JOURNAL OF LEARNING ANALYTICS

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

Big Data Characterization of Learner Behaviour in a Highly Technical MOOC Engineering Course

Kerrie A. Douglas
School of Engineering Education
Purdue University, USA
douglask@purdue.edu

Peter Bermel
School of Electrical and Computer Engineering
Purdue University, USA

Md Monzurul Alam
School of Electrical and Computer Engineering
Purdue University, USA

Krishna Madhavan
School of Engineering Education
Purdue University, USA

ABSTRACT: MOOCs attract a large number of learners with largely unknown diversity in terms of motivation, ability, and goals. To understand more about learners in highly technical engineering MOOCs, this study investigates patterns of learners' (*n* = 337) behaviour and performance in the Nanophotonic Modelling MOOC, offered through nanoHUB-U. The authors explored clusters of learner clickstream patterns using the k-means++ algorithm and found five clusters of learner behaviour, labelled according to learners' use of materials: *Fully Engaged*, *Consistent Viewers*, *One-Week Engaged*, *Two-Week Engaged*, and *Sporadic learners*. The Kruskal-Wallis nonparametric statistical test yielded a significant difference ($p < 0.01$) between learners' access of course materials in each cluster. The researchers then examined the participation and mean scores on course quizzes and exams for each learner group. One-Week Engaged learners, on average, scored significantly lower on the first week's assessment. Two-Week Engaged learners, on average, scored significantly lower on the second week's assessments. Other differences found in learners' participation and performance on quizzes and tests based on the five clusters are discussed. These findings suggest that some of the high dropout numbers in advanced MOOCs may be related to learners' performance on course assessments. In addition, integration of learner access to course material with course assessment scores provides a much richer understanding of learners in a MOOC.

Keywords: MOOCs, learning analytics, assessment

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

1 INTRODUCTION

Many education researchers strive to conduct research that will one day translate into widespread improved practice, informed by evidence. The implementation of massive open online courses, MOOCs, has occurred very differently, arguably in reverse fashion. The potential that MOOCs have to disseminate valuable knowledge and skills, without financial, location, or social-status barriers, is a very exciting promise, as noted in the media (e.g., Leckart, 2012). At no previous time in history has knowledge been so freely available. Yet, there remains limited information for institutions regarding the value of offering MOOCs (Hollands & Tirthali, 2014) and there is limited empirical information to inform pedagogy (Reich, 2015; Perna et al., 2014). While research related to online learning more broadly is quite vast, research on how learning can occur within the MOOC model is still relatively unexplored and many questions remain (Reich, 2015). Perna et al. (2014) note three areas of consensus in the emerging field: 1) MOOCs have very low completion rates, typically 5–12%; 2) how learners progress through MOOCs is not largely understood; and 3) the effects of individual course characteristics on learner outcomes is unknown.

With open access, thousands of learners may register with little or no commitment, entering and leaving the MOOC at will. This non-traditional behaviour in open access courses brings significant challenges to assessing what learning has indeed occurred. In addition, the application of basic instructional design principles is unclear, especially considering the diversity of learners in terms of educational background, usage, motivation, and intention for the courses (Douglas, Mihalec-Adkins, Hicks, Diefes-Dux, Bermel, & Madhavan, 2016). Given that good instructional design is based on understanding the learning environment (i.e., the context wherein the learning occurs), the learner, and the tasks to demonstrate learning (Ragan & Smith, 1999), it is imperative for researchers to understand learner behaviour more clearly in the MOOC environment, the types of learner needs, and how learners can demonstrate their learning.

To better understand behaviour and types of learners in MOOCs, we must create new methods of research that integrate analytics with traditional forms of assessment data. While clickstreams are not a measure of learning (Thille et al., 2014), learner access data can serve to identify groups of learners who utilize the materials differently. In addition, identification of usage patterns can serve as a grouping variable to allow deeper investigation into what are the underlying differences between learners. We consider this integration of learner analytics with forms of traditional educational research and assessment as a big data characterization of learner behaviour (Morabito, 2015) because the vast amount of data is used to reveal something of depth. The purpose of this research is to characterize learner patterns in a highly technical engineering MOOC.

The specific MOOC under study, Nanophotonic Modelling, was developed to provide graduate-level content related to the field of nanotechnology, which is constantly changing through research advancements (Roco, 2011), causing many traditional course textbooks to fall quickly out of date. In fact, many of the recent course topics are so new that the information is only available through conference proceedings and peer-reviewed journal publications. Developing courses such as the one on

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

nanophotonic modelling, nanoHUB-U, significantly decreases the time between research discovery and access to the information for engineering students and practitioners. As a first step to understanding the types of learners present within an advanced MOOC, we identify learner groups based on their course material usage. Next, using the emergent groups as a sorting variable, we explore each group's quiz and test scores to understand more about the similarity of learners within each group and aid in our classification of learners.

2 LITERATURE REVIEW

MOOCs tend to attract levels of enrollment that are quite different from traditional higher education learning environments. Anyone with Internet access can register for a course, whether it is out of passing curiosity or purely by accident (Liyanagunawardena, Parslow, & Williams, 2014). In a survey study of 34,779 MOOC learners, Christensen et al. (2013) found that the majority of learners were from developed countries, young, employed, and well educated. In addition, survey respondents were asked their reason for enrollment with 10 answer choices, including "other" and "none of the above." Overall, approximately 50% of respondents indicated that they enrolled for "curiosity, just for fun." Christensen and colleagues (2013) note that this result varied when considered by course. For example, 75% of respondents reported their reason for enrollment in humanities courses as "curiosity, just for fun." The lowest percentage of learners reporting enrollment out of curiosity was approximately 49% for the group of science, healthcare, and math courses. Thirty-nine percent from those courses report enrolling to gain specific skills related to job. However, they do not report findings specifically for science or engineering courses.

Kizilcec & Schneider (2015) studied learners' reported reasons for enrolling in MOOCs and found thirteen different reasons, varying from general interest to improving English. The researchers then turned the open-ended questions into a survey question, asking respondents to select "all that apply" on a different sample. In other words, respondents could pick as many reasons for enrolment as desired, without giving them a rank. The mean number of chosen categories was 6.3 out of the 13 reasons for enrollment. While it is clear that there are multiple reasons for signing up for a course, it is unknown what reasons are the most important or of value to the learners. In addition, reasons for enrollment were not examined in context of learner behaviour in the courses. It is unknown whether those few that enrol to increase their skills may seek deeper levels of learning or participate with course materials more fully.

2.1 Learner Completion

The openness of MOOCs presents additional challenges as learners may enter and leave a course without consequence. Learners are able to view materials based on topics of interest. It is unsurprisingly found that the actual percentage of learners who complete the course as per traditional definition of "completion" (i.e., successfully finish the majority of course activities and assessments), or even those who regularly engage in the course at all, can be very low. For example, Perna et al. (2014) studied learner usage in 16 MOOCs offered through Coursera, and found that fewer than half of registrants

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

(46%) accessed one or more video lectures. Liyanagunawardena, Adams, and Williams (2013) examined 45 MOOC-related research papers published between 2008 and 2012, and noted that the highest rate of successful completion for any course was 19.2%, but the vast majority reported less than 10%. In addition, Jordan (2014) considered completion as the percentage of course registrants who met criteria to earn a course certificate in 39 MOOCs offered on different platforms (e.g., edX, Coursera, Udacity), and found a typical completion rate to be 5%. Little is known about the differences between learners who leave the course after only a couple of weeks versus those who are more sporadic or those who complete all the aspects.

While consensus has emerged that completion rates are not fully appropriate in the MOOC environment (e.g., Kizilcec, Piech, & Schneider, 2013; DeBoer, Ho, Stump, & Breslow, 2014; Liyanagunawardena et al., 2014), few have suggested what metrics should be used to make evaluative conclusions. Indeed, given the lack of personal investment needed to sign up for a course, researchers and evaluators are not able to infer the quality of a MOOC based on completion rate alone. The openness and accessibility to MOOCs has even led some to argue that completion rates are "misleading and counterproductive indicators" of the quality and potential of MOOCs (Ho et al., 2015). DeBoer et al. (2014) elaborate on this point, asserting that completion rates must be re-conceptualized for a MOOC environment where learners may consider MOOCs less of a traditional course than a collection of learning resources. In that sense, access to MOOC materials becomes the desired goal rather than the course itself. Others point out that MOOCs did not create a need to redefine successful completion; rather they raise a longstanding issue in higher education: learners have different goals (Liyanagunawardena et al., 2014). Tinto (1975) points to the argument that it is inappropriate to categorize all learners who withdraw from a course as one group without distinguishing between academic failures and voluntary withdrawals. When Tinto's conceptualization is applied to MOOCs, it is important to understand the reasons behind disengagement and other non-traditional use patterns.

The level of importance to place on completion rates is partly dependent upon the intended design of the MOOC (Yuan & Powell, 2013). Indeed, MOOC completion rates become meaningful only when interpreted in the context of learner and stakeholder goals (Koller, Ng, Do, & Chen, 2013). Retention and successful completion, as evaluative metrics, also become important in contexts where MOOC stakeholders (e.g., faculty who teach, university administrators) desire learners to use the course materials more traditionally. Indeed, the MOOC phenomenon poses new challenges for educational researchers to determine what metrics of evaluation are appropriate. In order to begin to answer meaningful questions about the value of MOOCs, these challenges must be addressed.

2.2 Learner Usage

Reich (2015) points to four recent studies where learner activity in a course was positively correlated with higher final outcomes, such as course completion and exam scores (Collins, 2013; Murphy, Gallagher, Krumm, Mislevy, & Hafter, 2014; Reich et al., 2014; Wilkowski, Deutsch, & Russell, 2014). Given the discrepancy between enrollment and course activity, behaviour in MOOCs seems to indicate that many learners are taking advantage of MOOCs in the same way they use other online resources.

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

MOOC learners generally enter the course with a wide variety of participation objectives (Kizilcec & Schneider, 2015) making them far too heterogeneous to fully characterize simply by the percentage using course resources. These latent or hidden variables contribute to learner behaviour in the course and must be explored (Thille et al., 2014). Simply put, descriptive statistics present an incomplete picture of the true variance of learner behaviour.

Some researchers have considered the types of learners based on usage patterns as a method for understanding how learner needs can be met better (Kizilcec, Piech, & Schneider, 2013). Studying learner patterns in three computer science courses of varying difficulty, Kizilcec et al. (2013) propose four profiles: 1) completing 2) auditing, 3) disengaging, and 4) sampling. *Completing* is described as learners who complete most aspects of the course, including the assessments. This group was most similar to traditional classroom learners and varied in their performance on assignments. *Auditing* is described as the group of learners who followed along with the videos throughout the course duration, but who did assessments infrequently. *Disengaging* is described as the group of learners who participated in most aspects of the course initially, but then disengaged at some point in the first third of the course offering. *Sampling* is described as the group of learners who watched videos or other materials for only one or two assessment periods, many of whom only watched one video.

Similarly, Hill (2013a) identified four patterns of learners within Coursera MOOCs: 1) *Lurkers* who only sample items, or do nothing at all beyond registration, 2) *Drop-Ins* who become involved for a selected topic, 3) *Passive Participants* who "view a course as content and expect to be taught," tend to watch videos and maybe take quizzes, but do not usually participate in activities/discussion, and 4) *Active Participants* who participate in most aspects of the MOOC, including social media. Hill (2013b) then revised his finding to delineate between those who only register (*No-Shows*) and those who read content, but do not participate in discussions (*Observers*). A limitation of both Kizilcec et al. (2013) and Hill (2013b) is that neither fully describes the methods used to determine the ideal number of learner groups nor tests the significant difference between groups. The role of theory or rationale in developing a generalizable typology of learner usage is also not fully explicit and it is possible to generate several different visualizations of access that may or may not be helpful for MOOC instructional design.

To further research MOOC education and evaluation, we must explore whether the typology of learner patterns found in Kizilcec et al. (2013) or Hill (2013b) generalize to MOOCs regardless of content, or whether types of learner patterns vary based on course context. Furthermore, to understand the similar learners within each group, we must contextualize groups in terms of their performance on course assessments. This study identifies groups of learner behaviour patterns and delves deeper into their quiz and test assessment performance. Specifically, we ask the following research questions: 1) What are the clusters of learner usage patterns in a highly advanced MOOC engineering course? 2) How are participation and performance patterns similar within each cluster? 3) What are the differences in quiz and exam participation and performance between clusters? The specific course studied in this paper was entitled Nanophotonic Modelling, a highly technical, advanced engineering course provided through nanoHUB-U.

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

3 BACKGROUND

Unlike MOOCs offered on large platforms that do extensive marketing to drive high enrollment, a highly technical (graduate level) engineering course offered through an NSF-sponsored computational nanotechnology site provides the opportunity to explore meaningful results available from combining learner usage and assessment performance data. As Nanophotonic Modelling is highly specialized and not offered through a major MOOC platform, enrollment was substantially lower than typical MOOCs. Yet, compared to similar on-campus graduate courses that typically have 10 students, the enrollment (*n* = 337) was well beyond what would be plausible at one geographic location. The course was designed for a relatively narrow intended learner type, i.e., someone with a graduate-level understanding of physics and computational simulation. Ideally, the instructor would target engineering and science professionals who need nanotechnology skills and knowledge. By developing courses such as Nanophotonic Modelling, one of the major intended outcomes of nanoHUB-U is to significantly decrease the time between research discovery, to access of new information, to project applications for engineering students and practitioners. The sequence begins with the most basic concepts in order to scaffold more difficult methods later.

Table 1. Structure of Nanophotonic Modelling course

Week(s)	Major topic(s)
1	Photonic bandstructure solvers
2	Transfer matrix analysis; rigorous coupled wave analysis (RCWA)
3–4	Finite-difference time domain
5	Finite-element methods

Much like other nanoHUB-U offerings, the Nanophotonic Modelling course contains five lectures per week of approximately 20 minutes each, released weekly over the course of five weeks. This instructor-led section of the course is intended for learners to approach sequentially, as in a traditional course. During this time, a graduate teaching assistant and the professor regularly interact with their students through a message board. The instructor-led parts of the course conclude after the fifth week. Subsequently, the course material is archived and made freely available through a self-paced section of the course with no fixed end date. In addition to the lectures, five types of course materials are given for each lecture: 1) a set of lecture slides in PDF format, 2) a quiz, 3) an assignment, 4) a pdf assignment solution, and 5) a tutorial to provide an overview of the solution technique. Finally, a weekly exam takes place at the end of each of five weeks. Students who achieve an average of over 60% in all the graded

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

material by the end of the last week, including material not taken, receive a certificate of completion, known as a badge. In some cases, learners can also receive University or continuing education credit(s).

4 METHODS

4.1 Participants

In this section, we present the learner demographics in terms of their geographic locations, gender, and organizational affiliation. Often, MOOC researchers report the number of learners in the categories of registrants and starters (e.g., Perna et al., 2014). *Registrants* are the total unique individuals enrolled, while S*tarters* are the total unique individuals who actually access materials in the first week of the course. Complicating matters, however, learners may choose to join after the course has started. To account for this issue, we also create a third category of *Active Learners*, defined as the total unique individuals who participated in one or more activities at any point during the course, i.e., downloading or watching one or more course material item at least once during the time it was originally offered. Nanophotonic Modelling had 337 registrants from 48 countries, with 68% of participants from 45 countries. There were 226 *Active* Learners who utilized materials as the course progressed. Although the learners were from across the globe, some geographical clustering was observed in Anglophone countries: notably, the largest group of registrants (31%) were from the United States, with the second highest (17%) from India. Nearly the same percentage of *Active Learners* in the course came from the United States (31%) and India (16%). No other single country accounted for more than 5% of the learners. As shown in Table 2, 29% of the learners are from 34 countries, each country with only a few participants. Thirty-one learners earned a badge for the instructor-led portion of Nanophotonic Modelling.

Table 2. Geographic Distribution of Active Learners

Country	% Active
United States	31.4
India	15.9
Egypt	4.4
Canada	4.0
Mexico	2.7
Russian Federation	2.7
Brazil	2.2
Hong Kong	2.2
Bangladesh	1.8
China	1.8
Spain	1.8
Other (All < 2%)	29.2

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

4.2 Sources of Data

We collected learner behaviour and assessment data to develop a characterization of the learners from the Nanophotonic Modelling course. The learner behaviour is derived from typical clickstream data: number of lecture views, access to PowerPoint slides, tutorials, assignments, and assignment solutions. Clickstream data were recorded every time a learner clicks on any course resource. In this study, we identified 31 course materials made available every week, totalling 155 course materials during the five weeks of the course. Two types of course materials were made available to the learners every week: learning materials like lectures, homework, and tutorials, as well as assessment materials like quizzes and exams. Clickstream data captures the date and time when the learner accesses any material, but comes with some practical limitations. For example, learners may download any course material in PDF or video format or follow a YouTube link, but the clickstream data does not capture whether the learner saves the material or takes notes and then refers back to it.

All data were de-identified using an algorithm that assigned a random new unique code for each learner. The newly created code was used to create a unique 1:1 map between multiple data sets while greatly lowering risk of re-identification. Also, the actual calendar dates and access duration values were converted to ordinal numbers to protect the privacy of the learners. Still, this approach can capture the multiple access events to particular course materials by some learners, which may be useful in learner characterization.

Exam and quiz scores were collected for analysis. There was one quiz per lecture and it consisted of two questions, directly related to material covered in the lecture (multiple-choice, with four to five choices per item). In total, there were 25 quizzes and 50 questions to assess learners directly after presentation of the material. In addition, there were weekly exams, each with 10 questions aligned to the concepts covered in that week. In total, the five exams comprised 50 questions.

4.3 Identifying Learner Groups by Clustering

The research team analyzed the clickstream data by looking for patterns of course material access, which is the basis for identifying learner groups. The rationale for this approach is simple: by exploring the usage patterns and understanding more about learners within and between each pattern of behaviour, we can understand more about the diversity of learners' participation and achievement. It is important to understand that all the learners may not have the same learning objective, which may give rise to different patterns of access. Investigating the relationship between learners and course materials can give better insight into the learners (Hecking, Ziebarth, & Hoppe, 2014). Relationships between learners based on identifying similarities in course material usage patterns can help reveal commonalities of each group. Some researchers analyzed the learner groups using social network analysis methods to find groups with common interests (Harrer, Malzahn, Zeini, & Hoppe, 2007). Other researchers presented data mining clustering methods for web-based course material usage (Romero, Gutiérrez, Freire, & Ventura, 2008). While many clustering algorithms can be used, we used a k-means

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

clustering technique. The popularity of k-means clustering techniques rests on the ease of implementation, simplicity, efficiency, and empirical success (Jain, 2010). This approach helped to identify key patterns of course material usage among the learners, and then divide them into a corresponding set of learner groups.

4.3.1 Data Pre-Processing

Prior to applying the k-means++ (explained in next section) algorithm to our data, we identified the proper data sets suitable for the k-means++ algorithm. For instance, among the 155 course materials, lectures and tutorials had different video formats. More precisely, identical videos were provided in three formats: online streaming, downloadable MP4, and YouTube. The learner can get the same study material from any of those three video modes; therefore, accessing all three formats or just one does not reflect different learning intentions. Therefore, we combined all three formats of the same video into a single item. This approach results in 95 distinct course materials. Next, each course item is labelled with a unique ID number, according to the sequence of release. We generated a binary table with 226 rows, representing the material usage pattern of each active learner. The binary sequence of an individual had a numeric digit "1" if the person accessed the corresponding material and a "0" if s/he did not access it.

Next, we generated a dot plot from the binary table, as shown in Figure 1. Every dot represents a learner accessing the corresponding course material. The horizontal axis represents the material ID and the vertical axis represents the learner ID. It appears upon casual inspection that the material usage pattern is just random. Nevertheless, clustering algorithms can reveal unique patterns of content usage.

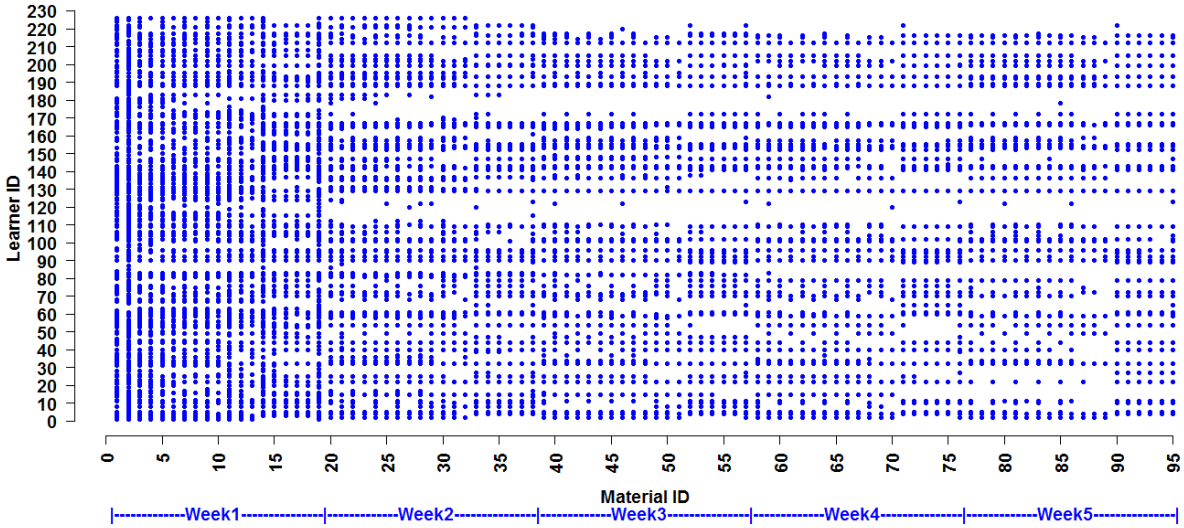


Figure 1: Learners Course Material Access Data

4.3.2 K-Means Clustering Algorithm

The k-means clustering algorithm is a well-established data mining research method for classification of data into distinct groups. As discussed in Section 4.3.1, our initial data is derived from the access

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

patterns of 226 active learners, captured as a 95-dimensional binary vector X. We select k cluster centres μ from X to minimize the sum of distances in the 95-dimensional space using the k-means algorithm (Jain, 2010). Intuitively, this corresponds to choosing group membership to maximize uniformity within each group, even if the overall data set is very heterogeneous. However, there are many examples of generating low-quality, non-reproducible clusters when using arbitrarily chosen initial cluster centres in k-means (Arthur & Vassilvitskii, 2007). Therefore, in this work, we follow the k-means++ heuristic method to reproducibly classify our learners (Arthur & Vassilvitskii, 2007). The probabilistic approach of choosing the initial cluster centres by the k-means++ algorithm improves the quality and reproducibility of the found cluster (Arthur & Vassilvitskii, 2007). The key steps are 1) to randomly select an initial cluster centre from *X*; 2) to weight the selection of subsequent cluster centres by their relative distances to other points in *X*; 3) to assign all points to the closest cluster centre; 4) to re-centre each cluster so formed; and 5) to repeat the last two steps until no further changes are seen.

We implemented the method outlined above using the function "kmeans," available in MATLAB. The reproducibility of the clusters by k-means++ algorithm depends on the initialization (Lisboa, Etchells, Jarman, & Chambers, 2013). Therefore, k-means++ clustering can be treated as a heuristic approach to divide the dataset into different groups. Nonetheless, the choice of groups was fairly uniform in the datasets examined in the next section.

5 RESULTS

In this section, we demonstrate how the clustering method shows different learner groups among the students. We also discuss the similarities and differences between the groups in terms of material usage pattern, as well as their performances in quizzes and exams. We explain what the clusters tell us about the learners and what the consequences are.

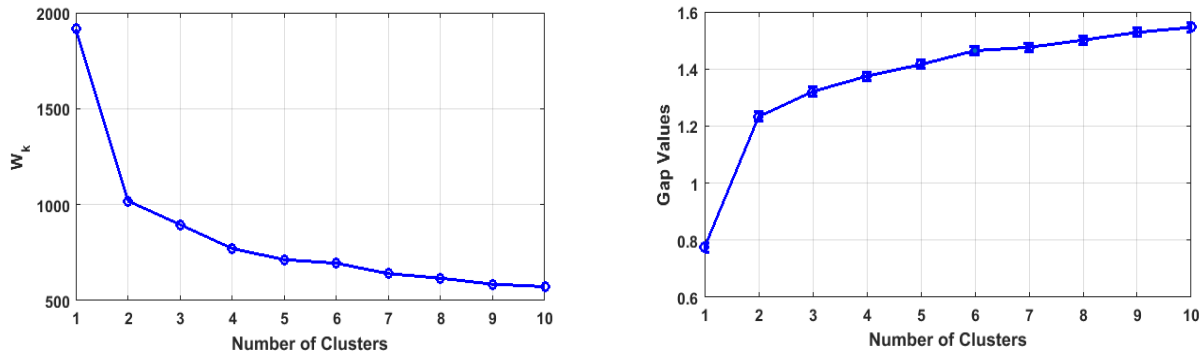
5.1 Identified Learner Groups

The k-means++ clustering technique found some interesting learner groups within the Nanophotonic Modelling course. Course material usage patterns by different clustered groups are presented in Figure 3. Every dot in Figure 3 indicates that the learner accessed that particular material; conversely a blank represents that the learner did not access that material. Note that only learner data is analyzed; registered but inactive learners are excluded from our analysis.

A common concern of clustering is the optimal number of clusters. We identified the optimal number both empirically and through rationale. The two empirical methods used were within-cluster error dispersion and gap statistics, as in Tibshirani, Walther, and Hastie (2001). Within-cluster error dispersion (W_k) is presented against the number of clusters in Figure 2a and the gap statistic is shown in Figure 2b. It is obvious in Figure 2a that the W_k value started to become flat at $k = 5$ for the first time; therefore, the optimal number of clusters is 5. Gap statistic evaluation was completed using the function "evalclusters," available in MATLAB. A plot of gap values against the number of clusters is shown in Figure 2b. The requirement for optimal number of cluster is $Gap(k) \geq Gap(k+1) - s_{k+1}$ where k is

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

the optimal number of clusters and $s_{k+1}$is the standard error of the clustering solution used in the Gap statistic method proposed in Tibshirani et al. (2001). Finally, the method identified the optimal number of cluster as 6.



(a) (b)

Figure 2: (a) Within-Cluster Error Dispersion; (b) Gap Statistic

The learners are clustered by the k-means ++ method with clusters numbering 6 and then 5. Although Gap statistic suggested that the number of clusters be six, few learners were actually placed in the sixth cluster, and they were very similar to learners in other clusters; i.e., the learners of cluster 6 were engaged for the first two weeks with some irregular use of course materials. Based on this rationale, we decided to cluster the learners into five groups in order to describe them according to their material usage pattern, rather than maintaining the sixth cluster group of two-week usage but not fully engaged

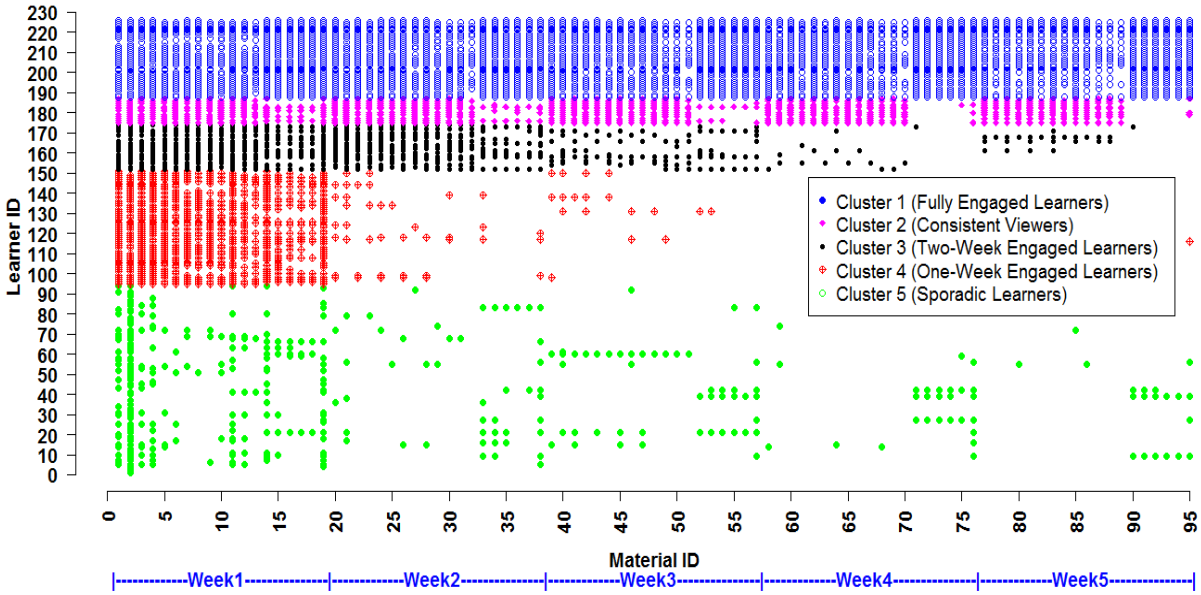


Figure 3: Learner Clustering by Course Materials Usage Pattern

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

The simulation for k-means++ clustering was performed by changing various numbers of clusters and settling finally on five distinct groups (clusters) of learners, as presented in Figure 3, with the x-axis representing unique course materials in chronological order, and the y-axis representing the new cluster-based ordering of each unique learner. Clusters 1 through 5 are coloured blue, magenta, black, red, and green, respectively. This clustering graph shows the patterns of activities of the learners much more clearly than in Figure 1. According to our reduced course material list, 19 items were released every week. The first 13 are study materials, while the last six are assessments (five quizzes and an exam), as indicated on the horizontal axis in Figure 3. The clustering graph highlights key patterns of the material usage present in each group of learners.

Based on the material usage pattern, we identified the five clusters and assigned labels to the found clusters, as descriptive of their usage to aid in interpretation: 1) *Fully Engaged Learners,* 2) *Consistent Viewers,* 3) *Two-Week Engaged Learners,* 4) *One-Week Engaged Learners,* and 5) *Sporadic Learners* respectively. *Fully Engaged Learners* access the study materials regularly, and attempt most quizzes and exams. *Consistent Viewers* actively access content over the entire course, but do not regularly access the course assessment materials. *Two-Week Engaged Learners* actively access materials in the first two weeks of the course, with a sharp reduction in their activities in subsequent weeks. *One-Week Engaged Learners* are quite interesting; they fully access the first week's materials, but do very little in the course afterwards. Finally, *Sporadic Learners* randomly access materials. We identified two learners from this group who attempted all five exams without accessing any of the study materials. Many of the *Sporadic Learners* accessed only assignments and none of the videos. The cluster-wise distribution of learners is presented in Table 3. *Sporadic Learners* form the biggest group among the five, accounting for 42%.

Table 3. Cluster Groups

Cluster	*n*	% of Learners
Fully Engaged Learners	39	17%
Consistent Viewers	13	6%
Two-Week Engaged Learners	23	10%
One-Week Engaged Learners	57	25%
Sporadic Learners	94	42%

5.2 Exam and Quiz Grades for Each Group

To better capture the similarities and differences within and between groups, we also examined the exam and quiz scores for each group. We calculated mean and standard deviation of the scores within each group to more thoroughly understand whether learners who behaved similarly within the course also performed similarly.

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

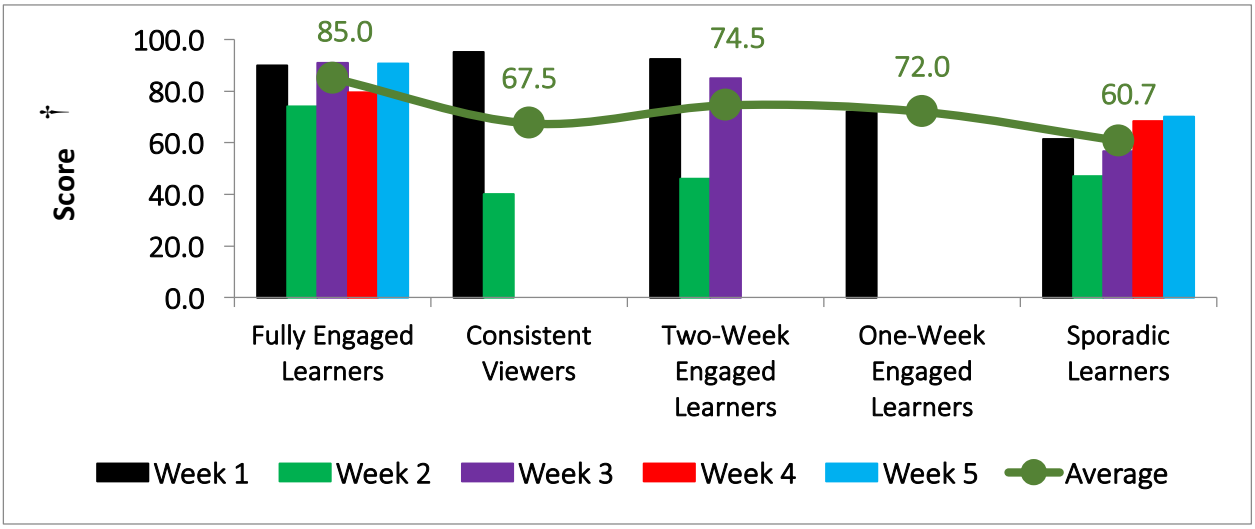


Figure 4: Scores Obtained by Different Groups in Weekly Exams. † Average among students who attempted the exam

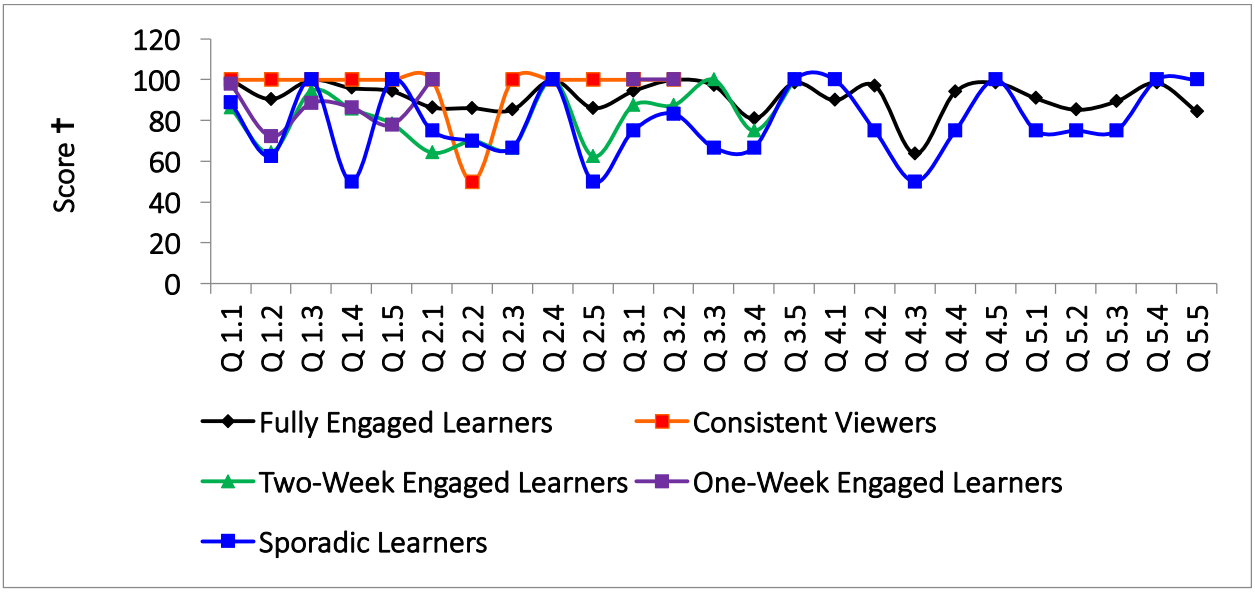


Figure 5: Scores Obtained by Different Groups in Quizzes

Figure 4 presents the group-wise performance in the weekly exams and quizzes, respectively. The mean scores were calculated by computing the average of scores for all learners who attempted the given assessment in each group. As shown, the first three learner clusters (*Fully Engaged, Consistent Viewers*, and *Two-Week Engaged*) scored 90 or above in the first exam, while the other two less-consistent groups performed poorly. *Fully Engaged* and *Consistent Viewers* scored above 90 consistently in the first week's quizzes, while the other groups were not consistent at all (Figure 5). All scores decreased in each learner group on the second exam (with the exception of *One-Week Engaged*, as they did not attempt any more exams after the first week).

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

We also considered learners' assessment-related behaviour over the duration of the course. We found that 90% of those *Fully Engaged* attempted all the exams and quizzes. By contrast, only about 15% of *Consistent Viewers* attempted first week's exam and quiz, and none attempted any thereafter. In the *Two-Week Engaged* group, only 35% attempted the first week's exam; participation gradually declined over each week thereafter. In the case of *One-Week Engaged*, 44% attempted a first week quiz, and fewer than 9% took the first exam. They did not attempt any other quizzes or exams after the first week. Less than 10% of the *Sporadic* group attempted all the exams and quizzes.

We now consider the variance of performance by compiling detailed statistics on the first week assessments for each group of learners, provided in Table 4. This work shows that *One-Week Engaged* and *Sporadic Learners* had a high standard deviation of scores in the first exam compared to other groups while *Two-Week Engaged*, *Consistent Viewers*, and *Fully Engaged Learners* scored in the first week's quizzes with high standard deviations, indicating the variance of performance of the learners in these groups was high. Overall, the learners within these three groups performed similarly to each other in the first week's exam compared to other groups. *Two-Week Engaged*, *One-Week Engaged, and Sporadic* learners performed similarly in the first week quizzes, in the B-range; *Fully Engaged* and *Consistent Viewers* earned very high percentages.

Table 4: Group-wise Mean and Standard Deviation (SD) of Assessment Scores

Assessment	Group Name	Number of Learners Attempted	% of Learners within Cluster	Mean (SD)
First Week Exam	Fully Engaged	37	95	90.0 (11.0)
	Consistent Viewers	2	15	95.0 (5.0)
	Two-Week Engaged	8	35	92.5 (9.7)
	One-Week Engaged	5	9	72.0 (20.4)
	Sporadic	7	7	61.4 (31.8)
First Week Quizzes	Fully Engaged	37	95	96.2 (13.3)
	Consistent Viewers	4	31	100.0 (0.0)
	Two-Week Engaged	11	48	82.9 (23.7)
	One-Week Engaged	25	44	86.0 (22.4)
	Sporadic	9	10	82.4 (23.0)

5.3 Statistical Significance of Clustering

Given that learner groups identified by the clustering technique performed differently in the course, there is some evidence of underlying learner characteristics (i.e., latent traits) that contribute to learner behaviour in the course. To check for the significance of our clustering solution, we performed non-parametric Kruskal-Wallis (1952) and Mann-Whitney U (Kirk, 2008) tests on the course material access pattern of each cluster. The usual technique for such measurements is the analysis of variance;

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

however, with uneven usage patterns of groups, the data did not meet assumptions of normality or similar group size. Therefore, we chose the non-parametric Kruskal-Wallis test (1952), which tests the hypothesis that the clusters have a significantly different distribution in the population (Stark, Woods, Thilaka, & Kumar, 2012). To allow for comparison between the clusters in a 95-dimensional space, the Kruskal-Wallis test was performed by summing each participants' access to all course materials. It was performed on all the clusters in a single calculation (i.e., rank ordering all points, calculating the sum of ranks for each cluster, and performing a χ^2 test with df=5–1=4, since we have five clusters. The results were significant, $H(4) = 191.88$; $p<0.01$, indicating that the different clusters have a less than 1% likelihood of being drawn from the same underlying distribution.

The Mann-Whitney U test is useful for testing the hypothesis that two groups are identical or not (Kirk, 2008). We performed the Mann-Whitney U test as post hoc test for the Kruskal-Wallis test as a follow-up to the finding that each cluster represents different learner types. Many researchers have applied this technique for statistical hypothesis testing. For example, researchers applied the Mann-Whitney U test to establish the impact of the neurotransmitter oxytocin on trust between humans (Kosfeld, Heinrichs, Zak, Fischbacher, & Fehr, 2005). In our study, we had 10 options for comparison. Finally, we calculated the effect sizes of all the comparison cases. The effect size was defined as $r = Z/\sqrt{N}$ where Z was the z-score found from the test and *N* was the sample size. It is important to note that we set the significance level for the test to 0.05, but the critical level of significance should be 0.05/10 = 0.005 because we had 10 cases. The significance levels ($p < 0.001$) of all the comparisons listed in Table 3 are well below the critical significance level, again implying that the clusters are very unlikely to be drawn from identical distributions. In addition, all the test cases had large effect sizes ($r < -0.5$), except for a medium effect size in comparing Clusters 1 and 2, which is consistent with our earlier findings. Taken together, these statistical results constitute significant evidence that clustering can provide new insight into the behaviour of distinct learner populations.

Table 3: Effect Size and Significance Level from the Mann-Whitney Test

Comparison	Effect Size (r)
Fully Engaged/Consistent Viewers	-0.49
Fully Engaged/Two-Week Engaged	-0.83
Fully Engaged/One-Week Engaged	-0.85
Fully Engaged/Sporadic	-0.80
Consistent Viewers/Two-Week Engaged	-0.77
Consistent Viewers/One-Week Engaged	-0.67
Consistent Viewers/Sporadic	-0.58
Two-Week Engaged/One-Week Engaged	-0.75
Two-Week Engaged/Sporadic	-0.69
One-Week Engaged/Sporadic Learners	-0.74

Note: $p < 0.001$ for all comparisons

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

6 DISCUSSION

From the results of the k-means++ clustering technique, we identified five learner groups according to their course usage patterns: *Fully Engaged, Consistent Viewers*, *One-Week Engaged*, *Two-Week Engaged,* and *Sporadic*. The results from our analysis are similar, but somewhat different from the patterns found by Kizilcec et al. (2013) and Hill (2013b). The *Fully Engaged* group is similar in usage to Kizilcec et al.'s *Completing group*. We chose to label the group of learners who participated in most course activities as *Fully Engaged* rather than *Completing* for two reasons: 1) the term "completion" has become somewhat controversial in the MOOC literature, and 2) learners in this group not only finished the course, they faithfully engaged in most or all aspects. The *Completing* group varied by course level, contributing to 27% of the learners enrolled in the high-school level course; 8% of the undergraduate-course level, and 5% of the graduate level course. Our Fully Engaged group made up 17% of learners in Nanophotonic Modelling. Of note, we identified groups of learners who fully engaged with the course early on and withdrew activity after one or two weeks. This is an important finding, especially in light of their performance on assessments prior to disengaging with the course.

Next, we examined how learners in differing behaviour groups participated in and performed on the course assessments. Together, this data allowed us a closer look at the varying learner groups within Nanophotonic Modelling and developed a characterization of the types of learners present. The major distinction between the groups was their usage of study and assessment materials, with remarkable uniformity within each group. Next, we summarize the characteristics of each learner group.

Fully Engaged learners represented approximately 17% of learners within Nanophotonic Modelling, somewhat higher than the typical completion rate (less than 10%) reported in the literature (Liyanagunawardena et al., 2013). It may be that because Nanophotonic Modelling is a highly technical course offered on nanoHUB.org, the learners attracted to the course may be somewhat different than those who register in a MOOC offered through a large platform, such as Coursera or edX. *Fully Engaged* learners earned completion certificates and consequentially are the most similar to traditional classroom learners. They tended to participate in most or all aspects of the course, including the weekly quizzes and exams. Ultimately, as a group, they performed very well, with an average final grade of 90. Unfortunately, many of the techniques commonly used to analyze course behaviour simply do not tell the story of those who wanted to complete all aspects of the course. More about this group of MOOC learners must be understood. These *Fully Engaged* learners get lost in the means and averages of the majority of learners in MOOCs, yet there is value in understanding their experiences. Within these smaller percentages of *Fully Engaged* learners, having access to the content may have the biggest impact, such as learners who intend to apply the information in their work projects or change dissertation topics based on this new understanding. While the majority of learners certainly did not utilize the course in a strict course format, some did and indeed performed quite well.

Consistent Viewers represented approximately 6% of learners within the course. Their usage patterns indicate that they had different goals than the Fully Engaged learners. *Consistent Viewers* accessed most

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

or all of the study materials, but did not attempt the majority of the quizzes and exams. Interestingly, in the first week, the 15% of the group who took the assessments performed very well on the exam (*M* = 95.0, *SD = 5.0*) and all the quizzes (*M = 100, SD = 0)*. Nevertheless, it is unknown why they chose to discontinue taking the assessments. One potential explanation is that they were interested in gaining the knowledge, but were not concerned about earning a certificate or achieving a high grade. Without assessment scores, there is no way to make inferences about the level of knowledge this group achieved. At the surface level, it is apparent that there was something about the learning opportunity that brought them back into the course each week. One potential explanation may be that this group included faculty, students, or working professional engineers who did not have time to devote to the assessments, but wanted to become familiar with the content. Along the same lines, it is unknown how or if the *Consistent Viewers* assessed their own learning or the depth at which they intended to learn the material. Interestingly, they faithfully followed along, which seems to indicate a high level of motivation and intention despite not using the assessment opportunities for feedback. Their behaviour brings up the undefined role of assessment in MOOCs for those not seeking certificates of completion. Tests and quizzes are often thought of as means to passing a formal class, yet in an open-online environment, assessments can be taken multiple times as a feedback mechanism. Rather than viewing assessment as a burden, assessments can help inform what areas may need extra attention to understand. The value these learners placed on the learning experience and how they differ from those in the *Fully Engaged* group are considerations for further research. Additionally, it is unknown whether this group of learners would return to reference their course materials for application or how they benefitted from the learning experience.

Two-Week Engaged learners represented approximately 10% of the learners. They were engaged in the course very actively during the first two weeks. As shown in Figures 5 and 6, their performance in the second exam and the second week's quizzes was very poor (*M = 46.0 and 71.0, SD = 19.5 and 24.7 respectively*). Only 8.7% of these learners subsequently attempted the third week's exam and quizzes. Considering their low scores, one possible explanation for their discontinued engagement with the course is perhaps related to the course being too difficult for them. It is unknown whether their challenges were related to course characteristics (such as pedagogy) or if they were not prepared for the content covered (i.e., did not have the necessary prerequisite knowledge to be successful).

One-Week Engaged learners made up approximately 25% of the learners. This is the second largest group among the learners. They attempted the first exam and the first week's quizzes but performed considerably lower on the exam (*M = 72.0, SD = 20.4*) than the first three clusters. These learners dropped the course immediately after the first week. They did not even access the study materials, even though they were all available online and could be downloaded for later access. One possible explanation may be that they recognized early on that they did not have the background needed for the course or the course was different from their expectations, so they chose to disengage.

The largest cluster was *Sporadic Learners*, comprising 42% of the total course learners. Their behaviours would appear random to an outside observer. They did not access any type of course materials regularly

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

at all. A few learners (7% of the cluster who took the first exam) attempted the first exam and the first week's quizzes without accessing the study materials and thus performed poorly (*M = 61.4* and *82.4, SD = 31.8* and *23.9,* respectively). The majority of *Sporadic Learners* were accessing the materials with no defined pattern. Some accessed several resources around one topic, others only accessed assignments without videos, while others took a few exams. One possible reason for this behaviour is that they were interested in specific topics rather than in learning all the material and earning a certificate. Considering that *Sporadic Learners* make up the largest group of learners, there is a need to better understand what appears as sporadic. It is likely that learners within this group have different interests, motivations, and intentions. It is possible that what appears as sporadic is actually planned or deliberate. There is a clear need to contextualize the clickstream and assessment data with a deeper understanding of the latent groups of learners. In order to make meaningful inferences about differences in behaviour, it is essential that future offerings of the course include pre-course surveys to capture self-reported learner interests and planned behaviour.

7 LIMITATIONS

This study was conducted post-hoc on event log and assessment data collected through the Nanophotonic Modelling instructor-led course offering. Learners expressed motivations and intentions in the course were not included in this study. As a first step to a bigger research agenda of learner-centred approaches to MOOC research, the current study finds groups of learners based on their patterns of usage and makes meaning of their learning outcomes in the course assessments. Furthermore, it is unknown how similar the pattern of Nanophotonic Modelling learner behaviour is to more general topic MOOCs, or other highly advanced courses that resemble MOOCs, retaining most but not all of their characteristics. In particular, we chose to represent the learner behaviour in five clusters, where a group of learners who engaged with most materials, but not all in the two weeks were combined with a group that accessed all materials in the first two weeks. In larger datasets, this group may become more pronounced. The k-means++ algorithm is not necessarily the best algorithm in all circumstances and should be compared with other approaches in future work (such as DBscan). Nonetheless, we find that within-cluster error dispersion drops by 63% when using five clusters, as shown in Figure 2a, which suggests that it can be a useful analytical approach. Another limitation of this study is treating the access of videos and assignments as a binary variable rather than capturing additional information associated with multiple views. Although it is possible to capture whether students click on any course element more than once while online, it is beyond the scope of this data set to capture whether and how often any downloaded material is used by the student while offline. Because of this constraint, multiple time access is not modelled. Furthermore, while we used k-means++ on the access patterns of videos and assignments, additional data collected may be worth adding to the investigation to see if the existing patterns persist or are modified in any significant fashion.

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

8 CONCLUSIONS

Nanophotonic Modelling is a very technical MOOC designed for learners of a specific background. Even within this highly targeted learner population, our results indicate five distinct groups of learners who utilized the course in different ways to meet their own learning goals and needs. By analyzing the patterns of access and performance for each group to the extent data was available, we were able to obtain key insights into their behaviour. For example, by traditional course standards, achievement of less than 60% of possible points results in a failing grade, and is considered unsuccessful. Our findings provide evidence that there are groups of MOOC learners clearly unconcerned with grade outcomes, but find enough value in the course materials to return intermittently. While certainly some learners performed poorly on an exam and then withdrew, not all of those who discontinued activity performed poorly. Therefore, it would be a mistake to infer that all who disengage from the course do so because they are unsuccessful in learning the content.

The openness of MOOCs provides some challenge to how assessment can be used. While one group of learners, *Fully Engaged*, completed all assessments, the majority of learners did not. This may be related to those who sought certificates versus those taking the course more informally. Some learners may not understand the formative role that assessment can play in learning. While *Fully Engaged* learners could have used the assessments as a way to self-monitor their own learning, learners who did not attempt the quizzes or exams did not seek the opportunity for external feedback of their learning. Instructors may point out the learning value from the assessments. Furthering this idea of would be to clearly identify assessments as formative (used to enhance or support learning) or summative (used to assess competency). As MOOC developers consider the role of assessment for non-certificate and certificate earners, two paths may be developed: one where assessment is a feedback mechanism where learners can take the same quiz or exam multiple times and another where additional summative assessments are used to verify the issuance of a certificate. In this way, the integrity of the certificate can be maintained, while still allowing learners to have access to the correct answers to the quizzes and tests.

The similarity of learners within each cluster both in terms of their course material usage and performance on assessments is confirmatory that there are hidden or latent learner traits and characteristics. These distinctive group characteristics and performance clearly demonstrate that simply reporting the overall course means in terms of completion and pass rate truly fails to account for the diversity of learners within the course. Furthermore, more research is needed to understand the latent variables that account for what is seen in groups of learner behaviour.

In an open-access environment, learners have a wide variety of intentions, and it is inaccurate to assume that all learners are committed to the course or that none of them are. There is simply too much heterogeneity to classify all MOOC learners in terms of their mean usage or performance. Furthermore, learners may not complete all aspects of the course and yet still gain valuable knowledge. Some learners may choose to utilize a MOOC more like a textbook, focusing on certain sections and not on others. For example, educational researchers may have several research methodology textbooks, containing a few

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

chapters that are extremely relevant for a researcher who frequently utilizes the concepts. In the same way, it is possible that some learners come in and out of a MOOC as needed. Before evaluative inferences can be made about the quality of any MOOC, more research is required to understand types of learners and their learning needs better.

While large studies with descriptive statistics based on learner behaviour are abundant in the MOOC literature, it is imperative that researchers first focus on identifying the groups of learners in the course. Lumping all learners together and calculating the overall course assessment scores does not adequately describe all that occurs within a course. With the variety of use patterns and the high level of not fully engaged learners, there is much noise in any statistical model that attempts to estimate based on overall course mean.

Additionally, the categories of learners as "Registrants" or "Starters" does not capture the variety of learners within a course. Additional categories based on learner patterns are needed. Others have pointed out that descriptive statistics in MOOC environments are not always helpful (e.g., Perna et al., 2014; Reich, 2014). We have demonstrated the use of clickstream data alongside course assessment to understand more about learners. Furthermore, overall course descriptive statistics are confounded by latent learner variables. However, by identifying types of learner behaviour and examining descriptive statistics within each course, we have a richer understanding of what occurred and how to make future curricular improvements based on that information.

Future research should focus on the value of the learning experience to each group of learners and how courses can be developed to incorporate this information. While descriptive reports about completion rates and access of course materials are one approach to making sense of the MOOC phenomenon, learner-centred, theory-driven approaches are needed to contextualize the meaning of the findings (Wiebe, Thompson, & Behrend, 2015). MOOC learners do not follow a normal distribution curve in terms of their behaviour in the course. However, there is strong evidence from Nanophotonic Modelling that there are groups of learners who are very similar. Once we begin to understand more about these groups of learners, we will have a more contextualized way to make evaluative and research inferences from the large datasets of clickstream behaviour. It may be appropriate to value MOOCs at least in part based on how learners benefit from *access* to the information, not only whether they follow the course in a traditional sense. There is also potential to break up courses into two or more modules with self-contained assessments so that students could select those most relevant to their interests. From this perspective, the next step in our research is to understand qualitatively how learners in each cluster felt about having this opportunity, and whether they benefited from the information and assessments in a quantifiable fashion.

9 ACKNOWLEDGEMENTS

This work was made possible by NSF Award EEC 1227110 - Network for Computational Nanotechnology Cyberplatform) and EHR 1544259. Any opinions, findings, and conclusions or recommendations

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms* (SODA '07), 7–9 January 2007, New Orleans, LA, USA (pp. 1027–1035). Philadelphia, PA: Society for Industrial and Applied Mathematics.

Collins, E. D. (2013). *SJSU Plus augmented online learning environment pilot project report.* Sacramento, CA: The Research and Planning Group for California Community Colleges.

Christensen, G., Steinmetz, A., Alcorn, B., Bennett, A., Woods, D., & Emanuel, E. J. (2013). The MOOC phenomenon: Who takes massive open online courses and why? http://dx.doi.org/10.2139/ssrn.2350964

DeBoer, J., Ho, A. D., Stump, G. S., & Breslow, L. (2014). Changing "course": Reconceptualizing educational variables for massive open online courses. *Educational Researcher*, *43*(2), 74–84. http://dx.doi.org/10.3102/0013189X14523038

Douglas, K. A., Mihalec-Adkins, B. P., Hick, N., Diefes-Dux, N., Bermel, P., & Madhavan, K. (2016, June). Learners in advanced nanotechnology MOOCs: Understanding their intention and motivation. *Paper presented at the 123rd Annual Conference of the American Society of Engineering Education*, 26–29 June 2016, New Orleans, LA, USA.

Harrer, A., Malzahn, N., Zeini, S., & Hoppe, H. (2007). Combining social network analysis with semantic relations. In C. Chinn, G. Erkens, & S. Puntambekar (Eds.), *Mice, Minds, and Society: Proceedings of the 8th International Conference on Computer-Supported Collaborative Learning* (CSCL 2007), (pp. 267–276). New Brunswick: International Society of the Learning Sciences.

Hecking, T., Ziebarth, S., & Hoppe, H. (2014). Analysis of dynamic resource access patterns in online courses. *Journal of Learning Analytics*, *1*(3), 34–60.

Hill, P. (2013, March 6). Emerging student patterns in MOOCs: A graphical view. *e-Literate* [Weblog post]. Retrieved from http://mfeldstein.com/emerging_student_patterns_in_moocs_graphical_view

Hill, P. (2013, March 10). Emerging student patterns in MOOCs: A (revised) graphical view. *e-Literate.* [Weblog post]. Retrieved from http://mfeldstein.com/emerging-student-patterns-in-moocs-a-revised-graphical-view/

Ho, A. D., Chuang, I., Reich, J., Coleman, C. A., Whitehill, J., Northcutt, C. G., Williams, J. J., Hansen, J. D., Lopez, G., & Petersen, R. (2015). Harvardx and mitx: Two years of open online courses fall 2012–summer 2014. Social Science Research Network (SSRN). http://dx.doi.org/10.2139/ssrn.2586847

Hollands, F. M., & Tirthali, D. (2014). *MOOCs: Expectations and reality.* New York: Center for Benefit-Cost Studies of Education, Teachers College, Columbia University.

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. http://dx.doi.org/10.1016/j.patrec.2009.09.011

Jordan, K. (2014). Initial trends in enrolment and completion of massive open online courses. *International Review of Research in Open and Distance Learning*, *15*(1), 133–160. http://dx.doi.org/10.19173/irrodl.v15i1.1651

Kirk, R. E. (2008). *Statistics: An Introduction, 5th ed*. Belmont, CA: Thompson/Wadsworth.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013, April). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 170–179. https://doi.org/10.1145/2460296.2460330

Kizilcec, R. F., & Schneider, E. (2015). Motivation as a lens to understand online learners. *ACM Transactions on Computer–Human Interaction (TOCHI)*, *22*(2), Article 6, 24 pages. https://doi.org/10.1145/2699735

Koller, D., Ng, A., Do, C., & Chen, Z. (2013). Retention and intention in massive open online courses: In depth. *Educause Review*, *48*(3), 62–63.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., & Fehr, E. (2005). Oxytocin increases trust in humans. *Nature*, *435*(7042), 673–676. http://dx.doi.org/10.1038/nature03701

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621.

Leckart, S. (2012, March 20). The Stanford education experiment could change higher learning forever. Wired.

Lisboa, P. J., Etchells, T. A., Jarman, I. H., & Chambers, S. J. (2013). Finding reproducible cluster partitions for the k-means algorithm. *BMC Bioinformatics*, *14*(1). http://dx.doi.org/10.1186/1471-2105-14-S1-S8

Liyanagunawardena, T. R., Adams, A. A., & Williams, S. A. (2013). MOOCs: A systematic study of the published literature 2008–2012. *International Review of Research in Open and Distance Learning*, *14*(3), 202–227. http://dx.doi.org/10.19173/irrodl.v14i3.1455

Liyanagunawardena, T. R., Parslow, P., & Williams, S. A. (2014). Dropout: MOOC participants' perspective. In U. Cress, C. Delgado-Kloos (Eds.), *Proceedings of the Second MOOC European Stakeholders Summit* (pp. 95–100). Lausanne, Switzerland: P.A.U. Education.

Morabito, V. (2015). Big data and analytics. *Strategic and Organisational Impacts*. New York: Springer. http://dx.doi.org/10.1007/978-3-319-10665-6

Murphy, R., Gallagher, L., Krumm, A. E., Mislevy, J., & Hafter, A. (2014). *Research on the use of Khan Academy in schools: Research brief*. Menlo Park, CA: *SRI Education*. Retrieved from http://www.sri.com/sites/default/files/publications/2014-03-07_implementation_briefing.pdf

Perna, L. W., Ruby, A., Boruch, R. F., Wang, N., Scull, J., Ahmad, S., et al. (2014). Moving through MOOCs: Understanding the progression of learners in massive open online courses. *Educational Researcher*, *43*(9), 421–432. http://dx.doi.org/10.3102/0013189X14562423

Ragan, T. J., & Smith, P. L. (1999). *Instructional design*. New York: Macmillan Publishing Company.

(2016). Big data characterization of learner behaviour in a highly technical MOOC engineering course. *Journal of Learning Analytics*, *3*(3), 170–192. http://dx.doi.org/ 10.18608/jla.2016.33.9

Reich, J. (2014). MOOC completion and retention in the context of student intent. *EDUCAUSE Review Online*. Retrieved from http://er.educause.edu/articles/2014/12/mooc-completion-and-retention-in-the-context-of-student-intent

Reich, J. (2015). Rebooting MOOC research. *Science*, *347*(6217), 34–35. http://dx.doi.org/10.1126/science.1261627

Reich, J., Emanuel, J., Nesterko, S. O., Seaton, D. T., Mullaney, T., Waldo, J., et al. (2014). *HeroesX: The Ancient Greek Hero: Spring 2013 Course Report*. HarvardX–MITx Working Paper Series. http://dx.doi.org/10.2139/ssrn.2382246

Roco, M. C. (2011). The long view of nanotechnology development: The National Nanotechnology Initiative at 10 years. *Journal of Nanoparticle Research*, *13*(2), 427–445. http://dx.doi.org/10.1007/s11051-010-0192-z

Romero, C., Gutiérrez, S., Freire, M., & Ventura, S. (2008). Mining and visualizing visited trails in web-based educational systems. In R. S. J. d. Baker, T. Barnes, & J. E. Beck (Eds.), *Educational Data Mining 2008: Proceedings of the 1st International Conference on Educational Data Mining* (EDM 2008), 20–21 June 2008, Montreal, QC, Canada (pp. 182–186). International Educational Data Mining Society.

Stark, H., Woods, J. W., Thilaka, B., & Kumar, A. (2012). *Probability, statistics, and random processes for engineers* (Vol. 76). Upper Saddle River, NJ: Pearson Education.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*(1), 89–125.

Thille, C., Schneider, E., Kizilcec, R. F., Piech, C., Halawa, S. A., & Greene, D. K. (2014). The future of data-enriched assessment. *Research & Practice in Assessment*, *9*, 5–16.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society B*, *63*(2), 411–423. http://dx.doi.org/10.1111/1467-9868.00293

Wiebe, E., Thompson, I., & Behrend, T. (2015). MOOCs from the viewpoint of the learner: A response to Perna et al. (2014). *Educational Researcher*, *44*(4), 252–254. http://dx.doi.org/10.3102/0013189X15584774

Wilkowski, J., Deutsch, A., & Russell, D. M. (2014, March). Student skill and goal achievement in the mapping with Google MOOC. *Proceedings of the 1st ACM conference on Learning@Scale* (L@S 2014), 3–10. https://dx.doi.org/10.1145/2556325.2566240

Yuan, L., & Powell, S. (2013). *MOOCs and open education: Implications for higher education.* Glasgow, UK: JISC CETIS. Retrieved from http://publications.cetis.ac.uk/2013/667