International Journal for the Scholarship of Teaching and Learning

Volume 4 | Number 2 Article 17

7-2010

Academic Analytics and Data Mining in Higher Education

Paul Baepler
University of Minnesota, baepl001@umn.edu

Cynthia James Murdoch
University of Minnesota, macal001@umn.edu

Recommended Citation
Baepler, Paul and Murdoch, Cynthia James (2010) "Academic Analytics and Data Mining in Higher Education," *International Journal for the Scholarship of Teaching and Learning*: Vol. 4: No. 2, Article 17.
Available at: https://doi.org/10.20429/ijsotl.2010.040217

Academic Analytics and Data Mining in Higher Education

Abstract
The emerging fields of academic analytics and educational data mining are rapidly producing new possibilities for gathering, analyzing, and presenting student data. Faculty might soon be able to use these new data sources as guides for course redesign and as evidence for implementing new assessments and lines of communication between instructors and students. This essay links the concepts of academic analytics, data mining in higher education, and course management system audits and suggests how these techniques and the data they produce might be useful to those who practice the scholarship of teaching and learning.

Keywords
Analytics, SoTL, Research, Data, Action, Visualization

Academic Analytics and Data Mining in Higher Education

Paul Baepler University of Minnesota Minneapolis,
Minnesota, USA
baepl001@umn.edu

Cynthia James Murdoch
University of Minnesota
Minneapolis, Minnesota, USA
macal001@umn.edu

Abstract

The emerging fields of academic analytics and educational data mining are rapidly producing new possibilities for gathering, analyzing, and presenting student data. Faculty might soon be able to use these new data sources as guides for course redesign and as evidence for implementing new assessments and lines of communication between instructors and students. This essay links the concepts of academic analytics, data mining in higher education, and course management system audits and suggests how these techniques and the data they produce might be useful to those who practice the scholarship of teaching and learning.

Keywords: analytics, SoTL, research, data, action, visualization

Introduction

Academic analytics is a new field that has emerged in higher education in the aftermath of the widespread use of data mining practices and "business intelligence tools" in business and marketing. It can refer broadly to data-driven decision making practices for operational purposes at the university or college level, but it also can be applied to student teaching and learning issues. For example, information culled from a course management system (CMS) can be quickly assessed for indicators of student failure, and early alerts can be sent to faculty or students to warn them of poor performance (Campbell, DeBlois, & Oblinger, 2007). The phrase “academic analytics” entered teaching conversations in 2005 but had previously been coined by the WebCT company (now Blackboard) to describe the data collection functions the CMS enabled. Today, much of the focus on academic analytics is on the *actions* that can be taken with real-time data reporting and with predictive modeling which helps suggest likely outcomes from familiar patterns of behavior. Given the desire for and usefulness of such information, perhaps it comes as no surprise that the phrase “action analytics” has emerged (Norris, Baer, Leonard, Pugliese, & Lefrere, 2008). Such phrasing is a strategic choice that points to the need for early student data, prompt analysis, and immediate access by students, faculty, and advisors who can make smart choices to influence learning. It also hearkens to a long history of “action research” and the imperatives of the scholarship of teaching and learning (SoTL) to influence change within the classroom and among learners (Cook, Wright, O’Neal, 2007).

https://doi.org/10.20429/ijsotl.2010.040217

Both academic analytics and data mining have emerged in the wake of higher education's ability to capture an increasing volume of data. The concept of data mining, upon which academic analytics is built, has existed in business for decades, but data mining in higher education surged around 1995, at the advent of the Internet. A greater interest in "educational data mining" emerged around 2004, and researchers in this field are closely aligned with analysts in the "Intelligent Tutoring System" and the "Artificial Intelligence in Education" communities (Winters, 2006). A distinction can be drawn, however, between academic analytics and data mining. Academic analytics is often thought of as hypothesis-driven, using a particular dataset to solve a practical academic problem, such as increasing student retention levels. Data mining, to continue the mineralogical metaphor, is thought of more as a kind of speculative prospecting for riches. A large field of data might unearth all kinds of insensible information that, when manipulated with data mining techniques, might present some useful insights. Researchers use data mining techniques to sift through data for implicit affinities and hidden patterns without a preconceived hypothesis. They wait for patterns to emerge.

Analytics is associated with a scientific, hypothesis-driven approach, while data mining has a legacy with strategic business techniques and marketing. The fact that the latter method—data mining—typically lacks a hypothesis to drive an investigation can seem troubling, but it's a distinction that might be rendered immaterial when it produces insights. That is to say, if a model works, even if one does not understand exactly how it works, the results may still be valuable even if they lack an originating hypothesis. A similar debate has already been waged—or some would say is still ongoing—in the life sciences, particularly in fields like genomics. In these fields, where, for instance, analysis of extremely complex gene sequences can lead to useful therapies, it may be more important to find a viable cure for a damaging disease than to fully comprehend why a protocol works. That is, "A test may be useful if patterns reliably discriminate, regardless of whether they can be understood and explained. Although ultimately important, explanation may be saved for later" (Ransohoff, 2004, p. 1028). Thus data mining may not provide causality, but in many instances, correlation might still yield interesting and powerful results. Applied to higher education, this might mean noticing a particular behavior in a CMS—for example, a student's posting more than "x" number of times in an online forum leads to a "y" gain in that student's final grades. The mechanism for this improvement may be purely speculative—greater student engagement, perhaps—but the results may encourage instructors to continue or initiate such online discussions.

In addition to academic analytics and data mining, CMS audits can provide useful insights into student behavior online. Institutions spend millions purchasing, implementing, and supporting commercial and open source CMSs, but little work has been done to analyze how these tools are actually used. Is the CMS largely an institutional lockbox for syllabi, grades, and course readings, or does it spur knowledge construction through meaningful peer-to-peer discussion and assessment? Retrospectively auditing CMS data—hardly the urgent act implied in *action* analytics—could nevertheless yield useful data about how to engage faculty in course redesign.

These three data gathering efforts—academic analytics, data mining, and CMS audits—point to a future of new evidence that can influence instructional decisions

and could serve as the basis of a robust SoTL agenda. The availability of early student assessment data can affect student motivation and retention. The possibility of creative visual representation of student interactions may point to otherwise difficult to envision online discussion patterns. The evidence provided by a thorough CMS audit might support shifts in faculty development initiatives that could emphasize the use of web 2.0 tools within or even outside of a CMS (Ajjan & Hartshorne, 2009). Although these attempts to quantify a small fraction of the student learning experience are still in their foundational stages, they present important opportunities for faculty developers who are not only interested in SoTL, but who would also find these data useful departure points to initiate discussions about assessment, student engagement, and evidence-based course redesign.

Academic Analytics

Academic analytics combines select institutional data, statistical analysis, and predictive modeling to create intelligence upon which students, instructors, or administrators can change academic behavior. The University System of Georgia undertook an early experiment using analytic techniques to develop an algorithm that could predict student completion and withdrawal rates in an online environment. Their results helped to confirm that it was possible, in this instance, to predict with up to 74% accuracy the likelihood that a student would successfully complete an online course. Both high school GPA and the SAT quantitative measure were demonstrated to be related to retention in online courses, and with additional information on locus of motivation (internal or external), and financial aid, the researchers were able to correctly classify student dropout (60%) and student completion (76%) (Morris, Wu, Finnegan 2005). By some definitions, this experiment would be an early example of academic analytics, though currently, more emphasis is being placed on "actionable intelligence," information that can be delivered early enough to make a difference in academic performance.

More recently, John Campbell and Kimberly Arnold at Purdue University have begun to move beyond the research and pilot phases of an exciting analytics project (Campbell, 2007; Iten, Arnold, & Pistilli, 2008; Arnold 2010). Building on decades of research showing that early and frequent assessment is not only a best practice but also a method for changing the studying habits of underperforming students in introductory courses, the team developed an early academic alert system. *Signals*, the new academic warning system, is integrated into the Blackboard CMS and draws from 20 discrete datapoints. Students log in and are presented with a simple panel of red, yellow, and green lights, indicating intuitively whether or not the student seems to be doing well. The *Signals* algorithm examines both academic performance data, such as quiz and test grades and evidence of student effort. For instance, depending upon the course structure, the program might factor in time spent reading online course material or performing practice assignments to gauge effort and motivation. The ability to splice together these different types of data, weight their predictive relevance accordingly, and present the information back to the student in a quick and simple manner makes this a valuable tool, particularly for early career students who may have trouble acclimating to a challenging academic workload. In Fall 2009, over 11,000 students were enrolled in these vast gateway courses, many of them in the STEM disciplines where there are many standardized exams and many quantifiable measures of student progress (Tally, 2009).

As analytic efforts like Purdue's unfold, and should these enterprise-scale implementations succeed to the same degree as their pilots, new opportunities for course redesign might arise. We have known that frequent and early assessment has an impact, but when we can express that impact in terms of student retention, a very recognizable value, it may be easier to encourage colleagues to redesign courses and integrate more formative assessment and supplemental instruction. While instructors have always been able to create more quizzes and follow-up study aids, they are generally hesitant to spend significant time creating teaching materials when there is little evidence that they will be used. Students need rapid feedback, and struggling students, in particular, need to be directed to salient content to help them redress their misconceptions and complete courses more successfully. Academic analytics can provide SoTL practitioners with a vital link between instruction, assessment, and student effort.

Data Mining in Higher Education

While there are many data mining techniques, most of the work that has been done in higher education falls into the categories of clustering, classification, visualization, and association analysis (Castro, Vellido, Nebot, & Mugica, 2007; McGrath, 2008; Romero & Ventura, 2007). This work remains largely exploratory, pointing to the potential of these forms of analyses more than their current application, yet their early findings show promise. As such, two examples stand in here for a growing number of experiments.

Researchers at the University of Florida aggregated all student activity in a graduate course with 67 students as expressed in the Moodle CMS data logs to see if this activity was predictive of a student's sense of community (Black, Dawson, & Priem, 2008). All students took the validated Classroom Community Scale (CSS) at the end of the course, and these scores were paired with each student's cumulative user log (the sum of all their "clicks" within the system). Although there were several significant limitations to the study, the researchers concluded that the total number of log entries or user events was a valid predictor of a sense of community within the course. Such information, should it prove valid through additional studies, might eventually help instructors unobtrusively construct a more accurate representation of, and tailored content for, their online students. Overall, the experiment suggests that it may be possible to measure some affective attributes among students with simple data logs. This development could augment data from student surveys or reduce the need to administer so many questionnaires at a time when survey fatigue has become a concern.

When instructors assign an online discussion, they tend to assess the forum based on the number and length of individual posts, or use a rubric to quickly assign a few points to each thread. These assessments become increasingly difficult as class sizes or the number of sections increases. And after scrutinizing every post, it is difficult to have a sense of how well the entire discussion has developed. Researchers at the University of Auckland, however, have begun to experiment with new automated modeling tools that create network maps of online discussions. Each map can be played as an animation or be viewed at a glance as a static image.

With a snapshot of the discussion's architecture—essentially a picture of how a dialogue developed –it becomes much more efficient to see how new threads emerge, whether they begin from an instructor's post or through an individual student's initiative. These maps also make it easier to pinpoint discussion leaders as posts start with or return to particular students. Essentially representing the timestamp and thread of a network of posts graphically, this potential alternative to examining posts individually, may provide instructors with a key tool for delivering a group discussion grade based on student-to-student interactions. Additionally, this kind of analysis could alter the assessment of individual students, automatically representing the answers to such question as: "When did the student make postings?" "Did the student respond to postings of other students?" "How immediate were those responses?" and "Did other students respond to this student?" (Dringus & Ellis, 2005; Kim, Chern, Feng, Shaw, & Hovy, 2006).

Open source CMSs present some of the best chances to analyze individual courses with data mining techniques. Researchers at Cordoba University in Spain, for instance, have developed some experimental data mining tools that are integrated directly into Moodle (Romero, Ventura, & Garcia, 2008). These tools, themselves built in an open source framework, KEEL, allow course designers to perform a series of analyses on a course or collection of courses. The tools primarily use classification tasks that involve decision trees, rule induction, neural networks, and statistical inference. While this tool is as yet unproven, classification data mining techniques have already been used to

- select student groups with similar characteristics and reactions to learning strategies;
- detect student misuse and lurking;
- identify students who, in multiple choice tests, are hint-driven or failure-driven in order to find common misconceptions;
- locate students who exhibit low motivation and find alternate means of reaching them; and,
- predict probable student outcomes.

One of the goals of this team's work is to make the data mining tools easy enough to be used by individual instructors so they can analyze their own courses. A secondary goal would be to make the tools sufficiently transparent so that students could analyze their own usage data.

Although not specifically for data mining per se, the Visual Understanding Environment (VUE), developed by Tufts University (http://vue.tufts.edu), already provides SoTL researchers with an intriguing set of visualization tools. VUE is essentially a concept mapping application, but it has the capability to import datasets and rss feeds and represent the data as a web of nodes and links, ostensibly a network diagram. (It also has a set of analysis tools and integrates with other applications like Zotero (a reference manager) and SEASR (a Mellon Foundation digital humanities tool). VUE is freely available and easy to use, and it could eventually help SoTL researchers explore formal and informal learning in social networks. (Before this can happen easily, though, academic technologists will still need to create an easy way to flow data out of a CMS, and, of course, researchers will need to seek proper IRB approval to work with it. When performed covertly, data mining, as Helen Nissenbaum reminds us, can rightly provoke a sense of

"privacy under assault" (44, Nissenbaum)). Of course, all of these analytical tools would be most practical when an instructor makes extensive use of Moodle or another CMS, perhaps in a fully-online or blended course.

Course Management System Auditing

Until recently, there have been very few studies that have attempted to gauge CMS use broadly on a campus (West, Waddoups, Kennedy, & Graham, 2007). Previous work has been devoted to large-scale faculty and student satisfaction surveys (Educause Center for Applied Research [ECAR], 2005) or more localized examinations of individual courses. In the latter case, the amount of data or number of measurable student actions is so small that it is difficult to generalize from the findings. What has been needed for course redesign purposes are more comprehensive studies and research that combine survey information with actual student user data so that perceptions of use can be matched with actual usage statistics.

Recently, over 36 million student and instructor events within the Blackboard CMS were analyzed in an institution-wide study at Brigham Young University (Griffiths & Graham, 2009). The investigation revealed that as much as 90% of student CMS use concentrated on just six tools, and much of this use was simple content delivery using the announcement and content tools. Concluding that the study pointed to a possible need for more targeted faculty training in the lesser-used tools, the authors also entertained the idea that a CMS at BYU may simply need to deliver the limited functionality that this subset of applications provides. In addition to generating a baseline of data for future research, they also confirmed the need to examine individual colleges, many of which exhibited distinct patterns of use in the CMS. These patterns, in turn, suggest that institutions may have signature pedagogical styles and discipline-specific technological needs (Gurung, Chick, & Ciccone, 2009).

A second large study, this one conducted at the University of Michigan to examine two years of user log data in a CMS as well as faculty and student surveys, sought to determine whether what faculty and students reported they valued was consistent with their actual use of a CMS (Lonn & Teasley, 2009). The researchers discovered that document management and broadcast-oriented communication tools were heavily used (95% of all user actions) and highly valued. Simultaneously, tools that are more interactive are rarely used and much more likely to be rated by both students and instructors as "not valuable." When specifically asked if information technology (IT) improves teaching and learning, instructors were more likely than students to agree. Students instead tended to value efficiency over more interactive tools such as chat, discussion, and wiki. The authors conclude, "As long as students fail to see the relevance of interactive tools for their learning or for instructors' teaching, they are likely to continue to view IT as merely a quick and accessible means to retrieve course documents and get messages from instructors." They suggest that because the interactive tools are still relatively new, instructors might gain from training that helps them design activities that facilitate peer evaluation and student questioning. CMS audits combined with large scale surveys can provide strong evidence and guidance for directing faculty in a manner that more closely aligns instructor pedagogical beliefs with skills that go beyond simple document management and broadcast communication. They also give us a baseline for

determining important gaps in our knowledge about how students and instructors use these large learning systems.

Conclusion

Kathleen McKinney (2007) outlines seven challenges and opportunities for the future of SoTL. She augmented this perceptive vision with an amalgamation of guidance from others who largely agreed that SoTL, to be truly effective, needed to influence curricular advancement. SoTL should precipitate action and enable instructional choices for both the student and the faculty. Academic analytics, educational data mining, and CMS audits, although in their incipient stages, can begin to sift through the noise and provide SoTL researchers with a new set of tools to understand and act on a growing stream of useful data.

References

Ajjan, H., & Hartshorne, R. (2009). Investigating faculty decisions to adopt Web 2.0 technologies: Theory and empirical tests. *Internet and Higher Education*, *11*(2), 71-80. Retrieved from http://www.elsevier.com/wps/find/journaldescription.cws_home/620187/descriptio#description

Arnold, K. E. (2010). Signals: Applying academic analytics. *Educause Quarterly, 33* (1) Retrieved from http://www.educause.edu/EDUCAUSE+Quarterly/EDUCAUSEQuarterlyMagazineVolum/SignalsApplyingAcademicAnalyti/199385

Black, E. W., Dawson, K., & Priem, J. (2008). Data for free: Using LMS activity logs to measure community in online courses. *Internet and Higher Education*, *11*(2), 65-70. http://www.elsevier.com/wps/find/journaldescription.cws_home/620187/description#description

Campbell, J. P. (2007). *Utilizing student data within the course management system to determine undergraduate student academic success: An exploratory study*. Unpublished doctoral dissertation, Purdue University.

Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. *Educause Review*, *42* (4), 40–42, 44, 46, 48, 50, 52.

Castro, F., Vellido, A., Nebot, A., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. *Studies in Computational Intelligence*, *62*, 183-221.

Cook, C. E., Wright, M., O'Neal, C. (2007). Action research for instructional improvement: Using data to enhance student learning at your institution. *To Improve the Academy, 25*, 123-138.

Dringus, L. P., & Ellis, T. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. *Computers & Education*, *45*, 141-160.

Educause Center for Applied Research. (2005). The promise and performance of course management systems in *ECAR Study of Students and Information Technology, 2005: Convenience, Connection, Control, and Learning*. Retrieved from http://www.educause.edu/ers0506.

Griffiths, M. E., & Graham, C. R. (2009). Patterns of user activity in the different features of the Blackboard CMS across all courses for an academic year at Brigham Young University. *MERLOT Journal of Online Learning and Teaching*, *5*(2). Retrieved from http://jolt.merlot.org/vol5no2/griffiths_0609.htm.

Gurung, R., Chick, N., & Haynie, A. (2009). *Exploring signature pedagogies: Approaches to teaching disciplinary habits of mind*. Sterling, VA: Stylus.

Iten, L., Arnold, K., & Pistilli, M. (2008). Mining real-time data to improve student success in a gateway course. Eleventh Annual TLT Conference, Purdue University, March 4.

Kim, J., Chern, G., Feng, D., Shaw, E., & Hovy, E. (2006). Mining and assessing discussions on the web through speech act analysis. *Proceedings of the International Semantic Web Conference (ISWC)06 Workshop on Web Content Mining with Human Language Technologies*.

Lonn, S., & Teasley, S. D. (2009). Saving time or innovating practice: Investigating perceptions and uses of Learning Management Systems. *Computers & Education*, *53*(3). 686-694.

McKinney, K. (2007). *Enhancing learning through the scholarship of teaching and learning: The challenges and joys of juggling*. Boston: Anker Publishing.

Morris, L. V., Wu, S., & Finnegan, C. (2005). Predicting retention in online general education courses. *The American Journal of Distance Education, 19*(1), 23-36.

Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford, CA: Stanford Law Books.

Norris, D., Baer, L., Leonard, J., Pugliese, L., & Lefrere, P. (2008). Action analytics. *Educause Review*, *43*(1), 42-67.

Ransohoff, D. F. (2004). Evaluating discovery-based research: When biologic reasoning cannot work. *Gastroenterology*,*127*, 1028.

Romero, C., Ventura, S., & Garcia, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368-384.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*, 135-146.

Tally, S. (2009, September 1). Signals tells students how they're doing even before the test. *Purdue University University News Service*. Retrieved from http://www.purdue.edu/newsroom/students/2009/aug/story-print-deploy layout_1_1336_1336.html

West, R. E., Waddoups, G., Kennedy, M. M., & Graham, C. R. (2007). Evaluating the impact from implementing a Course Management System. *Instructional Journal of Educational Technology & Distance Learning.* 4(2) Retrieved from http://www.itdl.org/Journal/Feb_07/article01.htm

Winters, T. (2006). *Educational data mining: Collection and analysis of score matrices for outcomes-based assessment*. Unpublished doctoral dissertation, University of California Riverside.