces systemwide approaches to analytics. A systemic approach ensures that support resources are systematized, rather than relying solely on faculty time and/or observation. Interventions, such as providing students with support resource recommendations or creating predictive models of learner success, are not possible without top-down support in a university.

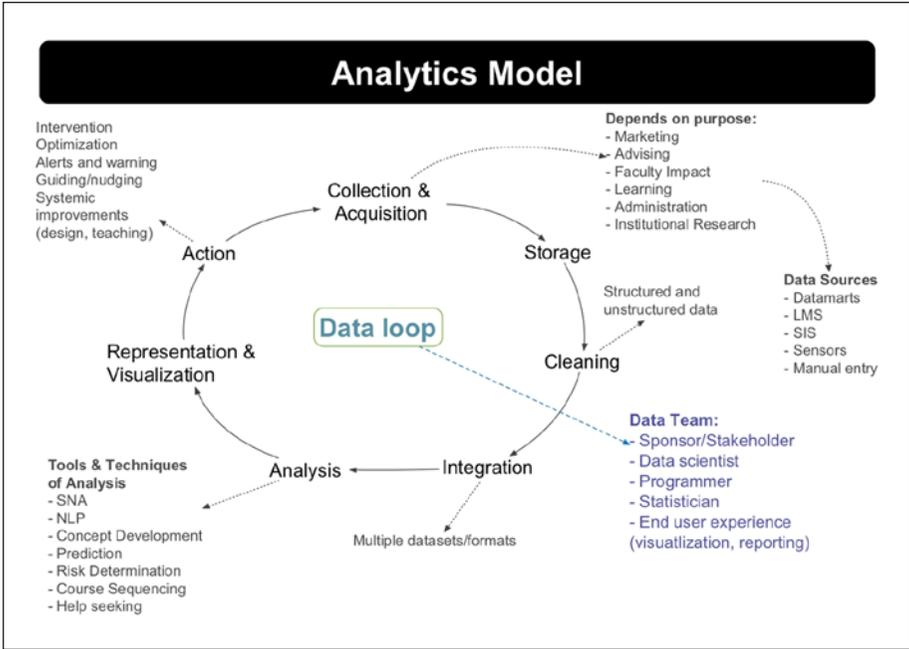


Figure 2. Learning analytics model.

LAM includes seven components: collection, storage, data cleaning, integration, analysis, representation and visualization, and action. These components are detailed in Figure 2. The importance for a data team is also highlighted. A systemic approach to analytics requires a combination of skills and knowledge that are likely not in the possession of a single individual.

Challenges

The most significant challenges facing analytics in education are not technical. Concerns about data quality, sufficient scope of the data captured to reflect accurately the learning experience, privacy, and ethics of analytics are among the most significant concerns (see Slade & Prinsloo, 2013). These challenges will become more prominent in LA as the field advances and analytics begins to form a greater part of educational research and how universities and schools track, monitor, and advise learners.

Data Quality and Scope

An important challenge for researchers involves increasing the scope of data capture through alternative collection models, such as wearable computing and mobile devices.

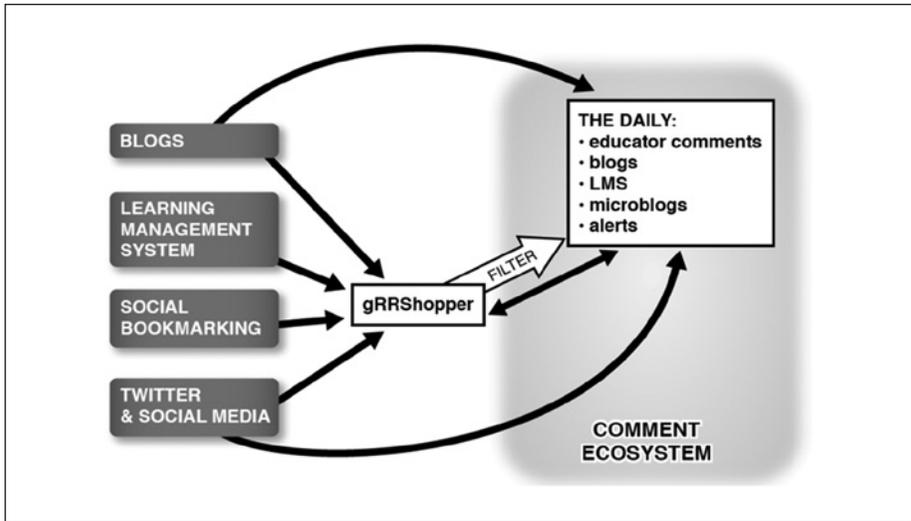


Figure 3. Distributed analytics approaches.

Data interoperability “imposes a challenge to data mining and analytics that rely on diverse and distributed data” (Bienkowski et al., 2012, p. 38) and needs to be addressed early in an analytics project to ensure that technical challenges surface early. As Verbert, Manouselis, Drachsler, and Duval (2012) state, “although an enormous amount of data has been captured from learning environments, it is a difficult process to make this data available for research purposes” (p. 145). Privacy concerns, diversity of data sets and sources, and lack of standard representation make sharing available data difficult.

Distributed and fragmented data present a significant challenge for analytics researchers. The data trails that learners generate are captured in different systems and databases. The experiences of learners interacting with content, each other, and software systems are not available as a coherent whole for analysis. Suthers and Rosen (2011) capture the challenge when stating “since interaction is distributed across space, time, and media, and the data comes in a variety of formats, there is no single transcript to inspect and share, and the available data representations may not make interaction and its consequences apparent” (p. 65).

Assessing interaction in distributed systems raises additional concerns for researchers as different identities in different software services make it difficult to determine how various identities map to a particular individual. Approaches to analytics in distributed systems require either building an infrastructure that aggregates data from multiple sources, such as gRSShopper, or developing a series of “recipes” for capturing and evaluating distributed data (Hawksey, 2012; Hirst, 2013). Figure 3 details the gRSShopper system, where multiple data sources, including blogs, LMS, and social media (essentially any software that offers an RSS feed), are aggregated and then

filtered based on course tag. Posts that do not include the course tag are not distributed to learners. Those that include the course tag are sent to learners in the form of a daily e-mail newsletter or as a web page.

Privacy

Privacy and data ownership concerns are not unique to analytics; any type of online or digital interaction produces a data trail, and ownership of that trail has not been decided either culturally (i.e., through norms and socially acceptable approaches to data use and analytics) or legally. Access to personal data “is generating a new wave of opportunity for economic and societal value creation” (World Economic Forum, 2011, p. 5). In higher education, this economic value can come from improved teaching and learning, reduced student attrition, and improved quality of support services. With interactions online reflecting a borderless and global world for information flow, any approach to data exchange and data privacy requires a global view (World Economic Forum, 2011, p. 33).

Further challenges around the use of analytics in education are reflected in the broader privacy and ethical concerns stemming from the rapid development of online technologies. In many areas, including copyright and intellectual property (IP) law, new opportunities with technology have not been fully addressed by the legal system. This shows a low level of “legal ‘maturity’” (Kay, Korn, & Oppenheim, 2012, p. 8), where legal systems have not yet advanced to address privacy, copyright, IP, and data ownership in digital environments. Privacy laws differ from nation to nation, and additional questions arise when, for example, a student from India takes an online course with a provider in the United States. In the near future, privacy rules and laws may require a harmonization similar to what has occurred for copyright and IP laws in many developed countries over the past several decades.

The importance of data ownership and learner control is reflected in the development of tools and initiatives, such as MyData Button, that “enable students to download their own data to create a personal learning profile that they can keep with them throughout their learning career.”⁹ However, ownership of, and access to, data is only one aspect that educators need to consider. The analysis of data presents a secondary concern. Who has access to analytics? Should a student be able to see what an institution sees? Given variations in privacy laws, should educators be able to see the analytics performed on students in different courses? On graduation, should analytics be made available to prospective employees? When a learner transfers to a different program or a different university, what happens to his or her data? How long does a university keep those data, and can they be shared with other universities? These and numerous equally intractable problems will need to be addressed.

One approach to consider in the privacy and ethics of analytics is to treat data as a transactional entity (such as money). It is conceivable that in the future, students will be encouraged to provide their data to the university in exchange for personalized support services. For some students, the sharing of personal data with an institution in exchange for better support and personalized learning will be seen as a fair value exchange.

The Dark Side

The potential of LA to provide educators with actionable insight into teaching and learning is clear. The implications of heavy reliance on analytics are less clear. Ellul (1964) stated that technique and technical processes strive for the “mechanization of everything it encounters” (p. 12). Ellul’s comments remind us of the need to keep human and social processes central in LA activities. The learning process is essentially social and cannot be completely reduced to algorithms. The difficulties of automating social systems have not prevented universities from making the attempt: “Most of our institutions of higher learning are as thoroughly automated as a modern steel plant” (Mumford, 1964, p. 274). The learning process is creative, requiring the generation of new ideas, approaches, and concepts. Analytics, in contrast, is about identifying and revealing what already exists. Self-organizing systems and software may in the future be capable of innovation in modeling learner creativity, but currently even agent-based simulations are rudimentary. The tension between innovation (generating something new) and analytics (evaluating what exists in data) is one that will continue to exist in the foreseeable future.

A Personal Reflection

In 2010, a small group of researchers and academics became involved in organizing the first LA conference, in Banff, Alberta, Canada.¹⁰ The conference was small, with 100 attendees. Initial planning for the event emphasized the interdisciplinary nature of analytics. Conference organizers explicitly sought out presentations that addressed both the technical/algorithmic as well as the social/pedagogical aspect of analytics, recognizing that LA requires considerations of the social aspects and activities that may not yet be quantifiable. The home disciplines of LA participants ranged widely, including education, statistics, computer science, information science, sociology, and computer-supported collaborative learning. Since then, additional fields, such as machine learning, AI, organizational theory, learning sciences, scientometrics, and psychology, have been represented. Neuroscience and neurocognition have not yet been represented but are important fields for future connections.

Looking forward, it is likely that analytics will continue to grow as a field. From the vantage of 2010, questions existed around whether LA would develop as an academic and research field or whether LA would be incorporated into existing fields. Since then, EDM has developed a formal organizational structure (International Educational Data Mining Society), and the Society for Learning Analytics Research (SoLAR) has also been established. Both societies host annual conferences and publish open-access journals. SoLAR has initiated outreach through doctoral seminars, distributed research lab, regional events, and data challenges. Continued growth in conference attendance and publication, as well as special issues and workshops in existing communities (HICSS and IEEE), indicates that interest is growing. EDUCAUSE has been one of the earliest and most active sources of LA research and dissemination of LA case implementation.

Interest in analytics is not confined to researchers. Corporate interest is high in analytics and learning. LMS providers are offering analytics in their software, and companies such as Pearson and McGraw-Hill are investing in or acquiring adaptive learning software (e.g., Knewton and Area9, respectively). The primary and secondary education markets are also being served by growing numbers of analytics providers.

In addition to addressing the challenges listed above, the future success of LA and EDM as research domains requires the development of academic programs to foster and develop new researchers as well as development of grant programs that target LA. Barry Wellman, during his keynote to the 2nd International Conference on Learning Analytics and Knowledge, drew parallels between his early work in social network analysis (1970s) and the energy currently evident in the LA field. For researchers involved in LA, this is an exciting reflection and offers a vision for LA to develop to a similar level of influence as social network analysis has in the academy and society today.

Conclusion

As a field, LA is still developing. Questions remain about how the field will emerge: Will it remain a distinct field of research, or will analytics practices be subsumed into other related fields? Analytics is already an existing core activity of researchers. The growth of available data, due to the prominence of online learning and digital technologies in education, forces educators to confront P. W. Anderson's (1972) observation: *More is different*. Managing large quantities of learner-generated data and gaining insight into the learning process through LA raise the profile of new tools and new techniques. With LA as a field now with its third annual conference, a journal, doctoral research lab, local and regional events, summer institutes, and special issues with established journals, current indications suggest that analytics will indeed establish its own identity as a distinct field.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. See <https://tekri.athabascau.ca/analytics/>.
2. See <https://www.educationaldatamining.org/>.
3. Commercial providers, such as Revolution Analytics (<http://www.revolutionanalytics.com/>), have started to develop enterprise-level R tools. Within education, however, open-source enterprise-level tools do not currently exist.

4. See <https://plus.google.com/+projectglass/about>.
5. See <https://quantifiedself.com/>.
6. See <https://www.edx.org/>.
7. Personal conversation with Anant Agarwal, January 31, 2013.
8. See <https://www.google.ca/insidesearch/features/search/knowledge.html>.
9. See <https://www.ed.gov/edblogs/technology/mydata/>.
10. See <https://tekri.athabascau.ca/analytics/>.

References

- Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, 4(2), 167-207.
- Anderson, P. W. (1972). More is different. *Science*, 177(4047), 393-396.
- Anderson, T. (Ed). (2008). *The theory and practice of online learning* (2nd ed.). Athabasca, Canada: Athabasca University Press.
- Andrews, R., & Haythornthwaite, C. (Eds.). (2007). *Handbook of e-learning research*. London, UK: Sage.
- Arnold, K. (2010). Signals: Applying academic analytics. *EDUCAUSE Review*. Retrieved from <http://www.educause.edu/ero/article/signals-applying-academic-analytics>
- 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). Retrieved from http://www.educationaldatamining.org/JEDM/images/articles/vol1/issue1/JEDMvol1Issue1_BakerYacef.pdf
- Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics*. Washington, DC: U.S. Department of Education. Retrieved from <http://www.ed.gov/edblogs/technology/files/2012/03/edm-la-brief.pdf>
- Börner, K. (2011). Plug-and-play macroscopes. *Communications of the ACM*, 54(3), 60-69.
- Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing domains of knowledge. *Annual Review of Information Science and Technology*, 37(1), 179-255.
- Brusilovsky, P. (2001). Adaptive hypermedia: From intelligent tutoring systems to web-based education. *User Modeling and User-Adapted Interaction*, 11(1/2), 87-110.
- Buckingham Shum, S. (2012). *Learning analytics*. UNESCO policy brief. Retrieved from <http://iite.unesco.org/pics/publications/en/files/3214711.pdf>
- Buckingham Shum, S., & Ferguson, R. (2012). Social learning analytics. *Educational Technology and Science*, 15(3), 3-26.
- Burns, H. L. (1989). Foundations of intelligent tutoring systems: An introduction. In J. J. Richardson & M. C. Polson (Eds.), *Proceedings of the Air Force Forum for Intelligent Tutoring Systems*. Retrieved from <http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA207096#page=16>
- Chen, C., & Paul, R. (2001). Visualizing a knowledge domain's intellectual structure. *Computer*, 34(3), 65-71.
- Choudhury, T., & Pentland, A. (2003, October). *Sensing and modeling human networks using the sociometer*. Paper presented at the Seventh IEEE International Symposium on Wearable Computers (ISWC'03), White Plains, NY.
- Chronicle of Higher Education. (2012). What you need to know about MOOCs. *Chronicle of Higher Education*. Retrieved <http://chronicle.com/article/What-You-Need-to-Know-About/133475/>

- Clow, D., & Makriyannis, E. (2011). iSpot analysed: Participatory learning and reputation. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. Retrieved from <http://dl.acm.org/citation.cfm?id=2090121&CFID=82269174&CFTOKEN=35344405>
- Contractor, N. S., Monge, P. R., & Leonardi, P. M. (2011). Multidimensional networks and the dynamics of sociomateriality: Bringing technology inside the network. *International Journal of Communication, 5*, 682-720.
- Cooper, A. (2012a). A brief history of analytics. *JISC CETIS Analytics Series, 1*(9). Retrieved from <http://publications.cetis.ac.uk/wp-content/uploads/2012/12/Analytics-Brief-History-Vol-1-No9.pdf>
- Cooper, A. (2012b). What is analytics? Definitions and essential characteristics. *JISC CETIS Analytics Series, 1*(5). Retrieved from <http://publications.cetis.ac.uk/wp-content/uploads/2012/11/What-is-Analytics-Vol1-No-5.pdf>
- Dawson, S., Bakharia, A., & Heathcote, E. (2010, May). *SNAPP: Realising the affordances of real-time SNA within networked learning environments*. Paper presented at Networked Learning—Seventh International Conference, Aalborg, Denmark.
- Denley, T. (2012). Austin Peay State University: Degree Compass. *EDUCAUSE Review*. Retrieved from <http://www.educause.edu/ero/article/austin-peay-state-university-degree-compass>
- Downes, S. (2005). E-learning 2.0. *eLearn Magazine*. Retrieved from <http://elearnmag.acm.org/featured.cfm?aid=1104968>
- Ellul, J. (1964). *The technological society*. New York, NY: Vintage Books.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *American Association for Artificial Intelligence, 17*(30), 37-54.
- Fischer, G. (2001). User modeling in human-computer interactions. *User Modeling and User-Adapted Interaction, 11*, 65-86.
- Fritz, J. (2010). Video demo of UMBC's "Check My Activity" tool for students. *EDUCAUSE Review*. Retrieved from <http://www.educause.edu/ero/article/video-demo-umbcs-check-my-activity-tool-students>
- Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science, 122*(3159), 108-111.
- Goldstein, P. J. (2005). Academic analytics: The uses of management information and technology in higher education. *ECAR Key Findings*. Retrieved from <http://net.educause.edu/ir/library/pdf/ERS0508/ekf0508.pdf>
- Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology, 78*(6), 1360-1380.
- Greller, W., & Drachler, H. (2102). Translating learning into numbers: A generic framework for learning analytics. *Educational Technology and Society, 19*(3), 42-57.
- Gruzd, A. (2010). Automated discovery and analysis of online social networks. Retrieved from <https://www.mitacs.ca/events/images/stories/focusperiods/social-presentations/gruzd-mitacs.pdf>
- Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems, 24*(2), 8-12. doi:ieeecomputersociety.org/10.1109/MIS.2009.36
- Hawkey, M. (2012). *CFHE12 analysis: Summary of Twitter activity*. Retrieved from <http://mashe.hawkeye.info/2012/11/cfhe12-analysis-summary-of-twitter-activity/>
- Haythornthwaite, C. (2002). Strong, weak, and latent ties and the impact of new media. *Information Society, 18*, 285-401.

- Haythornthwaite, C., & Andrews, R. (2011). *E-learning theory and practice*. London, UK: Sage.
- Hendler, J., & Berners-Lee, T. (2010). From the semantic web to social machines: A research challenge for AI on the World Wide Web. *Artificial Intelligence, 174*(2), 156-161.
- Herskovitz, Baker, R. S. J., Gobert, J., Wixon, M., & Pedro, M. S. (2013). Discovery with models: A case study on carelessness in computer-based science inquiry. *American Behavioral Scientist*. Advance online publication. doi:10.1177/0002764213479365
- Hey, T., Tansley, S., & Tolle, K. (Eds.) (2009). The fourth paradigm: Data-intensive scientific discovery. *Microsoft Research*. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf
- Hirst, T. (2013). *Exploring ONCourse data a little more*. Retrieved from <http://blog.ouseful.info/2013/01/19/exploring-oucourse-data-a-little-more/>
- Kay, D., Korn, N., & Oppenheim, C. (2012). Legal, risk, and ethical aspects of analytics in higher education. *JISC CETIS Analytics Series, 1*(6). Retrieved from <http://publications.cetis.ac.uk/wp-content/uploads/2012/11/Legal-Risk-and-Ethical-Aspects-of-Analytics-in-Higher-Education-Voll-No6.pdf>
- Kim, M., & Lee, E. (2012). A multidimensional analysis tool for visualizing online interactions. *Educational Technology and Society, 15*(3), 89-102.
- Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., & Haydock, M. (2011). Analytics: The widening gap. *MIT Sloan Management Review*. Retrieved from <http://sloanreview.mit.edu/reports/analytics-advantage/>
- Kuhn, T. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago, IL: University of Chicago Press.
- Latour, B. (2008). *Beware, your imagination leaves digital traces*. Retrieved from <http://www.bruno-latour.fr/sites/default/files/P-129-THES-GB.pdf>
- Lee, V. R., & Thomas, J. M. (2011). Integrating physical activity data technologies into elementary classrooms. *Technology Research and Development, 59*, 865-884.
- Macfadyen, L., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers and Education, 54*(2), 588-599.
- Macfadyen, L. P., & Dawson, S. (2012). Numbers are not enough: Why e-learning analytics fail to inform an institutional strategic plan. *Educational Technology and Society, 15*(3), 149-163.
- Madrigal, A. (2012). How Google builds its maps—and what it means for the future of everything. *Atlantic*. Retrieved from <http://www.theatlantic.com/technology/archive/2012/09/how-google-builds-its-maps-and-what-it-means-for-the-future-of-everything/261913/>
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. Retrieved from http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation
- Milgram, S. (1967). The small world problem. *Psychology Today, 2*, 60-67.
- Morin, E. (2008) *On complexity*. Cresskill, NJ: Hampton Press.
- Morris, L. V., Finnegan, C., & Wu, S. (2005). Tracking student behavior, persistence, and achievement in online courses. *Internet and Higher Education, 8*, 221-231.
- Mumford, L. (1964). The automation of knowledge. *AV Communication Review, 12*(3), 261-276.
- Nielsen, M. (2012). *Reinventing discovery*. Princeton, NJ: Princeton University Press.
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web*. Technical report, Stanford InfoLab. Retrieved from <http://ilpubs.stanford.edu:8090/422/>

- Rich, E. (1979). User modeling via stereotypes. *Cognitive Science*, 3, 329-354.
- Siemens, G., & Baker, R. S. J.d. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 252-254). New York, NY: ACM.
- Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*.
- Suthers, D., & Rosen, D. (2011). A unified framework for multi-level analysis of distributed learning. In *Proceedings of the 1st International Conference on Learning Analytics* (pp. 64-74). New York, NY: ACM.
- Verbert, K., Manouselis, N., Drachsler, H., & Duval, E. (2012). Dataset-driven research to support learning and knowledge analytics. *Educational Technology and Society*, 15(3), 133-148.
- Wellman, B. (1999). *Networks in the global village: Life in contemporary communities*. Boulder, CO: Westview Press.
- Wolpers, M., Jehad, N., Verbert, K., & Duval, E. (2007). Tracking actual usage: The attention metadata. *Educational Technology and Society*, 10(3), 106-121.
- World Economic Forum. (2011). *Personal data: The emergence of a new asset class*. Retrieved from http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf

Author Biography

George Siemens researches learning, technology, networks, analytics, and openness in education. He is the author of *Knowing Knowledge*, an exploration of how the context and characteristics of knowledge have changed and what it means to organizations today, and the *Handbook of Emerging Technologies for Learning*. *Knowing Knowledge* has been translated into Mandarin, Spanish, Italian, Persian, and Hungarian. He is the associate director of the Technology Enhanced Knowledge Research Institute at Athabasca University and a faculty member in the School of Computing and Information Services and the Centre for Distance Education. His research has received numerous national and international awards, including an honorary doctorate from Universidad de San Martín de Porres, for his pioneering work in learning, technology, and networks. He is a founding member of the Society for Learning Analytics Research (<http://www.solaresearch.org/>). In 2008, he pioneered massive open online courses (sometimes referred to as MOOCs) that have included more than 25,000 participants.