CHAPTER

13

Big Data in Education: New Efficiencies for Recruitment, Learning, and Retention of Students and Donors

Andy Peterson

VP for Educational Innovation and Global Outreach, Western Seminary, Charlotte, North Carolina

PREAMBLE

Predictive analytics with Big Data in education will improve educational programs for students and fund-raising campaigns for donors (Siegel, 2013). Research in both educational data mining (EDM) and data analytics (LA) continues to increase (Siemens, 2013; Baker and Siemens, 2014). The key elements of recruitment, learning, and retention can be tracked and increased over time in both cases for administrators and donor development. A few early examples of Big Data programs in these key areas of formal education will be reviewed. The paradigms of educational psychology can inform the construction of systems that maximize "engagement." Contemporary educational technology enables the personalization of education whether face to face or at a distance. More responsive evaluation systems will allow the school to demonstrate the effectiveness of its services. There is more opportunity for education as a sustainable enterprise with the implementation of Big Data, both locally and globally.

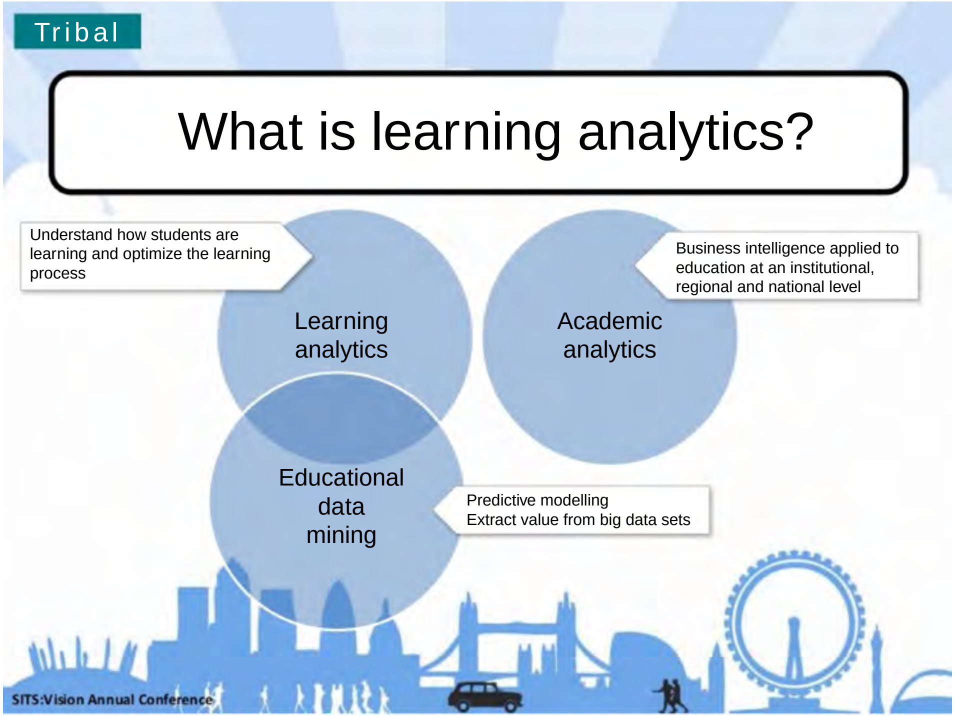
INTRODUCTION

A new and growing element of educational technology is the application of predictive analytics to model performance, retention, and overall learning experience for students in schools and colleges (Parmar et al., 2014). Many articles in education journals use the key phrase "Big Data" in reference to these analytic operations, rather than the actual size of their data sets. Many school administrators depend on patterns in their data discovered by

https://doi.org/10.1016/B978-0-12-416632-5.00013-X

this technology to develop personalized plans for student service and remedial actions and facilitate increase student success rates. Progress in instruction based on these patterns can be monitored and amplified with the right interventions and at the right time. Big Data analyses have great implications for education at all levels, ages and venues, but it is important to understand the potential opportunities and risks in these operations.

In addition to analyses of student performance and retention, studies of how students learn have given rise to the discipline of *learning analytics.* This new discipline includes individualized and personalized monitoring of the course and efficiency of the student learning experience. For example, an "engagement index" for students can be monitored and assessed to help increase student involvement in many aspects of education. Many studies have shown that student performance is related to the amount and diversity of student involvement in campus programs. The explosion of online learning programs provides an ideal platform for the practice of Big Data analytics. One of the most powerful results of these analytic studies is the provision of opportunities and metrics (which can become "drivers") of innovation in the striving for excellence in education. Fig. 13.1 shows the relationship between learning analytics, academic analytics, and educational data mining.



What is learning analytics?

Understand how students are learning and optimize the learning process

Learning analytics

Academic analytics

Business intelligence applied to education at an institutional, regional and national level

Educational data mining

Predictive modelling
Extract value from big data sets

FIG. 13.1 Relationships between learning analytics, academic analytics, and educational data mining. *From https://www.google.com/search?q=learning+analytics+images&biw=1536&bih=735&tbm=isch&imgil=xXRse5i__ZrNKM%253A%253Bk8Qz2vVajIjYzM%253Bhttp%25253A%25252F%25252Fdailygenius.com%25252Flevels-of-learning-analytics%25252F&source=iu&pf=m&fir=xXRse5i__ZrNKM%253A%252Ck8Qz2vVajIjYzM%252C_&usg=__PxEasCf9vDCTihUpAyv756g9wJY%3D&ved=0ahUKEwirqPuYy5LSAhWB6SYKHQMCCX4QyjcIRQ&ei=jYqkWOutA4HTmwGDhKTwBw#imgrc=hSt2DrJo1PqUtM.*

Learning analytics and educational data mining activities in Fig. 13.1 overlap to some extent, while academic analytics do not. Why is this so? To answer that question, we must define the three disciplines.

Learning Analytics

This activity is focused on understanding and optimizing interactions between students and educational system. These interactions include the measurement, analysis, and reporting of student actions for the purpose of understanding of patterns and optimization of the learning environment. These student actions can include the following:

- Time spent in an instructional activity
- Student performance by course and by term
- Interactions of students with the learning environment (e.g., e-book readings)

Educational Data Mining (Analytics)

Apart from the learning experience, this activity analyzes patterns across students and program areas and may include some data captured for use in learning analytics (note the overlap between these two activities). These patterns can include the following:

- Student achievement patterns between academic program areas
- Overall relationships between student performance and academic environment (e.g., student performance patterns related to library study facilities)
- Class tardiness related to the extent and location of bike paths
- Shares some data with learning analytics, hence the overlap in Fig. 13.1

Academic Analytics

These activities analyze nonperformance and nonenvironmental aspects of the educational experience and can include the following:

- Student recruitment
- Student admissions
- Student persistence
- Student retention
- Other administrative areas of the educational institution:
    - ○ Donor development
    - ○ Grant writing
    - ○ School administrative structure (e.g., college vs university organization)

Drivers for Innovation

In addition to needing to increase operational efficiencies, most schools are under increasing pressures and scrutiny to produce educational results leading to professional and financial success of students in the job market. In the face of the high cost of tuition, parents and students expect tangible results of the educational experience. Costs in education have increased at a rate even higher than in health care. Considering the large role of government in providing guaranteed student loans for education, the resulting massive indebtedness that students may generate may cause a burst in the "bubble" of education costs very soon similar to the bursts in the technology bubble in 2000 and the housing bubble in 2007. To prevent this bubble burst, educational institutions must become more efficient in student retention, more effective in fostering student performance and donor development, and increase the "bang for the buck" that students can reap from their educational experience. Big Data analytics can help significantly to make these innovations happen.

Yet, the human element of the teacher and the educational administrator is still central to the goal of making the result of great value to the student and the enterprise and for program sustainability. Student recruitment, performance, and retention will help to stabilize the program financially, which will facilitate learning success in a stable environment, and improve the educational experience by increasing student engagement in the program. In due course, this approach can be applied to donor development, too. The goal is to find patterns in the institutional data base to describe, a least mathematically, the most successful students and the best donors. Evidence for the most significant patterns in successful and unsuccessful students and donors is suggested. Big Data implementation, however, can function as disruptive innovation in the student body and the institution at large, because organizational stress can be as challenging as the technological systems. These benefits and challenges compose an exciting prospect for the future in education.

Future Scenarios

With the explosion of online learning in education, more digital data are captured than ever before. The online platform is especially appropriate for data analytics to analyze and model a growing volume and velocity of information provided by assignment submissions and student interactions (e.g., student forums). There is a huge potential to blend learning analytic operations with a learning management system (LMS) on campus. Of course, the appropriate IT topology is required to capture data in real time and provide output as needed for instructional and social purposes (e.g., student blogs).

Alexander (2014) suggests that there are three possible scenarios for the future of education over the next 10 years: (1) two cultures of service venue: online and on campus, (2) renaissance of social interaction, and (3) development of a "health-care nation." First is the articulation and development of online and on-campus education venues side by side, with management and decision-making powered by Big Data analytics. Secondly, the rich interaction between people as seen in the recent explosion of social media interactions constitutes a renaissance in human behavior that can fuel a more dynamic and interesting educational experience at all levels. And thirdly, the dominance of health care as an industry will attract many new students to the related disciplines. A more scientific and clinical curriculum will benefit greatly from the strategies and tactics that can be employed using increasing amounts of digital data collection and analysis.

The tsunami of data in all industries is beginning to spill over into even the most traditional institutions of education at all levels, producing new research and practice (Sawyer, 2014). Many educational projects with educational data mining and learning analytics are planning data-driven projects (Baker and Siemens, 2014). From K-12 to adult education, more gadgets are being used with the capacity to capture, track, and respond to the formal learning activities of students. These gadgets include the following:

- Cell phones
- Electronic tablets

- Personal computers
- Watches

It is imperative to bring the best practices of instructional design to this complex opportunity to compose it into a blessing and not a curse. It is important to realize the importance of and leverage the opportunities for more individualization and personalization of instruction. This practice of personalized education should foster the increased connection of the educational program with all of the stakeholders of formal education and informal learning activities. Many service and commodity vendors are jumping on this "bandwagon" to provide personalized goods and services keyed to the needs of the educational institutions.

Industry Vendors

Amazon is the most recent entrant to the list of vendors ready to help more people to use data analytics (cf. http://aws.amazon.com/machine-learning). They have tried to provide the hardware and software in the cloud to support analytics, together with step-by-step processes for setup, analysis, and understanding of the results. Amazon also provides cloud-based business management services to many educational institutions. Whether one is a parent, teacher, or administrator, the common goal should be to democratize powerful analytic applications designed to find and use key variables to predict academic and practical achievement outcomes. Likewise, Microsoft has added some tools that may help large organizations to find patterns useful to optimize instructional and administrative processes (e.g., the Microsoft Analytics Platform System). But before these tools can be used effectively by educational staff, concepts of effective data preparation must be applied to the input data before modeling can begin. Thus, the practice of data science in Big Data educational institutions will continue to be a combination of science and art, of which effective data preparation is one of the most important "artistic" components. Therefore, much practice is necessary to develop the artistic aspects of predictive analytics (e.g., knowing how to treat specific data preparation problems).

Critics like Gary King (of the Institute for Quantitative Social Science at Harvard University) are concerned that the publicity about Big Data is misleading for business and education (King, 2011). He is supportive of the application of Big Data in social research but remains cautious of the sometimes overstated claims of Big Data proponents. He and his team are writing reviews of Big Data reports and developing their own approach to innovative teaching and learning. They are concerned about "Big Data hubris" that could lead to overlooking basic practices for accurate measurement validity and reliability. Yet, King does join those who see great promise in the principles and new practices of Big Data (King and Maya, 2013).

Apple has become the world's most valuable company, partly by providing computers and related tools to education, both in and out of the classroom. In addition to the features of usability, reliability, fidelity, etc., Apple has provided closer integration among their tools than competing vendors (i.e., Microsoft). The core element of their products remains computational, but communication has been included over the past decade. Thus, we have the Apple "ecosystem" for education.

Academic Analytics

We will start a discussion of academic analytics, because this activity promises the relatively quick harvesting of "low-hanging fruit" in the form of retaining students who have

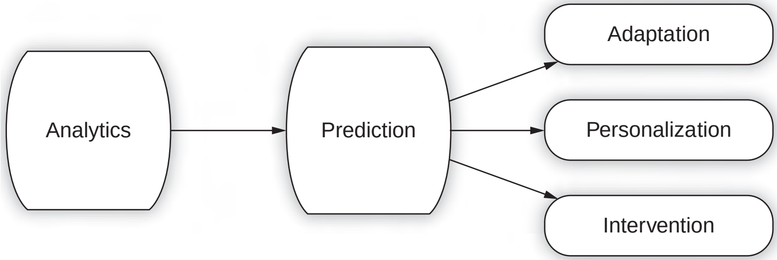


FIG. 13.2 The general process followed by the deployment of student recruitment and retention analyses. *From* *https://www.google.com/search?q=Educational+analytics+images&biw=1536&bih=735&tbm=isch&tbo=u&source=univ&sa=X&ved=0ahUKEwjdwoex0ZXSAhWE5CYKHdfuA54QsAQIGQ&dpr=1.25#imgrc=ofMEsBNy4d5W7M.*

a high probability of leaving before graduation (churning). Students are the "lifeblood" of a school. The primary goal of any school is to lose students, that is, to see them graduate. One of the primary strategies to accomplish this goal is to retain (not lose before graduation) the students they have. The next two most important strategies in administration are to gain new students (student recruitment) and enhance donor development. Apart from the order in which these activities take in an institution, we will discuss them in the logical order of occurrence, beginning with student recruitment, followed by student retention and donor development (beginning with alumni).

The basic process followed in all of these studies is shown in Fig. 13.2.

The general goals of an academic analytics process in deployment are (1) adaptation of old processes to new conditions, (2) personalization of processes to fit each student, and (3) intervention to improve the structure of the process or prevent an undesirable outcome (e.g., student attrition).

Student Recruitment

Analysis of Big Data is poised to make big contributions to predict many outcomes for both students and donors in K-12 and schools of higher education. The primary concerns in this environment are in the areas of student recruitment, retention, and performance. While the interest in these concerns is great in education at this time, case studies involving these analyses are relatively few in number at present, and most of them are in the early stages of implementation. Such that best practices are being developed in the process. But the promise provided by these studies is to help alleviate much of the stress related to soaring costs and to help to accrue the benefits that these analyses can provide for both undergraduate and graduate institutions. The technology of predictive analytics may serve also as a key element of successful recruiting for educational institutions. In addition to the United States, improvement of successful recruiting, retention, and performance tracking will be important worldwide in student management systems.

A good example of the use of predictive analytics for student recruitment was described by Goenner and Pauls (2006) at the University of North Dakota. They used inquiry data from prospects to find variables that were the best predictors of the probability of application and enrollment in the school. Zip codes were shown to contain very useful information related to student recruitment. Even with this limited assessment, an enrollment model with accurate predictions was shown to be effective, even though the included demographic data were relatively simple in nature.

The use of predictive analytics will increase with the acceleration of the use of Web marketing and the expansion of online sale “funnels” and mobile platforms, which are interactive and operate in real time. The goal in building such a funnel for marketing and admissions in student recruitment is to connect with prospects and add them to a prospect database, from which future mailing lists can be extracted. A next step might offer some free or low-cost item of value, if the individual registers with the academic site. Registered contacts are introduced to the blogs of the relevant talent of the school, and podcasts, audio, and/or video is offered on a weekly or monthly basis. Content is generated easily with the use of the interview format. Some e-books can be offered on an admission web page, which might interest prospective students. With all of this infrastructure in place, the prospect is ready to entertain more advanced and expensive items such as a kit to accomplish a major task or even a personal consult. Such steps of incremental information gain lead to the natural next step of filling out an application to the institution. Using mobile smartphones, all of these interactions can take place in the palm of one's hand. Even when prospects are mobile themselves, the filling of the prospect funnel, the vending of the products (educational programs), and, eventually, the actual course of study can transpire in real time. Increased engagement seems to be a primary result of these prospect touchpoints and can be a key element in student enrollment.

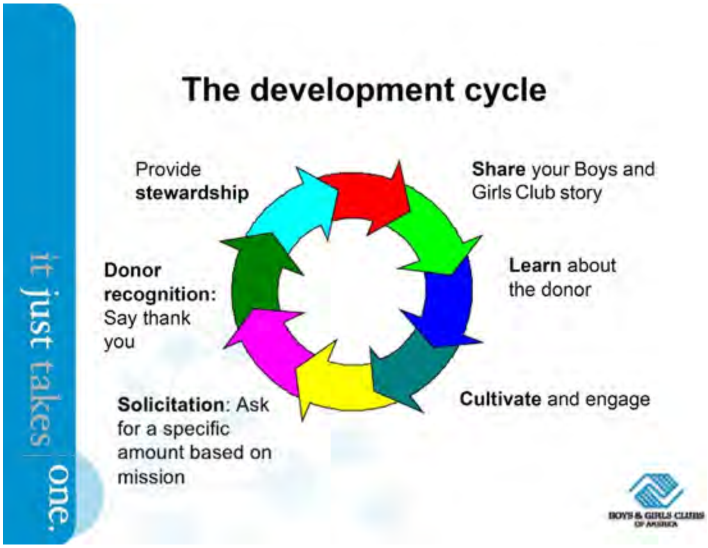
Student Retention

After students begin an educational program, a primary goal for both business and educational purposes is to retain the student in the program until graduation. Graduation is certainly the ultimate goal, but a secondary goal is persistence in the program. Some students leave for a term or two (stop-outs), while others *persist* in their enrollment from term to term. A better learning experience is fostered by greater persistence to graduation and provides a more consistent financial environment for the institution. If there is a well-designed Big Data program in place, then the patterns of successful students can be identified and “shaped” by future students. Beginning with hunches by the educators, data can be acquired and arranged for machine-learning algorithms to find significant combinations of variables for testing and review. It is important to note that these steps constitute much more than just a record of student withdrawal from the program. There is emotional baggage with “churn” (attrition) in education and in business. The primary focus of the administration on the data from successful students does not include these factors, but Big Data analyses can help to analyze them in a much more rigorous way that previously.

Donor Development (Including Alumni Relations)

This activity is discussed under academic analytics, because it includes financial contributions from former students (alumni), and from nonstudents, donor development is an important part of building a sustainable educational institution and includes donor recruitment, donor development (to donate more), and donor retention. There are many analogies between the use of Big Data for student analysis and donor analysis. Fig. 13.3 shows the typical donor development cycle followed by the Boys and Girls Clubs of America.

Each of these elements of the donor development cycle (cf. Fig. 13.3) may be embedded in each of the primary activities of educational donor development: (1) donor recruitment, (2) donor development, and (3) donor retention.



The development cycle

Share your Boys and Girls Club story

Learn about the donor

Cultivate and engage

Solicitation: Ask for a specific amount based on mission

Donor recognition: Say thank you

Provide stewardship

it just takes one.

FIG. 13.3 The donor development cycle used by the Boys and Girls Clubs of America. *From https://www.google.com/search?q=Educational+analytics+images&biw=1536&bih=735&tbm=isch&tbo=u&source=univ&sa=X&ved=0ahUKEwjdwoex0ZXSAhWE5CYKHdfuA54QsAQIGQ&dpr=1.25#tbm=isch&q=donor+development+images&imgrc=ahEqBlCo6in7lM.*

Donor Recruitment

The donor development programs can be early adopters of Big Data analytic processing. In order to fill in a "donor engagement index," various features can be tracked in the donor account. Such activities as responses on direct mail, phone, travel, and event attendance, volunteer can be tracked in the database. These metrics lead to segmentation categories. As appropriate, the standard measures of capacity are taken plus an affinity index for the school relationship. The whole program can then track, display, schedule, and monitor for solicitors and report to leadership for action. For donors (and students), engagement is the primary key to success.

Donor Development

The personal generosity theme and related studies have become important to most schools recently. Certainly, donors intend to give to educational institutions for a variety of reasons. Some want to have recognition. Some want to see progress in a particular mission. But personal growth of generosity and evaluation of the recipient's use of the funds are additional primary reasons for giving. Growth in the trait of generosity is the new goal of contemporary fund-raising. Just as a student can learn in a classroom or online, so, too, can a donor grow in the trait and joy of generosity by forms of engagement with the institution. This can be taught formally and within an immersive experience of high engagement of donors with the institutions to be recipient of the funds. Schools can foster personal growth in generosity as a trait along with the giving of funds. Again, engagement is the primary key for success.

Donor Retention

Like the acquisition of a new student for the school, the recruitment of a new donor is very expensive. Losing that person is disappointing and expensive. Customary estimates are that donor acquisition costs 10 times the expense that can be spent on retention of existing donors. As machine learning goes through the data about the donors, patterns emerge that become "red flags" when a donor is about to disengage. And once they have left, it is not easy to persuade them to come back. It is much better to do the outreach while the relationship is still warm. And once again, engagement is the primary key of success.

Educational Data Mining (Analytics)

The focus of this activity in education is not on individual student responses, but rather upon patterns of student response across the institution. The student response of greatest interest is student academic achievement, and it is the one to which most of the effort is devoted.

Student Achievement

Despite being a critic of the use of Big Data, King and Maya (2013) has collected three findings from social science research that highlight important factors for learning, which can be monitored and applied to student learning. These elements of a teaching/learning situation will be likely to lead to optimal learning marked by (1) motivation from social connections, (2) teaching by other learners, and (3) instant feedback. These activities can be tracked and measured for high-performance learning. Whether in an online or in a local community format, the great opportunities provided by predictive analytics to aid in the teaching of the subject matter and providing a forum for quick feedback can lead to significant improvement in the rate and depth of learning.

With the explosion in online and hybrid (online combined with on campus) education opportunities, data collection and its quick analysis become feasible, which can generate significant improvements in the quality of education. By selecting key variables to track in the midst of completing the course, analysts can identify other very important variables to observe. All variables can be submitted to software algorithms to find relevant patterns of successful and unsuccessful students. These patterns can be used to match future students to predict their success or failure in learning in an academic program.

Interactive multimedia can be combined with the personal contact of a mentor to provide a much richer learning environment than available in either venue separately. In addition, a certificate can be made available from the school via their mobile technology, together with appropriate methods for test proctoring. Such certificates of completion have been shown to be very effective in the certificate programs at the University of California, Irvine, in predictive analytics and data science to move graduates into predictive analytics jobs. Not many degree-granting institutions have implemented programs like this, possibly because of internal resistance to change in this direction.

The use of Big Data analyses in education can provide very precise information, which can be related to student performance. A good example of this promise is to use it to track student behavior in reading assigned e-textbooks and other e-books. Time on task, notation activities, and student collaboration are some of the variables that can be used to predict successful class performance. Unstructured data of social media can be connected to with data on in-class activities to define a series of best behavior patterns to promote student achievement. Since 2007, CourseSmart (now VitalSource) has been the only provider of digital course materials able to combine curriculum, content and delivery into a single solution (https://www.vitalsource.com/).

CourseSmart began as a consortium of publishers who aimed to find the most important predictor variables as students use e-textbooks, for example, note-taking, bookmarking, and time on task. Their goal was to derive a learning engagement measure with a proprietary algorithm and apply it to for each student in the class. The resource is the learner's work in the assigned e-textbooks. The original studies monitored various performance outcomes and retention rates as 100,000 e-books from major publishers (e.g., McGraw-Hill, Pearson, and Houghton Mifflin) were used by one million students. Since the purchase by VitalSource in 2014, a study of CourseSmart was conducted by Dr. Reynol Junco (Fellow of the Berkman Center for Internet and Society, Harvard University) who reported on the "learner engagement index" involving 76 faculty members, 26 administrators, and 3700 students, as measured by recorded work by students in e-textbooks. The related report commends this tool as a "significant step forward" in helping students to succeed in their coursework (cf. http://blog.reyjunco.com/wp-content/uploads/2010/03/FINAL-CourseSmart_Analytics_White_Paper.pdf).

VitalSource (CourseSmart) students work through a page viewer and thumbnail page views are shown in Fig. 13.4A and B (VitalSource, 2017). This approach leverages some of the skills and experience of students with playing video games on a computer.

Another example that is even more advanced technologically is the use of online interactive social simulations for training (cf. NexLearn, http://www.nexlearn.com/). In this study of student activities, student proceeds through a series of simulations of course-related activities, such as social interaction of daily problem solving or educational games. Machine-learning algorithms can uncover patterns that characterize the most successful achievers. Results can prompt various interventions in the course of the simulation or in other parts of the course. Interpretation of these results can improve current immersive learning programs by providing better content, more practical application, unlimited practice, and additional visualization.

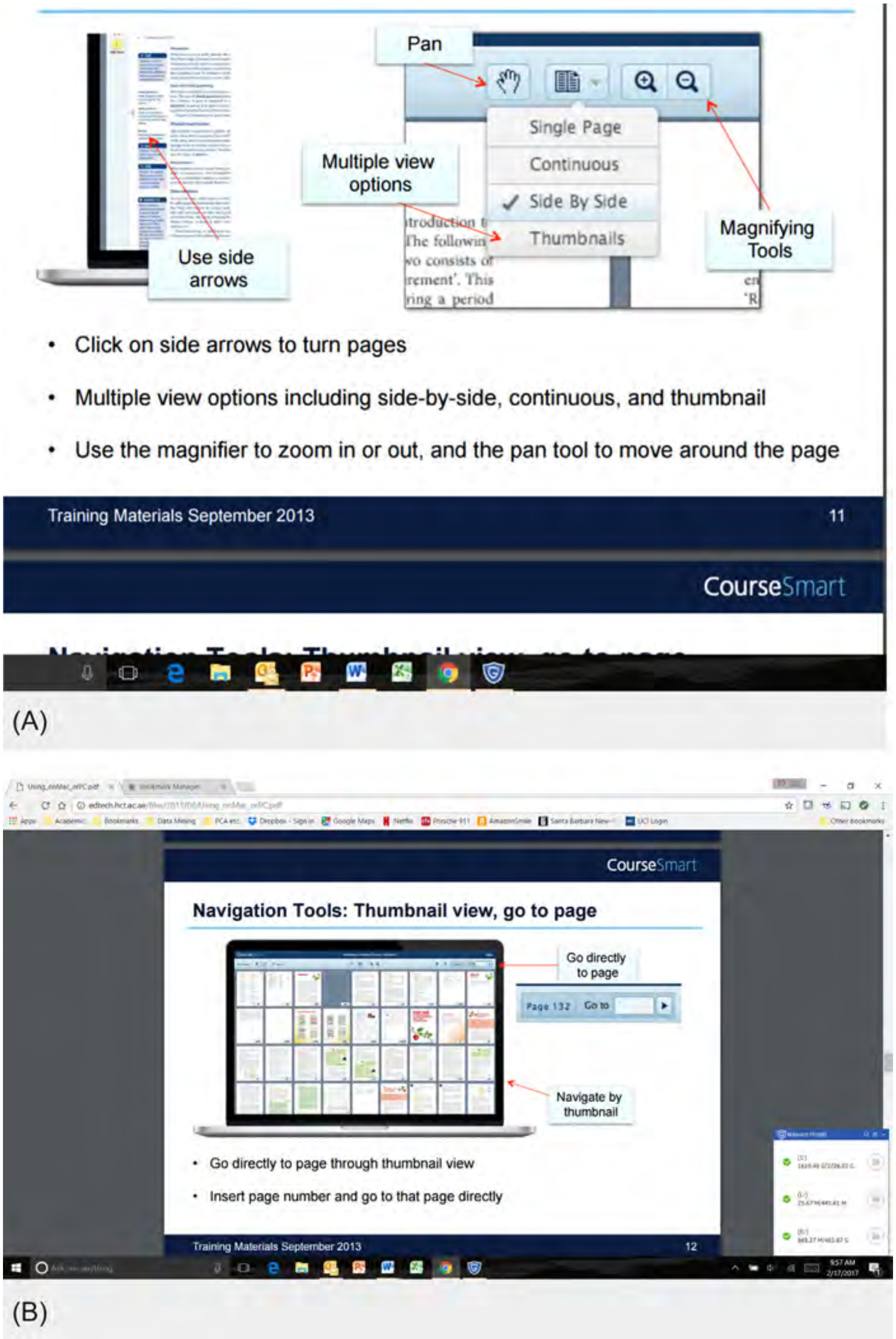
Whether with e-books or simulations, candidate predictor variables could include completeness, note-taking, bookmarking, printing, sharing, collaboration, speed, participation, episode scores, quiz scores, story character choices, and personal avatar choices. And the predicted outcomes could include quizzes (responses correct plus branched practice), exams (cumulative responses correct and course grade), surveys (activity liking, sphere liking, and confidence), magnitude scaling on application preference (not Likert-style surveys), field assignments, and future vocational choices and behavior.

Learning Analytics

Developments in education are driven primarily by trends in educational philosophy; there, a short discussion of the history of educational psychology is presented below as the context for discussion of how Big Data capabilities can contribute to the process and environment of learning.

Education Psychology—A History

Big Data in education is coming of age for reasons other than for student retention and education donor development. For example, it is critical to understand the progression of development of psychology in relation to technology. The application of this relationship



- Click on side arrows to turn pages
- Multiple view options including side-by-side, continuous, and thumbnail
- Use the magnifier to zoom in or out, and the pan tool to move around the page

Training Materials September 2013 11

CourseSmart

CourseSmart

Navigation Tools: Thumbnail view, go to page

- Go directly to page through thumbnail view
- Insert page number and go to that page directly

Training Materials September 2013 12

FIG. 13.4 (A) Page viewer in CourseSmart. Readers can select the single-page view, side-by-side view, and thumbnail view. (B) The thumbnail view of a CourseSmart e-textbook.

to instructional design is a major area of interest, and the understanding of this relationship requires a background in the historical phases of instructional psychology.

The historical phases of development of instructional psychology over the past 100 years have moved from behaviorism to cognitivism to constructivism. One important trend of change is the transition from an emphasis on overt behavior patterns through emphases on covert cognitive algorithms to an emphasis on subjective perception. The precision of behaviorism, the formalization of cognitivism, and the dynamics of constructivism provide the educational designer with a toolbox of key aspects for instruction.

Paradigms in Educational Psychology

BEHAVIORISM

Crosscurrents in educational psychology began early in the 20th century with the behaviorist paradigm. The study of behavior as habitual responses to specific stimuli was the dominant science for education until the 1980s. It produced programmed instruction with an emphasis on immediate correct responses.

COGNITIVISM

When key experiments began to show the inadequacy of a psychology without mind, educational researchers and practitioners became serious about the logical protocols of thinking as part of learning. While unobservable thinking was rejected in the past as not a subject matter for materialistic science, the necessity of hypothetical constructs like cognition became clear. Even for prediction and control with humans, thinking was a necessary variable in the lab and the field.

At the time of this crossing over to cognitive psychology, the advances in computer science had lead to the suggestion of artificial intelligence. The idea of information processing became the model for human thinking and a common phrase in every chapter of general psychology in the 1980s. But like behaviorism and despite the internal constructs about thinking, it remained a mechanistic approach. The ideas of insight, curiosity, challenge, etc. were still absent. Educational leaders from philosophy, psychology, and business were not satisfied without these concepts in education planning for the future.

CONSTRUCTIVISM

Going forward from the 1970s, Hubert Dreyfus at UC Berkeley continued to challenge the assumptions and conclusions from the AI researchers at MIT and Stanford. With his analysis from *What Computers Cant Do* still on the table, he looked for an education with more insight and action as one moved from novice to beginner to expert status. Another scholar who moved from behaviorism and cognitivism to being a constructivist was Omar K. Moore of Yale and the University of Pittsburgh. He was convinced by Gödel's incompleteness theorem that a strictly behavioristic model was impossible (1940). Based on a combination of insights from sociology, psychology, and logic, he executed a 20-year research program to study the best environments for learning including much interaction and data collection and analysis (cf. Moore, 1980).

Currently, the preeminent approach is to set up an environment to promote qualities of content, challenge, and collaboration. Predictive analytics can add metrics to this qualitative approach. The Big Data from the classroom, whether online or on campus, can provide information patterns that can be leveraged to develop a personalized instruction program (Miner et al., 2015).

Industrial Approaches

In the midst of this progressive development psychology, Apple embraced the approach of the proactive learner with an emphasis on challenge, collaboration, and subject content. John Couch, VP of education at Apple operationalized the insights of Dreyfus and Moore into the development of the Apple education ecosystem. With both computers and constructivist education, Apple has produced tools for computation coordinated with communication capabilities. Contemporary predictive analytics combines machine-learning analysis with user intuition to provide an approach to learning analogous to the way the human brain learns (see Chapter 19 for examples of how deep-learning technology can facilitate this process). The Big Data approach can coordinate well with the constructivist approach to educational psychology, particularly in immersive learning with interactive online social simulations.

In the clarifying environment program, Moore (1980) tested four assumptions about learning, all four of which can be quantified and tracked for review and intervention as needed while still allowing great flexibility:

1. Productive principle (heuristic learning should be included for more progress)
2. Perspectives principle (folk models of puzzles, chance, strategy, and aesthetics should be included for a range of angles on learning the content and more)
3. Personalization principle (responsive environments for exploration, discovery, feedback, interrelated domain, and reflexive opportunities should be designed)\
4. Autotelic principle (some learning opportunities in a responsive environment without systematic praise or punishment)

The Apple education ecosystem provides convenient tools to track and adapt to student performance. It allows for use of the assumptions of the clarifying environment program listed above. Big Data in this context provides an evidence-based approach to quantify learner achievement and evaluate educational success and training utility for business. It can be employed in conjunction with real-life mentoring for community application. Big Data becomes a resource for adaptive education and training. This means that the instruction can be adapted “on the fly” as the learner proceeds in the program. There can be dynamic and ongoing assessment within sessions and among students in this model. Summative assessment among various designs and approaches can be conducted to observe long-term student assessment for the individual and long-term program evaluation for the school or workplace.

The Technical Environment—How Does it Fit in?

Two areas of technology interface with the application of Big Data to answer academic questions and solve problems are (1) math and statistical analysis and (2) machine-learning techniques.

Math and Statistical Analysis

Naturally, there is a strong statistical analysis side to Big Data and theoretical context. Moreover, both research design and practical application are being affected by a new model of data collection and new ways of data analysis beyond the more superficial logs and the traditional parametric statistic analysis, where the underlying assumptions are often not met. Much more information can be pulled from the patterns of predictive analytics than the usual model of hypothesis testing of traditional parametric and nonparametric analysis in research, in which its averages and variances inject much noise into the data analysis (Marascuilo and McSweeney, 1977).

Here are some limitations to the use of traditional parametric statistical analysis. Violation of these assumptions can inject significant error in the estimation of parameters upon which the evaluation of statistical significance is based (Nisbet et al., 2009):

1. Assumption of linearity
   All variables are linear in their relationship to the target variable.
   There are no nonlinear effects (but most business applications and most responses are highly nonlinear)
2. Assumption of normality—using the bell curve to estimate probabilities and judge significance.
   Most business data distributions are highly nonnormal. Violation of this assumption might largely invalidate conclusions of analyses of these data distributions.
3. Assumption of independency
   Effects of each variable on the target variable is completely independent of effects of any other variable (this is almost never the case in business situations). Many nonindependent (interaction) effects among predictor variables may be the primary predictors of outcomes in business applications (see Chapter 2).
4. Assumption of homoscedasticity (equal variance throughout the range of a variable)

Parametric modeling algorithms (e.g., multiple linear regression) require much data preparation to enable the algorithm to sense the signal in your data, including the following:

1. Transformation of variable data distributions to approximate a normal distribution
2. Filling of missing values
3. Creation of separate "dummy" variables for each category in a categorical variable
4. Standardization of values to remove effects of very different scales among variables in the data set (which can cause significant bias in parameter estimation)

The process of searching for patterns in Big Data with analytics algorithms is better described as a "data-driven" process, rather than a hypothesis testing exercise. With hypothesis testing, experimental or field groups are compared according to some specified hypothesis as to whether or not there is a significant difference between the means of the samples. In Big Data, the researcher looks for multivariate collections of factors that make the best fit for a prediction of a relevant outcome. The subject pool is the entire population of participant rather than a small sample from which inferences are made. Thus, predictive analytics has as its purpose to find the best connections between predictor variables and predicted outcomes using all the available data directly.

Machine Learning Techniques

In addition to changes in educational psychology since the 19th century, there have been significant changes statistical analysis also. The initial move was from raw numbers to measures of central tendency such as mean, median, and mode and then to variability such as range, variance, and standard deviation. The field of psychometrics uses tests of reliability and concurrent and predictive validity, using these classical statistical methods. The significant differences between two study groups are evaluated using calculated probabilities and levels of significance (e.g., 95% level of confidence) based on the characteristics of the normal curve. Big Data analysis is more like factor analysis as it looks for exemplars by assessing collections of variables that compose patterns in a data set. Yet, intuition and teamwork are required often to recognize patterns with Big Data analytic techniques. The approach is to build a model from *all* the data as opposed to an *inference* from a limited sample to a population as in parametric statistics. This is the approach followed in analysis with machine-learning tools.

Machine-learning tools were developed in the artificial intelligence community in the search for the "intelligent" machine. AI investigators tried to mimic the way the human brain analyzes data and solves problems, which is very different from the way it is done in statistical analysis (see Chapters 1 and 2). Machine-learning tools build patterns of variables (fields in a record in a data table or row in a spreadsheet) in an input data in a manner similar to the way humans do it—case by case (or row by row). Stored patterns are used to match with other similar data with the same variables and output the fidelity (closeness) of the match. This process uses several partitions of the entire data set, rather than a sample of it (as in statistical analysis). With various tests on various combinations of the data, an analytic program moves through all of the data in the population rather than comparing smaller samples to each other. The machine-learning approach to analysis combs through all of the data to find patterns of relationship between predictor variables and an outcome (the "target" variable). This is different from the hypothesis testing approach to experimental design where means of groups are compared. This approach has been applied at the individual student level to study the task completion by students and at entire student population level in studies of student recruitment and retention.

It is clear that Big Data can help in many ways in the design, development, and evaluation of learning programs, but the application of this technology will always need the human touch and insight. The persistent hope for perfect prediction in education or perfect analog to human intelligence will always be limited by at least three factors:

1. The incompleteness theorem of logic and math implies that any algorithm cannot be both complete and consistent (cf. Gödel, 1940).
2. "Counter predictive effect" of humans for liberty maintains that once any theory is published, there will be those who manager to get around its predictive power (Donaldson and Scriven, 2003).
3. Creativity and innovation as the hallmark of good learning of expertise, Omar Moore (1980) and John Couch (Couch and Peterson, 1991) claim that the highest goal is for our learners to exceed our behavioral objectives

Notwithstanding, Big Data can be a great resource for adaptive education and training. It can help with ongoing and dynamic assessment within sessions and students. There is a need for summative quantitative and qualitative assessment among various designs and approaches. Big Data can improve long-term student assessment for the individual student and long-term program evaluation for the school or workplace. All the while, Big Data can comport with the major educational psychology paradigms for learning.

INDUSTRIAL INTEGRATION OF EDUCATIONAL PSYCHOLOGY AND BIG DATA ANALYTICS

The surprising result of the combination of Big Data and the revolution in educational technology is the increased opportunity to make education more personal. These new tools enable individual student needs and interests to be reflected in the form and structure of the assignments by putting more responsibility in the hands of the learner. The increasing affordability of servers for archive and processing, together with new devices for delivery and display, permits greater reliance on both quantitative and qualitative instructional venues. Big Data analytics can guide adaptive instruction for the individual or group to accomplish specific educational objectives, based on appropriate need assessment and situation analysis, according to accepted principles governing the nature of the learning environment.

Apple is a good example of a complete learning ecosystem composed of appropriate hardware and software devices with network connections that enable Big Data analytics to be effective. These hardware devices include desktop and laptop computers, mobile devices along with the projection platforms on television, projectors, wearables, and more. The software available in these learning ecosystems includes the operating system, various apps, iBook Author, and other programs on various Apple devices, connected to collections such as the iTunes store and iTunes U for education. In every case, data are used, generated, collected, and shared in a myriad of ways. The Apple ecosystem contains also a range of content and collaboration tools within iTunes U capacity for education using iPad apps or following iTunes U courses, including those written with iBook Author, a free application for making e-books with multimedia features.

The backbone of the development of online learning over the past decade has been the learning management systems (LMS). This program accommodates a roster of students and delivers their syllabus, course materials, discussion forums, learning activities, and grading structure. Both commercial systems (e.g., Blackboard and Canvas) and open systems (e.g., Moodle) are used in many current applications of distance education and blended learning on campus. They contain a record of interaction with all students and all instructors throughout the course. Another innovation for online and blended education is the electronic book. Learner data from e-books can be tracked as structured responses such as bookmarking or unstructured data such as textual notes. Some publishers are beginning to collaborate on ways to give feedback to students as they use electronic textbooks. The combination of the online platform recorded lectures and e-books allows for the package of educational materials known as a massive open online course (MOOC). Like iTunes U courses, MOOCs are free but are not accredited and normally are not given credit for completion by schools.

These individual tools may lead to the development of the next phase of Big Data in education with its application to "immersive learning." Analysis of online interactive video social simulations can be useful for predicting learner retention and real-time assessment and prediction of individual long-term outcomes. There will still be lectures and discussion forums with personal teachers in immersive learning environments, but online social simulations with avatars can be added in the same way that a textbook would be added to the course materials. Big Data will optimize applying immersive learning to global distance learning enterprises. Immersive learning is practiced in online interactive video social simulations, in which role playing increases learner engagement. This powerful educational dynamic is enhanced with the personalization of scripts and characters enabled by Big Data. Examples of instructional exercises in online immersive scenarios are available in a wide range of disciplines.

In conclusion, Big Data has the potential to improve immersive learning with better content from massive databases, more types of situations for application, unlimited opportunities for practice with appropriate feedback, interventions for student retention going forward, and additional full-bodied visualization for feedback and results from built-in predictive analytics. Big Data can help in many ways in the design, development and evaluation of immersive learning programs, but automation and application will always need the human touch.

The IT infrastructure to support Big Data is also very important to provide the structure and facilities to permit the potential benefits to be realized. Nisbet (2013) maintains that

> How you work it out in a specific educational context will be a much greater challenge. Schools must build systems to handle it. But, they have to build the system with the right architecture, or it won't work right. Data must be prepared properly; as much as 90% of the project time will be spend in data access, data integration, data cleansing, and other data preparation jobs, before the modeling can even begin. Some of that preparation is the subject of my Effective Data Preparation course at UC-Irvine. The big challenge for schools will be that last "mile" in the data pathway (analogous to the last "mile" in a telecommunications network). Those last "mile" problems in deployment can kill a project in any organization, particularly a school.

Big Data is more than just an IT project. The input formation is critical and cannot be haphazard. It is more than a post hoc data mart or a general dashboard. A robust educational psychology with the right statistical approaches and network logic is required to realize educational benefits from Big Data. The actual analytic software is available in many commercial and open-source packages. Examples of free open-source tools include Rattle for R, RapidMiner, and KNIME. Examples of commercial tools (that can be very expensive) include IBM SPSS Modeler, STATISTICA Data Miner, and SAS. Some commercial tools provide substantial educational discounts.

POSTSCRIPT

This chapter presents large amount of information on the role of Big Data in education. Some readers may ask "where do we begin?" One approach to answer that question is to present 12 steps for implementing Big Data analytics in education:

Step 1. Review strategy with your interdisciplinary team.

Evaluate your legacy business questions (sustaining operations), new questions for the current business (sustaining innovation), and new business opportunities (disruptive innovation).

Step 2. Audit your data.

List the database silos with customer data, product data, and interaction in the ecosystem.

Step 3. Survey your customers and graduates.

Step 4. Analyze your *predicted variables*.

Use an accepted method to rank predicted variables to show what is important thus far, for example, GPA and job placement.

Add new predicted variables, including aggregations (e.g., 5-year evaluations) and various abstractions (e.g., time since some action) and combinations of individual variables to derive new variables such as learner engagement.

Step 5. Use some feature selection tool to select those variables with the most potential to be powerful predictors, and generate a “short list” for submission to modeling algorithms.

Step 6. Plan dashboard metrics for stakeholders and decision-makers.

Step 7. Select software application.

Select analytic methods, algorithms, and machine learning.

Step 8. Construct hardware platform to hold the data mart to support Big Data analytics.

Step 9. Design the logical and physical structure of the data mart.

Step 10. Capture student responses and analyze input data.

Step 11. Do staff training for delegation of maintenance of the system.

Step 12. Build adaptive assessment by using interrelated content, self-pacing of questions, and immediate feedback to student and teacher.

References

Alexander, B., 2014. Higher education in 2024: glimpsing the future. EDUCAUSE Rev. 49, 91–98.

Baker, R., Siemens, G., 2014. Educational data mining and learning analytics. In: Sawyer, K. (Ed.), Cambridge Handbook of the Learning Sciences. second ed. Cambridge University Press, New York, NY.

Couch, J., Peterson, A., 1991. Multimedia curriculum development: a K12 campus prepares for the future. Technol. Horiz. Educ. J. 18 (7), 72–80.

Donaldson, S.I., Scriven, M. (Eds.), 2003. Evaluating Social Programs and Problems: Visions for the New Millennium. Erlbaum, Hillsdale, NJ.

Gödel, K., 1940. The Consistency of the Axiom of Choice and of the Generalized Continuum Hypothesis with the Axioms of Set Theory. Princeton University Press, Princeton.

Goenner, C., Pauls, K., 2006. A predictive model of inquiry to enrollment. Res. High. Educ. 47, 935–956.

Junco, R., 2014. Evaluating How the CourseSmart Engagement Index Predicts Student Course Outcomes. CourseSmart, San Mateo, CA.

King, G., 2011. Ensuring the data rich future of the social sciences. Science 331, 719–721.

King, G., Maya, S., 2013. How social science research can improve teaching. PS: Polit. Sci. Polit. 46 (3), 621–629.

Marascuilo, L.A., McSweeney, M., 1977. Nonparametric and Distribution-Free Methods for the Social Sciences. Wadsworth, Belmont, CA.

Miner, L., Bolding, P., Hilbe, J., Goldstein, M., Hill, T., Nisbet, R., Walton, N., Miner, G., 2015. Practical Predictive Analytics and Decisioning Systems for Medicine: Informatics Accuracy and Cost-Effectiveness for Healthcare Administration and Delivery Including Medical Research. Academic Press, New York, NY 1100 pp.

Moore, O.K., 1980. About talking typewriters, folk models, and discontinuities: a progress report on twenty years of research, development and application. Educ. Technol. 20 (2), 15–27.

Nisbet, R., 2013. Personal communication, regarding data preparation in the Predictive Analytics Certificate Program at the University of California, Irvine.

Nisbet, R., Elder, J., Miner, G., 2009. Handbook of Statistical Analysis & Data Mining Applications. Academic Press, New York, NY.

Parmar, R., Mackenzie, I., Cohn, D., Gann, D., 2014. The new patterns of innovation: how to use data to drive growth. Harv. Bus. Rev. January–February 2–11.

Sawyer, R.K. (Ed.), 2014. The Cambridge Handbook of The Learning Sciences. second ed.. Cambridge University Press, New York, NY.

Siegel, E., 2013. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Buy. John Wiley & Sons, Hoboken, NJ.

Siemens, G., 2013. Learning analytics: the emergence of a discipline. Am. Behav. Sci. 57 (10), 1380–1400.

VitalSource, 2017. Using CourseSmart eTextbooks on a Mac or PC. http://edtech.hct.ac.ae/files/2013/08/Using_onMac_orPC.pdf.

Further Reading

Christensen, C.M., 1997. The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail (Management of Innovation and Change). Harvard Business School, Boston, MA.

Christensen, C., Eyring, H., 2011. The Innovative University: Changing the DNA of Higher Education from the Inside Out. Jossey-Bass, San Francisco, CA.

Christensen, C., Horn, M., Johnson, C., 2008. Disrupting Class, Expanded Edition: How Disruptive Innovation Will Change the Way the World Learns. McGraw Hill, New York, NY.

Christensen, C., Grossman, J., Hwang, J., 2009. The Innovator's Prescription: A Disruptive Solution for Health Care. McGraw Hill, New York, NY.

Davenport, T., 2014. Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. Harvard Business Review Press, Boston, MA.

Davenport, T., Patil, D.J., 2012. Data scientist: the sexiest job of the 21st century. Harv. Bus. Rev. October 90–95.

Dreyfus, H., 1972/1992. What Computers Still Can't Do. MIT Press, New York, NY.

Dreyfus, H., 2008. On the Internet, second ed. Routledge, New York, NY.

Peterson, A.J., 1991. Evaluation of hypermedia and interactivity in the museum: a constructivist approach to instructional design. In: Hypermedia & Interactivity in Museums: Proceedings of an International Conference, Pittsburgh, PA, October 14–16. pp. 74–80.

Peterson, A.J., 2015. Big data for faculty development in research and teaching. Theol. Educ. 29 (2), 75–87.

Thompson, C., 2011. How Kahn Academy is changing the rules of education. Wired, July 15. http://aws.amazon.com/machine-learning/.

Tufte, E. Visual presentation of information best practices. http://www.edwardtufte.com/tufte/.