TEACHING DATA SCIENCE TO TEENAGERS

Amelia McNamara[1] and Mark Hansen[2]
[1]Department of Statistics, University of California, Los Angeles, CA, USA
[2]Department of Journalism, Columbia University, New York, USA
amelia.mcnamara@stat.ucla.edu

Teaching data science to secondary school students might seem outlandish, given how few programs of instruction currently exist, even at the post-graduate level. However, teenagers are surrounded by data, and, in one form or another, view and manage data every day. Many teenagers carry smart phones, which already automatically store, retrieve and analyze data, but which can also be used to collect data. We will discuss challenges faced by the Mobilize project, an NSF-funded project that utilizes participatory sensing to teach data science to secondary school students (ages 15-17). In particular, we'll talk about the challenges of including statistical computing as a natural and integrated part of the curriculum.

CONTEXT OF MOBILIZE

The Mobilize Project is a National Science Foundation grant focused on bringing data science, participatory sensing, and computational thinking to high school students. Through Mobilize, we have been exploring the spectrum of places where one can frame and study data. From data as computer science, to data as its own science, to data as a site for creativity and storytelling, we have been playing with the interactions one can have with data. Along the way, we have come to some realizations about the difficulties of teaching data to teenagers, working in multidisciplinary groups, and finding the "right" technology.

Our interpretation of data science is that it requires a mixture of computational thinking and statistical thinking, much like Robert Gould's 2010 suggestions for statistics and the modern student. Although he was largely addressing the undergraduate curriculum, Gould's suggestions apply to high school curriculum as well. For example, his goals include improving student access to data and integrating computation into the curriculum (Gould, 2010). Mobilize attempts to address both of these goals, particularly by the use of participatory sensing as a data collection method. Participatory sensing allows humans to become active observers of their everyday life, and record these observations as data.

The technical team on the grant has created an app for recording data (Tangmunarunkit, et al., 2013). The app makes it simple to specify survey questions and deploy them on any internet-enabled phone or through the web browser on a personal computer. The data are time- and location-tagged by the devices. So far, our campaigns have been sociological (e.g. sleep patterns, advertising in neighborhoods, snacking habits). Campaigns utilize the participatory sensing platform around a common investigative theme, and include all aspects of the data "life cycle": collection, hypothesis generation, display, exploration, analysis, etc.

Because teenagers are especially likely to be involved in participatory cultures (Jenkins, Clinton, Purushotma, Robison, & Weigel, 2009), the use of smart phones as data collection devices allows us to leverage an existing behavior and allow students to become content creators of data and analysis products. So far, the implicit interest in learning more about one's self has given a good payoff for doing data analysis on the participatory sensing data.

TOOLS, TRAINING, AND ENGAGEMENT

However, while our assumptions about participatory sensing seem to be somewhat validated, many of our other assumptions have not been. One area where this has led to much iteration is our choice of data analysis tool. From R to Deducer to RStudio, we have tried to use free and open source programs to reduce the financial burden on schools that want to participate in our curricula. All three of these tools rely to some degree on the programming language R, which is free and open source, in addition to being the programming language of choice of professional statisticians. By using a real language, we're not limited by features implemented in a user-focused program for learning statistics, and students can go as far into programming as they want. However, with each tool we have tried, there have been struggles with the limited time available

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

for teacher training and frustrations with the tools in general. In addition to the deeper analysis tools mentioned above, our technical team has developed several dashboard applications for quicker analysis, which we use in separate contexts.

Pre-pilot: R

During our pre-pilot year, the tool of choice was R, in the standard R Graphical User Interface (GUI). This proved to be too much material to teach and learn during the short period of professional development, especially given the relatively complex tasks and data types we hoped to cover. As a result, teachers were very frustrated with the tool.

Year 1: Deducer

For the 2011-2012 school year, we moved to Deducer, a menu-driven GUI for R (Fellows, 2012). Deducer was slightly easier for the teachers to pick up, but its menu driven paradigm proved difficult for a number of reasons. First, it was very hard to document, because each step was an additional click. Then, when teachers wanted to go beyond the features already supported, the development time necessary to implement the new feature was very limiting. Finally, Deducer is standalone software, which means it must be installed on each computer individually. Anyone who has spent time doing technical support for computers can probably picture the multitude of problems this presented.

Moving Forward: R in RStudio

Learning from the experience of Deducer, we chose to move back toward R in the 2012-2013 school year and moving forward. However, instead of using the standard R GUI (a desktop application) we chose to use the server-side version of RStudio (RStudio, 2014). RStudio offers many great features, like code completion, file management, and code history. For introductory college users, we have found that the support RStudio provides makes it much easier to learn R. RStudio can run as a desktop application, but by installing it on a server, we were able to manage all installations from a central location, and provide quick bug fixes to all users at once. Access was through the browser, so students could access it from any computer with internet access. Files were stored on the server, so they could not be lost to nightly hard drive wipes. This eliminated the technical difficulties we encountered with Deducer.

Besides using RStudio, we made several modifications to smooth out the implementation. We beefed up our documentation to include demonstration videos, a wiki page, and several overview pdf documents. This is of course in addition to the built-in R documentation, and our existing curricular resources. We also created an additional package to simplify the amount of code necessary to complete the curricular tasks such as map-making and text analysis (McNamara & Molyneux, 2013). This paradigm has proven to be the most popular with teachers, although the challenges of using R are still present.

Dashboards

Our goal is to get students doing their own authentic analysis as quickly as possible, but it is also useful to have a quick payoff between collecting data and seeing the results. To that purpose, Jeroen Ooms has developed several tools for simple visualization of the student-collected data. The first is an embedded visualization tool within our data access platform, Ohmage (Tangmunarunkit et al., 2013). It allows students to select one or two variables to analyze, and then produces an appropriate type of plot. Ooms has also produced a standalone data visualization tool that is currently custom-tailored to the data students are producing (Tangmunarunkit et al., 2013). The standalone tool is an interactive map with associated time series plots, bar charts, donut charts, and a view of the raw data (including the photographs taken by students). It allows users to engage with the data and ask more complex questions than the embedded visualization tool, and even do a form of subsetting, but it obviously cannot perform every analysis one could think of.

These tools have been invaluable for teachers and students participating in our projects. They allow students to begin working with data without having to learn many skills first. For this reason, they are an excellent first foray into the world of data analysis. They allow students to start asking questions of the data, and answering them; but they prescribe what a person can do with

them. There is limited support for subsetting, for example, and one can only browse the raw data, not manipulate it. In professional development sessions where we have only introduced teachers to the dashboards, they do quickly get frustrated by the limitations of the tool and begin asking how they can answer additional questions. This provides a clear segue into standalone analysis platforms.

CONCLUSION

The Mobilize Project seeks to engage high school students in authentic data science experiences. Students learn exploratory data analysis driven by questions they can answer about themselves, using tools that professional data scientists and statisticians use.

However, the tool itself has been a struggle to find, and training teachers to teach data science is difficult. They may have little to no experience with statistics or computing, and learning both at once is simultaneously motivating and frustrating. It is clear that longer professional development would help with this issue, but resources of time and money are limiting. The question is, what would help alleviate this issue? Members of the Mobilize team are conceptualizing and building new tools to help step learners (both teachers and students) from an exploratory setting much like our dashboard into a more computational setting like R in RStudio, but it is clear that even the best tool cannot solve all the issues present. More creativity is needed to determine what supports would be beneficial.

REFERENCES

Burke, J., Estrin, D., Hansen, M., Parker , A., Ramanathan, N., Reddy, S., & Srivastava, M. B. (2006). Participatory Sensing. *UCLA: Center for Embedded Network Sensing.*

Fellows, I. (2012). Deducer: A Data Analysis GUI for R. *Journal of Statistical Software. 49*(8), 1-15.

Gould, Robert. (2010). Statistics and the Modern Student. *International Statistical Review, 78*(2), 297-315.

Jenkins, H., Clinton, K., Purushotma, R., Robison, A.J., & Weigel, M. (2009). Confronting the Challenges of Participatory Culture: Media Education for the 21st Century. *The MIT Press*, 2009.

McNamara, A., & Molyneux, J. (2013). MobilizeSimple. GitHub Respository, https://github.com/mobilizingcs/MobilizeSimple

R Core Team (2013). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing.* Vienna, Austria. http://www.R-project.org

RStudio (2014). RStudio: Integrated Development Environment for R. *RStudio.* Boston, MA. http://www.rstudio.com/

Tangmunarunkit, H., Hsieh, C. K., Jenkins, J., Ketcham, C., Selsky, J., Alquaddoomi, F., … Estrin, D. (2013). Ohmage: A General and Extensible End-to-End Participatory Sensing Platform. *ACM Transactions on Intelligent Systems and Technology.*