Enhancing Teaching and Learning Through Educational Data Mining and

Learning Analytics

[draft for public comment]

U.S. Department of Education
Office of Educational Technology

Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief

Submitted to:

Bernadette Adams
Office of Educational Technology
U.S. Department of Education
400 Maryland Avenue, SW
Washington, DC 20202

Prepared by:

Marie Bienkowski
Mingyu Feng
Barbara Means

April 10, 2012

Contents

Introduction 1
Scenarios 5
Educational Data Mining and Learning Analytics Research 7
  Educational Data Mining 9
  Learning Analytics 12
  Visual Data Analytics 14
Data Flow in a Typical Adaptive Learning System 17
Educational Data Mining and Learning Analytics Applications 25
  User Knowledge Modeling 27
  User Behavior Modeling 28
  User Experience Modeling 29
  User Profiling 31
  Domain Modeling 32
  Learning System Components and Instructional Principle Analysis 32
  Trend Analysis 33
  Adaptation and Personalization 34
Implementation Challenges and Considerations 35
  Technical Challenges 36
  Limitations in Institutional Capacity 38
  Privacy and Ethics Issues 39
Recommendations 41
  Educators 42
  Researchers and Developers 44
  Collaborations Across Sectors 45
Conclusion 47
References 49
Recommended Reading 55
  Selected Websites 57

Acknowledgments

This issue brief was developed under the guidance of Karen Cator and Bernadette Adams of the U.S. Department of Education, Office of Educational Technology.

At SRI International, Marianne Bakia provided advice and insightful feedback on drafts of the report. Yukie Toyama (now at the University of California, Berkeley) provided research assistance. The report was edited by Mimi Campbell. Kate Borelli produced graphics and layout, assisted by Vickie Watts.

The authors made their best attempt to incorporate the thoughts and experiences of the experts interviewed for this report, Shelby Andrews (Onsophic Inc.), Linda Chaput (Agile Minds Inc.), Michael Freed and Dror Oren (SRI International), Dave Gutelius (Jive Software), Michael Jahrer and Andreas Toescher (Commendo Inc., Austria), Jeff Murphy (Florida Virtual School), Peter Norvig (Google Inc.), Sunil Noronha (Yahoo! Research Labs), Ken Rudin (Zynga Inc.), Steve Ritter (Carnegie Learning, Inc.), Bror Saxberg and David Niemi (Kaplan Inc.), and Chuck Severance (University of Michigan and Sakai Inc.).

The authors are grateful for the deliberations of our technical working group (TWG) of academic experts in educational data mining and learning analytics. These experts provided constructive guidance and comments for this issue brief. The TWG comprised Ryan S. J. d. Baker (Worcester Polytechnic Institute), Gautam Biswas (Vanderbilt University), John Campbell (Purdue University), Greg Chung (National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles), Alfred Kobsa (University of California, Irvine), Kenneth Koedinger (Carnegie Mellon University), George Siemens (Technology Enhanced Knowledge Research Institute, Athabasca University, Canada), and Stephanie Teasley (University of Michigan, Ann Arbor).

Introduction

As more of our commerce, entertainment, communication, and learning are occurring over the Web, the amount of data generated by online activities is skyrocketing. Commercial entities have led the way in developing techniques for harvesting insights from this mass of data for use in identifying likely consumers of their products, in refining their products to better fit consumer needs, and in tailoring their marketing and user experiences to the preferences of the individual. More recently, researchers and developers of online learning systems have begun to explore analogous techniques for gaining insights from learners' activities online.

This issue brief describes data analytics and data mining in the commercial world and how similar techniques (learner analytics and educational data mining) are starting to be applied in education. The brief examines the challenges being encountered and the potential of such efforts for improving student outcomes and the productivity of K-12 education systems. The goal is to help education policymakers and administrators understand how data mining and analytics work and how they can be applied within online learning systems to support educational decision making.

Specifically, this issue brief addresses the following questions:

- What is educational data mining, and how is it applied? What kinds of questions can it answer, and what kinds of data are needed to answer these questions?
- How does learning analytics differ from data mining? Does it answer different questions and use different data?
- What are the broad application areas for which educational data mining and learning analytics are used?
- What are the benefits of educational data mining and learning analytics, and what factors have enabled these new approaches to be adopted?

Online learning systems refer to online courses or to learning software or interactive learning environments that use intelligent tutoring systems, virtual labs, or simulations. Online courses may be offered through a learning/course management system (such as Blackboard, Moodle, or Sakai) or a learning platform (such as Knewton and DreamBox). Examples of learning software and interactive learning environments are those from Kaplan, Khan Academy, and Agile Mind. When online learning systems use data to change in response to student performance, they become *adaptive* learning environments.

- What are the challenges and barriers to successful application of educational data mining and learning analytics?
- What new practices have to be adopted in order to successfully employ educational data mining and learning analytics for improving teaching and learning?

Sources of information for this brief consisted of

- A review of selected publications and fugitive or gray literature (Web pages and unpublished documents) on educational data mining and learning analytics
- Interviews of 15 data mining/analytics experts from learning software and learning management system companies and from companies offering other kinds of Web-based services
- Deliberations of a technical working group of eight academic experts in data mining and learning analytics.

Learning management systems, or LMS, are suites of software tools that provide comprehensive course-delivery functions—administration, documentation, content assembly and delivery, tracking and reporting of progress, user management and self-services, etc. LMS are Web based and are considered a platform on which to build and deliver modules and courses. Open-source examples include Moodle, Sakai, and ILIAS.

This issue brief was inspired by the vision of personalized learning and embedded assessment in the U.S. Department of Education's National Education Technology Plan (NETP, U.S. Department of Education, 2010a). As described in the plan, increasing use of online learning is offering opportunities to integrate assessment and learning so that information useful for improving future instruction can be gathered in nearly real time:

> When students are learning online, there are multiple opportunities to exploit the power of technology for formative assessment. The same technology that supports learning activities gathers data in the course of learning that can be used for assessment…. An online system can collect much more and much more detailed information about how students are learning than manual methods. As students work, the system can capture their inputs and collect evidence of their problem-solving sequences, knowledge, and strategy use, as reflected by the information each student selects or inputs, the number of attempts the student makes, the number of hints and feedback given, and the time allocation across parts of the problem. (U.S. Department of Education 2010a, p. 30)

While students can clearly benefit from this detailed learning data, the NETP also describes the potential value for the broader education community through the concept of an interconnected feedback system:

> The goal of creating an interconnected feedback system would be to ensure that key decisions about learning are informed by data and that data are aggregated and made accessible at all levels of the education system for continuous improvement. (U.S. Department of Education 2010a, p. 35)

The interconnected feedback systems envisioned by the NETP rely on online learning systems collecting, aggregating, and analyzing large amounts of data and making the data available to many stakeholders. These online or adaptive learning systems will be able to exploit detailed learner activity data not only to recommend what the next learning activity for a particular student should be, but also to predict how that student will perform with future learning content, including high-stakes examinations. Data-rich systems will be able to provide informative and actionable feedback to the learner, to the instructor, and to administrators. These learning systems will also provide software developers with feedback that is tremendously helpful in rapidly refining and improving their products. Finally, researchers will use data from experimentation with adaptive learning systems to test and improve theories of teaching and learning.

In the remainder of this report, we

1. present scenarios that motivate research, development, and application efforts to collect and use data for personalization and adaptation;

2. define the research areas of educational data mining and learning analytics and describe the research goals they pursue and the questions they seek to answer about learning at all levels of the educational system;

3. present an abstracted adaptive learning system to show how data are obtained and used, what major components are involved, and how various stakeholders use such systems;

4. examine the major application areas for the tools and techniques in data mining and analytics, encompassing user and domain modeling; and

5. discuss the implementation and technical challenges and give recommendations for overcoming them.

Scenarios

Online consumer experiences provide strong evidence that computer scientists are developing methods to exploit user activity data and adapt accordingly. Consider the experience a consumer has when using Netflix to choose a movie. Members can browse Netflix offerings by category (e.g., Comedy) or search by a specific actor, director, or title. On choosing a movie, the member can see a brief description of it and compare its average rating by Netflix users with that of other films in the same category. After watching a film, the member is asked to provide a simple rating of how much he or she enjoyed it. The next time the member returns to Netflix, his or her browsing, watching, and rating activity data are used as a basis for recommending more films. The more a person uses Netflix, the more Netflix learns about his or her preferences and the more accurate the predicted enjoyment. But that is not all the data that are used. Because many other members are browsing, watching, and rating the same movies, the Netflix recommendation algorithm is able to group members based on their activity data. Once members are matched, activities by some group members can be used to recommend movies to other group members. Such customization is not unique to Netflix, of course. Companies such as Amazon, Overstock, and Pandora keep track of users' online activities and provide personalized recommendations in a similar way.

We are getting very close to a time when personalization will become commonplace in learning. Imagine an introductory biology course. The instructor is responsible for supporting student learning, but her role has changed to one of designing, orchestrating, and supporting learning experiences rather than "telling." Working within whatever parameters are set by the institution within which the course is offered, the instructor elaborates and communicates the course's learning objectives and identifies resources and experiences through which those learning goals can be attained. Rather than requiring all students to listen to the same lectures and complete the same homework in the same sequence and at the same pace, the instructor points students toward a rich set of resources, some of which are online and some of which are provided within classrooms and laboratories. Students learn the required material by building and following their own learning maps.

Suppose a student has reached a place where the next unit is population genetics. In an online learning system, the student's dashboard shows a set of 20 different population genetics learning resources, including lectures by a master teacher, sophisticated video productions emphasizing visual images related to the genetics concepts, interactive population genetics simulation games, an online collaborative group project, and combinations of text and practice exercises. Each resource comes with a rating of how much of the population genetics portion of the learning map it covers, the size and range of learning gains attained by students who have used it in the past, and student ratings of the resource for ease and enjoyment of use. These ratings are derived from past activities of all students, such as "like" indicators, assessment results, and correlations between student activity and assessment results. The student chooses a resource to work with, and his or her interactions with it are used to continuously update the system's model of how much he or she knows about population genetics. After the student has worked with the resource, the dashboard shows updated ratings for each population genetics learning resource; these ratings indicate how much of the unit content the student has not yet mastered is covered by each resource. At any time, the student may choose to take an online practice assessment for the population genetics unit. Student responses to this assessment give the system—and the student—an even better idea of what he or she has already mastered, how helpful different resources have been in achieving that mastery, and what still needs to be addressed. The teacher and the institution have access to the online learning data, which they can use to certify the student's accomplishments.

Capturing the Moment of Learning by Tracking Game Players' Behaviors

The Cyber-enabled Teaching and Learning through Game-based, Metaphor-Enhanced Learning Objects (CyGaMEs) project was successful in measuring learning using assessments embedded in games. CyGaMEs quantifies game play activity to track timed progress toward the game's goal and uses this progress as a measure of player learning. CyGaMEs also captures a self-report on the game player's engagement or flow, i.e., feelings of skill and challenge, as these feelings vary throughout the game play. In addition to timed progress and self-report of engagement, CyGaMEs captures behaviors the player uses during play. Reese et al. (in press) showed that this behavior data exposed a prototypical "moment of learning" that was confirmed by the timed progress report. Research using the flow data to determine how user experience interacts with learning is ongoing.

This scenario shows the possibility of leveraging data for improving student performance; another example of data use for "sensing" student learning and engagement is described in the sidebar on the moment of learning and illustrates how using detailed behavior data can pinpoint cognitive events.

The increased ability to use data in these ways is due in part to developments in several fields of computer science and statistics. To support the understanding of what kinds of analyses are possible, the next section defines educational data mining and learning analytics and describes the techniques they use to answer questions relevant to teaching and learning.

Educational Data Mining and Learning Analytics Research

Using data for making decisions is not new; companies use complex computations on customer data for *business intelligence* or *analytics*. Business intelligence techniques can discern historical patterns and trends from data and can create models that predict future trends and patterns. Analytics, broadly defined, comprises applied techniques from computer science, mathematics, and statistics for extracting *usable information* from very large datasets.

An early example of using data to explore online behavior is *Web analytics* using tools that log and report Web page visits, countries or domains where the visit was from, and the links that were clicked through. Web analytics are still used to understand and improve how people use the Web, but companies have now developed more sophisticated techniques to track more complex user interactions with their websites. Examples of such tracking include changes in buying habits in response to disruptive technology (e.g., e-readers), most-highlighted passages in e-books, browsing history for predicting likely Web pages of interest, any changes in game players' habits over time. Across the Web, social actions such as bookmarking to social sites, posting to Twitter or blogs, and commenting on stories can be tracked and analyzed.

Analyzing these new logged events requires new techniques to work with unstructured text and image data, data from multiple sources, and vast amounts of data ("big data"). Manyika et al. (2011) defined big data as "…datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze." Big data captured from users' online behaviors enables algorithms to infer the users' knowledge, intentions, and interests and to create models for predicting future behavior and interest.

Data are often put into a *structured format,* as in a relational database. Structured data are easy for computers to manipulate. In contrast, ***unstructured data*** have a semantic structure that is difficult to discern computationally (as in text or image analysis) without human aid. As a simple example, an e-mail message has some structured parts—To, From, and Date Sent— and some unstructured parts—the Subject and the Body.

Machine learning approaches to data mining deal with unstructured data, finding patterns and regularities in the data or extracting semantically meaningful information.

Research on machine learning has yielded techniques for knowledge discovery or data mining that discover novel and potentially useful information in large amounts of unstructured data. These techniques find patterns in data and then build *predictive models* that probabilistically predict an outcome. Applications of these models can then be used in computing analytics over large datasets.

Knowledge discovery in databases is an interdisciplinary area focusing on methodologies for extracting useful knowledge from data. The extraction of knowledge from data draws on research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing to deliver advanced business intelligence and Web discovery solutions.

http://researcher.ibm.com/view_pic.php?id=144

Two areas now developing that are specific to the use of big data in education are *educational data mining* and *learning analytics.* Although there is no hard and fast distinction between these two fields, they have had somewhat different research histories and are developing as distinct research areas. Generally, *educational data mining* is looking for new patterns in data and developing new algorithms and/or new models, while *learning analytics* is applying known predictive models in instructional systems. We discuss each in turn below.

Educational Data Mining

Educational data mining is emerging as a research area with a suite of computational and psychological methods and research approaches for understanding how students learn. New computer-supported interactive learning methods and tools—intelligent tutoring systems, simulations, games—have opened up opportunities to collect and analyze student data, to discover patterns and trends in those data, and to make new discoveries and test hypotheses about how students learn. Data collected from online learning systems can be aggregated over large numbers of students and can contain many variables that data mining algorithms can explore for model building.

Educational data mining (EDM) develops methods and applies techniques from statistics, machine learning, and data mining to analyze data collected during teaching and learning. EDM tests learning theories and informs educational practice.

Learning analytics applies techniques from information science, sociology, psychology, statistics, machine learning, and data mining to analyze data collected during education administration and services, teaching and learning. Learning analytics creates applications that directly influence educational practice.

Educational data mining researchers have been organizing yearly international conferences since 2008. The *Journal of Educational Data Mining* was launched in 2009, and in 2011 the International Educational Data Mining Society (http://www.educationaldatamining.org) was founded by the International Working Group on Educational Data Mining.

Just as with early efforts to understand online behaviors, early efforts at educational data mining involved mining website log data (Baker and Yacef 2009), but now more integrated, instrumented, and sophisticated online learning systems provide more kinds of data. Educational data mining generally emphasizes reducing learning into small components that can be analyzed and then influenced by software that adapts to the student (Siemens and Baker 2012). Student learning data collected by online learning systems are being explored to develop predictive models by applying educational data mining methods that classify data or find relationships. These models play a key role in building adaptive learning systems in which adaptations or interventions based on the model's predictions can be used to change what students experience next or even to recommend outside academic services to support their learning.

An important and unique feature of educational data is that they are hierarchical. Data at the keystroke level, the answer level, the session level, the student level, the classroom level, the teacher level, and the school level are nested inside one another (Baker 2011; Romero and Ventura 2010). Other important features are time, sequence, and context. Time is important to capture data such as length of practice sessions or time to learn. Sequence represents how concepts build on one another and how practice and tutoring should be ordered. Context is important for explaining results and knowing where a model may or may not work. Methods for

hierarchical data mining and longitudinal data modeling have been important developments in mining educational data.

Educational data mining researchers view the following as the goals for their research:

1. Predicting students' future learning behavior by creating student models that incorporate such detailed information as students' knowledge, motivation, metacognition, and attitudes
2. Discovering or improving domain models that characterize the content to be learned and optimal instructional sequences
3. Studying the effects of different kinds of pedagogical support that can be provided by learning software
4. Advancing scientific knowledge about learning and learners through building computational models that incorporate models of the student, the domain, and the software's pedagogy.

To accomplish these four goals, educational data mining research uses the five categories of technical methods (Baker 2011) described below.

1. ***Prediction*** entails developing a model that can infer a single aspect of the data (predicted variable) from some combination of other aspects of the data (predictor variables). Examples of using prediction include detecting student behaviors (e.g., when they are gaming the system, engaging in off-task behavior, or failing to answer a question correctly despite having a skill), developing domain models, and predicting and understanding student educational outcomes (Baker, Gowda, and Corbett 2011).
2. ***Clustering*** refers to finding data points that naturally group together and can be used to split a full dataset into categories. Examples of clustering applications are grouping students based on their learning and interaction patterns (Amershi and Conati 2009) and grouping users for purposes of recommending actions and resources to similar users.
3. ***Relationship mining*** involves discovering relationships between variables in a dataset and encoding them as rules for later use. For example, relationship mining can identify the relationships among products purchased in online shopping (Romero and Ventura 2010).

Educational Data Mining and Educational Data

"Educational data mining…exploits statistical, machine-learning, and data-mining… algorithms over...different types of educational data.… EDM is concerned with developing methods to explore the unique types of data in educational settings and, using these methods, to better understand students and the settings in which they learn.... EDM seeks to...develop computational approaches that combine data and theory to transform practice…."

Romero and Ventura 2010, p.601

"Whether educational data is taken from students' use of interactive learning environments, computer-supported collaborative learning, or administrative data from schools and universities, it often has multiple levels of meaningful hierarchy, which often need to be determined by properties in the data itself, rather than in advance. Issues of time, sequence, and context also play important roles in the study of educational data."

- *Association rule mining* can be used for finding student mistakes that co-occur, associating content with user types to build recommenders for content that is likely to be interesting, or making changes to teaching approaches (e.g., Merceron and Yacef 2010).
- *Sequential pattern mining* builds rules that capture the connections between occurrences of sequential events, for example, finding temporal sequences such as student mistakes followed by help seeking.

Key educational applications of relationship mining include discovery of associations between student performance and course sequences and discovering which pedagogical strategies lead to more effective or robust learning.

4. ***Distillation for human judgment.*** This technique involves depicting data in a way that enables a human to quickly identify or classify features of the data. This area of educational data mining improves machine-learning models because humans identify patterns in student learning, behavior, or collaboration data. This approach overlaps with visual data analytics, described below.
5. ***Discovery with models.*** This technique involves using a validated model of a phenomenon (developed through prediction, clustering, or manual knowledge engineering) as a component in further analysis. For example, Jeong and Biswas (2010) built models that categorized student activity from basic behavior data: students' interactions with a game-like learning environment that uses learning by teaching. A sample student activity discerned from the data was "map probing." A model of map probing was then used within a second model of learning strategies and helped researchers study how the strategy varied across different experimental states. Discovery with models supports discovery of relationships between student behaviors and student characteristics or contextual variables, analysis of research questions across a wide variety of contexts, and integration of psychometric modeling frameworks into machine-learned models.

Education data mining technologies and applications can build models to answer such questions as

- What sequence of topics is most effective for a specific student?
- What student actions are associated with more learning (e.g., higher course grades)?
- What student actions indicate satisfaction, engagement, learning progress, etc.?
- What features of an online learning environment lead to better learning?
- What will predict student success?

Learning Analytics

Learning analytics is becoming defined as an area of research and application and is related to academic analytics, action analytics, and predictive analytics.[1] Beginning in 2010, a yearly conference has been held on learning analytics (the International Conference on Learning Analytics and Knowledge), and in 2011 a professional society was founded, the Society for Learning Analytics Research (www.solaresearch.org).

Learning analytics draws on a broader array of academic disciplines than educational data mining, incorporating concepts and techniques from information science and sociology in addition to computer science, statistics, psychology, and the learning sciences. Unlike educational data mining, learning analytics generally does not emphasize reduction of learning into components but instead seeks to understand entire systems and support human decision making instead of pursuing automated responses.

Learning analytics emphasizes measurement and data collection as activities that institutions need to undertake and understand and focuses on the analysis and reporting of the data. Unlike educational data mining, learning analytics does not generally address the development of new computational methods for data analysis but instead addresses the application of known methods and models to answer important questions that affect student learning and organizational learning systems. The 2011 Horizon Report describes the goal of learning analytics as enabling teachers and schools to tailor educational opportunities to each student's level of need and ability (Johnson et al. 2011). Unlike educational data mining, which emphasizes automated responses to students, learning analytics enables tailoring of responses, such as through adapting instructional content, intervening with at-risk students, and providing feedback.

Defining Learning Analytics

"Learning analytics refers to the interpretation of a wide range of data produced by and gathered on behalf of students in order to assess academic progress, predict future performance, and spot potential issues. Data are collected from explicit student actions, such as completing assignments and taking exams, and from tacit actions, including online social interactions, extracurricular activities, posts on discussion forums, and other activities that are not directly assessed as part of the student's educational progress. Analysis models that process and display the data assist faculty members and school personnel in interpretation. The goal of learning analytics is to enable teachers and schools to tailor educational opportunities to each student's level of need and ability.

"Learning analytics need not simply focus on student performance. It might be used as well to assess curricula, programs, and institutions. It could contribute to existing assessment efforts on a campus, helping provide a deeper analysis, or it might be used to transform pedagogy in a more radical manner. It might also be used by students themselves, creating opportunities for holistic synthesis across both formal and informal learning activities."

Johnson et al. 2011, p. 28

[1] Academic analytics is described in Goldstein (2005). The term "learning analytics" began to be used in 2009–2010. Differences among these terms are not important for purposes of this brief. The interested reader may wish to consult Elias (2011).

Technical methods used in learning analytics are varied and draw from those used in educational data mining. Additionally, learning analytics may employ

- *Social network analysis* (e.g., analysis of student-to-student and student-to-teacher relationships and interactions to identify disconnected students, influencers, etc.), and
- *Social or "attention" metadata* to determine what a user is engaged with.

As with educational data mining, visualization of the analytics is critical to generate actionable analyses and often are represented as "dashboards" that show data in an easily digestible form.

Learning analytics systems apply models to answer such questions as

- When are students ready to move onto the next topic?
- When are students falling behind in a course?
- When is a student at risk for not completing a course?
- What grade is a student likely to get without intervention?
- What is the best next course for a given student?
- Should a student be referred to a counselor for help?

A key application of learning analytics is monitoring and predicting students' learning performance and spotting potential issues early so that interventions can be provided to identify students at risk of failing a course or program of study (EDUCAUSE 2010; Johnson et al. 2011). Several learning analytics models have been developed to identify student risk level in real time to increase the students' likelihood of success. Examples of such systems include Purdue University's Course Signals system (Arnold, 2010) and the Moodog system being used at the course level at the University of California, Santa Barbara and at the institutional level at the University of Alabama (EDUCAUSE, 2010). Higher education institutions have shown increased interest in learning analytics as they face calls for more transparency and increased scrutiny of their recruitment and retention practices.

Data mining analysis of student behavior variables in online courses has revealed differences between successful and unsuccessful students (as measured by final course grades) in terms of such variables as level of participation in discussion boards, number of e-mails sent, and number of quizzes completed (Macfayden and Dawson 2010). Analytics based on these student behavior variables can be used in feedback loops to provide more fluid and flexible curricula and to support immediate course alterations (e.g., sequencing of examples, exercises, and self-assessments) based on analyses of real- time learning data (Graf and Kinshuk, in press).

Visual Data Analytics

> ***Visual data analysis*** a way of discovering and understanding patterns in large datasets via visual interpretation. It is used in the scientific analysis of complex processes. As the tools to interpret and display data have become more sophisticated, models can be manipulated in real time, and researchers are able to navigate and explore data in ways that were not possible previously. Visual data analysis is an emerging field, a blend of statistics, data mining, and visualization, that promises to make it possible for anyone to sift through, display, and understand complex concepts and relationships.
>
> *Johnson et al. 2010, p. 7*

Visual data analysis blends highly advanced computational methods with sophisticated graphics engines to tap the ability of humans to see patterns and structure in complex visual presentations (Johnson et al. 2010). Visual data analysis is designed to deal with very large heterogeneous and dynamic datasets collected from complex systems to help expose patterns, trends, exceptions, and more. A variety of techniques and tools are emerging to enable analysts to easily interpret all sorts of data. For instance, visual interactive principal components analysis (finding the components of a dataset that reduce many variables into few) is a technique once available only to statisticians that is now commonly used to detect trends and data correlations in multidimensional data sets. Gapminder (http://www.gapminder.org/), for example, uses this approach in its analysis of multivariate datasets over time. Websites such as ManyEyes (http://www-958.ibm.com/software/data/cognos/manyeyes/) offer tools for any user to create visualizations (map based, text-based clouds and networks, and charts and graphs) of personal datasets. Early in its release, the creators of ManyEyes discovered that it was being used to do visual analytics, check for data quality, characterize social trends, and reveal personal and collective sentiments or advocate for a position (Viégas et al. 2008). Like ManyEyes, other online services such as Wordle and Flowing Data accept uploaded data and allow the user to configure the output to varying degrees. To facilitate the development of this field, the National Visualization and Analytics Center was established to provide strategic leadership and coordination for visual analytics technology and tools nationwide.

The 2010 Horizon Report (Johnson et al. 2010) describes the promise of visual data analysis (in the 4- to 5-year time frame) for teaching undergraduates to model complex processes in subjects such as quantum physics. Visual data analysis may also help expand our understanding of learning because of its ability to support the search for patterns. It may be applied, for example, to illustrate the relationship among the variables that influence informal learning and to "see" the social networking processes at work in the formation of learning communities.

Currently, the tools, techniques, and high-resolution displays that enable people to interactively manipulate variables or zoom through the analysis results are still found mostly in research settings. Because interpreting data generated for visual data analysis requires analytical knowledge, researchers have thus far been the major population to use this method. Nevertheless, sites such as GapMinder offer teacher professional development and data aimed at educators.

Social Explorer, for example, offers tools for exploring map-based census and demographic data visualizations and is used by both researchers and educators. In the future, advances in visual data analytics and human-computer interface design may well make it feasible to create tools such as ManyEyes that policymakers, administrators, and teachers can use.

We have described the promise of educational data mining (seeking patterns in data across many student actions), learning analytics (applying predictive models that provide actionable information), and visual data analytics (interactive displays of analyzed data) and how they might serve the future of personalized learning and the development and continuous improvement of adaptive systems. How might they operate in an adaptive learning system? What inputs and outputs are to be expected? In the next section, we answer these questions by giving a system-level view of how these techniques could improve student performance by creating feedback loops.

Data Flow in a Typical Adaptive Learning System

Online learning systems—learning management systems, learning platforms, and learning software—have the ability to capture streams of fine-grained learner behaviors, and the tools and techniques described above can operate on the data to provide a variety of stakeholders with feedback to improve teaching, learning, and educational decision making. To demonstrate how such adaptive systems operate, we describe a prototypical learning system with six components (Exhibit 1):

- A content management, maintenance, and delivery component interacts with students to deliver individualized subject content and assessments to support student learning.
- A student learning database (or other big-data repository) stores time-stamped student input and behaviors captured as students work within the system.
- A predictive model combines demographic data (from an external student information system) and learning/behavior data from the student learning database to track a student's progress and make predictions about his or her future behaviors or performance such as future course outcomes and dropouts.
- A reporting server uses the output of the predictive model to produce dashboards that provide visible feedback for various users.
- An adaption engine regulates the content delivery component based on the output of the predictive model to deliver material according to a student's performance level and interests, thus ensuring continuous learning improvement.
- An intervention engine allows teachers, administrators, or system developers to intervene and override the automated system to better serve a student's learning.

Exhibit 1.
The Components and Data Flow Through a Typical Adaptive Learning System

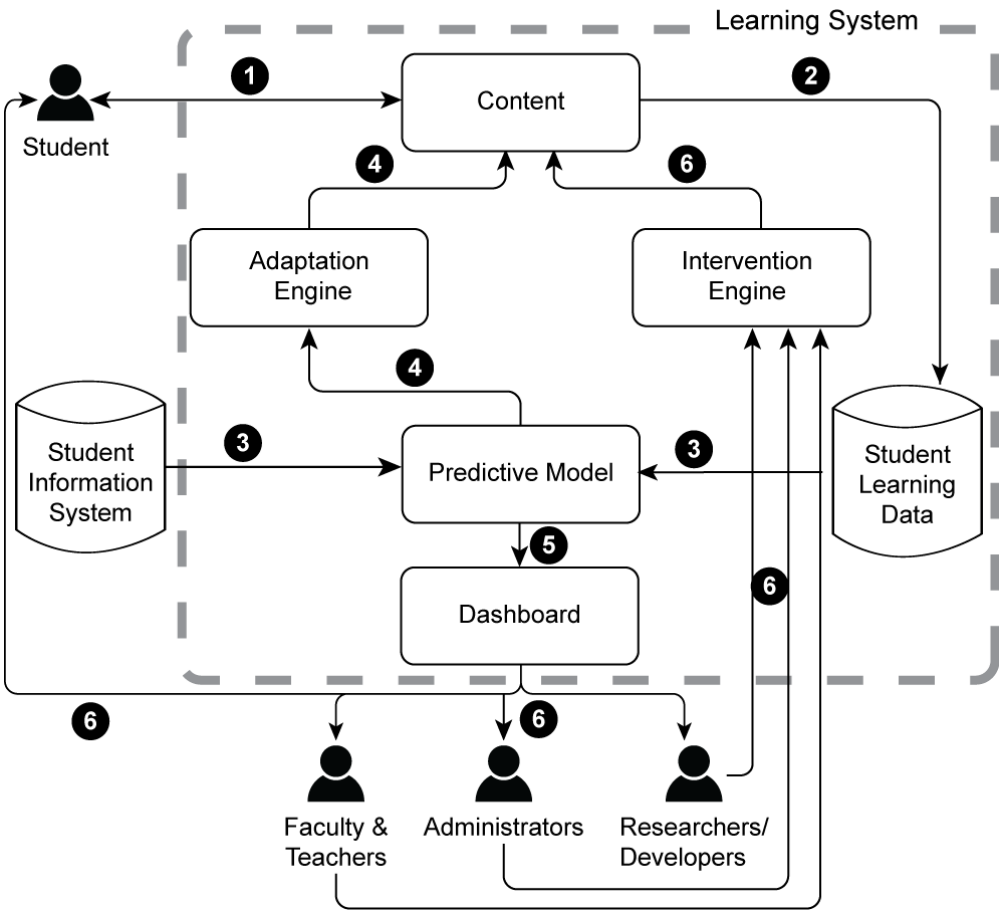


Exhibit reads: The data flow is shown through a box and arrows diagram with a content box on the top with an arrow to a student and two engines underneath shown as boxes: an adaptation engine and an intervention engine, with arrows up to the content box and down to a predictive model. The predictive model is connected to two databases with arrows. On the right is the student learning database and on the left is the student information system. Below the predictive model and connected with an arrow is a dashboard that is shown connected with arrows to faculty and educators, administrators, and researchers and developers.

In addition to these six internal components (shown inside the dashed line in Exhibit 1), an adaptive learning system often uses the student information system (SIS) that is maintained by a school, district, or institution as an external data source. Student profiles from the SIS are usually downloaded in batch mode, as they do not change often, and then are linked with performance data in the student learning database using student identifiers. Student profiles contain background information on students that can be used to group them into specific categories or to provide more variables that might suggest a particular student is at risk.

The numbers in Exhibit 1 signify the data flow that creates feedback loops between the users and the adaptive learning system. The data flow starts with Step 1, students generating inputs when interacting with the content delivery component. (In the future, a student may have a portable learning record that contains information from all past interactions with online leanring systems.) The inputs are time-stamped and cleaned as necessary and stored in the learning database according to predefined schema (Step 2). At certain times (not synchronized with student learning activities), the predictive model fetches data for analysis from both the student learning database and the SIS (Step 3). At this stage, different data mining and analytics tools and models might be applied depending on the purpose of the analysis. Once the analysis is completed, the findings are fed into the content delivery component via the adaptation engine, and these findings are used to adjust computer tutoring and teaching strategies (Step 4). The findings may also flow to the dashboard (Step 5). In the last step in the data flow, various users of the system examine the reports for feedback and respond in ways appropriate for their role (Step 6).

These last steps complete the feedback loop as various stakeholders receive information to inform their future choices and activities. *Students* receive feedback on their interactions with the content they are learning through the adaptive learning system. The feedback typically includes the percentage correct on embedded assessments and lists of concepts they have demonstrated mastery on, but it can also include detailed learning activity information (e.g., hints requested and problems attempted). Detailed learning information for one student can be compared with that for students who earned high grades so that students can adjust their learning with the system accordingly.

Teachers receive feedback on the performance of each individual student and of the class as a whole and adjust their instructional actions to influence student learning. By examining the feedback data, instructors can spot students who may need additional help or encouragement to spend more time on the content and identify areas where the class as a whole is struggling. The latter area can be addressed during class time when the instructor can respond to questions and address student misconceptions. For the former areas, teachers may choose to intervene with the system to adjust student learning pace or may assign additional learning materials targeting the skills that are not yet mastered (see Case Study 1 sidebar).

Measuring Student Effort

Learning software collects such data as minutes spent on a unit, hints used, and common errors and aggregates these data across many students in a school or schools in a district (Koedinger, McLaughlin, and Heffernan 2010). Using these measures, teachers can distinguish between students who are not trying and those who are trying but still struggling and then differentiate instruction for each group.

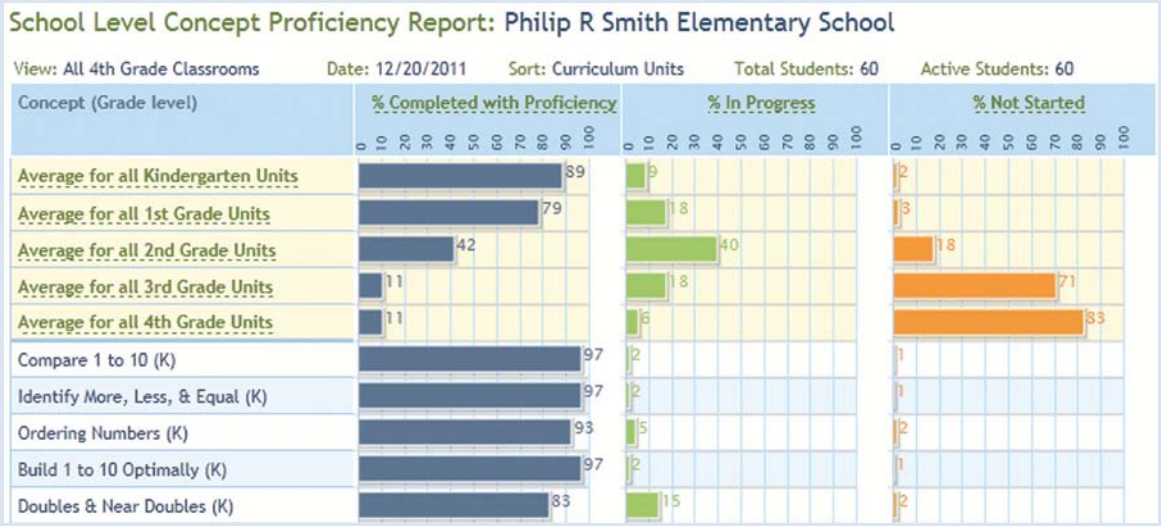
Administrators can look at detailed data across different classes to examine progress for all students at a school. Typically, the detailed learning data the system provides can be disaggregated by student subgroup (for example, to see how students without a course prerequisite do or to compare males' and females' progress in the course), by instructor, or by year. Learning system data can support analyses of how well students learn with particular interventions and how implementation of the intervention could be improved. Using the data, administrators can set policies, implement programs, and adapt the policies and programs to improve teaching, learning, and completion/retention/graduation rates.

Tracking Topic Mastery

Learning systems typically track the state of student mastery at the skill or topic level (e.g., the quadratic equation) and can provide this information to students so they know what to study and to teachers so they know the areas where they should concentrate further instruction.

Aggregating Data for Administrators in a Dashboard

Administrators can benefit from dashboards that show data across teachers and classes at a school or across schools. Shown below is a dashboard from a commercial provider, DreamBox, which captures concept proficiency. District administrators can use data from this kind of dashboard as a basis for determining whether a particular learning intervention is effective at promoting student learning, even at the level of individual concepts.



Researchers can use the same fine-grained data to experiment with learning theories and to examine the effectiveness of different types of instructional practices and different course design elements. *Learning system developers* can conduct rapid testing with large numbers of users to improve online learning systems to better serve students, instructors, and administrators. Researchers using online learning systems can do experiments in which many students are

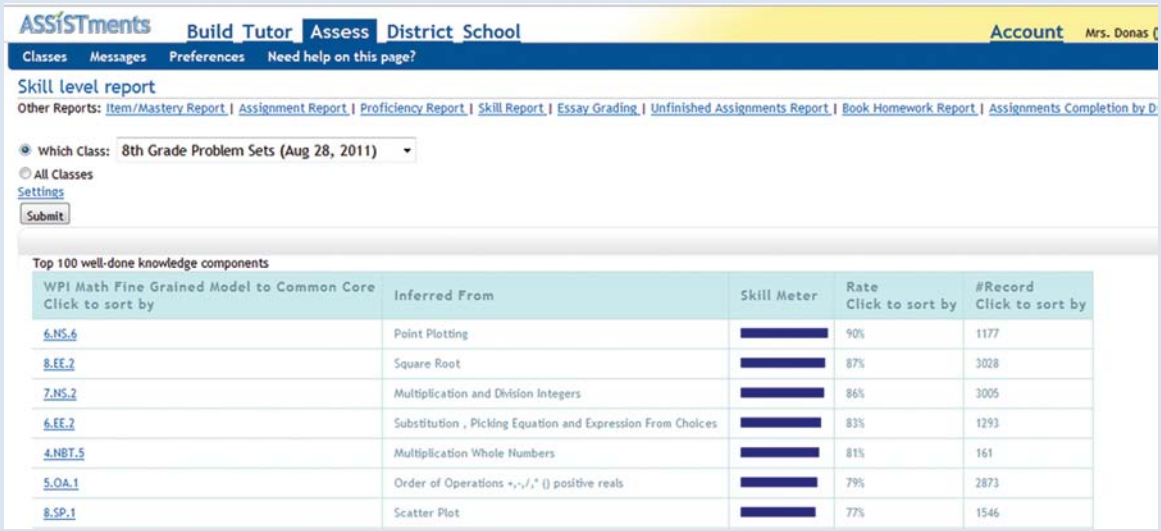
assigned at random to receive different teaching or learning approaches, and learning system developers can show alternative versions of the software to many users: version A or version B. This so-called "A/B testing" process can answer research questions about student learning such as, Do students learn more quickly if they receive a lot of practice on a given type of problem all at once ("massed practice") or if practice on that type of problem is spaced out over time ("spaced practice")? What about students' retention of this skill? Which kind of practice schedule is superior for fostering retention? For what kind of students, and in what contexts? Researchers involved with the Open Learning Initiative at Carnegie-Mellon University have a similar vision of interlocking feedback systems that is guiding their work in developing online courses (Bajzek et al. 2008) (see Case Study 2 sidebar).

Case Study 1. Fine-grained Data Collection and Use: ASSISTments

Fine-grained student data can be structured into meaningful chunks to provide evidence of student problem-solving sequences, knowledge state, and strategy. An example of this use of fine-grained data that is in wide-scale use is the ASSISTments tutoring system, currently used by more than 20,000 students in the New England area. Designed by researchers at Worcester Polytechnic Institute and Carnegie Mellon University, ASSISTments combines online learning assistance and assessment activities. ASSISTments tutors students on middle school math concepts while they practice on problems and provides educators with a detailed assessment of students' developing math skills.

When students respond to ASSISTments problems, they receive hints and tutoring to the extent they need them. At the same time, ASSISTments uses information on how individual students respond to the problems and how much support they need from the system to generate correct responses as assessment information. Each week, when students work on ASSISTments, it learns more about their abilities and thus can provide increasingly appropriate tutoring for each student and can generate increasingly accurate predictions of how well the students will do on the end-of-year standardized test. In fact, the ASSISTments system, taking into account information on the quantity and quality of help that students request, has been found to be more accurate at predicting students' performance on the state examination than the number of items students get correct on benchmark assessments (Feng, Heffernan, and Koedinger 2009).

The ASSISTments system gives educators detailed reports of students' mastery of over 100 middle school math skills, as well as their accuracy, speed, help-seeking behavior, and number of problem-solving attempts. The system can identify the difficulties that individual students are having and the weaknesses demonstrated by the class as a whole so that educators can tailor the focus of their upcoming instruction.



Case Study 2. Meshing Learning and Assessment in Online and Blended Instruction

The online learning systems being developed through the Open Learning Initiative (OLI) at Carnegie Mellon University illustrate the new advances that allow integration of learning and assessment systems. The OLI team set out to design learning systems incorporating the learning science principle of providing practice with feedback. In the OLI courses, feedback mechanisms are woven into a wide variety of activities. A biology course, for example, has the following components:

- Interactive simulations of biological processes that students can manipulate; the student's interaction with the simulation is interspersed with probes to gauge his or her understanding of how it works
- *Did I Get This?* quizzes after presentation of new material so that students can check for themselves whether or not they understood, without any risk of hurting their course grade
- Short essay questions embedded throughout the course material that call on students to make connections across concepts
- *Muddiest Point* requests that ask students what they thought was confusing.

Tutored problem solving gives students a chance to work through complex problems and get scaffolds and hints to help them. The students receive feedback on their solution success after doing each problem, and the system keeps track of how much assistance students needed for each problem as well as whether or not they successfully solved it.

When OLI courses are implemented in a blended instruction mode that combines online and classroom learning, the instructors can use the data the learning system collects as students work online to identify the topics students most need help on so that they can plan upcoming classroom activities on those misconceptions and errors (Brown et al. 2006). OLI is now doing R&D on a digital dashboard to give instructors an easy-to-read summary of the online learning data from students taking their course.

OLI has developed learning systems for engineering statics, statistics, causal reasoning, economics, French, logic and proofs, biology, chemistry, physics, and calculus. A study contrasting the performance of students randomly assigned to the OLI statistics course with those in conventional classroom instruction found that the former achieved better learning outcomes in half the time (Lovett, Meyer and Thille 2008).

Educational Data Mining and Learning Analytics Applications

Educational data mining and learning analytics research are beginning to answer increasingly complex questions about what a student knows and whether a student is engaged. For example, questions may concern what a short-term boost in performance in reading a word says about overall learning of that word, and whether gaze-tracking machinery can learn to detect student engagement. Researchers have experimented with new techniques for model building and also with new kinds of learning system data that have shown promise for predicting student outcomes. In previous sections, we presented the research goals and techniques used for educational data mining and learning analytics. In this section, we present broad areas of applications that are found in practice, especially in emerging companies. These application areas were discerned from our review of the published and gray literature and were used to frame our interviews with industry experts. They represent the broad categories in which data mining and analytics can be applied to online activity especially as it relates to learning online. This is in contrast to the more general areas for big data use, such as health care, manufacturing, and retail (see Manyika et. al 2011).

These application areas are variations of user modeling (knowledge, behavior, experience and profiling), domain modeling (key concepts, knowledge components), and trend analysis for time and sequence analysis. Each of them uses different sources of data to answer questions. Exhibit 2 lists data sources that have been used thus far. We describe each area below, using examples from industry practice (developers) and academic research (researchers).

Exhibit 2
Application Areas and Types of Data Used

Application Area	Description of Category	Type of Data Needed for Analysis
User modeling	Knowledge (what content a user knows, e.g., procedural knowledge, higher order thinking)	...the interaction between the student and the system such as correctness of responses, time spent before making an attempt, hints requested, repetitions of wrong answers, and errors made; system's feedback ...the skills that were practiced: what skills and previous opportunities
User modeling	Behavior, motivation	...the interaction between the student and the system such as correctness of responses, time spent before making an attempt, hints requested, repetitions of wrong answers, and errors made' system's feedback ...the changes in the classroom/school context during the investigation period of time Student's performance level on domain knowledge inferred from system work or collected from other sources such as standardized test
User modeling	Experience, satisfaction	...response to surveys or questionnaires ...choice, behavior, or performance in subsequent learning unit/courses
User profiling	What groups do users cluster into?	...the interaction between the student and the system
Domain modeling	How is content decomposed into components and sequenced?	...response (including correctness) and actions on atomic learning objects over time A domain model taxonomy Association structure among problems, and with skills
Learning component analysis; instructional principle analysis; curriculum evaluation	Which components are effective at promoting learning? What learning principles work well? How effective are whole curricula?	...responses (including correctness) and actions on atomic learning objects over time A domain model taxonomy Association structure among problems and with skills
Trend analysis	What changes over time and how?	Varies depending on what information is of interest; typically would need at least 3 data points longitudinally to be able to discern a trend Data collected are enrollment records, degrees, completion, student source, and high school data in consecutive years
Recommendation	What next actions can be suggested for the user?	Varies depending on the actual recommendation given May need to collect historical data about the user and also related information on the product or service to be recommended Need student's academic performance record.
Improvement	How should the user experience be changed for the next user?	
Adaptation	How can the user experience be altered, most often in real time?	

User Knowledge Modeling

Tailoring Learner Feedback

Adaptive learning systems can provide tailored feedback that gives guidance based on analysis of fine-grained data. Knewton's Math Readiness system uses analytics to deliver only the content each student needs, and skips concepts the student has already shown he or she understands.

Researchers and developers build and tune user models that represent a collection of user-specific data, especially skills and knowledge. User models are used to customize and adapt the system's behaviors to users' specific needs so that the systems "say" the "right" thing at the "right" time in the "right" way (Gerhard 2001). Inferring what a user knows (i.e., user knowledge modeling) requires looking at accumulated data that represents the interactions between students and the learning system.[2] Knowledge can be inferred from interactions such as correctness of student responses alone or in a series, time spent on practice before attempting to answer a question, number and nature of hints requested, repetitions of wrong answers, and errors made. Such "inferences" can be made by a predictive computer model or by a teacher looking at student data on a dashboard.

Advancing Instruction

Many learning technology experts are enthusiastic about the possibility of data completely driving the student's experience. By tracking a student's mastery of each skill, a learning system can give just the right amount of instruction. Other experts caution against allowing analytics to completely determine what problems or skills students practice next or whether they advance to the next topic. Automatically holding a student back on the assumption that difficulty with one topic will preclude making progress on another may not be the best course of action (Means, Chelemer, and Knapp 1991).

User knowledge modeling has been adopted to build adaptive hypermedia, recommender systems, expert systems, and intelligent tutoring systems. In intelligent tutoring systems, user knowledge models direct key operations such as deciding which problems to give students. A popular method for estimating students' knowledge is Corbett and Anderson's knowledge tracing model (Corbett and Anderson 1994), an approach that uses a Bayesian-network-based model for estimating the probability that a student knows a skill based on observations of him or her attempting to perform the skill. More recently, Baker and colleagues proposed a new method for knowledge tracing using a machine learning approach to make contextual estimations of the probability that a student has guessed or slipped. Incorporating models of guessing and slipping into predictions of student future performance was shown to increase the accuracy of the predictions by up to 48 percent (Baker, Corbett, and Aleven 2008).

[2] Even though one could envision that continuous knowledge modeling could supplant traditional assessments, our technical working group still saw a need for end-of-course or state-level assessments as a check on this more local and possibly more formative type of assessment.

Student knowledge modeling is a common component of commercial learning software. How these models are used to adapt instruction varies, however. For example, one company builds dynamic student models for determining a student's readiness to move to new learning content and then advancing the student automatically. Other companies resist automatic advancement, and instead their systems offer suggestions to teachers after detecting a student's placement. Other companies are trying a fully automated approach: If students are performing above average, they receive suggestions to move on to new content; otherwise, they are encouraged to consolidate current skills and work on prerequisites. As an example, learning software can collect such data as how many minutes are spent on a unit, how many hints were used, and common errors. The data for an individual student can then be compared against a model built from a large number of students. Our expert from Agile Mind, a learning software company, explained that this analysis enables teachers to distinguish between students who are not trying and those who are trying but still struggling. This information then helps teachers use different instructional strategies for these two groups of students. Agile Mind, however, cautions against allowing the data to drive what a student sees next or allowing the data to prevent a student from advancing because, according to the data, he or she has not achieved "mastery." Not enough is known about the dependencies among topics to make these decisions in a completely automated manner.

In contrast, the Onsophic Inc. online learning platform collects data at a very granular level (per topic) for each student and detects student mastery at this topic level (e.g., quadratic equation) rather than at the course level. Plans are to provide detailed feedback, such as, "A week ago, you were 'yellow' on a prerequisite but now you are struggling on this topic. We suggest that you make sure you have a solid foundation on this topic through practicing on the prerequisite."

User Behavior Modeling

User behavior modeling in education often characterizes student actions as on or off task and can be used as a proxy for student engagement. It relies on the same kinds of learning data used in predicting user knowledge plus other measures such as how much time a student has spent online (or on the system), whether a student has completed a course, documented changes in the classroom or school context, and sometimes a student's level of knowledge as inferred from his or her work with the learning system or from other data sources, such as standardized test scores, attendance, and tardiness. Baker and colleagues have conducted a series of studies on detecting and adapting to students' off-task behaviors (called gaming the system) in adaptive learning systems that teach algebra (Baker et al. 2004, 2006). They found that gaming behaviors (such as clicking until the system provides a correct answer and advancing within the curriculum by systematically taking advantage of regularities in the software's feedback and help) were strongly associated with less learning for students with below-average academic achievement levels. In response, they modified the system to detect and respond to these students and provide

them with supplementary exercises, which led to considerably better learning. Much of this research has been done with software for teaching mathematics, but similar research has been done in unscripted environments that are more open ended than the well-defined domain of mathematics. For instance, Blikstein (2011) has presented an automated technique and a case study to assess, analyze, and visualize behaviors of students learning computer programming.

Online learning systems log student data that can be mined to detect student behaviors that correlate with learning. Macfayden and Dawson (2010) analyzed learning management system tracking data from a Blackboard Vista-supported course and found variables that correlated with student final grade. Fewer than five variables were found to account for 30 percent of the variation in student final grades, and the model could correctly pick 81 percent of students who failed the course.

Not all learning software companies have adopted user behavior modeling. Those that have collect and provide data to teachers to help them diagnose student learning issues. Carnegie Learning reported that its user behavior modeling was able to detect shifts in the classroom, such as the use of a substitute teacher, a teacher's lack of attention to an online learning system, or a classroom visit by a trainer for the learning system. Social gaming companies, such as Zynga, try to predict what users want and will do next in a game, to find out how to make games more fun and get users more engaged. Others companies, such as Onsophic, Inc. are testing whether capturing on- and off-task behaviors can help them understand online learning through addressing such questions as, Does more interaction between the student and the system lead to increased learning? Do people learn more from items they show interest in? What patterns of interactions are associated with more learning?

User Experience Modeling

User experience modeling—ascertaining whether a student is satisfied with the learning experience—can be judged by students' responses to follow-up surveys or questionnaires and by their choices, behaviors, performance, and retention in subsequent learning units or courses.

User experience modeling has been most popular in Web-based applications such as online shopping. Some of our interviewees' companies model user experience through methods other than data mining. Zynga explicitly asks users for their reactions via a survey, conducts user studies, or has humans conduct postmortem analyses (like Google's researchers who look at failed searches). Zynga described an extended approach to user experience modeling: A sample of users can be surveyed about their experience, and then their behavior can be correlated with their survey results as a way to confirm what they said. Zynga is also experimenting with a more leading-edge approach: analyzing free-text responses given by users in responding to a survey (this is most useful when the sample of users is large, e.g., 250,000 users).

Compared with commercial and industrial applications of user experience modeling, less work has been done in education to use analytics to improve students' learning experience and foster their success and retention rate. Dawson and colleagues (Dawson, Heathcote, and Poole 2010) examined how effective higher education institutions have been in harnessing the data-capture mechanisms from their student information systems, learning management systems, and communication tools for improving student learning experiences and informing practitioners of the achievement of specific learning outcomes. They found that if the multiple means through which students engage with university systems are considered, individual activity can be tracked throughout the entire student life cycle—from initial admission through course progression and finally graduation and employment transitions. The combined data captured by various systems build a detailed picture of the activities that students, instructors, service areas, and the institution as a whole undertake and can be used to improve relevance, efficiency, and effectiveness in a higher education institution.

User experience, as measured by retention, is important for companies offering commercial online courses. Kaplan, Inc. uses retention to judge whether its product is meeting customer needs. Kaplan has experimented with course redesigns using analytics. In one experiment, it collected time spent on redesigned course components, periodic surveys of students' motivation state during the course, and learning performance. Kaplan then looked at instructor satisfaction, student satisfaction, performance on embedded learning assessments, whether the student passed the course, and whether the student was retained until the next semester. Kaplan was able to conduct A/B testing and ascertained that the new course was better overall. But this was only visible via multiple measures: Instructors preferred the redesign; students did better on the assessments, spent more time on the materials, and were more likely to pass and take the next course. Yet students reported liking the old version more.

User Profiling

A user profile is a collection of personal data describing the essential characteristics of a user. User profiling refers to the process of constructing and applying student or group profiles using data mining and machine learning algorithms. Because students differ in their preferences, interests, background, and even goals for learning, the long-term objective of user profiling is often to provide adapted and personalized learning environments for individuals or groups of students to maximize learning effectiveness and efficiency.

Profiling technologies can be applied in a variety of domains and for a variety of purposes. Knowledge about customer behavior and preferences is of great interest to the commercial sector. With profiling technologies, companies can predict the behavior of different types of customers. Marketing strategies such as personalized advertising can then be tailored to the people fitting these types.

In education, data mining techniques such as classification and clustering are often used to categorize (or profile) students based on the kinds of personal learning data described above, on student demographic data, or both. Kardan and Conati (2011) proposed a user modeling framework that relies on interaction logs to identify different types of learners, as well as their characteristic interactions with the learning system. This information would then be used to classify new learners, with the long-term goal of providing adaptive interaction support when behaviors detrimental to learning are detected or learning ways to support engaged behavior. Classification can also group students together into study groups or other joint learning activities.

Gaming companies automatically cluster users into groups using behavioral data and use different strategies with each group to increase engagement and reduce drop-offs in playing. These groups emerge from the data and are often named based on human interpretations of the emergent patterns, for example, casual players, weekenders, social players, big spenders, decorators, and the like. In practice, these user groups may not always be informative or actionable, although groupings based on purchasing habits have proven useful for recommendation services. One of the learning companies we interviewed was hesitant to provide automatic recommendations for students based on profiles, believing that evidence for the effectiveness of such adaptations is not sufficient. Instead, this company has found that concentrating on assignments, concept strands, standards, and students who do or do not have mastery of the concepts in a standard is more fruitful than classifying students into groups based on learner types. In contrast, another company interviewed for this report is working to classify users based on understandings, learning trajectories, motivation, and possibly even cultural background. They are researching how this helps teachers differentiate instruction.

Domain Modeling

A domain model is often created to represent the key concepts of a subject or topic area. The domain model also identifies the relationships among all the key concepts or units of study. Domain modeling in educational data mining and learning analytics studies how learning is affected by differences in how a topic is decomposed into key concepts at a particular grain size, and differences in presenting sequences of related concepts. For domain modeling, the required data are a taxonomy of the domain, associations among skills, such as prerequisites, user responses (including correctness), and actions over time on individual learning resources (such as a unit concept like multiplication of whole numbers).

Domain modeling has been adopted as an approach to fine-tune learning systems to better serve learning and instruction. For instance, Martin et al. (2011) described three studies to demonstrate how learning curves can be used to drive changes in the user model for personalized learning environments. Learning curves for subsets of the domain model were shown to yield insight into the appropriateness of the model's structure and granularity. They also used learning curves to analyze large amounts of user data to fine-tune a system's domain model.

In the education industry, some learning software companies have the goal of collecting data on "atomic learning objects" (i.e., objects that teach one concept that cannot be decomposed) and creating linking relationships among topics based on user tags or other actions. They intend to pair this technique with a feature that enables users to improve on any automatically built relationships or to create their own taxonomies.

Learning System Components and Instructional Principle Analysis

The type of analysis examines components of a learning system and types of instructional practices adopted at various time points or for various student groups to address such questions as

- Which learning components are effective at promoting learning?
- Does a newly developed curriculum enable more learning than an alternative?
- What types of instructional practice are more effective in promoting learning (e.g., massed practice vs. spaced practice)?

Answering these questions entails collecting data such as student input and response correctness, student actions on learning system components over time, when and to which group a specific instructional strategy was applied, and students' performance on pre/posttests and/or delayed tests or their standardized test results.

Because studying the effectiveness of different learning system components and instructional practices can contribute to the design of better learning systems and has strong implications for student learning, it has been a key area of interest for educational data mining and analytics researchers, as evidenced by widely cited papers in using educational data mining to study and improve online courses (Baker and Yacef 2009). For example, researchers and educators from Carnegie Learning, Inc. and Carnegie Mellon University have been working to build cognitive models of mathematics, which have become the basis for middle-school and high-school curricula incorporating the Cognitive Tutor, an intelligent tutoring system. In these systems, complex tasks are decomposed into individual knowledge components, and a model is used to follow students' actions as a basis for diagnosing their strategy in solving a problem. Each action that the student takes is associated with one or more skills. Therefore, researchers have been able to use Cognitive Tutor data to dynamically evaluate the effectiveness of instruction at a more detailed level. Evaluations and improvements have been conducted over the past 15 years (Ritter, et al. 2007). To discover which pedagogical support is most effective, Beck and Mostow (2008) proposed learning decomposition an alternative to traditional A/B testing methods.

As a type of relationship mining, learning decomposition involves fitting exponential learning curves to performance data and relating student success to the amount of each type of pedagogical support a student has received (with a weight for each type of support). The weights indicate how effective each type of pedagogical support is for improving learning. Another company uses data from many teachers to identify the pedagogical patterns of effective teachers, i.e., teachers whose students learn the most or are most engaged. The company is training other teachers in the same techniques and studying what happens in the learning system when these other teachers adopt those patterns.

Trend Analysis

Trend analysis in general refers to the practice of collecting information and attempting to spot a sequential pattern, or trend, in the information over time. Web-based companies use trend analysis to predict what users might be searching for, or interested in, or how user participation ramps up or falls off. In education, trend analysis helps answer such questions as what changes have occurred in student learning over time and how learning has changed. At the school level, trend analysis can be used to examine test scores and other student indicators over time to help administrators determine the impact of policies. In educational data mining, trend analysis often refers to techniques for extracting an underlying pattern, which might be partly or nearly completely hidden by data that does not contribute to the pattern (i.e., noise). Although the actual data needed for trend analysis vary depending on what information is of interest, typically longitudinal data from at least three points in time are required.

As an example of trend analysis, the Postsecondary Education Commission of California provides a trend analysis tool online at http://www.cpec.ca.gov/OnLineData/Mining.asp. This tool can be used to examine the commission's database tables to identify trends. It can also be used to discover anomalies with the data such as large numerical differences between consecutive years and gaps when no data were reported. Visitors can generate customized reports on enrollment, degree completion, student home school, and high school data.

Adaptation and Personalization

Personalization, as defined in the NETP (U.S. Department of Education, 2010a), indicates adaptive pacing, styling instruction to learning preferences and tailoring content to learners' interest. We use *adaptation* to indicate the changes a system (interface or behavior) or instructor makes in response to students, thereby personalizing their experience. Adaptation and personalization address questions such as, How should the user experience be changed for this user? How can user experience be altered to best serve individual users in real time? User classification techniques and trend/sequence analysis are often applied to create models for adapting instruction to students' needs. These adaptations may include recommendations or feedback to students about their best next actions and changes to their experience with an online learning system (such as different content, more practice, or signals about their progress through a course).

To adapt instruction or personalize student learning experiences, data such as sequences of student activity, information on the problems or steps a user has attempted, and student demographic information are often collected and used to create a personal profile for each system user. Researchers from Austria (Köck and Paramythis 2011) investigated the monitoring and interpretation of sequential learning activities to improve adaptation and personalize educational environments. They analyzed student problem solving data from the Andes tutoring system (VanLehn et al. 2005) by first converting activity sequences in the raw data into chain-like models and then clustering sequences to detect problem-solving styles. These models are used to adapt the tutoring system to students' preferred learning methods.

This section has described a broad set of general applications that exploit educational data mining and learning analytics techniques to adapt and personalize learning and improve teaching. These represent the promise of educational data mining and learning analytics, with the caveat that some are still in the research stage. The next section examines challenges and considerations to bring these techniques into K-12 and higher education.

Implementation Challenges and Considerations

New technology start-ups founded on big data (e.g., Knewton, Desire2Learn) are optimistic about applying data mining and analytics applications—user and domain modeling and trend analysis—to adapt their online learning systems. Companies that "own" personal data (e.g., Yahoo!, Google, LinkedIn, Facebook) have supported open-source developments of big data software (e.g., Apache Foundation's Hadoop, hadoop.apache.org) and encourage collective learning through public gatherings of developers to train them on the use of these tools (called hackdays or hackathons). The big data community is, in general, more tolerant of public trial-and-error efforts as they push data mining and analytics technology to maturity.[3] What is the gap between the big data applications in the commerce, social, and service sectors and K–20 education? The 2012 Horizon Report's short list of projects to watch in higher education shows learning analytics in the 2- to 3-year range for widespread adoption (Horizon Project Shortlist 2012). Given that learning analytics practices have been applied primarily in higher education thus far, the time to full adoption may be longer in different educational settings such as K–12 institutions.

This section describes the challenges in implementing data mining and learning analytics within K–20 settings. Experts pose a range of implementation considerations and potential barriers to adopting educational data mining and learning analytics, including technical challenges, institutional capacity, and ethical issues. Successful application of educational data mining and learning analytics will not come without effort, cost, and a change in educational culture to more frequent use of data to make decisions (U.S. Department of Education 2010b).

[3] As an example, consider the contrasting cases we described for user profiling. One learning company believed it was ineffective, while another was willing to experiment with it as a differentiator for its company.

Technical Challenges

Online learning technologies offer researchers and developers opportunities for creating personalized learning environments based on large datasets that can be analyzed to support continuous improvement. However, these benefits depend on managing all the data that can now be captured in real time across many students. A challenge for successful implementation of educational data mining and learning analytics techniques is having sufficient technical resources for using big data and incurring the expenses associated with software services and storage in either remote servers provided by a company or local servers. Although data mining and analytics are used in some courses and institutions, computer scientists are still working on reducing the computer memory requirements needed to support advanced algorithms, and some are not optimistic about the near-term resolution of this issue.

In response to the big data challenge, a few key issues should be considered when implementing data mining and analytics. These include choosing what data collect, focusing on the questions to be answered, and making sure that the data align with the questions. This includes being strategic about what data to collect and carefully studying the analytic techniques needed to answer the most pressing questions. One expert stressed the importance of starting out by understanding what questions data mining and analytics can answer: "If you have 100 people working, I would allocate 99 for identifying what questions to answer and one for [the technical process of] data mining."

Lack of data interoperability[4] among different data systems imposes a challenge to data mining and analytics that rely on diverse and distributed data. Over time, piecemeal purchases of software can lead to increased decentralization of the source of education data, such as student information systems, teachers' online grade books, homework submission systems, and publishers' online assignments, homework help, and assessments. Systems provided by commercial vendors may not give the data they collect back to the teacher, district, or parents, so understanding who will own the data is an important consideration.

Lack of consistency among government data systems requires data translation and alignment. For example, suppose a state wished to mine data for the factors that influence high school graduation and dropouts. Graduates and dropouts may be defined differently across districts' data systems, and potential different factors either may not be collected (e.g., year a student was retained in the same grade) in many districts or defined in the same way across districts (e.g., codes for misbehavior). If data could be aligned, techniques for handling *hierarchical* data are

[4] Data interoperability refers to a property of a system whose input/output data flow and formats are completely understood by others so that data from such systems can be integrated or exchanged seamlessly for analysis.

still under development. Such inconsistencies make it difficult to conduct comprehensive mining and analytics that involve correlating numerous variables.

Researchers in educational data mining and learning analytics seek to make claims about student learning of topics or concepts based on a student's interaction with an online learning system. These claims can be validated by comparing scores on assessments and course grades. Going beyond one dataset to combining multiple sources of data (e.g., multiple tests, both teacher-made and standardized; behavioral assessments; or online behavior tracking) in order to provide an integrated view of a student's progress is not a straightforward task. Existing datasets may not have been designed to support creating profiles of student behaviors and, for example, may leave out data that could be an important variable in a model. Combining disparate data sources to make claims about student learning is known to be fraught with difficulties in assessment and, when used for high-stakes actions, must meet appropriate standards for valid student assessment.

Limitations in Institutional Capacity

Technical challenges can be overcome through research, development, and testing, and computing and storage can be budgeted as part of an institution's infrastructure costs. However, implementing data mining and learning analytics in K–20 institutions have costs that go beyond simply computing and storage. Significant human resources also are needed for data preparation, processing, and analysis. Integration of existing data systems such as grade books and student information systems can be expensive, and the requirements can exceed the capabilities of the information technology department of a single institution. Our experts reported that at least 70 percent and often 80 to 85 percent of the effort in data analytics is devoted to data cleaning, formatting, and alignment and suggested that education has the further complication of needing to move data across different levels of the system, back and forth between classroom, school, district, and state databases.

If technical challenges can be overcome and data can be prepared and analyzed, smart consumers are needed to use the data. Today, teachers and school leaders are surrounded by many data reports and often are frustrated by how much work is required to sort the useful from the useless. Data dashboards need to be adapted to everyday users. Successful adoption and implementation of data mining and analytics in classrooms, schools, districts, and other institutions require that education researchers and software developers obtain a good understanding of the challenges from the users' perspective. This will enable them to pose questions that matter to teachers and other users and to frame findings in a thoughtful, informative way that highlights and recommends clear actions.

In reports about the newest technologies for adaptation, personalization, and recommendation, the role of human judgment is sometimes underemphasized (with the exception of visual data analytics). All the experts we consulted emphasized the key role that people play in many steps of the

Open Research Questions

What is the right amount of data to collect?

Experts from the learning analytics field tend to favor a top-down approach: Meaningful questions should be posed to drive the data collection and analysis. They advocate a targeted strategy of collecting the right data in the right form at the outset. In contrast, data mining researchers favor a bottom-up approach supported by a more inclusive data collection strategy. They believe that collecting more data allows for exploratory data mining approaches in which a main question drives analysis, but the large amount of data collected supports finding unexpected patterns. Solutions from commercial companies have also shown promise in a middle ground, such as collecting dense usage data from a randomly selected sample of users to inform product improvement.

What is the right data structure?

Given the heterogeneous (many data sources) and hierarchical (multiple levels) nature of educational data, determining data structures and data formats that accurately represent an event under consideration become key. A basic data format may be a "learning transaction" generated by the system, the student, or the interactions between the two. The best data structure and analytic techniques are determined by the types of problem to be solved. Answering a focused question takes extensive data cleaning and extraction, and it is very important to have the best analytic algorithm. Pattern-seeking approaches such as outlier detection (e.g., to detect atypical student behavior such as novice mistakes or irregular learning), on the other hand, require less data cleaning and can employ a coarser algorithm.

data mining and analytics process. Smart data consumers can help determine what questions to address, what data to collect, and how to make reports meaningful and actionable. They can even help interpret data, discern and label patterns, and guide model building. Data mining and analytics technology play a supporting role in the essentially human and social effort of making meaning out of experience. One expert stressed that data mining and analytics do not give answers when just unleashed on a big data warehouse. Instead, the recommendation was to approach the problem in an informed way, considering what can be acted on, what evidence can came from data analysis, and what early pilots of the data mining and analytics applications reveal.

Smart data consumers must learn to keep an open mind to what the data say. Data mining and analytics techniques can confirm or disconfirm teachers' and students' beliefs about student knowledge, abilities, and effort. Sometimes, these beliefs are not consistent with the data: Teachers may believe particular students are more or less capable than they are, and students may report spending more time and effort on learning than they actually do. For example, one company found in an A/B study it conducted on the use of visualizations that students were more engaged when complex visualizations were included in the software. Students identified complexity as a source of their engagement, but teachers thought the visualizations were too complex, underestimating what the students were capable of understanding.

Privacy and Ethics Issues

It has been acknowledged for many years (e.g., Kobsa 1990) that personalized interaction and user modeling have significant privacy implications because personal information about users needs to be collected to customize software to individuals. Press coverage and recent Federal Trade Commission rulings have highlighted online companies' privacy protection lapses. Data mining researchers have exposed obvious weaknesses, e.g., querying a social network for registered e-mail addresses on a large scale (Balduzzi et al. 2010).[5] Consumer surveys (ChoiceStream 2005) often show that online users value personalized content but are also concerned about their privacy on the Internet. At the same time, privacy versus personalization is not a simple trade-off: A more complete set of factors includes personal and community attitudes, how far the disclosed information differs from the norm, and even how much users know about what was disclosed and how much control they have over it (Kobsa, 2007).

[5] Starting with a list of about 10.4 million e-mail addresses, Balduzzi and colleagues were able to automatically identify more than 1.2 million user profiles associated with the addresses. By searching through these profiles, they collected publicly available personal information about each user. After being exposed, this vulnerability was repaired.

Whereas educational institutions have always had requirements to protect student and teacher privacy, methods for collecting sufficient data for predictive models while preventing inadvertent identity disclosure in data mining and analytics are not widespread. Experts we spoke to believed that both formal and informal controls are needed: formal controls through legislation and informal controls through a code of ethics voluntarily adopted by organizations that use data. Recent guidance on the *Family Educational Rights and Privacy Act* (FERPA) has helped clarify the rights of institutions to use detailed and longitudinal student data for research, accountability, and school improvement. New amendments to the existing regulations increase access to data for research and evaluation (including sharing across levels, such as from high school to college) while maintaining student privacy and parents' rights (Data Quality Campaign, 2011).

Educational data mining and learning analytics make predictions and recommend actions based on increased visibility into student actions, and these give rise to a number of social and ethical concerns. Experts cited the ethical obligation to act on the knowledge about students gained through data mining. Educational data analysts should share their insights with those who can benefit from them (for example, students, teachers, and school districts) and what is shared must be framed in a way that benefits rather than harms. For example, is it useful to share with a particular student that he has only a 20 percent chance of success in a course given his past performance? What is the impact of this finding on the classroom and on the teacher's practices? What will happen to the student-teacher relationship once such results are released?

Policymakers bear an ethical responsibility to investigate the validity of any predictive model that is used to make consequential decisions about students. Policymakers must be able to explain the evidence for predictions and the actions taken by the computer system on the basis of learning analytics. Analysts conducting data mining may discover patterns or associations that were previously unknown and that involve sensitive information (e.g., teacher performance or student's family situation), and validating them with external observations and further data collection may be needed.

Recommendations

Educational institutions pioneering the use of data mining and learning analytics are starting to see a payoff in improved learning and student retention (Koedinger, Baker, et al. 2010). As described by an Institute of Education Sciences practice guide (Hamilton et al. 2009), working from student data can help educators both track academic progress and understand which instructional practices are effective. The guide describes also how students can examine their own assessment data to identify their strengths and weaknesses and set learning goals for themselves. Recommendations from this guide are that K–12 schools should have a clear strategy for developing a data-driven culture and a concentrated focus on building the infrastructure required to aggregate and visualize data trends in timely and meaningful ways. The vision that data can be used by educators to drive instructional improvement and by students to help monitor their own learning is not new (e.g., Wayman, 2005). However, the feasibility of implementing a data-driven approach to learning is greater with the more detailed learning microdata generated when students learn online, with newly available tools for data mining and analytics, with more awareness of how these data and tools can be used for product improvement and in commercial applications, and with growing evidence of their practical application and utility in K–12 and higher education. There is also substantial evidence of effectiveness in other areas such as energy and health care (Manyika et al. 2011).

Internet businesses—both providers of general commodities and services and learning software companies—have discovered the power of using data for rapid improvement of their practices through experimentation and measurement of change that is *understandable* and that leads to *actionable* next steps. The key for data analysis consumers, such as students, parents, teachers, and administrators, is that the data are presented in such a way that they are clearly answering a question being asked and point toward an action that is within the data consumer's repertoire.

Specific recommendations for educators and researchers and developers are presented in the remainder of this section.

Educators

Stakeholders in the K–12 and higher education sectors should increase the use of educational data mining and learning analytics to improve student learning. Our recommendations to facilitate adoption are as follows:

Educators should develop a culture of using data for making instructional decisions. This brief builds on the recommendations of the U.S. Department of Education (2009) report calling for development of the mind-set that using data can drive school improvement. Educators need to experience having student data that tells them something useful and actionable about teaching and learning. This means that instructors must have near-real-time access to easy-to-understand visual representations of student learning data at a level of detail that can inform their instructional decisions. Scores on an achievement test taken 6 months ago do not tell a teacher how to help a particular student tomorrow. The kinds of data provided to instructors need to be truly helpful in making instructional decisions, and instructors will need to come to these learning data with a different mind-set than that engendered by data systems geared to serving purposes of accountability.

Districts and institutions of higher education need to understand that their information technology department is part of the effort to improve instruction but is not the only responsible department. Establishing a data-driven culture requires much more than simply buying a computer system. District staff from the information technology department need to join with assessment, curriculum, and instruction staff as well as top decision makers and work together to iteratively develop and improve data collection, processing, analysis, and dissemination. A U.S. Department of Education report (Hamilton, et al. 2009) suggests that districts foster a culture of using data by beginning with such questions as, Which instructional materials or approaches have been most effective in promoting student learning of this area of math content? Are there differences in course success rates for students coming in to our high schools from different feeder schools? Are there teachers who are particularly successful in terms of their students' learning gains whose practice might serve as a model for others?

Understand all details of a proposed solution. When purchasing learning software or learning management systems, districts should demand details about the kinds of learning analytics the system will generate and make sure the system will provide teachers and school leaders with information they can use to improve teaching and learning. What are the analytics based on? Have these measures been validated? Who gets to see the analytic data and in what format, and what do they have to do to gain access? If students, teachers, and district administrators will use visualizations or other reports from a data mining or an analytics package, they should evaluate the solution to make sure the data are presented in a comprehensible way. Give teachers the opportunity to ask questions about data mining and analytics that go beyond the numbers, colors, or charts and instead probe the value that the analytics system will bring to them and the steps

they can take in response to the data the system will give them. Any predictive models proposed for consequential use (such as assigning students to services or qualifying them for advanced courses) should be transparent and backed up by solid empirical evidence based on data from similar institutions.

Start small and leverage the work of others. It can be tempting to latch on to a solution that promises to integrate all data systems to support powerful learning analytics. But the experience of districts pioneering the use of data-driven decision making suggests that there are no easy turnkey solutions (Hamilton et al. 2009). Districts and higher education institutions typically have much more data than they actually use to inform their actions. Part of the problem is that data reside in multiple systems in different formats. The development of standards for educational information systems, software to facilitate integration of data from multiple systems, and models of putting a lightweight data dashboard on top of different data systems are all active areas of technology development. At the present time, however, districts typically incur significant costs when trying to integrate data across different systems. In addition to technology and user interface development costs are the costs involved in developing staff capacity for using data in smart ways. Adoption should be conceptualized as a set of processes and ongoing investments rather than a one-time purchase of a single product or technology. Data mining and analytics can be done on a small scale. In fact, starting with a small-scale application can be a strategy for building a receptive culture for data use and continuous improvement that can prepare a district to make the best use of more powerful, economical systems as they become available. Starting small can mean looking at data from assessments embedded in low-cost or open learning systems and correlating those data with student grades and achievement test scores. Open educational software systems that provide analytics are listed in the Resources section at the end of this report.

Help students and parents understand the source and usefulness of learning data. As colleges and schools move toward the use of fine-grained data from learning systems and student data aggregated from multiple sources, they need to help students understand where the data come from, how the data are used by learning systems, and how they can use the data to inform their own choices and actions. Feedback is an important variable in changing behavior, and research on systems like Purdue's Signals suggests that many students will respond appropriately in the face of feedback that they understand. Similarly, parents can help their children make smarter choices if they have access to student data and understand how the data are generated and what they mean.

Researchers and Developers

R&D in educational data mining and learning analytics occurs in both academic and commercial organizations. Research and development are tightly linked, as the field seeks to understand basic processes of data interpretation, decision making, and learning and to use those insights to develop better systems. We encourage the R&D community to consider these recommendations.

Conduct research on the usability and impact of alternative ways of presenting fine-grained learning data to instructors, students, and parents. Data visualizations provide an important bridge between technology systems and data analytics, and determining how to design visualizations that practitioners can easily interpret is an active area of research. Solving this problem will require identifying the kinds of choices or decisions that teachers, students, and parents want to make with fine-grained learning data and the time pressure and cognitive load factors present when different kinds of decisions are made.

Develop decision supports and recommendation engines that minimize the extent to which instructors need to actively analyze data. The teacher in a truly instrumented classroom would have much more than access to student scores on state and district tests. Diagnostic real-time assessment tools and decision support systems would enable the instructor to work with automated systems to make decisions "on the fly" to improve instruction for all students (Crawford, Schlager, Penuel, and Toyama 2008). But conscious labor-intensive processing of data is not possible under the time constraints of efficient classroom management. To support teachers in the act of instruction, we need decision supports and recommendation systems that link student learning profiles to recommended instructional actions and learning resources. We give such tools to physicians and military decision makers; education is no less complex and no less important.

Continue to perfect the anonymization of data and tools for data aggregation and disaggregation that protect individual privacy yet ensure advancements in the use of educational data. Recent modifications to FERPA guidance has removed ambiguities that made some states and school districts reluctant to make student data available for research purposes even when information that could be used to identify individual students was removed. Much remains to be done, however, in figuring out how to support aggregation and disaggregation of student data at different levels of the education system (classroom, school, district, state) in ways that make it possible to combine data from different sources yet protect student privacy. Alternative approaches to solving this problem should be tried and compared, seeking further clarifying FERPA guidance for each case.

Develop models for how learning analytics and recommendation systems developed in one context can be adapted and repurposed efficiently for other contexts. Differences in educational contexts have made it a challenge to transfer developed predictive models across

educational settings. Because students, administrative policies, course programs (e.g., 4-year vs. community colleges), and/or adopted learning systems often vary among institutions, student learning data that can be collected changes too. Thus, a model developed for one institution usually cannot be applied directly and efficiently to another without research into whether it must be changed for the new context (Lauría and Baron 2011). Understanding how this process can become more efficient will be key to scaling up the use of learning analytics.

Collaborations Across Sectors

As noted above, building the capacity of educational organizations to use data mining and analytics meaningfully is a major undertaking. This section addresses R&D collaborations that can aid the process. Our advisors recommended collaboration among learning system designers (often commercial entities), learning scientists, and educators. Learning product designers want access to the knowledge base built by academic researchers. Policymakers want findings about student learning and clear-cut guidelines for practice (e.g., O'Neil 2005). As we move from print to digital classrooms, learning products will change rapidly, and academic institutions and policies must respond accordingly. We anticipate that the next 5 years will bring an increase in models for collaboration between learning system designers, researchers, and educators. Possibilities for such collaborations include the following:

- Learning labs where commercial designers can make data from their learning systems available to the research community, as is being done through the Pittsburgh Science of Learning Center's Data Shop (Koedinger, Baker, et al. 2010)
- Partnerships between research organizations and education organizations to improve research-based products. For example, the Strategic Education Research Partnership (SERP) is an organization that stimulates innovation in education through sustained collaboration among distinguished researchers, educators, and designers. Under SERP, researchers built a set of in-depth partnerships with large school systems and developed tools and interventions in Boston and San Francisco to help middle and high school teachers, particularly those in science, social studies, and other content areas, incorporate academic vocabulary into their teaching.
- Organizational structures that bring together people with the requisite expertise from industry, academia, and school systems in a sustained interaction to improve learning systems. The newly funded Digital Promise (http://www.digitalpromise.org/) has the goal of fostering sustained investments in such partnerships, which are much more likely to have an impact than simply publishing research and expecting that the commercial sector will incorporate it into products.

Conclusion

Working with big data using data mining and analytics is rapidly becoming common in the commercial sector. Tools and techniques once confined to research laboratories are being adopted by forward-looking industries, most notably those serving end users through online systems. Higher education institutions are applying learning analytics to improve the services they provide and to improve visible and measurable targets such as grades and retention. K–12 schools and school districts are starting to adopt such institution-level analyses for detecting areas for improvement, setting policies, and measuring results.

Now, with advances in adaptive learning systems, possibilities exist to harness the power of feedback loops at the level of individual teachers and students. Measuring and making visible students' learning and assessment activities open up the possibility for students to develop skills in monitoring their own learning and to see directly how their effort improves their success. Teachers gain views into students' performance that help them adapt their teaching or initiate interventions in the form of tutoring, tailored assignments, and the like. Adaptive learning systems enable educators to quickly see the effectiveness of their adaptations and interventions, providing feedback for continuous improvement.

Open source tools for adaptive learning systems, commercial offerings, and increased understanding of what data reveal are leading to fundamental shifts in teaching and learning systems. As content moves online and mobile devices for interacting with content enable teaching to be always on, educational data mining and learning analytics will enable learning to be always assessed. Educators at all levels will benefit from understanding the possibilities of the developments that we have described in the use of big data.

References

Amershi, S., and C. Conati. 2009. "Combining Unsupervised and Supervised Classification to Build User Models for Exploratory Learning Environments." *Journal of Educational Data Mining* 1 (1).

Arnold, K. E. 2010. "Signals: Applying Academic Analytics. *EDUCAUSE Quarterly* 33 (1). http://www.educause.edu/EDUCAUSE+Quarterly/EDUCAUSEQuarterlyMagazineVolum/SignalsApplyingAcademicAnalyti/199385

Bajzek, D., J. Brooks, W. Jerome, M. Lovett, J. Rinderle, G. Rule, and C. Thille. 2008. "Assessment and Instruction: Two Sides of the Same Coin." In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2008*, edited by G. Richards. Chesapeake, VA: AACE, 560–565.

Baker, R. S. J. d. 2011. "Data Mining for Education." In *International Encyclopedia of Education*, 3rd Ed., edited by B. McGaw, P. Peterson, and E. Baker. Oxford, UK: Elsevier.

Baker, R. S. J. d., A.T. Corbett, and V. Aleven, 2008. "More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing." *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, 406-415.

Baker, R. S. J. d., A.T. Corbett, K. R. Koedinger, and I. Roll. 2006. "Generalizing Detection of Gaming the System Across a Tutoring Curriculum." *Proceedings of the 8th International Conferenceon Intelligent Tutoring Systems*, 402-411.

Baker, R. S., A. T. Corbett, K. R. Koedinger, and A. Z. Wagner. 2004. "Off-Task Behavior in the Cognitive Tutor Classroom: When Students 'Game The System.'" *Proceedings of ACM CHI 2004: Computer-Human Interaction,* 383-390.

Baker, R. S. J. d., S. M. Gowda, and A. T. Corbett. 2011. "Automatically Detecting a Student's Preparation for Future Learning: Help Use Is Key." *Proceedings of the 4th International*

*Conference on Educational Data Mining*179–188. http://users.wpi.edu/~rsbaker/PFL-EDM-2011-v19.pdf

Baker, R. S. J. D., and K. Yacef. 2009. "The State of Educational Data Mining in 2009: A Review and Future Visions." *Journal of Educational Data Mining* 1 (1): 3–17.

Balduzzi, M., C. Platzer, T. Holz, E. Kirda, D. Balzarotti, and C. Kruegel. 2010. *Abusing Social Networks for Automated User Profiling*. Research Report RR-10-233 - EURECOM, Sophia Antipolis; Secure Systems Lab, TU Wien and UCSB.

Beck, J. E., and J. Mostow. 2008. "How Who Should Practice: Using Learning Decomposition to Evaluate the Efficacy of Different Types of Practice for Different Types of Students." *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, 353–362.

Blikstein, P. 2011. "Using Learning Analytics to Assess Students' Behavior in Open-Ended Programming Tasks." *Proceedings of the First International Conference on Learning Analytics and Knowledge*. New York, NY: ACM Press.

Brown, W., M. Lovett, D. Bajzek, and J. Burnette. 2006. "Improving the Feedback Cycle to Improve Learning in Introductory Biology Using the Digital Dashboard." In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2006I,* edited by G. Richards. Chesapeake, VA: AACE, 1030–1035.

ChoiceStream. 2005. *ChoiceStream Personalization Survey: Consumer Trends and Perceptions*. http://www.choicestream.com/pdf/ChoiceStream_PersonalizationSurveyResults2005.pdf

Corbett, A. T., and J. R. Anderson. 1994. "Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge." *User Modeling and User-Adapted Interaction* 4 (4): 253–278. doi: 10.1007/BFO1099821

Crawford, V., M. Schlager, W. R. Penuel, and Y. Toyama. 2008. "Supporting the Art of Teaching in a Data-Rich, High-Performance Learning Environment." In *Data-Driven School Improvement*, edited by E. B. Mandinach and M. Honey. New York, NY: Teachers College Press, 109-129.

Data Quality Campaign. 2011. *U.S. Department of Education Final FERPA Regulations: Advisory and Overview*. Education Counsel, LLC for the Data Quality Campaign on December 2, 2011. http://www.educationcounsel.com/files/FERPA_Analysis_Final.pdf

Dawson, S., L. Heathcote, and G. Poole. 2010. "Harnessing ICT Potential: The Adoption and Analysis of ICT Systems for Enhancing the Student Learning Experience." *International Journal of Educational Management* 24 (2), 116–128.

EDUCAUSE. 2010. *Next Generation Learning Challenges: Learner Analytics Premises*. http://www.educause.edu/Resources/NextGenerationLearningChalleng/215028

Elias, T. 2011. *Learning Analytics: Definitions, Processes and Potential*. http://learninganalytics.net/LearningAnalyticsDefinitionsProcessesPotential.pdf

Feng, M., N. T. Heffernan, and K. R. Koedinger. 2009. "User Modeling and User-Adapted Interaction: Addressing the Assessment Challenge in an Online System That Tutors as It Assesses." *The Journal of Personalization Research* (UMUAI journal) 19 (3): 243–266.

Gerhard, F. 2001. "User Modeling in Human-Computer Interaction." *User Modeling and User-Adapted Interaction* 11: 65–68.

Goldstein, P. J. 2005. *Academic Analytics: The Use of Management Information and Technology in Higher Education.* EDUCAUSE Center for Applied Research. http://net.educause.edu/ir/library/pdf/ECM/ECM0508.pdf

Graf, S., and Kinshuk. In press. "Dynamic Student Modeling of Learning Styles for Advanced Adaptivity in Learning Management Systems." *International Journal of Information Systems and Social Change*.

Hamilton, L., R. Halverson, S. Jackson, E. Mandinach, J. Supovitz, and J. Wayman. 2009. *Using Student Achievement Data to Support Instructional Decision Making* (NCEE 2009-4067). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Horizon Project Shortlist. 2012. *NMC Horizon Project Short List,* 2012 Higher Education Ed. http://www.nmc.org/news/download-horizon-project-2012-higher-ed-short-list.

Jeong, H., and G. Biswas. 2010. "Mining Student Behavior Models in Learning-by-Teaching Environments." In the *Proceedings of the 1st International Conference on Educational Data Mining*, Montréal, Québec, Canada.

Johnson, L., A. Levine, R. Smith, and S. Stone. 2010. *The 2010 Horizon Report.* Austin, TX: The New Media Consortium. http://wp.nmc.org/horizon2010/

Johnson, L., R. Smith, H. Willis, A. Levine, and K. Haywood. 2011. *The 2011 Horizon Report*. Austin, TX: The New Media Consortium. http://net.educause.edu/ir/library/pdf/HR2011.pdf

Kardan, S., and C. Conati. 2011. *A Framework for Capturing Distinguishing User Interaction Behaviours in Novel Interfaces*. International Conference on Educational Data Mining, Eindhoven, the Netherlands.

Kobsa, A. 1990. "User Modeling in Dialog Systems: Potentials and Hazards." *AI & Society* 4 (3): 214–240.

———. 2007. "Privacy-Enhanced Personalization." *Communications of the ACM* 50 (8), 24–33.

Köck, M., and A. Paramythis. 2011. "Activity Sequence Modeling and Dynamic Clustering for Personalized E-Learning. *Journal of User Modeling and User-Adapted Interaction* 21 (1-2), 51–97.

Koedinger, K. R., R. Baker, K. Cunningham, A. Skogsholm, B. Leber, and J. Stamper. 2010. "A Data Repository for the EDM Community: The PSLC DataShop." In. *Handbook of Educational Data Mining*, edited by C. Romero, S. Ventura, M. Pechenizkiy, and R.S.J.d. Baker. Boca Raton, FL: CRC Press, 43–55.

Koedinger, K., E. McLaughlin, and N. Heffernan. 2010. "A Quasi-experimental Evaluation of an On-line Formative Assessment and Tutoring System." *Journal of Educational Computing Research* 4: 489–510.

Lauría, E. J. M., and J. Baron. 2011. *Mining Sakai to Measure Student Performance: Opportunities and Challenges in Academic Analytics*. http://ecc.marist.edu/conf2011/materials/LauriaECC2011-%20Mining%20Sakai%20to%20Measure%20Student%20Performance%20-%20final.pdf

Lovett, M., O. Meyer, and C. Thille. 2008. "The Open Learning Initiative: Measuring the Effectiveness of the OLI Statistics Course in Accelerating Student Learning." *JIME Special Issue: Researching Open Content in Education*.

Macfayden, L. P., and S. Dawson. 2010. "Mining LMS Data to Develop an 'Early Warning' System for Educators: A Proof of Concept." *Computers & Education* 54 (2): 588–599.

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. 2011. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Institute. http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

Martin, B., A. Mitrovic, K. Koedinger, and S. Mathan. 2011. "Evaluating and Improving Adaptive Educational Systems with Learning Curves." *User Modeling and User-Adapted Interaction* 21 (3): 249–283.

Means, B., C. Chelemer, and M. S. Knapp (Eds.). 1991. *Teaching Advanced Skills to at-Risk Students: Views from Research and Practice*. San Francisco, CA: Jossey-Bass.

Merceron, A., and K. Yacef. 2010. "Measuring Correlation of Strong Symmetric Association Rules in Educational Data." In *Handbook of Educational Data Mining*, edited by C. Romero, S. Ventura, M. Pechenizkiy, and R. S. J. d. Baker. Boca Raton, FL: CRC Press, 245–256.

O'Neil, H. F. 2005. *What Works in Distance Learning: Guidelines*. Greenwich CT: Information Age Publishing.

Reese, D. D., R. J. Seward, A. Harrison, L. McFarland, B. Hitt, and B. G. Tabachnick. In press. "The Moment of Learning: Quantitative Analysis of Exemplar Gameplay Supports CyGaMEs Approach to Embedded Assessment." In *Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives*, edited by D. Ifenthaler, D. Eseryel, and X. Ge. New York, NY: Springer.

Ritter, S., J. Anderson, K. Koedinger, and A. Corbett. 2007. "Cognitive Tutor: Applied Research in Mathematics Education." *Psychonomic Bulletin & Review* 14 (2): 249–255.

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

Siemens, G., and Baker, R. S. J. d. 2012. 2012. "Learning Analytics and Educational Data Mining: Towards Communication and Collaboration." LAK12: 2nd International Conference on Learning Analytics & Knowledge, 29 April – 2 May, Vancouver, BC, Canada

Siemens, G., and P. Long. 2011. "Penetrating the Fog: Analytics in Learning and Education." *EDUCAUSE Review* 46 (5).

U.S. Department of Education. 2010a. *National Education Technology Plan*. http://www.ed.gov/technology/ netp-2010.

———. 2010b. *Use of Education Data at the Local Level: From Accountability to Instructional Improvement.* Washington, DC: U.S. Department of Education.

VanLehn, K., C. Lynch, K. Schulze, J. A. Shapiro, R. H. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill. 2005. "The Andes Physics Tutoring System: Lessons Learned." *International Journal of Artificial Intelligence in Education* 15 (3): 147-204.

Viegas, F. B., M. Wattenberg, M. McKeon, F. Van Ham, and J. Kriss. 2008. "Harry Potter and the Meat-Filled Freezer: A Case Study of Spontaneous Usage of Visualization Tools."

Proceedings of the 41st Annual Hawaii International Conference on System Sciences, 159–159.

Wayman, J. C. 2005. "Involving Teachers in Data-Driven Decision Making: Using Computer Data Systems to Support Teacher Inquiry and Reflection." *Journal of Education for Students Placed At Risk 10*(3): 295–308.

Recommended Reading

Coley, T. 2010. *Defining IT's Role in Mission-Critical Retention Initiatives*. EDUCAUSE Quarterly. 33 (4). Presents a method for adoption of a data culture with leadership from institutional Information Technology departments. Gives examples of early indicators, early alerts, and aligning separate data systems and people. http://www.educause.edu/EDUCAUSE+Quarterly/EDUCAUSEQuarterlyMagazineVolum/DefiningITsRoleinMissionCritic/219108

Goetz, T. 2011, July. *Harnessing the Power of Feedback Loops*. *Wired Magazine*. Gives explanations and examples of simple feedback loops to improve human behavior. Stresses real-time feedback. http://www.wired.com/magazine/2011/06/ff_feedbackloop/all/1

Ferguson, R. 2012. *The State Of Learning Analytics in 2012: A Review and Future Challenges*. Technical Report KMI-12-01, Knowledge Media Institute, The Open University, UK. Reviews the last decade of work on learning analytics, including factors that influenced its development, and looks at future challenges. http://kmi.open.ac.uk/publications/techreport/kmi-12-01

Johnson, L., A. Levine, R. Smith, and S. Stone. 2010. *The 2010 Horizon Report*. Austin, TX: The New Media Consortium. Horizon reports identify and describe emerging technologies likely to have an impact on college and university campuses within the next five years. This issue includes visual data analysis as an emerging technology. http://www.nmc.org/pdf/2010-Horizon-Report.pdf

Johnson, L., R. Smith, H. Willis, A. Levine, and K. Haywood. 2011. *The 2011 Horizon Report*. Austin, TX: The New Media Consortium. Horizon reports identify and describe emerging technologies likely to have an impact on college and university campuses within the next five years. This issue includes learning analytics as an emerging technology. http://www.nmc.org/pdf/2011-Horizon-Report.pdf

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. 2011. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. McKinsey Global Institute. Looks at innovation and competitive advantages for industries using big data, including health care, retail, and use of personal location

data. http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

Patil, D.J. 2012, September. *Building data science teams*. @dpatil shares his advice on "what data scientists add to an organization, how they fit in, and how to hire and build effective data science teams." He also presents highlights of how Internet companies use big data. http://radar.oreilly.com/2011/09/building-data-science-teams.html

Romero, C. R., and S. Ventura. 2010. "Educational Data Mining: A Review of the State of the Art." *IEEE Transactions on Systems, Man and Cybernetics*, *Part C: Applications and Reviews* 40 (6): 601-618.
In the introduction, Romero and Ventura describe different types of data mining techniques, both classical and emergent, used for educational tasks by different stakeholders.

Romero, C., S. Ventura, M. Pechenizkiy, and R.S.J.d. Baker (eds.). 2010. *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press.
This book provides a thorough and technical overview of the current state of knowledge in educational data mining. It helps education experts understand what types of questions data mining can address and helps data miners understand what types of questions are important in educational design and educational decision making.

Siemens, G., and Baker, R. S. J. d. 2012. "Learning Analytics and Educational Data Mining: Towards Communication and Collaboration." LAK12: 2nd International Conference on Learning Analytics & Knowledge, 29 April – 2 May, Vancouver, BC, Canada.
This paper presents an updated and nuanced distinction between the fields.

Siemens, G., and P. Long. 2011. *Penetrating the Fog: Analytics in Learning and Education*. *EDUCAUSE Review* 46, 5. Gives a broad discussion of what analytics can do in learning. http://www.educause.edu/EDUCAUSE+Review/EDUCAUSEReviewMagazineVolume46/PenetratingtheFogAnalyticsinLe/235017

Siemens, G. et al. 2011. *Open Learning Analytics: an integrated & modularized platform*. Society for Learning Analytics Research (SoLAR). Concept paper on an open learning analytics architecture that raises the need for openness in learning algorithms so that different school settings (cultural or otherwise) can adjust how content is personalized. http://solaresearch.org/OpenLearningAnalytics.pdf

Selected Websites

Visualization and Data Exploration

http://www-958.ibm.com/software/data/cognos/manyeyes/. ManyEyes lets users explore existing visualized datasets and upload their own for exploration. Users can comment on visualizations or create topic areas for discussion. Visualization types are organized by how they show the data (e.g., "See the parts of a whole" and "See the world") and datasets can be numerical, textual, or spatial.

http://hint.fm/. Data visualization meets art in this site showing work by Fernanda Viégas and Martin Wattenberg.

http://research.uow.edu.au/learningnetworks/seeing/snapp/index.html. SNAPP is a tool for visualizing networks resulting from the posts and replies to discussion forums as a measure of student interactions.

http://www.socialexplorer.com/. Social Explorer is an online tool that allows map- and report-based visualizations of census data and demographic information. Flexible enough for use in sectors ranging from education to journalism.

http://www.tableausoftware.com/products/public. Tableau Software offers a free data visualization tool that is used by companies, individuals, and journalists. Visualizations are stored on the Tableau Public site but are embeddable into blogs or Websites.

Online Learning Systems with Analytics

http://www.assistments.org. The ASSISTments online platform helps teachers write assessments and then see reports on how their students performed. Students can get immediate tutoring while they are being assessed.

http://wayangoutpost.com/. Wayang Outpost is an intelligent tutoring system that helps middle- and high-school students study for standardized tests and adjusts instruction as they progress.

http://oli.web.cmu.edu/openlearning/forstudents/freecourses. The Open Learning Initiative (OLI) offers open and free courses on subjects such as biology, programming, chemistry, and statistics. Both students and instructors get timely and targeted feedback.

http://www.khanacademy.org/. Khan Academy provides a library of videos, worked examples, and practice exercises, organized into knowledge maps, for self-paced learning in many topic areas. Khan Academy keeps track of students' progress and shows at-a-glance displays for students, parents, and educators.