MOBILIZE: A DATA SCIENCE CURRICULUM FOR 16-YEAR-OLD STUDENTS

Robert Gould and the Mobilize Team[i]
Department of Statistics, UCLA
Los Angeles, CA 90095, USA
rgould@stat.ucla.edu

The Mobilize Project was funded by the U.S. National Science Foundation in 2011 to develop "computational thinking" in secondary math and science classrooms through the use of participatory sensing, a data collection paradigm. This evolved into creating a year-long course Introduction to Data Science, which prepares students to think critically and constructively with data of many types and forms using the statistical software R. This course is currently taught in roughly 50 classrooms in a large urban school district in which a majority of students are below poverty levels. This talk will provide an overview of this curriculum and discuss the successes and the on-going challenges to developing "computational thinking with data".

THE NEED FOR DATA SCIENCE

Data lie at the heart of statistics. In the United States, at the secondary level, statistical inference is also at the heart of statistics, and so the data used in statistics curricula are selected in order to help students to learn statistical inference. (For instance, the Guidelines for Assessment and Instruction in Statistics Education K-12 (GAISE K-12) describe three developmental levels that culminate in statistical inference (Franklin 2005).) Data that support statistical inference are often collected from random samples (or are assumed to be thus collected), or through experiments employing random assignment, and occasionally through observational studies. The datasets themselves are often small in size and have a limited number of variables; in many textbooks, only the variables needed to solve a problem are provided. Presumably, one reason for this is that students are meant to focus on the subtle and difficult conceptual underpinnings of statistical inference, and not worry too much about managing the data themselves.

Today's students live in a world surrounded by data, and only rarely are these data of the neat, spreadsheet-like, ready-for-inference variety that we show them in the classroom. The data that students encounter are often complexly structured, non-randomly sampled, and not part of a formal experiment or study. Exposing students to these "every day" data types is necessary not only to motivate them through exposure to data concerning to issues of relevance, but also to prepare them for working and living in a society that is shpaed by data in complex and sometimes unwelcome ways. (For example, see the many examples provided in O'Neill (2017)).

THE MOBILIZE PROJECT

The Mobilize Project, funded by the National Science Foundation in 2010, is a partnership between several academic units within the University of California, Los Angeles (UCLA) and the Los Angeles Unified School District (LAUSD). The UCLA departments of computer science and statistics, and Center X – a unit within the Graduate School of Education and Information Sciences that oversees teacher preparation and professional development – teamed with LAUSD to explore ways in which the data collection paradigm known as participatory sensing could be used to bolster STEM (Science, Technology, Engineering and Math) in high school classrooms.

Participatory Sensing

The central notion to participatory sensing is that a device almost all of us carry with us, the smart phone, is "a special and unprecedented tool for engaging participants in sensing their local environment" (Goldman, et. al 2008). Participatory sensing attempts to create communities united around the common goal of collecting, sharing, and analyzing data for a shared purpose. (Burke, et. al 2006). The Mobilize Project created an extensive technology suite to implement participatory sensing in the high school math and science classrooms. This suite includes apps that can be downloaded on smart phones, a system that ensures data transmitted from the apps are secure and viewable only when shared by the collector and then only by students enrolled in the class, management tools for the teacher to monitor and manage data activity, and a "dashboard" visualization tool (Tangmunarunkit et. al 2015).

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

A demo is available at https://sandbox.mobiliznngcs.org/#demo.) Figure 1 shows a snapshot of one component of the dashboard called the "trashboard". Students (and in the case of this particular data set, some teachers) collected data every time they discarded an object. The dashboard is interactive; clicking on any one component updates the display to show only data whose values match those selected. For example, to generate figure 1, we clicked on "landfill" under the "What type" display so that we now view data only for items that were designated to be recyclable. The map shows where these data were recorded; integers inside of circles indicate the number of data points at that location, and can be "zoomed" in to see more detail. We see that a slight majority of recyclable items were deposited into recycling bins, and that often in these cases there were no recycling bins visible ("# Recycle Bins").

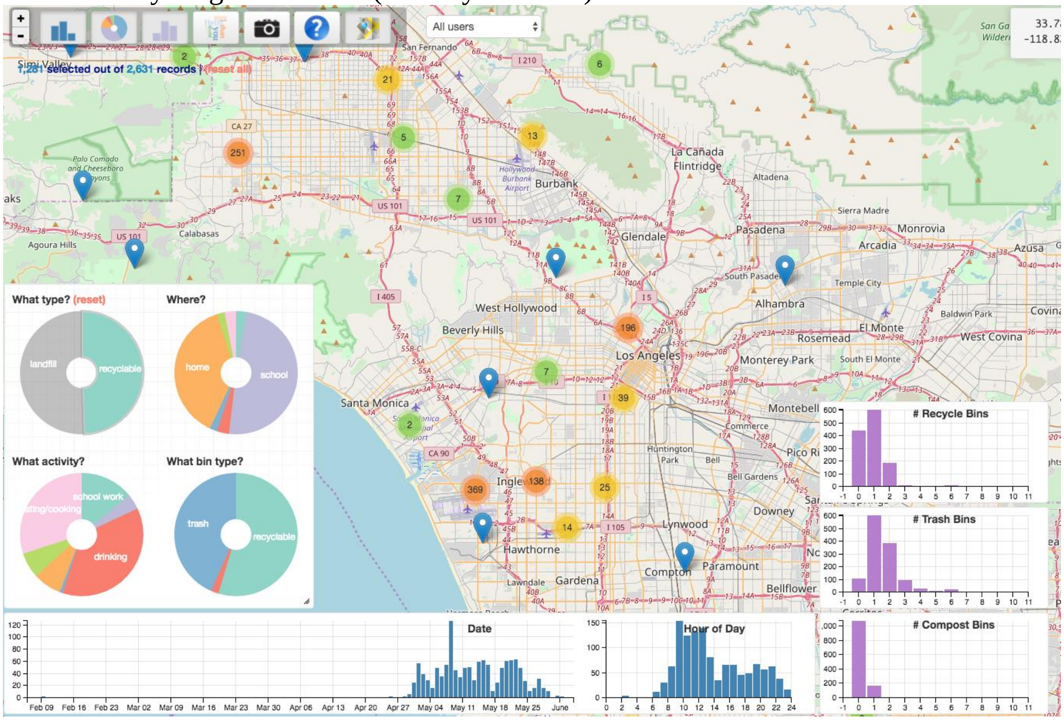

Figure 1. The "trashboard" provides an interactive multivariate view of the data collected through a participatory sensing campaign in which students examined their patterns of waste disposal.

INTRODUCTION TO DATA SCIENCE

The Introduction to Data Science (IDS) course was developed because the Mobilize team recognized limitations to implementing participatory sensing within math and science classes. One problem noted was that participatory sensing is time consuming. Time is needed to install and learn the technology (time for both students and teachers); at least several days are required to collect data; time is needed to teach students how to analyze data. (Without instruction in data analysis, we found that students' interpretations of the dashboard displays were often quite superficial. An illustrative example would be along the lines of "We noticed that the '1' bin was the tallest in the 'Number Recycle Bins' graph.")

The natural solution, then, was to create a year-long class that would allow students to dive deep into data and learn not just data analysis, but also the role that data play in our culture and daily lives. LAUSD, our partnering school district, is the nation's second largest district (with roughly 650,00 students) and serves a population that is largely impoverished, has a high percentage of English learners, and is underrepresented in the sciences and mathematics. With LAUSD's assistance, the course was approved by the California to partially satisfy the mathematics requirement for admission to the state's two public university systems (the Universty of California and the California State University). As a consequence, students can take IDS instead of Intermediate Algebra, a course that had historically suffered from high failure rates that fell disproportionately on underrepresented minority groups (Burdman 2015).

IDS Design

IDS lessons come in two flavors. Most meetings are "regular" class sessions in which students engage in active, inquiry-oriented lessons in order to develop conceptual understanding and learn fundamental terms. Roughly once per week, students meet in a computer lab and engage in data analysis exercises. These labs teach students to "code" in the R (R Core Team 2017) language via the Rstudio (Rstudio Team 2015) interface. A package, mobilizR, was written to unify the R syntax and to provide "wrapper" functions for common analyses such as creating word clouds or drawing maps. (https://github.com/mobilizingcs/mobilizr). The mobilizR package was based on mosaic (Pruim, Kaplan & Horton 2015), a package designed to help undergraduate students implement R in statistics, math, and science courses.

While several of the topics in IDS are rarely, if ever, taught in introductory level statistics courses (see *Topics* below), much of the content would not be out of place in a typical high school statistics course, although in IDS these topics are often augmented with a computational component. For example, while IDS covers regression, it's focus is not on inference or on understanding least squares, but is primarily on using regression to make predictions, and on understanding how predictive success is measured, and understanding how predictive success may be improved.

For those topics that are more traditional, IDS relies on the American Statistical Association's Guidelines for Assessment and Instruction in Statistics Education (GAISE) K-12 Report (Franklin et. al 2007) and situates the content firmly in levels A and B (roughly, beginning and intermediate). As a consequence, while there is some statistical inference, it is at the informal level (Makar & Rubin 2009). For example, students learn to perform permutation tests to generate null sampling distributions, but do not learn the formal vocabulary. Instead, the goal is to develop an understanding of what it means for outcomes to happen "by chance" and to use the generated distribution to express whether or not an outcome seems surprising.

A central component of the IDS curriculum is what we've called the Data Cycle (see Figure 2). The Data Cycle is a graphical representation of the four-step investigation process as defined by GAISE K-12, which is itself based on the more detailed PPDAC cycle of Wild & Pfannkuch (1999). The Data Cycle modifies these four steps to adjust for a more data-centered course by replacing the "collect data" step with the more generic and data science-relevant "consider data" step.

The Data Cycle was introduced into the curriculum to correct issues identified through student work from the original Mobilize participatory sensing exercises. The Data Cycle is intended to serve as a reminder to both teachers and students that a statistical investigation consists of more than simply typing the correct code and printing out a graph. It reminds students that an analysis serves a purpose: to answer questions. Posing productive statistical questions turned out to be a challenge for both teachers and students. We found that many of the mathematics teachers we worked with were unfamiliar with the notion of posing questions that could be answered with data (i.e., "statistical investigation questions", using Arnold's (2013) terminology). In one case study, we saw evidence that suggests that teachers' failure to phrase productive questions early in an investigation may be detrimental to the investigation (Gould, Bargagliotti, Johnson 2017). Therefore, a significant amount of professional development time is devoted to this topic.

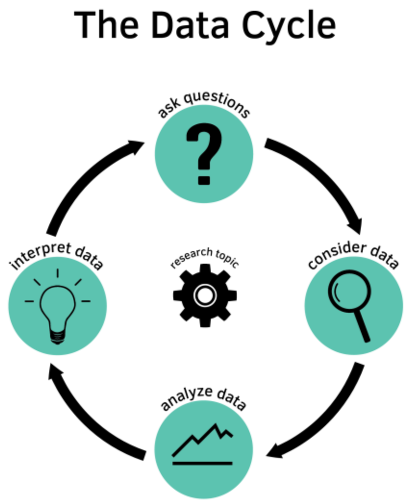

Figure 2. The Data Cycle

Data

The IDS curriculum engages students in four participatory sensing "campaigns" and culminates in a classroom-designed campaign. The first campaign, Food Habits, asks them to record data about the food they eat during one week. The Time Use asks them about whatever activity they are doing when a pre-set alarm on their phone rings. A third asks about their stress level (also at arbitrarily chosen moments of the day), and a third collects data around water use (a big topic in Southern California). Finally, the class also designs its own campaign.

Data generated by these campaigns are similar in many ways to the data we now routinely encounter in our every day lives. They are from non-random samples, of many types (including locations, dates, and images), and complexly structured. While we've found that these exercises can be quite rich, particularly in helping students understand data collection issues, data privacy issues, and understanding how data are represented in graphs, the actual data collected can be sparse and does not always provide the opportunity for rich findings. For this reason, the curriculum is richly seasoned with a wide number of additional data sets, many pulled from open data (the American Time Use Survey (https://www.bls.gov/tus/) , the Youth Behavior At Risk Survey (https://www.cdc.gov/healthyyouth/data/yrbs/data.htm) and others scraped from various sites (movie rating sites, for example). Sometimes, the participatory sensing campaign prepares students to better understand the other data sets used in the curriculum. For example, students engage in the Time Use campaign to prepare them to analyze the American Time Use Survey data.

Topics

The course is organized into four units, each roughly nine weeks long. Each unit has at least one "capstone" assignment intended to provide an opportunity to synthesize the various concepts and techniques introduced. The general themes of the units are, in order, "Focus on Data" (data organization, data types, exploratory/descriptive statistics), "Informal inference", (probability and randomization testing), "Data Collection" (observational studies, randomized assignment, surveys, sensors, participatory sensing, scraping data from HTML tables), "Predictions" (multivariate regression). Further details of the curriculum are provided at https://mobilizingcs.org. A few topics that we feel are unusual for a high school introductory statistics course include: data and privacy, writing questions for surveys, scraping data from html tables on the internet, classification and regression trees, k-means.

CONCLUSION

The IDS course is, at the time of writing, being taught by 44 teachers in the Southern California area, including 17 teachers from districts other than LAUSD. Interest continues to grow, but there are real challenges. First among them is teacher preparation. The Mobilize Project has used the NSF-funding to support a substantial series of professional development sessions. For all teachers, the technology is new and requires learning R, a notoriously difficult statistical analysis language. Many teachers are unfamiliar with data analysis and fundamental concepts of statistics, and so they must learn those. And many teachers are unfamiliar with, or out of practice with, an inquiry-based pedagogy.

Evaluation has been another interesting issue. For evaluation purposes, we administered the Levels of Conceptual Understanding of Statistics (LOCUS) Beginning/Intermediate instrument pre- and post (Jacobbe, Case, and Whitaker & Foti 2014). While this was a good fit for the statistical content of the course, it fails to capture the computational thinking that we desired to develop and doesn't assess statistical thinking when it is done through a language such as R. And, while computational thinking assessments exist (for example, see https://www.sri.com/work/projects/principled-assessment-computational-thinking-pact), they do not assess computational thinking with data, and so are not good measures of learning in this course.

A third area of challenge is the computational labs, which strive to simultaneously teach the R language and data analysis. Our external evaluators, embedded researcher, and classroom teachers report that while the classroom activities work well at including all students, once students enter the computer lab, engagement declines. A common scenario is that one or two students who "get it" provide answers to those who don't. There have also been cases in which students working at the computer discover novel solutions or pose interesting questions, but their discoveries and ideas are not communicated to the rest of the classroom (Olivares Pasillas 2017). To improve this situation, the Mobilize Team is revising the labs to include and emphasize "pair programming", a programming practice that has been adapted for classroom instruction (McDowell, Werner, Bullock, and Fernald 2002).

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant Number 0962919. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

IDS is extremely grateful for its teacher leaders who have shaped the curriculum into what it is, and who continue to prepare future data science teachers: Pamela Amaya, Monica Casella’s, Heidi Estevez, Joy Min, Robert Montgomery, Roberta Ross, Carol Sailor, and Alma Villegas-Torres. We are also, of course, extremely grateful for former principal investigators who got the project off the ground: Deborah Estrin (Computer Science), Mark Hansen (Statistics), Jane Margolis (Education), Jody Priselac (Education), and Todd Ullah (LAUSD).

REFERENCES

Burdman, P. (2015). *Degrees of freedom: Diversifying math requirements for college readiness and graduation (Report 1 of a 3-part series)*. Oakland, CA: LearningWorks and Policy Analysis for California Education, PACE. Retrieved from Institute of Education Sciences ERIC collection. (ERIC Number: ED564291).

Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., Srivastava, M.B., (2006). Participatory Sensing. *WSW'06 at SenSys*., Boulder, CO

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report*. Alexandria, VA: American Statistical Association.

Goldman, J., Shilton, K., Burke, J., Estrin, D., Hansen, M., Ramanathan, N., Reddy, S., Samanta, V., Srivastiva, M. (2008). Participatory Sensing: A citizen-powered approach to illuminating the patterns that shape our world. Retrieved from http://escholarship.org/uc/item/19h777qd

Gould, R., Bargagliotti, A., Johnson T (2017). An Analysis of Secondary Teachers' Reasoning with Participatory Sensing Data. *Statistics Education Research Journal*, *16*(2). Retrieved from https://iase-web.org/documents/SERJ/SERJ16(2)_Gould.pdf

Jacobbe, T., Case, C., Whitaker, D., & Foti, S. (2014). Establishing the content validity of the LOCUS assessments through evidence centered design In K. Makar & R. Gould (Eds.) *Proceedings of the 9th International Conference on Teaching Statistics*.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal*, *8*(1), 82-105. Retrieved from https://www.stat.auckland.ac.nz/~iase/serj/SERJ8(1).pdf#page=85

McDowell, C., Werner, L, Bullock, H., and Fernald, J. (2002). The effects of pair-programming on performance in introductory programming course. *Sigcse '02 Proceedings of the 33rd SIGCSE technical symposium on Computer science education*, (pp. 38-42), Cincinnati, KY. https://doi.org/10.1145/563517.563353

Olivares Pasillas, Maria (2017). *Toward Critical Data Scientific Literacy: An Intersectional Analysis of the Development of Student Identities in an Introduction to Data Science Course*. Doctoral dissertation, UCLA, 2017.

Pruim, R., Kaplan, D., Horton, N., Creativity, M., & Minimal, R. (2015). Mosaic: Project MOSAIC statistics and mathematics teaching utilities. R package version 0.10.0. [Computer Software.]. Retrieved from https://cran.r-project.org/web/packages/mosaic/index.html

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc. [computer software]. Boston, MA. Retrieved from http://www.rstudio.com/

Tangmunarunkit, H., Hsieh, C.K., Longstaff, B., Nolen, S., Jenkins, J., Ketcham, C., Selsky, J., Alquaddoomi, F., George, D., Kang, J. & Khalapyan, Z. (2015). Ohmage: A general and extensible end-to-end participatory sensing platform. *ACM Transactions on Intelligent Systems and Technology (TIST), 6*(3), 38.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223-248.

[i] Suyenn Monaca-Machado, LeeAnn Trusela, James Monteux, Terri Johnson, Amelia McNamara, Jeroen Ooms, Jane Margolis, Jody Priselac, Joanna Goode, Derrick Chao, Hongsuda Tangmunarunkit, Steve Nolen, Kapeel Sable, Shuhao Wu, Maria Olivares