Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 23 Number 15, October 2018 ISSN 1531-7714

Overview of Data Mining's Potential Benefits and Limitations in Education Research

Emi Iwatani, *Digital Promise*

Education researchers are increasingly interested in applying data mining approaches, but to date, there has been no overarching exposition of their methodological advantages and disadvantages to the field. This is partly because the use of data mining in education research is relatively new, so its value and consequences are not yet well understood. Yet statisticians, sociologists and those who study computer-based education *have* discussed the methodological merits of data mining in education research. This article brings together their perspectives, providing an interdisciplinary overview of potential benefits and drawbacks. The benefits, regardless of scholar background, largely emphasize the speed and ease with which data mining approaches can help explore very large datasets. Perceived drawbacks, however, differ based on disciplinary expertise. For example, statisticians question data mining's exploratory nature and non-reliance on sampling theory, while sociologists raise concerns about an excessive reliance on data in research designs and in understandings of education.

Data mining is a process of systematically, and automatically or semi-automatically, uncovering patterns in data (Witten, Frank and Hall, 2011). It is typically conducted on very large datasets that would be difficult to examine sufficiently through traditional descriptive and inferential approaches. Data mining has become a hot topic in education research, which prompts the question: What benefits and concerns have scholars identified with using data mining in education research? This overview addresses the question by (1) identifying a representative sample of scholarly views on the value of data mining to education research, (2) closely examining the discourse to understand the context, motivations, and distinctness of various views, and (3) synthesizing these views comprehensively and succinctly. Relevant articles were identified through a search of peer-reviewed works concerning "data mining" in the ERIC database in August 2015. ERIC, sponsored by the U.S. Department of Education, was chosen because it is considered "the premier national bibliographic database of education literature" (University of Pittsburgh University Library System, 2015), and it only includes references related to education. The search was restricted to works published between 2005 and 2015, identifying 137 academic journal articles and 1 ERIC document. Among them, there were 13 substantive conceptual or theoretical discussions about the value of data mining as a methodology for education research. Key conceptual papers cited by these articles, some outside education, were also examined when appropriate.

After relevant articles were identified, each was read carefully to understand its main claims about the utility of data mining in education research and justifications for each claim. In the initial detailed read, care was taken to retain similar-sounding arguments and understand as much as possible about their assumptions and implications. While scholars generally agreed on the potential benefits of data mining, they raised a wide array of concerns based on their disciplinary expertise. Scholars with a statistics background, for example, identified different problems than those trained in sociology of science or learning analytics. Understanding arguments and counterarguments in each of these disciplines generally required additional rounds of careful reading and mapping out logical dependencies within and across the disciplines. Thus, one of the main

contributions of this article is that it comprehensively surveys perspectives from many different subspecialties of education research, incorporating concerns from different disciplines and making them accessible to a broader audience.

The articles were generally optimistic about how data mining could contribute methodologically to education and education research; a few seemed overly optimistic (AlShammari, Aldhafiri, & Al-Shammari, 2013; ElAtia, Ipperciel, & Hammad, 2012), and a few were critical (Gaševic, Dawson, & Siemens, 2015; Reimann, Markauskaite, & Bannert, 2014). There was general consensus on what data mining is and why it is used, and a shared sense of inevitability about its widespread use in education. Several compared and contrasted data mining to traditional statistics (Grover & Mehra, 2008; Zhao & Luan, 2006), which turned out to be an important theoretical framework through which to understand the purported benefits and drawbacks of data mining.

Potential benefits of using data mining in education research

Most scholars were optimistic about the benefits data mining could confer to the field. An important reason for this enthusiasm was that, in theory, data mining may lead to deeper understandings of individual learners, which in turn can improve their learning experiences (Berland, Baker, & Blikstein, 2014; Papamitsiou & Economides, 2014). Since learning involves multiple and complex pathways, approaches that can help detect such patterns could be especially valuable (Berland et al., 2014; Martin & Sherin, 2013). Data mining may even be necessary as educational datasets become larger and more complex. Some have pointed out that given the increasing size of available educational datasets, we cannot afford not to mine data (Grover & Mehra, 2008).

Data mining also may offer a unique contribution that differs from traditional statistical methods. In contrast to traditional statistical approaches, which were designed to analyze small samples, data mining is designed to efficiently analyze very large datasets (Grover & Mehra, 2008). This allows data mining to provide information when and how it is needed (Berland et al., 2014; Luan & Zhao, 2006), and detect *unexpectedly* useful information (ElAtia et al., 2012; Thuneberg & Hotulainen, 2006). Data mining also requires fewer statistical assumptions, making it easier and more flexible to employ for analysis. Decision trees, for example, do not require the typical parametric assumptions of linearity, normality, and homogeneity of variance. In addition, being less hypothesis-driven, data mining allows one to examine data without a heavy reliance on theoretical frameworks. As explained below, this can benefit a field like education where theoretical frameworks are not as strongly established (at least compared to the natural sciences) (Luan & Zhao, 2006).

Another unique benefit to data mining is that it can help analyze non-traditional forms of data in efficient and effective ways. Data mining can be applied to data on text, location, audio, images, interactions, and social relations (Grover & Mehra, 2008; Papamitsiou & Economides, 2014). This may help expand the analytic scope of traditionally qualitative education sub-fields. Lang and Baehr (2012) used text-mining to better understand the relationship between writing composition instruction and student performance. Through data mining, they could analyze larger quantities of text data than typical in writing composition education research and have more confidence in their results.

However, scholars do caution against the blind use of data mining in education research. Concerns arise from considerations of traditional statistical principals, sociology of science, and from examinations of recent activities in learning analytics and educational data mining. Concerns from the perspectives of traditional statistics are discussed first, as these are fundamental yet complex, and likely to be widely shared by many education researchers who have considered mining data.

Concerns from the perspectives of traditional statistics

> *Despite its obvious connection to statistics, data mining, which often employs "exotic" algorithms and seems to be operating mostly in a black box, has produced a fairly high level of discomfort in the statistical community. The major criticism of data mining centers on the lack of theory in the search for best predictions and, therefore, that too much power is given to the computer. This is directly contradictory to the traditional understanding of data analysis… (Zhao & Luan, 2006, p. 8)*

Data mining has been criticized in several ways, one of which is having insufficient regard of traditional statistical theory. Hand (1998, 2000) and Zhao and Luan (2006) described and addressed these types of concerns

by contrasting data mining with traditional statistical approaches. Statistics use data to confirm a statement nested within a theoretical framework, beginning with a null hypothesis about a population and employing a random sample of that population to either reject or fail-to-reject the hypothesis. Included variables in a statistical model are also selected based on theory. Data mining, on the other hand, "shares a similar philosophical root" with exploratory data analysis, which is not as focused, or dependent, on theory confirmation (Zhao & Luan, 2006, p. 11). Its goal is typically to find immediately actionable information that accurately predicts the behavior of a particular group of customers, students, or patients, *rather than* providing the best possible theoretical explanation of a complex social phenomenon. As such, data mining does not necessitate a well-defined background theory against which a model is selected and results are interpreted; although as Zhao and Luan (2006) emphasize, data mining still requires a great deal of sound human input. As the leading data mining frameworks (Chapman et al., 2000; Fayyad, Piatetsky-Shapiro, & Smyth, 1996; SAS Institute, 1998) make explicit, the researcher's understanding of the research context and dataset are critical to effective data mining. However, "compared with [traditional] statistics, data mining is less confined in presumptions about the relations among variables," and therefore it "[leaves] ample space for discoveries that might not occur otherwise" (Zhao & Luan, 2006, p. 11).

This difference in the importance of theory underscores traditional statisticians' concerns about data mining (Grover & Mehra, 2008; Zhao & Luan, 2006). Data mining activities are typically not well grounded in prior research, and therefore have less to contribute in terms of theory confirmation or explanation. They often do not assume a sampling theory, so they cannot make convincing statistical generalizations about a larger population. Without reliance on background and sampling theories, there is no hypothesis testing or significance values (often construed as "statistical rigor") attached to results. Finally, data mining may inflate the possibility of erroneously concluding that a finding is significant or important (inflation of Type I error). Such an error can be made because the data miner has very little theoretical grounding and does not know what is or is not significant with respect to what is already known. It can also occur if the data miner repeatedly explores the same data, using different methods or conditions, which increase the possibility of interpreting a spurious relationship as valid or important.

However, as Zhao and Luan (2006) explain, data mining's limitations are not necessarily devastating, and traditional statistics also help illuminate why. First, while theory can guide observations and provide a level of comfort that important findings actually exist, it can also blind researchers to seeing what is important, or even guide them in the wrong direction. John Tukey made the analogy of a data analyst as a detective "open to a wide range of ideas, possibilities, and idiosyncrasies," and a (traditional) statistician as a judge "examining and testing clearly identified hypotheses" (Tukey, 1962, summarized by Zhao & Luan, 2006, p. 11). To build on Tukey's analogy, detectives with strong preconceived notions about how criminals think and act can miss important clues that don't align with their preconceptions, or weigh too heavily the evidence that strongly supports their particular viewpoints, failing to resolve a case. This also applies in social scientific research, where it is not always prudent to have too many assumptions about what exists and how things work. When it comes to understanding a phenomenon that has insufficient theorizing, the atheoretical nature of data mining can be a strength, rather than a weakness.

That data mining is not based on sampling theory is also not particularly concerning if the technique is used primarily to build specific models that reflect local conditions, rather than to build global understandings. When companies and institutions mine data, their purpose is typically to predict information about their own clients and guide near-future decision making. Such organizations generally do not care whether that information is true more generally, across an entire industry, and therefore have no need to acquire random samples of companies. Zhao and Luan (2006) add that generating global, rather than specific, models is "an ambitious and even unrealistic task" (p. 12). They remark:

> A model is a simplification of reality, and a global model excludes low-level details, focusing only on a high level of abstraction that summarizes the data structure because it assumes homogeneity within the population. A globally generalizable model usually contains less detailed information than a specific model. But reality is extremely complicated, especially for social sciences, and fraught with difficulties and ambiguities stemming from deficiencies in measurement, design, and analysis. (Zhao & Luan, 2006, p. 12)

These authors argue that this general and crude nature of traditional statistical models explains the low threshold of acceptability of statistical models and why it is not uncommon for social scientists to present results that explain less than 20 percent of the variance in a dependent variable. The contrast between data mining and traditional statistics then, is not simply that the latter attains more generalizable knowledge. Rather, it is a tradeoff where, "typical statistical regression model uses a few variables to generalize to an entire population, [while] data mining provides the potential to take advantage of information at a more detailed and specific level" (Zhao & Luan, 2006, p. 12).

There is another way to think about the role of sampling in the data mining context: That as long as there is information and computing power necessary to analyze the entire population, there simply is no need for sampling. Traditional statistics was developed, in large part, as a pragmatic and economical means to understanding a phenomenon—it provided justification for making claims even if one looked only at a very small piece of it. Until well into the 1970s, most statistical analyses were conducted by hand (Zhao & Luan, 2006), which meant there was a significant limitation to how much information one could reasonably include in an analysis. Data collection and storage were expensive, especially before the use of electronic databases and online communication became routine, prohibiting analyses of rich population data. Over the past several decades, population information has become increasingly available, as has computing power. School districts, institutes of higher education, state and local education, health and social services departments, and criminal justice systems, now often have electronic records of every person who has been part of their systems. Many research questions that may have once required sampling no longer require it because data are available for the entire population. Although data mining does not require users to adhere to a sampling theory, it is not a serious concern as long as data is mined from all or most of the population that one hopes to understand.

Concern about a lack of statistical significance attached to data mining results is a variant of the sampling concern. Statistical significance is a measure of uncertainty associated with sampling error. In some instances, there is no need to assess the possibility that the results are due to sampling error, e.g., when: (i) there is information on the entire population, (ii) there is a large sample that adequately represents the population, (iii) there is a large enough sample such that nearly any difference turns out to be statistically significant, and/or (iv) there is no interest in generalizing conclusions far beyond existing data. However, if the above conditions are not met—i.e., if the researcher seeks to generalize conclusions far beyond a small, potentially unrepresentative sample—trusting data mining results wholesale, without regard to the possibility of sampling error, would be problematic.

The most serious concern about data mining from the perspective of traditional statistics is the inflation of Type I error due to data dredging. As Hand (1998) describes:

> [Data mining] has a derogatory connotation because a sufficiently exhaustive search will certainly throw up patterns of some kind—by definition data that are not simply uniform have differences which can be interpreted as patterns. The trouble is that many of these "patterns" will simply be a product of random fluctuations, and will not represent any underlying structure. ... To statisticians, then, the term data mining conveys the sense of naïve hope vainly struggling against the cold realities of chance. (p. 112)

The possibility of model over-fit and Type I error increases when data mining is used to build precise models for local use (rather than less precise models for global understanding). Cross-validation of the results within and/or across datasets and across algorithms are essential to data mining, as is checking the feasibility of the model with domain experts (Luan & Zhao, 2006; Provost & Fawcett, 2013; Witten et al., 2011). Restricting model specificity during the model creation stage (e.g., using stopping rules or pruning when creating decision trees) is another way in which model over-fit can and should be addressed.

In summary, through the lens of traditional statistics, data mining may be problematic because of its atheoretical nature, non-reliance on sampling theory, and increased possibility of Type I error. The concerns are not insurmountable, yet they need to be understood and considered when employing data mining techniques.

Concerns from sociology of science

> *All told, the generation, accumulation, processing and analysis of digital data is now being touted as a potential panacea for many current educational challenges and problems.* (Selwyn, 2015, p. 67)

A concern from sociology of science is that data mining contributes to tunnel-vision of education data, which has serious repercussions. In a discussion of the significance, merit and demerits of data mining and data-driven approaches in education, Selwyn (2015) raises concerns from a sociological (and the newly emerging "digital sociological"[1]) perspective, regarding the "datafication" of education, or the increased data-reliance in designs and understandings of education. Several of these concerns pertain directly to data mining's usefulness in education research. The first is that increased data-reliance may cause people to regard complex social and educational problems as complex but solvable *statistical* problems. Focusing too much on available data may prevent education researchers from considering important and relevant nuances, contextual factors, causal factors, and counter-narratives. Selwyn describes:

> The recording of social 'facts' into digital data, therefore, implies that some qualities and characteristics will be made better known than others. For example, as Ruppert (2012) notes, the core sociological constructs of race, social class, gender, sexuality and so on, do not translate easily into data categories, despite their constant use within data collection and analysis. Often digital data can be said to support little more than 'surface' understandings of sociological entities (Savage, 2009). … Much of the depth that is lacking from digital data could be argued to include issues of historical context and connections with past events, individualist and humanist accounts of the social, and an underpinning sense of moral knowledge (see Barnes, 2013; Ruppert, 2013). (p. 75)

Along the same lines, increased interest in data mining could consciously or subconsciously lure education researchers toward an unhealthy reductionism: regarding teaching and learning primarily in terms of easily operationalized attributes for practicality or other reasons. Worries about unhealthy reductionism and brute operationalization of complex constructs are not unique to data mining. However, the increased volume, variety, and velocity of data processing (the classic descriptors of "big data," per Laney (2001)) increases attention and reliance on data-driven approaches, and therefore increases the magnitude of this concern. Important factors related to learning, such as social interactions, agency, perception, attitudes, race, gender, historical context, cultural beliefs, are difficult to operationalize, and quality data will always be difficult and time-consuming to collect. As Selwyn (2015) and Manovich (2012) note, we do not want to neglect studies on "deep data" on just a few cases by focusing too much on "surface data" about many cases.

In addition, data mining raises concerns about differential power dynamics among those who analyze and are analyzed, and those who can and cannot analyze. Selwyn (2015), drawing from Lupton (2013), Manovich (2012), and Ruppert (2012), suspect that data, and the ability to use data, is a form of power that has the potential to be distributed inequitably and misused. It is conceivable that machine learning specialists involved in educational data mining come to obtain a disproportionate amount of power in deciding what happens in education (even if they are not familiar with many aspects of the field), simply because of their technical knowledge of manipulating large educational datasets. Governments, education policy makers, school districts, researchers and companies may provide machine learning specialists with more funding, attention, and voice than is ultimately good for teachers and students.

Open-access data and privacy are related concerns for education researchers as they further explore big data in education (ElAtia et al., 2012). Open access would protect data from concentrating in the hands of the few, while privacy would provide some protection of those who are analyzed from those in power.

Concerns from Learning Analytics and Educational Data Mining

Those with direct experience or familiarity with current data mining practices have raised similar concerns. Educational data mining (EDM) and learning analytics are emerging and overlapping interdisciplinary fields, which harness knowledge from large educational datasets. Relatively speaking, EDM is more interested in finding new patterns, and/or developing new algorithms, while learning analytics applies the patterns

[1] This emerging subfield of sociology, and sociology of technology, begins with the assumption that data is political, value-laden and power-conferring in nature, rather than objective, neutral and unproblematic. It also pays close attention to how data shapes and are shaped by social interests.

to improve teaching and learning (Bienkowski, Feng, & Means, 2012).

Reflecting on how EDM and related e-research methods have analyzed self-regulated learning, Reimann et al. (2014) noticed that many studies tend to assume a "flat ontology" that relies heavily and makes assumptions about simple user behaviors such as clicking, logging in, moving their eyes, typing, and uttering. For example, while their previous study found that "reading" for successful students was more strongly associated with "monitoring" and "elaboration" than with "repeating," their models lacked explanatory power across contexts and different student dispositions because their theoretical framework was ontologically impoverished. Reiman et al.'s general cautionary point was that "big data" and "more data" are not identical with conceptually rich data and deep data. They suggested enriching the EDM research ontology to include social structures and a range of cognitive and non-cognitive processes, which extend beyond physical observable behaviors such as clicking and typing. This would also involve collecting richer data, meaning data may be acquired from multiple sources, and analyzed in a way that respects ontological complexity (these researchers suggest system dynamics and agent-based modeling).

Martin and Sherin (2013) raised similar concerns in their introduction to a special issue on learning analytics of the *Journal of Learning Sciences*. Their assessment of the EDM and learning analytics fields was cautiously optimistic and based on the potential utility of these methods, rather than their actual results:

Although the educational data mining and [learning analytics] communities have produced a steady stream of interesting results, work in education has far to go in order to reap the benefits for student learning… (pp. 511-512).

Their discussion on the potential of learning analytics to learning science researchers, while on-the-whole positive, cautioned that there is increased temptation to conduct research on topics where big data are easy to collect: While learning analytics can be conducted on traditional data, "when we apply [learning analytics], we are more likely to restrict our study to learning activities that are conducted using computers" (p. 515). Like Reimann et al. (2014), they urged learning analytics researchers to look beyond mouse clicks and key presses, to continue to research learning in a broad range of settings, and to ensure research questions guide methodology rather than the other way around.

Progress in learning analytics has been difficult because of its interdisciplinary nature (Gaševic et al., 2015). Consider, for example, an initiative to improve academic success by providing students with timely, automated feedback about their coursework. For such an initiative to work, good analytics, a user-friendly implementation platform, and high-quality feedback are needed. The success of learning analytics depends upon substantive collaboration among machine learning scientists, education practitioners, and educational researchers, making such initiatives riskier and more expensive.

A final concern that those in EDM and learning analytics raise pertains to unintended negative consequences for students. Corrin and de Barba (2014) found that high-achieving students tended to underperform in a class, when dashboards informed them of their standing relative to the class mean. Along the same lines, learning analytics researchers have worried that constant reminders of poor performance may cause undue distress to students, and/or diminish the quality of teaching and learning such that it becomes narrowly focused on improving superficial metrics (Gaševic et al., 2015). Of course, conducting data mining in educational research *per se* is unlikely to be a direct cause of such consequences. However, just as educational and psychological assessment developers must carefully consider the unintended negative consequences of the instruments they develop (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Linn, 1997; Messick, 1975), quantitative education researchers should take care to minimize the negative implications of their research.

Implications

There is great optimism and momentum for data mining applications that investigate the nature of learning and education. The ability to analyze a large amount of data quickly provides the possibility to find undiscovered relationships among teaching and learning variables that are useful or important. Data mining also allows researchers to analyze visual, audio, and text data without extensive recoding.

Concerns about data mining are not devastating, but they do provide guidance to those who hope to use it for research. Researchers should be principled in their use of this approach. It is possible to mine data with hardly any knowledge of the domain from which the data come—however, such reckless application is likely to be a hindrance to the field. While it is neither necessary nor always desirable for data miners to take a rigorous hypothesis-driven approach, the methodology and interpretation of results should be well informed by what is known (or anticipated) in the field. Data mining can be used for prediction, theory development, or hypothesis generation: The specific objective should determine the method, rather than conversely. Special attention should be paid to sampling, over-fit avoidance, and predictor set completeness.

Like any tool, the utility of data mining depends largely on the skill and imagination of the user. And like any tool, it may be used for a variety of goals and purposes. The verdict is still out on how useful data mining can be in educational research; even in learning analytics and educational data mining, convincing applications of data mining are still rare. As educational researchers explore the utility of data mining, they should maintain a balanced perspective, inform others even of null-results and unintended downstream consequences, and be vigilant in pursuing questions with answers worth knowing.

References

Asterisks (*) denotes articles that were analyzed in the literature review.

AlShammari, I. A., Aldhafiri, M. D., & Al-Shammari, Z. (2013). A meta-analysis of educational data mining on improvements in learning outcomes. *College Student Journal, 47*(2), 326-333. *

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, D.C: National Research Council and National Academy of Education.

Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning*, 19(1-2), 205-220. *

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. Washington DC: U.S. Department of Education, Office of Educational Technology.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-dm 1.0: Step-by-step data mining guide. Retrieved from https://the-modeling-agency.com/crisp-dm.pdf

Corrin, L., & de Barba, P. (2014). Exploring students' interpretation of feedback delivered through learning analytics dashboards. Paper presented at the Australasian Society for Computers in Learning in Tertiary Education, Dunedin, NZ.

ElAtia, S., Ipperciel, D., & Hammad, A. (2012). Implications and challenges to using data mining in educational research in the canadian context. *Canadian Journal of Education*, 35(2), 101-119. *

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge discovery and data mining: Towards a unifying framework. Paper presented at the KDD.

Gaševic, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. TechTrends: *Linking Research and Practice to Improve Learning*, 59(1), 64-71.*

Grover, L. K., & Mehra, R. (2008). The lure of statistics in data mining. *Journal of Statistics Education*, 16(1).*

Hand, D. J. (1998). Data mining: Statistics and more? The *American Statistician*, 52(2), 112-118. doi:10.1080/00031305.1998.10480549

Hand, D. J. (2000). Data mining: New challenges for statisticians. *Social Science Computer Review*, 18(4), 442-449.

Laney, D. (2001). 3d data management: Controlling data volume, velocity and variety. META Group Research Note, 6, 70. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lang, S., & Baehr, C. (2012). Data mining: A hybrid methodology for complex and dynamic research. *College Composition and Communication*, 64(1), 172-194.*

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, 16(2), 14-16. doi:10.1111/j.1745-3992.1997.tb00587.x

Luan, J., & Zhao, C.-M. (2006). Practicing data mining for enrollment management and beyond. *New Directions for Institutional Research* (131), 117-122.*

Lupton, D. (2013). Digital sociology: Beyond the digital to the sociological. Paper presented at the The Australian Sociological Association 2013 Conference, Melbourne.

Manovich, L. (2012). Trending: The promises and challenges of big social data. In G. K. Matthew (Ed.), Debates in the digital humanities (pp. 460-475). Minneapolis, MN: University of Minnesota Press.

Martin, T., & Sherin, B. (2013). Learning analytics and computational techniques for detecting and evaluating patterns in learning: An introduction to the special issue. *Journal of the Learning Sciences*, 22(4), 511-520.*

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30(10), 955-966. doi:10.1037/0003-066X.30.10.955

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

Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. Sebastopol, CA: O'Reilly Media.

Reimann, P., Markauskaite, L., & Bannert, M. (2014). E-research and learning theory: What do sequence and process mining methods contribute? *British Journal of Educational Technology*, 45(3), 528-540.*

Ruppert, E. (2012). The governmental topologies of database devices. *Theory, Culture & Society*, 29(4-5), 116-136.

SAS Institute, I. (1998) SAS institute white paper, data mining and the case for sampling: Solving business problems using SAS enterprise miner software. Carey, NC: SAS Institute.

Selwyn, N. (2015). Data entry: Towards the critical study of digital data and education. *Learning, Media and Technology*, 40(1), 64-82.*

Thuneberg, H., & Hotulainen, R. (2006). Contributions of data mining for psycho-educational research: What self-organizing maps tell us about the well-being of gifted learners. *High Ability Studies*, 17(1), 87-100.*

Tukey, J. W. (1962). The future of data analysis. 1-67. doi:10.1214/aoms/1177704711

University of Pittsburgh University Library System. (2015, Jul 13, 2015). Education databases. Retrieved from http://pitt.libguides.com/educationdatabases

Witten, I. H., Frank, E., & Hall, M. A. (2011). Data mining : Practical machine learning tools and techniques. Burlington, MA: Morgan Kaufmann.

Zhao, C.-M., & Luan, J. (2006). Data mining: Going beyond traditional statistics. *New Directions for Institutional Research* (131), 7-16.*

Citation:

Iwatani, Emi. (2018). Overview of Data Mining's Potential Benefits and Limitations in Education Research. *Practical Assessment, Research & Evaluation*, 23(15). Available online: http://pareonline.net/getvn.asp?v=23&n=15

Corresponding Author

Emi Iwatani
Digital Promise
2955 Campus Dr. Suite 110
San Mateo, CA 94403

email: eiwatani [at] digitalpromise.org