Situating Big Data Across Heterogeneous Data Sets of Game Data Exhaust, Class Assessment Measures, and Student Talk

Lauren Wielgus, University of Wisconsin–Madison
Craig G. Anderson, University of Wisconsin–Madison
John Binzak, University of Wisconsin–Madison
Jennifer Dalsen, University of Wisconsin–Madison
Pasqueline Scaico, Universidade Federal da Paraíba
David Azari, University of Wisconsin–Madison
Vanessa Meschke, University of Wisconsin–Madison
Matthew Berland, University of Wisconsin–Madison
Kurt Squire, University of Wisconsin–Madison
Constance Steinkuehler, University of Wisconsin–Madison

Abstract: This project seeks to marry theories of situated cognition to the big data movement by connecting clickstream data from technologies in isolation to key forms of multimodal data available from their contexts of use. Using a data corpus gathered from a five-day implementation of the STEM game *Virulent* (targeting cellular biology), we are combining multiple analytic strategies commonly considered incommensurate including educational data mining, qualitative coding, discourse analysis, natural language processing, and standard classroom assessments. In this paper, we review the project goals and preliminary findings, and discuss the benefits and drawbacks to analysis across heterogeneous data sets. Our goal is to provide a more complete model for big data analysis, one that includes both talk and play data equally or, where not possible, identify find its limitations so that future "data rich" attempts on learning might be better informed by the limitations of technology-rich but talk-poor data sets.

One of the defining questions for education over the next decade is, how do we shift education from a data poor to a data rich activity? (T. Kalil, personal communication, September 1, 2013). "Big data" techniques (the capture, curation, storage, and analysis of massive, complex data sets spanning large numbers of individuals) have made progress in areas such as content knowledge, inquiry practice, and, to a lesser extent, interest, but significant work remains in areas such as identity, participation, and epistemology – domains historically studied through discourse analysis and other forms of qualitative methods. The majority of data streams used in big data analyses are "data exhaust" from technologies considered largely in isolation. We know, however, that such technologies are sociotechnical artifacts (Bijker, 1995) whose potential for learning, like that of any instructional tool, is highly influenced by its context of use. Whether it's a textbook, calculator, or high-end 3-D graphical data display, a tool is only as good as the activities and practices in which it is embedded. Thus, if we want to catalyze progress toward more expanded frameworks for learning goals that include tricky variables such as identity and dispositions, then, we must include not only the data streams from technology and tool use but also talk and interaction data that surround it. And we would be wise to build on the last several decades of discourse and content analytic techniques used routinely in more qualitatively oriented research.

This project seeks to marry theories of situated cognition to the big data movement by connecting clickstream data to key forms of multimodal data available from their contexts of use. Using a data corpus gathered from a five-day game-based implementation of the STEM game *Virulent* (targeting cellular biology), we are combining multiple analytic strategies commonly considered incommensurate, including educational data mining, learning analytics, quantification of qualitative coding, natural language processing, and standard classroom assessments. The data include: clickstream telemetry data; individual and group discourse; individual and curricular artifacts; classroom assessments; and online forum postings. In this paper, we describe the context of our investigation, including the game used and its attending curricula, and then detail our data collection methods thus far. We discuss the benefits and drawbacks to analysis across heterogeneous data sets and our current attempts to develop a more complete model for big data analysis, one that includes both talk and play data equally or, where not possible, identify its limitations so that future "data rich" attempts on learning are better informed by the limitations of technology-rich but context-poor (and talk-poor) data sets.

The Game & Activities

Virulent (Figure 1) is a real-time strategy game about viral infection and the cellular immune system. Players control the fictional "Raven" virus and guide it to infect a host organism. *Virulent* allows players to 'hijack' cellular parts such as mitochondria, protein receptors, and nuclear pores, in an approximation of a viral life cycle. Players also utilize host cell ribosomes to enable translation of viral mRNA fragments into proteins to use as power-ups.

Additionally, *Virulent* features a graphical almanac where players read about different units, enemies, and cellular structures found in the game, such as proteasomes, budding sites, and viral genomes.

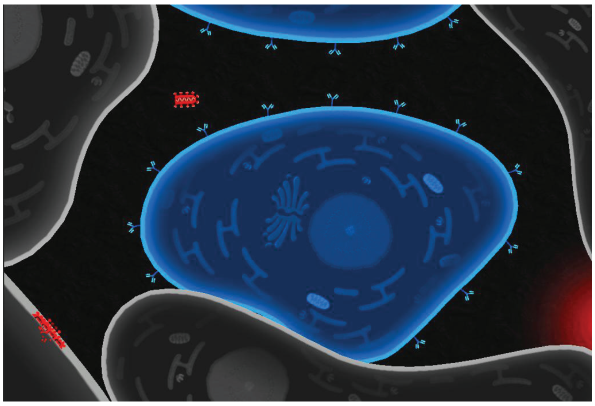

Figure 1: Screenshot of *Virulent* gameplay.

We developed a five-day role-playing curriculum based on the game focusing on a fictional outbreak of the Raven virus. The overarching narrative of the role-playing unit was that the Center for Disease Control (CDC) needed young innovative scientists (the students) to help them figure out how to stop a dangerous outbreak of the virus. At the beginning of the unit, the class received a pre-recorded video conference call from actors pretending to be CDC scientists and were given the mission of using their "digiscope" (their iPad with the Virulent game software on it) to figure out how the virus attacks cells and develop a strategy for stopping it. The curriculum unit lasted five days total (90 minutes per session). Each day, students received updates on the status of the outbreak from the "CDC" (Figure 2). Participants were divided into small "research teams" of 3-4 students; each team was responsible for using the game to create a model of how the virus and cell interact. Each team had to present their model to their peers and to the CDC scientists (via video). On the fifth and final day, the CDC gave teams three plausible intervention ideas to fight the outbreak of the virus (develop a vaccine, use an RNA interference treatment, or inhibit the cell's mitochondria) along with outside information sources about each solution (a news article about vaccines, an adapted science journal article about RNA interference, and a mock textbook page about mitochondria). Participants then attempted to reach consensus in a whole class debate as to which solution they should recommend to the CDC based on their model and the outside resources given and converged on a final course of action.

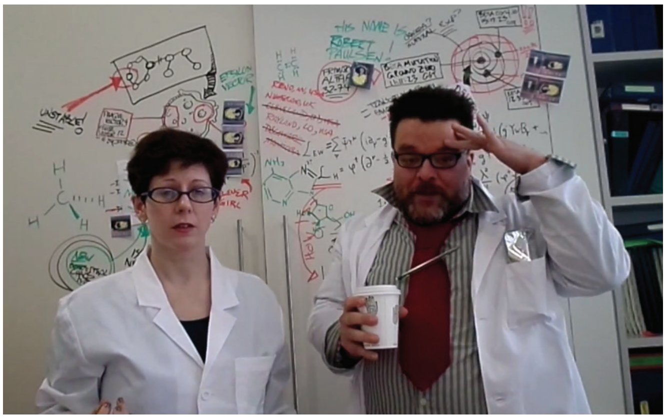

Figure 2: Still image from "CDC" Skype call featuring actors playing concerned scientists.

Data Collection

Data was collected in three separate contexts: as part of a spring break camp run by GLS called Game-a-Palooza (*N*=34 across 3 cohort groups), within in a single class at a local private school's extended day program (*N*=11), and as a unit within a larger summer camp run by the Boys and Girls Club (*N*=27). Across all three contexts, participants were separated into their research teams (*n*=3-4) with each team supported by a facilitator. We assembled teams semi-randomly (based on age) but allowed players to switch teams to enable participants to play with friends or siblings also enrolled in the event. Over the course of five 90-minute sessions, the research team was the primary context for interaction.

Across these nested groupings, we collected multiple data streams from each participant. For each individual participant, across all five 90-minute sessions, we recorded all in-game telemetry data via our backend data framework, the Assessment Data Aggregator for Gaming Environments (ADAGE). ADAGE (Owen & Halverson, 2013) is a tool for capturing and tagging clickstream data to correspond with moves, decisions, and events within the game, allowing for analysis of player actions in-game to assess learning and triangulate those in-game play patterns against other external measures. We also collected the in-room complement to this telemetry data stream: the complete stream of verbal data via lavalier audio recorders. We collected individual pre-/post-assessments, including: Likert items measuring attitudes about, interest in, and confidence in science; multiple-choice and open-ended items testing specific *Virulent*-related content; and creating a drawing of a scientist. We also collected all individual paper artifacts such as diagrams, worksheets and models. For each team, we collected: photographs of their models of the virus (to capture model creation and revision over time) and their short video explaining and justifying their model. For each cohort (*n*=15-20), we video-recorded classroom presentations of models of the virus and debate of proposed solutions for the epidemic, and we did daily interviews with the teacher. We also performed 28 stimulated recall interviews with participants several weeks after the unit was completed to understand participants' persistent understanding of and attitudes towards the content and to specifically prompt discussion of model revisions over the five days. We are using the Qualitative Data Analysis Software MAXQDA as the central data hub for our qualitative data.

Data Analysis

Four primary forms of analysis comprise our overall strategy: educational data mining and learning analytics, the quantification of qualitative codes, natural language processing modeling of student discourse and text, and pre/post traditional assessments.

Big data efforts in educational research fall under the general rubric of educational data mining (EDM) or learning analytics (LA), where EDM describes broad analysis of a large set of learners (>1K) and LA describes deeper analysis of a smaller set. EDM has predicted how in-game behavior relates to cognitive-affect and performance metrics (Baker, D'Mello, Rodrigo & Graesser, 2010); assessed intuitive and formal understanding of specific physics content (Clark & Martinez-Garza, 2012); and demonstrated student learning of programming through strategic game mechanics (Berland, Martin, Benton, Petrick Smith, & Davis, 2013). Our ADAGE data framework guides game developers to articulate the content model of their games through a metadata tagging process that facilitates data mining and exploration as well as matching play patterns to key events in game play (Owen, Shapiro & Halverson, 2013). Variables for analysis for Virulent are included in *Table 1*. Note that the variables listed do not entail specific research questions, but require additional referential work in order to answer questions of interest.

Virulent Telemetry Variables
Play Time Totals (in- and out-of-session)
Levels Played, Level Success (pass/fail, time taken, number of attempts)
Scores per Level (units lost, units destroyed)
Challenge Completions (judged by coin rewards criteria)
Cell Resource Use (mRNA creation, protein creation)
Time Spent on Level Intro Panel (instructions)
Time Spent in Almanac, Almanac Entries Referenced
Spawn, Death, and Collision Locations of Game-Controlled and Player-Controlled Resources
Target Selection by Game-Controlled Resources (whenever a game-controlled resource targets a player-controlled resource)

Table 1. Telemetry variables for analysis in *Virulent*.

Two forms of contextual data analysis are being conducted: The UW-Madison team focuses on quantified qualitative coding of the data corpus while a second analytic team at Arizona State University, led by Danielle McNamara, focuses on natural language processing analyses. Analysis in Madison includes an “a priori” six construct coding scheme – based on the National Academy of Science’s “six strands of science learning” framework (Bell & Lewenstein, 2009): (a) interest, (b) content knowledge, (c) inquiry practice, (d) epistemological disposition, (e) longer term participation in the field, and (f) identity development within the domain – and “ground up” constructs related to the six primary themes that arise from the data. We assess aggregated patterns through code counts (where possible) and examine interrelationships among our codes, noting any discursive features (e.g., keywords, grammar) that cluster with codes which could be explored for natural language processing at ASU. The ASU lab applies techniques from natural language processing to find patterns within the textual data corpus in terms of lexical density and sophistication, syntactic complexity, cohesive devices, and other discourse features using tools including the Linguistic Inquiry Word Count (LIWC; Pennebaker, Booth, & Francis, 2007) and Coh-Metrix (Graesser, McNamara, & Kulikowich, 2011; McNamara, Graesser, McCarthy, & Cai, in press). Such tools have been used in a variety of contexts to understand the processes involved in both discourse comprehension and production. When used together, they can be extremely powerful, capturing a wide range of psychological and linguistic attributes of text (e.g., Crossley & McNamara, 2012; Varner, Roscoe, & McNamara, 2013).

Results from these two approaches to contextual data analysis can be merged in order to assess how and where patterns of discourse features and human-coded trajectories of learning converge. We know that some features of discourse and artifacts that are crucial to context and meaning can be meaningfully quantified (e.g., discourse markers, Schiffrin, 1993; quantifiable codes, Chi, 1997) for comparison to patterns in telemetry data sets while the frequency of other structures is less meaningful (e.g., singular utterances that signal a crucial transition in an interaction may only appear once but once is enough, Gee, 2010). Thus, our task is to determine fruitful areas where patterns uncovered through human interpretation co-locate with patterns discerned through NLP (Natural Language Processing) and other automated means. This combination of analyses enables to cast a wide, empirically robust net on socio-material interactions.

Discussion

Data Collection Challenges

Collecting a rich, multi-modal dataset presents a series of difficult challenges. First, assembling the right team of researchers and designers is non-trivial. Data collection alone required master teachers, trained and supported facilitators, an outreach coordinator, and several undergraduate interns whose job was only to keep track of students and their data as they moved about the facilities. Analysis of the corpus will require experience in EDM, inferential statistics, qualitative inquiry, quantified codes, NLP, discourse analysis, and both game and curricular design.

Although data were carefully labeled throughout the event and we had a dedicated researcher whose sole focus was data collection and organization, lossy data was unavoidable. To facilitate the tracking of each individual participant across the data set, we used lavalier name-badges with an integrated audio recorder and unique QR codes in place of logins, as logins (and passwords) can be used inconsistently. Despite efforts to collect complete data streams on each and every student, absences were common – and because much of the activity was group-based, we had a constant renegotiation of groups, roles, and membership. One researcher tracked the makeup of each group on each day, but analyses will have to be mindful of the flux of participants between/across groups. Such shifts in team composition generate noise in the data corpus for both assessed impact and learning progressions, but are unavoidable in a voluntary event. Future analyses need to include how team composition, stability, and cohesion impact attitudinal and conceptual change.

There are also basic limitations to telemetry data that can be collected on iPads. We intended to log the positions of all player-controlled and game-controlled objects at all times given that players manage and manipulate a whole system of assets. However, due to the large number of moving objects on-screen during higher levels and an iPad’s limited processing power, recording that volume of data was impossible. As a result, we opted to only record the paths of player-controlled objects, with the position of game-controlled objects recorded only upon a significant event (e.g., an object appears on screen). This reduction still resulted in over two million lines of data but sufficiently reduced the load on the iPads.

Additionally, the transcription and integration of 60 separate audio files, one of our largest and richest sources of data in the study, is non-trivial. Each participant was issued a small recorder the size of a USB drive that stayed in their name badge throughout the event, ensuring that every small group or whole cohort discussion was recorded from multiple points. This system created much-needed redundancy in our recordings, allowing us to get around ill-placed recorders and fidgety subjects by providing us multiple copies to select from. The resulting data corpus, however, is roughly 10,000 minutes of audio data that must be transcribed and stitched together into single coherent transcripts of student talk.

Data Analysis Challenges

Our first challenge to analysis of this large corpus is choosing the right unit of analysis given the various configurations of group, cohort, and class. The data collection strategies used enable analysis on the individual level, group level (*n*=3-4), cohort level (*n*=15), and event level (*n*=20-45). If we examine gameplay data only at the individual level, we may miss effects that players had on each other as they played the game together. Across all analyses, we need to carefully consider the nestedness of any data points, which can be influenced by team dynamics, facilitator styles, and overall event context.

A second challenge to analysis is the resourcefulness required to clean the telemetry data. No matter how well the data architecture is designed, in practice its first several runs generate repeated events that need to be found and removed. Play times can be skewed by players simply leaving their iPads running. Moreover, we know from the last decade of games based research that players often have unique game-play strategies, but identifying them within the datasets can be a daunting challenge. In the case of Virulent and given data collection constraints on the iPad, we can only collect the positions of player controlled objects and not game controlled objects, making it even more difficult to determine play-styles based on moment-by-moment choices in relation to the current in-game context of events.

A third and perhaps most crucial challenge to analysis is the complexity of our initial plan for quantifying qualitative data. Our initial "a priori" coding scheme is based on all six strands of the National Academy of Science's (Bell & Lewenstein, 2009) informal science learning framework, and thus errs on the side of comprehensiveness rather than depth. This scheme will likely need to be iterated and dramatically reduced in scope if we hope to have any sense of reliability not just within-analysis but between analysis and depth. Our "ground up" strategy for generating additional related codes is again comprehensive and needs to be narrowed to key discourse markers or other reliable (and ideally countable) cues in order to intersect in meaningful ways with both the EDM/LA and NLP data analyses. With such a large corpus, any attempt to analyze data line by line, turn by turn, and frame by frame will have to be conducted on small, strategically sampled excerpts of data, not the corpus overall. Addressing both challenges will require multiple iterations, cross-disciplinary communication, and patience.

There is an inescapable tension in combining rich qualitative data with quantitative data. In order to link the qualitative audio data to in-game clickstream data and, ideally, some score on a pre/post test, it needs to be quantified in some way. The parts of the audio data that are easiest to quantify, such as a count of vocabulary words used by a participant, may not be the most meaningful. The game almanac is a case in point. In Virulent, the almanac provides players with just-in-time information about the cell, virus, and immune system and all uses of this almanac are recorded by ADAGE. One obvious path for analysis across our disparate strategies is to compare each individual's in-game almanac use to players' vocabulary in pre-/post-assessments and interviews and key word counts from audio data. Such analyses will allow us to more accurately account for individual vocabulary use and development over time and its relation to gain on pre/post assessments. Whether such vocabulary are the constructs most worth investigation is an open question.

Conclusion

This project uses big data research techniques to provide a quantifiable approach to understanding learning in context. Current efforts struggle to measure learning in complex learning environments at scale due to difficulties in capturing how learners interact with a broad range of tools and resources across time. Game-based learning approaches, in addition to providing a rich data exhaust, can demonstrate for learners how and why science is useful and provide them experiences using science to solve problems, which are pedagogical approaches known to increase participation in science (Bell & Lewenstein, 2009). Assessing learning in action through games rather than *post hoc* enables educators to collapse formative and summative assessments into routine assessments within an integrated learning activity, providing better learning data and more valid claims about what students know. If a defining question for education now is how to move education from a data poor to a data rich activity, and we stay committed to frameworks for learning that include more than simply declarative knowledge and rote skills, then we must take seriously the creation of big data systems that can handle richer data and richer educational constructs. The proliferation of digital devices and distribution platforms makes the rapid expansion of a system inevitable; the next stage in this line of inquiry is to consider the full ecosystem of learning to take into account context and not just technology.

References

Baker, R .S. J. d., D'Mello, S. K., Rodrigo, M. M. T., Graesser, A. C. (2010) Better to be frustrated than bored:

The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies, 68*(4), 223-241.

Bell, P., & Lewenstein, B. (2009). *Learning science in informal environments*. National Academies Press.

Berland, M., Martin, T., Benton, T., Petrick Smith, C., & Davis, D. (2013). Using learning analytics to understand the learning pathways of novice programmers. *Journal of the Learning Sciences*, *22*(4), 564-599.

Bijker, W. E. (1995). *Of bicycles, bakelites, and bulbs: Toward a theory of sociotechnical change*. Cambridge, MA: MIT Press.

Chi, M. T. H. (1997). Quantifying qualitative analysis of verbal data: A practical guide. *Journal of Learning Sciences, 6*(3), 271-315.

Clark, D. B., & Martinez-Garza, M. (2012). Prediction and explanation as design mechanics in conceptually-integrated digital games to help players articulate the tacit understandings they build through gameplay. In C. Steinkuehler, K. Squire & S. Barab (Eds.), *Games, learning, and society: learning and meaning in the digital age* (pp. 279-305). Cambridge: Cambridge University Press.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading, 35*(2), 115-135.

Gee, J. P. (2010). *An Introduction to discourse analysis: Theory and method (3rd Edition*). London: Routledge.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix providing multilevel analyses of text characteristics. *Educational Researcher*,*40*(5), 223-234.

Owen, V. E., & Halverson, R (2013). ADAGE (Assessment Data Aggregator for Game Environments): A click-stream data framework for assessment of learning in play. Proceedings *from Games+Learning+Society Conference 9.0*. Pittsburgh: Lulu Press.

Owen, V. E., Shapiro, R. B., & Halverson, R. (2013). Gameplay as assessment: Analyzing event-stream player data and learning using GBA (a Game-Based Assessment Model). Proceedings from *Tenth International Conference on Computer-Supported Collaborative Learning.* Madison, WI.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic inquiry and word count: LIWC [PC software]. Austin, TX: Pennebaker Conglomerates, Inc.

Schiffrin, D. (1993). *Approaches to discourse: Language as social interaction*. Cambridge MA: Blackwell.

Varner, L., Roscoe, R., & McNamara, D. (2013). Evaluative misalignment of 10th-grade student and teacher criteria for essay quality: an automated textual analysis. *Journal of Writing Research, 5*, 35-5. 2