Data Analytics: Techniques and Applications to Transform Army Learning

Rodney Long
U.S. Army Research Laboratory
Orlando, FL
Rodney.A.Long3.Civ@mail.mil

Mike Smith, Sue Dass, Clarence Dillon, Katherine Hill
ICF International
Fairfax, VA
Mike.Smith@icfi.com, Sue.Dass@icfi.com, Clarence.Dillon@icfi.com

ABSTRACT

Data analytics is the scientific process of transforming data into insight for making better decisions and is used in industry to improve organizational decision-making and in the sciences to verify or disprove existing models or theories. Current data analytic models have begun to make an impact on the way that courses are designed, run, and evaluated, although little progress has been made towards the design of a structured method to categorize and implement data measurements as they relate to the Army Learning Model (ALM) goals. The following paper describes ongoing work with the U.S. Army Research Laboratory to examine data analytics as it relates to the design of courses, evaluation of individual and group performances, and the ability to tailor the learning experience to achieve optimal learning outcomes. This paper describes: a) the methodology for research and evaluation; b) the fields of Learning Analytics and Educational Data Mining; c) data analytics methods and techniques relevant to learning systems; and d) a framework for applying data analytic methods and techniques for learning via three illustrative use cases. Ultimately, the goal of this paper will be to provide a vision for successful application of these techniques within the Army learning community and higher education.

ABOUT THE AUTHORS

Rodney Long is a Science and Technology Manager at the Army Research Laboratory in Florida and is currently conducting research in adaptive training technologies. Mr. Long has a wide range of simulation and training experience spanning 28 years in the Department of Defense (DoD) and has a Bachelor's Degree in Computer Engineering from the University of South Carolina and Master's degree in Industrial Engineering from the University of Central Florida.

Mike Smith has over 11 years of experience in data analytics, strategic planning, and risk assessment. Mr. Smith currently advises several DoD clients on how to adapt emerging analytics practices to improve their organizational performance. Mr. Smith has a Bachelor of Arts in International Economics from Longwood University and a Master of Public Policy from Georgetown University and is a Certified Analytics Professional (accredited by INFORMS).

Sue Dass has over 13 years designing, developing, and managing instructional design projects. Dr. Dass is familiar with the many advanced learning technologies available having co-designed an electronic performance support tool to help faculty explore, select, and implement learning technologies based on learning objectives. Dr. Dass has a B.S. in Civil Engineering, a M.Ed. in Instructional Design, and a Ph.D. in Education.

Clarence Dillon has over 20 years of experience as a project manager, strategist, and analyst. He authored the ontology for the DoD strategic planning scenarios and established the first collaborative, semantic web platform in DoD. Mr. Dillon has been conducting graduate research in social complexity at George Mason University's Krasnow Institute of Advanced Study, holds a Bachelor of Arts in International Affairs and a Masters of Social Science in International Relations.

Katherine Hill is a Business Systems Analyst at ICF International focusing on data analytics and business process improvement. Ms. King has a Master’s degree in Business Administration from George Mason University and a Bachelor’s of Science in Foreign Service degree from Georgetown University.

Data Analytics: Techniques and Applications to Transform Army Learning

Rodney Long
U.S. Army Research Laboratory
Orlando, FL
Rodney.A.Long3.Civ@mail.mil

Mike Smith, Sue Dass, Clarence Dillon, Kate King
ICF International
Fairfax, VA
Mike.Smith@icfi.com, Sue.Dass@icfi.com, Clarence.Dillon@icfi.com

1. INTRODUCTION

Big data, the contemporary use of parallel processing to derive value from large-scale, heterogeneous data sets, has begun a transformational shift across society that has already changed the way business operates and academia evaluates performance, and promises to reshape society at large. Computers have self-learned that cats are important in online videos (Markoff, 2012), cars can self-drive (Greenough, 2015), and both the component tools and practices that underlie these innovations are available to the public—for free. Exponential increases in processing power and data availability continue to drive the creation of qualitatively new analytic methods, tools, and techniques that have transformative implications for learning research and practices. Within the past 10 years, the field of data analytics has emerged as a synthesis of computer science and statistics, now both necessary for dealing with complex, data-intensive challenges. What these developments ultimately mean for the Army learning community will come down to the practitioner's ability to develop valuable learning products using data-intensive methods.

This paper explores the impact of data analytics as it relates to the design of courses, evaluation of individual and group performance, and the ability to refine the learning experience to achieve optimal learning outcomes across institutions and the Army at large. This paper describes: a) the methodology for research and evaluation; b) the fields of Learning Analytics and Educational Data Mining; c) data analytics methods and techniques relevant to learning systems; and d) a framework for applying data analytic methods and techniques for via three illustrative use cases. Ultimately, the goal of this paper will be to provide a vision for successful application of these techniques within the Army learning community and higher education.

2. METHODOLOGY

The authors focused their research efforts on discovery and evaluation of data analytics methods and techniques that have direct potential to improve Army Learning practices and initiatives. The vast scope of data analytics disciplines precludes an exhaustive key-word search methodology, so the authors developed an iterative research process to identify where methods and techniques were being researched and applied in the fields of Learning Analytics and Educational Data Mining, which will be reviewed in Section 3. As described in Figure 1, these fields provided examples of current applications and a context for reviewing the broader literature on data analytics. In turn, the focused research into data analytics provided a broader perspective and potential for new methods that can be applied in a learning context. Using this methodology, research began by identifying key Learning Analytics and Educational Data Mining sources. Of primary interest were Siemens (2012); Springer's compilations on both disciplines (Peña-Ayala, 2014; Larusson & White 2014); Siemens and Baker (2012); and Papamitsiou and Economides (2014). Forward- and backward-searching was utilized to expand the bibliography for sources that were primarily published in academic journals.

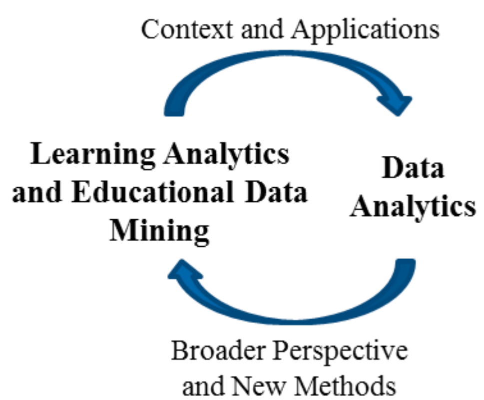

Figure 1. Research Methodology Overview

Data analytics research is complicated by the fact that many innovations and emerging methods and techniques are developed by practitioners and industry partners which are not always published in academic journals. The authors utilized key data mining books used in educational settings, including Provost and Fawcett (2013), Chambers and Dinsmore (2014), and O'Neil and Schutt (2014), as well as a review of topics from data analytics websites, including O'Reilly Media, KDD Nuggets, Data Science Central, and MIT Technology Review, to focus on the set of topical research areas in Table 1. In cases where websites were utilized to identify innovative concepts, the authors identified subsequent publications that supported the ideas presented in the article. In all cases, the authors endeavored to strike the right balance between emerging and potentially experimental applications in the commercial world with established practices being studied and applied in an academic setting. While there are common underlying techniques in each of these areas, for example classification methods are used in several of these areas, after several iterations these groups were selected as a useful and succinct method for binning research based on intended applications.

Table 1. Methods and Techniques

Recommender Systems
Similarity Grouping
Predictive Modelling
Social Network Analysis
Natural Language Processing & Text Mining
Machine Learning
Big Data Tools and MapReduce

3. LEARNING ANALYTICS AND EDUCATIONAL DATA MINING

Data analytics began in the business industry to support data-driven decision-making that relied on sophisticated algorithms (Papamitsiou & Economides, 2014; Scheffel, Drachsler, Stoyanov, & Specht, 2014). Educational Data Mining (EDM) evolved from general data mining to explore methods as applied specifically to educational data to meet educational goals (Chatti, Dyckhoff, Schroeder, & Thüs, 2012). EDM is focused on using big data to inform education using tailored data mining analytics that initially rely on automated discovery to find patterns. In contrast, Learning Analytics (LA) leverages and furthers EDM results to improve education using data mining analytics, as well as other techniques that initially rely on human interpretation to first define the pattern. The connection between EDM and LA is the use of models whether as predictive, descriptive, or prescriptive in nature as shown in Figure 2.

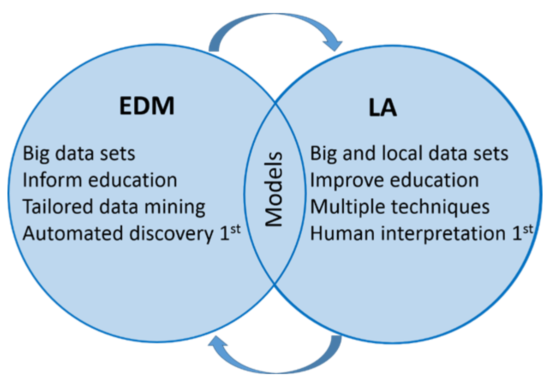

Figure 2. Learning Analytics & Educational Data Mining

Learning analytics (LA) is a fast growing, new multidisciplinary field that also supports educational research and educational understandings (Ferguson, 2012; Gašević, Dawson, & Siemens, 2015; Johnson, Adams Becker, Cummins, Estrada, Freeman, & Ludgate, 2013; Scheffel et al., 2014). LA is said to draw from the learning sciences, data processing, web analytics, psychology, philosophy, sociology, linguistics, information visualization, adaptive and adoptive systems, and recommender systems (Chatti et al., 2012; Dawson, Gašević, Siemens, & Joksimovic, 2014; Ferguson, 2012; Gašević et al., 2015; Scheffel et al., 2014). The field of LA continues to evolve from a focus on technological perspectives to a focus on educational perspectives with an anticipation that this evolution will further extend into other disciplines as appropriate (Dawson et al., 2014; Ferguson, 2012).

Together, "LA and EDM constitute an ecosystem of methods and techniques (in general procedures) that successively gather, process, report and act on machine-readable data on an ongoing basis in order to advance the educational environment and reflect on learning processes." (Papamitsiou & Economides, 2014, p. 49). Some researchers perceive this multidisciplinary ecosystem as a positive in that each community adheres to different standards and values in determining what is important to the community and what constitutes good research (Siemens & Baker, 2012).

LA and EDM stakeholders cover a broad range of roles. Ifenthaler and Widanapathirana (2014) has developed a hierarchical framework to represent group roles. As shown in Figure 3, the mega-level analytics represents governance, basically reviewing data at the national and international level by policy and decision-makers. The macro-level analytics explores student data across regional and state administrators to evaluate and benchmark against educational goals. The meso-level analytics explores student data at the institution level. Institutions may be interested in predicting college acceptance, improving student retention, improving college success, and identifying at-risk students (Campbell, DeBlois, & Oblinger, 2007). Lastly, the micro-level addresses individual student and small group performance where focus is on student performance and hence benefits the student and supports faculty. Learning analytics is said to focus on the micro-level at the benefit of the student and faculty while academic analytics support the meso- and macro-levels whereby the benefactors are administrators, funders, marketers, policy makers, and governments (Siemens et al., 2011).

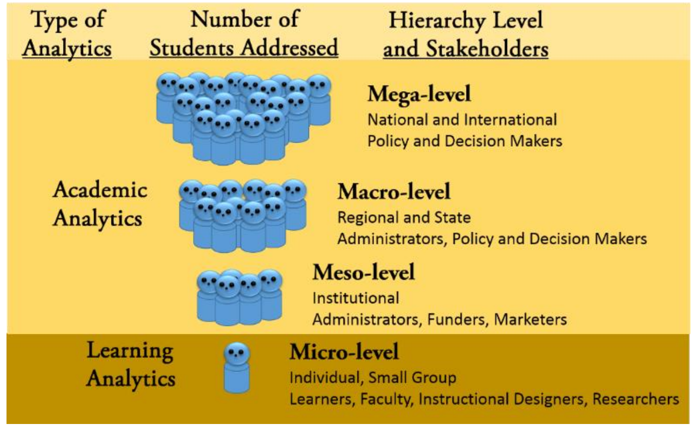

Figure 3. Hierarchy of LA and EDM Stakeholders

4. DATA ANALYTICS METHODS AND TECHNIQUES

Data analytics, as defined by the Institute for Operations Research and Management Science (INFORMS), is the scientific process of transforming data into insight for making better decisions. Data analytics is used in industry to improve organizational decision-making and in the sciences to verify or disprove existing models or theories. Modern analytic techniques draw from multiple disciplines such as statistics, artificial intelligence, software engineering and others to solve data-intensive problems and generate novel insights and products. This section describes methods and techniques described in Table 2 along with potential learning applications for each; each of these areas will also be addressed in the context of potential use cases in Section 5.

Table 2. Description of Data Analytic Methods and Techniques

Predictive Modelling	Development of a statistical relationship between input and output variables used to predict a future outcome.
Similarity Grouping	Grouping and retreiving similar data objects based on measure of their statistical distance.
Recommender Systems	Determines the affinity between a user and a content point with the goal of pointing the user to his desired content.
Social Network Analysis	Evaluates the relationships and structures of a network to identify central and isolated individuals and groups and make predictions.
Natural Language Processing & Text Mining	Methods for tagging, evalauting similarity, deriving meaning, and processing text data; often an input to other methods and techniques.
Machine Learning	Training computers as intelligent agents to extract models from data without explicit programming; used to augment other methods and techniques.
Big Data Tools and MapReduce	Programming standards and tools popularized by Google; widely used as a method to process large-scale, heterogeneous datasets

Predictive Modeling

Predictive modeling in data analytics refers to the use of statistical techniques that allow analysts to leverage the relationship of input and output variables to predict a future outcome. These relationships allow analysts to predict, classify, and act on an outcome prior to the occurrence of an event. Additionally, beyond simply outcome prediction, the use of predictive modeling allows analyst to identify the components that influence a specific outcome, enabling an organization to take targeted action to improve performance. When applied to the educational setting, these techniques prove particularly powerful in applications that range from predicting an at risk students in need of intervention to predicting the cost of a future training based on the proposed module components.

Much of the current literature on predictive modeling in the academic setting describes its use determining student success. Kongsakun (2013) describes the use of linear regression models in tangent with clustering techniques to produce a prediction model called e-Grade that predicts a student's likely final course score before and after midterm exams. This model uses predictor variables such as attendance at the first class and previous GPA to assess a student's possible course grade. Students with at risk predicted course grades are then targeted for intervention by tutors and professors. Smith and Lange (2010) describe a similar process in predicting student success in online community colleges in the United States. Similarly, Hung and Zhang (2003) describe the use of regression models to predict academic performance of students based on online learning logs. These logs contained a range of data, such as reading or posting in online discussion boards, which could be used as input variables to predict the success of undergraduate students. Likewise, predictive classification techniques were applied by Iam-on (2014) to predict student drop-out rate.

Alternatively, Hutzler, David, Avigal, and Azoulay (2014) describe the use of decision trees to predict and classify the level of difficulty of test questions provided for reading comprehension based on a set of training data in which the test question difficulty for each question has been previously assessed. In addition, the decision tree model used nine predictor variables converted into a numeric output ranging from question style (e.g., 1= multiple choice, 2 = open-ended) to presentation of information (e.g., 1= textual, 2= graphic). Using the decision tree model, the learning organization was able to predict the difficulty of new questions and create well-balanced and aligned exams.

Similarity Grouping

Similarity grouping in data analytics consists of using automated methods to identify and quantify meaningful segments in data based on statistical attributes. Clustering, one of the most popular techniques, uses a variety of attributes to cluster information into meaningful categories based on groupings of high similarity. In clustering techniques, data points grouped together not only share a high similarity within their assigned cluster grouping, but also display a high degree of dissimilarity from other clustered groupings. Due to the unsupervised nature of this technique, data clusters are not pre-defined but are formed as dictated by the attributes of the provided dataset—a method referred to as unsupervised learning. This would allow analysts using clustering analysis to draw insights from complex data sets that can be applied in the education context through the grouping of similar students, courses and course materials.

Similarity grouping has been used in a variety of contexts when applied to learning analytics and educational data mining. Govindarajan, Somasundaram, Kumar, and Kinshuk (2013) describe the use of continuous clustering techniques to organize students into similar grouping for the purpose of providing targeted learning objects. Using this clustering technique, students that share similar knowledge gaps based on previous course work or based on in-module knowledge checks could be grouped together to form follow-on classes, break out groups with the instructor, or additional modules in asynchronous training. Likewise, Kumar and Ramaswami (2010) highlight the valuable insights a training institution can gain into the make-up of their student population by clustering by variables outside of course performance including social, cultural, and economic measures. Finally, Abukhousa and Atif (2014) describe the use of cluster models to group similar students into communities of practice that allow like-minded students to share their own industry knowledge outside of the formal course setting.

Valsamidis et al. (2012) and Potok and Palathingal (2005) describe the use of cluster models to group similar courses and course materials with the purpose of gaining a deeper understanding of content or characteristic overlap and for use in combination of additional analytical techniques such as recommender engines. Valsadmis et al. (2012) demonstrate that using metadata and weblog data of course concepts, training organizations can group similar courses outside of intuitive categories for targeted action. For example, apart from simply subject areas, courses can be grouped into high or low activity groups, allowing organizations to prioritize courses in need of revision or understand the characteristics that define a popular curriculum (Valsamidis et al., 2012). Similarly, Potok and Palathingal (2005) provide an example of using cluster analysis to organize large libraries of supporting documents to provide students an easier entry into choosing supporting course materials.

Recommender Engines

Recommender engines in data analytics have become a key tool in bringing users the information they seek, at times even before the user has come to the point of beginning their search. Applied to a learning environment, recommender engines support students in finding the appropriate course, instructors in choosing the most relevant material, and course developers in choosing the best aligned learning object within large educational databases, with relative ease and efficiency. This ability to find the right educational resource at the right time becomes especially appropriate in today's changing learning environment as web-based training content databases grow and learners find they are unable to devise the search terms necessary to sufficiently filter results through a simple query. Moreover, “users do not have a precise enough understanding of what they want to formulate specific queries” (Kumar et al. 2007) and may become overwhelmed at the ‘hit-shock’ of receiving an overwhelming number of search results.

Recommender engines determine the affinity between a user and a content point with the goal of pointing the user to his desired content (Lu and Sindhwani, 2012). To meet this goal, recommender engines generally apply two primary analytical techniques, content-based filtering and collaborative filtering, to provide smart recommendations to individuals or groups of users. As education increasingly moves to a learner-centric approach (Imran et al., 2015) where the educational path is driven by the learner, recommender engines can assist students, instructors and teaching organizations in connecting the right content to the right user at the right time.

Kumar et al. (2007) describes the use of recommendation techniques to assist learners in choosing specific learning objects within large databases. Through this technique, the associated course metadata is used to recommend courses of aligned keywords based on a community-filtering model. Alternatively, El-Bishouty et al. (2014) present the use of recommendation techniques to assist learners in choosing learning content based on the student’s individual learning style captured via completed survey. Lastly, Niemann and Wolpers (2013) describe the use of weblog data, rather than content tags that can be time consuming and costly to add to learning objects, to recommend course objects to targeted learners.

Beyond recommendation of content, these models can be used in a range of additional functions to assist the learner or educational provider. Imran et al. describes the use of recommendation systems in assisting students in self-directed learning programs to select the most relevant tasks (as opposed to content or materials) for their desired training outcome (Imran et al., 2015). Within this system, Imran et al. argue for the use of a rule-based recommendation system based on past performance of similar users and course difficulty levels.

Social Network Analysis

Social Network Analysis (SNA) is a principal method for any quantitative social research because it focuses on relationships---the essence of society. SNA is a formal perspective of a social environment expressed as patterns or regularities in relationships among interacting units (Wasserman & Faust, 2009). There exists a vast literature for SNA from both a mathematical-theoretic perspective---graph theory and statistics---as well as social, application-level perspective. Many of the techniques developed to analyze social networks can be applied to analysis of other networks; almost anything that can be described by relationships. Most of the common SNA measures are simple calculations or statistics about the ratios of nodes to edges for sub-components of a network.

One example of how SNA is being applied to learning analytics is the Social Networks Adapting Pedagogical Practice (SNAPP) program lead by Shane Dawson from the University of Wollongong (Australia), which uses student social networks to inform instructors about challenges and opportunities among their students in real-time (Dawson 2009; Dawson & Heathcote, 2010)[1]. The research for this program takes advantage of Dawson's long treatment of the subject of computational learning analytics and SNA, in particular. Other research in this area has included: learner isolation (McDonald et al., 2005); how networks impact creativity (Dawson 2009); and how SNA supports instructor's view of the "big picture" of large classrooms (including recommendations on content scaffolding as class size grows (Brooks, et al., 2007).

Natural Language Processing and Text Mining

Natural Language Processing (NLP) and Text Analysis (TA) are analytic methods used to extract information from (typically) unstructured texts. Both NLP and TA are special applications of machine learning; popular algorithms can be found in texts on machine learning and artificial intelligence. Though there is a distinction between the purpose that NLP and TA each serve, these fields share some of the same evolutionary roots. Both can also be applied to big data, with appropriate modifications. The TA techniques focus on text as raw data (word counts, semiotics, roots, etc.) while NLP attempts to extract meaning from sentences or whole documents (semantics, rhetoric’s, hermeneutics) by leveraging the structure of the text's language. Many techniques have been codified in software libraries and are commonly used as sub-processes, together with other machine learning algorithms.

[1] SNAPP - Social Networks Adapting Pedagogical Practice at www.snappvis.org/

Computational analysis techniques that evolved from this approach can be applied directly to learning analytics. Some practically achievable objectives are: to evaluate media resources that represent or partially represent learning objects and catalog it into a machine-readable data; real-time (or near real-time) evaluation of students' (virtual) classroom discussions, questions, assignments, etc. for sentiment, structure, content and complexity; and to support careful reading of students' written products for in-depth evaluation, to detect plagiarism, and so forth.

Techniques such as these are already being applied by learning analytics researchers. In 2001, Wang, et al. published a paper on successful discourse analysis of online classroom chats to predict student performance. Wang and his cohort had to code their data by hand. As computer science advances, the possibility of automated coding becomes more feasible. Just seven years after Wang's paper, Rose, et al. (2008) published a study comparing computer-based NLP to hand coding, showing between 42% and 97% accuracy over various measures of similarity, using a variety of NLP methods. Data analysis tools and technology have continued to improve since 2008 and now offer a wider selection of methods than Rose, et al (2008) had available.

State of the art TA and NLP methods have improved to the point that in some disciplines computer-automated coding is no more prone to error than human coding (Leetaru & Schrodt, 2013). However, some features of natural language, such as sentiment and humor, are difficult to detect with common algorithms and simple machines (Davidov, Tsur, & Rappoport 2010). Statements that seem to have a positive sentiment analytically actually represent negative sentiment. Tsvetovat (2011) calls this the "flat tire problem"[2]. Similarly, while it is currently possible to train a computer to recognize images in pictures or video, they do not recognize the meanings in those images. Thus, despite advancements, many media types still need to be hand-coded.[3]

Machine Learning

Machine learning is a field of practice adapted from the artificial intelligence discipline that focuses on training computers as intelligent agents to extract models from data without explicit programming. While considerable overlap exists across each of the aforementioned techniques, the primary focus in machine learning is on improvement of a machine’s ability to model the world based on experience, as opposed to a more generalized development of model(s) where the primary agent is the analyst. While machine learning experts tend to have experience in both computer science and data analytics, computer science tends to be the primary field of machine learning practitioners, as opposed to analytics practitioners that tend to have backgrounds in statistics, operations research and business.

As a method for applying automated agents to solve computational problems, machine learning is not a data analytic technique in and of itself. Rather it should be seen as a tool for tackling problems that would be intractable or inefficient without computer assistance. The greater the complexity of the problem set, the more appropriate to focus on applying machine learning to assist in dimensionality reduction or identifying patterns that may not be readily apparent (Marsland, 2009). While the techniques themselves are multidisciplinary and challenging to master, they have demonstrated exceptional value and predictive efficacy in a number of disciplines.

As a supporting method, machine learning has been applied to a wide array of learning applications in conjunction with many of the techniques outlined above. One example of this is a popular application known as Support Vector Machines (Steinward and Christmann, 2008).[4] This machine learning classifier works similarly to other predictive modelling methods by making a binary class prediction for a given subject using non-linear functions fit the model across many dimensions. As a consequence, the predictions are highly accurate but extremely difficult to interpret, a common issue among machine learning applications. In support of the Army Research Laboratory, Charles River Analytics has applied Support Vector Machines to conduct automatic classification of training documentation according to their developmental categories (e.g., Bloom’s taxonomy) at a level of accuracy that rivals human

[2] Consider the message, "Great, I got a flat tire." Humans detect sarcasm in this sentence and understand the negative sentiment. Computers match the word "great" from a lexicon of positive words and interpret a positive sentiment.

[3] Closed captioning and transcript text can be analyzed using common NLP and TA methods. The results can be mapped to the video using Continuous Media Markup Language (CMML), which synchronizes the text to video by dividing the video into short segments with time stamps.

[4] For a concise overview, see: http://pages.cs.wisc.edu/~jerryzhu/cs540/handouts/hearst98-SVMtutorial.pdf

annotation. Other applications include those displayed in the Knowledge Discovery and Data-mining (KDD) Cup 2010, a data mining competition that ranked participants based on the accuracy of a successful prediction model for students in basic math course (e.g., Algebra I).[5] Many of the top scoring models utilized ensemble methods, techniques that weight the predictions of numerous weaker predictive models to generate an overall score that is more accurate than the individual predictions.

Big Data Tools and MapReduce

We have already alluded that data analytics with big data is a special case, though it is inextricably linked to the methods already discussed. Big data refers to data sets that are too large or complex to fit in the memory of typical computer workstations and laptops or through traditional relational database methods. The most commonly applied solution, invented by Jeffrey Dean and Sanjay Ghemawat (2004) at Google, is to break the big data into small chunks for processing with many computers in parallel--the “Map” function--then, “Reduce” the results from those parallel computations into a common set of results. MapReduce is only appropriate for computing tasks that can be mapped into chunks—what computer scientists call, "embarrassingly parallel problems."

Big data is commonly characterized by the volume, velocity, and variety of data collected from people's everyday activities and computer interactions, referred to as data exhaust. Big data will become more prevalent as the "internet of things" records even more of our everyday activities. The bulk of current big data stores have been collected from online interactions through social media and online shopping: the movies we watch, the books we purchase, the music we collect, where we take pictures, things we "Like" or share with our "followers." The scale of big data--the volume and velocity---comes primarily from automatic collection of online events. Variety of big data comes largely from aggregating the various data streams.

The decision to apply big data techniques to learning analytics requires an up-front decision to capture and store the interactions students have as part of their computer-based education and training, including their interactions and communications wherever possible. Big data collection and storage requires unique (though ubiquitous) computing resources. Collection mechanisms need to be written into the learning platform software. Storage and computation typically happen "in the cloud" to provide data surety and rapid analysis. A big data collection strategy might include traditional events like the web page visits, media downloads, forum post metadata, and so on. But, it could also include things like location tracking to discover whether students participate more from work and home, or from libraries and coffee shops. It could measure activity intensity, like whether students download their assignments iteratively over several days or all at once (and whether that is the day after class or the afternoon before class). The benefit of big data for instructors, institutions and policy-makers are manifold. It provides an ability to discover ways that students approach learning and a passive feedback mechanism to judge changes in policies at all levels.

5. DATA ANALYTICS USE CASES

For purposes of this research, an application model for evaluating use cases in the context of a learning system modules was developed as shown in Figure 4. Applications should be conceptualized in the context of formal, online learning to provide the most data-rich context for evaluating data analytic methods, though this is not intended to preclude application to informal or offline settings. The three overarching use cases explored are: (1) content development; (2) real-time analysis; and (3) post-evaluation. For the first use case, Content Development, the instructor or instructional designer is assumed to search within an existing repository for content to create a new online learning experience aligned with the desired learning outcomes best suited for the target audience. The desired content may be for an entire course, for a topic, or simply for an instructional activity.

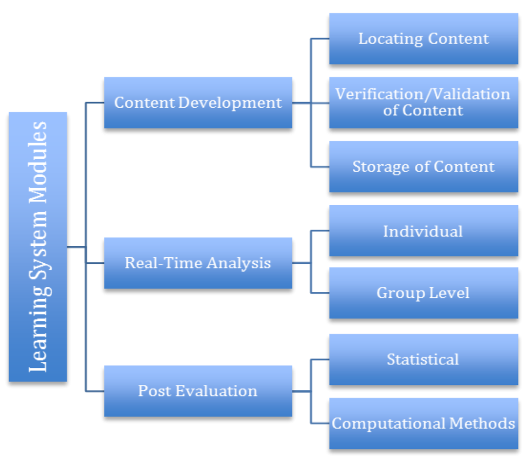

Figure 4. Data Analytics Application Model

[5] Many participants published papers available at: http://pslcdatashop.web.cmu.edu/KDDCup/workshop/

In the second use case, Real-Time Analysis, the system is performing analysis of learner performance based on responses and interactions to support the learner as well as the instructor. Real-time analysis could auto-generate remedial support for an at-risk student or could identify problematic or difficult areas within the course. The third use case, Post Evaluation, focuses on rolling-up individual and course insights developed during Real-Time Analysis across courses to provide organizational level insights.

This paper focuses on the Real-Time Analysis Use Case as an illustrative example because it could inherently include aspects of content development (providing additional resources) as well as post evaluation (since data collected in this use case will be used to support Meso-level applications and higher, refer to Figure 3). In this use case, the system is performing a real-time analysis of learner performance based on real-time responses and interactions. Multiple data analytic techniques could be used to support real-time analysis as indicated in Table 3. Data sources could include: amount of time online; time spent reviewing resources; tracking resources reviewed; tracking the sequence content was accessed; evaluating assignments such as submitted papers or blog/wiki postings; and social networking activities (emails and text chat usage). Real-time analysis can serve multiple purposes. For the individual learner, the analysis could identify learners at risk of poor performance. Identifying at-risk learners affords the opportunity to provide additional support and remediation as appropriate. For the instructor, the real-time analysis across the class could identify problematic or difficult areas within the course. Training modifications might include content re-organization or augmentation, updating learning objectives, or changing learning activities to improve learner performance.

Table 3. Data Analytics in the Context of Real-Time Analysis

Method	Use Case 2: Real-Time Analysis
Recommender Engine	• Identify additional content or resources to import into course • Locate other learners in network to reach out to and connect with
Machine Learning	• Refine content segmentation and classification • Develop learning topics based on current needs
Predictive Modeling	• Assess at-risk based on assignments, exams, individual exam questions • Alter course pace based on learner progress • Support intelligent tutoring and instructional intervention
NLP and text mining	• Analyze student assignments, identify plagiarism, support automated feedback • Evaluate communications for sentiment, e.g., boredom, frustration, engagement
Social Network Analysis	• Assess level of engagement, isolation, and influence • Support evaluation of instructor and student performance • Monitor peer relationships for requesting help
Similarity Grouping	• Understand learner population segments, composition, demographics • Cluster to identify support for populations and subgroups • Explore data to support other data analytic techniques
Big Data & MapReduce	• Could be utilized for highly variable data types • Would support real-time analysis for high-velocity, high-volume data

At a basic level, real-time analysis could be a simple time online to be compared to time online of a past successful learner. However, interpretation of the real-time analysis needs to be accomplished with caution. In the case of limited time online, remedial content would not necessarily resolve the problem unless the limited time online reflected the learner's avoidance of the course due to not understanding content. Alternatively, poor performance could be due to lack of motivation, time management skills, or self-regulation skills. These problems would also not be resolved through remedial content. Therefore, it is likely a mix of real-time analyses across multiple variables that will be required to not only assess and predict learner performance but to then provide the right solution. The focus of this use case is however on real-time analysis of learner performance and not determining the appropriate support to success.

For example, social network analysis could be used to assess engagement as measured by a learner's level of activity in terms of volume (count) and frequency (timeline) of contributions to online activities such as forums and wiki edits. For one example of how SNA is combined with other data analytic technologies to support online education, consider the Wikispaces website (wikispaces.com) and their Engagement dashboard element in Figure 5. This gauge shows each classroom participant, their relative connectivity, and their level of activity (forum

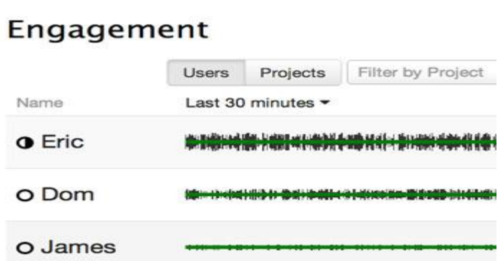

Figure 5. Learner Engagement Dashboard

contributions, wiki edits, etc.) expressed as volume and frequency. The SNA metric of connectedness informs the viewer whether a student's activity is "just noise" or if fellow students are listening. Exploring these additional activities through sentiment analysis could also determine if the initial posting was perceived as positive or negative. The horizontal timeline indicates the frequency (horizontal axis) and volume (vertical axis) of the learner's activity over the last 30 minutes, in this case. This type of dashboard could be used to integrate the results of multiple data analytic techniques to provide a more comprehensive real-time analysis and interpretation tool for both the learner and the instructor

As noted previously, real-time analysis could indicate the course may need additional content to support difficult content. These supplemental instructional materials could be provided across a range of automated to completely author-defined methods. Recommender systems, such as the ones used by Amazon and Netflix, could be utilized to automatically suggest materials for remediation. In this case, we assume instructor mediated provision of materials. For example, the instructor may search a learning repository through a dashboard relying on different criteria to identify and assess the appropriateness of the retrieved content as shown in Figure 6. Appropriateness might be: based on learner needs (relevance and context); based on course content needs (content type, learning difficulty, learning domain, subject domain, and learning level); and based on instructional needs (quality, instructional activity, duration, and format). These three themes representing the learner, the content, and the instruction are the important elements in the instructional design of a course (see for example, Anderson, Krathwohl, Airasian, Cruikshank, Mayer, Pintrich, Raths, & Wittrock, 2001, Gagne, Briggs, & Wager, 1992; Merrill, 1994; Morrison, Ross, & Kemp, 2007).

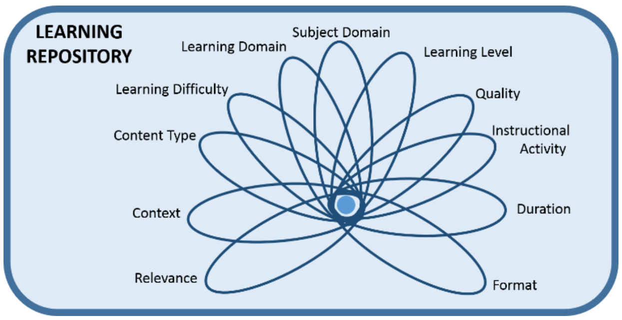

Figure 6. Content Evaluation Criteria

Based on instructor input for the criteria, the system may return multiple options. Therefore, a systematic presentation may afford a more effective and efficient means to select the most appropriate content. As shown in Figure 7, each circle on the graph represents a potential content that the system has found based on instructor input parameters. Pop-up tips from a rollover interaction could summarize the data according to the criteria while selecting (clicking) the data point could reveal additional information such as when the content was uploaded to the repository and by whom. In this conceptualized dashboard, the instructor could select three variables to review. All data would initially be presented on the graph but the instructor could then filter the data to refine the selection. For example, the instructor could select didactic or role-play or whatever the available instructional activity types were for the system-identified potential content. The instructor could vary selections until a reasonable amount of recommended content could be reviewed.

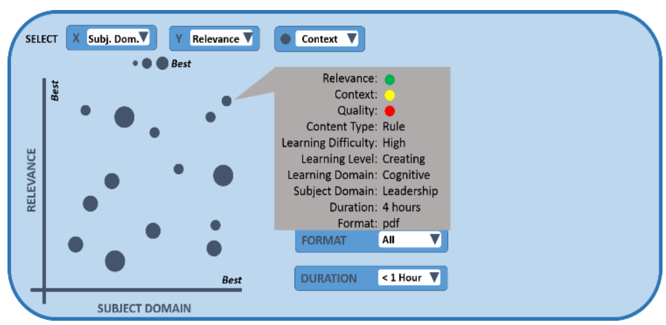

Figure 7. Content Selection Interface with Pop-up Tip

These dashboard views provide a foundation for developing advanced data analytics. For Real-Time analysis, the focus should always be on supporting student learning, i.e., efficient and effective attainment of learning objectives. For example, while these views are described in the context of instructor-student interaction, they also provide a workflow foundation and conceptual basis for automated remediation and support and combination of distinct methods. For example, predictive modeling can be utilized in this context to identify students in most need of personalized intervention based on their risk of failure. Similarity grouping can be utilized to recommend study groups of students with like characteristics and backgrounds. NLP and TA can be overlaid to identify and focus on levels of frustration on message boards or other social media around specific topics. These examples provide an illustrative use case in the context of Real-Time analysis, but can easily be extended into Content Development prior to instruction and Post Evaluation after completion.

6. CONCLUSIONS

This paper summarized the fields of learning analytics and educational data mining that are relevant to learning systems. A framework to evaluate data analytic techniques for learning systems was developed and explored through the identification of three use cases. The use cases focused on content development, real-time analysis, and post evaluation. Expanding on the Real-Time Analysis Use Case, several applicable data analytic techniques were identified and discussed in terms of their application to learning analytics. A potential instructor interface was also presented to support an efficient content selection process.

Each of these techniques and applications serve as examples of the potential value of data analytics to support Army learning. Recent military operations in Iraq and Afghanistan have again demonstrated that the Army's ability to learn and adapt are critical to success, particularly in the Irregular Warfare environments that today's operations consist of (for an overview of this literature see Nagl, 2002). The ability of the organization to succeed ultimately rests on the back of the soldiers that execute the mission, and the training and education they receive. The growing pace of change in technology only makes it more imperative to apply the best methods and techniques available to help prepare our soldiers to go into harm's way. In addition, our adversaries are not standing still. As presented in Peter Senge's The Fifth Discipline (1990): "The only sustainable competitive advantage is an organization's ability to learn faster than the competition." This paper has described the state of the art that is available for the Army and its educational institutions to adapt to learning practices. The challenges next steps are taking the art of the possible and adapting the most relevant and beneficial practices to Army learning challenges. The risk of falling behind in a chaotic global environment makes success an imperative.

ACKNOWLEDGEMENTS

The authors would like to thanks Janet Weisenford and Paul Cummings of ICF International for their inputs and critiques, as well as for helping to make the scope of data analytics in a learning context more manageable. The authors would also like to thank Amy Bisker of ICF International for her assistance with editing, formatting, citations, and revisions.

REFERENCES

Abukhousa, E., & Atif, Y. (2014). Big Learning Data Analytics Support for Engineering Career Readiness. 2014 International Conference on Interactive Collaborative Learning (ICL), (December), 663–668.

Brooks, C., Liu, W., Hansen, C., McCalla, G., & Greer, J. (2007). Making Sense of Complex Learner Data. In Assessment of Group and Individual Learning through Intelligent Visualization Workshop (AGILeViz). Suppementary Proceedings of the 13th International Conference of Artificial Intelligence in Education. (pp. 1–42).

Campbell, J. P., DeBlois, P. B., & Oblinger, D. (2007). Academic analytics: A new tool for a new era. EDUCAUSE Review, 42(4), 40–57.

Chambers, M., & Dinsmore, T. (2014). Advanced Analytics Methodologies: Driving Business Value with Analytics. Pearsons FT Press.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. International Journal of Technology Enhanced Learning, 15(3), 58-76.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon. In *Fourteenth Conference on Computational Natural Language Learning* (pp. 107–116). Retrieved from http://eprints.pascal-network.org/archive/00007069/

Dawson, S. (2009). Seeing the learning community. *British Journal of Educational Technology*, *41*(5), 736–752.

Dawson, S., Bakharia, A., & Heathcote, E. (2010). SNAPP: Realising the affordances of real-time SNA within networked learning environments. *Proceedings of the 7th International Conference on Networked Learning*, 125–133. Retrieved from http://www.lancs.ac.uk/fss/organisations/netlc/past/nlc2010/abstracts/PDFs/Dawson.pdf

Dawson, S., Gašević, D., Siemens, G., & Joksimovic, S. (2014). Current state and future trends: A citation network analysis of the learning analytics field. Proceedings of the 3rd International Conference on Learning Analytics and Knowledge (LAK '14) ACM, New York. Retrieved from https://files.semtech.athabascau.ca/public/papers/lak_ca14.pdf

Dean, J., & Ghemawat, S., (2004). MapReduce: Simplifed Data Processing on Large Clusters. Google, Inc.

El-Bishouty, M. M., Chang, T.-W., Graf, S., & Chen, N.-S. (2014). Smart e-course recommender based on learning styles. *Journal of Computers in Education*, *1*, 99–111. http://doi.org/10.1007/s40692-014-0003-0

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. International Journal of Technology Enhanced Learning, 4(5/6) pp. 304–317.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. Tech Trends, 59(1), 64-71.

Govindarajan, K., Somasundaram, T. S., Kumar, V. S., & Kinshuk. (2013). Continuous clustering in big data learning analytics. Proceedings - 2013 IEEE 5th International Conference on Technology for Education, T4E 2013, 61–64. http://doi.org/10.1109/T4E.2013.23

Greenough, J. (2015, June 15). The self-driving car report: Forecasts, tech timelines, and the benefits and barriers that will impact adoption. Business Insider Intelligence. Retrieved from: http://www.businessinsider.com/report-10-million-self-driving-cars-will-be-on-the-road-by-2020-2015-5

Hung, J., & Zhang, K. (2003). Data Mining Applications to Online Learning, 1–8.

Hutzler, D., David, E., Avigal, M., & Azoulay, R. (2014). Learning Methods for Rating the Difficulty of Reading Comprehension Questions. 2014 IEEE International Conference on Software Science, Technology and Engineering, 54–62. http://doi.org/10.1109/SWSTE.2014.16

Iam-on, N. (2014). Using Cluster Ensemble to Improve Classification of Student Dropout in Thai University, 452–457.

Ifenthaler, D., & Widanapathirana, C. (2014). Development and validation of a learning analytics framework: Two case studies using support vector machines. *Technology, Knowledge and Learning*, *19*, 221–240. http://doi.org/10.1007/s10758-014-9226-4

Imran, H., Belghis-Zadeh, M., Chang, T.-W., Kinshuk, & Graf, S. (2015). Emerging Issues in Smart Learning. Lecture Notes in Educational Technology, 337–344. http://doi.org/10.1007/978-3-662-44188-6

Johnson, L., Adams Becker, S., Cummins, M., Estrada, V., Freeman, A., and Ludgate, H. (2013). NMC Horizon Report: 2013 Higher Education Edition. Austin, Texas: The New Media Consortium.

Kongsakun, K. A. N. (2013). An Improved Recommendation Model Using Linear Regression and Clustering for a Private University in Thailand. Proceedings of the 2013 International Conference on Machine Learning and Cybernetics, 14–17.

Kumar, V., Nesbit, J., Winnie, P., Hadwin, A., Jamieson-Noel, D. Han, K. (2007). Quality Rating and Recommendation of Learning Objects. E-Learning Networked Environments and Architectures, Advanced Information and Knowledge Processing, 337–373.

Kumar, S. P., & Ramaswami, K. S. (2010). Fuzzy K- means cluster validation for institutional quality assessment. 2010 International Conference on Communication and Computational Intelligence (INCOCCI), 628–635.

Larrusson, J., & White, Brandon, editors, (2014). Learning Analytics from Research to Practice. Springer.

Lü, L., & Sindhwani, V. (2012). Recommender systems. Encyclopedia of Machine Learning, 1–9. Retrieved from http://dl.acm.org/citation.cfm?id=245121

Markoff, J. (2012, June 25). How many computer to identify a cat? 16,000. The New York Times. Retrieved from: http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html?_r=0

Marsland, S. (2009). Machine Learning: An Algorithmic Perspective. Chapman & Hall/CRC, Cambridge. http://cogsci.fmph.uniba.sk/~farkas/courses/CompInt/marsland.ml-alg-perspect.09.pdf

Martens, D., & Provost, F. (2011). Pseudo-social network targeting from consumer transaction data (No. 11-05). Working Paper. New York. Retrieved from http://archive.nyu.edu/handle/2451/31253

McDonald, B., & Noakes, N. (2005). Breaking down learner isolation: How social network analysis informs design and facilitation for online learning. In AERA, Montreal, (pp. 1–30). Montreal. Retrieved from http://cpsquare.org/wp-content/uploads/2008/07/stuckey-etal-aera-sna.pdf

Nagl, J. (2002). Counterinsurgency Lessons from Malaya and Vietnam: Learning to Eat Soup with a Knife. Praeger Publishers, United States.

Niemann, K., & Wolpers, M. (2013). Usage context-boosted filtering for recommender systems in TEL. Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8095 LNCS, 246–259. http://doi.org/10.1007/978-3-642-40814-4_20

O'Neil, C., & Schutt, R. (2014). Doing Data Science. O'Reilly Media, United States.

Papamitsiou, Z., & Economides, A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Educational Technology & Society, 17 (4), 49–64.

Peña-Ayala, A., editor (2014). Educational Data Mining Applications and Trends. Springer.

Potok, T. E., & Palathingal, P. (2005). Document clustering using particle swarm optimization. Proceedings 2005 IEEE Swarm Intelligence Symposium, 2005. SIS 2005, 185–191. http://doi.org/10.1109/SIS.2005.1501621

Provost, F., Dalessandro, B., & Hook, R. (2009). Audience Selection for On-line Brand Advertising: Privacy-friendly Social Network Targeting. Brand, 707–715. http://doi.org/10.1145/1557019.1557098

Provost, F., Fawcett, T. (2013). Data Science for Business. O'Reilly Media, United States.

Rose, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2008). Analyzing Collaborative Learning Processes Automatically: Exploiting the Advances of Computational Linguistics in Computer-Supported Collaborative Learning. *International Journal of Computer-Supported Collaborative Learning*, *3*(3), 237–271. http://doi.org/10.1007/s11412-007-9034-0

Scheffel, M., Drachsler, H., Stoyanov S., & Specht, M. (2014). Quality indicators for learning analytics. Educational Technology & Society, 17 (4), 117–132.

Senge, P. (1990). The Fifth Discipline: The Art and Practice of the Learning Organization. Random House: United States.

Siemens, G. (2012). Learning Analytics: Envisioning a Research Discipline and a Domain of Practice, (May).

Siemens, G., & Baker, R. S. J. d. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 252-254). ACM. Vancouver, British Columbia, Canada

Siemens, G., Gašević, D., Haythornthwaite, C., Dawson, S., Buckingham, S., Ferguson, R., Duval, E., Verbert, K., & Baker, S. J. d. (2011). Open Learning Analytics: An integrated & modularized platform. Society for Learning Analytics Research, Open Learning Analytics. Retrieved from http://solaresearch.org/initiatives/ola/

Smith, V. C., & Lange, A. (2010). Predictive Modeling to Forecast Student Outcomes and Drive Effective Interventions. Council for the Study of Community Colleges, 16(3), 51–61.

Steinward, I., & Christmann, A. (2008). Support Vector Machines. Springer.

Valsamidis, S., Kontogiannis, S., Kazanidis, I., Theodosiou, T., & Karakos, A. (2012). A Clustering Methodology of Web Log Data for Learning Management Systems. Educational Technology & Society, 15, 154–167.

Wang, A. Y., Newlin, M. H., & Tucker, T. L. (2001). A discourse analysis of online classroom chats: Predictors of cyber-student performance. *Teaching of Psychology*, *28*(3), 222–226. http://doi.org/10.1207/S15328023TOP2803_09

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. PNAS, 112(4), 1036–1040. http://doi.org/10.1073/pnas.1418680112