Foundations of dynamic learning analytics: Using university student data to increase retention

de Freitas, S., Gibson, D., Du Plessis, C., Halloran, P., Ambrose, M., Dunwell, I. and Arnab, S.

Original citation & hyperlink:
de Freitas, S., Gibson, D., Du Plessis, C., Halloran, P., Ambrose, M., Dunwell, I. and Arnab, S. (2015) Foundations of dynamic learning analytics: Using university student data to increase retention. British Journal of Educational Technology, volume 46 (6): 1175-1188
http://dx.doi.org/10.1111/bjet.12212

DOI 10.1111/bjet.12212
ISSN 0007-1013
ESSN 1467-8535

Publisher: Wiley

Foundations of Dynamic Learning Analytics: Using University Student data to increase retention

Authors & Short Biographies:

Sara de Freitas[1] is Associate Deputy Vice Chancellor (Teaching and Learning), she publishes widely in the area of technology enhanced learning and higher education policy. David Gibson[1] is Associate Professor of Learning Engagement and publishes in simulations and games. Coert Du Plessis[3] is a Partner in Deloitte and has responsibility for a team of data analysts. Pat Halloran[1] work in Curtin Teaching and Learning and is a specialist in assessment. Ed Williams[3] work with Deloitte Analytics and has expertise in data analytics. Matt Ambrose[3] has completed his PhD in computer science and has expertise in data analysis. Ian Dunwell[4] is a Senior Research at Coventry where he has developed international expertise in data analysis in game systems. Sylvester Arnab[4] work at the Serious Games Institute and has a specialism in educational technology. Jill Downie[5] is Deputy Vice Chancellor for Education and supports the learning analytics programme at Curtin.

[1]Curtin Teaching and Learning, Curtin University, Perth WA

[3]Deloitte Analytics, Perth, WA

[4]Coventry University, Coventry, UK

[5]Vice Chancellory, Curtin University

Abstract

With digitisation and the rise of e-learning have come a range of computational tools and approaches that have allowed educators to better support the learners' experience in schools, colleges and universities. The move away from traditional paper-based course materials, registration, admissions and support services to the mobile, always-on and always accessible data has driven demand for information and generated new forms of data observable through consumption behaviors. These changes have led to a plethora of data sets which store learning content and track user behaviours. Most recently, new data analytics approaches are creating new ways of understanding trends and behaviours in students that can be used to improve learning design, strengthen student retention, provide early warning signals concerning individual students and help to personalise the learner's experience. This paper proposes a foundational learning analytics model (LAM) for higher education that focuses on the *dynamic* interaction of stakeholders with their data supported by visual analytics, such as self-organizing maps, to generate conversations, shared inquiry and solution-seeking. The model can be applied for other educational institutions interested in using learning analytics processes to support personalized learning and support services. Further work is testing its efficacy in increasing student retention rates.

An Introduction to Learning Analytics

While data analytics capabilities have been developing over the last ten to twenty years there has broadly been a disconnect between business intelligence and the use of data for supporting learning-based hypotheses. For example, while data has been gathering in educational databases, the capability and know-how for using it to advance learning and improve the student experience has barely begun and only rarely been investigated (Ferguson, 2012). With the build-up of data from learning management systems, customer relationship management systems and student-based systems, some universities have begun to investigate how to increase student retention, improve student-centred services and to develop more interactive learning experiences.

The move of universities for example towards a more service-centred ethos often driven by rising costs of education and the introduction of student fees, has created an environment in which data has taken on an increasing value as a proactive tool for ensuring student recruitment, for lifting the quality of service delivery, and finding new ways to make cost savings throughout the sector. In general few of these studies have been visible in the literature and while some of this data has invariably been left undisclosed due to competitive advantage, overall the comparatively recent nature of the software tools and the high costs of data analysis, as

well as the lack of interoperability of datasets and diversity of vendor offerings have left much of this evidence untapped and unpublished.

Recently the central teaching and learning research facilities at Curtin University have undertaken a study that brought together large and diverse datasets (*n*=51,182 with 61 million data elements) to explore the causes and conditions of retained and non-retained students. In the course of the study the research team has utilized various methodologies and approaches to ensure that the dataset can be used to investigate a number of inter-related hypotheses to confront a number of ideas and biases about the influences on student retention. A related intention was to establish a model so that steps could be taken to improve provision for vulnerable groups of students who are most likely to leave university in advance of course completion, often as early as first semester of their studies. It is important to point out that this was not 'singular hypothesis-driven' research as is often the case. Researchers are often channeled into thinking that all questions can be answered in binary terms or in 2 by 2 matrices in an experimental or quasi-experimental framework, or that the alternative is narrative-based qualitatively rich description. Our methods assume that driving factors in real complex systems require handling multiple variables and multiple hypotheses with the substantial aide of computational resources for machine learning. This paper outlines the foundations of those dynamic methods.

We first outline the main areas of consideration of the study, and provide background from a previous study undertaken in 2010 (Deloitte, 2010). The paper introduces a new learning analytics model (LAM) for building capacities in universities for improving the student experience, increasing student retention and providing an evidence-based structure for educational course design and support built around the personalized learner. This view of the student as the unit of analysis is a fundamental shift to a grain size of 'one', which is possibly unique for a retention study of this size. The model process outlined here will guide the creation of a dashboard system for admissions, tutors and student support services.

Learning Analytics: Background and review of the literature

Learning analytics is emerging as a key area of study in education science. However searches for journal articles in the field reveal that this is a study area in its infancy with few scientific studies currently available and few theoretical pieces published at the present time. Early papers include reviews of the field e.g. Siemens and Long (2011); Ferguson, (2012); and Buckingham Shum and Ferguson (2012). Other papers such as: Buckingham Shum and colleagues (2012) have considered the broader issues around the use of 'big data' systems for supporting learning. The first Association for Computing Machinery international conference in learning analytics was held in 2011 (e.g. Duval, 2011; De Liddo et al., 2011) and the earliest models of learning analytics begun with a social learning analytics approach developed by Ferguson and Buckingham Shum (2012). There is also evidence in the literature of new analytics tools being developed (Ali et al., 2012) and there are some early indications of how learning analytics might support personalization of the learning experience. Work by one of the authors has described a more general paradigm shift from knowledge based learning approaches to more experiential learning experiences that utilize mobile and immersive content that can be built around the learner (de Freitas, 2014). Another of the authors has been developing theory for the analysis of user behavioral data from digital learning experiences while acquiring analysis experience with big data sets (Gibson & Clarke-Midura, 2013).

While the rise in higher education costs, the introduction of student fees and greater use of digital media rich content such as game environments and social software communities have substantively altered the education landscape, the emergence of e-learning, online learning and reusable learning content have offered educators new and scalable delivery modes for education that may have substantial benefits for traditional campus based learners. However, the paradigm shift to immersive and personalized learning experiences necessitates dramatic changes in the infrastructure of our learning institutions. Gibson (2012) suggests a more responsive infrastructure is needed that is resilient to rapid changes and consequent employment re-profiling to ensure that higher education has the staff needed to teach in a more dynamic and flexible context. Educational systems need to be more receptive and adaptable to new markets of independent learners and learners from non-traditional pathways and higher education institutions in particular need to be prepared to alter their infrastructure to

provide open access, technology enhanced, personalized learning and support services to meet the demands for universal education.

In addition, while the debates about moving towards research intensivity are critical for competing on an international stage, most universities will also need to focus upon ensuring that the quality of their teaching and learning services is maintained and improved upon in order to increase student success and satisfaction. In this study for example, we estimated a financial impact of over two million dollars saved for every fifty undergraduate students retained into year two onward until graduation. On the way to bottom line impacts, data analytics can play a significant role in ensuring that current students get the highest quality of teaching and learning and the best support and personalised service to ensure that they are retained, achieve to their highest capabilities and enter into the workforce ready to compete on the global stage.

Curtin Student Retention Study

In line with other universities internationally, the issue of student retention is understood as a complex one involving support, social and pedagogic as well as performance factors. Curtin University, along with other universities plans to reduce student attrition rates and increase year on year retention rates to 90% by 2017. As part of this ambitious target, over the last four years, Curtin has been undertaking a student retention study designed to analyse the main causes of drop out (attrition). This in part led to the Transforming Teaching and Learning effort, firstly, to redesign teaching, learning and student support systems and processes and secondly, to actively increase retention rates through better scaffolding, learning and student life service provision through an early warning system. The student retention research project has involved two studies over a four year period, working closely with Deloitte Data Scientists, the initial study in 2010 resulted in a new methodology for exploring existing student data. The second study undertaken in 2013 has resulted in the development of a highly granular dataset for retention analysis and reporting and a Student Discovery Model for modelling student retention over time. The Student Discovery Model uses a semi-supervised neural network algorithm to construct and place students into behavioural profile groupings. Future work will involve elaboration of the learning analytics strategy for the university with an ongoing implementation plan ensuring that the outcomes and implications of the research can be embedded university-wide.

Here we detail some of the findings of the second study. Additional papers will be published that consider the strategic goals of the study and additional work utilising the dataset will be undertaken by research groups across the university. The next section will focus upon the main methodology and approach of the study and lead to the initial model of a learning analytics framework and process-based model in the following section.

Retention Study: Methodology

The Learning Analytics Model (LAM) is a mixed methods methodology (Creswell, 2003) based upon qualitative focused workshops designed to extract hypotheses through engagement with students and staff, iterated with quantitative data discovery and statistical modelling by an expert team. In the 2013 retention study, netted two hundred fifty-six hypotheses distilled from the qualitative data collection phase. Following a review of the available data and as a result of its iterative and dynamic continuous reconsideration, fifty retention hypotheses were selected as possible retention drivers and indicators. These hypotheses were assessed within the Student Discovery Model (SDM). Where required, additional insights were generated through traditional statistics analysis and data mining methods of the modelling data set. The methodology is designed to produce a qualitatively driven and successively refined quantitative study of a very large dataset (see: *Figure 1*).

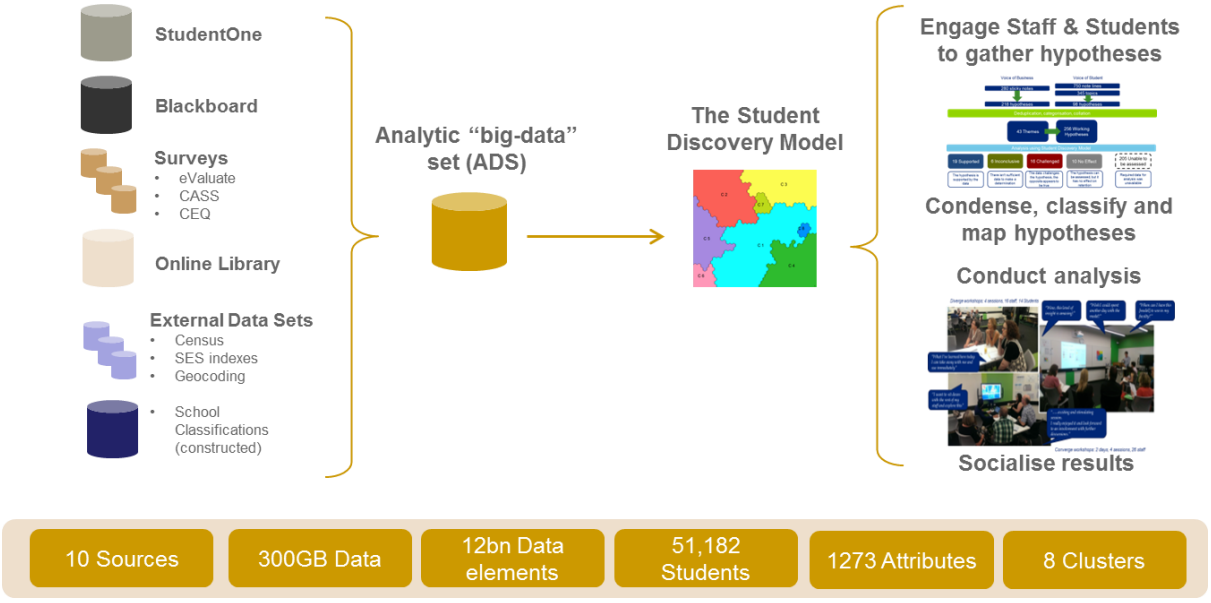

Figure 1: Study methodology: Using Qualitative and Quantitative approaches

The SDM was developed to allow us to map attributes of behaviour represented in the Analytic Data Set (ADS) and to assess the different hypotheses and factors affecting retention. The initial model contained 1,272 measures of behaviour (modelling attributes). Based on the iterative LAM process of expert and stakeholder interactions as well as upon traditional statistical considerations such as data reduction of multi-collinear effects, the final model used 273 of these measures (training attributes) to place students within the Self Organising Map (SOM). The cluster process identified eight mathematically calculated behaviour types. After the SOM algorithm described below had classified the students into behaviour groupings, stakeholders were asked to name the most salient features of each group to provide a short-hand name would capture that sub-group, such as: the first year experience, 'at risk' group, international and mainstream students. These clusters were then held constant during hypothesis testing and real-time exploration of the factors and attributes associated with retention.

The Student Discovery Model is the product of applying a semi-supervised Kohonen learning algorithm using training attributes that were sourced and optimised for modelling from various Curtin student information systems. Supervised machine learning is guided by a target variable that one seeks to explain, hold constant, or discover components of (as in multifactorial analysis, but conducted with nonlinear and categorical methods). Unsupervised machine learning methods are guided by a metric of organization or fitness, so that all the variables are considered against that metric, rather than in relationship to one or a subset of the variables. By 'semi-supervised' we mean a mixed methodology that cycles between supervised and unsupervised machine learning methods along with continuous human guidance as the dimensionality of the variable space is reduced, particularly as the space is discovered to have multi-colinearities and other data issues that impact on prediction, modeling and interpretation of the data. Because we manually include or exclude data based on context (e.g. their weighting or whether we train on them or not) it is referred to as semi-supervised training. The hypotheses generated by actively engaged stakeholders help the expert team make those tuning decisions, which in turn help to maximize map coverage and avoid data elements that overtrain the model on particular variables.

Derived from the *divergent thinking* workshops (three with students and one with staff), 51 hypotheses were produced. These hypothesis were clustered and the top five were then tested against the Student Discovery Model and brought back to the stakeholder groups for confirmation and further exploration.

The dataset was drawn from a variety of sources of student information across the campus. The university learning management system (Blackboard) provided data tracking students use of course work and materials producing 50 billion elements of data; the student management system (Student One) producing hundreds of thousands of rows of authoritative student data including registration and academic results data; unit evaluation

(eVALUate) and CASS data from student surveys; additional data used included socioeconomic data from recent 2011 Australian census, and geo-location data of students for testing hypotheses related to distance to campus impacts on study time and as a measure of work-life-school balance. All the data was cleared for use and utilized under Curtin research ethics procedures and processes including careful blind anonymisation of data, and secure use of datasets in accordance with data protection legislation.

Privacy and anonymisation are key aspects of any big data project and this project is one of the largest undertaken at a higher education institution involving data over a three year period. Student IDs have been anonymised in a two-step process to further secure and permanently maintain privacy. We are taking great care to ensure that all stakeholders (e.g. students, parents, public, staff, researchers) feel that we have not and will not cross the 'creepy line' of privacy (Wolverton, 2013) which is increasingly a feature of modern digital society. The final modelling dataset file contains over 1270 attributes for over 51,000 students in scope, and is the same source used to construct the SDM. Data was also checked to ensure rigor and accuracy, and hypotheses were used to build up an understanding of the underlying SDM. Evidence in support of confirming or disconfirming hypotheses was gathered in an *iterative process designed to confirm or dispel widely held ideas using multiple rather than a single data sources.* All hypotheses in the LAM model are used as guidance for data methods and conclusions are provisional, subject to further confirmation or disconfirmation against the wider retention literature as well as our own future studies.

Curtin Retention Study: Overview

To understand the scope of this study, it is important to point out the re-conceptualization of subjects, not as a group to be studied, but as a unit-of-one with a time-based personalized journey of touch points with the university. The students' journey starts sometime before we get to know them. Our information about the student begins with first contact in marketing, where we may have vast numbers but very little information, and progresses through the admissions process, where both the university and the student must make a decision whether to engage with each other. During these pre-enrolment phases, our knowledge of each individual student begins to grow and helps both parties make a decision about whether to engage in more depth. In future research, we will turn attention to these earlier phases to study the impact of new methods of outreach via processes such as games, challenges and MOOCs and their related learning analytics. When enrolment occurs, our information about the student rapidly expands each year until completion of a degree program. This phase is the scope of the current study. Eventually, we intend to extend our relationship from that point onward with alumni, some of whom may want retraining, advanced degrees, assistance with employment networking, or to foster contacts with international business, government and other students. This student experience timeline might last from four to forty years or more. The current study concentrated on a small window on that journey, but note that even within the three-year timeframe of the study, there is a journey, there is changing and deepening knowledge of the student each year, and therefore in a real sense, there are different subjects each year that need to be reconciled back to the unit-of-one unit of analysis.

There are more than 500,000 students in potential scope for this or any historical learning analytics study at Curtin. So, while the process of retention is taken to begin earlier than enrolment (e.g. whether the student is about to make a good decision and is well matched to the institution's expectations for students), this study focused on enrolled students in a three-year time window and the data available to support this aspect of the student's journey (see: *Figure 2*). In scope for analysis therefore, were 51,181 students who:

- Had at least one active enrolment in a unit during 2010-2013;
- Were enrolled in an undergraduate course;
- Were based on the Bentley, Perth campus.

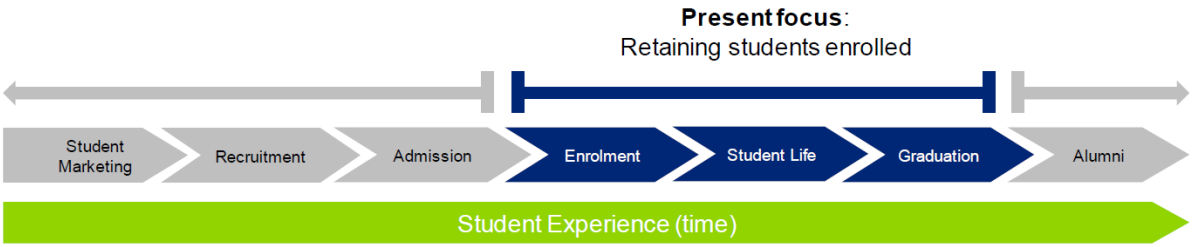

Figure 2: Curtin Retention study within the journey of the student's experience.

Curtin Retention Study: Results & Discussion

The 2013 study provided a number of results to be documented here, based on the SOM produced by data mining, machine learning and statistical decision making guided by the overarching group inquiry into student retention issues. For our purposes here and as a model for how data analytics can be used more generally in higher education, we have extracted some of the most salient results in this section. The focus here will be on retention results, but over time we intend to apply LAM to a host of other questions with impacts on teaching, learning and higher education policy and leadership.

The five main hypotheses clusters selected were around: previous education, cultural background, online engagement, student's mindset, and similarity with the rest of the cohort (Figure 3). After testing we found the hypothesis that:

1) *being an international student could affect attrition adversely* was found to have an inverse effect within the model. International students had a higher retention rate than local students. This finding confirms findings in the literature (Olsen, 2007).

2) *greater use of online materials relative to the students immediate cohort in addition to onsite attendance decreases attrition* was supported by the analysis.

3) *students from private school are better prepared to study and will have higher retention* was found to have no effect with the model. In fact, retention rates for students from both public and private schools were almost identical on the whole. We observed in the SOM four smaller subsets of behaviours where there are some differences that may be investigated for specific micro targeting.

4) *students that who are happy with their academic performance are less likely to attrite* was supported by the study.

5) The hypothesis that: *students closer to the average age of their cohort are more likely to be retained* is supported by the study. Although interestingly students who take a gap year had slightly higher retention rates.

Theme	Hypothesis	Conclusion
Cultural Background	*International students with greater perceived cultural differences are less likely to assimilate and more likely to attrite*	No Effect
Online Engagement	*Use of online lecture materials in addition to onsite attendance decreases chance of attrition*	Supported
Previous Education	*Students from private schools are more motivated to study and will have higher retention*	No Effect
Student Mindset	*Students that are happy with their academic performance are less likely to attrite*	Supported
Cohort	*Students closer to the average age of their cohort are more likely to be retained*	Supported

Figure 3: Results of selected conclusions of 5 retention hypotheses tested against the Student Discovery Model.

In this section we briefly discuss some of the findings and our observations, summarized by new themes.

Blended learning: Use of Blackboard resources and library resources was observed to be associated with higher rates of graduation. Library resource use is not as widespread as Blackboard use. Comparing on-campus use to off-campus use did not indicate there was a meaningful difference in relation to retention. Increased Blackboard use at all distances increases rates of graduation, however we did not observe an increased effect in rates of graduation at greater distances from campus. Of concern, the general pattern of observations shows once distance to campus exceeds 30 kms graduation rates begin to fall.

We did not extensively study students who were 'only online.' We will conduct that study later. However, since the majority of units of study have an LMS site as well as a face-to-face component, and some programs that are largely online still require someone to set foot on the Bentley campus at some point in time, we did include a great number of online students in the study – as long as they had taken at least one unit on the Bentley campus. As expected, comparing the attrition and graduation rates for both external and online study methods shows higher rates of attrition and lower rates of graduation for students who spend more time online. We observed increased rates of graduation as the regularity of LMS use increased. It should be noted that as a cohort people who use the LMS perform better than those who don't. Comparing the student's age, socioeconomic score and LMS usage led to the observation that for non-mature age students, socioeconomic scores has little to no impact on LMS usage. Age was found to be more a determinant of online engagement than socioeconomic factors. Interestingly, for mature age students socioeconomic factors were observed to have an additional effect. The combination of mature age and low socioeconomic score resulted in lower observed LMS usage than just mature age as a factor alone.

Retention rates: Bear in mind that our definition of retention is 'lifetime retention' so comparisons to 'year-on-year' retention rates in other research should be made with caution. In our study and using our lifetime retention definition, Commonwealth supported students show higher rates of attrition compared to International Fee paying students over the students entire undergraduate study period. Also, Domestic Fee paying students show lower rates of attrition compared to Domestic Commonwealth supported students. It should be noted that there are only a small number of domestic fee paying students in scope. Numbers are as follows: international (7,916 students, 82% of commencements have completed a course), domestic HECS/HELP (17,073 students, 52% of commencements have completed a course), domestic fee paying (1,786 students, 75% of commencements have completed a course). This supports the hypothesis that fee payment engenders higher levels of retention. Part-

time students also had the lowest levels of retention when measured over the entire study period, clearly then, motivation or impetus to study and social interactions are central components of retention.

Academic performance: As to be expected, there is a strong relationship between a student's marks and academic performance with their chances of graduation. Students with average marks in the 50's are up to 20% less likely to graduate compared to students with average marks in the 60's or 70's. This relationship is strong in all periods, but becomes stronger as the students progress further through their course. The relationship is not linear, but rather functions with a peak at about 80. We observed that above the peak, attrition starts to increase again. This cohort of high performing students that are leaving will require a unique retention response.

International students: Looking at the graduation and attrition statistics for international students, the hypothesis generated by the stakeholders did not match the data observations. International students with behaviour profiles distinct from domestic students are *more* likely to graduate and less likely to attrite compared to international students who behave and make decisions more like domestic students. This, we believe is a new finding that further explains the difference, but recognizes that some international students who behave like their Australian counterparts then begin to share the lower retention rates of Australians. It should be noted that international students as a group have higher graduation rates than domestic students over the entire undergraduate study period (82%).

Gap year students: Regional gap year was tested as another hypothesis. Investigating gap year in general, the indications are that students who commence University ages 19 or 20, as opposed to school leaving age of 18 are less likely to attrite. However this is skewed by international students who are an intrinsically older cohort. Removing international students reduces the performance between gap and non-gap year students, however gap year students still show higher rates of graduation.

Public versus Private school background: We did not observe any impact on lifetime retention based on the student's high school being with Public or Private. Graduation rates in decreasing order are: International (82.2%), domestic public school (58.3%), domestic private school (58.0%), domestic other and unknown (47.0%)

While these observations and conclusions are in need of further validation through other replicated and comparable studies and against the main literature findings, this study does illustrate how quantitative and qualitative approaches can be combined to cross-validate endogenous hypotheses as well as literature and theory-based findings. While this study's approach relied upon hypotheses derived from workshop activities, other approaches could start from hypotheses or assumptions derived from the literature.

Learning Analytics Model (LAM) for Higher Education

We have developed a Learning Analytics Model (LAM) set of principles that can be utilised by other higher education institutions to use the dynamic analytics process to improve decision-making and business practices.

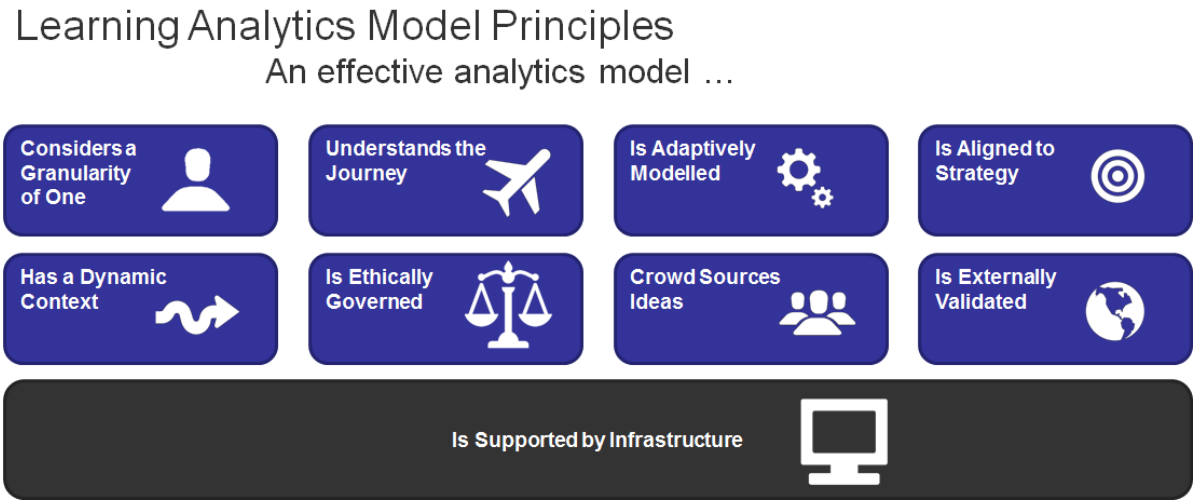

Figure 4: The Learning Analytics Model Principles

1) *Develop a learning analytics strategy*. Having a university-wide analytics strategy will ensure join up of datasets across the institution as part of clustering business intelligence, teaching and learning practices and support services data to inform development around key objectives and mission (e.g. improving learner retention, enriching the student experience, improving the quality of teaching and learning, personalizing support). The components of the strategy include people with the right skills, processes and governance to manage data analytics projects, a culture of collaboration, an infrastructure that is capable of collecting, filtering and storing massive data, and a robust 'sensor net' of data that provides windows into student behaviors, attributes and key events.

2) *Commit to create an infrastructure for big data integration*. Learning analytics capabilities are powerfully linked to the institution's information architecture. A precondition for the consideration of continuous data processing and analysis by key stakeholders for example, which is at the heart of the LAM model, is access to near real-time, always-on, data collection in a fully integrated enterprise system. While large banks and IT companies are accustomed to using SAP systems and utilizing customer relationship management tools for data analysis, many universities have yet to integrate their datasets to allow for maximizing their understanding of their business processes and quality of service provision. Well-connected and maintained data sets are important, so the work of business intelligence groups to clean up existing data is essential, as is a logical and well maintained IT architecture to support efficiencies in achieving accurate and rigorous data collection. More generally, the costs of moving large amounts of data even within cloud infrastructures is significant, therefore systems approaches that bring the code to the data are beneficial. This requires integrating not only the data sources, but the analytics processes and tools into a seamless whole.

3) *Learner-centred service ethos – the unit of one*. Key to our model of the analytics system, the adaptive response of the whole university system is intended to enhance the individual's learning and life experience. Adoption of a learner-centred service ethos is seen as necessary to develop more personalized systems and more effective and immersive learning experiences that make maximum use of data via learning analytics, since at its root, the most important analytics decision concerns how a particular student is experiencing the university. This approach needs to be given a strong emphasis upon defining and developing over time an adaptive and dynamic (rather than only predictive) student experience. The systems need to sit around and flex with the learner rather than be static and embedded due to an institutional driver. This also implies that universities need their analytics capabilities to help them become more flexible and dynamic institutions, able to develop and evolve strong working partnerships with both internal and external organizations and service providers.

4) *Dynamic look at the students' learning journey* (e.g. marketing, admissions, recruitment, enrollment, retention, graduation, employment). It is critical to move from a mass of student data produced by and then

limited by group means to a new more 1:1 personalized approach to data collection for each individual student. Systems and processes also need to evolve that provide solid links to the student's lifecycle taking on board pre-university and post-university trajectories and adapting individually according to each student, triggering supportive behaviours in the institution as a response to knowledge gained about the student, knowledge that changes over time. For example, from this study, we will next move toward a student retention response system that attempts to make use of all available attributes that define the differences among students in order to focus upon the subgroups with the highest risk of attrition.

5) *Adaptively model user behaviour*. Once individual data is gathered, it needs to be modelled adaptively to the user's behaviour and identity. In the evidence-centered design framework (Mislevy, Steinberg, & Almond, 2003) the user model sends signals to the task model, which chooses the next best task for the user, based on a model of how the user's performance data provides evidence of expected performance. The adaptation of the system is several-fold; it is to choose the next best learning engagement, it is to shape the long term learning model of this user, and it is to subject the current model to learning analytics evaluation of the import of one's actions compared to an expert model of performance. We intend to replicate this adaptive engine on the scale of the university as a whole. The drive towards an adaptive university system needs to take into consideration the purpose of the university: to learn, educate and research and needs to filter its own behaviours according to these requirements.

6) *Linking learning analytics within a wider dynamic context.* All data sources that can be tied to an individual student are potentially valuable nodes in a network for analysis. For example, learning performance data (e.g. grades and online resources utilization patterns), buying patterns on campus, (e.g. economic factors) and business drivers (e.g. student intake numbers), socio-economic status, general market datasets (e.g. geo-demographic profiling), and census data may all be important related drivers of behavioral similarities among students. A wider context also means pulling data in from different domains (e.g. student learning domain, student management domain, load planning domain, student projected feedback based on surveys). With such large, complex and highly varied datasets streaming into collections at a fast rate, it is essential to use agile, mixed methodologies for data analysis. A wider contextual frame utilizes all variables as part of the individual's characteristics (e.g. the individual in a set of psychological, social and environmental mutually-reinforcing relationships), and then applies iterative qualitative and quantitative methods to mine the data for patterns of similarity that subdivide the contexts for greater targeting of student behavior types (e.g. within a group of Australian-born students, those with high grades but low satisfaction).

7) *Qualitative driven crowd-sourced hypotheses formation*. An earlier study in 2010 found that 'buy-in' from key players was hard to garner without a clear stakeholder user group associated with and owning the research (Deloitte, 2010). Based on this lesson, the LAM model uses qualitatively driven crowd sourced ideation to generate multiple hypotheses, and has been found to be more effective for building stakeholder engagement and understanding than pure research-focused or quantitative-focused approaches. In our second study, initial '*divergent thinking*' workshops and focus group activities involved staff, students, management and administrative groups in understanding the project's aims and offering over 200 testable hypotheses. A core team then tested those hypotheses against the fine tuned data model, and a final phase of '*convergent thinking'* workshops brought new people into the discussion to explore how hypotheses are supported (or not) by the data, and to invite further open-ended data exploration. The level of stakeholder engagement in 2013 was an order of magnitude higher, more intense, and more widely dispersed across the university than earlier, and the resulting positive impacts can already be felt.

8) *Rigorous view of ethics and adherence to highest standards of ethical procedures*. It is essential that all data be treated with the highest ethical standards, ensured both through university and national guidelines. The crucial elements of ethics and data are: anonymity of data, identity protection and secure data storage as required by law. We have developed a process for doubly-removed anonymity by first securing the student identity codes in the data preparation files, then having those codes hashed again by the data preparation team, so that to recover an identity would take a two-step process with two entities involved in un-encryption of the data. However, since the high resolution of the final Analytic Data Set (ADS) data set still

presents minor risks (e.g. some subgroups in the final model might have small cell sizes and thus lead to potential identity risks of group membership) we thus cannot make the ADS data set or the modelling file available as a public resource. More research is needed to ensure that *if* public data resources are needed, then additional mechanisms may be needed to protect privacy of students. In the meantime, we are openly seeking collaborative research opportunities both within Curtin University and across the higher education sector where we can ensure appropriate data safeguards.

9) *External as well as internal review and cross-validation processes*. The need to continuously review findings against received literature, assumptions and hypotheses is critical and is within the time-honored traditions of all sciences. In addition, internal data reconciliation and cross-validating processes that are prevalent in machine learning helps ensure that false positives are discovered and removed and their ill-founded causal conclusions are not adopted into the training attributes for modeling or the resulting policy and response-action systems. We recommend external evaluation and different validation approaches be used within any study and across studies, and be applied continuously so that as contexts change, so do the findings.

Conclusions and Future Study

The 2013 student retention study provides significant outcomes confirming that a set of indicators for the likelihood of attrition for certain students can be built to provide a simple dashboard for tutors, admissions and student support services, to create adaptive response services for supporting vulnerable students. The iterative approach to learning analytics using the LAM principles orchestrated a timely process with a capable set of tools for identifying user groups with similar behavioral patterns, modelling user requirements and indicating business trends. Based on those groups, patterns and trends, the study provides a foundational framework for creating more effective pedagogies, service mapping and ongoing gap analyses for real-time adaptation of the university's curriculum and student life experience. The dynamic approach also shows how data-driven systems can be co-designed and constructed more closely in line with user group development.

Here we posit the need for a paradigm shift in business intelligence from traditional predictive analysis based on means to what we term *'dynamic learning analytics*,' which captures the notion of a substantial change from the methodologies typically used. The newer method utilizes crowd ideation, group and individually driven hypotheses formation from across a wide spectrum of perception and student engagement in the university, and a continuous interactive link between qualitative and quantitative methodologies, bridged by the modeling exercise supported by nonlinear and cluster-based data methods. In addition, we posit here that changes in the educational landscape favour this more fluid and dynamic analytics approach, because the large, quickly amassing and highly varied datasets will continue to drive the need to bring information together in a highly variable and rapidly changing environment.

Working with stakeholders from industry, student populations, academic research fields and service facilities also can provide broad and sometime contradictory targets and trajectories so focus group mediation sessions and cross-cutting interest areas are useful tools for assimilating the direction of analysis across an institution and it is critical to develop an analytics strategy to accompany these types of studies.

Improvement stimulated by the dynamical learning analytics strategy espoused here will continue to provide a vital and productive resource for the university, impacting upon increasing student retention rates, and shaping the development of appropriate and well-timed adaptive curriculum content, academic guidance and student support services. The resulting data resource will provide us with a foundation for investigating a wide range of hypotheses developed for a range of different applications in core business support, setting academic and institutional priorities and creating new technical and human resource innovations for supporting the student experience. The LAM model will continue to be tested within different frames of reference both within SOM and within the wider data set.

We caution again that ethical considerations are at the heart of all big data studies, maintaining the highest level of caution around data usage and data bias. Since ethics procedure are constantly in flux, and are not always

considered at the heart of research studies in the digital media sphere, issues of data ownership and access are not always easily reconciled in practice and new research is needed here to ensure that guidelines are developed in parallel with institution-wide studies and access to datasets of this scale. While all cross-disciplinary study is welcomed in this ongoing project, adherence to core values and ethical standards are needed that match priorities regardless of academic disciplinary background.

Finally, we envisage the evolution of an advanced and adaptive user model that can draw in from other datasets dynamically while providing a simple dashboard for support services and academic staff to apply new approaches with students individually and in groups. Our vision of learning in the future will draw in pedagogic, employability, work-integrated learning and skills-focused components to ensure a better integration of work, study and life to ensure that student time is well focused and used within a framework of seamless systems that will support and guide students through the best matching set of learning experiences adapted to their strengths, interests and aspirations.

References

Ali, L., Hatala, M., Gašević, D., & Jovanović, J. (2012). A qualitative evaluation of evolution of a learning analytics tool. Computers & Education, 58(1), 470-489.

Buckingham Shum, S., Aberer, K., Schmidt, K., Bishop, S., Lukowicz, P., Anderson, S., Charalabidis, Y., Domingue, J., de Freitas, S., Dunwell, I., Edmonds, B., Grey, F., Haklay, M., Jelasity, M., Karpištšenko, A., Kohlhammer, J., Lewis, J., Pitt, J., Sumner, R. & Helbing, D. (2012). Towards a global participatory platform: democratising open data, complexity science and collective intelligence. European Physical Journal Special Topics, 214(1), 109-152.

Buckingham Shum, S., & Ferguson, R. (2012). Social Learning Analytics. Educational Technology & Society, 15(3), 3-26.

Creswell, J. (2003). *Research design: Qualitative, quantitative and mixed methods approaches*. Thosand Oaks, CA: Sage Publications.

de Freitas, S., (2013). MOOC: The Final Frontier for Higher Education. Coventry. Last accessed online 12th January 2013 at: https://www.dropbox.com/s/pv51ml5zc0kscx7/MOOCs_The_Final_Frontier_report%20%282%29.pdf

de Freitas, S. (2014) Education in Computer Generated Environments. London & New York: Routledge.

Deloitte. (2010). Student retention analytics in the Curtin business school. Bentley, WA.

Duval, E. (2011, February). Attention please!: learning analytics for visualization and recommendation. In Proceedings of the 1st International Conference on Learning Analytics and Knowledge (pp. 9-17). ACM.

Ferguson, R. (2012). The state of learning analytics in 2012: A review and future challenges. Knowledge Media Institute, Technical Report KMI-2012-01.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *... Journal of Technology Enhanced Learning*, *4*(5-6), 304–317. Retrieved from http://inderscience.metapress.com/index/W1QP4L6217K0Q2PV.pdf

Ferguson, R., & Buckingham Shum, S. (2012, April). Social learning analytics: five approaches. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (pp. 23-33). ACM.

Gibson, D. (2012). Game Changers for Transforming Learning Environments. In F. Miller (Ed.), *Transforming Learning Environments: Strategies to Shape the Next Generation (Advances in Educational Administration, Volume 16)* (pp. 215 – 235). Emerald Group Publishing Ltd. doi:10.1108/S1479-3660(2012)0000016014

Gibson, D., & Clarke-midura, J. (2011). Some Psychometric and Design Implications of Game-Based Learning Analytics. In *Cognition and Exploratory Learning in the Digital Age*. Forth Worth: CELDA-IADIS.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE, 78, 9, 1464-1480*.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspective*, *1*(1), 3–62.

Olsen, P. (2007). *Staying the course : Retention and attrition in Australian universities Findings* (pp. 1–16). Sydney. Retrieved from http://www.spre.com.au/download/AUIDFRetentionResultsFindings.pdf

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. Educause Review, 46(5), 30-32.

Wolverton, T. (2013, May 28). Google flirts with what chairman Eric Schmidt once called "the creepy line." *Denver Post*, p. na. Denver. Retrieved from http://www.denverpost.com/ci_23335230/google-flirts-what-chairman-eric-schmidt-once-called

Acknowledgements: The authors wish to acknowledge the assistance of Michelle Rodgers (Director of Student Support Services, Curtin University) for her inputs with the project.

Open Data: Due to data protection legislation, the large amount of source data used for this study cannot be made publicly available. The study was undertaken in line with Curtin University ethics procedures and guidelines.

Ethics statement: Ethics approval for research on the 'foundations of learning analytics' study covering this publication was granted using approval number CTL 1_2014.

Conflicts of Interest statement: The authors are unaware of any conflicts of interest with this study.