Educational Data Mining and Learning Analytics

Ryan S.J.d. Baker, Teachers College, Columbia University
George Siemens, Athabasca University

## 1. Introduction

During the last decades, the potential of *analytics* and *data mining* —methodologies that extract useful and actionable information from large datasets--has transformed one field of scientific inquiry after another (cf. Summers et al., 1992; Collins et al., 2004). Analytics has become a trend over the last several years, reflected in large numbers of graduate programs promising to make someone a master of analytics, proclamations that analytics skills offer lucrative employment opportunities (Manyika et al., 2011), and airport waiting lounges filled with advertisements from different consultancies promising to significantly increase profits through analytics. When applied to education, these methodologies are referred to as *learning analytics* (LA) and *educational data mining* (EDM). In this chapter, we will focus on the shared similarities as we review both parallel areas, while also noting some important differences.

Using the methodologies we describe in this chapter, one can scan through large datasets to discover patterns that occur in only small numbers of students or only sporadically (cf. Baker et al., 2004; Sabourin et al., 2011); one can investigate how different students choose to use different learning resources and obtain different outcomes (cf. Beck et al., 2008); one can conduct fine-grained analysis of phenomena that occur over long periods of time (such as the move towards disengagement over the years of schooling -- cf. Bowers, 2010); and one can analyze how the design of learning environments may impact variables of interest through the study of large numbers of exemplars (cf. Baker et al., 2009). In the sections that follow, we argue that learning analytics has the potential to substantially increase the sophistication of how the field of learning sciences understands learning, contributing both to theory and practice.

**The emergence of analytics**

Compared to sciences such as physics, biology, and climate science, the learning sciences are relatively late in using analytics. For example, the first journal devoted primarily to analytics in the biological sciences, *Computers in Biology and Medicine*, began publication in 1970. By contrast, the first journal targeted towards analytics in the learning sciences, the *Journal of Educational Data Mining*, began publication in 2009, although it was preceded by a conference series (commencing in 2008), a workshop series (commencing in 2005), and earlier workshops in 2000 and 2004. There are now several venues that promote and publish research in this area -- currently including the *Journal of Educational Data Mining*, the *Journal of Learning Analytics*, the *International Conference on Educational Data Mining*, the *Conference on Learning Analytics and Knowledge* (referred to below as “LAK”), as well as a growing emphasis on research in this area at conferences such as the *International Conference on Artificial Intelligence in Education, ACM Knowledge Discovery in Databases,* the *International Conference of the Learning Sciences,* and the annual meeting of the *American Educational Research Association*.

The use of analytics in education has grown in recent years for four primary reasons: a substantial increase in data quantity, improved data formats, advances in computing, and increased sophistication of tools available for analytics.

*Quantity of Data*

One of the factors leading to the recent emergence of learning analytics is the increasing quantity of analyzable educational data. Considerable quantities of data are now available to scientific researchers through public archives like the Pittsburgh Science of Learning Center DataShop (Koedinger et al., 2010). Mobile, digital, and online technologies are increasingly utilized in many educational contexts. When learners interact with a digital device, data about that interaction can be easily captured or "logged" and made available for subsequent analysis. Papers have recently been published with data from tens of thousands of students. With the continued growth of online learning (Allen and Seamen, 2013) and the use of new technologies for data capture (Choudhury and Pentland, 2003), even greater scope of data capture during learning activities can be expected in the future, particularly as large companies such as Pearson and McGraw-Hill become interested in EDMand  Massive Online Open Courses (MOOCs) and providers such as Coursera, edX, and Udacity generate additional data sets for research (Lin, 2012).

*Data formats*

Baker recalls his first analysis of educational log data; almost two months were needed to transform logged data into a usable form. Today, there are standardized formats for logging specific types of educational data (cf. Koedinger et al., 2010), as well as considerable knowledge about how to effectively log educational data, crystallized both in scientific publications and in more informal knowledge that is disseminated at conferences and in researcher training programs like the Pittsburgh Science of Learning Center Summer School and the Society for Learning Analytics Research open online courses[1].

*Increased processing/computation power*

The increase in attention to analytics is also driven by advances in computation (Mayer, 2009). Smart phones today exceed the computational power of desktop computers from less than a decade ago, and powerful mainframe computers today can accomplish tasks that were impossible only a few years ago.  Increases in computational power support researchers in analyzing large quantities of data, and also help to produce that data, in fields such as healthcare, geology, environmental studies, and sociology.

*Development of Analytics Tools*

Some of the most significant advances have been in supporting the management of large data sets, making it possible to store, organize, and sift through data in ways that make it substantially easier to analyze. Google developed MapReduce to address the substantial challenges of managing data at the scale of the internet (Dean and Ghemawat, 2008), including

[1] https://learn.canvas.net/courses/33

distributing data and data-related applications across networks of computers; previous database models were not capable of managing web-scale data. MapReduce led to the development of Apache Hadoop, now commonly used for data management.

In addition to tools for managing data, an increasing number of tools have emerged that support analyzing it. In recent years, the sophistication and ease of use of tools for analyzing data make it possible for an increasing range of researchers to apply data mining methodology without needing extensive experience in computer programming. Many of these tools are adapted from the business intelligence field, as reflected in the prominence of SAS and IBM tools in education, tools that were first used in the corporate sector for predictive analytics and improving organizational decision making by analyzing large quantities of data and presenting it in a visual or interactive format (particularly valuable for scenario evaluation). In the early 2000s, many analytics tools were technically complex and required users to have advanced programming and statistical knowledge. Now, even previously complex tools such as SAS, RapidMiner, and SPSS are easier to use and allow individuals to conduct analytics with relatively less technical knowledge. Common desktop software, such as Microsoft Excel, has also incorporated significantly improved visualization and analytics features in recent years. Other tools such as Tableau Software are designed to support the use of analytics tools without advanced technical knowledge. As easier-to-use tools emerge, it will make EDM and LA accessible to a larger number of learning sciences researchers.

**Developing Research Communities**

The two research communities we review in this chapter, educational data mining and learning analytics, have adopted complementary different perspectives on the analysis of educational data. Siemens and Baker (2012) noted that the two communities have considerable overlap (in terms of both research and researchers), and that the two communities strongly believe in conducting research that has applications that benefit learners as well as informing and enhancing the learning sciences. They also described some of the differences:

1 Researchers in EDM are more interested in automated methods for discovery within educational data; researchers in LA are more interested in human-led methods for exploring educational data. This difference approximately tracks the relationship between data mining and exploratory data analysis, in the wider scientific literature. Automated methods for discovery can help to achieve the best possible prediction; human-led methods of discovery can result in more interpretable and more understandable models of phenomena.

2 Researchers in EDM emphasize modeling specific constructs and the relationships between them; researchers in LA emphasize a more holistic, systems understanding of constructs. This difference parallels long-standing differences in approaches among learning sciences researchers. EDM research in this fashion more closely ties to theoretical approaches such as Anderson's ACT-R Theory (Anderson & Lebiere, 1998) or the PSLC Theoretical Framework (Koedinger, Corbett, & Perfetti, 2012); LA research more closely ties to theory that attempts to understand systems as wholes or that take a situationalist approach (Greeno, 1998; also see Nathan and Sawyer, this volume, on elemental and systemic approaches in learning

sciences). That said, We believe that researchers from each of these traditions may find methods emerging from each community to be useful.

3 Researchers in EDM look to applications in automated adaptation, such as supporting learners through having educational software identify a need and automatically change to personalize the learner's experience (cf. Arroyo et al., 2007; Baker et al., 2006; Corbett & Anderson, 1995); researchers in LA look to ways to inform and empower instructors and learners, such as informing instructors about ways that specific students are struggling, so that the instructor can contact the learner (cf. Arnold, 2010) . In this case, EDM and LA methods are each suited to both types of use, and the differences in focus are primarily due to the applications that were historically of interest to researchers in each community.

Both EDM and LA have a strong emphasis on connection to theory in the learning sciences and education philosophy. Most researchers that publish at the EDM and LA conferences use theory from the learning sciences and education to guide their choice of analyses, and aim to contribute back to theory with the results of their analyses. The theory-oriented perspective marks a departure of EDM and LA from technical approaches that use data as their sole guiding point (see, for example, Anderson's argument in 2008 that big data will render the scientific method obsolete: "But faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete.").

## 2. Key methods and tools

The methodologies used in EDM and LA have come from a number of sources, but the largest two sources of inspiration for the area have been methods from data mining and analytics in general, and from psychometrics and educational measurement. In many cases, the specific characteristics of educational data have resulted in different methods playing a prominent role in EDM/LA than in data mining in general, or have resulted in adaptations to existing psychometric methods. In this section, we survey some of the key methodologies in the field, and discuss a few examples of how these methodologies have been applied. This review draws on past reviews (cf. Baker & Yacef, 2009; Romero & Ventura, 2010; Ferguson, 2012; Siemens & Baker, 2012), but extends them to incorporate recent developments.

### 2a. Prediction methods

One of the most prominent categories of EDM methods in the review by Baker & Yacef (2009), and continuing to this day, is **prediction**. In prediction, the goal is to develop a model which can infer a single aspect of the data (the **predicted variable**, similar to dependent variables in traditional statistical analysis) from some combination of other aspects of the data (**predictor variables**, similar to independent variables in traditional statistical analysis). Developing a prediction model depends on knowing what the predicted variable is for a small set of data; a model is then created for this small set of data, and statistically validated so that it can be applied at greater scale. For instance, one may collect data on whether 1000 students dropped out of college, develop a prediction model to infer whether a student will drop out of

college, validate it on sub-sets of the 1000 students that were not included when creating the prediction model, and then use the model to make predictions about new students. As such, prediction models are commonly used to either predict future events (cf. Dekker et al., 2009; Feng et al., 2009; Ming & Ming, 2012), or to predict variables that are not feasible to directly collect in real-time – for example, collecting data on affect or engagement in real-time often requires expensive observations or disruptive self-report measures, whereas a prediction model based on student log data can be completely non-intrusive (cf. Baker et al., 2004; D'Mello et al., 2008; Sabourin et al., 2011).

These methods have successfully supported interventions to improve student outcomes. For example, the *Purdue Signals* project used prediction models to identify students who were at risk for dropout in courses and university programs. The use of Purdue Signals resulted in more help-seeking, better course outcomes, and significantly improved retention rates (Arnold, 2010).

Three types of prediction models are common in EDM/LA: **classifiers, regressors, and latent knowledge estimation**. In **classifiers**, the predicted variable can be either a binary (e.g. 0 or 1) or a categorical variable. Some popular classification methods in educational domains include decision trees, random forest, decision rules, step regression, and logistic regression. In **regressors**, the predicted variable is a continuous variable, e.g. a number. The most popular regressor in EDM is linear regression (note that linear regression is not used the same way in EDM/LA as in traditional statistics, despite the identical name).

A third type of prediction model that is important in EDM/LA (which is actually just a special type of classifier) is **latent knowledge estimation**. In latent knowledge estimation, a student's knowledge of specific skills and concepts is assessed by their patterns of correctness on those skills (and occasionally other information as well). The models used in online learning typically differ from the psychometric models used in paper tests or in computer-adaptive testing, because with an interactive learning application, the student's knowledge is continually changing. A wide range of algorithms exist for latent knowledge estimation; the two most popular are currently Bayesian Knowledge Tracing (BKT -- Corbett & Anderson, 1995) and Performance Factors Analysis (PFA -- Pavlik, Cen, & Koedinger, 2009), which have been found to have comparable performance in a number of analyses (see review in Pardos et al., 2011). Knowledge estimation algorithms increasingly underpin intelligent tutoring systems, such as the Cognitive Tutors currently used for Algebra in 6% of U.S. high school classrooms (cf. Koedinger & Corbett, 2006).

### 2b. Structure Discovery

**Structure discovery algorithms** attempt to find structure in the data without an a priori idea of what should be found, a very different goal than in prediction. In prediction, there is a specific variable that the EDM/LA researcher attempts to model; by contrast, there is not a specific variable of interest in structure discovery. Instead, the researcher attempts to determine what structure emerges naturally from the data. Common approaches to structure discovery in

EDM/LA include **clustering**, **factor analysis**, **social network analysis**, and **domain structure discovery**.

In **clustering**, the goal is to find data points that naturally group together, splitting the full data set into a set of clusters. Clustering is particularly useful in cases where the most common categories within the data set are not known in advance. If a set of clusters is well-selected, each data point in a cluster will generally be more similar to the other data points in that cluster than data points in other clusters. Clusters have been used to group students (cf. Beal, Qu, & Lee, 2006) and student actions (cf. Amershi & Conati, 2009). For example, Amershi & Conati (2009) found characteristic patterns in how students use exploratory learning environments, and used this information to identify more and less effective student strategies.

In **factor analysis,** a closely related method, the goal is to find variables that naturally group together, splitting the set of variables (as opposed to the data points) into a set of latent (not directly observable) factors. Factor analysis is frequently used in psychometrics for validating or determining scales. In EDM/LA, factor analysis is used for dimensionality reduction (e.g., reducing the number of variables) for a wide variety of applications. For instance, Baker et al. (2009) used factor analysis to determine which design choices are made in common by the designers of intelligent tutoring systems (for instance, tutor designers tend to use principle-based hints rather than concrete hints in tutor problems that have brief problem scenarios).

In **social network analysis (SNA),** models are developed of the relationships and interactions between individual actors, as well as the patterns that emerge from those relationships and interactions. A simple example of its use is in understanding the differences between effective and ineffective project groups, through visual analysis of the strength of group connections (cf. Kay et al., 2006). SNA is also used to study how students' communication behaviors change over time (cf. Haythornthwaite, 2001), and to study how students' positions in a social network relate to their perception of being part of a learning community (cf. Dawson, 2008). This is valuable information because patterns of interaction and connectivity can indicate prospect of academic success as well as learner sense of engagement in a course (Macfadyen and Dawson, 2010; Suthers and Rosen, 2011).

SNA reveals the structure of interactions, but does not detail the nature of exchanges or the impact of connectedness. Increasingly, network analysis is paired with additional analytics approaches to better understand the patterns observed through network analytics; for example, SNA might be coupled with discourse analysis (see Enyedy and Stevens, this volume; also Buckingham Shum and Ferguson, 2012).

**Domain structure discovery** consists of finding the structure of knowledge in an educational domain (e.g., how specific content maps to specific knowledge components or skills, across students). This could consist of mapping problems in educational software to specific knowledge components, in order to group the problems effectively for latent knowledge estimation and problem selection (cf. Cen, Koedinger, & Junker, 2006), or could consist of mapping test items to skills (cf. Tatsuoka, 1995). Considerable work has recently been applied

to this problem in EDM, for both test data (cf. Barnes, Bitzer, & Vouk, 2005; Desmarais, 2011), and for tracking learning during use of an intelligent tutoring system (Cen, Koedinger, & Junker, 2006).

### 2c. Relationship Mining

In **relationship mining**, the goal is to discover relationships between variables in a data set with a large number of variables. Relationship mining has historically been the most common category of EDM research (Baker & Yacef, 2009), and remains extremely prominent to this day. It may take the form of attempting to find out which variables are most strongly associated with a single variable of particular interest, or may take the form of attempting to discover which relationships between any two variables are strongest. Broadly, there are four types of relationship mining: **association rule mining**, **correlation mining**, **sequential pattern mining**, and **causal data mining**.

In **association rule mining**, the goal is to find if-then rules of the form that if some set of variable values is found, another variable will generally have a specific value. For instance, Ben-Naim and colleagues (2009) used association rule mining to find patterns of successful student performance in an engineering simulation, to make better suggestions to students having difficulty about how they can improve their performance. In **correlation mining**, the goal is to find positive or negative linear correlations between variables (using post-hoc corrections or dimensionality reduction methods when appropriate to avoid finding spurious relationships). An example can be found in Baker et al. (2009), where correlations were computed between a range of features of the design of intelligent tutoring system lessons and students' prevalence of gaming the system (intentionally misusing educational software to proceed without learning the material), for example finding that brief problem scenarios lead to a greater proportion of gaming behavior than either rich scenarios or having no scenario at all (just equations to manipulate). In **sequential pattern mining**, the goal is to find temporal associations between events. One successful use of this approach was work by Perera et al. (2009), to determine what path of student collaboration behaviors leads to a more successful eventual group project. In **causal data mining**, the goal is to find whether one event (or observed construct) was the cause of another event (or observed construct), for example to predict which factors will lead a student to do poorly in a class (Fancsali, 2012). What all of these methodologies share is the potential to find unexpected but meaningful relationships between variables; as such, they can be used for a wide range of applications, generating new hypotheses for further investigation, or identifying contexts for potential intervention by automated systems.

### 2d. Distillation of data for human judgment

For data to be useful to educators, it has to be timely. When educators have immediate access to visualizations of learner interactions or misconceptions that are reflected in students' writing and interaction, they can incorporate those data quickly into pedagogical activity. For this reason, one methodology that is common in LA is the **distillation of data for human judgment**. There has been a rich history of data visualization methods, which can be leveraged

to support both basic research and practitioners (teachers, school leaders, and others) in their decision-making. For example, visualizations of student trajectories through the school years can be used to identify common patterns among successful and unsuccessful students, or to infer which students are at-risk, sufficiently early to drive intervention (Bowers, 2010). Some of the visualization methods that have been used in education include **heat maps** (which incorporate much of the same information as scatterplots, but are more scalable – cf. Bowers, 2010), **learning curves** (which show performance over time – cf. Koedinger et al., 2010), and **learnograms** (which show student alternation between activities over time – cf. Hershkovitz & Nachmias, 2008).

### 2e. Discovery with Models

In **discovery with models** (Baker & Yacef, 2009; Hershkovitz et al., in press), the results of one data mining analysis are utilized within another data mining analysis. Most commonly, a model of some construct is obtained, generally through prediction methods. This model is then applied to data in order to assess the construct the model identifies. The predictions of the model are then used as input to another data mining method. There are several ways that discovery with models can be conducted.

Perhaps the most common way that discovery with models is conducted is when a prediction model is used within another prediction model. In this situation, the initial model's predictions (which represent predicted variables in the original model) become predictor variables in the new prediction model. In this way, models can be composed of other models, or based on other models, sometimes at multiple levels. For instance, prediction models of student robust learning (cf. Baker, Gowda, & Corbett, 2011) have generally depended on models of student meta-cognitive behaviors (cf. Aleven et al., 2006), which have in turn depended on assessments of latent student knowledge (cf. Corbett & Anderson, 1995), which have in turn depended on models of domain structure (cf. Koedinger et al., 2012).

A second common way that discovery with models is conducted is when a prediction model is used within a relationship mining analysis. In this type of research, the relationships between the initial model's predictions and additional variables are studied. This enables a researcher to study the relationship between a complex latent construct (represented by the prediction model) and a wide variety of other variables. One example of this is seen in work by (Beal, Qu, & Lee, 2008), who developed a prediction model of gaming the system and correlated it to student individual differences in order to understand which students are most likely to engage in gaming behavior.

It is worth noting that the models used in discovery with models do not have to be obtained through prediction methods. These models can also be obtained through other approaches such as cluster analysis or **knowledge engineering** (Feigenbaum & McCorduck, 1983; Studer, Benjamins, & Fensel, 1998)**,** where a human being rationally develops a model rather than using data mining to produce a model. The merits of knowledge engineering versus data mining

for this type of analysis are out of scope for this chapter; greater discussion of this issue can be found in (Hershkovitz et al., in press).

### 2f. Tools for Conducting EDM/LA Methods

In recent years, dozens of tools have emerged for data mining and analytics, from both the commercial and academic sectors. The majority of papers published in the proceedings of the the EDM and LAK conferences, or published in the *Journal of Educational Data Mining*, and allied communities (such as Artificial Intelligence in Education, Intelligent Tutoring Systems, and User Modeling and Adaptive Personalization) use publicly available tools including RapidMiner, R, Weka, KEEL, and SNAPP. These tools include algorithms that implement the methods discussed above, and provide support for readying data for use within these methods. They also provide support for conducting statistical validation of model appropriateness for a range of uses (RapidMiner is particularly flexible for conducting sophisticated validation, leading to it becoming increasingly popular among EDM researchers), and for visualizing data.

In addition to these general-purpose tools, other tools are available for special purposes. For example, two competing packages are available for estimating student knowledge with Bayesian Knowledge Tracing (e.g., Chang et al., 2006; Baker et al., 2010). Tools for supporting the development of prediction models, by obtaining data on the predicted variable through hand-annotating log files, have also recently become available (Rodrigo et al., 2012). Tools for displaying student learning over time and the pattern of student performance for different problems or items have been embedded into the Pittsburgh Science of Learning Center's DataShop, a very large public database on student use of educational software (Koedinger et al., 2010).

Some analytics tools are open source, allowing any researcher to develop add-ons that increase the core functionality of the tool. For example, R (http://www.r-project.org) is an open source statistical computing environment, and a very popular tool for statistical and advanced analytics (Muenchen, 2012). A valuable feature for R users is the ability for researchers to create specific R packages to address research needs in specific fields. Weka (Witten & Frank, 2005; http://www.cs.waikato.ac.nz/ml/weka) is also open source, and a number of its tools for data mining have been incorporated into other tools, such as RapidMiner. Open platforms like R and Weka allow researchers to scrutinize the methods and algorithms used by other researchers.

Commercial tools are driving the administrative use of analytics in many schools and districts. Enterprise tools such as IBM Cognos, SAS, and analytics offerings by learning management system providers such as Blackboard, and student systems, such as Ellucian, enable an integrated research/application approach. The questions being asked by administrators and educators can sometimes differ in focus from those being asked by EDM researchers. For example, a researcher may look for patterns in data, test algorithms, or develop analytics models to understand what contributed to learning success. In contrast, institutional analytics activities are likely to focus on improving learner success and providing support programs.

3. Impacts on learning sciences

Educational data mining and learning analytics have had several recent impacts on the learning sciences.

One area where these methods have been particularly useful is in research on *disengagement* within educational software. Prior to the development of analytics, disengagement was difficult to measure (Corno & Mandinach, 1983), but EDM and LA methods have produced models that can infer disengaged behaviors in a fine-grained fashion. Automated detectors of a range of disengaged behaviors have been developed using prediction methods or knowledge engineering methods, including detectors of gaming the system (Baker, Corbett, & Koedinger, 2004; Walonoski & Heffernan, 2006; Beal, Qu, & Lee, 2008; Muldner et al., 2011), off-task behavior (Baker, 2007; Cetintas et al., 2010), carelessness (San Pedro et al., 2011a), and inexplicable behaviors (Sabourin et al., 2011; Wixon et al., 2012).

These detectors have been used to study the relationship between these behaviors and learning (Baker et al., 2004; Walonoski & Heffernan, 2006; Baker, 2007; Cocea et al., 2009; Baker, Gowda, & Corbett, 2011), including study of how behaviors lead to differences in learning (e.g., Cocea et al., 2009), and how seemingly disengaged behaviors might in some cases, paradoxically, reflect deep engagement (Shih, Koedinger, & Scheines, 2008). They have also been used to understand what affect is associated with these behaviors (Sabourin et al., 2011; San Pedro et al., 2011b), and which learner individual differences are associated with these behaviors (Walonoski & Heffernan, 2006; Beal, Qu, & Lee, 2008).

Detectors have also been embedded into intelligent tutors that adapt based on student disengagement. For instance, Baker et al. (2006) built an automated detector of gaming the system into a software agent that provides alternative exercises to students who game, both giving students an alternate way to learn material bypassed by gaming and making gaming behavior more time-consuming. Arroyo et al. (2007) used an automated detector of gaming to provide students with information on their recent gaming and to provide an opportunity to give meta-cognitive messages on how to use the software more effectively. Each of these approaches resulted in less gaming, and better learning.

EDM and LA methods have similarly been useful in understanding student learning in various collaborative settings. Collaborative learning behaviors have been analyzed in order to determine which behaviors are characteristic of more successful groups and more successful learners, in multiple contexts, including computer-mediated discussions (McLaren et al., 2007, 2010), online collaboration using software development tools (Kay et al., 2006), and interactive tabletop collaboration (Martinez et al., 2011). For instance, Prata and colleagues (2012) found that students who were contradicted by their partners when they were incorrect tended to learn more than students whose partners chose not to correct them. Dyke and colleagues (2012) found that off-topic discussions during collaborative learning are more harmful to learning within

some parts of the learning process than during other parts – specifically, off-topic discussion is more harmful when learning basic facts than during discussion of problem-solving alternatives. Models based on student contributions to online discussion forums have even been able to predict those students' final course grades (Ming & Ming, 2012). This work has been embedded into automated agents that scaffold more effective collaboration (McLaren et al., 2010; Dyke et al., 2012), and tools to support instructors in scaffolding their students' collaboration (Martinez et al., 2012).

## 4. Impacts on practice

We have often observed a positive feedback loop between research and practice in EDM and LA--with research discoveries leading to changes in practice, which in turn lead to the possibility of studying new issues. One example of this can be seen in research over the years on student knowledge in Cognitive Tutors (cf. Koedinger & Corbett, 2006). In the mid-1990s, mastery learning (where a student keeps receiving problems of a certain type until he or she successfully demonstrates mastery) was introduced, based on assessments of student mastery from the EDM algorithm Bayesian Knowledge Tracing (Corbett & Anderson, 1995). A prerequisite structure – e.g. an ordering of which content must be learned prior to other content because it is needed to understand the later content – was developed for the content in Cognitive Tutors and applied in the design of tutor lessons. However, freedom was given to instructors to deviate from the planned prerequisite structure in line with their pedagogical goals – in other words, if an instructor thought that a prerequisite topic was not actually needed for a later topic, they could skip the prerequisite topic. This enabled later analyses of the pedagogical value of the prerequisite structure (Vuong, Nixon, & Towle, 2011). These results found that it is disadvantageous to students when prerequisite structure is ignored by instructors.

The development of analytics around social learning and discourse have been important in increasing awareness of the impact of social dimensions of learning and the impact of learning environment design on subsequent learning success. Several open online courses, such as etmooc (http://etmooc.org) and edfuture (http://edfuture.mooc.ca) have incorporated principles from social network theory (and related analytics) in the design of distributed, networked learning systems, in contrast with more centralized platforms such as learning management systems. For example, the analytics options available to researchers in a platform where data collection is centralized (as in an LMS) differ from the analytics approaches possible when data is distributed and fragmented across numerous technical and social spaces. A predictive learner success tool such as Purdue Signals draws data from LMS interactions and from the student information system. In contrast, learning in distributed social networks produces data that reflects learner interest and engagement across multiple spaces that are under the control of individuals. As analytics of learning move into a broader range of settings—such as informal interactions through peer networks in universities, workplace learning, or lifeline learning--EDM and LA can help to evaluate how learning happens across various settings and how patterns of engagement or predictors of success differ in distributed versus centralized learning systems..

## 5. EDM/LA and the Learning Sciences: to the future

Educational data mining and learning analytics, despite being new research areas, have already made contributions to the learning sciences and to practice. The current trend suggests that this contribution will continue, and even increase in the years to come.

One key trend is that these methods have been applied to an ever-widening range of data sources. Much of the early work in EDM was conducted within intelligent tutoring systems (as described in Koedinger and Corbett, 2006) and much of the work in LA began in web-based e-Learning and social learning environments. In recent years, this has extended to a wider variety of educational situations, including data from student collaboration around learning resources (Martinez et al., 2012), science simulations (Sao Pedro et al., 2013), teacher newsgroups (Xu & Recker, 2011), and school district grade data systems (Bowers, 2010).

A second key trend, which can be seen in the examples above, is the use of EDM methods to answer an expanding range of research questions, in an expanding range of areas represented in this handbook: computer games (Steinkuller and Squire, this volume; cf. Hernandez, Sucar, & Conati, 2009; Kerr & Chung, 2012), argumentation (Andriessen & Baker, this volume; cf. Lynch, Pinkwart, Ashley, & Aleven, 2008; McLaren, Scheuer, & Miksatko, 2010), computer-supported collaborative learning (Stahl et al., this volume;  Kay et al., 2006; cf. Dyke et al., 2012), learning in virtual worlds (Kafai and Dede, this volume; cf. Sil et al., 2012), and teacher learning (Fishman et al., this volume; cf. Xu & Recker, 2011).

As EDM and LA become used in a wider variety of domains, by researchers from a wider variety of disciplines, and within learning systems of a wider variety of types, we will see the potential of these approaches for enhancing both practice and theory in the learning sciences. As this occurs, there will be opportunities to conduct finer-grained, broader-scale research in the learning sciences, benefiting the field and the learners impacted by developments in the field. .

6. Acknowledgements

We would like to thank Lisa Rossi and Keith Sawyer for helpful comments and suggestions.

7. References

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence and Education*, 16, 101-128.

Allen, I.E., Seaman, J. (2013). Changing Course: Ten years of tracking online education in the United States. Sloan Consortium. Available from: http://sloanconsortium.org/publications/survey/changing_course_2012

Amershi, S., Conati, C. (2009). Combining Unsupervised and Supervised Machine Learning to Build User Models for Exploratory Learning Environments. *Journal of Educational Data Mining*, 1(1), 71-81.

Anderson, J.R., Lebiere, C. (1998) *Atomic Components of Thought.* Mahwah, NJ: Lawrence Erlbaum Associates.

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired.*

Arnold, K.E. (2010). Signals: Applying academic analytics. *Educause Quarterly*, 33, 1-10.

Arroyo, I., Ferguson, K., Johns, J., Dragon, T., Meheranian, H., Fisher, D., Barto, A., Mahadevan, S., and Woolf. B.P. (2007). Repairing Disengagement with Non-Invasive Interventions. *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, 195-202.Baker, R.S., Corbett, A.T., Koedinger, K.R. (2004). Detecting Student Misuse of Intelligent Tutoring Systems. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, 531-540.

Baker, R.S., Corbett, A.T., Koedinger, K.R. (2004). Detecting Student Misuse of Intelligent Tutoring Systems. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, 531-540.

Baker, R.S.J.d. (2007). Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems. *Proceedings of ACM CHI 2007: Computer-Human Interaction*, 1059-1068.

Baker, R.S.J.d., Corbett, A.T., Gowda, S.M., Wagner, A.Z., MacLaren, B.M., Kauffman, L.R., Mitchell, A.P., Giguere, S. (2010). Contextual Slip and Prediction of Student Performance After Use of an Intelligent Tutor. *Proceedings of the 18th Annual Conference on User Modeling, Adaptation, and Personalization*, 52-63.

Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., Evenson, S.E., Roll, I., Wagner, A.Z., Naim, M., Raspat, J., Baker, D.J., Beck, J. (2006). Adapting to When Students Game an Intelligent Tutoring System. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 392-401.

Baker, R.S.J.d., de Carvalho, A.M.J.A., Raspat, J., Aleven, V., Corbett, A.T., Koedinger, K.R. (2009). Educational Software Features that Encourage and Discourage "Gaming the System". *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 475-482.

Baker, R.S.J.d., Gowda, S.M., Corbett, A.T. (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.

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

Barnes, T., D. Bitzer, and Vouk, M. (2005). Experimental analysis of the q-matrix method in knowledge discovery. *Proceedings of the 15th International Symposium on Methodologies for Intelligent Systems*, May 25-28, 2005, Saratoga Springs, NY.

Beal, C.R., Qu, L., & Lee, H. (2006). Classifying learner engagement through integration of multiple data sources. Paper presented at the *21st National Conference on Artificial Intelligence (AAAI-2006)*, Boston, MA.

Beal, C.R., Qu, L., Lee, H. (2008). Mathematics motivation and achievement as predictors of high school students' guessing and help-seeking with instructional software. *Journal of Computer Assisted Learning*, 24, 507-514.

Beck, J.E., Chang, K.-m., Mostow, J., Corbett, A.T. (2008) Does help help? Introducing the Bayesian evaluation and assessment methodology. *Proceedings of Intelligent Tutoring Systems, ITS 2008,* 383–394. Ben-Naim, D., Bain, M., Marcus, N. (2009). User-Driven and Data-Driven Approach for Supporting Teachers in Reflection and Adaptation of Adaptive Tutorials. *Proceedings of the 2nd International Conference on Educational Data Mining*, 21-30.

Bowers, A.J. (2010). Analyzing the Longitudinal K-12 Grading Histories of Entire Cohorts of Students: Grades, Data Driven Decision Making, Dropping Out and Hierarchical Cluster Analysis. *Practical Assessment, Research & Evaluation (PARE)*, 15(7), 1-18.

Buckingham Shum, S., Ferguson, R., (2012). Social Learning Analytics. *Educational Technology and Society*, 15(3), 3-26.

Cen, H., Koedinger, K., and Junker, B. (2006). Learning Factors Analysis - A general method for cognitive model evaluation and improvement. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 164-175.

Cetintas, S., Si, L., Xin, Y., & Hord, C. (2010). Automatic Detection of Off-Task Behaviors in Intelligent Tutoring Systems with Machine Learning Techniques. *IEEE Transactions on Learning Technologies*, 3(3), 228-236.

Chang, K.-m., J. Beck, J. Mostow, and A. Corbett. (2006). A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 104-113, Jhongli, Taiwan.

Choudhury, T., Pentland, A. (2003). Sensing and modeling human networks using sociometer. *Proceedings of the Seventh IEEE International Symposium on Wearable Computers (ISWC'03).*

Cocea, M., Hershkovitz, A., Baker, R.S.J.d. (2009). The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate? *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 507-514.

Collins, F.S., Morgan, M., Patrinos, A. (2004). The Human Genome Project: Lessons from Large-Scale Biology. *Science*, 300, 5617, 286-290.

Corbett, A.T., Anderson, J.R. (1995). Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4, 253-278.

Corno, L. & Mandinach, E.B. (1983). The role of cognitive engagement in classroom learning and motivation. *Educational Psychologist*, 18(2), 88-108.

D'Mello, S.K., Craig, S.D., Witherspoon, A., McDaniel, B., and Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and UserAdapted Interaction*, 18, 45-80.

Dawson, S. (2008). A study of the relationship between student social networks and sense of community. *Educational Technology & Society*, 11(3), 224-238.

Dean, J., Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1).

Dekker, G., Pechenizkiy, M., and Vleeshouwers, J. (2009). Predicting students drop out: A case study. *Proceedings of the 2nd International Conference on Educational Data Mining, EDM'09*, 41-50.

Desmarais, M.C. (2011). Conditions for effectively deriving a q-matrix from data with non-negative matrix factorization. In C. Conati, S. Ventura, T. Calders, and M. Pechenizkiy, editors, *4th International Conference on Educational Data Mining, EDM 2011*, 41-50, Eindhoven, Netherlands.

Dyke, G., Adamson, D., Howley, I., Rosé, C.P. (2012). Towards Academically Productive Talk Supported by Conversational Agents. *Intelligent Tutoring Systems*, 531-540.

Dyke, G., Kumar, R., Ai, H., Rosé, C. P. (2012) Challenging Assumptions: using sliding window visualizations to reveal time-based irregularities in CSCL processes. *International Conference of the Learning Sciences (ICLS 2012)*, Sydney, 363—370.

Fancsali, S. (2012) Variable Construction and Causal Discovery for Cognitive Tutor Log Data: Initial Results. *Proceedings of the 5th International Conference on Educational Data Mining*, 238-239.

Feigenbaum, E. A., McCorduck, P. (1983) *The fifth generation: Artificial Intelligence and Japan's Computer Challenge to the World.* Reading, MA: Addison-Wesley.

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the Assessment Challenge in an Intelligent Tutoring System that Tutors as it Assesses. *User Modeling and User-Adapted Interaction*, 19, 243-266.

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. http://kmi.open.ac.uk/publications/techreport/kmi-12-01

Greeno, J.G. (1998) The situativity of knowing, learning, and research. *American Psychologist, 53* (1), 5-26.

Halevy, A.Y., Norvig, P., Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2), 8–12.

Haythornthwaite, C. (2001). Exploring Multiplexity: Social Network Structures in a Computer-Supported Distance Learning Class. *The Information Society: An International Journal*, 17 (3), 211-226.

Hernanzez Y., Sucar E., and Conati C. (2009). Incorporating an Affective Behaviour Model into an Educational Game. *Proceedings of FLAIR 2009, 22nd International Conference of the Florida Artificial Intelligence Society*, ACM Press.

Hershkovitz, A., Baker, R.S.J.d., Gobert, J., Wixon, M., Sao Pedro, M. (in press) Discovery with Models: A Case Study on Carelessness in Computer-based Science Inquiry. To appear in *Amercian Behavioral Scientist*.

Hershkovitz, A., Nachmias, R. (2008) Developing a log-based motivation measuring tool. *Proceedings of the 1st International Conference on Educational Data Mining,* 226-233.

Kay, J., Maisonneuve, N., Yacef, K., Reimann, P. (2006) The Big Five and Visualisations of Team Work Activity. *Proceedings of the International Conference on Intelligent Tutoring Systems*, 197 – 206.

Kerr, D., Chung, G.K.W.K. (2012). Identifying Key Features of Student Performance in Educational Video Games and Simulations through Cluster Analysis. *Journal of Educational Data Mining*, 4(1), 144-182.

Koedinger, K. R., & Corbett, A. T. (2006). Cognitive Tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.) *The Cambridge Handbook of the Learning Sciences* (pp. 61-78). New York: Cambridge University Press.

Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. (2010). A Data Repository for the EDM community: The PSLC DataShop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press, 43-56.

Koedinger, K.R., Corbett, A.T., Perfetti, C.The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning.*Cognitive Science, 36 (5),* 757-798.

Koedinger, K.R., McLaughlin, E.A., Stamper, J.C. (2012). Automated Student Model Improvement. *Proceedings of the 5th International Conference on Educational Data Mining*, 17-24.

Lin, L (2012). edX platform integrates into classes http://tech.mit.edu/V132/N48/801edx.html

Lynch, C., Ashley, K., Pinkwart, N., & Aleven, V. (2008) "Argument Graph Classification with Genetic Programming and C4.5" In R.S.J.d. Baker, T. Barnes, and J. E. Beck (Editors). *Educational Data Mining 2008: Proceedings of the 1st International Conference on Educational Data Mining*, Montréal, Québec, Canada, June 20-21 2008. (p 137-146)

Macfadyen, L. P., Dawson, S. (2010). Mining LMS data to develop an “early warning system” for educators: A proof of concept. *Computers & Education*. 588-599

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A.H. (2011). Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute.

Martinez, R., Yacef, K., Kay, J., Kharrufa, A., AlQaraghuli, A. (2011) Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop. *Proceedings of the 4th International Conference on Educational Data Mining,* 111-120.

Martinez, R., Yacef, K., Kay, J., and Schwendimann, B. (2012). An interactive teacher’s dashboard for monitoring multiple groups in a multi-tabletop learning environment. *Proceedings of Intelligent Tutoring Systems*, 482-492. Springer.

Mayer, M (2009). The physics of big data: http://www.parc.com/event/936/innovation-at-google.html

McLaren, B.M., Scheuer, O., & Mikšátko, J. (2010). Supporting collaborative learning and e-Discussions using artificial intelligence techniques. *International Journal of Artificial Intelligence in Education (IJAIED)*, 20(1), 1-46.

McLaren, B.M., Scheuer, O., DeLaat, M., Hever, R., DeGroot, R., & Rosé, C.P. (2007). Using Machine Learning Techniques to Analyze and Support Mediation of Student E-Discussions. *Proceedings of the 13th International Conference on Artificial Intelligence in Education (AIED 2007)*.

McLaren, B.M., Scheuer, O., Miksatko, J. (2010). Supporting Collaborative Learning and E-Discussions Using Artificial Intelligence Techniques. *International Journal of Artificial Intelligence in Education*, 20, 1-46.

Ming, N.C., Ming, V.L. (2012). Predicting Student Outcomes from Unstructured Data. *Proceedings of the 2nd International Workshop on Personalization Approaches in Learning Environments*, 11-16.

Muenchen, R. A. (2012). The popularity of data analysis software. http://r4stats.com/articles/popularity/

Muldner, K., Burleson, W., Van de Sande, B., VanLehn, K. (2011). An analysis of students’ gaming behaviors in an intelligent tutoring system: Predictors and impacts. *User Modeling and User-Adapted Interaction*, 21(1-2), 99-135.

Pardos, Z.A., Baker, R.S.J.d., Gowda, S.M., Heffernan, N.T. (2011). The Sum is Greater than the Parts: Ensembling Models of Student Knowledge in Educational Software. *SIGKDD Explorations*, 13 (2), 37-44.

Pavlik, P.I., Cen, H., Koedinger, K.R. (2009). Performance Factors Analysis -- A New Alternative to Knowledge Tracing. *Proceedings of AIED2009*.

Perera, D., Kay, J., Koprinska, I., Yacef, K., and Zaiane, O.R. (2009). Clustering and Sequential Pattern Mining of Online Collaborative Learning Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(6), 759-772.

Prata, D., Letouze, P., Costa, E., Prata, M., Brito, G. (2012). Dialogue Analysis in Collaborative Learning. *International Journal of e-Education, e-Business, e-Management, and e-Learning*, 2 (5), 365-372.

Reimann, P., Yacef, K., Kay, J. (2011). Analyzing Collaborative Interactions with Data Mining Methods for the Benefit of Learning. *Computer-Supported Collaborative Learning Series*, 12, 161-185.

Rodrigo, M.M.T., Baker, R.S.J.d., McLaren, B., Jayme, A., Dy, T. (2012). Development of a Workbench to Address the Educational Data Mining Bottleneck. *Proceedings of the 5th International Conference on Educational Data Mining*, 152-155.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state-ofthe-art. *IEEE Transaction on Systems, Man and Cybernetics, part C: Applications and Reviews*, 40(6), 610–618

Sabourin, J., Rowe, J., Mott, B., Lester, J. (2011). When Off-Task in On-Task: The Affective Role of Off-Task Behavior in Narrative-Centered Learning Environments. *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, 534-536.

San Pedro, M.O.C., Baker, R., Rodrigo, M.M. (2011). Detecting Carelessness through Contextual Estimation of Slip Probabilities among Students Using an Intelligent Tutor for Mathematics. *Proceedings of 15th International Conference on Artificial Intelligence in Education*, 304-311.

San Pedro, M.O.C., Rodrigo, M.M., Baker, R.S.J.d. (2011). The Relationship between Carelessness and Affect in a Cognitive Tutor. *Proceedings of the 4th bi-annual International Conference on Affective Computing and Intelligent Interaction*.

Sao Pedro, M.A., Baker, R.S.J.d., Gobert, J., Montalvo, O., Nakama, A. (2013). Leveraging Machine-Learned Detectors of Systematic Inquiry Behavior to Estimate and Predict Transfer of Inquiry Skill. *User Modeling and User-Adapted Interaction*, 23 (1), 1-39.

Shih, B., Koedinger, K., Scheines, R. (2008). A Response Time Model for Bottom-Out Hints as Worked Examples. *Proceedings of the 1st International Conference on Educational Data Mining*, 117-126.

Siemens, G., Baker, R.S.J.d. (2012). Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*.

Sil, A., Shelton, A., Ketelhut, D.J., Yates, A. (2012). Automatic Grading of Scientific Inquiry. *The 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, 22-32.

Studer, R., Benjamins, V.R., Fensel, D. (1998) Knowledge engineering: Principles and methods. *Data and Knowledge Engineering (DKE),* 25(1–2),161–197.

Summers, D.J., et al. (1992). Charm Physics at Fermilab E791. *Proceedings of the XXVIIth Recontre de Moriond, Electroweak Interactions and Unified Theories*, Les Arcs, France, 417-422.

Suthers, D., Rosen, D. (2011). A unified framework for multi-level analysis of distributed learning. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. 64-74

Tatsuoka, K.K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment*, 327–359. Hillsdale NJ: Erlbaum.

Vuong, A., Nixon, T., and Towle, B. (2011). A method for finding prerequisites within a curriculum. *Proceedings of the 4th International Conference on Educational Data Mining*, 211-216.

Walonoski, J.A., & Heffernan, N.T. (2006). Detection and analysis of off-task gaming behavior in intelligent tutoring systems. In Ikeda, Ashley & Chan (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, (pp. 382-391), Jhongli, Taiwan. Berlin: Springer-Verlag.

Wixon, M., Baker, R.S.J.d., Gobert, J., Ocumpaugh, J., Bachmann, M. (2012). WTF? Detecting Students who are Conducting Inquiry Without Thinking Fastidiously. *Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UMAP 2012)*, 286-298.

Witten, I.H., Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques.* San Francisco, CA: Morgan Kaufmann.

Xu, B., Recker, M. (2011). Understanding Teacher Users of a Digital Library Service: A clustering approach. *Journal of Educational Data Mining*, 3 (1), 1-28.