INTRODUCING SECONDARY SCHOOL STUDENTS TO BIG DATA AND ITS SOCIAL IMPACT: A STUDY WITHIN AN INNOVATIVE LEARNING ENVIRONMENT

Einat Gil and Alison L. Gibbs
The Department of Statistical Sciences at the University of Toronto
egil767@gmail.com, alison.gibbs@utoronto.ca

In this paper we report on a study of an innovative curriculum unit in which secondary school students learned about Big Data using real multivariate data with local and global contexts. The unit was designed to promote students' covariational reasoning and engagement with big data concepts, while experiencing how statistical tools can be used to investigate trends and relationships to make decisions that can positively impact society. Inspired by notions of Classroom of the Future, the program used Interactive Orchestrated Learning Space (IOLS) and knowledge community approaches to create a multi-disciplinary learning environment. Students' learning was investigated with mixed methods research tools. Findings from this study, including knowledge gains, are presented together with examples of students' use of covariational reasoning to create arguments for change in a large city.

INTRODUCTION

Current developments in the availability of data, including new sources, and the role information plays in scientific, business, government, and day-to-day life have created new challenges in statistics education. The influx of data through social networks, business, and governmental websites and increasing numbers of websites and tools for global monitoring in different disciplines reflect the growing interest in data-based decision-making and awareness of the opportunities of big data (Agrawal et al., 2012). Yet development of programs to introduce big data to students is still in its early stages in tertiary education, let alone for secondary school students. This paper presents the design of a curriculum for 12th grade students that uses innovative interactive technology and a collaborative approach to knowledge construction to introduce ideas about big data and promote covariational reasoning, relying on data and applications with important social impact to motivate learning.

Big data is often characterized by the four V's (Beyer & Laney, 2012): Volume – sheer size, Velocity – ongoing streaming of information leading to rapid accumulation and need for constant updating, Variety – different data types or data originating from different sources, resulting in data which are often unstructured, and Veracity – accuracy of the data. Some authors (e.g., Thirunarayan, & Sheth, 2013) have suggested an additional V, Value, which we find to be of particular relevance to the educational context. Value refers to the ability to extract meaningful information from the data. While this is often mentioned in the context of economic value to a business, it also includes the role big data can have in providing scientific and social insights.

Placing more emphasis on larger samples together with introducing multivariate data and data visualization was suggested by Ridgeway (2015). Gould (2010) recommended that statistics education aim at producing citizen statisticians that create and analyze available and complex data. Furthermore, a study using varied streaming mid-size mobile data was pioneered in a school in the context of democratic participation (Philip, Schuler-Brown & Way, 2013). It is evident that connecting big data to education at the high school level requires simplification both in concepts and processing tools using a variety of technological and contextual scaffolds.

Covariational reasoning is of particular importance in big data where the goal is often to discover hidden structure in the data. In big data, the number of variables is often larger than the number of observations (Franke et.al, forthcoming) and it is necessary to understand if and when this structure can be seen across subgroups (Fan, Han & Liu, 2014). Thus facility in covariational reasoning and the use of modeling to explore associations between variables are necessary to develop deeper understanding of data (Lehrer & Schauble, 2010).

Covariational reasoning plays an important role in different disciplines, including science and mathematics (Garfield & Ben-Zvi, 2008). The features attributed to engaging in covariational reasoning include looking at the shape and strength of the relation and generalizing and contextualizing these observations (Watkins, Scheaffer, & Cobb, 2004). It is one of the topics in statistics that challenges students. Among the difficulties is correct interpretation of a negative covariation, especially if it is counter to a prior belief, unidirectional and casual misconception, and a tendency to fit a linear model, even in the case of a non-linear or no existing relationship (Moritz, 2004; Batanero, Estepa & Godino, 1997; Casey, 2015).

With these theoretical challenges, our aim was to engage students in experiencing and investigating real, new sources of data, both big data and smaller data with characteristics of big data, while encouraging students' understanding of how they might use data in the future to influence the development of a better society.

METHODS

In early 2015, 55 students from two 12th grade classes in an Ontario secondary school for above average students participated in a three week unit in the course Mathematics for Data Management. The program was designed to allow multi-disciplinary investigation of big data and social data derived from big data or with some of the characteristics of it, and to support the development of covariational reasoning in multivariate data. Table 1 shows the curricular unit relating to both statistics and contextual aspects of the five activities.

Big Data Interactive (second activity) included a special class design - Interactive Orchestrated learning Space (IOLS; Gil & Slotta, 2015; Figure 1) inspired by notions of Classroom of the Future (e.g. Slotta, 2010). The IOLS takes into account the classroom as a physical space (Slotta, 2010) and operates in a transformed regular classroom using laptops, projectors, large displays (foam boards and a television) and a SMART board. In the IOLS, students' learning is orchestrated through two sets of four interdisciplinary stations illustrating the use of big data in a variety of important societal contexts and that are designed to promote collaborative knowledge

Table 1: Unit activities and statistical context

#	Focus	Statistics	Context
1	*Trends in the World of Data* Introduction to data inquiry	Data analysis / experience association in a multivariate context	World countries with GapMinder
2	*Big Data Interactive* Experiencing and learning about Big Data in Interactive Orchestrated Learning Space	Learning about Big Data	Big data resources on Social networks, Health, Ecology, Surveillance & privacy
3	*The Faces of Big Data* Discussion about aspects of Big Data;	Characteristics of Big data	Students responses about big data resources (act. 2)
	Introduction of statistical tool (iNZight)	Learning tools and developing facility in investigating association in a multivariate context	Exercise data
4	*Toronto Explorations* Exploring data with iNZight	Investigating association in a multivariate context	City of Toronto multivariate data
5	*Students' Big Data Pitch* (Presentation of inquiry)	Presenting findings from investigation	City of Toronto multivariate data

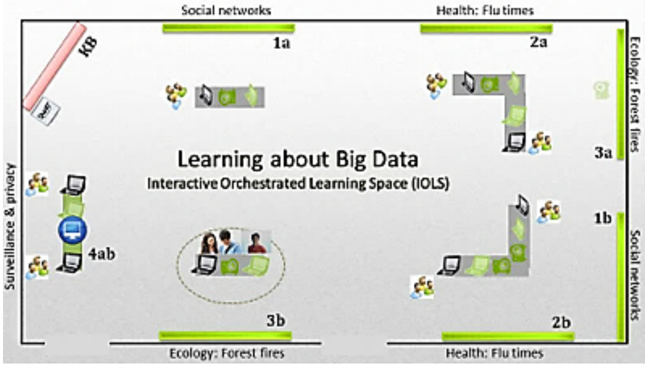

Figure 1. Interactive orchestrated learning space (Activity 2; Gil & Slotta, 2015)

building. The stations are Station 1ab Social networks: Learning about the data collected by Twitter; Station 2ab Health: Comparing models of Google Flu Trends and CDC data; Station 3ab Ecology: Investigating visual forest fire data, and Station 4ab Surveillance and Privacy: Confronting ethical aspects of big data through a Hollywood movie. In each station the students interact with the content using a group laptop and a station laptop, discuss and answer questions relating to the big data resource and submit their responses using Google forms embedded in the course website. A knowledge base (KB) aggregating the Google form results is streamed and presented both on the SMART board and a KB page in the website.

In activities 4 and 5, the students' task is to investigate a topic of their choice first visually using spatial visualization tools on the Wellbeing Toronto website, and then in iNZight, investigating associations among variables related to aspects of urban life including crime, transportation, ecology, and health. The data used in the 2015 program were a real 2011 dataset measured on 140 neighborhoods of Toronto, and were in certain respects representational of big data consisting of various observations on 2.6 million Toronto residents. The students prepared a pitch that is a "well-supported argument based on the Toronto Wellbeing data," to the newly elected city mayor (Gil & Gibbs, 2015). The pitch was presented to the class, teacher, school principal and researchers as the final phase of the unit.

Processing, visualization and analysis of big data requires knowledge of advanced computational and statistical methods. In order to make big data accessible to secondary school students, the following design considerations guided our approach: 1. Students interact with big data in an informal conceptual way avoiding the computationally and mathematically sophisticated layer; 2. Data are represented through visualizations (e.g. GapMinder, Global Forest Watch) and using tools recommended for secondary school students for the analysis of mid-size data (iNZight); 3. The unit begins and ends with the investigation of smaller data that represent aspects of big data, supporting activities for learning about big data.

Using a mixed-methods approach, gains in students' knowledge about big data and covariational reasoning were examined from a pre-post test, students' shared documents, and group presentations. The pre- and post-test focused on general statistical knowledge and covariational reasoning (9 questions), and knowledge about big data (3 questions). The test was checked and updated for content validity (Fraenkel & Wallen, 2009) with two external experts from statistics/statistics education. While the completion rate for the pre-test was 90%, completion was lower for the post-test in the open covariation questions but was still very high for the post-test big data questions. Consequently, the quantitative analysis was carried out on the big data questions only, for which 39 students out of the 55 completed both pre- and post-test.

Students' covariational reasoning in both the first activity and final pitch were coded and compared in relation to known characteristics of association: trend, scatter and outliers (Wild & Seber, 2000) to identify changes in reasoning. For this purpose, we created additional categories relating to known characteristics of the trend: shape, slope, direction, strength and cluster. This analysis is further explained and elaborated elsewhere (Gil & Gibbs, submitted). The analysis included triangulation between the two researchers when needed.

The program and study were conducted within regular class periods (72 minutes) over three weeks using an interactive technology-enhanced inquiry-based curriculum. The researchers served as designers, teachers and researchers with the support and assistance of the class teacher.

RESULTS

In this section, findings are presented in two parts: learning about big data in the IOLS (activity 2) and an example of covariational reasoning in the context of creating social impact (activities 4-5).

Learning about big data in IOLS

Enabled by the class design, the technology used, and big data resources, findings from the quantitative and qualitative analyses of the pre- and post-tests and qualitative analyses of postings on the knowledge board suggest significant knowledge gains about big data.

Question 7 on the pre- and post-tests measured students' familiarity with Big Data. 87.2% of the 39 students that completed both tests reported greater familiarity with big data, a concept that was new to many of them. A Wilcoxon signed-rank test on a 5-point Likert scale ranging from

1= 'not at all' to 5 = 'very familiar' indicated that post-test responses were statistically significantly higher than pre-test responses (z=595, p<0.001).

On the pre- and post-tests, students were also asked to describe big data. Their responses were examined for relating to the five V's of big data to give a more objective view of their knowledge gains. Significant improvements were found for volume, variety and value. Students increasingly related to value by mentioning applications, examples and uses of big data such as discovering trends and correlation in a variety of settings. Students related very little to velocity and veracity aspects of big data both in the pre & post-test.

To illustrate students' understanding of big data, Table 2 shows some examples from their contributions to the knowledge board. These are responses to the question 'What did I learn about big data?' the final question in all four stations in the 'Big data Interactive' activity. These examples were chosen as some of the most interesting insights in the station and to illustrate student's emerging understanding of the value of big data. In Station 3, the students' emergence of the local versus global view is evident in the recognition that big data allows, in the case of global forest fire data, to help predict from one area to another. In Station 2, the students also refer to the global view in their perception that big data can be used to create arguments for infrastructure in preparation for control of the spread of disease. In Station 1, students in Group 6 commented that data from social media, collected without design and without apparent social purpose, can be used to investigate worldwide trends with some potential economic benefits. In Station 4, the students recognized some of the ethical issues that are prevalent in the collection of big data and the corresponding need to protect information, a topic that was prevalent in the news in Canada at the time of the study.

Table 2: Examples from groups' insights about big data from the four stations (Class 1)

Station	What did you learn about big data?
3 Ecology	*Instead of just looking at a single country/location that has forest fires and trying to find the causes and predict when the next one will happen, we can analyze data on a more global scale to determine the larger trends and how fires in one place can help us predict or be more ready for a forest fire in another area.* (Group 6)
2 Health	*- It can help society prepare for large sources of disease and help prevent spreading from individual to individual.* *- It can also help societal infrastructure deal with sources of disease in a timely fashion to mitigate the effect of disaster.* (Group 7)
1 Social network	*Even social media can be used to investigate larger trends all over the world, and even whom people are following. What may seem as just a social media site can be analyzed to benefit businesses.* (Group 6)
4 Surveillance and privacy	*Big data is very versatile and can affect everyone. Big data can also include personal information that must be protected. Therefore big data could often result in ethical issues.* (Group 3)

Thus, findings from both quantitative and qualitative analyses provide evidence for the contribution of the design, enabled by big data resources, to facilitate learning about big data in a way that promoted students' understanding of the impact on society of modern sources of data coupled with statistical reasoning.

Covariational reasoning with social impact

As mentioned in the Methods section, activities 4-5 were designed to promote covariational reasoning in the context of a group investigation of the Wellbeing Toronto data. Students explored the data visually using spatial display tools on the website and additional visualization and statistical analysis tools in iNZight. They used their insights about the data to support a pitch suggesting changes needed in their city.

A brief analysis of 20 randomly selected pre-post tests indicated no progress in the students' covariational reasoning. This might be attributed, at least in part, to students' apparent unwillingness to write at length during the post-test, which was required in this part of the test. Qualitative analysis of a few groups of students showed progress from the initial to final activities in the unit, measured by the coding of their presentations according to a covariational coding scheme (elaborated in Gil & Gibbs, submitted). In the first activity, one of the groups did not present even bivariate understanding, relating very locally to a certain area in the Gapminder graph (discussing a local rise in childbirth after the great depression in the US). However, in the final pitch, a group that included one of the students from this group was identified by their peers as giving the best pitch. Her team prepared background information about cervical cancer and its prevalence in society. They presented visualization, both spatially and in scatterplots, and results of analyses carried out using iNZight, reporting on the positive relation between cervical screening and house prices in Toronto: higher neighborhood incidence of cancer screening is associated with higher prices. They then suggested an explanation and a practical application - to lower the price of cervical cancer checks. In their pitch they used concepts relating to covariational reasoning such as trend, scatter and outliers and related to the direction and shape of the trend. Thus, although not taking the relative neighborhood size into account as suggested in the class feedback, their pitch demonstrated a much more advanced stage of covariational reasoning then the presentation at the beginning of the unit.

CONCLUSION

Introducing big data and reasoning about multivariate data at the secondary school level addresses some of the challenges presented by the "data revolution" (Ridgway, 2015). Moreover, using real data to investigate questions of global and municipal significance is of paramount importance to secondary school students.

This study presents the design and learning gains of a technology-supported instructional unit using an innovative learning trajectory and an in-class design that requires only commonly available platforms and statistics tools. The unit was designed to expose secondary students to big data and to promote covariational reasoning using multivariate big data and data sets with some of the characteristics of big data. Data, resources, and unit activities were designed to motivate students through the societal impacts of the analyses and resulting conclusions. The study provides evidence for gains of knowledge and insight about big data from both quantitative and qualitative analyses, and advancement in covariational reasoning from qualitative analyses.

This study might also have implications beyond secondary school, suggesting the design of a learning trajectory for the introduction of big data to post-secondary students that can be used at any level, since the barrier of prerequisite knowledge of advanced computational and statistical methods has been removed, and in a variety of disciplines, because of its emphasis on the value of big data in society.

REFERENCES

Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., & Widom, J. (2012). Challenges and opportunities with big data. *A community white paper*.

Batanero, C., Estepa, A., Godino, J. D. (1997). Evolution of students' understanding of statistical association in a computer-based teaching environment. In J. B. Garfield & G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics: Proceedings of the* 1996 *IASE Round Table Conference* (pp. 191-205). Voorburg, The Netherlands: International Statistical Institute.

Beyer, M. A. & Laney, D. (2012). *The importance of big data: A definition*. Stamford, CT: Gartner.

Casey, S. A. (2015). Examining Student Conceptions of Covariation: A Focus on the Line of Best Fit. *Journal of Statistics Education*, *23*(1).

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review*, *1*(2), 293-314.

Franke, B., Plante, J.-F., Roscher, R., Lee, A., Smyth, C., Hatefi, A., Chen, F., Gil, E., Schwing, A., Selvitella, A., Hoffman, M. M., Grosse, R., Hendricks, D., & Reid. (forthcoming). Statistical Inference, Learning and Models in Big Data. *International Statistical Review.*

Fraenkel, J. R., & Wallen, N. E. (2009). The nature of qualitative research. *How to design and evaluate research in education, seventh edition. Boston: McGraw-Hill*, *420*.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning connecting research and teaching practice*. Springer.

Gil, E., & Gibbs, A. (2015). *Big data in Math for Data Management – activities and teachers' guide*, The Fields Institute for Research in Mathematical Sciences and The University of Toronto [draft].

Gil, E., & Gibbs, A. (Submitted). Promoting modeling of covariational reasoning among secondary school students in the context of big data. *Statistics Education Research Journal*.

Gil, E., & Slotta, J. D. (2015, June). Knowledge Community and Inquiry about big data among high school students with Interactive Orchestrated Learning Space. Proceedings of the *Eleventh International Conference on Computer Supported Collaborative Learning* (CSCL, June 2015), Gothenburg, Sweden: The International Society of the Learning Sciences.

Gould, R. (2010). Statistics and the modern student. *International Statistical Review*, 78(2), 297–315.

Lehrer, R., & Schauble, L. (2010). What kind of explanation is a model? In M.K. Stein, L. Kucan (Eds.), *Instructional explanations in the disciplines* (pp. 9-22). Springer US.

Moritz, J. (2004). Reasoning about covariation. In D. Ben-Zvi and J. Garfield (Eds.). *The challenge of developing statistical literacy, reasoning and thinking* (pp. 227-255). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Philip, T. M., Schuler-Brown, S., & Way, W. (2013). A framework for learning about big data with mobile technologies for democratic participation: Possibilities, limitations, and unanticipated obstacles. *Technology, Knowledge and Learning*, *18*(3), 103-120.

Ridgway, J. (2015). Implications of the Data Revolution for Statistics Education. *International Statistical Review*, Doi:10.1111/insr.12110.

Slotta, J. D. (2010). Evolving the classrooms of the future: The interplay of pedagogy, technology and community. In K. Makital-Siegl, F. Kaplan, Z. J., & F. F. (Eds.), *Classroom of the Future: Orchestrating collaborative spaces* (pp. 215–242). Rotterdam: Sense.

Thirunarayan, K., & Sheth, A. (2013, November). Semantics-empowered approaches to big data processing for physical-cyber-social applications. In*Proc. AAAI 2013 Fall Symp. Semantics for Big Data*.

Watkins, A. E., Scheaffer, R. L., & Cobb, G. W. (2004). Statistics in action: Understanding a world of data. Emeryville, CA: Key Curriculum Press.

Wild, C. J., & Seber, G. A. (2000). Chance Encounters: A First Course in Data Analysis and Inference. John Wiley & Sons Inc.