International **R**eview of **R**esearch in **O**pen and **D**istributed **L**earning
Volume 16, Number 1

February – 2015

Big(ger) Data as Better Data in Open Distance Learning

Paul Prinsloo, Elizabeth Archer, Glen Barnes, Yuraisha Chetty and Dion van Zyl
University of South Africa

Abstract

In the context of the hype, promise and perils of Big Data and the currently dominant paradigm of data-driven decision-making, it is important to critically engage with the potential of Big Data for higher education. We do not question the potential of Big Data, but we do raise a number of issues, and present a number of theses to be seriously considered in realising this potential.

The University of South Africa (Unisa) is one of the mega ODL institutions in the world with more than 360,000 students and a range of courses and programmes. Unisa already has access to a staggering amount of student data, hosted in disparate sources, and governed by different processes. As the university moves to mainstreaming online learning, the amount of and need for analyses of data are increasing, raising important questions regarding our assumptions, understanding, data sources, systems and processes.

This article presents a descriptive case study of the current state of student data at Unisa, as well as explores the impact of existing data sources and analytic approaches. From the analysis it is clear that in order for big(ger) data to be better data, a number of issues need to be addressed. The article concludes by presenting a number of theses that should form the basis for the imperative to optimise the harvesting, analysis and use of student data.

Keywords: Big Data, learning analytics, student success

Introduction

> Technology is neither good nor bad; nor is it neutral...technology's interaction with the social ecology is such that technical developments frequently have environmental, social, and human consequences that go far beyond the immediate purposes of the technical devices and practices themselves.
>
> Melvin Kranzberg (1986, p. 545)

The view that Big Data is neither inherently good nor bad, increasingly finds its way into the current discourses in higher education. There is therefore a need for critical scholarly engagement amidst claims that "Big Data represents a paradigm shift in the ways we understand and study our world" (Eynon, 2013, p. 237). Much of the discourses regarding Big Data in higher education focus on increasing efficiency and cost-effectiveness (Altbach, Reisberg & Rumbley, 2009), amidst concerns regarding privacy, surveillance, the nature of evidence in education, and so forth (Biesta, 2007, 2010; Eynon, 2013; Prinsloo & Slade, 2013; Wagner & Ice, 2012). While Big Data has been lauded as "chang(ing) everything" (Wagner & Ice, 2012), and "revolutionis(ing) learning" (Van Rijmenam, 2013), boyd and Crawford (2012, 2013) question, quite rightfully, whether Big Data is an unqualified good.

The potential, perils, and the harvesting and analysis of Big Data in general, and in higher education in particular are, therefore, still relatively unexplored (Lyon, 2014; Siemens & Long, 2011). There is a need to establish a common language (Van Barneveld, Arnold, & Campbell, 2012) as well as resolve a number of conceptual, ethical and practical matters (Andrejevic, 2014; boyd & Crawford, 2012; Bryant & Raja, 2014; Clow, 2013a, 2013b, 2013c; Morozov, 2013a, 2013b; Mayer-Schönberger, 2009; Mayer-Schönberger & Cukier, 2013; Puschmann & Burgess, 2014; Richards & King, 2013; Slade & Prinsloo, 2013; Siemens, 2013).

While higher education institutions always had access to relatively large data sets and tools for analysis, there is an increasing amount of digital student data that can be harvested and analysed (Swain, 2013), as well as increased technological and analytical capabilities (Wishon & Rome, 2012). Describing learning analytics as the "new black" (Booth, 2012), and student data as the "new oil" (Watters, 2013) may resemble "techno-solutionism" (Morozov, 2013b) and a certain "techno-romanticism" in education (Selwyn, 2014). Despite various claims regarding the success of learning analytics to improve student success and retention (e.g., Arnold, 2010; Clow 2013a, 2013b, 2013c), Watters (2013) warns that "the claims about big data and education are incredibly bold, and as of yet, mostly unproven" (par. 17). Without ascribing to a technological pessimism, we have to critically explore the current belief in higher education that harvesting and analysing increasingly bigger data sets, will, *necessarily* improve student success and retention (boyd & Crawford, 2012; Briggs, 2014; Epling, Timmons & Wharrad, 2003).

This article does not question the potential of big(ger) data to be better data but would rather present a number of theses that may ensure that big(ger) data will result in a more thorough understanding of the complexities of student success and retention, and more appropriate, timeous and effective institutional support.

Method and Context

This descriptive and interpretive case study explores the proposition of big(ger) data within the context of a mega open, distance learning institution (ODL), the University of South Africa (Unisa). With more than 360,000 students registered for a variety of qualifications (more than 1,200) and selecting from a range of close to 3,000 courses– the size of available data may already qualify as 'big, ' though we acknowledge that compared to definitions of Big Data (e.g., Kitchen, 2013) the data harvested, analysed and used by Unisa currently falls short.

As a descriptive case study, the issues and theses raised do not claim to be original, nor generalisable to other higher education institutions, but rather linked to a number of theoretical propositions or theses (Yin, 2009). The value of descriptive case studies lies in its potential to extrapolate from a specific phenomenon a number of abstract interpretations of data and propositions for theoretical development (Mills, Eurepos & Wiebe, 2010). As an interpretive case study we attempt to not only describe a specific case of the use of student data, but interpret the case to move to a temporary conceptual framework, propositions or theses (Thomas, 2011). Our aim is therefore to advance phronesis or practical knowledge.

As Unisa increasingly moves into digitised and online learning, the amount of data available, as well as increased analytical capability provides ample foundation for an intensification of data harvesting and analysis efforts. Some of the questions integral to our exploration relate to how optimally we use data that we already have access to and whether more and/or different data will support our endeavours in finding the Holy Grail of a unified theory of student engagement, retention and success. How do we distinguish between signals and noise (Silver, 2012)? We accept that harvesting more and different data may hold potential, but if we do not think critically about institutional support and operational integration with regard to the harvesting and analysis of data, we may never realise the potential of bigger data.

All of the authors except one, who is a researcher in open distance learning, are involved in institutional research at Unisa, with the mandate to provide high quality, relevant and timely information, analysis and institutional research, making the institution more intelligible to itself. As such, the question regarding Big Data is central to the daily praxis of the researchers.

The authors engaged with literature and the current status of student data at Unisa in a systematic way over the course of six months. Several meetings were held, and the analysis in this

article resulted from the field notes from these meetings. Construct validity was ensured by engaging with multiple sources of evidence. The team confirmed internal validity by explanation building and addressing rival explanations as suggested by Yin (2009). The resulting theses were compared with the literature and available evidence in the field.

This paper employs the insights, experience and thoughts of these researchers engaging with the often undocumented realities of engaging with big(ger) data within a mega open, distance learning institution. The aim of the paper is to move beyond the discussion of whether or not big(ger) data is better data towards the more practical questions of: in order for big(ger) data to be better data, what needs to be in place?

Big Data

Big Data refer to "the capacity to search, aggregate and cross-reference large data sets" (boyd & Crawford, 2012, p. 663) and should be explored not only for its potential but also to question its capacities, its socio-political consequences and the need for critique (Lyon, 2014). "Big Data is, in many ways, a poor term" (boyd & Crawford, 2012, p. 1) and increasingly refers to metadata or "data about data" (Lyon, 2014, p. 3). Rob Kitchen (in Lyon, 2014, p. 5) describes Big Data as having the following characteristics:

> huge volume, consisting of terabytes or petabytes of data; high velocity, being created in or near real time; extensive variety, both structured and unstructured; exhaustive in scope, striving to capture entire populations of systems; fine-grained resolution, aiming at maximum detail, while being indexical in identification; relational with common fields that enable the conjoining of different data-sets; flexible, with traits of extensionality (easily adding new fields) and scalability (the potential to expand rapidly)

While most of the current discourses emphasise the increasing amount of data, the 'real' value (and peril) in Big Data lies in its networked and relational nature (Bauman & Lyon, 2013; boyd & Crawford, 2012; Marwick, 2014; Mayer-Schönberger & Cukier, 2013; Solove, 2001) with "at least three significant actors in this drama, government agencies, private corporations and, albeit unwittingly, ordinary users" (Lyon, 2014, p. 3). "It is the kind of data that encourages the practice of apophenia: seeing patterns where none actually exist, simply because massive quantities of data can offer connections that radiate in all directions" (boyd & Crawford, 2012, p. 2).

In the context of Big Data, there is talk of the "age of analytics" (Tene & Polonetsky, 2012, p.1), and increasingly, the "algorithmic turn", "the algorithm as institution" (Napoli, 2013, p. 1), the

"threat of algocracy" (Danaher, 2014) and "algorithmic regulation" (Morozov, 2013a, par.15). In these instances, algorithms have regulative, normative, and cultural-cognitive dimensions in the intersection between algorithm and institution where code becomes law (Napoli, 2013, referring to Lessig, 2006). A number of authors interrogate the potential of Big Data through the lens of "societies of control" (Deleuze, 1992, p. 4) (also see Henman, 2004). Big Data and its algorithms resemble a possible "gnoseological turning point" in our understanding of knowledge, information and faculties of learning where bureaucracies increasingly aspire to transform and reduce "ontological entities, individuals, to standardized ones through formal classification" into algorithms and calculable processes (Totaro & Ninno, 2014, p. 29).

A number of authors (boyd & Crawford, 2013; Lyon, 2014; Richards & King, 2013) therefore posit some provocations regarding Big Data that demand critical reflection. The increasing reliance on Big Data questions our traditional assumptions about knowledge in the context of Big Data's claim to "objectivity and accuracy [which] are misleading." We also need to realise that "not all data are equivalent" (boyd & Crawford, 2013, pp. 3-12). There is also the potential that Big Data will create new divides and be employed to perpetuate and increase existing inequalities and injustices (Andrejevic, 2014; boyd & Crawford, 2013; Couldry, 2014; Lyon, 2014; Richard & King, 2013).

There are also claims that Big Data "is about *what*, not *why*. We don't always need to know the cause of the phenomenon; rather, we can let data speak for itself" (Mayer-Schönberger & Cukier, 2013, p. 14). And herein lays the dilemma. *Data cannot speak for itself*. boyd and Crawford (2012) refer to this assumption as "an arrogant undercurrent in many Big Data debates" (p. 4) and Gitelman (2013) states that raw data "is an oxymoron."

Amidst the hype and relative current scarcity of evidence regarding the impact of learning analytics on increasing the effectiveness of teaching and learning (Altbach, Reisberg & Rumbley, 2009; Clow, 2013a; Watters, 2013), it is necessary to problematise the relationship between Big Data and education. Selwyn (2014) remarks that educational technology "needs to be understood as a knot of social, political, economic and cultural agendas that is riddled with complications, contradictions and conflicts" (p. 6). There are also concerns about education's current preoccupation with evidence-based teaching and learning (Biesta, 2007, 2010).

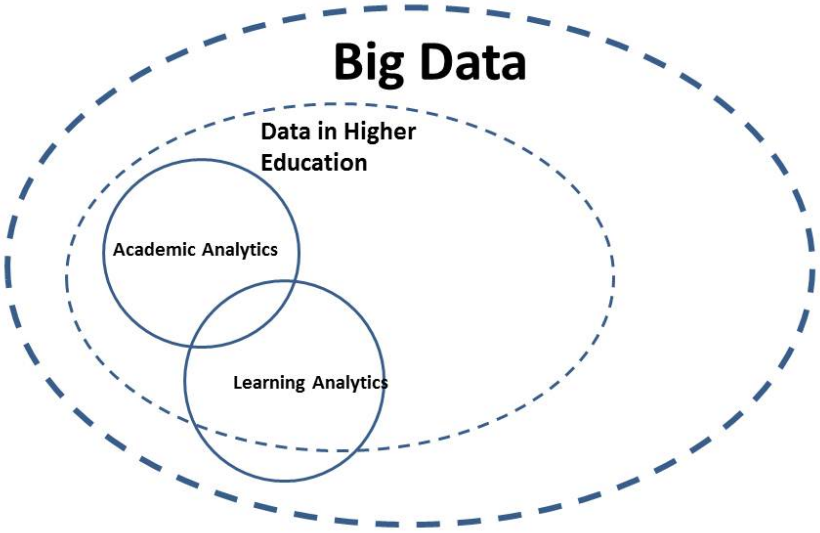
In the following section we therefore attempt to chart the relationship of learning analytics with academic analytics in the context of the bigger picture of data in education and Big Data.

Learning Analytics, Academic Analytics and Big Data

This article is not primarily concerned with the distinctions, scope and definitions of the different terminology used in the context of Big Data in higher education such as business, academic and learning analytics (see for example Buckingham Shum & Ferguson, 2012; Campbell, DeBlois, & Oblinger, 2007; Clow, 2013a; Siemens & Long, 2011; Van Barneveld et al., 2012).

We therefore adopt the joint *use* (not definition) of learning and academic analytics (as proposed by Siemens & Long, 2011). Learning analytics concerns itself with how students learn, integrating many aspects of students' transactions with amongst others, the learner management system, while areas such as admission, course success, graduation, employment and citizenship are better encompassed under the broader definition of academic analytics (see Siemens & Long, 2011). The notion of Big Data incorporates all sorts of electronic interactions even beyond the higher education environment, incorporating different elements of students' digital identities and personal learning environments (PLEs). Interestingly, as various sources of information start to resemble "electronic collages" and an "elaborate lattice of information networking" (Solove, 2004, p. 3), learning analytics may potentially overlap with academic analytics, the available data in the context of higher education, and data available 'outside' the strict boundaries of higher education (see Diagram 1).

Diagram 1: Mapping Learning and Academic Analytics in the Context of Big Data



It is crucial to delimit the scope of this article's focus on the potential of big(ger) data as better data as focusing only on the data available in, harvested and analysed by higher education institutions. This delimitation is, however, somewhat problematic, as the lattices of information networking (Solove, 2004) do not necessarily care about the (relatively) artificial boundaries between data 'in' higher education and data 'outside' of higher education. There is evidence that the nature of these boundaries may, actually, become increasingly artificial as student data becomes a potential income stream for institutions, and secondly, as student learning moves

beyond the strict confines of institutional learning management systems (LMS) (Prinsloo & Slade, 2013) (also see Bennet, 2001; Giroux, 2014; Lyon, 2014; Marx & Muschert, 2007).

An Overview of the Current State of Student Data at Unisa

The current state of the harvesting and use of student data at Unisa should be understood in terms of the institution's understanding of student retention and success (the student journey or walk), and secondly in the context of existing data sources and analytical approaches. The aim of this section is therefore not to offer immediate and possible solutions to some of the challenges experienced, but to contextualize and highlight the current state.

Student Data and the Student Journey

Unisa understands and predicts student retention, dropout and student success based on a socio-critical understanding of student success as described by Subotzky and Prinsloo (2011). This model draws attention to five student data elements, which can be harvested to support institutional planning and decision-making and to determine possible student support interventions. These include admission, learning activity, course success, graduation, and employment and citizenship. Data, which pertains to *learning activities* is perhaps the one area where learning analytics finds it home given that it is concerned with how students learn. The other areas (admission, course success, graduation, and employment and citizenship) are better encompassed under the broader definition of academic analytics (see Siemens & Long, 2011).

Admissions and Registrations

A variety of demographic data is captured at registration about the student, including attributes such as gender, ethnicity, age and educational background, which can be used by researchers to create student profiles linked to success indicators such as exam success and graduation, as well as learning analytics data, which are available on the learning management system (LMS). The integrity of the data, is of course something that needs to be regularly monitored. There are important profiling elements, which are not currently available in the databases, such as living standard measures and access to ICTs – these are instead sourced through research surveys involving samples of students and not the entire student population.

Learning Activities

This aspect of the student journey includes, but is not limited to, downloading study materials and learning objects, preparing and submitting assignments, engaging with e-tutors and e-mentors, engaging with a range of learning facilitators via the LMS through discussions groups and blogs, uploading assignments as they prepare for examinations and digitised learning activity indicators such as tutorial attendance. The Unisa LMS is a key data source, which can be used for harvesting and making sense of student data, and discussions are underway to make relevant data

available for research and analysis purposes. It is important to note that while most of the above data can be made available, the data currently reside in different systems, which are not integrated. This presents a challenge and lengthens the time of gaining access to relevant data.

Considering the fact that much of students' learning and learning activities may actually take place *outside* of the institutional LMS in digital and non-digital personal learning networks (PLNs) or personal learning environments (PLEs), this poses some additional challenges to the harvesting and analysis of data, on a number of levels (e.g., interoperability, appropriateness of data, etc.). Slade and Prinsloo (2013) also recognised the potential loss of gaining a "holistic picture of students' lifeworlds" (p. 1515) due to the "inability to track activity outside of an institution" (also see Prinsloo & Slade, 2014).

Course Success

Course success (as measured by passing the examination) remains but one element of measured success (Subotzky & Prinsloo, 2011). The examination process is closely monitored and reported on, covering aspects of exam admission, exam absence, and success in terms of number of students that passed relative to the number that sat for the examination. Additional information is also available on deferred examination statistics. All these data provide a comprehensive view of the examination event. These data can also be extended to compute other cohort aspects of success including examination passes relative to enrolments and intake. A number of reporting repositories are available to academics for the dissemination of examination statistics.

Recent engagement in this area suggests that attrition during the examination event is relatively small and focused attention is now moving to retention prior to the examination. The combination of academic and real-time learning analytics will allow us to get a sense of students' losing or changing momentum as they progress, whether by analysing their engagement in online forums, and/or submission of assignments.

Retention

At present, most of the analysis on retention is focused on various forms of student cohort analyses that are conducted and reported on at various decision making forums. While cohort retention and throughput information is readily available, the issue of the level of granularity of the data is under question and automated 'drill through' reports are not yet in place. These analyses remain plagued by data structure issues and the lack of a nuanced definition of 'drop-out'. A pertinent issue is that student dropout is only picked up when students fail to submit assignments, sit for the examination or re-register. Currently there is no mechanism in place to record, analyse and respond to *real-time* data, for example, where students do not log onto the LMS in more than a week, which would then enable early identification of students at risk of drop-out and the ability to provide personalised support and feedback. (For a critical appraisal of the use of learning analytics to increase student retention, see Clow, 2013b.)

Graduation

Analysis of graduation data provides another lens and level of granularity on student retention. Graduation data are available for analysis and reporting purposes to a range of stakeholders. However, this is a complex process as graduation data can be reported from a number of different perspectives, which are mutually exclusive and result in different outcome indicators.

Employment and Citizenship

Though the Alumni database at the institution stores data on graduates and their contact details, it is almost impossible to keep this database current. As an important data source, an up-to-date and information-rich Alumni database may provide crucial data for analyses such as measuring the impact of qualifications on graduate employment trajectories, as well as provide opportunities for post-enrollment guidance and lifelong learning.

Engaging with the Student Data Elements

Existing Data Sources and Analytical Approaches

Existing data sources.

The existing data sources lie in a number of disparate operational systems maintained by various functional units within the institution. Broadly, the roles of Data Owners, Data Custodians and Data Stewards have been identified and communicated, however, the need for cross-functional data access makes operating within these specific roles problematic. A far greater level of systems integration needs to be achieved if these roles are to be meaningfully applied. The disparate systems typically have grown from the need of different functional areas to have customised functionality built into operational systems. These developments take place without a reporting or analytic objective in mind with the result that levering sensible analytics from these sources is a problem.

The disparate nature of systems at Unisa has not embraced the concept of central warehousing, with data stores remaining fragmented into the major areas of (1) student administration, (2) learner management, (3) human resources, (4) finance and (5) space and the built environment. Not only are these areas supported by different operational systems, they have produced separate data warehouses and even separate business intelligence frameworks. While this is not a problem for operational needs, it remains a huge problem for integrated analysis and reporting, which requires the integration of these data sets.

Considering the velocity, agility and complexity of Big Data in the context of disparate existing data sources, it is clear that in order for big(ger) data to be better data, Unisa will have to rethink

the governance and integration of these different sources. Not only does Big Data increasingly allow the same data to be used for different purposes (Lyon, 2014), but with each recursion, there is an increasing threat of the loss of original context. With each new configuration and combination the data may "be given new meanings that are cut off from the values that once made sense of them and the identifiable subjects whose activities generated them in the first place" (Lyon, 2014, p. 6) (also see Amoore, 2011).

Analytical Approaches

The disparate nature of the various data sources typically requires technically skilled staff within the institution to be commissioned to translate the 'business rules' of the organisation into extracts that afford data migrated from an operational system towards an analytic framework. Data extracts vary in complexity depending on the alignment of business rules to analytic opportunities. In many cases, time-based analyses of the raw data are not possible as there are no date stamps on certain areas of the data. For example, the need to analyse the academic drift in the offerings of the institution requires the Programme Qualification Mix (PQM) or academic structure to be stored in a way that differentiates time (per year). In Unisa's case, the academic structure can be overwritten each year with the new (or combinations of) offerings with the result that the history of the data set is lost. The warehousing project will then have to take snapshots at designated points in the year to store this information in order to conduct the analytics.

Once on the warehouse a number of 'views' of the data are required primarily to address the performance of analytics and are set up to constitute a 'reporting layer'. Since the core business of the institution is undergraduate teaching and learning, more emphasis is placed on the various student views. The term used here does not refer to a specific table on a database but rather a design feature that can be applied to any technical environment. Typically, views are determined by a particular granularity (level of detail) of the data, or could also be designed around a common 'currency' in the data. For example, a course view has a significantly different granularity of student information compared with a programme or qualification view. Currency would refer to the underlying common element linking aspects of the analytics, for example the course could form a common currency for assignment information, tutor information and student information.

Broadly speaking, student views lie separately in the areas of assignments, courses, qualifications, tutors and lecturers. All metrics are governed by the structure within which they reside. Courses and qualifications reside within the academic structure or curriculum, but are also closely associated with the organisational structure at the time. Changes in these areas over time need to be addressed in an automated way in order to retain the integrity of analytics. Challenges in these areas emanate when the objectives of analytics are not clearly defined as an institutional take on reporting requirements. A good example of this would be the creation or dissolution of a faculty, school or college. These major changes occur relatively frequently in the current higher education environment and have major impacts on the intelligence of the institution. An institutional decision is required where reporting is either done employing the current structure or not. Many institutions choose to link all dependencies to the current structure and to transpose the history

in the data to that structure. In the case of our current example, if a college is created in a particular year, the history of all the related metrics (enrolments, exam statistics, assignments, etc.) are linked to the college and reported as if it were in existence in history. The question which needs to be considered is 'at what point is historical data irrelevant?'

The design of the reporting layer takes cognisance of the vast variety of analytics to be performed, from automated summary tables that underpin aggregated dashboards to the need for extended data mining of all attributes in these views to pick up underlying trends. The main purposes for the creation of the reporting layer is to support the user requirements of analysts and researchers, this being different from Management Information referred to above. Data sets are provided to these users either via direct access to the warehouse environment (SQL Server database), by provisioning front-end tools linked to these sources (MS Excel pivots, IBM SPSS, etc.) and also via automated web applications (Business Intelligence tools, customised net applications, proprietary tools, etc.).

Ensuring Big(Ger) Data as Better Data: A Number of Theses

Having described the current state of student data in a descriptive case study, we now propose a number of theoretical theses to inform further research and policy development. The theses are not mutually exclusive and the sequencing of the theses indicate a relative weighting in importance in order to address the research focus of this article, namely: In order for big(ger) data to be better data, what are the issues that we need to consider?

Thesis 1: A Critical View of Big Data

Amidst the hype and promise of Big Data in higher education, it is crucial to critically question and engage not only with its potential and its perils, but also with our assumptions about student engagement, retention and success. In order for big(ger) data to be better data in the context of Unisa, it is clear that a skeptical (Selwyn, 2014) and most probably a disenchanted view of Big Data's potential and promise is necessary (Couldry, 2014) (also see boyd & Crawford, 2012, 2013; Lyon, 2014 and Richards & King, 2013). In following Gitelman (2013) we believe that raw data "is an oxymoron", and contrary to Mayer-Schönberger and Cukier (2013) we do not believe that data speaks for itself. We need to consider 'what' and the 'why' (also see Kitchen, 2013). Considering that data is increasingly used to determine access to support and resources (Henman, 2004), we need to be cognisant of the impact and unintended consequences of our assumptions underpinning the algorithmic turn in higher education. Being skeptical about the potential of more data to be necessarily better data, we also need to critically engage with why much of the data harvested and analysed are not used (Macfadyen & Dawson, 2012).

Not only does such a critical and even skeptical view of big(ger) data herald and necessitate the need for a paradigm shift, such an approach instigates a number of other paradigm shifts with regard to institutional processes, skills-sets, systems and institutional knowledge.

Thesis 2: Process Changes

A combination of systems is used at different levels to harvest the data, but first let us determine the sources. The main source and management of data occur at the enterprise system level (Oracle student, Oracle HR, LMS, XMO, etc.). At this level, the data entry points are established and processes are put in place to capture and manage data input. It is at this level where boundaries and roles need to be clearly delineated to ensure the correct access is given to those that need update rights on the data. In many cases, we have staff with access to update data elements that are not specific to their environment and operational needs.

From these systems, a very limited quantity of Management Information (MI) is provided, mainly because these systems are in-house developments, which were not designed with a reporting perspective in mind. We make a distinction between MI and analytics in this article and submit that any respectable proprietary or in-house developed system should cater for all the needs of the user to perform the operational functions of their respective departments. As we move towards embracing analytics and the possible benefits of big(ger) data as better data, a paradigm shift will be required as it is this very operational and transactional data that will provide the basis for real time analytics.

The process of harvesting is important as during this process attention is given to the quality of the data. Procedures in place rely on a series of 'exception' or 'error' reports being run to determine data quality issues. These processes also allow for feedback to the source and responsible person to address data quality issues. This process is an iterative one of checking, feedback, fixing and re-extracting the data before release to users in the reporting layer.

Thesis 3: Skills and Capabilities Shift

The harvesting and analysis of available data by institutional researchers within Unisa is imperative as a contributor, firstly, towards understanding student success and secondly, to support strategic and operational decision making in order to ensure institutional sustainability and growth. However, where institutional researchers have dwelled within a paradigm of more traditional research designs and methods, Big Data offer significant new challenges from a data harvesting and mining perspective.

The availability of such Big Data is, however, still only in its infancy at Unisa. In order for Unisa researchers to harvest and analyse such data in future to create valuable and actionable insights, a new range of skills and competencies will be required. Traditional research methods often focus on data generation due to a scarcity or lack of data. This required researchers to focus and develop their skillsets around areas such as the science of measurement (are we measuring what

we are supposed to measure), data collection (to ensure representativeness) and analysis (can we make inferences about the larger population).

The promise of Big Data offers a new world and requires a new skillset for analysts. Big data emanate from multiple sources, but require some manipulations to answer specific research questions. Researchers and stakeholders must now scope information needs to more exact questions. But with more data comes more noise and also missing values (i.e., gaps in data). The 'new' institutional researcher must have skills to find or direct others to meaningful big data and make relevant connections. Programming skills in new languages combined with multi-source data mining, statistical modelling and prediction are now required. While the development of algorithms will be a part of this process, from an institutional and pedagogical perspective, an understanding of what drives student learning and success will remain key. Institutional researchers must balance the "what" provided by the patterns in data with the "why" which require more in-depth investigation through traditional research approaches. Identifying correlations alone will therefore not be sufficient. See Table 1 for a contrasting view between traditional approaches to institutional research and data analyses and modern analytics.

Table 1

Shifts from Traditional Institutional Research Skills to the Demands of Modern Analytics

	More Traditional Institutional Research	Modern Analytics
Data	• Scarcity of data • Focus on data generation • Generally historical, trend or snap shot orientation	• Big Data • Multiple data sources • Generally trend and real-time data.
Skillset emphasis	• Science of measurement including instrument design and construct measurement • Data collection to ensure representativeness and generalisations • Analysis for inference about the larger population	• Finding meaning in Big Data and making relevant connections • Programming skills • Multi-source data mining • Statistical modelling • Development of algorithms
Approach	• Multitude of possible questions narrowed down to more specific research questions	• Big Data, not tailored to any questions, narrowed down to information needs for specific questions
Main complexities	• Limited granularity and ability to segment due to small numbers • Representativeness of sample and ability to make inferences about the larger population	• More data = more noise, difficulty to determine which data is meaningful and what the patterns are reflecting • Missing values • Data Quality
Driver	• Understanding of what drives student learning and success	

Will the future institutional researcher be one that is described as statistician, mathematician, computer scientist, database administrator, coder, hardware guru, systems administrator, researcher and interrogator, all in one? Or will the individual make way for a more team based approach? A literature search points to a lack of research investigating the demands of Big Data on the skills and capabilities of institutional researchers. The convergence of traditional

institutional research skills, data science, analytical services and organisation intelligence therefore becomes a key area of consideration that should be taken cognisance of.

Thesis 4: Systems Evolution

The move toward central warehousing is imperative if the potential of Big Data and enterprise-wide analytics is to be achieved. A systematic mapping of the data elements needed for reporting and analysis is required, this in conjunction with an overview approach to systems architecture. Very little opportunity is taken to look at reporting and analytic requirements from an institutional perspective, which is exacerbated by the lack of clear roles and responsibilities, poor communication between affected parties and 'turf wars'.

Unisa should move towards the deployment of an 'information control centre', which has analytics as the primary objective, and which is built from a solid base of technical hardware, software and skilled personnel. The use of sophisticated software to bridge the gaps and boundaries of storage and location of data needs to be investigated. The movement away from traditional development thinking needs to happen in order to be agile in this area. We remain trapped in the mindset that requires a particular process to be followed: (1) data elements are identified for a particular purpose, (2) the metrics and measures for the element then defined, (3) the specifications are written to enhance a particular system for the purposes of storage, (4) development takes place to enhance the system, (5) the implementation of the upgrade is commissioned through extensive, often invaluable consultation, and (6) the process of capture and monitoring is documented and applied.

Alignment of processes and systems will allow Unisa to not only use historical data in more effective (and appropriate) ways, but ensure an agile information architecture to optimise the potential and guard against the perils of Big Data.

Thesis 5: Institutional Relevance, Context, and Knowledge

With the possibilities of working with Big Data, traditional research methods will have to evolve to face this new reality of Big Data. The challenge lies also not only with the technical aspects of finding, organising and combining the often unstructured data, but with the contextual insights needed to interpret and apply the knowledge and intelligence gained. Contextual intelligence speaks to an in-depth understanding of the institution and how it functions, and will include knowledge of its operational and strategic objectives and direction. Herein lays one of the major challenges with bigger data. All this will require new ways of thinking on various levels. Importantly, researchers and analysts will need to balance the promise of Big Data, and the various opportunities it presents to uncover patterns in data and employ advanced analytics (the "what"), with an ongoing search for the drivers of student learning, success, retention, dropout and throughput (the "why") in order to provide relevant intelligence with maximum impact. A stark reality within the Unisa context, which also faces institutional researchers globally, is the pertinent question of "So what"? If all the analytics and research do not result in actionable interventions by the university, in our case to strengthen support to students towards success,

then how relevant are we? This is a sobering thought, which also speaks to the tension between the provisioning of intelligence and influencing action.

Limitations to this Study

We acknowledge the concerns regarding case study methodology of a lack of rigor, and the impossibility to generalise from a single case (Yin, 2009). The value this case study adds is an attempt to generalise to "theoretical propositions, and not to populations or universes" (Yin, 2009, p. 15). The theses proposed in the preceding section therefore could be used as a heuristic framework for engaging with the state and use of student data in other higher education contexts.

(In)Conclusion

Despite and amidst the fact that "big data is all the rage" (Richards & King, 2013, p. 41) and various claims that Big Data will revolutionise and transform the way we live, work, and think (Mayer-Schönberger & Cukier, 2013), there are also many authors who take a more sceptical and critical approach to Big Data (Amoore, 2011; boyd & Crawford, 2012, 2013; Bryant & Raja, 2014; Couldry, 2014; Lyon, 2014; Richards & King, 2013).

In the current context of persisting concerns about student success and retention in higher education and in particular ODL (e.g., Subotzky & Prinsloo, 2011) and the prevailing logic of evidence-based decision-making and the "algorithmic turn" (Napoli, 2013, p. 1) and "algorithmic regulation" (Morozov, 2013a, par.15) in higher education, a critical interrogation of the potential of Big Data is called for.

This article did not attempt to question the potential of Big Data for higher education but rather raised the question: "In order for big(ger) data to be better data, and to result in more effective and appropriate teaching, learning and support, what are the issues that we need to consider?" In the context of Unisa as a mega ODL institution, this question necessitated a sober, and somewhat sceptical (if not disenchanted), view of the practical implications to realise big(ger) data's potential. We concluded the article with proposing five theses that, when considered, can increase the realisation of Big Data's potential in a specific context, that of Unisa.

Though some of the issues may be more applicable to higher and distance education contexts in developing world contexts, we propose the concluding theses to inform further research, contemplation and consideration.

Acknowledgement

The authors would like to acknowledge the comments and critical input received from the reviewers during the review process. Their critical and supportive engagement allowed the authors to reconsider and rework the value proposition of the article.

References

Altbach, P.G., Reisberg, L., & Rumbley, L.E. (2009). *Trends in global higher education: Tracking an academic revolution. A report prepared for the UNESCO World Conference on Higher Education*. Paris. UNESCO. Retrieved from http://atepie.cep.edu.rs/public/Altbach,_Reisberg,_Rumbley_Tracking_an_Academic_Revolution,_UNESCO_2009.pdf

Amoore, L. (2011). Data derivatives: On the emergence of a security risk calculus for our times. *Theory Culture Society, 28*, pp. 24—43. DOI: 10.1177/0263276411417430

Andrejevic, M. (2014). The Big Data divide. *International Journal of Communication*, 8, 1673—1689.

Arnold, K. (2010, March 3). Signals: Applying academic analytics. *EDUCAUSEreview* [online]. Retrieved from http://www.educause.edu/ero/article/signals-applying-academic-analytics

Bauman, Z., & Lyon, D. (2013). *Liquid surveillance*. Cambridge, UK: Polity.

Bennett, C. J. (2001). Cookies, web bugs, webcams and cue cats: Patterns of surveillance on the world wide web. *Communications of the ACM, 3*, 197–210. DOI:10.1023/A:1012235815384.

Biesta, G. (2007). Why "what works" won't work: Evidence-based practice and the democratic deficit in educational research. *Educational Theory, 57*(1), 1–22. DOI: 10.1111/j.1741-5446.2006.00241.x.

Biesta, G. (2010). Why 'what works' still won't work: From evidence-based education to value-based education. *Studies in Philosophy of Education, 29*, 491–503. DOI 10.1007/s11217-010-9191-x.

Bollier, D. (2010). The promise and peril of Big Data. *Eighteenth Annual Aspen Institute Roundtable on Information Technology*. Retrieved from http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf

Booth, M. (2012, July 18). Learning analytics: The new black. *EDUCAUSEreview* [online]. Retrieved from http://www.educause.edu/ero/article/learning-analytics-new-black

boyd, D., & Crawford, K. (2012). Critical questions for Big Data. *Information, Communication & Society, 15*(5), 662—679. DOI: 10.1080/1369118X.2012.678878

boyd, D., & Crawford, K. (2013). Six provocations for Big Data. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431

Briggs, S. (2014, January 13). Big data in education: Big potential or big mistake? [Web log post]. Retrieved from http://www.opencolleges.edu.au/informed/features/big-data-big-potential-or-big-mistake/

Bryant, A., & Raja, U. (2014). In the realm of Big Data. *First Monday*, 19(2-3). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/4991

Buckingham Shum, S., & Ferguson, R. (2012). Social learning analytics. *Educational Technology & Society*, 15, 3–26.

Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. *EDUCAUSEreview*, *42*, 40–57. Retrieved from http://net.educause.edu/ir/library/pdf/ERM0742.pdf

Clow, D. (2013a). An overview of learning analytics. *Teaching in Higher Education*, *18*(6), 683—695. DOI: 10.1080/13562517.2013.827653

Clow, D. (2013b, November 13). Looking harder at Course Signals. [Web log post]. Retrieved from http://dougclow.org/2013/11/13/looking-harder-at-course-signals/

Clow, D. (2013c, December 10). InFocus: Learning analytics and Big data. [Web log post]. Retrieved from http://dougclow.org/2013/12/10/infocus-learner-analytics-and-big-data/

Couldry, N. (2014). A necessary disenchantment: Myth, agency and injustice in a digital world. *The Sociological Review*, 1—18. DOI: 10.1111/1467-954X.12158. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/1467-954X.12158/full

Danaher, J. (2014). Rule by algorithm? Big data and the threat of algocracy. *Institute for Ethics and Emerging Technologies.* [Web log post]. Retrieved from http://philosophicaldisquisitions.blogspot.com/2014/01/rule-by-algorithm-big-data-and-threat.html

Deleuze. G. (1992). Postscript on the societies of control. *October*, 59 pp. 3-7.

Epling, M., Timmons, S., & Wharrad, H. (2003). An educational panopticon? New technology, nurse education and surveillance. *Nurse Education Today*, *23*, 412—418.

Eynon, R. (2013). The rise of Big Data: What does it mean for education, technology, and media research? *Learning, Media and Technology*. DOI: 10.1080/17439884.2013.771783

Giroux, H. (2014, February 10). Totalitarian paranoia in the post-Orwellian surveillance state. [Web log post]. Retrieved from http://www.truth-out.org/opinion/item/21656-totalitarian-paranoia-in-the-post-orwellian-surveillance-state

Gitelman, L. (Ed.). (2013). *"Raw data" is an oxymoron*. London, UK: MIT Press.

Henman, P. (2004). Targeted!: Population segmentation, electronic surveillance and governing the unemployed in Australia. *International Sociology*, *19*, 173-191. DOI: 10.1177/0268580904042899

Kitchen, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, *3*, 262-267. SOI: 10.1177/2043820613513388

Kranzberg, M. (1986) Technology and history: Kranzberg's laws. *Technology and Culture*, *27*(3), 544—560.

Lyon, D. (2014). Surveillance, Snowden, and Big Data: Capacities, consequences, critique. *Big Data & Society*, July-September pp. 1—13. DOI: 10.1177/20253951714541861

Macfadyen, L.P., & Dawson, S. (2012). Numbers are not enough. Why e-learning analytics failed to inform an institutional strategic plan. *Educational Technology & Society*, *15*(3), 149-163.

Marwick, A.E. (2014, January 9). How your data are being deeply mined. [Web log post]. *The New York Review of Books*. Retrieved from http://www.nybooks.com/articles/archives/2014/jan/09/how-your-data-are-being-deeply-mined/?pagination=false

Marx, G. T., & Muschert, G. W. (2007). Personal information, borders, and the new surveillance studies. *Annual Review of Law and Social Science*. DOI:10.1146/annurev.lawsocsci.3.081806.112824.

Mayer-Schönberger, V. (2009). *Delete. The virtue of forgetting in the digital age*. Princeton, NJ: Princeton University Press.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data. A revolution that will transform how we live, work, and think*. New York, N.Y.: Houghton Miffling Harcourt Publishing Company

Mills, A.J., & Durepos, G., & Wiebe, E. (2010). *Encyclopedia of case study research*. Thousands Oaks, CA: SAGE Publications.

Morozov, E. (2013a, October 23). The real privacy problem. *MIT Technology Review*. Retrieved from http://www.technologyreview.com/featuredstory/520426/the-real-privacy-problem/

Morozov, E. (2013b). *To save everything, click here*. London, UK: Penguin Books.

Napoli, P. (2013). The algorithm as institution: Toward a theoretical framework for automated media production and consumption. In *Media in Transition Conference* (pp. 1–36). DOI: 10.2139/ssrn.2260923

Prinsloo, P., & Slade, S. (2013). An evaluation of policy frameworks for addressing ethical considerations in learning analytics. *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, pp. 240—244. Retrieved from http://dl.acm.org/citation.cfm?id=2460344

Puschmann, C., & Burgess, J. (2014). Metaphors of Big Data. *International Journal of Communication, 8*, pp. 1690—1709.

Richards, N.M., & King, J.H. (2013). Three paradoxes of Big Data. *Stanford Law Review* [Online], pp. 41—46. Retrieved from http://www.stanfordlawreview.org/online/privacy-and-big-data/three-paradoxes-big-data

Selwyn, N. (2014). *Distrusting educational technology. Critical questions for changing times.* New York, NY: Routlegde.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSEreview* [online], September/October, 31—40. Retrieved from http://www.elmhurst.edu/~richs/EC/OnlineMaterials/SPS102/Teaching%20and%20Learning/Penetrating%20the%20Fog.pdf

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioural Scientist, 57*(10), pp. 1380—1400.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail — but some don't.* London, UK: Penguin Books.

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioural Scientist*, 57(1) 1509–1528.

Solove, D.J. (2001) Privacy and power: Computer databases and metaphors for information privacy. *Stanford Law Review*, 53(6), 1393—1462.

Solove, D.J. (2004). *The digital person. Technology and privacy in the information age.* New York, NY: New York University Press.

Stiles, R.J. (2012). Understanding and managing the risks of analytics in higher education: A guide. *EDUCAUSE*. Retrieved from https://net.educause.edu/ir/library/pdf/EPUB1201.pdf

Subotzky, G., & Prinsloo, P. (2011). Turning the tide: A socio-critical model and framework for improving student success in open distance learning at the University of South Africa. *Distance Education, 32*(2), 177—193. DOI:10.1080/01587919.2011.584846.

Swain, H. (2013, August 5). Are universities collecting too much information on staff and students? [Web log poast]. *The Guardian.* Retrieved from http://www.theguardian.com/education/2013/aug/05/electronic-data-trail-huddersfield-loughborough-university

Tene, O., & Polonetsky, J. (2012). Big data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology and Intellectual Property, 239*, 1–36. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2149364

Thomas, G. (2011). *How to do your case study. A guide for students and researchers*. London, UK: SAQGE.

Totaro, P., & Ninno, D. (2014). The concept of algorithm as an interpretive key of modern rationality. *Theory Culture Society, 31*, pp. 29—49. DOI: 10.1177/0263276413510051

Van Barneveld, A., Arnold, K.E., & Campbell, J.P. (2012). Analytics in higher education: Establishing a common language. *EDUCAUSE Learning Initiative, 1*, 1—11. Retrieved from http://sites.ewu.edu/elearningservices/files/2012/06/Analytics-in-Higher-Education-Establishing-a-Common-Language-ELI3026.pdf

Van Rijmenam, M. (2013, April 30). Big data will revolutionise learning. [Web log post]. Retrieved from http://smartdatacollective.com/bigdatastartups/121261/big-data-will-revolutionize-learning

Wagner, E., & Ice, P. (2012, July 18). Data changes everything: Delivering on the promise of learning analytics in higher education. *EDUCAUSEreview* [online]. Retrieved from http://www.educause.edu/ero/article/data-changes-everything-delivering-promise-learning-analytics-higher-education

Watters, A. (2013, October 13). Student data is the new oil: MOOCs, metaphor, and money. [Web log post]. Retrieved from http://www.hackeducation.com/2013/10/17/student-data-is-the-new-oil/

Wishon, G.D., & Rome, J. (2012, 13 August). Enabling a data-driven university. *EDUCAUSEreview* [online]. Retrieved from http://www.educause.edu/ero/article/enabling-data-driven-university

Yin, R.K. (2009). *Case study research. Design and methods,* 4th edition, Applied Social Research Methods Series, Volume 5. London, UK: Sage.

Athabasca University