INTEGRATING BIG DATA INTO THE SCIENCE CURRICULUM

Daniel Kaplan, Paul Overvoorde, and Elizabeth Shoop
Department of Math, Statistics, and Computer Science,
Macalester College, Saint Paul, Minnesota, USA
kaplan@macalester.edu

The contemporary practice of science widely involves "big data": data with many cases and many variables and often with a distributed structure. The science curriculum, however, does not. Our approach to bringing big data to the science curriculum involves developing computational approaches that are powerful, engaging, and can be learned quickly with little background. We're developing a 10-class-hour course (1 credit hour)—called Data and Computing Fundamentals (DCF)—to get students and faculty involved with big data, and able to tackle problems of realistic complexity.

The contemporary practice of science widely involves "big data": data with many cases and many variables and often with a distributed structure. The science curriculum, however, does not. This is largely because both students and instructors lack the skills needed to deal with data: bringing together data from different sources, wrangling it into a format suitable for analysis, carrying out that analysis and presenting the results graphically. Without skills in big data, students and faculty cannot use it in their science courses; without it being used in science courses, students and faculty don't have the opportunity to develop skills in big data. Science curricula are already crowded, so it's difficult to introduce a new pre-requisite topic for studying science. Indeed, over the last half century, computational topics have barely made it into the foundational courses in science.

Our approach to bringing big data to the science curriculum involves developing computational approaches that are powerful, engaging, and can be learned quickly with little background. We're developing a 10-class-hour course (1 credit hour) --- called Data and Computing Fundamentals (DCF) --- to get students and faculty involved with big data, and able to tackle problems of realistic complexity. The course is short enough to make it feasible to require for all science students.

The approach we're taking starts by developing students' ability to parse scientific graphics, for example by understanding how variables are encoded in glyphs. (A glyph is the basic data-information carrying unit in scientific graphics. Other graphical units are spaces representing variables and guides such as axis scales or outlines in a map.) In this analytic phase, students learn how to map backwards from graphs to the tabular form of data. This is followed by a synthetic stage, where students take sometimes large data in tabular form and construct appropriate glyphs to show. Appropriate software makes this straightforward: the skill and creativity come in constructing the map from variables to the various graphical components and choosing an appropriate mode from a small set of possibilities: glyphs show single cases, glyphs show network connections, glyphs show summary constructions such as density or models.

From there, we move to the "back story" of data. This refers to how the graph-ready data is constructed from the collected data, which may be in a very different form. Central to this part of the story are the methods of data wrangling and the relational operators: group, select, and join.

Finally, we study methods that relate not to individual cases in the data table but to collective properties of many cases. This can be very simple, e.g. the properties that go into a box-and-whiskers plot. We also introduce simple model-building techniques, such as regression or smoothers. Unsupervised clustering and (automatic) methods of dimension reduction are included here, as are measures of precision (often depicted as "error bars" or "error bands").

Of course, there's only so much that can be done with 10 hours of classroom time and approximately 25 hours of out-of-classroom time. We'll describe how we balance the development of concepts and the understanding of operations with the use of technology to carry out the operations and provide access to notes and other classroom materials that support the course, as well as the main software used: R.

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 [© 2014 ISI/IASE]