Exploratory learning analytics methods from three case studies

David Gibson
Curtin University

Sara de Freitas
Murdoch University

Brief outlines of *exploratory analysis methods* (analysis designed to develop hypotheses) from three research projects illustrate the size, scope, variety and increased resolution that are becoming increasingly available at the unit of analysis for research in the learning sciences. The tools and methods applied in these studies are briefly outlined, which enable researchers to deal with complexity in time and event structures involving complex data in learning analytics projects. In particular, the transformation of data involving both reduction methods and pattern aggregation into *motifs* were found to be crucial for data interpretation. The article describes *data mining* with a *self-organizing map*, involving *unsupervised machine learning* and *symbolic regression* and combining exploratory analysis methods to achieve *causal explanations*.

Keywords: Learning analytics, exploratory analysis methods, game-based learning

Introduction

Assessment and learning analytics challenges have dramatically increased since new digital performance affordances and user interfaces have become more complex. The increased complexity is due in part to technology's capabilities and roles in presenting interactive learning experiences and collecting rich data (de Freitas, 2014; Quellmalz et al., 2012) which is leading to the infusion of data science methods and techniques into learning and behavioral science research (Gibson & de Freitas, [forthcoming] 2013; Gibson, 2012). These changes require new quantitative methods as well as a reconceptualization of mixed methods (Tashakkori & Teddlie, 2003) that include domain experts as well as stakeholders in the construction of knowledge of such complex systems.

The case examples in this brief paper share some common features and challenges for learning analytics needed with big data sets. In particular, they come from systems where information is flowing rapidly, is highly varied in format and grain size, and is available in relatively high resolution, that is, the unit of analysis is quite small compared to the aggregated level where decisions need to be made; see for example (IBM, n.d.). They each involve high-resolution data collection compared to the previous norms in educational research. For example, instead of a small number of data points per subject – a pre-test, post-test and a handful of other data - these projects have several hundred data points per subject. The cases also illustrate the challenge of varied formats and grain size, in part because of role of technology as an interactive agent in the production of data; some of the data is produced by a complex interaction between the learner and the digital learning environment as well as from co-production of data by the learner in an environment and social context. These co-production situations produce highly variable and diverse data sources that must be reconciled and interwoven. Lastly, the data in each of the example cases needs to be analyzed relatively quickly in order for people to react, or the learning environment to adapt, to the learner, ideally before the interaction concludes.

Exploratory data analysis is research undertaken in order to develop hypotheses and to better understand the structure of information and physical relationships of some complex system. The data is usually unstructured, and the questions of interest might be ill formed, undefined or broad. In the 1970's John Tukey of the Bell Labs first promoted the idea of increasing the use of exploratory research methods to suggest hypotheses to test as a counterbalance to the overemphasis in research placed on confirmatory data analysis and hypothesis testing (Tukey, 1977). Elsewhere, we have asserted that this rebalancing is still very much needed in transformative educational research (Gibson & Knezek, 2011; Gibson, 2012; Kozleski, Gibson, & Hynds, 2012) and here we assert that perhaps nowhere is this more evident in research on teaching and learning than in learning analytics. The three cases presented next illustrate why.

Case 1: Virtual Performance Assessment

Case 1 comes from a purpose-built game of the Virtual Assessment Project at the Harvard Graduate School of Education (Gibson & Clarke-Midura, 2013), which has a finite but large array of learning affordances and an educative aim to examine whether the game is able to assess middle school students' abilities to design a scientific investigation and construct a causal explanation (Clarke-Midura, Mayrath, & Dede, 2010). The assessments start out with one of two problems that students must solve: Why is there a frog with six legs? What is causing the population of bees to die? The assessments were designed in the Unity game engine (http://unity3d.com/) and have the look and feel of a videogame (
Figure 1) in which students make choices, talk to people and use environmental and laboratory resources while trying to solve the mysteries.

Figure 1. Screenshots of two Virtual Performance Assessments

Participant actions (e.g. opening a page, saving a note) were time-stamped and labeled as *events*. Analytic data from two pilots consisted of 1987 users (423,616 event records, 205 records per subject) in the frog assessment and 1958 users (396,863 event records) in the bee assessment. In addition to the raw performance data on all events, the analysis also included demographic information about students (age, gender, class, teacher), the starting prediction for the cause of the problem, raw event data (e.g. up to when a student made their final claim about the problem) and human-scored constructs of designing a causal explanation and designing a scientific investigation.

The purpose of the exploratory analysis was to search for *patterns of action* that might relate to the performance of the user *correlated with the student's final claim*. Could the action log and score data tell us about the user's performance? Ultimately can performance in a virtual performance assessment replace performance on a test?

A variety of exploratory methods were used. *Empirical probabilities* (the relative frequency of an outcome in relation to a number of trials) were derived from simple counts of key actions such as the student's prediction of the likely cause, and the final choice of a claim about the cause based on evidence collected during the game. This method led to evaluation claims about how many students changed their minds from pre to post, which actions correlated most highly with success and failure, and what sources of evidence were most likely to be used by the most and least successful. *Cluster analyses* (grouping similar data objects near to each other) were used to discover whether particular sets of resources and actions influenced the final claims. This method led to a methodological finding that only clusters of event-pairs and event-triads (not the raw events alone) held value for analysis. The analogue of the raw event in this instance of data mining might be taking the average of a non-normal distribution in a traditional statistical analysis, which tends of obscure rather than reveal important summary information about the data. Finally, *symbolic regression* (which searches the space of mathematical expressions to find a class of equations that best fits a given dataset) was used to discover whether underlying dynamic network relationships could be represented by predictive mathematical expressions and *network analyses* (which uses graph theory to understand relationships in data) were used to visualize the networks based on those regressions. This method led to the finding that certain key actions by the most successful groups were unique to those groups, which means that in the future, the game authors could use that knowledge to prompt students to try certain actions in order to increase their chance of joining the most successful students.

Case 2: Attrition in a Massively Open Online Course

Case 2 concerns a Curtin University learning analytics project aimed at discovering characteristics of attrition in a massively open online course or MOOC. A 'funnel of participation' has also been observed in online courses (Clow, 2013) and MOOCs in particular have been found to have only 7-10% of participants completing the

experiences (e.g., Daniel, 2013).
Data sources included performance data on 507 participants in a massively open online course was examined to determine what regimes of behavior might be evident through the exploratory method of symbolic regression (Schmidt & Lipson, 2009). Data was also collected on a survey of 64 students at the beginning of Curtin's Astronomy MOOC and were combined with data from the activity, grade and completion date for 177 individuals (31,000 records, 175 records per subject).

The purpose of the analysis was to help determine whether Curtin's offering would have similar attrition as expected in other kinds of online courses (e.g. Simpson, 2012) and if so, what kinds of action patterns in the data (e.g. doing assignments, contributing to discussions, using resources) might help predict when someone was about to drop out?

Exploratory methods used in this study included initial visualizations as data summaries to discover elements of the structure of the data (Figure 2). For example, how did the number of activities completed relate to completion status and final grade? This method led to finding three regimes: 1. people who completed less than 20 activities and whose final grade was unrelated to the number activities; 2. people whose final grade was positively related to the number of activities they completed; and 3. people who completed nearly all activities and whose grade was above average as a result.

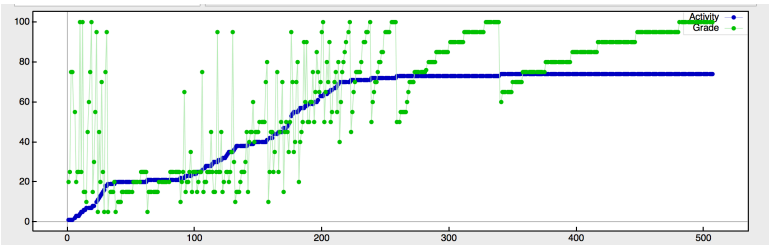

Figure 2. Visualization of 507 MOOC students sorted by activity level (solid line) juxtaposed with their final grade (connected dots)

A symbolic regression was conducted to discover whether a predictive equation could be found for the relationship of number of activities completed to the final grade. For group two, people (n=195) who completed between 20 and 69 out of 74 total activities, a strongly predictive relationship was found to the final grade (Figure 3).

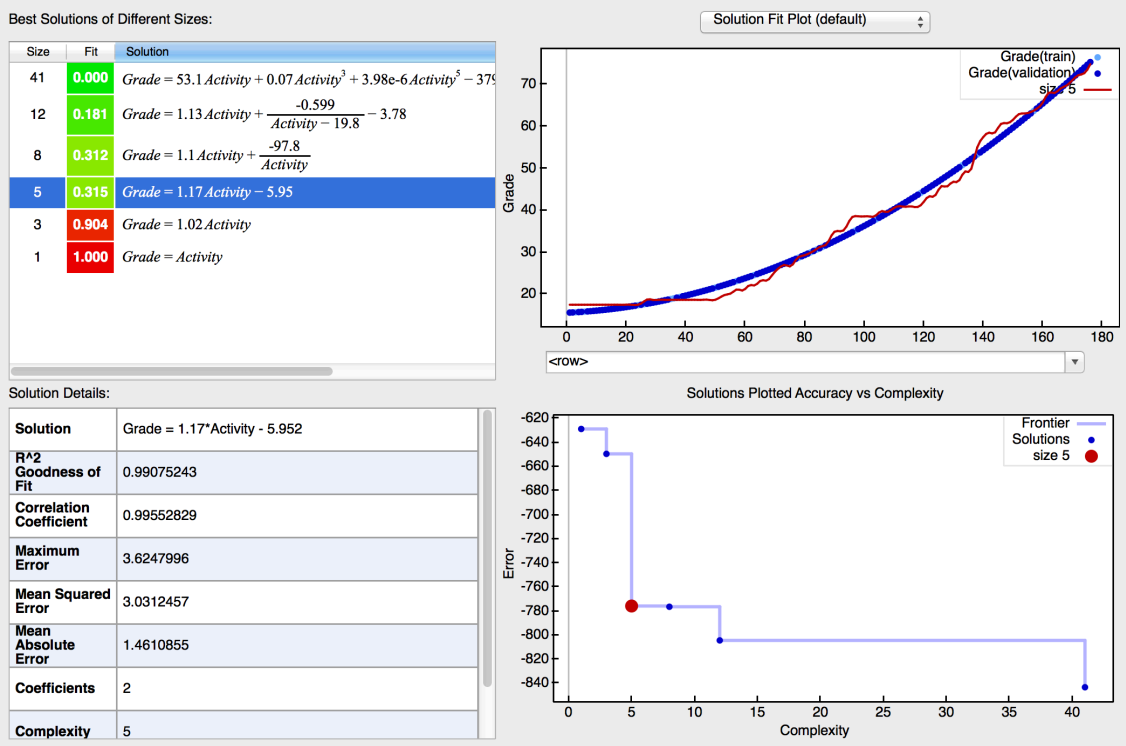

Figure 3. Symbolic regression results for n=195 people who completed between 20 and 69 out of 74 activities. The selected predictive equation (Grade = 1.17 Activity – 5.95) has a corr = .99 and r^2 = .99. The result was cross-correlated between training and validation subsets as well on a new population.

Case 3: Retention in Higher Education

In the third case, the purpose of the quantitative analysis was *to better understand* and *characterize similar behavioral groups* among students who leave and those who stay in higher education. We wanted to explore why some students succeed in higher education courses of study while others drop out, in the hopes that the university can develop better intervention methods that provide appropriate support and academic guidance.

Data sources included several large sets of anonymous, post-hoc course and unit attendance and grade data from the university's student data application, the learning management system, post-unit student evaluation surveys, online library use, interviews and focus group sessions that generated over 240 hypotheses, concentrating on a four-year period of the university's recent history. Over 1200 attributes for each student were considered (62 million elements), which led to a need for *data reduction* methods and sensitivity to *multicollinearity* (e.g. the combined effect of linked variables that lead to overstating their influence and drowning out the influence and details of other variables). A 13 million-element 'big data' set was extracted and served as the basis for a *semi-supervised machine learning model* that mapped clusters of similar behaviors, which was trained on 52,000 students with over 250 attributes.

The analytics methodology followed *a staged process of data acquisition, preparation, discovery and analysis*, followed by the creation of a *self-organizing map* (SOM) that was further shaped by the analytics team. An SOM uses a neighbourhood clustering function to preserve the topological properties of the input space. The staged research process engaged a wide spectrum of key stakeholders, with several diverse data sources and over 240 hypotheses statements coming from and going back into a series of public engagements across the university. The data sources were combined and transformed into an analytic data set (ADS) and that set was then used to build the SOM. The method allowed initial in-depth testing of 50 of the hypotheses that were immediately amenable to data discovery and exploratory analyses.

Discussion & conclusion

The records-per-unit-of-analysis varied from 175 to 250 in these cases while the total data-element sizes of the data based varied from 31,000 to 13 million. So the record sizes and resolution greatly, but similar exploratory analysis tools and methods were useful in each case, varying only slightly depending upon the purposes of the research. For example, some form of data mining (e.g. clustering, machine learning, symbolic regression, network analysis, a self-organizing map) was used in all cases, and some of these methods used in combination also proved useful for constructing evidence that supported causal explanations in Case 1 and 2.

Transforming data for data mining in all cases involved both reduction moves and intermediate pattern aggregations, which had important implications for data interpretation. For example, the clustering in Case 1 was ineffective until subject domain experts identified a two or three-element chain of actions called a motif. In Case 3, reductions followed traditional lines (e.g. searching for multicollinearities), but in addition, the relatively novel use of a self-organizing map (SOM) in educational research is itself a form of motif creation via transformation that led to multiple hypotheses generation; the unsupervised and supervised phases of the machine learning method chose a different subset of attributes for each cell of the map to maximize the local similarity of neighboring groups, which led to deeper questions about how and why the clusters form and how they relate to the prediction variable of the probability of attrition. The advantage of unsupervised machine learning in this case is that no prior hypotheses or assumptions creep into the model, and the resulting best-fit map is then available for multiple hypothesis testing as overlays on the SOM, while the supervised reduction of the dataset allowed finer and finer grain sizes of the underlying population similarities. This method allowed us to test quite a large number of hypotheses compared to traditional research methods.

The exploratory analysis stance does not attempt to determine a correlational strength of a well-formed and presupposed hypothesis, but instead allows the data to speak for itself. The methods described here are recommended for a toolkit for learning analytics.

References

Clarke-Midura, J., Mayrath, M., & Dede, C. (2010). Measuring Inquiry : New Methods , Promises & Challenges. *Library*, 2, 89–92.

De Freitas, S. (2014). *Education in computer generated environments*. London New York: Routledge.

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

Gibson, D., & Clarke-Midura, J. (2013). Some Psychometric and Design Implications of Game-Based Learning Analytics. In *Cognition and Exploratory Learning in the Digital Age*. Forth Worth: CELDA-IADIS.

Gibson, D., & de Freitas, S. (2013). *Technology-based assessment of work integrated learning*. In Press, (special issue).

Gibson, D., & Knezek, G. (2011). Game Changers for Teacher Education. In P. Mishra & M. Koehler (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference 2011* (pp. 929–942). Chesapeake, VA: AACE.: AACE.

IBM. (n.d.). Big Data. Retrieved from http://www-01.ibm.com/software/au/data/bigdata/

Kozleski, E., Gibson, D., & Hynds, A. (2012). Changing complex educational systems: Frameworks for collaborative social justice leadership. In C. Gersti-Pepin & J. Aiken (Eds.), *Defining social justice leadership in a global context* (pp. 263–286). Charlotte, NC: Information Age Publishing.

Quellmalz, E., Timms, M., Buckley, B., Davenport, J., Loveland, M., & Silberglitt, M. (2012). 21st century dynamic assessment. In M. Mayrath, J. Clarke-Midura, & D. Robinson (Eds.), *Technology-based assessments for 21st Century skills: Theoretical and practical implications from modern research (*pp. 55–90). Charlotte, NC: Information Age Publishers.

Schmidt, M., & Lipson, H. (2009). Symbolic regression of implicit equations. Genetic Programming Theory and Practice, 7(Chap 5), 73–85.

Tashakkori, A., & Teddlie, . (2003). *Handbook of Mixed Methods in Social & Behavioral Research* (p. 768). SAGE. Retrieved from http://books.google.com/books?id=F8BFOM8DCKoC&pgis=1

Tukey, J. (1977). *Exploratory Data Analysis*. NYC: Addison-Wesley.

Contact author: David Gibson,david.c.gibson@curtin.edu.au

Please cite as: Gibson, D., & de Freitas, S. (2014). Exploratory learning analytics methods from three case studies. In B. Hegarty, J. McDonald, & S.-K. Loke (Eds.), *Rhetoric and Reality: Critical perspectives on educational technology. Proceedings ascilite Dunedin 2014* (pp. 383-388).

Note: All published papers are refereed, having undergone a double-blind peer-review process.