Teaching Statement

K. Jarrod Millman

My teaching focuses on developing computational reproducibility, mathematical insight, and statistical rigor.

Introduction

Before graduate school, I spent over a decade helping PhD students, postdoctoral researchers, and faculty analyze neuroscience data (e.g., electrophysiology, imaging, and genomics data). I cofounded the Neuroimaging in Python project, a collection of software packages for the analysis of neuroimaging data in the Python programming language. I helped create the Collaborative Research in Computational Neuroscience website, the central data sharing resource of a joint NSF/NIH program to better integrate theoretical and experimental neuroscience. I have also been extensively involved with developing Python software packages for array (NumPy), image (scikit-image), and graph/network (NetworkX) processing as well as fundamental algorithms for scientific computing (SciPy).

Given my background in scientific and statistical computing, I taught three courses in the statistics department at the start of my PhD. Much of the motivation for my teaching approach is discussed in "Developing Open Source Scientific Practice,"[1] which I cowrote with Fernando Pérez (an assistant professor in Statistics and co-founder of Project Jupyter).

Concepts in Computing with Data (Summer 2014) **Course Website**

Concepts in Computing with Data is one of three core courses for the undergraduate major in statistics and is usually taken after the other two. It is a computationally intensive applied statistics course focused on organizing, manipulating, and visualizing data with a focus on the R statistical computing environment. I covered the following topics: Unix command line, R programming, simulation, random number generation, exploratory data analysis, dimension reduction, clustering, classification, regression, and hypothesis testing.

Computational reproducibility was a central theme of my course. At the start, I introduced version control with Git and GitHub. Then I immediately covered functions and automated unit tests. These three concepts were used extensively throughout the course. All coursework was assigned and submitted using Git. Assignments were presented as unimplemented functions with tests. Grading was automated using separate tests.

I used a microarray data set from the Duke/Potti scandal as a running example. The data set was used in a retracted study that claimed to demonstrate the efficacy of a simple test using microarray genetic analysis for personalised cancer treatment. The data set was simple enough that over the semester, students learned both how to replicate the original analysis as well as identify problems with the data and analysis. To motivate the example at the start of the semester, I had students watch a *60 minutes* investigation of the scandal, titled Deception at Duke, which received a 2012 Peabody Award.

Reproducible and Collaborative Statistical Data Science (Fall 2015) **Course Website**

Reproducible and Collaborative Statistical Data Science is a project-based course that introduces students to reproducible and collaborative statistical research, applied to real scientific data. Working with Philip Stark (chair of Statistics at the time) and building on two pilot courses, I wrote the proposal for an upper division undergraduate (STAT 159) and a graduate (STAT 259) version of the course the previous academic year. The proposal was approved by campus and I taught it the first semester it was offered. It is now offered every fall and the scientific content varies by instructor.

[1] Millman, K. Jarrod, Fernando Pérez. "Developing Open Source Scientific Practice." In V. Stodden, F. Leisch, and R. D. Peng, editors, Implementing Reproducible Research, pages 149–183. Chapman and Hall/CRC, 2014.

My course centered around a semester-long group project involving exploring a published FMRI paper and the accompanying data.[2] During the first third of the course, teams formed, choose a published FMRI paper with publicly-available data, and proposed projects. Two of my colleagues, Matthew Brett and Jean-Baptiste Poline, who were then at the Brain Imaging Center, my teaching assistant Ross Barnowski, and I worked with the teams to iteratively define the scope of their projects, to answer questions, and to provide feedback. Teams worked on their projects for nearly three months.

Statistic MA Capstone Project (Spring 2016) **Course Website**

Statistic MA Capstone Project is a course requirement for students in the MA program in Statistics during the final semester. It is organized around data sets and not around a prespecified set of lecture topics. Data sets vary by instructor.

During the first part of the course, there were two small group projects and one individual project. For the first small group project, the primary data source was Twitter. This project gave them practice using Python for text mining. The primary data source for the second small group project was financial time series data. This project involved predict stock price movements using machine learning. A wireless sensor network, which captured spatial and temporal information (e.g., temperature, humidity) from the microclimate around a coastal redwood tree was the primary data source for the individual project. Working with wireless sensor networks exposed students to messy, incomplete, and inconsistent data.

The second part of the course involved one large group project. The primary data source was mouse behavioral data from the Tecott Lab at UCSF. The lab had recently developed a method for continuous high-resolution behavioral data collection and analysis, which enabled them to observe and study the structure of spontaneous patterns of behavior in the mouse.

Conclusion

From my experience working with neuroscientists and the open source scientific Python community, I learned the importance of computational tools and practice. I saw that researchers were rarely trained in computation, scientific coding, or data management. This made it hard for researchers to document and replicate analyses.

In addition to the courses discussed above, I have taught numerous bootcamp and workshops. I have also been a teaching assistant for undergraduate-level introductions to R, SAS, and Python programming as well as probability and statistics in biology and public health and a graduate-level course on statistical computing covering both programming concepts (e.g., data structures, flow control) and statistical computing concepts (e.g., numerical linear algebra, simulation studies, numerical optimization).

[2] Millman, K. Jarrod, Matthew Brett, Ross Barnowski, and Jean-Baptiste Poline. "Teaching Computational Reproducibility for Neuroimaging." Frontiers in Neuroscience 12 (2018): 727.