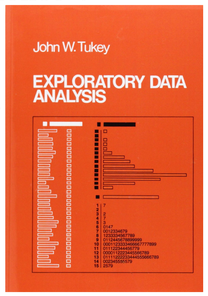
Tools for Teaching Data Science

Amelia McNamara[1] and Mark Hansen[2]
[1]University of California--Los Angeles
[2]Columbia University

What do we mean by "data science"?

Essentially, statistics. But traditional statistics teaching has focused on hand-computation, formulas, and toy data sets. We want students to be engaged fully with data-- using exploratory data analysis (Tukey, 1977), re-randomization, and visualization to explore their data. Ideally, they will be involved in the entire process, from data collection to presentation.



What level of teaching are we talking about?

Most of our suggestions are applicable across all levels, but they are more likely to have been implemented at the graduate and undergraduate level, particularly in courses designated as statistics. We envision the same tools and techniques becoming available for students in many disciplines and fields. For example, many of our recommendations come from our experience on the Mobilize project, which brings data science into high school classrooms. Additionally, we realize that there are a growing number of fields that use data and need appropriate tools. For example, journalists are beginning to embrace "data journalism" but often do not know where to begin.

Mobilize 2013 professional development teachers

Tools for teaching data science should be:

Tools for doing data science. We believe that if a student is going to make the investment of time to learn data science, they should be able to use the tool they learned on in the real world. Examples of tools that don't fit this criteria include TinkerPlots and Fathom.

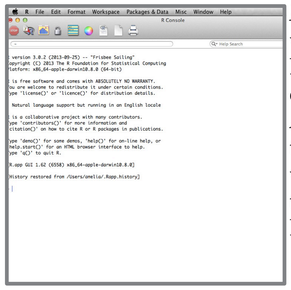
Free (and open source). Especially given our experience with Mobilize, we are sensitive to the prohibitive cost of many tools used for doing data science. Especially when teaching high school students, it's important to have a free tool, to avoid putting financial pressure on schools. We also believe in the importance of open source software, which often goes along with being free. Examples of tools that don't fit this criteria include SAS, SPSS and Stata.

Easy to get started in. Many tools for doing data science are overly complex for learners, at least at first. Instead, the tool should have a "low threshold" for beginners (Ioannidou et al, 2011). Examples of tools that don't fit this criteria include Python, R, SAS, SPSS, and Stata.

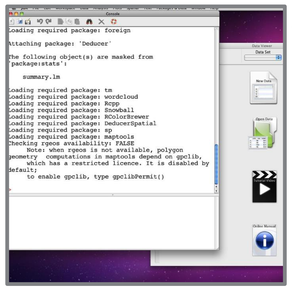
Extensible. This is similar to the "tools for doing data science" criteria above. A tool for learning data science should have a "high ceiling," meaning that it can be extended to do new tasks not already in the system (Ioannidou et al, 2011). Examples of tools that don't fit this criteria include TinkerPlots, Fathom, SAS, SPSS, and Stata.

Case study: mobilize

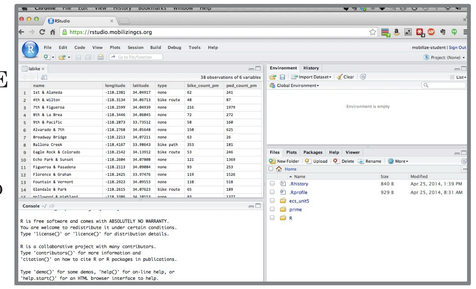
We have been working on the Mobilize project, an NSF-funded grant that develops high school curriculum, software for data collection and analysis, and teacher training, for the last four years, and over iterations of the project we have come to many of the conclusions enumerated here. In chronological order, here is a case study from the Mobilize project.



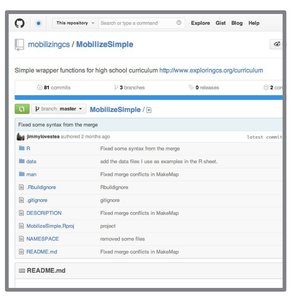
Pre-pilot year: during this year of the grant, we taught preliminary PD to a small group of high school computer science teachers. Our tool of choice was R, in the standard R GUI. Although they taught computer science, the syntax of R overwhelmed the teachers, especially when trying to do tasks like map-making and text analysis. It failed because it did not have a low threshold for beginners.



First grant year: learning from the experience of the pre-pilot, we chose to use the Deducer GUI for R (Fellows, 2012). Deducer provides a menu- and button-driven interface to R. However, this did not necessarily make it easier to use. Each task took a series of clicks through menus and wizards, and documentation was nearly impossible. In addition, the software had many bugs and was difficult to install on school computers. Teachers disliked it even more than the standard R GUI. This tool was first and foremost not a tool for doing data science, it was hard to get started in, and it was almost impossible to extend. When we wanted to program additional features ourselves, we found that we needed the developer's help, and there was no way that a teacher or a student could have built new extensions.



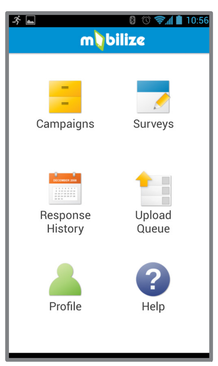
Second grant year: because neither we nor the teachers liked Deducer, we moved back to R in the second grant year. But, we chose to use the IDE RStudio to support learning. RStudio offers many advantages, including code completion, integrated help and visual previews of data (much like an Excel spreadsheet). Additionally, it can be installed on a server, rather than on individual computers. This meant that teachers did not have to worry about installing software, and students could access the tool from any internet-enabled computer.



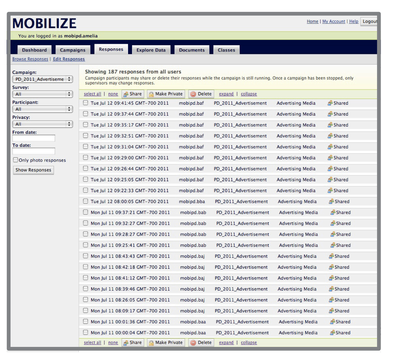
And beyond: In addition to the switch to RStudio, we wrote an additional R package called MobilizeSimple, which wrapped up difficult tasks into simpler helper functions.
Currently, we're having a discussion about which R syntax is the easiest for learning, the model-based syntax, which uses ~, or the indexing operator syntax, which uses $.

```
> plot(labike$bike_count_pm, labike$ped_count_pm)
> xyplot(bike_count_pm~ped_count_pm, data=labike)
>
```

Terms:
GUI: Graphical User Interface
IDE: Integrated Development Environment
PD: Professional Development



Participatory sensing: in addition to the data analysis tool, we train teachers and students on the use of a participatory sensing system called Ohmage. Ohmage allows students to easily deploy data-collection surveys on their mobile phones, and collect time- and location-tagged survey data on issues they care about.
Past surveys using the app have focused on snacking habits and advertising in the students' communities. Data is automatically synched with a server and aggregated over the class. Students can then download and analyze their own data. We have found it is engaging for anyone (but particularly teenagers) to analyze data that they can "see themselves" in.

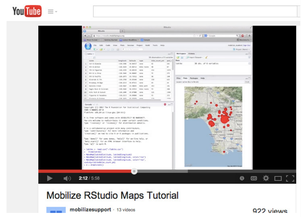


Other than software, what tools are needed?

Teacher training. In our experience on the Mobilize project, one of the major limiting factors is teachers' experience. We want them to be able to teach their students about data science using R, but first, they must feel comfortable with the statistical concepts, interpretation of analytical results, programming syntax, etc. In fact, even the presence of uncertainty (i.e. there isn't one "right answer") can be a source of anxiety. We regularly hold week-long professional development sessions, but a week is not enough.

Curriculum. Again, for teachers to feel comfortable teaching this material, they need scaffolding. Curriculum is a good way to scaffold the process, but it needs to be carefully constructed in order to provide support without removing the spirit of discussion, inquiry, and uncertainty we would like to see.

Documentation.
This goes for all of the above-- documentation is necessary for the elements taught to teachers, of the curriculum, and above all, for the software tool itself. We have tried a variety of methods of documentation, including a wiki (McNamara, 2013), YouTube videos (Mobilize Support, 2013), internal package documentation, and paper handouts.



Data. Obviously, to do data science, you need data. But it is difficult to locate data that is interesting to students, is "appropriate" for the school setting, and has enough substance to provide learning opportunities. Even thinking of appropriate participatory sensing campaigns can be challenging.

Future directions: there is much more to be done in the field of data science education. Specifically, there is a need for better tools that scaffold the flow from easy-to-use introductory functionality to more advanced data analytics. We are working on a project to illustrate some capabilities of such a system, but is just that, an illustration.

The problem of finding good data is also hard. Perhaps we as a community could develop and maintain a list of interesting and appropriate data.

Finally, there is a need for more randomized studies of statistics and data science education techniques. We have some intuition that the model-based R syntax is easier to learn, but have not done any studies yet.

References:

Ian Fellows. "Deducer: A Data Analysis GUI for R." *Journal of Statistical Software*, **49**(8). 2012.

Andi Ioannidou et al. "Computational thinking patterns." In *American Educational Research Association*, April 2011.

Amelia McNamara. "Mobilize Wiki: RStudio." http://wiki.mobilizingcs.org/rstudio, 2013.

Mobilize Support. "RStudio Tutorials." https://www.youtube.com/channel/UCefSNiXgoyPBd_zs9MPElGg, 2013.

John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley Publishing Company, 1977.