IRIE
International Review of Information Ethics Vol. (07/2014)

Jeffrey Alan Johnson:

The Ethics of Big Data in Higher Education

Abstract:

Data mining and predictive analytics—collectively referred to as "big data"—are increasingly used in higher education to classify students and predict student behavior. But while the potential benefits of such techniques are significant, realizing them presents a range of ethical and social challenges. The immediate challenge considers the extent to which data mining's outcomes are themselves ethical with respect to both individuals and institutions. A deep challenge, not readily apparent to institutional researchers or administrators, considers the implications of uncritical understanding of the scientific basis of data mining. These challenges can be met by understanding data mining as part of a value-laden nexus of problems, models, and interventions; by protecting the contextual integrity of information flows; and by ensuring both the scientific and normative validity of data mining applications.

Agenda:

Author:

Jeffrey Alan Johnson, Ph.D.:

- Assistant Director of Institutional Effectiveness and Planning, Utah Valley University, 800 W. University Pkwy., Orem, UT 84058, United States
- ☎ +1 801 696 5088, ✉ jeffrey.johnson@uvu.edu
- Relevant publications:
  - "From Open Data and Data Privacy to Data Justice," Midwest Political Science Association Annual Meeting, April 2013.
  - "The Illiberal Culture of E-democracy," *Journal of E-Government* 3:4 (Spring 2007), 85-111.

Big Data in Higher Education

Data mining and predictive analytics are increasingly used in higher education to classify students and predict student behavior. Institutions of higher education, in some cases working with commercial providers, have begun to use these methods to recommend courses, monitor student progress, individualize curriculum, and even build personal networks among students. Data mining, as a major part of business intelligence, is held to be part of a radically different future for higher education in which the ability to predict individual outcomes revolutionizes management and allows institutions to better understand their students and their needs by taking advantage of the vast trove of data that institutions generate in their operations.[1]

These techniques encompass practices and methods that are quite different from and present different challenges to its users than do inferential research methods—often called "academic analytics."[2] There are four key differences:

1. Data mining eschews the hypothetico-deductive process, relying instead on a strictly inductive process in which the model is developed a posteriori from the data itself.

2. Data mining relies heavily on machine learning and artificial intelligence approaches, taking advantage of vastly increased computing power to use brute-force methods to evaluate possible solutions.

3. Data mining characterizes specific cases, generating a predicted value or classification of each case without regard to the utility of the model for understanding the underlying structure of the data.

4. Data mining aims strictly at identifying previously unseen data relationships rather than ascribing causality to variables in those relationships.[3]

These reasons highlight the specific value of a strictly inductive, non-hypothesis driven approach: data mining works for the quite different purposes for which it was designed.[4] The aim of data mining is to identify relationships among variables that may not be immediately apparent using hypothesis-driven methods. Having identified those relationships it is possible to take action based on the fact that the relationships predict a given outcome.

The growing interest in data mining is spurred, in part, by the increasing quantity of data available to institutional researchers from transactional databases, online operations, and data warehousing.[5] Baker suggests four areas of application: building student models to individualize instruction, mapping learning domains, evaluating the pedagogical support from learning management systems, and scientific discovery about learners.[6] Kumar and Chadha suggest using data mining in organizing curriculum, predicting registration, predicting student performance, detecting cheating in online exams, and identifying abnormal or erroneous data.[7] More recent

[1] Baker, "Data Mining for Education"; Stirton, The Future of Institutional Research – Business Intelligence.

[2] Baepler and Murdoch, "Academic Analytics and Data Mining in Higher Education."

[3] Two Crows Corporation, Introduction to Data Mining and Knowledge Discovery.

[4] Baker, "Data Mining for Education."

[5] Baepler and Murdoch, "Academic Analytics and Data Mining in Higher Education."

[6] Baker, "Data Mining for Education."

[7] Kumar and Chadha, "An Empirical Study of the Applications of Data Mining Techniques in Higher Education."

applications have embraced such suggestions, exploring course recommendation systems, retention, student performance, and assessment.[8]

In spite of significant methodological problems with these pilot studies, however, data mining is gaining hold operationally at the institutional level, predicting student success and personalizing content in online and traditional courses; making Netflix-style course recommendations, monitoring student progress through their academic programs, and sometimes intervening to force student action; modeling campus personal networks and student behavior with an eye toward identifying lack of social integration and impending withdrawal from the institution based on facilities usage, administrative data, and social network data. Admissions and recruiting are also growth areas for data mining. [9]

Challenges of using Big Data in Higher Education

Consequentialism: The Immediate Challenge

Nearly from its inception, data mining has raised ethical concerns. Once implemented, a series of challenges for both the individuals who are the subjects of data mining and the institution that bases policy on it arise as consequences. The most prominent of these are the related problems of privacy and individuality. The privacy of subjects in a data mining process is primarily a factor of information control: a subject's privacy has been violated to the extent that the opportunity for consent to collection or use of information is absent or in which personal information flows are used in ways that are incompatible with their social context.[10] The potential of data mining to violate personal privacy spans a range of applications. Mining data allows one to infer information about the data subject that some would not be comfortable divulging themselves, and worse allows for the manipulation of or discrimination against the subject, for example, by price discrimination and restrictive marketing.[11] These risks are very much present in higher education applications of data mining. Course recommendation or advising systems that consider student performance are a way of developing a comprehensive picture of student performance, in essence, an electronic reputation that the institution maintains and makes available to faculty and staff through dashboard and stoplight processes and administrative rules. Arizona State University's effort to identify students who intend to transfer is clearly not information that students would consistently want to divulge, as one ASU student reported.[12]

Privacy concerns can easily give way to challenges to individuality. To be sure, such challenges are not new; older techniques that describe central tendencies and typical relationships can easily be seen as contributing to a collectivization of subject, where all are treated identically based on the assumption that they are all "typical" students. Data mining can go far toward overcoming this because it recognizes and models diversity among subjects.[13] But while academic analytics tends to collectivize the students by treating them all identically to the central tendency case, data mining has a tendency to disaggregate the whole individual into nothing more than the sum of a specified set of characteristics. Data mining can create group profiles that become the persons represented, treating the subject as a collection of attributes rather than a whole individual and interfere with

[8] Ayesha et al., "Data Mining Model for Higher Education System"; Baradwaj and Pal, "Mining Educational Data to Analyze Students' Performance"; Llorente and Morant, "Data Mining in Higher Education"; Vialardi et al., "Recommendation in Higher Education Using Data Mining Techniques"; Zhang et al., "Use Data Mining to Improve Student Retention in Higher Education: A Case Study."

[9] Deliso, How Big Data Is Changing the College Experience; Parry, Colleges Mine Data to Tailor Students' Experience; Parry, College Degrees , Designed by the Numbers.

[10] Nissenbaum, Privacy in Context: Technology, Policy, and the Integrity of Social Life; van Wel and Royakkers, "Ethical Issues in Web Data Mining."

[11] Danna and Gandy, "All That Glitters Is Not Gold: Digging Beneath the Surface of Data Mining."

[12] Parry, College Degrees , Designed by the Numbers.

[13] Thomas and Galambos, "What Satisfies Students? Mining Student-opinion Data with Regression and Decision Tree Analysis."

treating the subject as more than a final predictive value or category.[14] Course recommendation systems are just such a case; students are encouraged to do what students like them have done before. Austin Peay's system does not consider student interests, while Arizona State's eAdvising system is built specifically to identify students whose "ambitions bear no relation to their skills."[15] This suggests that the students, far from being understood as individuals, are simply bundles of skills that need to be matched to an outcome.

At its extreme, data mining can undermine individuals' autonomy. Broadly speaking, autonomy can be understood as the ability to critically reflect on and act so as to realize or modify one's preferences, particularly preferences among conceptions of the good. This raises the questions of whether coercion and paternalism are ever justified, questions that are often addressed on the basis of a principle of preventing harm to others, furthering ends that the objects of the paternalism values themselves, or addressing a limited capacity for autonomy on the part of the object.[16] ASU's system of compelling students making insufficient academic progress to change their major is very much coercive, explicitly denying students to opportunity to exercise their own agency.

A softer but probably more common interference with autonomy is seen in Austin Peay's course recommendation system. This system is intended not just to provide information and advice—acting as an informed advisor might—but to remedy poor decision-making by students:

> *"[Provost Tristan] Denley points to a spate of recent books by behavioral economists, all with a common theme: When presented with many options and little information, people find it difficult to make wise choices. The same goes for college students trying to construct a schedule, he says. They know they must take a social-science class, but they don't know the implications of taking political science versus psychology versus economics. They choose on the basis of course descriptions or to avoid having to wake up for an 8 a.m. class on Monday. Every year, students in Tennessee lose their state scholarships because they fall a hair short of the GPA cutoff, Mr. Denley says, a financial swing that 'massively changes their likelihood of graduating.'"*[17]

This is a classic example of paternalism, the "use of coercion to make people morally better."[18] In this case, the institution compels students to be "wise" (in the administration's understanding of what a wise student would do, i.e., make choices that keep their GPA up, maintain financial aid, and maximize the probability of graduating on time) rather than allowing students to pursue a course that reflects their own "unwise" identities and interests (understood implicitly as ignorance and laziness).

An especially complicated form of interference is the creation of disciplinary systems, wherein the control of minutiae and constant surveillance lead subjects to choose the institutionally preferred action rather than their own preference, a system that generally disregards autonomy. Classifying students and communicating the classification to the professor used at Rio Salado College is virtually identical to Foucault's example of the Nineteenth Century classroom[19] and could be expected to have similar effects: encouraging conformity to a set of behaviors and values that the institution has conceived of as successful.

This is not to say that these violations of student autonomy are inherently unacceptable, or that one might, though conscious attention to autonomy in system design, craft systems that promote autonomy and provide informed advising without paternalism or disciplinarity. One might justify these interferences with autonomy as preventing waste of taxpayers' money (a harm to the taxpayer, arguably), as furthering the educational ends that students presumably have when they enroll, or as guidance for those who are still not fully mature or

[14] van Wel and Royakkers, "Ethical Issues in Web Data Mining."

[15] Parry, College Degrees , Designed by the Numbers.

[16] Dworkin, "Autonomy."

[17] Parry, College Degrees , Designed by the Numbers.

[18] Dworkin, "Autonomy," 363.

[19] Foucault, Discipline and Punish: The Birth of the Prison, 146-149.

lacking information about the consequences of a decision. But it remains necessary to provide such a justification in each case, as violations of the principle of autonomy are generally justified only as exceptions to the broad aim of allowing each person the maximum autonomy consistent with all others also having such autonomy. Such justifications are not present in campus implementations of data mining, as Delany's statement above shows: paternalism is not justified, but rather is itself the justification for interfering in student autonomy. Nor will systems that promote autonomy be realized without paying close attention to the ways existing systems curtail it.

Scientism: The Deep Challenge

The consequential challenges of data mining are the most prominent ones, but they are not the only ones. In fact, the most difficult challenges may be ones of which institutional researchers are least aware. In the process of designing a data mining process, institutional researchers build both empirical and normative assumptions, meanings, and values into the data mining process. These choices are often obscured by a strong tendency toward scientism among data scientists. For philosophers of science and technology, the term refers (almost always critically) either to the claim that the natural sciences present both epistemologically and substantively the only legitimate way of understanding reality, or to instances of scientific claims being extended beyond the disciplinary bounds in which the claim can be supported.[20] Such perspectives introduce the temptation to uncritically accept claims that purport to have scientific backing. Scientism has a long tradition in the social sciences, and especially in the study of education.[21] Critics of scientism in education see a fetishization of the scientific method, which manifests itself in contemporary policies such as *No Child Left Behind* in the United States and the PISA testing regime internationally and mandates "scientific" evidence of effectiveness as an authoritative practice of politics.[22] The preponderance of such methods in education research—and especially in the kinds of studies produced by institutional research offices—point to the assumption that traditional scientific methods are the ideal approach to understanding contemporary higher education.

Scientism is a trap that, if not avoided, can do substantial harm to students. But unfortunately, current examples of data mining in higher education have embraced, rather than rejected, scientism. A non-scientistic perspective critically evaluates methods and evidence before taking action upon it. But the casual attitudes toward causality and the ignorance of even statistical uncertainty in the academic literature on data mining in higher education suggest that the authors have taken an uncritical attitude toward the underlying science of data mining. Assuming that the relationships uncovered by data mining are inherently causal and reasonably certain can lead to ineffective actions and actions that reinforce rather than interdict causal mechanisms. Rio Salado College's lack of success with intervention after having identified a relationship between first-day login and online course success is telling. The institution assumed that the relationship; encouraging students to log in on the first day would thus increase their likelihood of success. The encouragement, in the form of a welcome email, had no effect, supporting an alternative explanation that sees both course success and first-day login as caused by students' self-motivation. While this intervention is unlikely to harm, at the least an opportunity has been missed to make an effective intervention.

The problem of scientism in data mining goes deeper than just poor methodology. Part of the scientist epistemology is the claim that science is objective, and thus it—and its products—are value-neutral. But one of the key recent findings in both the philosophy and the sociology of science is the value-ladenness of science and technology. This is more than just claims of biases in scientific inquiry that deviate from the norms of such inquiry; it is an inherent feature of science and technology that they embody and embed values as they are created within a complex web of technical and social interdependencies.[23] Design intent and assumptions about user behavior are especially significant sources of embedded values in technologies. The connection between

[20] Peterson, "Demarcation and the Scientistic Fallacy."

[21] Hyslop-Margison and Naseem, Scientism and Education Empirical Research as Neo-liberal Ideology.

[22] Baez and Boyles, The Politics of Inquiry.

[23] Nissenbaum, Privacy in Context: Technology, Policy, and the Integrity of Social Life, 4-6.

technological artifact and social purpose suggests that data mining applications in higher education are best understood as part of a problem-model-intervention nexus: In developing models data miners link their own meanings, values, and assumptions to similar ones taken from the problem and the intended intervention. The values embedded in a model nexus become part of the institutional context.

Vialardi and colleagues note that predictive analytic models "are based on the idea that individuals with approximately the same profile generally select and/or prefer the same things."[24] This very behaviorist model of human nature is at the foundation of every data model. While it is generally reasonable, one should note that it directly contradicts the rational utility maximizer model of human nature used in microeconomics or the habitual perspective of behavioral economics, and has very different implications for interventions. This is especially problematic in that interventions often incentivize behavior, a prescription best suited for rational utility maximizers. Similar processes embed more specific values in specific models. Most models are developed with both a problem and an intervention in mind, as can be seen in Austin Peay Provost Tristan Denley's description of the university's course recommendation system presented in section 2.1. The wisdom of a student's choice and the difficulty of making such a choice under these circumstances is part of the model; what it is to predict is not just a choice that the student will like but one which will be, from the institution's perspective, wise in the sense that it conforms to a utility function that values high grades and rapid progress toward graduation.

Practical Ethics for Ethical Data Mining

The importance of these questions, unfortunately, has not been matched by general solutions. But the above analysis suggests the formation of several fragmentary perspectives that can, if not provide solutions, lead data users in higher education to ask questions that will help refine their practices. The first step is to re-think what is meant by data mining, considering it as part of a broad technosocial system, a nexus of problem, model, and intervention built around assumptions, meanings, and values. In practice, this means thinking in terms of policies in which data mining will play a role and not merely in terms of mathematical modeling techniques. The ethical questions presented in data mining will be clearer when building a data mining model is situated in relation to the perceived need for the policy, the interventions that are proposed, the expected outcomes of the policy, and the ways in which the policy will be evaluated; problems such as incompatibilities between the assumptions of the data model and those of the intervention will only be apparent from this perspective.

The empirical and normative problems presented by scientism run parallel to each other, a parallel that suggests a path toward addressing the challenge. In both cases, the question is one of whether the model's conclusion supports the interpretation given it. This is a familiar problem to empirical researchers in higher education: the problem of validity. One can thus think of scientism as an attitude that compromises (or, at the least, assumes rather than demonstrates) the validity of the problem-model-intervention nexus either empirically or normatively. Kane presents an approach to validating measures based on a series of inferences from observation to construct that can serve as a model for data mining applications.[25] In developing or applying a data mining process researchers should ask themselves if the chain of inference from problem to model to implementation is sound, both scientifically and normatively. Where it is, ethical problems originating in scientism are likely to be averted. Where it is clearly flawed, researchers should change their approaches. But most importantly, where there are gaps in the reasoning researchers should identify the assumptions that allowed those gaps to be bridged uncritically and then subject those assumptions to critical analysis. Practiced iteratively, this approach can minimize the effects of scientism in data mining.

Consequential challenges are more directly amenable to analysis if only because many of them have been addressed in other contexts. One approach to these problems that allows for a comprehensive analysis without an extensive technical background in ethics is to consider the contextual integrity of data mining practices.[26]

[24] Vialardi et al., "Recommendation in Higher Education Using Data Mining Techniques," 191.

[25] Kane, "Validation."

[26] Nissenbaum, Privacy in Context: Technology, Policy, and the Integrity of Social Life.

As technosocial systems, the context of information flows is as much a defining feature of data exchange and use as the content of that information flow. While Nissenbaum intended contextual integrity to be a way of addressing privacy concerns, it can be expanded to include respecting the integrity of the individual and of the university. As actors within the informational context, changes to how the actors understand themselves are equivalent to changes in the actors, and the actors' goals and values are themselves part of the context whose integrity is to be maintained. Contextual integrity can thus be used to understand a broad range of ethical problems in the flow of information; changes in this context that are not supported by its underlying norms are violations of the contextual integrity of the information flows, and require justification distinct from that which justifies existing practices.

Conclusion

There is no question that data mining can be useful in higher education. Many students' struggles with courses and programs have revealed the need for guidance that is tailored to their unique circumstances. Processes that replace general theories of how all students behave with ones that recognize their diversity while basing decisions on rigorous, reliable processes are a central tool in promoting academic success and graduation. With a wide range of social actors recognizing (for better or worse) that allowing large numbers of students to fail is an inefficient use of resources, the potential of data mining to improve the efficiency of higher education cannot be dismissed.

But that efficiency comes with risks; the "brave new world" of Shakespeare can easily become Huxley's *Brave New World*. Data mining done well presents challenges to both individuals and institutions, and because of scientistic attitudes it is often done poorly at great cost, both practically and morally. Institutional researchers must minimize this risk. To do so, institutional researchers must understand data mining as part of a technosocial whole that spans the entire policy process. They must ensure the contextual integrity of information flows to protect the actors involved in data mining. And they must ensure both the scientific and the normative validity of the data mining process. Done properly, institutional research can secure these gains without compromising its commitment to the good of students.

References

Ayesha, Shaeela, T Mustafa, AR Sattar, and MI Khan. "Data Mining Model for Higher Education System." Europen Journal of Scientific Research 43, no. 1 (2010): 24–29.

Baepler, Paul, and Cynthia James Murdoch. "Academic Analytics and Data Mining in Higher Education." International Journal for the Scholarship of Teaching and Learning 4, no. 2 (2010). http://academics.georgiasouthern.edu/ijsotl/v4n2/essays_about_sotl/PDFs/_BaeplerMurdoch.pdf.

Baez, Benjamin, and Deron Boyles. The Politics of Inquiry: Education Research and the "Culture of Science." Albany, NY: State University of New York Press, 2009.

Baker, Ryan S.J.d. "Data Mining for Education." In International Encyclopedia of Education, edited by B. McGaw, P. Peterson, and E. Baker. 3rd Edition. Oxford: Elsevier, 2010. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Data+Mining+for+Education#3 http://users.wpi.edu/ rsbaker/Encyclopedia Chapter Draft v10 -fw.pdf.

Baradwaj, Brijesh Kumar, and Saurabh Pal. "Mining Educational Data to Analyze Students' Performance." International Journal of Advanced Computer Science and Applications 2, no. 6 (2011): 63–69.

Danna, Anthony, and OH Gandy. "All That Glitters Is Not Gold: Digging Beneath the Surface of Data Mining." Journal of Business Ethics 40, no. 4 (2002): 373–386.

Deliso, Meredith. How Big Data Is Changing the College Experience, 2012. http://www.onlinedegrees.org/how-big-data-is-changing-the-college-experience/.

Dworkin, Ronald. "Autonomy." In Edited by Robert E. Goodin & Phillip Pettit, A Companion to Contemporary Political Philosophy. Cambridge, Massachusetts: Basil Blackwell, 1995.

Foucault, Michel. Discipline and Punish: The Birth of the Prison. Second Vintage Ed. New York: Vintage Books, 1995.

Hyslop-Margison, Emery J, and M. Ayaz Naseem. Scientism and Education Empirical Research as Neo-liberal Ideology. Dordrecht: Springer, 2007. http://public.eblib.com/EBLPublic/PublicView.do?ptiID=337528.

Kane, M. T. "Validation." In Educational Measurement, edited by R. L. Brennan, 17–64. Fourth Edi. Westport, Connecticut: American Council on Education/Praeger, 2006.

Kumar, Varun, and Anupama Chadha. "An Empirical Study of the Applications of Data Mining Techniques in Higher Education." International Journal of Advanced Computer Science and Applications 2, no. 3 (2011): 80–84.

Llorente, Roberto, and Maria Morant. "Data Mining in Higher Education." In New Fundamental Technologies in Data Mining, edited by Kimito Funatsu, 201–220. New York: InTech, 2011. http://www.intechopen.com/books/new-fundamental-technologies-in-data-mining/data-mining-in-higher-education.

Nissenbaum, Helen. Privacy in Context: Technology, Policy, and the Integrity of Social Life. Staford, California: Stanford Law Books, 2010.

Parry, Marc. College Degrees , Designed by the Numbers, 2012. https://chronicle.com/article/College-Degrees-Designed-by/132945/.

———. Colleges Mine Data to Tailor Students' Experience, 2011. https://chronicle.com/article/A-Moneyball-Approach-to/130062/.

Peterson, Gregory R. "Demarcation and the Scientistic Fallacy." Zygon 38, no. 4 (December 2003): 751–761. doi:10.1111/j.1467-9744.2003.00536.x.

Stirton, E. Rob. The Future of Institutional Research – Business Intelligence, 2012. https://www.airweb.org/eAIR/specialfeatures/Pages/default.aspx.

Thomas, EH, and Nora Galambos. "What Satisfies Students? Mining Student-opinion Data with Regression and Decision Tree Analysis." Research in Higher Education 45, no. 3 (2004): 251–269.

Two Crows Corporation. Introduction to Data Mining and Knowledge Discovery. Third. Potomac, MD: Two Crows Corporation, 2005. http://www.twocrows.com/intro-dm.pdf.

Van Wel, Lita, and Lambèr Royakkers. "Ethical Issues in Web Data Mining." Ethics and Information Technology 6, no. 2 (2004): 129–140. doi:10.1023/B:ETIN.0000047476.05912.3d.

Vialardi, Cesar, Javier Bravo, Leila Shafti, and Alvaro Ortigosa. "Recommendation in Higher Education Using Data Mining Techniques." In Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, Proceedings, edited by T. Barnes, M Desmarais, C. Romero, and S. Ventura, 190–199. Cordoba, Spain: International Working Group on Educational Data Mining, 2009. http://www.educationaldatamining.org/EDM2009/uploads/proceedings/vialardi.pdf.

Zhang, Ying, Samia Oussena, Tony Clark, and Hyeonsook Kim. "Use Data Mining to Improve Student Retention in Higher Education: A Case Study." In Proceedings of the 12th International Conference on Enterprise Information Systems, Volume 1, edited by Joaquim Filippe and Jose Cordiero, 190–197. Funchal, Madeira, Portugal: SciTePress, 2010. http://eprints.mdx.ac.uk/5808/1/%5BCam_Ready%5DICEIS2010 Use Data Mining_Ying.pdf.