Toward a Bidirectional View of Causality in Big Data Analytics: The Case of Learning Analytics

Marta Stelmaszak Rosa
Portland State University, stmarta@pdx.edu

Alexi Aaltonen
Temple University

Toward a Bidirectional View of Causality in Big Data Analytics: The Case of Learning Analytics

Aleksi Aaltonen, Fox School of Business, Temple University (aleksi@temple.edu)
Marta Stelmaszak, London School of Economics and Political Science (m.stelmaszak1@lse.ac.uk)

The effects of big data are felt across industries, shaping managerial practices, structures and organising largely through the new forms of analytics and modelling data enable (Couldry & Powell 2014; George et al. 2014). Currently, much of the managerial literature on big data assumes a relatively straightforward epistemic relationship operating from phenomena to data and from data to analytics and analytical insights, supporting increasingly detailed representations and attempts to model business processes and environment. At the same time, more critical researchers from a variety of fields, from information systems (Alaimo 2014; Alaimo & Kallinikos 2017; Constantiou & Kallinikos 2015; Ekbia et al. 2015) through sociology (van der Vlist 2016; Iliadis & Russo 2016) to media studies (Puschmann & Burgess 2014; Gregg 2015) have problematized the operation of big data analytics, suggesting that big data often incorporate various social, cultural and technical biases, and may even come to shape the phenomena they are used to describe.

Relying on both managerial and critical perspectives on big data analytics, we argue that to fully seize the opportunities big data analytics offers, a more elaborate, bidirectional view of causality associated with big data analytics is needed. In particular, we draw from an established theory of reactivity (Espeland & Sauder 2007) from sociology to lay bare the mechanisms by which big data analytics shape environments that it initially may only purport to describe. Such patterns unavoidably affect the usability of and uses to which big data analytics can be put in each context. A robust understanding of feedback mechanisms involved in big data analytics, which we hope to contribute to, could help solve some of the persistent problems in analytics, reconcile conflicting results, and thus enable a more potent use of big data.

To investigate the reactive character of big data analytics, we designed a mixed methods case study of a sophisticated learning analytics system used by a top UK business school to collect, analyse and utilize data about student learning. Our aim is to specify and validate feedback mechanisms at work, including the unintended consequences for the organisation, and thus contribute to a bidirectional view of causality in big data analytics.

Big Data

A central assumption or, for that matter, a belief about big data which prompted many commentators to herald a new era is that new forms of data and proliferating data sources can, in a relatively unmediated way, capture and describe reality bringing "the end of theory" (Anderson 2008) as though they fully represented the world (Ekbia et al. 2015, boyd & Crawford 2012; Kitchin 2014). At the same time, there are various strands of literature pointing to the reductive character of big data. In particular, social data, as a type of big data capturing human interaction and behaviour in social online environments, are created as an effect of "how technology translates social interaction into computable objects" (Alaimo & Kallinikos 2017: 9). The production of social data, including their analytics, relies on the processes of encoding, that is the formalisation as objects of users and their actions and connections along the lines of pre-established actions which entails "the programmed disaggregation of individual users in countable actions" (Alaimo & Kallinikos 2017: 12), aggregation, and computation that make it possible to measure and classify users and their actions. Importantly, several researchers pointed out that data and analytics seem to be a continuation of a much longer history of statistics, measurement, and calculations, claiming that data have history (Barnes & Wilson 2014) and rely heavily on the mechanism of commensuration (Kitchin & McArdle 2016), classification (Bowker & Star 1999: 10) and calculation (Power 1997), that is subjecting number-based data to calculative practices which entail “a progressive reduction of complexity” (Starr 1980: 40).

Different strands of literature which we bring together indicate that the relationship between the world of people, their actions or behaviours and big data produced out of them as users and their clicks is one of reduction, or reductive representation through a series of mechanisms of production, rather than a simple, uncomplicated equivalence.

Big Data Analytics

Terms such as “business analytics”, “business intelligence” and “big data” are often used interchangeably to refer to similar topics (Holsapple et al. 2014, Bayrak 2015). We refer to all such analytics as *big data analytics* and further focus on its subfield known as learning analytics, that is “the measurement, collection, analysis, and reporting of data about learners and their contexts, for the purposes of understanding and optimizing learning and the environments in which it occurs” (Siemens 2013: 1328). Learning analytics exemplifies epistemological issues related to big data, but it also allows us to study big data analytics in a setting that should be as informed about the managerial issues of big data as possible, that is, business school education.

Virtual Learning Environments (VLE) typically track behaviour such as reading and writing of resources, taking tests, performing tasks or communicating (Mostow et al. 2005) in the form of time-stamped clicks. These data can then be combined with information about users’ profiles, academic results and interaction data (Romero et al. 2008). The data are then subject to a range of pre-processing and aggregation tasks typical of data analytics (Romero et al. 2008), and finally they are displayed through dashboards to different types of users. The goals of learning analytics are built upon the assumption that VLE data can be associated with effective learning leading to better feedback and assessment, as well as improved capacity for interventions (Macfadyen et al. 2014). However, despite persistent efforts, researchers often have to agree that “one key finding from the work was largely a null one” (Clow 2014: 51). Researchers in the field of learning analytics have struggled to identify statistically significant and theoretically robust correlations (e.g. Ruipérez-Valiente et al. 2015).

At the same time, researchers are beginning to notice how learning analytics disrupts the current ways of how teaching staff work (Piety & Hickey 2014) and that teaching has to be changed to accommodate analytical procedures and data production, leading to a transformation of how schooling is done (Sellar 2015). Overall, analytics changes the practices of interacting with, understanding and using data (Sellar 2015), while at the same time introducing new practices around educational measurement (Sellar 2014).

As we have laid out, despite numerous voices pointing towards the (sometimes unintended) consequences of big data analytics, we do not yet have a complete view of potential mechanisms of how analytics can feed back into what big data attempt to capture. This is an important shortcoming in literature not only because we currently do not have a full picture of the potentially recursive nature of the relationship between big data and what they purport to describe, but also because of some puzzling and conflicting findings from practice.

Theory of Reactivity

We are not the first ones to posit that measurement can lead to reactive effects, and we are also not the first ones to suggest that big data are a form of measurement, as discussed above. The theory of reactivity, which we employ, has been developed through Espeland and Sauder’s seminal study of law school rankings (2007). Since then, reactivity has been deployed to study the rankings of other educational institutions (Goglio 2016; Hazelkorn 2011), corporate reputation rankings (Sekou Bermiss et al. 2014), or valuation online (Orlikowski & Scott 2014; Jeacle & Carter 2011).

As proposed by Espeland and Sauder, measurement and public measures tend to lead to reactivity: individuals altering their behaviour in reaction to being evaluated, observed or measured. While the authors do not question the value of reductive representation afforded by rankings, they are primarily preoccupied by the way rankings become reactive, that is feed back into the schools they are supposed to only rank in unintended ways. According to the theory, reactive measurement impacts organisations through four main mechanisms: commensuration, self-fulfilling prophecies, reverse engineering, and narratives. The four mechanisms are summarised in Table 1.

Mechanism	Operation	Effects
COMMENSURATION	Transformation of different qualities into a common metric (Espeland & Stevens 1998), translating complex processes into single figures (Miller 2001), often relying on simplification and normalisation (Sauder & Espeland 2009).	Changing locus of attention by altering relationships (Espeland & Stevens 1998), creating visibility and invisibility (Espeland & Stevens 1998).
SELF-FULFILLING PROPHECY	Reactions to measures which confirm the expectations embedded in measures (Espeland & Sauder 2007) which in turn encourage behaviour that conforms to them (Espeland & Sauder 2016).	Performing to a measure as seen in the case of US law schools (Espeland & Sauder 2007; Sauder & Lancaster 2006; Stake 2006).
REVERSE ENGINEERING	Working backward through the construction of a completed measure to understand how it works (Espeland & Sauder 2016).	Actors stop thinking about the institution as a whole, but rather as a collection of discrete, measurable units whose functioning can be changed according to the formula.
NARRATIVE	A story featuring characters, events, scenes and plots involving a conflict or problem (Espeland & Sauder 2016), can be celebratory or defensive, often including causal explanations for changes.	Repeated at various levels of seniority and across many functions, narratives become powerful vehicles of identity and influence actions and behaviours in line with the predominant narrative.

Table 1. Reactive mechanisms (Espeland & Sauder 2007, Sauder & Espeland 2009)

The four mechanisms described in Table 1 offer a promising starting point to develop a bidirectional view of causality in big data analytics albeit we expect the mechanisms to operate differently in other contexts. We use the theory of reactivity as our scaffolding, guiding our data collection, coding, and analysis, and enabling us to lay bare the unintended feedback mechanisms of the phenomenon under study.

Research Design and Expected Contribution

The research uses a sequential mixed-methods approach (Venkatesh et al. 2013) to first explore, and then develop and validate a theory on how reactive big data analytics works in a VLE. The approach allows to combine intensive understanding of specific causal mechanisms in operation with their validation and the assessment of effect sizes using standard quantitative techniques. In short, we believe the mixed methods approach offers a good methodological fit with the research problem (Edmondson & McManus 2007).

We first carried out 30 semi-structured interviews between June and September 2017 with members of academic, teaching, administrative and systems development staff to develop a rich understanding of the nature of analytics use in the organization. In the evidence, we looked for cues about the mechanisms suggested by the theory of reactivity. The evidence from interviews was complemented by the analysis of documentary evidence and, in particular, user interfaces of the learning platform and its analytics system. Our preliminary findings contain promising indications pointing towards the reactive character of big data analytics in this context, with the presence of four mechanisms initially identified.

The qualitative analysis fed into modelling prospective reactive mechanisms in the studied setting, which will then be explored using trace data from the VLE, relying on a large clickstream, student profile and assessment performance dataset spanning a period of 12 months, which we intend to subject to several computational analyses. This is expected to contribute to solving some methodological issues in big data analytics, laying grounds for further studies into bidirectional causality between phenomena and big data analytics at a significant detail. Further, we are hoping to understand how bidirectional causality of big data analytics impacts organisations and their management in a digital environment.

We are open to discussing the promising research design employing both a qualitative approach to studying big data, as well as a computational analysis of a large dataset. We are particularly interested in exploring the ways in which the quantitative part of the study can be used to further test and expand the theory. Finally, we are looking for a discussion of our theoretical proposals and how the bidirectional causality of big data analytics can be further embedded in the study of digital organisations.

References

Alaimo, C. & Kallinikos, J., 2017. Encoding the Everyday: Social Data and its Media Apparatus. In C. Sugimoto, H. Ekbia, & M. Mattiolo, eds. *Big data is not a monolith: policies, practices, and problems*. pp.1–29, Cambridge, MA: The MIT Press.

Alaimo, C., 2014. Thesis. Computational consumption: Social Media and the construction of Digital Consumers. pp.1–220. Available at: http://etheses.lse.ac.uk/975/. London School of Economics and Political Science.

Anderson, C., 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*. Available at: https://www.wired.com/2008/06/pb-theory/. Accessed on: 17/08/2016.

Barnes, T.J. & Wilson, M.W., 2014. Big Data, social physics, and spatial analysis: The early years. *Big Data & Society*, April-June, pp.1–14.

Bayrak, T. 2015. A Review of Business Analytics: A Business Enabler or Another Passing Fad. *World Conference on Technology, Innovation and Entrepreneurship*. Procedia – Social and Behavioral Sciences, 195, pp.230–239.

Bowker, G.C. & Star, S.L., 1999. *Sorting things out: classification and its consequences*, Cambridge: MIT Press.

boyd, d. & Crawford, K., 2012. Critical Questions for Big Data. *Information, Communication & Society*, 15(December 2013), pp.662–679.

Clow, D., 2014. Data Wranglers: Human interpreters to help close the feedback loop. In *LAK'14*, pp.49–53.

Constantiou, I. & Kallinikos, J., 2015. New games, new rules: big data and the changing context of strategy. *Journal of Information Technology*, 30(1), pp.44–57.

Couldry, N. & Powell, A., 2014. Big Data from the bottom up. *Big Data & Society*, July-Decem, pp.1–5.

Edmondson, A., McManus, S. E. 2007. Methodological Fit in Management Field Research. *Academy of Management Review*, 32:4, pp.1155–1179.

Ekbia, H. et al., 2015. Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology*, 66(8), pp.1523–1545.

Espeland, W. N., Sauder, M. 2016. *Engines of Anxiety: Academic Rankings, Reputation, and Accountability*, New York: Russel Sage Foundation.

Espeland, W.N. & Sauder, M., 2007. Rankings and Reactivity: How Public Measures Recreate Social Worlds. *American Journal of Sociology*, 113(1), pp.1–40.

Espeland, W.N. & Stevens, M.L., 1998. Commensuration as a Social Process. *Annual Review of Sociology*, 24, pp.313–343.

George, G., Haas, M. & Pentland, A., 2014. Big Data and Management. *Academy of Management Journal*, 57(2), pp.321–326.

Goglio, V. 2016. One size fits all? A different perspective on university rankings. *Journal of Higher Education Policy and Management*, 38:2, pp.212–226.

Gregg, M., 2015. Inside the Data Spectacle. *Television & New Media*, 16(1), pp.37–51.

Hazelkorn, E. 2011. *Rankings and the Reshaping of Higher Education. The Battle for World-Class Excellence*, London: Palgrave Macmillan.

Holsapple, C., Lee-Post, A., Pakath, R. 2014. A unified foundation for business analytics. *Decision Support Systems*, 64 pp.131–141.

Iliadis, A. & Russo, F., 2016. Critical data studies: An introduction. *Big Data & Society*, 3(2),

pp.1–7.
Jeacle, I., Carter C. 2011. In TripAdvisor we trust: Rankings, calculative regimes and abstract systems. *Accounting, Organizations and Society*, 36, pp.293-309.
Kitchin, R. & McArdle, G., 2016. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, 3(1), pp.1–10.
Kitchin, R., 2014. *Data devolution. Big data, open data, data infrastructures & their consequences*, London: Sage.
Macfadyen, L.P. et al., 2014. Embracing Big Data in Complex Educational Systems: The Learning Analytics Imperative and the Policy Challenge. *Research & Practice in Assessment*, 9, pp.17–29.
Miller, P., 2001. Governing by Numbers: Why Calculative Practices Matter. *Social Research*, 68(2), pp.379–396.
Mostow, J. et al., 2005. An Educational Data Mining Tool to Browse Tutor-Student Interactions: Time Will Tell! *Proceedings of the Workshop on Educational Data Mining (2005)*, pp.15–22.
Orlikowski, W.J. & Scott, S. V, 2014. What Happens When Evaluation Goes Online? Exploring Apparatuses of Valuation in the Travel Sector. *Organization Science*, 25(3), pp.868–891.
Piety, P.J. & Hickey, D.T., 2014. Educational Data Sciences: Framing Emergent Practices for Analytics of Learning, Organizations, and Systems. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge - LAK '14*. pp.193–202.
Power, M., 1997. *The audit society: Rituals of verification*, Oxford: Oxford University Press.
Puschmann, C. & Burgess, J., 2014. Metaphors of Big Data. *International Journal of Communication*, 8, pp.1690–1709.
Romero, C., Ventura, S. & García, E., 2008. Data mining in course management systems: Moodle case study and tutorial. *Computers and Education*, 51(1), pp.368–384.
Ruipérez-Valiente, J.A. et al., 2015. ALAS-KA: A learning analytics extension for better understanding the learning process in the Khan Academy platform. *Computers in Human Behavior*, 47, pp.139–148.
Sauder, M. & Espeland, W.N., 2009. The Discipline of Rankings : Tight Coupling and Organizational Change. *American Sociological Review*, 74, pp.63–82.
Sauder, M. & Lancaster, R., 2006. Do Rankings Matter? The Effects of U.S. News & World Report Rankings on the Admissions Process of Law Schools. *Law & Society Review*, 40(1), pp.105–134.
Sekou Bermiss, Y., Zajac E. J., Brayden G. K. 2014. Under Construction: How Commensuration and Management Fashion Affect Corporate Reputation Rankings. *Organization Science,* 25(2), pp.591–608.
Sellar, S., 2014. A feel for numbers: affect, data and education policy. *Critical Studies in Education*, 56(1), pp.131–146.
Sellar, S., 2015. Data infrastructure: a review of expanding accountability systems and large-scale assessments in education. *Discourse: Studies in the Cultural Politics of Education*, 36(5), pp.765–777.
Siemens, G., 2013. Learning Analytics: The Emergence of a Discipline. *American Behavioral Scientist*, 57(10), pp.1380–1400.
Stake, J.E., 2006. The Interplay Between Law School Rankings, Reputations, and Resource Allocation: Ways Rankings Mislead. *Indiana Law Journal*, 81, pp.229–270.
Starr, P., 1980. The Sociology of Official Statistics. In W. Alonso & P. Starr, eds. *The politics of numbers*. New York: Russell Sage Foundation, pp.7–58.
van der Vlist, F., 2016. Accounting for the social: Investigating commensuration and Big Data practices at Facebook. *Big Data & Society*, January-June, pp.1–16.

Venkatesh, V., Brown, S. A., Bala, H. 2013. Bridging the Qualitative-Quantitative Divide: Guidelines for Conducting Mixed Methods Research in Information Systems. *MIS Quarterly*, 31:1, pp.21–54.