First Monday, Volume 22, Number 4 - 3 April 2017

PEER-REVIEWED JOURNAL ON THE INTERNET

Big data and learning analytics: Singular or plural?

by Anna Wilson, Terrie Lynn Thompson, Cate Watson, Valerie Drew, and Sarah Doyle

Abstract

Recent critiques of both the uses of and discourse surrounding big data have raised important questions as to the extent to which big data and big data techniques should be embraced. However, while the context-dependence of data has been recognized, there remains a tendency among social theorists and other commentators to treat certain aspects of the big data phenomenon, including not only the data but also the methods and tools used to move from data as database to data that can be interpreted and assigned meaning, in a homogenizing way. In this paper, we seek to challenge this tendency, and to explore the ways in which explicit consideration of the plurality of big data might inform particular instances of its exploitation. We compare one currently popular big data-inspired innovation — learning analytics — with three other big data contexts — the physical sciences, business intelligence and public health. Through these comparisons, we highlight some dangers of learning analytics implemented without substantial theoretical, ethical and design effort. In so doing, we also highlight just how plural data, analytical approaches and intentions are, and suggest that each new big data context needs to be recognized in its own singularity.

Contents

1. Introduction

One result of the increasing role of computers, the Internet and digital data across many forms of human endeavor is the widespread uptake of the notion of "big data." This term has gained prominence in academic and applied research, in public debate and in political and policy rationales. More recently, it has been the subject of some important critiques, especially in relation to perceptions of objectivity/neutrality (boyd and Crawford, 2012; Vis, 2013; Markham, 2013); the related notion that high statistics data inevitably leads to more reliable, better understanding (boyd and Crawford, 2012); ethical difficulties around surveillance and inequities of access (boyd and Crawford, 2012; Crawford, *et al.*, 2014; Nunan and Di Domenico, 2013); limitations on access and issues of ownership (Zelenkauskaite, 2016); and the creation of a scholarly divide in relation to research capacities, due to technical expertise, cost of access and the proprietary status of data (Zelenkauskaite and Bucy, 2016). Less frequently, but equally importantly, critiques have been directed at a lack of theorization around big data (Boellstorff, 2013; Markham, 2013); questions around what shapes big (social) data (Markham, 2013; Vis, 2013); and the problems of context-dependence (boyd and Crawford, 2012).

Despite these important critiques, the notion of big data still seems to exert a peculiarly attractive force, leading many businesses, institutions and policy-makers to approach it as if it were some sort of panacea that will inevitably increase economic competitiveness and cure social and organizational ills. This may in part be because, despite the recognition of different stakeholders and contexts by some authors (boyd and Crawford, 2012; Helles and Jensen, 2013; Markham, 2013; Zelenkauskaite and Bucy, 2016) there is still a tendency to talk about "big data" and "big data methods" in a homogenizing way. In particular, there is often an underlying implication that — whether you are "for" or "against" the exploitation of big data in a particular context — its transformation from database to a more structured, processed and interpretable form is both automatic and unproblematic. In this paper, we argue that to speak in this way masks another important aspect of the big data discourse that needs to be opened up to critique. Taking the current proliferation of big data-inspired learning analytics programs in the higher education and, increasingly, school sectors as our starting point, we take up boyd and Crawford's (2012) challenge to 'ask difficult questions of big data's models of intelligibility" [1]. We do so by considering three different big data contexts and by comparing them to learning analytics.

2. Big data: Singular or plural?

Our case for reflecting more deeply on the notion of big data might be encapsulated in the way that the word “data” itself is sliding from a plural to a singular noun. We believe that a similar conceptual slippage, which positions big data as something singular, is misguided and potentially misleading. By failing to distinguish between fundamentally different types of data and context, it permits thinking to be side-stepped in favor of the machine-performance of automated, and automatically-reached for, regression analyses on large data sets. It also lends the results of these the weight of rigorous, well-designed scientific analyses. We suggest that the current rush to implement learning analytics is a good example of where a homogeneous conception of big data and big data methods may lead to unintended and potentially negative consequences. To explain our position, we first need to clarify what is generally meant by “big data”.

There have been several definitions of big data put forward. Some are based on technical requirements: for example, Provost and Fawcett (2013) define big data as ‘datasets that are too large for traditional data-processing systems and that therefore require new technologies’ [2], while Carter (2011) and Gartner (2013) define big data as data whose volume, velocity and variety require the development of ‘cost-effective, innovative forms of information processing for enhanced insight, decision-making and process automation’ [3]. De Mauro, *et al.* (2015) note that dig data is used to describe any of ‘social phenomenon, information assets, data sets, analytical techniques, storage technologies, processes and infrastructures’ [4] before offering what they assert is a consensual definition: ‘Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value’ [5].

Variety, in these definitions, may be taken as a reference to the multi-parameter nature of big data data sets, as information relating to a range of variables may be associated with each individual case (member of the sample or population). That is, the data are about something which can be characterized by a set of variables. Such definitions do not limit big data to include only quantitative information, but the tendency to assume that big data analyses will be algorithmic and automated — that the data will be parsed by software rather than by individual researchers “reading” every case — means that big data is usually conceived of as comprising sets of numerical metrics or indicators.

Moving beyond purely technical definitions, boyd and Crawford (2012) offer a definition that has been widely taken up by authors interested in the uses of big data in a variety of fields (see, for example, Crampton, 2015; Lewis and Westlund, 2015; Ovadia, 2013; Shin and Choi, 2015; Wamba, *et al.*, 2015). They define big data as:

> a cultural, technological, and scholarly phenomenon that rests on the interplay of:
>
> 1. Technology: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
>
> 2. Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
>
> 3. Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy. [6]

While such a definition offers a substantial elaboration on those that rely only on technical factors, it is still an intentionally broad definition, allowing for data generated in a myriad of ways, for a myriad of purposes, in a myriad of contexts. In their introduction to *First Monday’s* 2013 special edition dedicated to big data, Helles and Jensen (2013) further emphasize the non-specificity of the term, noting that:

> Data are made in a process involving multiple social agents — communicators, service providers, communication researchers, commercial stakeholders, government authorities, international regulators, and more. Data are made for a variety of scholarly and applied purposes ... And data are processed and employed in a whole range of everyday and institutional contexts. [7]

However, despite this clear recognition that big data are not a singular phenomenon but rather are intrinsically plural, there remains a tendency to talk about big data and big data methods in homogenizing or monolithic terms. In its most overt and extreme form, this attitude is exemplified in Anderson’s (2008) much-quoted assertion that we are now in a world:

> where massive amounts of data and applied mathematics replace every other tool that might be brought to bear ... Who knows why people do what they do? The point is they do, and we can track it and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves. [8]

Even those determined to critique such an overtly anti-theoretical (or at best behaviouristic) stance often neglect the implications of the sheer variety of data types, forms and contexts caught up in the big data net. Thus even as they critique the big data discourse, boyd and Crawford (2012) speak of '[c]omputer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and other scholars' [9] in the same breath, and group together 'genetic sequences, social media interactions, health records, phone logs, government records [10] as 'digital traces left by people' [11]. Similarly, they speak of automated analyses carried out by 'algorithms that can extract and illustrate large-scale patterns' [12], and despite taking her starting point that 'data is a deceptively easy term to toss around' [13], Markham refers to the 'shift to computational tools and methods for analyzing large sets of information' [14].

Such discussions leave rather unexamined the variation in these algorithms, tools and methods, and perhaps more importantly, the reasons for this variation. They not only gloss over the fact that there are different schools within the field of statistical analysis (frequentist, Bayesian) which are underpinned by what are effectively different epistemologies (Vanderplas, 2014) — they also fall into one of the very traps they wish to avoid, by assuming or implying that, while the contexts in which big data are generated and collected may vary, and the interests and power relationships that determine their interpretations and use may need to be challenged, the methods used to process them are both universal and unproblematic. This contributes to the mythology that is the third part of boyd and Crawford's (2012) definition. In fact, the methods used to analyze large statistical data sets are themselves multiple, rather than singular. The type and context of the data govern the choices of the type of relevant analysis (for example, regression, significance testing, network analysis, frequency domain analysis, edge detection, decomposition, deconvolution, or machine learning using *e.g.*, artificial neural nets). What is more, in traditional (regression) analyses, although algorithms can be written to extract patterns, the researcher or programmer constructing the algorithm must have imagined the pattern to be possible; and while advances in machine learning mean that algorithms can discover as well as illustrate patterns, the data on which they operate need to be understood in detail if spurious patterns are to avoided. In the following, we consider the case of automated analyses of electronic trace data left by students engaged in formal learning — so-called learning analytics — in an attempt to illustrate how the nature of data in a particular big data context, the intentions behind gathering and analyzing the data, and the appropriate analysis methods, all merit careful consideration.

3. Learning analytics: A case in point

'Learning analytics' is the name for a big data-inspired innovation (Baker and Inventado, 2014) being implemented in educational systems in many developed countries. It describes attempts to use digital data about students' backgrounds and learning behaviors in online contexts to monitor and predict student performance.

Following the widespread adoption of learning management systems (LMSs), systems vendors and educational institutions realized that they gave access to large quantities of trace data about students' online actions. In the higher education sector, LMSs were initially designed as content management systems through which resources were made available to students, allowing digital provision of course outlines and lecture notes, podcasts and vodcasts of lectures and other pre-recorded instructor-produced resources. Over recent years they have increasingly taken on some of the features of social media platforms, with the intention of facilitating social learning and self-expression through *e.g.*, discussion forums, blogs and cooperative tasks such as the creation and population of wikis. As Web analytics developed, the suppliers of LMSs began to incorporate tracking and reporting functions within their products (Sclater, 2014). Universities and other educational institutions increasingly view these as providing metrics that directly relate to student learning — that, is, as being meaningful analytics of learning.

The Society for Learning Analytics Research (https://solaresearch.org) defines learning analytics as 'the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs' [15]. In practice, however, learning analytics is used to label something rather more specific: what Clow (2013) describes as 'the application of ... big data techniques' [16] to data including 'demographic information, online activity, assessment data and final destination data' [17] to produce metrics which may be presented to student or tutor/instructor in order that they may improve their learning and teaching respectively. Although it is not inevitable that data gathered with the intention of analyzing learning be purely quantitative, it is clear from the emphasis on measurement and metrics that current conceptions of learning analytics rely on numerical data. Indeed, even where attempts are made to include qualitative data, they end up being translated into numbers. For example, in Australia, the University of New England developed a system allowing students to register their emotional reaction to units and subjects using emoticons — which were then counted, and effectively assigned numerical values, in order to be used in a learning analytics system intended to safeguard retention (Nelson and Creagh, 2013).

In much of the writing on the topic, there seems to be an implicit assumption that the mass gathering of what amounts to surveillance data is as inevitable in educational contexts as it is in matters of state security. The technical ease and low cost of tracking and storing online activity records, coupled with the normalization of metrics as the primary means of guaranteeing evidence-based practice, have led to a rapid and widespread adoption of systems that allow detailed tracking of students' activities in institutional LMSs. High hopes for the transformational power of this new evidence base are frequently expressed; for example, in their recent review, Greller and Drachsler (2012) suggest that 'these educational datasets offer unused opportunities for the evaluation of learning theories, learner feedback and support, early warning systems, learning technology, and the development of future learning applications' [18].

As is the case with other forms of high volume internet trace data, learning analytics data are often presented as part of an undifferentiated trend, with Clow (2013), for example, relating learning analytics data to data from research in the physical sciences, geographical location data and data in business intelligence. Like boyd and Crawford (2012), he seems to imply that there is a singular thing (big data) of which LMS data is one instance. He also implies that there are automatically applicable 'big data techniques' [19] that can be used to make sense of any such data. Echoing Anderson (2008), Clow (2013) suggests there is a feeling that '[the] volume and scope of data can be so large that it is possible to start with a data-set and apply computational methods to produce results and to seek an interpretation or meaning only subsequently' [20]. But as we have already suggested, there are many and varied approaches to analyzing large statistical data sets, the appropriateness of each depending on the nature of the data and what is hoped will be learned from it. Careful decisions must therefore be made about how learning analytics data are to be analyzed and what metrics might prove meaningful, reliable and significant.

In the following, we consider three different types of big data in some detail, in an attempt to draw out the ways in which data and analytics in an educational context differ from and are similar to other types of data processing. We consider big data in three contexts: (i) the physical sciences, using the example of data generated at the Large Hadron Collider; (ii) business intelligence; and (iii) public health. The first comparison is intended to highlight the substantial differences between two data contexts and associated analytical approaches. The second comparison is intended to highlight how wholesale transfer of analytic approaches from one context to another might call for serious caution. The third comparison is intended to highlight a possible elision between the macro and the micro — between the plural of population and the singular of the individual. Together, the comparisons illustrate some of the potential dangers of a discourse that presents plurality — whether in data types, intentions, or methods — as singularity.

3.1. Big data in the physical sciences

The first type of big data that we shall consider — one which is perhaps least similar to learning analytics data — is that generated in some branches of the physical sciences. Despite their frequently being lumped together under the big data label (see, for instance, boyd and Crawford, 2012; Clow, 2013), data such as those generated at the Large Hadron Collider have little in common with the data collected in learning analytics, social media analysis or business intelligence beyond their digital nature, the large number of parameters that may be associated with each case, the high volumes collected and the need for substantial computing processing power leading to developments in approaches to computer programming such as the use of massively parallel systems.

One of the most important differences between the big data and analysis of the Large Hadron Collider, and that comprising learning analytics, is that between intentional and incidental observation. Scientists involved in the design and implementation of detection and data acquisition systems at the Large Hadron Collider have well-developed, detailed and precise predictive theories describing what occurs when two accelerated, high-energy ions collide at the focal point of one of the detector systems. These theories predict both the types of particles that may be created during such a collision and the various combinations, permutations and correlations of particles that may be emitted as these initial products decay. Typically, the particles created in the initial collision exist for such short timescales that they decay almost instantaneously; they are not directly observed. The detection systems are designed to identify and characterize the longer-lived particles produced in the subsequent decays — proxies for the initially created particle — as efficiently and inclusively as possible. The data processing stage is designed to take the information provided by the detectors and work backwards, using the rules and variables provided by the underlying theories, to reconstruct what happened in the initial collision and infer what particles might have been initially produced.

These experiments are designed to produce and detect well-understood signals, proxies that indicate what was occurring in the inaccessibly short time period when, perhaps, a Higgs boson was fleetingly brought into existence. The software and algorithms used in the data analysis are rigorously tested using simulated data, which can only be generated because of the scientists' robust understanding of the ways that data are generated from electronic signals produced following interactions between particles and material elements of the detector system. The intention is to falsify or generate support for hypotheses of an extremely narrow, well-defined nature, which may be crucial steps to the falsification or support of a higher level of theory (the Standard Model).

This contrasts with the processing of student data in learning analytics. First, the LMSs through which much learning analytics data are obtained are not designed to generate and collect signals that we know *a priori* correlate with that mysterious education Higgs boson, observable or provable learning. They are designed, first and foremost, as content management systems (databases or repositories of learning resources), and second, as quasi-social media platforms. The massive and lengthy collaborative design process that preceded the construction of the Large Hadron Collider (see, for example, Science & Technology Facilities Council (STFC), n.d.) and its associated experimental stations initially involved groups of experts deciding what needs to be observed in order to probe something that is itself directly non-observable. The subsequent design and construction processes deliberately aimed at achieving these goals. The collection of incidental data such as click records and page hits generated by student visits to pages on course/module sites is hardly comparable.

Another important difference lies in the nature of the populations that are being sampled. During the LHC experiments, the same collision conditions are repeated again and again, between particles that are themselves fundamentally identical, so that the range of possible outcomes can be studied. This is in stark contrast with students' experience of learning, where each student progressing through a module is unique, and each time a student accesses an online resource is also unique, surrounded by a past and future of different prior experiences, potential connections and intentions.

There are significant differences in relation to the analytical techniques required to process the data, too. As described above, the techniques used to analyze LHC data are reconstructive — they seek to identify something that has already occurred. The aim is usually to identify statistically infrequent occurrences within a much larger population of much more frequent, generally well-understood (and thus uninteresting) events. The data sets are big often not because the scientists wish to gather more data on a previously identified phenomenon or trend, but because they are interested in a one in a trillion or rarer event. The vast majority of the data are, ultimately, filtered out or discarded, and the most common form of statistical analysis is statistical testing of actual observations against theoretical or simulated predictions. In contrast, the techniques used to analyze LMS data attempt to be predictive and prescriptive — to take patterns of behavior exhibited by a single individual over time and use these to predict whether the student is "at risk", and possibly to generate a prescription for remedial action.

Thus although it is true to say that statistical methods are used in both big science data analytics and learning data analytics, the comparison pretty much ends there. However, explicitly making the comparison serves to surface some important points — for example, the crucial role of theory and experimental design when analyzing multi-parameter data in the hope of observing correlations, and the differences between statistical data sets obtained by repeating identical conditions many times, compared to those where every member of the set is unique and inherently unrepeatable.

3.2. Big data in business intelligence

The previous section set out to remind readers of just how differentiated big data activities can be. In this section, we compare learning analytics to perhaps their closest big data relative. Learning analytics implementations are mostly borrowed from data collection and processing models developed in the field of business intelligence. Indeed, one of the most attractive arguments for learning analytics is the notion of personalized learning (a term that now resonates with personalized advertising). By putting the learner firmly at the center of their own learning, personalized learning is expected to 'reduce delivery costs while at the same time creating more effective learning experiences, accelerating competence development, and increasing collaboration between learners' [21]. In particular, personalization is expected to accommodate 'the diversity of learning' [22] more effectively than 'current learning environments' [23]. Who would argue against an innovation that accomplished this? The problem, however, is that the translation from personalized advertising to personalized learning is not as transparent as it might seem. To explain, we first consider some of the ways in which business intelligence analytics operate to produce personalized advertising.

Business intelligence analytics has arisen from the availability to online businesses and service providers of vast amounts of data about their users' online activities. Its approaches are predicated on business, economic and marketing theories, including the (not completely unreasonable) assumption that users' interests and habits can be inferred from their past activities, and that sufficient similarities persist among specific, identifiable groups characterized by factors such as culture, socio-economic status, age and gender so as to allow inferences to be broadly applicable to other members of the same group. Such theories thus mix some *a priori* assumptions about correlations between different variables with an openness to and search for new and emerging patters. Analyses are based on both simple frequency models and on more complex multi-parameter regression models indicating correlation between one online behavior (*e.g.*, the lead author uses her mobile phone to check the weather forecast for Stirling) and another (the lead author checks train times between Stirling and Edinburgh using the same device) to infer a likely characteristic (the lead author is or soon will be in the Stirling area — a probability that can be inferred even though she has disabled geolocation data on her phone).

Business intelligence analytic systems collect large amounts of data about individual users, based on unique identifiers such as their IP address; their computer's MAC address; the use of personal accounts held with companies such as Google, Facebook, eBay and online retailers; loyalty card schemes; and the terms-and-conditions that users probably do not read when downloading an app to their mobile device, that in fact mean they agree to share a substantial amount of information stored on that device with the provider of the app, and even to cede control of some device functionalities. These allow for personalized advertising — such as advertisements for businesses and services in the local area (childcare in nearby Falkirk flashing up on the lead author's National Rail Enquiries query), or targeted special offers, such as advertising encouraging customers to return to regular purchases of something they had stopped buying (perhaps because they had started to acquire it from a competitor).

The field of business intelligence studies has also produced a number of techniques for identifying users' likely socio-cultural constituency, allowing the preference and habits of groups that a user may identify or share values with to be suggested to that user as new services to use or products to buy. Add to the lead author's already-discovered Stirling location her recent Google searches and one may infer that she is reading *Romola* and trying to recall the lyrics of a Hazel O'Connor track, which may indicate that she is female and of a certain age. This allows for further targeted advertising (perhaps this is why National Rail Enquiries is so insistent on informing her about childcare), but also for the now ubiquitous recommender systems that tell us that "people who bought this also bought ... ."

There are several aspects to this that might be causes for concern in educational contexts. For example, one of the consequences of all this data gathering is that more and more data are gathered: if, for example, Google's data becomes apparently more valuable the bigger it becomes, then Google is likely to engage in closer and closer surveillance of users of all of its services. This increases the pertinence of questions about the ethics of surveillance, as raised by boyd and Crawford (2012) and others; questions that may be doubly important where use of the surveilling system cannot be opted out of (as in the case of students' use of LMSs); where no permission has been given by explicit, even if uninformed, acceptance of terms & conditions; and where users of the systems may be minors (as in schools, colleges and less frequently universities).

Another reason to think carefully about the transfer of business intelligence models to learning analytics relates to underlying purpose and intended beneficiary. (Analogous concerns have been raised (Zelekauskaite, 2016) in a critique of social media big data for users of commercial media platforms.) At their core, business intelligence processes are designed to get shoppers to both buy more and use more services from the companies collecting the user data and doing the analysis (or from those who pay them to do so). The underlying premise is that user behaviors are both predictable and malleable: that is, that customers can be nudged or persuaded into behaviors that benefit the seller. Let us be clear: the aim of business intelligence analytics is to maximize sales and profit. Targeted advertisements may mean that we spend less time browsing, comparing the products made by competing brands, for example, and so have more time to do other things (or to buy more products). Thus there may be some benefit to the analyzed user — but it is a concomitant benefit, not the designed aim of the analytical process. This kind of thinking, however, underpins a great deal of the promotion of learning analytics — it is assumed that, just as customers can be guided towards better shopping habits, students can be guided towards better learning habits. 'Just as Amazon.com uses the data from our purchase history to make suggestions about future purchases, so can learning analytics allow us to suggest new learning opportunities or different courses of action to our students' (Dietz-Uhler and Hurn, 2013).

One might also question how effective these personalization systems are — why, the lead author has to ask, would a childless person want to know about childcare? (Even more inappropriate targeted advertising based on personal health data is described by Ebeling, 2016.) Research suggests that the effectiveness is strongly dependent on the choice of algorithm used: for example, one study found that the recommender systems used by two online book retailers allow those retailers to increase prices as well as sales volume (Pathak, *et al.*, 2010), while another showed that one widely used algorithm has no effect at all on sales volume (Lee and Hosanagar, 2014). This highlights the fact that the computational techniques themselves are complex, varied and not always well-proven. And if automatically produced recommendations are not reliably helpful, do we risk recommending actions that are in fact detrimental to an individual student's learning? Concerns have also been raised that automated personalization in the online business context may result in increased insularity and fragmentation, leading users to more of the same rather than exposing them to new ideas and cultures beyond their existing interests and biases (Sunstein, 2007). Empirical evidence, however, suggests that personalization instead results in increased homogenization (Hosanagar, *et al.*, 2013) — but by guiding a large number of people to a limited number of popular resources, rather than by increasing the range of resources and ideas that users are exposed to (Lee and Hosanagar, 2014). Similar concerns have been expressed regarding learning analytics, with even strong proponents of learning analytics noting that 'aligning and regulating performance and behavior of individual teachers or learners against a statistical norm without investigating the reasons for their divergence may strongly stifle innovation, individuality, creativity and experimentation' [24].

It may be that desires to optimize users' shopping behaviors have something in common with desires to optimize students' learning behaviors, but it is not immediately obvious how deep such a comparison runs, or how desirable it might be. In many ways, the use of business intelligence-derived analytical algorithms and automated feedback procedures presupposes that there are such things as generically (or group-specifically) desirable learning behaviors; that we know or can tell from click data what those behaviors are; and that they can be (reasonably accurately) measured by the data available to analytical implements attached to or embedded in LMSs. These assumptions are somewhat problematic as we have shown in our recent work experimenting with learning analytics based on a socio-material perspective (Wilson, *et al.*, 2017). Proponents of large-scale, institutional uptake of learning analytics packages based on business intelligence approaches might argue that some studies have shown improved student outcomes when such systems are implemented. However, the findings in this regard are highly variable (Wilson, *et al.*, 2017), and where interventions are unambiguously successful they appear to be based on measurements of attendance or activity rather than learning. Purdue's Course Signals, for example, which has been shown to improve student retention by alerting students who were relatively inactive on their course sites compared to their peers (Arnold and Pistilli, 2012), might more accurately be described as activity analytics than learning analytics. The very fact that increased attention or participation can lead to improved performance suggests that the problems, if any, experienced by the students at risk in these situations had little to do with learning per se, and perhaps originated in factors such as competing commitments, time management, or feelings of alienation. If a warning that you are not spending enough time online is sufficient to improve your commitment and performance, you are probably not experiencing significant conceptual difficulties with the material you are trying to learn. This raises questions not only about the data processing algorithms, but also about the nature of 'learning data' itself.

Thus, the transfer of the statistical procedures and analytical algorithms developed for business intelligence to the context of learning in higher education may be technically straightforward, but there are surely some differences between the underlying intentions and nature of the data that might give pause to institutions thinking about implementing them.

3.3. Big data in public health

The final area in which mass data gathering is informing decision-making, policy implementation and intervention that we consider is public health. While the commoditization of healthcare data and its use in personalized, targeted marketing has been powerfully described by Ebeling (2016), here we focus on non-commercial interventions aimed at specific target groups and populations rather than individuals. Analytics in this context, like those employed in business intelligence, use multi-parameter regression in the hope of correlating particular health outcomes (disease occurrence and health risks) with different physical, lifestyle and socio-economic indicators. The data relating to these variables are gathered at the individual level, from national health service records or social surveys, but aggregated statistically. The idea is that the correlations identified in the analyses allow experts (health practitioners, service managers, public servants involved in allocating health budgets, and so on) to direct interventions where they are most needed. Interventions might be targeted geographically — for example, locally organized walking groups in areas where levels of obesity are high — or at particular social subgroups or subcultures — for example, anti-drug campaigns targeted at vulnerable youth groups. Recently, it has been suggested that internet trace data might also be incorporated into public health analyses and decision-making — for example, Brownstein, *et al.* (2009) suggested using regional variations in the frequencies of Google searches for particular terms, such as "diarrhea" and "food poisoning", for early disease detection.

Like business analytics, such public health interventions are predicated on the idea that human behavior can be largely predicted by socio-cultural group; and that humans can be persuaded to alter their behavior. However, unlike business analytics, which generally seeks to get users to do more of the same, the alterations aimed for in data-driven public health initiatives are likely to go against the inertia of habit and peer pressure, and require conscious commitment to change.

The main part of this thinking that has been assumed in learning analytics approaches is that particular (online) activities combined with demographic data might be indicators for learning outcomes. In common with health interventions, learning analytic implementations are intended to produce changed behaviors that result in better outcomes. One important difference is that, at least to date, big data in public health has not used the individualized tracking and recommendation techniques developed for business intelligence to nudge at the level of the individual person, while learning analytics seeks to do just this. Even those authors who advocate putting the combined might of personal health records and social determinant data in the hands of patients to increase personalized care do not envisage automated systems to deliver that care; rather they see such combined records as enhancing doctor-patient interactions and relationships (Murdoch and Detsky, 2013). This highlights an important distinction between information and tests generated or carried out at the population level and those carried out at the level of the (unique) individual — another example of the possible confusion between the plural and the singular in learning analytics approaches.

Another important thing to note is that the most successful public health campaigns have tended to revolve around issues that are presented as rather black and white (Hornik, 2002; Randolph and Viswanath, 2004) — where there is a binary choice, with one decision leading to increased likelihood of illness and death — and where campaigns have been supplemented by punitive rates of taxation. The anti-smoking campaign is a case in point (Townsend, *et al.*, 1994). More complex programs have yielded less impressive, occasionally ambiguous results. For example, concerns about growing levels of obesity led to the introduction of nutrition labelling on food packages — a form of guidance or feedback to the consumer that might be compared to the guidance derived from learning analytics. However, a substantial meta-review of studies carried out in the U.S. and Northern Europe (Cowburn and Stockley, 2005) suggested that consumers were not making use of the information being provided. The authors of this review suggest that this was due to difficulties in interpreting the data — what might now be termed as a need for nutritional data literacy. The multiple traffic light system currently in use in the U.K., Australia and elsewhere was introduced in response to these findings; however, point-of-sale studies in major food retailers have shown that this intentionally more user-friendly guidance has had no impact on sales of healthy versus unhealthy foods (Sacks, *et al.*, 2009; Sacks, *et al.*, 2011). Other studies have shown that while consumers have a good understanding of the information conveyed in nutrition labels, understanding and choice do not correlate (Grunert, *et al.*, 2010). Obesity levels — the primary target of this data-driven intervention — have not apparently fallen as a result. This has led to increased calls for taxation and/or regulation — that is, moves to change and regulate the providers, rather than rely on informed behavior change among consumers. The lesson for learning analytics may be that they should be used to change behavior at the level of provision, as much as or more than at the level of consumption.

4. Conclusions

In this paper, we have used three comparisons to draw out ways in which a particular and currently popular use of big data — learning analytics — differs from other big data contexts. We have tried to highlight the lessons for learning analytics that emerge from such explicit comparisons.

First, the comparison with big data in the physical sciences highlighted the difference between intentional and incidental data collection; the need for robust theoretical underpinnings to measurements that rely on proxies rather than direct observation; the further need for robust theoretical underpinnings to explain or predict the presence and absence of correlations between different variables; and the dangers of comparing the infinitely differentiated interactions between students and learning resources with the endlessly repeatable interactions between fundamentally identical particles.

Second, the comparison with big data in business intelligence highlighted the ethical issues around collecting data when users have no choice to opt out and do not give explicit or informed consent; issues around who business intelligence algorithms such as recommender systems are intended to benefit; the question of whether recommender systems are reliable (and if they are not, might they end up recommending behavior that is in fact detrimental to students); and questions of whether the digital trace data used in Learning Analytics are actually traces of learning at all.

Third, the comparison with big data in public health highlighted the possible elision between the levels of population and individual; and the question of whether analytics are best used to change the behavior of consumers (students) or the conditions in which they find themselves.

Through these comparisons, we have sought to draw attention to the need to recognize the plurality of both big data, and methods for processing and analyzing big data; and the singularity of every big data context. We hope that this will encourage more nuanced discussions of big data, and more thoughtful analyses of the different contexts in which large volumes of data may be available and the different uses to which they might be put.

About the authors

Anna Wilson is a post-doctoral researcher in the Division of Sociology at Abertay University.
E-mail: a [dot] wilson [at] abertay [dot] ac [dot] uk

Terrie Lynn Thompson is a Lecturer in Digital Media Professional Education in the Faculty of Social Sciences at the University of Stirling
E-mail: terrie-lynn [dot] thompson [at] stir [dot] ac [dot] uk

Cate Watson is a Professor in Professional Education in the Faculty of Social Sciences at the University of Stirling.
E-mail: cate [dot] watson [at] stir [dot] ac [dot] uk

Valerie Drew is a Senior Lecturer in Professional Education in the Faculty of Social Sciences at the University of Stirling.
E-mail: v [dot] m [dot] drew [at] stir [dot] ac [dot] uk

Sarah Doyle recently graduated with a PhD from the Faculty of Social Sciences at the University of Stirling.
E-mail: sarah [dot] doyle [at] stir [dot] ac [dot] uk

Notes

[1] boyd and Crawford, 2012, p. 666.

[2] Provost and Fawcett, 2013, p. 54.

[3] Gartner, 2013, n.p.

[4] De Mauro, *et al.*, 2015, p. 101.

[5] De Mauro, *et al.*, 2015, p. 103, original capitalization.

[6] boyd and Crawford, 2012, p. 663.

[7] Helles and Jensen, 2013, n.p. We note that Helles and Jensen's comments apply equally to data in general, and are not restricted to digital big data.

[8] Anderson, 2008, n.p.

[9] boyd and Crawford, 2012, p. 663.

[10] *Ibid.*

[11] boyd and Crawford, 2012, p. 663.

[12] boyd and Crawford, 2012, p. 664.

[13] Markham, 2013, n.p.

[14] *Ibid.*

[15] Quoted in Siemens and Gaevi, 2012, p. 1.

[16] Clow, 2013, p. 684.

[17] Clow, 2013, p. 685.

[18] Greller and Drachsler, 2012, p. 43.

[19] Clow, 2013, p. 684.

[20] *Ibid.*

[21] Greller and Drachsler, 2012, p. 42.

[22] Greller and Drachsler, 2012, p. 53.

[23] *Ibid.*

[24] Greller and Drachsler, 2012, p. 47.

References

C. Anderson, 2008. "The end of theory, will the data deluge makes the scientific method obsolete?" *Edge* (30 June), at http://www.edge.org/3rd_culture/anderson08/anderson08_index.html, accessed 18 March 2017.

K.E. Arnold and M.D. Pistilli, 2012. "Course signals at Purdue: Using learning analytics to increase student success," *LAK '12: Proceedings of the Second International Conference on Learning Analytics*

and Knowledge, pp. 267–270.
doi: http://dx.doi.org/10.1145/2330601.2330666, accessed 20 March 2017.

R.S. Baker and P.S. Inventado, 2014. "Educational data mining and learning analytics," In: *Learning analytics: From research to practice*. New York: Springer, pp. 61–75.
doi: http://dx.doi.org/10.1007/978-1-4614-3305-7_4, accessed 20 March 2017.

T. Boellstorff, 2013. "Making big data, in theory," *First Monday*, volume 18, number 10, at http://firstmonday.org/article/view/4869/3750, accessed 20 March 2017.
doi: http://dx.doi.org/10.5210/fm.v18i10.4869, accessed 20 March 2017.

d. boyd and K. Crawford, 2012. "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," *Information, Communication & Society*, volume 15, number 5, pp. 662–679.
doi: http://dx.doi.org/10.1080/1369118X.2012.678878, accessed 20 March 2017.

J.S. Brownstein, C.C. Freifeld and L.C. Madoff, 2009. "Digital disease detectionharnessing the Web for public health surveillance," *New England Journal of Medicine*, volume 360 (21 May),pp. 2,153–2,157.
doi: http://dx.doi.org/10.1056/NEJMp0900702, accessed 20 March 2017.

P. Carter, 2011. "Big data analytics: Future architectures, skills and roadmaps for the CIO," *IDC white paper*, at http://www.sas.com/resources/asset/BigDataAnalytics-FutureArchitectures-Skills-RoadmapsfortheCIO.pdf, accessed 20 March 2017.

D. Clow, 2013. "An overview of learning analytics," *Teaching in Higher Education*, volume 18, number 6, pp. 683–695.
doi: http://dx.doi.org/10.1080/13562517.2013.827653, accessed 20 March 2017.

G. Cowburn and L. Stockley, 2005. "Consumer understanding and use of nutrition labelling: A systematic review," *Public Health Nutrition*, volume 8, number 1, pp. 21–28.
doi: http://dx.doi.org/10.1079/PHN2004666, accessed 20 March 2017.

J.W. Crampton, 2015. "Collect it all: National security, big data and governance," *GeoJournal*, volume 80, number 4, pp. 519–531.
doi: http://dx.doi.org/10.1007/s10708-014-9598-y, accessed 20 March 2017.

K. Crawford, M.L. Gray and K. Miltner, 2014. "Critiquing big data: Politics, ethics, epistemology," *International Journal of Communication*, volume 8, at http://ijoc.org/index.php/ijoc/article/view/2167, accessed 20 March 2017.

A. De Mauro, M. Greco and M. Gimaldi, 2015. "What is big data? A consensual definition and a review of key research topics," *AIP Conference Proceedings*, volume 1644, number 1, at http://aip.scitation.org/doi/abs/10.1063/1.4907823, accessed 20 March 2017.
doi: http://dx.doi.org/10.1063/1.4907823, accessed 20 March 2017.

B. Dietz-Uhler and J.E. Hurn, 2013. "Using learning analytics to predict (and improve) student success: A faculty perspective," *Journal of Interactive Online Learning*, volume 12, number 1, pp. 17–26, and at http://www.ncolr.org/issues/jiol/v12/n1/using-learning-analytics-to-predict-and-improve-student-success, accessed 20 March 2017.

M.F.E. Ebeling, 2016. *Healthcare and big data: Digital specters and phantom objects*. New York: Palgrave Macmillan.
doi: http://dx.doi.org/10.1057/978-1-137-50221-6, accessed 20 March 2017.

R.C. Hornik, 2002. "Public health communication: Making sense of contradictory evidence," In: R.C. Hornik (editor). *Public health communication: Evidence for behavior change* Mahwah, N.J.: L. Erlbaum Associates, pp. 1–22.

Gartner, 2013. "Big data," *IT glossary*, at http://www.gartner.com/it-glossary/big-data/, accessed 20 March 2017.

W. Greller and H. Drachsler, 2012. "Translating learning into numbers: A generic framework for learning analytics," *Educational Technology & Society*, volume 15, number 3, pp. 42–;57, and at http://ifets.info/journals/15_3/4.pdf, accessed 20 March 2017.

K.G. Grunert, J.M. Wills and L. Fernández-Celemn, 2010. "Nutrition knowledge, and use and understanding of nutrition information on food labels among consumers in the UK," *Appetite*, volume 55, number 2, pp. 177–189.
doi: http://dx.doi.org/10.1016/j.appet.2010.05.045, accessed 20 March 2017.

R. Helles and K.B. Jensen, 2013. "Making data — Big data and beyond: Introduction to the special issue," *First Monday*, volume 18, number 10, at http://firstmonday.org/article/view/4860/3748, accessed 20 March 2017.
doi: http://dx.doi.org/10.5210/fm.v18i10.4860, accessed 20 March 2017.

K. Hosanagar, D. Fleder, D. Lee and A. Buja, 2013. "Will the global village fracture into tribes? Recommender systems and their effects on consumer fragmentation," *Management Science*, volume 60, number 4, pp. 805– 823.
doi: http://dx.doi.org/10.1287/mnsc.2013.1808, accessed 20 March 2017.

D. Lee and K. Hosanagar, 2014. "Impact of recommender systems on sales volume and diversity," *Proceedings of ICIS 2014: International Conference on Information Systems (Auckland, New Zealand)*, at http://aisel.aisnet.org/icis2014/proceedings/EBusiness/40/, accessed 20 March 2017.

S.C. Lewis and O. Westlund, 2015. “Big data and journalism: Epistemology, expertise, economics, and ethics,” *Digital Journalism*, volume 3, number 3, pp. 447–466.
doi: http://dx.doi.org/10.1080/21670811.2014.976418, accessed 20 March 2017.

A.N. Markham, 2013. “Undermining ‘data’: A critical examination of a core term in scientific inquiry,” *First Monday*, volume 18, number 10, at http://firstmonday.org/article/view/4868/3749, accessed 20 March 2017.
doi: http://dx.doi.org/10.5210/fm.v18i10.4868, accessed 20 March 2017.

T.B. Murdoch and A.S. Detsky, 2013. The inevitable application of big data to health care, *Journal of the American Medical Association*, volume 309, number 13 (3 April), pp. 1,351–1,352.
doi: http://dx.doi.org/10.1001/jama.2013.393, accessed 20 March 2017.

K.J. Nelson and T.A. Creagh, 2013. *A good practice guide: Safeguarding student learning engagement*. Sydney: Australian Government, Office for Learning and Teaching, and at http://www.olt.gov.au/system/files/resources/CG10_1730_Nelson_Good_Practice_Guide_2012.pdf, accessed 20 March 2017.

D. Nunan and M. Di Domenico, 2013. “Market research and the ethics of big data,” *International Journal of Market Research*, volume 55, number 4, pp. 505–520, and at https://www.mrs.org.uk/ijmr_article/article/98860, accessed 20 March 2017.

S. Ovadia, 2013. “The role of big data in the social sciences,” *Behavioral & Social Sciences Librarian*, volume 32, number 2, pp. 130–134.

B. Pathak, R. Garfinkel, R. Gopal, R. Venkatesan and F. Yin, 2010. “Empirical analysis of the impact of recommender systems on sales,” *Journal of Management Information Systems*, volume 27, number 2, pp. 159–188.
doi: http://dx.doi.org/10.2753/MIS0742-1222270205, accessed 20 March 2017.

F. Provost and T. Fawcett, 2013. *Data science for business: What you need to know about data mining and data-analytic thinking*. Sebastopol, Calif.: O’Reilly.

W. Randolph and K. Viswanath, 2004. “Lessons learned from public health mass media campaigns: Marketing health in a crowded media world,” *Annual Review of Public Health*, volume 25, pp. 419–437.
doi: http://dx.doi.org/10.1146/annurev.publhealth.25.101802.123046, accessed 20 March 2017.

G. Sacks, M . Rayner and B. Swinburn, 2009. “Impact of front-of-pack ‘traffic-light’ nutrition labelling on consumer food purchases in the UK,” *Health Promotion International*, volume 24, number 4, 344–352.
doi: https://doi.org/10.1093/heapro/dap032, accessed 20 March 2017.

G. Sacks, K. Tikellis, L. Millar and B. Swinburn, 2011. “Impact of ‘traffic-light’ nutrition information on online food purchases in Australia,” *Australian and New Zealand Journal of Public Health*, volume 35, number 2, pp. 122–126.
doi: https://doi.org/10.1111/j.1753-6405.2011.00684.x, accessed 20 March 2017.

N. Sclater, 2014. “Effective learning analytics: Using data and analytics to support students” (3 October), at http://analytics.jiscinvolve.org/wp/2014/10/03/engagement-reporting-tools-for-blackboard-and-moodle/, accessed 20 March 2017.

D.-H. Shin and M.J. Choi, 2015. “Ecological views of big data: Perspectives and issues,” *Telematics and Informatics*, volume 32, number 2, pp. 311–320.
doi: http://dx.doi.org/10.1016/j.tele.2014.09.006, accessed 20 March 2017.

Science & Technology Facilities Council (STFC), n.d. “Large Hadron Collider,” at http://www.stfc.ac.uk/research/particle-physics-and-particle-astrophysics/large-hadron-collider//a>, accessed 20 March 2017; ATLAS, at http://www.stfc.ac.uk/research/particle-physics-and-particle-astrophysics/large-hadron-collider/atlas/, accessed 20 March 2017; CMS, at http://www.stfc.ac.uk/research/particle-physics-and-particle-astrophysics/large-hadron-collider/cms/, accessed 20 March 2017.

C.R. Sunstein, 2007. *Republic.com 2.0*. Princeton, N.J.: Princeton University Press.

J. Townsend, P. Roderick and J. Cooper, 1994. “Cigarette smoking by socioeconomic group, sex, and age: Effects of price, income, and health publicity,” *British Medical Journal*, volume 309, number 6959 (8 October), pp. 923–927.
doi: https://doi.org/10.1136/bmj.309.6959.923, accessed 20 March 2017.

J. Vanderplas, 2014. “Frequentism and bayesianism: A practical introduction” (11 March), at http://jakevdp.github.io/blog/2014/03/11/frequentism-and-bayesianism-a-practical-intro/, accessed 20 March 2017.

F. Vis, 2013. “A critical reflection on big data: Considering APIs, researchers and tools as data makers,” *First Monday*, volume 18, number 10, at http://firstmonday.org/article/view/4878/3755, accessed 20 March 2017.
doi: http://dx.doi.org/10.5210/fm.v18i10.4878, accessed 20 March 2017.

S.F. Wamba, S. Akter, A. Edwards, G. Chopin and D. Gnanzou, 2015. “How ‘big data’ can make big impact: Findings from a systematic review and a longitudinal case study,” *International Journal of Production Economics*, volume 165, pp. 234–246.
doi: http://dx.doi.org/10.1016/j.ijpe.2014.12.031, accessed 20 March 2017.

A. Wilson, C. Watson, V. Drew, T.L. Thompson and S. Doyle, 2017. “Learning analytics: Challenges and limitations,” *Teaching in Higher Education*, under review.

A. Zelenkauskaite, 2016. “Remediation, convergence, and big data: Conceptual limits of cross-platform social media,” *Convergence*.
doi: http://dx.doi.org/10.1177/1354856516631519, accessed 20 March 2017.

A. Zelenkauskaite and E.P. Bucy, 2016. “A scholarly divide: Social media, big data, and unattainable scholarship,” *First Monday*, volume 21, number 5, at http://firstmonday.org/article/view/6358/5511, accessed 20 March 2017.
doi: http://dx.doi.org/10.5210/fm.v21i5.6358, accessed 20 March 2017.

Editorial history

Received 6 February 2017; revised 16 March 2017; accepted 20 March 2017.

Big data and learning analytics: Singular or plural?
by Anna Wilson, Terrie Lynn Thompson, Cate Watson, Valerie Drew, and Sarah Doyle.
First Monday, Volume 22, Number 4 - 3 April 2017
http://firstmonday.org/ojs/index.php/fm/rt/printerFriendly/6872/6089
doi: http://dx.doi.org/10.5210/fm.v21i14.