Data and analytics

17

This chapter focuses upon data and analytics, which both have the potential to provide value to universities as they seek to reshape their offer and to differentiate themselves from their peers. So-called "big" data and analytics have been two of the hottest buzzwords in the technology press in recent years, and they have also received a lot of coverage in the business press more generally. Analytics is often considered as part of "big data," but is an area of research and development in its own right, as well as being used in many areas of the commercial world. It includes the analysis that is done to data in order to gain insights from it.

As with other technology-driven change, data and analytics hold the potential to both disrupt current business models and also to offer new opportunities, either for new products or for products to be adapted to new usage models. Used well, they may deliver benefits; but used badly, they may be costly and even potentially misleading or even damaging.

Big data is something of a catch-all term that has been somewhat overused in recent years, and misused to capture a whole set of different meanings. In this text, we will not discuss big data specifically, but instead focus upon educational data more generally—that is, the potential to gather and hold information about people, content, and other objects (such as software tools) in digital systems.

The obvious partner to data, then, is the analysis of these data—or more specifically, the new affordances of specialist analytic systems and tools, which have been designed to make it easier to interrogate and analyze the data in a meaningful and useful way.

Taken together, data and analytics have huge potential to offer added value to higher education providers, in particular when potentially working with students who are located away from the university ("virtual" students), and also when providing services to large numbers of students, such as through MOOCs, where Online learning and MOOCs, in particular, make it easier for universities to gather data about learning behaviour and learners.

It may seem that analytics is mostly about getting the right analysis tools and putting them into place. However, for many institutions the far greater challenges are the underlying data that are used to inform analysis, and the systems that hold the data. This is not a new challenge for universities; for many years, they have invested in developing information systems that are appropriate to external reporting requirements on student numbers and also developing appropriate financial and human resources systems. These long-term challenges are thrown into sharp relief, however, when the institution wishes to put analytics in place, in particular to carry out analysis of data that is held in a number of different places, and to carry out real-time analysis of it.

To MOOC or Not to MOOC. http://dx.doi.org/10.1016/B978-0-08-100048-9.00017-2

At Purdue University in the United States, acquiring the data that were needed to be able to carry out analysis was the biggest first hurdle in the implementation of the *Course Signals* analytics system: "our biggest difficulty was the acquisition of data. As a siloed university, different data were owned by different stewards, and compiling all the data needed for the student success algorithm was an uphill battle that took well over a year" (Pistilli, Arnold, & Bethune, 2012).

Purdue is not alone in perceiving that many of its data systems operate in separate silos; this is a common problem in other universities. Ownership of systems and the data in them adds an additional level of complexity; data systems have often been developed for a particular purpose, and the data are often perceived to be sensitive and in need of careful security protection, and these often-legitimate concerns can slow progress to implementation substantially. For this reason, it is important to get appropriate understanding of and buy-in to the MOOC development by management at all levels.

There is a need for clear institutional strategies and policies for MOOC data, in order to get greatest benefit from the investment in MOOCs; indeed, data about learners and their behavior are one of the benefits accrued through investment in MOOCs by institutions (Hollands & Terthali, 2014).

MOOCs and analytics

Analytics are of particular importance to MOOCs. There are characteristics of MOOCs that make it a very different experience to a traditional course delivery situation. Firstly, you do not know your students and because they have no formal relationship with the university, the usual ways that you might gather information about the student, through their interaction with various parts of the university, such as the library or student services, are not available to you. Secondly, the students are likely to join the course for a short period of time, perhaps 8 or 10 weeks in total, which minimizes still further the opportunities to build up a picture. Thirdly, you are likely to have far larger numbers of students taking the MOOC than you have on any other courses. This final point presents a big challenge, when you want to build up a picture of "how the students are doing," but also a significant opportunity, as there will be so many more students engaged with the materials, carrying out assessments, and so on, and so there are opportunities to build up significant quantities of data about the students, their learning behaviors, and so on.

So teachers are faced with a different situation than usual, one where the normal processes for eliciting student feedback will not be possible. Perhaps more importantly, there are exciting, new opportunities that emerge because there is a large cohort of students learning solely online, where their learning journey will create a trail of data as they progress. The teacher is able to find out much more about the students—who they are, why they joined the course, and even why they choose to leave—than they might do normally. Data that are gathered from online learning systems can also enable the teacher to get a much clearer picture than usual of how the students are progressing through the course, and so make it possible to identify where there are

particular problem areas, where students behave differently than you expected, spending more or less time on an activity, for example. The teacher will be able to engage with discussions about course topics as they happen, and respond with guidance or extra resources. The data will come to the course team in "real time" as the course progresses, so they will be able to have a live picture of what is happening on the course, and respond to it, rather than waiting for the end of course evaluation. So one of the benefits of analytics, if used well, is that they can help course teams to design better courses and better learning environments for the students.

Those working at the cutting edge in the field of online education also have grander ambitions for how analytics can be used. Researchers at Stanford University, for example, hope that the rich source of data about how learners interact with materials online may lead to the development of much more sophisticated online learning environments, which may be able to adapt according to how the learner progresses, and particularly to help them with areas where they have difficulties. As researchers of how analytics can help the design of better MOOC experiences, their aim is to "provide educators, instructional designers, and platform developers with insights for designing effective and potentially adaptive learning environments that best meet the needs of MOOC participants" (Kizilcec, Piech, & Schnieder, 2013).

There is also the value that can be reaped from gathering data from courses that are repeated more frequently than usual—perhaps several times per year—to build up longitudinal pictures of key data such as the makeup of student populations, their previous interaction with the university, their progression through the materials, completion rates, and so on. Again, these data are not available for traditional courses but offer lots of possibilities for the institution to monitor course engagement much more closely than usual, and to make decisions about the institution's course offerings in a much more informed and responsive way.

Let us look at analytics more closely, both in terms of general online learning and in terms of MOOCs. Analytics has been used to support institutional strategies for a number of years and this experience can help to guide the use of analytics in MOOCs; in particular, in the support of strategies to help students and increase student retention (or completion). Innovative work is taking place in traditional teaching situations, where the results from analytic tools are used to guide interventions by teaching and support staff. At the University of Bolton in the UK, analytics is being used to drive face-to-face personal tutoring strategies as well as to help retention of students (Powell & MacNeill, 2012).

In the approach taken at the University of Bolton, analytic and information presentation tools, in the form of information dashboard, brings together data that would previously have been held in separate information silos, and possibly not interrogated at all.

In this example, the information provided is about the student's previous and current educational achievements, and the analytic tools automatically analyze the data and provide reports in order to highlight those students whom it predicted may do less well in their studies, or be at risk from dropping out early:

"An information dashboard allows staff, through a series of web-links, to access more detailed information about a student's attendance, and previous and current

educational performance. An algorithm is used to highlight students most at risk of leaving the university, initially based on a combination of their UCAS points and attendance. As the student's career progresses, the UCAS data are replaced by data on their assessment outcomes in module examinations in an attempt to provide a more relevant snapshot of their likelihood of either leaving the university or performing poorly in assessments.

A weekly e-mail summarizes these data and is sent to relevant curriculum managers and personal tutors that flags students those are of concern using a traffic light system where red indicates urgent action is required" (Kraan & Sherlock, 2013).

Not all analytics traditions and tools are relevant to all areas where analytics might be applied. A business intelligence tool that is fine tuned for financial market predictive insights will be of limited use when analyzing student engagement in Virtual Learning Environment (VLE) forums, for example.

In the US, analytics have for several years been a key focus of innovation and development in both traditional education and e-learning systems, as issues of students completing their course are a particularly key concern for many institutions, and analytics are being applied in an attempt to improve the situation. The learning that has been gained from using analytics in both traditional learning and blended learning scenarios can be brought to bear in the MOOC context.

Purdue University has created its own *Course Signals* system, a real-time analytics system which is used to analyze data from a number of sources and to predict students' success with the course.

This example is of particular relevance to MOOCs as it applies to a context of gateway courses that are taken by large numbers of students who then wish to progress on to other courses at the university. In this situation, the teacher is not able to develop personal knowledge of the whole cohort and the students may be particularly challenged to build up their knowledge of an unfamiliar subject area and pass an assignment, with defined time constraints. This has some similarities with the MOOC context.

The *Course Signals* system uses data that is gathered in Purdue University's Course Management system (in this case, Blackboard) which monitors student interaction with the course materials and assignments. It analyzes the data and then gives the teacher a "traffic light" rating (red, yellow, or green) which indicates how likely the student is to pass the course. The teacher is then able to intervene by providing additional resources to the student or direct support. In one experiment, the system predicted that 66–80% of students might need additional support (Pistilli et al., 2012).

Research carried out in 2012 found that those students who were supported by the system did significantly better than others, with average grades higher and lower levels of dropout (Pistilli et al., 2012). Results so far also indicate that students who have been supported with *Signals* will graduate more quickly than their peers: "preliminary results suggest that Course Signals students graduate sooner, allowing them to enter the work world before many of their classmates" (Pistilli et al., 2012). If this trend continues, use of the *Course Signals* analytics system will be really significant for the University and the way that it supports its students.

Analytics in MOOCs

If we apply these kinds of approaches to a MOOC that is being run by a university, similar tools could be applied to try to predict which of the students are likely to complete or not, but the analytics system would need to be "fed" by different sources of data. The university is unlikely to have access to data about the student's prior academic achievements in any level of detail, for example, so for a MOOC, we may instead have to rely upon more usage data about what resources the student has accessed, and for how long, during their engagement with the MOOC material.

At present—and despite early promises—the MOOC systems that are widely used do not include sophisticated data analysis tools like *Course Signals*, or similar tools. In a recent presentation, the CEO of Coursera, Nancy Koller, described the current provision of Coursera data as "one big spreadsheet" (Koller, 2014). In the same presentation, Nancy Koller did however emphasize that both collection of data and analytic tools or data dashboards are a key priority for the development of the Coursera system (Koller, 2014).

There is real potential for those institutions that are able to understand more about their MOOC learners, either individually or as a whole population. Some of the key areas where analysis of data may provide value from MOOCs are described below. Note that these are just illustrative; there are likely to be many more opportunities for data analysis that arise as analytic tools become more sophisticated and as a particular institution builds up their own data set, and can look at patterns of data over time.

Key areas where universities may benefit from using data collection and analytics as part of their MOOC strategy are described below.

1. Student retention and completion of the course

Using analytics will help the university to understand the overall picture of how many students sign up for a MOOC, how many proceed through the course, and how many complete. If the right data are stored, it can also be analyzed at a more granular level to look at how many students have completed particular exercises.

If data can be stored about specific students, than as with the examples above, it can be used to predict future student behavior, for example to predict dropout by particular individuals. More likely, the behavior of the student population as a whole will help the course designers to think about how to design the course, and support activities, in order to keep more students and get them past key points during the course when dropout occurs, by understanding more about why people leave the course at particular points (see below).

2. Improving MOOC content

Analytic data can be used to help course designers to understand more about how the MOOC content is being used and then improve it. Two examples of how data may help to improve MOOCs are given below:

1. Understanding whether the learning design for the MOOC works—is there a particular point in the MOOC at which students leave the MOOC, and is it possible to establish why this is?

Is it because the MOOC has failed to engage students through boring or badly presented materials; whether the course description was misleading and so students have discovered that the course is not what they were expecting; or some other reason?

2. Identifying badly written or designed materials—for example, is there a particular assessment exercise which many learners fail or do not complete, and which may prove to be challenging to large numbers of students? In which case, can the course team plan an intervention, such as an additional support segment, to help students to understand the materials? If analysis is carried out during the course and regular reports produced for teaching and support staff, it may be possible to make these interventions during the course rather than at the end.

We need, however, to bear in mind that data and analytics can only take us so far in understanding the perceptions and behavior of students, and in attempting to predict future requirements and behaviors. Data are only as useful as the questions that are asked of it—and the tools that are applied to it. Although sophisticated analytic tools are available in other fields, in educational situations, the tools are still relatively unsophisticated, and there is little consistent practice that has taken place so far.

This is particularly the case for data in MOOCs, where the potential for analysis of data is still a long way from being fulfilled in reality. This is particularly the case for the large MOOC platforms, which gather immense amounts of data but at present do very little with it. As Hollands and Terthali (2014) note: "while there is no doubt that the MOOC platforms collect oceans of data, it is apparent that interpreting these data and querying them … is a work in progress."

The difficulty seems to be that the major platforms have planned to collect data from the use of MOOCs, but did not consider in advance the analytical questions that universities might want to ask of the data. The data that are stored is that which is automatically stored for any Web site, not specifically for educational purposes. So the data are largely unstructured and may be stored in separate files, and a lot of additional work is needed in order to draw useful conclusions from it. Hollands and Terthali give the example of a researcher who is trying to analyze the user data from some edX MOOCs to find out what percentage of the learning materials had been read by each student, in order to inform the future design of the MOOC. This query is one which we might expect to be fairly commonplace, but in this case, it could not be answered without the researcher carrying out a great deal of bespoke, and time-consuming analysis of the data by designing and running their own computational queries using a programming language. The report estimates that it took 9 months of a computer scientist's time to create and conduct this and other standard queries that were needed to carry out standard data analysis of a single MOOC (Hollands & Terthali, 2014).

So although the MOOC platforms promise much for the data that are collected, the reality is likely to be very different. In the case studies that were carried out, universities reported planning to carry out their own additional analysis of the data that were returned to them by the MOOC platforms, but that this substantial undertaking had not yet taken place and was subject to investment of human resources into specific local information systems and their own sets of queries.

In summary, it is perhaps most accurate to say that the data collection and analysis tools are not yet sufficiently mature to be easily applied to standard queries about learner behavior in MOOCs, so institutions need to plan additional local manipulation of the data.

There are two other, specific issues to consider related to analytics and data storage and management.

- Data ownership, privacy, and rights issues

All data that are created by tracking the behavior of users of the MOOC materials is subject to complex data ownership, privacy and rights' issues. For those institutions that are working with a third-party MOOC platform rather than collecting and storing data locally, this is more complex, as they are likely to be subject to privacy laws and rights issues in more than one domain. We recommend that the university's legal team is consulted before learner data is collected or analyzed, and also before access to the data by third parties is permitted, and that full consideration is also given to the ethical issues of using learner data in ways with which participants may not be familiar (Neal, 2014).

- The limitations of "hard" data

Though we advise that part of the MOOC strategy should address the issue of data and analytics, it is important to bear in that there are limits to what can be concluded from the collection and interrogation of "hard" data. Particularly given the difficulties reported above in conducting even simple queries, it is important not to be drawn into a belief that data and analytic reports can provide all the answers to questions such as: how successful is the MOOC? Who is using it and what are they learning? What works well and what does not?

A strategy for the qualitative analysis of user perceptions needs to be put in place alongside the strategy for data and analytics, as part of the overall evaluation strategy, as previously discussed. As recent research has shown, for example, two similar populations of MOOC students, studying the same materials, can achieve very different results from the same assessments (Halawa, 2014). All "hard" data need to be complemented and informed by qualitative data, such as data from interviews, focus groups, and user diaries. This should be considered as part of the MOOC strategy.

References

Halawa, S., Greene, D., & Mitchell, J. (2014). *Dropout Prediction in MOOCs using Learner Activity Features*. Proceedings of the European MOOC Summit (EMOOCs 2014), Feb 10-12, 2014, Lausanne, Switzerland, Online at http://web.stanford.edu/~halawa/cgi-bin/files/emoocs2014.pdf.

Hollands, F. M., & Terthali, D. (2014). *MOOCs: Expectations and reality*. Educause, Online at http://www.educause.edu/library/resources/moocs-expectations-and-reality.

Kizilcec, R. F., Piech, C., & Schnieder, E. (2013). *Deconstructing disengagements: Analyzing learner subpopulations in massive open online courses*. ACM, Online at http://www.stanford.edu/~cpiech/bio/papers/deconstructingDisengagement.pdf.

Koller, N. (2014). The online revolution: education for everyone. In *Paper presented at the university college access lecture, 17 June 2014, at the T.S.Eliot lecture theatre*. University of Oxford.

Kraan, W., & Sherlock, D. (2013). Analytics tools and infrastructure *CETIS analytics series* (Vol. 1). No.11. CETIS. Online at http://publications.cetis.ac.uk/wp-content/uploads/2013/01/Analytics-Tools-and-Infrastructure-Vol-1-No11.pdf.

Neal, C. (2014). *MOOCs lead to massive data collection*. The Heartland Institute, Online at http://news.heartland.org/newspaper-article/2014/12/20/moocs-lead-massive-data-collection.

Pistilli, M. D., Arnold, K., & Bethune, M. (2012). *Signals: Using academic analytics to promote student success*. Educause Review Online, Online at http://www.educause.edu/ero/article/signals-using-academic-analytics-promote-student-success.

Powell, S., & MacNeill, S. (2012). Institutional readiness for analytics *CETIS analytics series* (Vol. 1). No.8. CETIS. Online at http://publications.cetis.ac.uk/wp-content/uploads/2012/12/Institutional-Readiness-for-Analytics-Vol1-No8.pdf.