The role of a reference synthetic data generator within the field of learning analytics

Berg, A.M.; Mol, S.T.; Kismihók, G.; Sclater, N.

DOI
10.18608/jla.2016.31.7

Publication date
2016

Document Version
Final published version

Published in
Journal of Learning Analytics

JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

The Role of a Reference Synthetic Data Generator within the Field of Learning Analytics

Alan M. Berg
ICT Services, University of Amsterdam, NL
a.m.berg@uva.nl

Stefan T. Mol
Gábor Kismihók
Center of Job Knowledge Research, Amsterdam Business School, University of Amsterdam, NL

Niall Sclater
Sclater Digital Ltd, UK

ABSTRACT: This paper details the anticipated impact of synthetic "big" data on learning analytics (LA) infrastructures, with a particular focus on data governance, the acceleration of service development, and the benchmarking of predictive models. By reviewing two cases, one at the sector-wide level (the Jisc learning analytics architecture) and the other at the institutional level (the UvAInform learning analytics project at the University of Amsterdam), we explore the need for an on-demand tool for generating a wide range of synthetic data. We argue that the application of synthetic data will not only accelerate the creation of complex and layered learning analytics infrastructure, but will also help to address the ethical and privacy risks involved during service development.

Keywords: Learning analytics, simulation, synthetic data, student consent service, Jisc learning analytics architecture

1 INTRODUCTION

There is growing interest in deploying learning analytics services at educational institutions. Stimulating the interest in developing and deploying learning analytics services are a number of successful examples that have affected student learning. Of these, Course Signals is arguably the best known (Arnold & Pistilli, 2012). Another example is the Open Academic Analytics Initiative (OAAI) led by Marist College (Jayaprakash, Moody, Lauria, Regan, & Baron, 2014). Building on early work in LA, Siemens et al. (2011) proposed developing an overarching framework for learning analytics. An all-encompassing framework would need to include the following: 1) the collection of data, 2) dealing with crucial issues such as data governance and ethics, 3) pre-processing of the data, 4) sharing of the data models, 5) predictive modelling, 6) interventions including dashboards and other strategies, and the measurement of their impact on the learning process. The conversation about this open learning analytics framework is ongoing and influencing the design of major learning analytics services such as Jisc's Open Learning Analytics Architecture (Sclater, Berg, & Webb, 2015) and the Apereo (2015a) Learning Analytics Initiative. These frameworks have many interrelated components, and they digest a rich variety of data. In this paper, we

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

will explore the roles synthetic data and the associated software that generates the data can play in helping to develop these emerging Big Data learning analytics service infrastructures.

Through the mechanism of a systematic literature review, we explore whether synthetic data approaches have been fully utilized in general, and specifically in the field of learning analytics. Are there already significant examples of synthetic data generation and usage whose methodologies are ready to apply within the field of learning analytics? Can we argue for a common unifying approach to the generation of synthetic data specific to learning analytics through the means of a reference synthetic data generator?

2 LITERATURE REVIEW

2.1 General Usage of Synthetic Data

Synthetic data is primarily used to avoid accidental disclosure or reconstruction of information; for example, as part of national microdata sets (Kinney et al., 2011). There are numerous methods to limit the risk (Matthews & Harel, 2011) such as using example data, fitting predictive models with the example data, and then generating replacement data from the tuned model. Synthetic data enables the rapid prototyping of services before the "real" big data has been amassed or made available to an application. Its availability supports proof of concept, security testing, practising, and training around data governance processes, boundary testing, user testing of visualizations, and interoperability testing of different architectural components, as well as many other applications.

Synthetic data, also known as simulated data, has been heavily researched and successfully applied across a broad range of scientific fields, including economic calculations as part of national micro-datasets (Kinney, Reiter, & Miranda, 2014); house occupancy for urban planning; transportation planning (Beckman, Baggerly, & McKay, 1996; Rich & Mulalic, 2012); deterioration of sewage systems (Scheidegger & Maurer, 2012); support of fraud detection systems (Barse, Kvarnstrom, & Jonsson, 2003); security testing of defense in-depth strategies (Boggs, Zhao, Du, & Stolfo, 2014); workload generation for cloud computing (Bahga & Madisetti, 2011); simulating real time network traffic (Botta, Dainotti, & Pescapé, 2012); weather behaviour, such as precipitation (Abtew, Moras, & Campbell, 1990; Piantadosi, Boland, & Howlett, 2009) and wind (Liang et al., 2013); the number of solar-power cells delivered in a year for a given location (Celik, 2003); and for realistic workload generation for YouTube (Abhari & Soraya, 2010). Within the field of bioinformatics, synthetic data has been used for the design and analysis of structure-learning algorithms (Van den Bulcke et al., 2006).

In the field of data mining, synthetic data has been used to generate and benchmark text-mining algorithms and tools (Eno & Thompson, 2008; Jeske, Lin, Rendon, Xiao, & Samadi, 2006); for building and testing Information Discovery Systems (Lin et al., 2006); selecting feature set discovery algorithms (Bolón-Canedo, Sánchez-Maroño, & Alonso-Betanzos, 2013); testing the scalability of big data infrastructures, for example by populating and testing the performance of databases of various types (Gray, Sundaresan,

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

Englert, Baclawski, & Weinberger, 1994; Tzouramanis, Vassilakopoulos, & Manolopoulos, 2002; Lo, Cheng, Lin, Hon, & Choi, 2014); for evaluating visual-analytics techniques (Maciejewski et al., 2009); for the generation analysis of social networks (Barrett et al., 2009); and to create training datasets for handwriting recognition (Varga & Bunke, 2008).

A recent review of learning analytics in UK higher and further education suggests that the emerging market for learning analytics products is highly fragmented (Sclater, 2014). Therefore, a great challenge for institutions is the risk of vendors developing and marketing similar systems that tackle different parts of the learning analytics infrastructure, but have not been made interoperable. Within this context, synthetic data has the potential to accelerate the development of big learning analytics infrastructure and methods and avoid unnecessary delays by early disclosure of realistically distributed, descriptive data that has the property of minimal risk of accidental disclosure (Matthews & Harel, 2011). The data can form the basis of benchmarks as it, and the systems developed towards their generation, can be shared freely as part of that benchmark. One focus of such benchmarks will be to support decision makers in choosing between a series of similarly visually appealing products. However, it should be noted that the challenge of bias in the generated data could lead to poor decision making. Consider the problem of class imbalance and the need to oversample minority populations (He, Bai, Garcia, & Li, 2008). Clearly labelling the degree of bias of the benchmarks in order to assist decision makers will be a challenge.

Another grand challenge for large organizations is to centralize data and, by implication, their governance (Ebner, Taraghi, Sarantie & Schon, 2015). This centralization allows universities to analyze a wider range of datasets for a broader audience with the support of central data governance to deploy learning analytics services across departmental boundaries. There is a risk of an emerging divide in the quality of these services between those organizations that strive for data centralism and those that do not (Berg, 2015). This divide has previously been reported from within the business context with suggestions for accelerating progress through business culture transformation, centralization of data, and the use of standards (Kiron, Shockley, Kruschwitz, Finch, & Haydock, 2011).

A survey on the subject of data quality management for big data analytics (Kwon, Lee, & Shin, 2014) discovered a positive relationship between a firm's competence in maintaining quality (i.e., consistency and completeness) and the firm's adoption intention for big data analytics. Synthetic data can be used to either replace missing data (completeness) or support the disambiguation process (consistency). For example, when using a broad range of social media as part of a learner's experience, there is a risk of students using multiple credentials. We might name ourselves jdberg1892 for our twitter account and john.doe.berg.1 for our LinkedIn account. Synthetic data has been applied in the development of disambiguation methodologies to define strategies to resolve this issue (Ferreira, Gonçalves, Almeida, Laender, & Veloso, 2012).

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

2.2 Usage within the Field of Learning Analytics

Ferguson (2012) noted that one of the challenges for learning analytics is to develop methods of working with a wide range of datasets in order to optimize learning environments. We argue that synthetic data supports the creation and refinement of processes prior to the data from multiple silos being freely and fully available. For example, it can be generated and utilized while waiting for approval from multiple ethics boards or working through politically sensitive data ownership issues. Synthetic data will also support researchers who do not have access to rich data sources, allowing them to tune and tweak their methodologies so that they can interact efficiently with "elite" researchers in more advantageous data centralized environments.

There is a close relationship between the Educational Data Mining and Learning Analytics communities, (Siemens & Baker, 2012); many methodologies and practices are shared between them. As evidenced in the last section, synthetic data generators are already applied in many data-mining contexts. A concrete example is the application of synthetic data to sparse probit-factor analysis to test the efficacy of estimating a learner's knowledge of the concepts within specific problem domains (Waters, Lan, & Studer, 2013).

There is also evidence of the use of synthetic data as part of the process of disseminating and practising learning analytics methodologies. For example, this occurred at a data manipulation hackathon (University of Michigan, 2015a), and is part of the training materials within a learning analytics MOOC (University of Michigan, 2015b; Koester, 2015).

The EP4LA Ethics and Privacy Workshop Series (Sherlock, 2014) is a set of interrelated workshops discussing a broad range of issues including, but not limited to data ownership, data degradation, anonymization of data, data security, data sharing, danger of linking datasets for privacy, context integrity, approaches to informed consent in the times of big data, expected changes to privacy due to big data, cross-cultural studies on privacy, transparency (purpose of analysis, raw data access, opt-out), and ethical considerations for learning analytics. As discussed in the introduction, synthetic data will play an alleviating role for issues across these themes.

Verbert, Manouselis, Drachsler, and Duval (2012) applied a framework mapping the high-level properties of datasets against their LA objectives. Through this utilitarian optic, the authors reviewed a range of datasets and their relevance for application within the field of learning analytics. They noted, "our endeavors to collect and share datasets for research remain quite challenging" (p. 145) and described how a number of datasets were made open. By modelling closed datasets, synthetic data generation can extend the range of open datasets available for characterization and experimentation.

The Apereo Learning Analytics Initiative (LAI) is applying synthetic data for performance testing its reference learning analytics infrastructure and the test plans (Apereo, 2015b). This was also used to

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

populate a Learning Record Store, a secure central repository for learners' activity streams, with example data to allow data scientists to experiment with learning analytics related visualizations while dashboard building (SoLAR, 2015). There is currently a discussion within the Apereo LAI community on the subject of extending the test plans to reflect emerging practices around xAPI recipes. For example, deriving tests based on recipes expressed in the connected learning analytics toolkit (Kitto, 2015).

Synthetic data generation has the potential to support large-scale, complex, and thus "big" learning analytics services such as a layered set of national or institutional services. In the next sections, we reflect on the opportunities for the application of synthetic data to big services. First we reference Jisc's Open Learning Analytics Architecture (Sclater, 2015b), which has been designed to allow universities and colleges in the UK to engage with learning analytics using a freely provided hosting service. Next, we look at the institutional level via the UvAInform project at the University of Amsterdam (Kismihók & Mol, 2014). Here a coordinated set of pilots is being carried out to develop a wider understanding of the value of learning analytics services within the university. The analysis of these two endeavours is followed by a review of the trend of increased sharing and richness of learning activity data outside the control of learner-centric organizations. We examine the implications and discuss the need for a reference implementation of a synthetic data generator.

3 THE JISC OPEN LEARNING ANALYTICS ARCHITECTURE

In response to requests for the provision of basic services to help institutions adopt learning analytics in the UK higher and further education sectors (Sclater, 2014), Jisc has developed an open learning analytics framework and is commissioning associated software components from a range of vendors (Sclater, 2015b). In summary, data sources — initially primarily from the virtual learning environment and the student information system — are extracted into a "learning records warehouse" which contains both unstructured and structured data, including learning records in the xAPI format (Tin Can[1]). Furthermore, there may also be "self-declared" data from students, such as e-portfolio content or data from wearable devices.

A *learning analytics processor* carries out the predictive analytics and provides the results to *staff dashboards*. A *student app* enables students to view their own analytics, set targets for learning, log their learning activities, and compare their engagement and attainment with others. Meanwhile an analytics based *alert and intervention system* prompts staff and students in the case of certain specific situations, such as a student's engagement signalling that they are at risk of dropout. This system also helps to manage any subsequent interventions with students. Students are also given a degree of control over what is done with their data by means of a *student consent service*. Note that the dashboard and app are relatively unintelligent, which allows different visualization tools to be slotted in. The (potentially quite complex) processes of managing alerts and interventions take place in the alerts and intervention system.

[1] http://tincanapi.com/

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

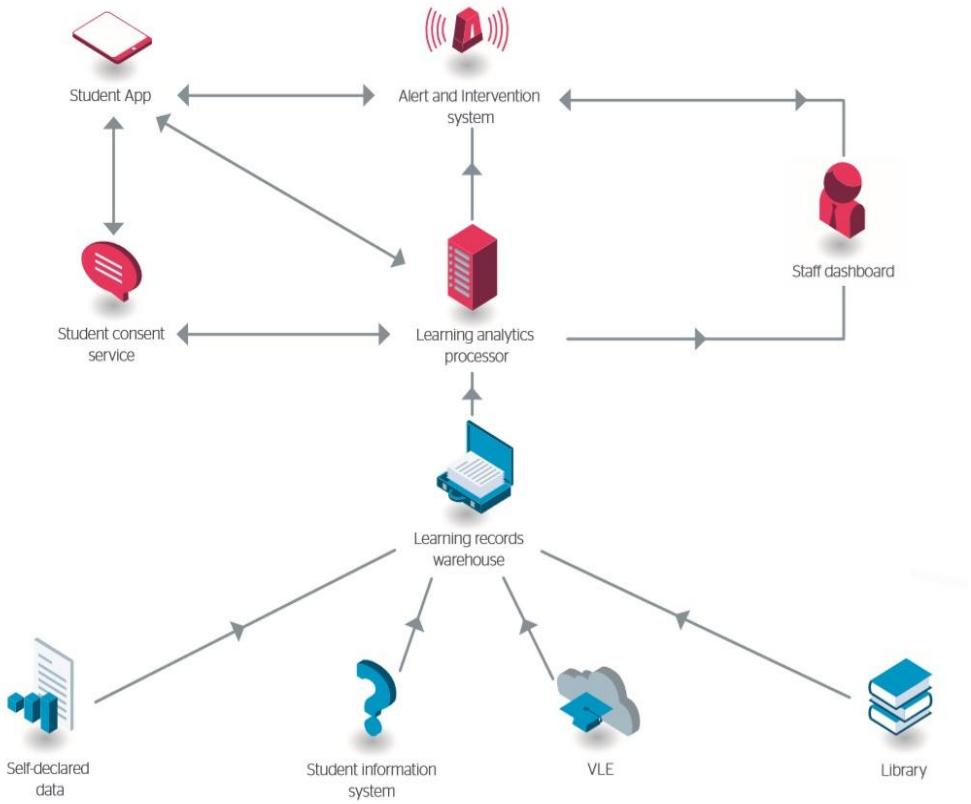

Figure 1: An overview of Jisc's learning analytics architecture.

A number of issues arise when considering the use of synthetic data within this architecture:

Security testing: The complexity of the various systems involved and the "big" data that they create (including self-declared data) suggest that a wide range of synthetic data will be required in order to carry out security testing prior to real data being entrusted to the infrastructure or its components.

Interoperability testing: A variety of modular systems from different vendors is being commissioned at different levels of the architecture in order to provide a cohesive overall learning analytics service for institutions. Each one of these systems could potentially (at any point in the lifecycle of the open learning analytics framework) be replaced by one from a different vendor. Thus, a core set of synthetic data is essential in order to ensure that data can pass interoperably through the different levels of the framework — so that alternative tools at each level can be tested quickly and effectively. The use of real data in a development or acceptance environment involves a significantly enhanced risk of unintended disclosure. This is because the lower quality of alpha and beta software and the number of actors involved in these non-production environments polynomially enhance the opportunities for attacks (due to the increased number of viable combinations of interactions with the system) compared to the more stable and locked down production environments.

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

An initial dataset of learning activity data for interoperability testing is being generated from a set of Moodle courses developed to explain aspects of the architecture. While this is "real" data provided automatically by users of the courses, it provides a useful basis for the generation of a larger scale synthetic dataset.

Benchmarking for predictive models: In order to compare and contrast different predictive models, a set of uniform benchmarks will be required. The benchmarks do not just include ways of comparing, but should also include example datasets or methods of generating realistic datasets on demand. Synthetic data enables those without access to full and "rich" datasets to compare their services to those where they are available. Synthetic datasets avoid concerns of disclosure or partial coverage.

Ethical and legal compliance: Learning analytics systems need to be tested with cohort data either real or synthetic. Testing may be across different institutions using the products of multiple vendors. The key ethical and legal issues arising in the literature around learning analytics are summarized by Sclater (2015a) and addressed in Jisc's Code of Practice for Learning Analytics (Sclater & Bailey, 2015). Using synthetic data can help to avoid ethical and legal issues, in particular breaching the privacy of "real" users and the need for institutions to adhere to strict data protection regulations. European legislation, for example, prohibits the transfer of personally identifiable data outside the European Economic Area except in strictly controlled circumstances; the use of synthetic data means that researchers can collaborate internationally without needing to be concerned about breaching such laws.

Staff training: A paper reporting the experiences from the deployment of analytics services noted that the "the initiative was hamstrung by a lack of availability of data management experts who could devote the amount of time necessary to produce and disseminate the datasets in a form that the researchers could use on an ongoing basis" (Buerck & Srikanth, 2014, p. 133). For staff to interact with and maintain complex systems requires training. Applying synthetic data to the systems again avoids privacy issues and allows data to be created that model the full range of outcomes, some of which may not yet have been created by "real" students. For example, the full ontology and the recipes describing new types of interactions have yet to be defined for learning management systems (Kitto, Cross, Waters, & Lupton, 2015). The synthetic data themselves can be considered self-descriptive, giving the trainees valuable context information during simulations of their working environment.

Additional services: New big data services will emerge. A tried and tested set of synthetic data will enable these to be tested alongside existing services.

4 UvAINFORM

The University of Amsterdam initiated the UvAInform project in 2013 (Kismihók & Mol, 2014) in order to coordinate strategically institution-wide learning analytics services. The project has evolved from one that initially took a centralized approach to the development and implementation of these services. It is now

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

more devolved, involving seven different pilot projects across the various faculties of the university. The objective is to gain experience, learn lessons, and develop expertise across the university. Furthermore, the project initiated the development of an open source Learning Record Store (LRS) to collect student activity (Apereo, 2015c), which in combination with a data warehouse and an open source Extract Transform and Load layer (Roldan, 2015) aimed to unlock the large number of data silos within the university, many of which were never developed specifically for learning analytics purposes.

Given the data-driven nature this endeavor (at least for the time being), the UvAinform pilot project leaders are not always able to fully articulate their intentions and/or desires. With no firm policy framework in place to guide and direct data governance, the vision of a fully data saturated LRS remains elusive. Budgetary and political constraints meant that instead of developing an overall strategy for a university-wide learning analytics framework, a less ambitious approach needed to be taken. This approach entailed having seven faculty-level pilots, which set out their requirements regarding connecting specific data sources to the LRS.

During the initial stages of the UvAInform project, 61 different information systems (IS) that use and store education-related data have been identified (Kismihók & Mol, 2014). Some of these systems are core elements of educational activities (such as the university's Learning Management System); some are minor software targeting a specific educational or administrative aim (e.g., faculty level thesis administration). Some of them are well integrated, but most systems exist as "islands" or data silos, without communicating with any other IS. Furthermore, silo gatekeepers are understandably wary of granting access to "outsiders" to "their" data sources (Kismihók & Mol, 2014).

The LRS was populated with activity data from the university's LMS (Blackboard) combined with data from the Student Information System (SIS) and the timetabling system. The range of data is expected to increase rapidly and to include more sources centred on the group activities found in flipped classrooms, especially video clips and forums. Even though the current pilots only use these three data sources, a number of UvAInform project members were facing challenges, including:

- How to transmit large amounts of data from the three sources to the LRS
- How to transmit large amounts of data from the LRS to the dashboards associated with the pilot projects
- On what basis should partners set the technical requirements of data management for the seven pilot projects? Should the infrastructure be centralized or decentralized? The pilots by their nature are short term and use relatively few resources but if successful may need to be generalized or scaled-up quickly.
- Testing the scalability of both the LRS and the pilot systems in terms of data processing and data management
- Lack of experience within the organization about data delivery. How to share the data with authorized users ethically and technically? How should the university govern such authorization?

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

- A lack of empowerment and influence on the part of key UvAInform stakeholders to evangelize and facilitate the cultural change associated with this challenging data-governance issue

A communication tool such as the learning analytics readiness instrument (LARI), a survey to measure institutional readiness would have helped us to understand and communicate where to focus our efforts (Arnold, Lonn, & Pistilli, 2014). However, the UvAInform pilots were a necessary precursor to developing an institutional culture of greater data-driven decision making.

4.1 Transparency of Educational Data Management

A study revealed that 22 internal stakeholder groups have an interest in the UvAInform project (Szörényi & Kismihók, 2015). This puts management in a difficult position since it is close to impossible to meet the requirements of all stakeholder groups. With few exceptions, all of these groups have claims on educational data. They are either data creators (students, teachers), data managers (technical support, IS management), or policy and decision makers (legal, ethical boards, and management bodies that use educational data for their decisions). The majority of these stakeholders face issues with overall educational data management, such as:

- Overseeing the data management processes within the organization
- Obtaining a clear picture of precisely what individual-level data is being recorded
- Knowing what is happening with the individual data (which IS at the university uses what data and how)
- Finding the barrier between the data the university is responsible for and the data that does not fall under its authority (for instance social media data, data generated by mobile devices, or data mirroring labour market information in a student goal-setting application; see Kobayashi, Mol, & Kismihók, 2015)
- Deciding under what circumstances data can leave the premises of the university. Ongoing research at the University of Amsterdam has revealed, for instance, that students have little idea about how their educational data is being managed by IS vendors and the government (Stuurman, 2015).

4.2 Empowerment of Learning Analytics Research

According to the lessons we learnt during the UvAInform pilots, learning analytics researchers and teachers involved in experiments around learning analytics have limited possibilities to pilot their software and algorithm prototypes. Lack of access to relevant data sources, due to the aforementioned characteristics of the local information architecture and its decision-making loops, can impede the progress of research. There is a clearly articulated need for a “data sandbox” that accurately models the data structures and data types of the various ISs in the organization. Breaking down data silos takes time. However, synthetic data will allow researchers and affiliated technical staff to build the services before the politics, ethics, legal, and data-cleaning issues have been resolved, or even find out what data exists.

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

This allows research to work in parallel with those processes, significantly decreasing the time to delivery of services to the target audience.

To summarize, we believe that two key lessons may be drawn from the UvAInform Learning Analytics Program. First, data centralism is key to developing a learning analytics framework and facilitating the development of learning analytics services at the university level. Second, a central body in the organization, such as the Centre for Data Governance and Innovation, could serve as a hub for facilitating the discourse and innovation around the ethical and privacy concerns raised by creating large-scale learning analytics frameworks. The Centre is a natural part of the organization to curate the synthetic data and benchmarks.

Although not ideal, the risk of an emerging digital divide for researchers with data centralism and those without is diminished as methodologies can be tested with synthetic data and used to cross-validate learning analytics projects.

5 Cross-Institutional Adoption of LA

The previous two sections have explored the value of synthetic data within real world situations. The UvAInform project took place in a typical university that wanted to research the requirements and impact of learning analytics. All the data used comes from within the university. Meanwhile the Jisc learning analytics architecture is a prototype for regional or national services. Here the activity data comes primarily from the participating organizations that consume the services, and remains under each institution's control. It is not yet possible to quantify the amount of self-declared data that will be provided by students using the student app or other input mechanisms. The volume and complexity of this "big" self-declared data will increase as the service matures and the service providers explore new ways of utilizing it.

A further theme to explore is the trend towards the use of learning activity data outside organizational boundaries. Online learning occurs, of course, not just within the organization's systems, but can take place within a wide variety of social media platforms and other web-based systems. This has an impact on the availability, the quality of learning activity data, and the increasing richness of the data that learning analytics services can utilize. It implies that the synthetic data generator needs to be flexible and cover an ever-increasing set of rich data sources. This section details the pressures, and briefly examines synthetic data's role for these sources.

Learning activity data provides challenges for a university's data governance processes. One of these is that students and teachers are regularly engaged in learning activities outside the sphere of control of the educational body or regional services in which they are embedded. There is an incremental loss of access to data caused by the increasing number of *globalized* services (such as MOOCs, Google Docs and Twitter) used. This lack of control over the data by the institution may increase the legal and ethical risks for data

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

subjects, particularly students.

The globalization of services is due to a number of trends, which include:

Pedagogical practices centred on blended learning and the flipped classroom (Bishop & Verleger, 2013). These are actively engaging groups of students with social media. The application of services such as Facebook (Junco, 2012; Ahn, 2013), Twitter (Junco, Heiberger, & Loken, 2011), YouTube (Ammari, Lau, & Dimitrova, 2012), and other social media has been shown to improve student engagement. Although there are pitfalls, such as the quality of the content provided (Duncan, Yarwood-Ross, & Haigh, 2013), these engagement tactics imply that a considerable percentage of the activity data has escaped central control. A natural consequence is that it might be feasible to collect some of the data for some of the students, but not the full range. Learning analytics services will need to replace some of the missing data to optimize the value of the collected data. The impact of example replacement methods to fill in the gaps is discussed by Farhangfar, Kurgan, and Dy (2008).

Traditional LMSs have tended to fully integrate the majority of their functionality such as wikis, forums, chat, polls, and resource areas within one application. The higher education sector is moving away from the concept of a monolithic LMS where all the services are contained in one application to a **thinner LMS** that orchestrates and enhances learning partially through a series of external tools fulfilling specific functionality (Dagger, O'Connor, Lawless, Walsh, & Wade, 2007). In general, the trend is towards thinner LMSs orchestrating a collection of third-party services. The design practice supports scalability and eases the effort to migrate and support third-party specialization. IMS Global's Learning Tools Interoperability (LTI) protocol allows a standalone application to appear to be working within different LMSs. The number of tools mentioned on the LTI conformance page (IMS Global, 2015) evidences the popularity of this approach. The Caliper sensor API[2] builds on this approach and allows for the collection of activity data in a standard format from a range of systems. IMS Global is working on an LTI compatible extension to track activity with a standardized ontology. The authors expect there to be a data quality divide between applications that apply learner activity standards, such as Caliper and xAPI (Kevan & Ryan, 2015; ADL, 2015), and non-standards-based applications. A synthetic data generator with generic capabilities to generate output for these standards will by default cover a wide and increasing range of compatible tools.

Dahlstrom, Brooks, and Bichsel report for the American higher education sector that "the average age of an LMS is eight years, and 15% of U.S. institutions are planning to replace their LMS within the next three years" (2014, p. 3). Although the velocity of replacement of a full LMS is relatively slow, the use of standards enables the incremental diversification of feature sets outside the LMS and therefore wider diffusion of learner activity. Dahlstrom et al. also note, "User satisfaction is highest for basic LMS features and lowest for features designed to foster collaboration and engagement" (2014, p. 4). If user satisfaction is the dominant driver, then expect an increasing range of applications used to foster better engagement

[2] http://www.imsglobal.org/IMSLearningAnalyticsWP.pdf

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

and collaboration. Therefore, the application of popular social media will continue to increase. A synthetic data generator will need to be flexible to adapt quickly to the sector-wide incremental evolution of LMS services.

MOOCs: There are differences between US and European attitudes towards the take-up of MOOCs, with European Universities planning more significant adoption (Jansen & Schuwer, 2014). There is a difference, for example, in the social media systems used locally e.g., Xing (Statista, 2015). Further variations include distribution across the alphabet of surnames (and one would therefore assume login names; ISOGG, 2015), the language of the content within the MOOCS, and the demographic weighting of the students and teachers. All these factors will influence the way that synthetic data is generated.

MOOCDb (Veeramachaneni & Dernoncourt, 2013) is an MIT project that enables researchers and practitioners to share MOOC data in a common format. If European adoption is a significant trend influencing the overall use of MOOCs and a representative portion of the activity is shared via MOOCDb then we should use the MOOCDb dataset to shape the synthetic data generator's output.

Cloud services enable outsourcing of what were traditionally considered core services such as e-mail (e.g., Google Mail) and LMSs (e.g., Canvas, Apereo OAE). Bedrossian et al. noted, "The economies of scale, resiliency, flexibility and agility provided by cloud computing are rendering the construction and maintenance of on-premises data centers obsolete" (2014, p. 2). However, there are significant concerns about security in the cloud and potential solutions such as trusted third parties (Zissis & Lekkas, 2012) that will impact the availability and practices surrounding activity data.

Universities are increasingly using **federated identity management** to share services and enable students to learn across organizations. For example, SURF (2015), the Dutch higher education federation, lists over sixty services and itself is attached to an overarching hub of federations known as Edugain.[3] As the popularity of the federative approach to services widens, organizations will need to share their activity data and uniformly apply student consent rules. Synthetic data will allow researchers to simulate the impact of adoption of different consent processes.

Devices in general such as activity trackers, the Internet of Things (Swan, 2012), room occupancy, brain computer interfaces (BCI), EEG devices for emotion mapping, occupancy sensors, house networks and car networks may play a role in supporting learning. **BYOD** (Bring Your Own Device) policies at institutions are encouraging the use of tablets, smartphones, smart watches, and e-readers with Wi-Fi connectivity, and enabling the viewing of content and interaction in different ways to desktop computers. For example, third-party apps allow you to monitor your heartbeat through the camera on your smartphone or use it as a clicker device. These new applications have the potential to impact course design. The generated data can then be fed back into predictive models, which then trigger interventions. This increases the range of

[3] https://technical.edugain.org/status.php

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

data sources relevant for learning analytics services. Complexity leads to insecurity and an increase in the risk of successful implementation of learning analytics services. The complexity and range of interactions possible in the scenarios mentioned make it difficult to secure the data.

Appropriate use of BYOD can improve the grades of students (Cristol & Gimbert, 2013). Thomson (2012) noted that we should not focus on issues such as whether to allow people to use their iPads at work. Rather, focusing on solutions is the bigger business challenge — enabling technology for competitive advantage. We should take into account the concerns of the consumer (student, teacher, etc.). Lebek, Degirmenci, and Breitner (2013) surveyed 151 employees and found that security aspects and the legal situation worry employees more than their individual privacy. The implication of these concerns is that we should focus our efforts on full end-to-end testing before considering building a sophisticated student consent service.

Hashizume, Rosado, Fernández-Medina, & Fernandez (2013) identify the main vulnerabilities for cloud computing. The list of vulnerabilities and countermeasures should be considered a limited subset of all the possible attack vectors. The value of personally identifiable information (PII) is high and, as has been seen in recent high-publicity data breaches (e.g., BBC, 2015), significant reputational impact occurs when the data is accidentally disclosed. The complex technical infrastructures involved require frequent expert testing to minimize the risk of exposure. Synthetic data again can perform a vital role by allowing early testing before the systems are fully secured. Furthermore, the data itself can be considered a form of documentation; by exchanging synthetic data, developers have more opportunity to validate the end-to-end processes of their software and to test its performance. Synthetic data generation is also applicable for the multiple new Internet-connected devices that are emerging. Anderson, Kennedy, Ngo, Luckow, and Apon (2014) note that research on Internet of Things data can be constrained by concerns about the release of privately owned data, and have therefore implemented a synthetic data generator to help diminish this issue.

6 DISCUSSION

In this paper, we reviewed a university project based on faculty pilots, and a national infrastructure that has the potential to become a template for further large-scale projects. We then looked at some of the challenges for the sharing of learning experience outside traditional data silos, with the data being spread across legal and geographical boundaries. Under these pressures, it is difficult to fully optimize data-driven analytics services with a set of real “big” data. We argue for a comprehensive, realistic, shared set of synthetic data generated through an easy-to-apply tool. The synthetic data should encompass all systems with which the student or teacher interacts. This will enable practitioners to prioritize the data requirements and governance around learning analytics services. The tool will empower designers to explore a full range of possible services without the barrier of gathering data from multiple and idiosyncratic infrastructures.

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

For governance processes, a simple solution is to ignore the external data and consume only the data from internal data silos. However, the Jisc infrastructure empowers students to incorporate self-declared data. This strategy will be put under pressure as external learner activity increases and predictive models and associated interventions using this external data are energetically adopted. It is not only a loss of control of the activity data that requires careful examination of data governance, it is also the increasing number of third parties involved, scattered across many geographical locations. These organizations are under the authority of a number of legal frameworks driven by different cultures of ethics and privacy. The self-declared approach neatly avoids complex policy decisions and supports fine-grained student consent. Delegation also avoids a significant degree of central administrative effort. This delegation empowers the student to choose to share the data. However, if only a portion of the students within a cohort connect their external data, this will cause issues with the coverage of the values returned from predictive models. Synthetic data can play a role in replacing the missing data (Baraldi & Enders, 2010); for example, replacing missing data with mean values or estimates from regression models. Synthetic data can also support simulations to estimate the thresholds set for when the volume of student self-declared data is acceptable as an input to student retention systems.

Even if we design in well-articulated governance processes, if we cannot secure to a high degree of certainty the data within the boundaries of trusted parties, wherever the learning experience takes place, then the governance process is flawed. For large service providers (Google, Amazon, Microsoft, etc.) individual universities will not be able to exert enough pressure to achieve reasonable data governance processes. For a sector-wide, global data governance body that represents the concerns of universities, the collective influence over third parties is significantly greater. For example, it could recommend standards that government procurement agencies should adhere to, and define sector-wide policies and best practices around the full end-to-end process.

Meanwhile, the more complexity there is, the more testing is required to manage risks and deliver stable and secure services. Synthetic data naturally supports data-driven testing. In the medical field synthetic data has been used to generate patient records that collectively simulate the outbreak of infectious diseases (Buczak, Babin, & Moniz, 2010) avoiding privacy and anonymization issues. Buczak et al. note that there is no consistent set of test data and that only a small number of institutions have a full set of data. We argue that the same conditions currently exist within the field of LA.

Large-scale infrastructures being built for learning analytics services deliver wider opportunities, such as academic analytics services focused on the management of institutions. Promising for curriculum design is the work at The Open University UK (Rienties, Toetenel, & Bryan, 2015) where individual learning trajectories are aggregated to look at learning design patterns. The aggregation across curricula is not possible without central control of learner data. Once research leads to services, universities with data centralism will have significant advantages, such as the early exploration of a richer and more representative set of data, compared to the unconsolidated universities. A broad range of realistic synthetic data will enable researchers to design and test their research practices and algorithms,

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

enhancing the degree of potential co-operation across an emerging divide.

Standards such as xAPI and Caliper enable the storage of learning activity data in well-defined formats. However, the recipes around how to use those data structures do not yet cover the majority of learning scenarios and are not widely adopted. An example of defining relevant recipes is that of Kitto et al. (2015). However, this research needs further expansion and adoption of recipes to cover a much greater range of situations. The lack of a fully documented and accepted range of recipes risks inconsistent application, implying greater effort in consolidating activity datasets, increasing costs, and potentially slowing down research projects. Once a range of recipes has been accepted, adaption of tools such as the simple Apereo (2015b) stress test plans will allow for the generation of a wider set of reference datasets. This approach easing the issues mentioned in previous sections such as accidental disclosure or the inability to test complex infrastructures.

7 CONCLUSION

The literature review showed that synthetic data generation is widely applied outside the field of learning analytics. Because educational data mining and learning analytics research are closely related, synthetic methodologies are, to an extent, already embedded within specific learning analytics research methods. There is a small set of clearly applied applications within the field, such as in the performance testing of learning record stores and supporting training exercises through MOOC courses.

We have discussed the significant drivers for increasing the richness of learning activity, and hence the increasing production of learning activity data. This is due to pressures such as the adoption of online teaching methodologies and the increasing range of online services. Meanwhile many universities are expanding the use of analytics services. This combined with potentially highly rich datasets is increasing the need for synthetic data generation. The complexity of interactions and range of possible data sources, combined with the need to avoid accidental disclosure, require a synthetic data generator that is easy to extend, simulating a wide range of real datasets.

The current state of benchmarking for big data where “workloads currently discussed in the testing and benchmarking community do not capture the real complexity of big data” (Alexandrov, Brücke, & Markl, 2013, p. 1) argues for continued research specifically around the theme of capturing the richness and range of the datasets. As a community, we should consider building or adopting an easy-to-use, easy-to-extend synthetic data generator that generates realistic learning activity data. As a standards-based learner activity collection is increasingly adopted within higher education, synthetic xAPI data generation will become increasingly necessary. The xAPI recipes mentioned by Kitto et al. (2015) are a starting point for a generator. The improvement of the test plans held by the Apereo Foundation is a potential solution for a reference implementation.

The generation of rich datasets for testing learning analytics applications requires coordination across the

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

community of researchers and developers in higher education and liaison with vendors. A significant opportunity exists to work collaboratively towards generating standards-based synthetic datasets to ensure robust, secure, scalable architectures, and valid learning analytics.

8 ACKNOWLEDGEMENTS

We would like to acknowledge Professor Robin Boast and thank him for his valuable comments.

REFERENCES

Abhari, A., & Soraya, M. (2010). Workload generation for YouTube. *Multimedia Tools and Applications, 46*(1), 91–118. http://dx.doi.org/10.1007/s11042-009-0309-5

Abtew, W., Moras, R., & Campbell, K. (1990). Synthetic precipitation data generation. *Computers & Industrial Engineering, 19*(1–4), 582–586. http://dx.doi.org/10.1016/0360-8352(90)90185-O

ADL (Advanced Distributed Learning). (2015). xAPI specification. *Produced by the Experience API Working Group in support of the Office of the Depute Assistant Secretary of Defense (Readiness) Advanced Distributed Learning Initiative.* Retrieved from https://github.com/adlnet/xAPI-Spec/blob/master/xAPI.md

Ahn, J. (2013). What can we learn from Facebook activity? *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge*, 135–144. http://dx.doi.org/10.1145/2460296.2460323

Alexandrov, A., Brücke, C., & Markl, V. (2013). Issues in big data testing and benchmarking. *Proceedings of the 6th International Workshop on Testing Database Systems*, (article 1). New York: ACM Press. http://dx.doi.org/10.1145/2479440.2482677

Ammari, A., Lau, L., & Dimitrova, V. (2012). Deriving group profiles from social media to facilitate the design of simulated environments for learning. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 198–207. http://dx.doi.org/10.1145/2330601.2330650

Anderson, J. W., Kennedy, K. E., Ngo, L. B., Luckow, A., & Apon, A. W. (2014). Synthetic data generation for the Internet of things. *Proceedings of the 2014 IEEE International Conference on Big Data* (pp. 171–176). Washington, DC, USA: Institute of Electrical and Electronics Engineers.

Apereo. (2015a). Learning Analytics Initiative (LAI). Available at https://confluence.sakaiproject.org/display/LAI/Learning+Analytics+Initiative

Apereo. (2015b). Synthetic data generator. Retrieved from https://github.com/Apereo-Learning-Analytics-Initiative/LRSLoadTest

Apereo. (2015c). Larissa LRS. Retrieved from https://github.com/Apereo-Learning-Analytics-Initiative/Larissa

Arnold, K. E., Lonn, S., & Pistilli, M. D. (2014). An exercise in institutional reflection: The learning analytics readiness instrument (LARI). *Proceedings of the 4th International Conference on Learning Analytics and Knowledge,* 163–167. http://dx.doi.org/10.1145/2567574.2567621

Arnold, K. E., & Pistilli, M. D. (2012). Course Signals at Purdue: Using learning analytics to increase student success. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 267–270. http://dx.doi.org/10.1145/2330601.2330666

BBC. (2015, 11 November). TalkTalk hack to cost up to £35m. *BBC News*. Retrieved from http://www.bbc.co.uk/news/uk-34784980

Bahga, A., & Madisetti, V. K. (2011). Synthetic workload generation for cloud computing applications.

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

Journal of Software Engineering and Applications, 4(7). 396–410. http://dx.doi.org/10.4236/jsea.2011.47046

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*(1), 5–37. http://dx.doi.org/10.1016/j.jsp.2009.10.001

Barrett, C. L., Beckman, R. J., Khan, M., Kumar, V. S. A., Marathe, M. V., Stretz, P. E., Lewis, B. (2009). Generation and analysis of large synthetic social contact networks. *Proceedings of the 2009 Winter Simulation Conference* (WSC), 13–16 December, Austin, TX, USA (pp. 1003–1014). IEEE. http://dx.doi.org/10.1109/WSC.2009.5429425

Barse, E. L., Kvarnstrom, H., & Jonsson, E. (2003). Synthesizing test data for fraud detection systems. *Proceedings of the 19th Annual Computer Security Applications Conference* (CSAC), 384–394. http://dx.doi.org/10.1109/CSAC.2003.1254343

Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice, 30*(6), 415–429. http://dx.doi.org/10.1016/0965-8564(96)00004-3

Bedrossian, A. (2014, November). Cloud strategy for higher education: Building a common solution. *ECAR Research Bulletin*. Retrieved from https://library.educause.edu/~/media/files/library/2014/11/erb1413-pdf.pdf

Berg, A. (2015). Learning analytics, of standards architectures and grand challenges. [Web log post, June 10]. Retrieved from https://blog.surfnet.nl/en/learning-analytics-over-standaarden-architectuur-en-grote-uitdagingen/

Boggs, N., Zhao, H., Du, S., & Stolfo, S. (2014). Synthetic data generation and defense in depth measurement of web applications. *Proceedings of the 17th International Symposium on Research in Attacks, Intrusions and Defenses (Lecture Notes in Computer Science, Vol. 8688),* 234–254. http://dx.doi.org/10.1007/978-3-319-11379-1_12

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems, 34*(3), 483–519. http://dx.doi.org/10.1007/s10115-012-0487-8

Botta, A., Dainotti, A., & Pescapé, A. (2012). A tool for the generation of realistic network workload for emerging networking scenarios. *Computer Networks, 56*(15), 3531–3547. http://dx.doi.org/10.1016/j.comnet.2012.02.019

Buczak, A. L., Babin, S., & Moniz, L. (2010). Data-driven approach for creating synthetic electronic medical records. *BMC Medical Informatics and Decision Making, 10*–59. http://dx.doi.org/10.1186/1472-6947-10-59

Buerck, J. P., & Srikanth, P. (2014). A resource-constrained approach to implementing analytics in an institution of higher education: An experience report. *Journal of Learning Analytics, 1*(1), 129–139. Retrieved from https://epress.lib.uts.edu.au/journals/index.php/JLA/article/view/3244

Celik, A. N. (2003). Long-term energy output estimation for photovoltaic energy systems using synthetic solar irradiation data. *Energy, 28*(5), 479–493. http://dx.doi.org/10.1016/S0360-5442(02)00140-8

Cristol, D., & Gimbert, B. (2013). Academic achievement in BYOD classrooms. Paper presented at the 12th World Conference on Mobile and Contextual Learning (mLearn 2013), 22–24 October, Doha, Qatar. Retrieved from http://www.qscience.com/doi/pdfplus/10.5339/qproc.2013.mlearn.15

Dagger, D., O'Connor, A., Lawless, S., Walsh, E., & Wade, V. P. (2007). Service-oriented e-learning platforms: From monolithic systems to flexible services. *IEEE Internet Computing, 11*(3), 28–35. http://dx.doi.org/10.1109/MIC.2007.70

Dahlstrom, E., Brooks, C., & Bichsel, J. (2014). *The Current Ecosystem of Learning Management Systems in*

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

Higher Education: Student, Faculty, and IT Perspectives. Research report. Louisville, Co: ECAR. Retrieved from https://library.educause.edu/~/media/files/library/2014/9/ers1414-pdf.pdf

Duncan, I., Yarwood-Ross, L., & Haigh, C. (2013). YouTube as a source of clinical skills education. *Nurse Education Today, 33*(12), 1576–1580. http://dx.doi.org/10.1016/j.nedt.2012.12.013

Ebner, M., Taraghi, B., Sarantie, A. & Schon, S. (2015). Seven features of smart learning analytics lessons learned from four years of research with learning analytics. *elearning Papers, 40.* Retrieved from http://www.openeducationeuropa.eu/en/article/Assessment-certification-and-quality-assurance-in-open-learning_From-field_40_3

Eno, J., & Thompson, C. W. (2008). Generating synthetic data to match data mining patterns. *IEEE Internet Computing, 12*(3), 78–82. http://dx.doi.org/10.1109/MIC.2008.55

Farhangfar, A., Kurgan, L., & Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, *41*(12), 3692–3705. http://dx.doi.org/10.1016/j.patcog.2008.05.019

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*(5–6), 304–317. http://dx.doi.org/10.1504/IJTEL.2012.051816

Ferreira, A. A., Gonçalves, M. A., Almeida, J. M., Laender, A. H. F., & Veloso, A. (2012). A tool for generating synthetic authorship records for evaluating author name disambiguation methods. *Information Sciences*, *206*, 42–62. http://dx.doi.org/10.1016/j.ins.2012.04.022

Gray, J., Sundaresan, P., Englert, S., Baclawski, K., & Weinberger, P. J. (1994, June). Quickly generating billion-record synthetic databases. *ACM SIGMOD Record, 23*(2), 243–252. http://dx.doi.org/10.1145/191843.191886

Hashizume, K., Rosado, D., Fernández-Medina, E., & Fernandez, E. (2013). An analysis of security issues for cloud computing. *Journal of Internet Services and Applications*, *4*(5), 1–13. http://dx.doi.org/10.1186/1869-0238-4-5

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Proceedings of the IEEE International Joint Conference on Neural Networks*, 1322–1328. Retrieved from http://140.123.102.14:8080/reportSys/file/paper/manto/manto_6_paper.pdf

IMS Global. (2015). IMS Certified Product Directory. Retrieved from http://www.imsglobal.org/cc/statuschart.cfm

ISOGG (International Society of Genetic Genealogy). (2015). Surname mapping. Retrieved from http://www.isogg.org/wiki/Surname_mapping

Jansen, D., & Schuwer, R. (2014). *Institutional MOOC strategies in Europe: Status report based on a mapping survey conducted in October–December 2014*. Retrieved from the European Association of Distance Teaching Universities website http://www.eadtu.eu/documents/Publications/OEenM/Institutional_MOOC_strategies_in_Europe.pdf

Jayaprakash, S. M., Moody, E. W., Lauria, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics, 1*(1), 6–47. Retrieved from https://epress.lib.uts.edu.au/journals/index.php/JLA/article/view/3249

Jeske, D., Lin, P., Rendon, C., Xiao, R., & Samadi, B. (2006). Synthetic data generation capabilities for testing data mining tools. *Proceedings of MILCOM 2006 – 2006 IEEE Military Communications conference*, 1–6. http://dx.doi.org/10.1109/MILCOM.2006.302440

Junco, R. (2012). The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement. *Computers and Education, 58*(1), 162–171. http://dx.doi.org/10.1016/j.compedu.2011.08.004

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

Junco, R., Heiberger, G., & Loken, E. (2011). The effect of Twitter on college student engagement and grades. *Journal of Computer Assisted Learning, 27*(2), 119–132. http://dx.doi.org/10.1111/j.1365-2729.2010.00387.x

Kevan, J. M., & Ryan, P. R. (2015). Experience API: Flexible, decentralized and activity-centric data collection. *Technology, Knowledge and Learning*, *21(*1*),*143-149. http://dx.doi.org/10.1007/s10758-015-9260-x

Kinney, S. K., Reiter, J. P., & Miranda, J. (2014). SynLBD 2.0: Improving the synthetic longitudinal business database. *Statistical Journal of the IAOS, 30*(2), 129–135. http://dx.doi.org/10.3233/SJI-140808

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., & Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic longitudinal business database. *International Statistical Review*, *79*(3), 362–384. http://dx.doi.org/10.1111/j.1751-5823.2011.00153.x

Kiron, B. D., Shockley, R., Kruschwitz, N., Finch, G., & Haydock, M. (2011). Analytics: The widening divide. *MIT Sloan Management Review*, *53*(2), 1–21.

Kismihók, G., & Mol, S. T. (2014). Barriers to adoption for learning analytics at a Dutch University. Presented at the Learning Analytics Summer Institute, Utrecht, the Netherlands. Retrieved from https://lasiutrecht.files.wordpress.com/2014/06/uvainform-presentation-lasi-utrecht-2014.pdf

Kitto, K. (2015). CLRecipe: A xAPI based connected learning recipe for use with the CLA toolkit. Retrieved from https://github.com/kirstykitto/CLRecipe

Kitto, K., Cross, S., Waters, Z., & Lupton, M. (2015). Learning analytics beyond the LMS. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge*, 11–15. http://dx.doi.org/10.1145/2723576.2723627

Kobayashi, V. B., Mol, S., & Kismihók, G. (2015). Labour market driven learning analytics. *Journal of Learning Analytics, 1*(3), 207–210. Retrieved from https://epress.lib.uts.edu.au/journals/index.php/JLA/article/view/4194

Koester, B. (2015). University of Michigan source code and data associated with practical learning analytics course. Retrieved from https://github.com/bkoester/PLA

Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management, 34*(3), 387–394. http://dx.doi.org/10.1016/j.ijinfomgt.2014.02.002

Lebek, B., Degirmenci, K., & Breitner, M. H. (2013). Investigating the influence of security, privacy, and legal concerns on employees' intention to use BYOD mobile devices. *Proceedings of the 19th Americas Conference on Information Systems*, 1–8. Chicago, IL: Americas Conference on Information Systems. Retrieved from http://aisel.aisnet.org/amcis2013/ISSecurity/GeneralPresentations/8/

Liang, L., Zhong, J., Liu, J., Li, P., Zhan, C., & Meng, Z. (2013). An implementation of synthetic generation of wind data series. *Proceedings of the 2013 IEEE PES Innovative Smart Grid Technologies Conference*, 1–6. Washington, DC: Institute of Electrical and Electronics Engineers. http://dx.doi.org/10.1109/ISGT.2013.6497844

Lin, P. J., Samadi, B., Cipolone, A., Jeske, D. R., Cox, S., Rendon, C., & Xiao, R. (2006). Development of a synthetic data set generator for building and testing information discovery systems. *Proceedings of the 3rd International Conference on Information Technology: New Generations* (ITNG 2006), 707–712. Nevada, USA: IEEE. http://dx.doi.org/10.1109/ITNG.2006.51

Lo, E., Cheng, N., Lin, W. W. K., Hon, W. K., & Choi, B. (2014). MyBenchmark: Generating databases for query workloads. *The VLDB Journal*, *23*(6), 895–913. http://dx.doi.org/10.1007/s00778-014-0354-1

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

Bishop, J., Verleger, M. (2013). The flipped classroom: A survey of the research. *Proceedings of the 120th Annual Conference of the American Society for Engineering Education* (ASEE 2013), 23–26 June, Atlanta, GA, USA: American Society for Engineering Education. Retrieved from http://www.asee.org/public/conferences/20/papers/6219/view

Maciejewski, R., Hafen, R., Rudolph, S., Tebbetts, G., Cleveland, W. S., Grannis, S. J., & Ebert, D. S. (2009). Generating synthetic syndromic-surveillance data for evaluating visual-analytics techniques. *IEEE Computer Graphics and Applications*, *29*(3), 18–28. http://dx.doi.org/10.1109/MCG.2009.43

Matthews, G. J., & Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys, 5*, 1–29. http://dx.doi.org/10.1214/11-SS074

Piantadosi, J., Boland, J., & Howlett, P. (2009). Generating synthetic rainfall on various timescales-daily, monthly and yearly. *Environmental Modeling and Assessment, 14*(4), 431–438. http://dx.doi.org/10.1007/s10666-008-9157-3

Rich, J., & Mulalic, I. (2012). Generating synthetic baseline populations from register data. *Transportation Research Part A: Policy and Practice*, *46*(3), 467–479. http://dx.doi.org/10.1016/j.tra.2011.11.002

Rienties, B., Toetenel, L., & Bryan, A. (2015). "Scaling up" learning design. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge (LAK '15)*, 315–319. http://dx.doi.org/10.1145/2723576.2723600

Roldan, M. C. (2015). Pentaho Data Integration (Kettle) Tutorial. Retrieved from http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+%28Kettle%29+Tutorial

Scheidegger, A., & Maurer, M. (2012). Identifying biases in deterioration models using synthetic sewer data. *Water Science & Technology*, *66*(11), 2363–2369. http://dx.doi.org/10.2166/wst.2012.471

Sclater, N. (2014). Taking analytics to the next stage. *Effective Learning Analytics*. Jisc. [Web log post, September 16] Retrieved from http://analytics.jiscinvolve.org/wp/2014/09/16/taking-learning-analytics-to-the-next-stage/

Sclater, N. (2015a, March 3). A taxonomy of the ethical and legal issues of learning analytics v0.1. *Effective Learning Analytics.* Jisc. [Web log post]. Retrieved from http://analytics.jiscinvolve.org/wp/2015/03/03/a-taxonomy-of-ethical-legal-and-logistical-issues-of-learning-analytics-v1-0/

Sclater, N. (2015b, April 4). Explaining Jisc's open learning analytics architecture. *Effective Learning Analytics.* Jisc. [Web log post]. Retrieved from http://analytics.jiscinvolve.org/wp/2015/04/04/explaining-jiscs-open-learning-analytics-architecture/

Sclater, N., & Bailey, P. (2015). *Code of practice for learning analytics*. Jisc. Retrieved from http://www.jisc.ac.uk/guides/code-of-practice-for-learning-analytics

Sclater, N., Berg, A., & Webb, M. (2015). Developing an open architecture for learning analytics. In *Proceedings of the 21st Congress of European University Information Systems (EUNIS 15)* (pp. 303–313). Dundee, Scotland: European University Information Systems Organisation. Retrieved from http://www.eunis.org/wp-content/themes/eunis/assets/EUNIS2015_Book_of_Abstracts.pdf

Sherlock, D. (2014). Ethics and Privacy in Learning Analytics (#EP4LA). [Web log post] Retrieved from the Learning Analytics Community Exchange (LACE) website http://www.laceproject.eu/ethics-privacy-learning-analytics/

Siemens, G., & Baker, R. S. J. d. (2012). Learning analytics and educational data mining. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12),* 252–254. http://dx.doi.org/10.1145/2330601.2330661

Siemens, G., Gašević, D., Haythornthwaite, C., Dawson, S., Shum, S. B., & Ferguson, R. (2011). *Open*

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

learning analytics: An integrated & modularized platform proposal to design, implement and evaluate an open platform to integrate heterogeneous learning analytics techniques. Retrieved from the Society for Learning Analytics Research website http://classroom-aid.com/wp-content/uploads/2014/04/OpenLearningAnalytics.pdf

SoLAR (Society for Learning Analytics Research). (2015). First annual open learning analytics hackathon: From research to delivery, building production worthy dashboards. [Web log post] Retrieved from http://lak15.solaresearch.org/hackathon

Statista. (2015). Leading social networks ranked by number of visitors in Germany in September 2013 (in millions). [Data] Retrieved from the Statistics Portal http://www.statista.com/statistics/432478/social-networks-number-of-visitors-germany/

Stuurman, S. C. (2015*). Students' attitudes and awareness regarding public universities' data policies: An empirical study* (Unpublished bachelor thesis). University of Amsterdam, Netherlands.

SURF. (2015). Cloud services connected to SURFconext. [Web site] Retrieved from https://www.surf.nl/en/services-and-products/surfconext/cloud-services-connected-to-surfconext/index.html

Swan, M. (2012). Sensor mania!: The Internet of things, wearable computing, objective metrics, and the quantified self 2.0. *Journal of Sensor and Actuator Networks, 1*(3), 217–253. http://dx.doi.org/10.3390/jsan1030217

Szörényi, D., & Kismihók, G. (2015). Contribution of normative stakeholder theory to the improvement of factors affecting educational data warehousing implementation success. In A. M. Teixeira, A. Szucs, I. Mazar (Eds.), *Proceedings of the Conference on Expanding Learning Scenarios: Opening Out the Educational Landscape* (pp. 27-28). Barcelona, Spain: European Distance and E-Learning Network.

Thomson, G. (2012). BYOD: Enabling the chaos. *Network Security*, *2012*(2), 5–8. http://dx.doi.org/10.1016/S1353-4858(12)70013-2

Tzouramanis, T., Vassilakopoulos, M., & Manolopoulos, Y. (2002). On the generation of time-evolving regional data. *GeoInformatica, 6*(3), 207–231. http://dx.doi.org/10.1023/A:1019705618917

University of Michigan. (2015a). Digital innovation greenhouse MOOC data hackathon. [Web log post, October 20] Retrieved from http://digitaleducation.umich.edu/event/digital-innovation-greenhouse-mooc-data-hackathon/

University of Michigan. (2015b). Practical learning analytics MOOC. [Online Course] Retrieved from https://www.coursera.org/course/pla

Van den Bulcke, T., Van Leemput, K., Naudts, B., van Remortel, P., Ma, H., Verschoren, A., Marchal, K. (2006). SynTReN: A generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics, 7*(43). http://dx.doi.org/10.1186/1471-2105-7-43

Varga, T., & Bunke, H. (2008). Perturbation models for generating synthetic training data in handwriting recognition. *Studies in Computational Intelligence*, *90*, 333–360. http://dx.doi.org/10.1007/978-3-540-76280-5_13

Veeramachaneni, K., & Dernoncourt, F. (2013). MOOCdb: Developing data standards for MOOC data science. *AIED 2013 Workshops Proceedings (Volume 1): Workshop on Massive Open Online Courses (moocshop)* (pp. 17–24). Retrieved from http://ceur-ws.org/Vol-1009/aied2013ws_volume1.pdf - page=22

Verbert, K., Manouselis, N., Drachsler, H., & Duval, E. (2012). Dataset-driven research to support learning and knowledge analytics. *Journal of Educational Technology & Society, 15*(3), 133–148.

Waters, A. E., Lan, A. S., & Studer, C. (2013). Sparse probit factor analysis for learning analytics.

(2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics*, *3*(1), 107–128. http://dx.doi.org/10.18608/jla.2016.31.7

Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013) (pp. 8776–8780). Vancouver, BC, Canada: Institute of Electrical and Electronics Engineers. http://dx.doi.org/10.1109/ICASSP.2013.6639380

Zissis, D., & Lekkas, D. (2012). Addressing cloud computing security issues. *Future Generation Computer Systems*, *28*(3), 583–592. http://dx.doi.org/10.1016/j.future.2010.12.006