Augmenting the Vocabulary Used to Describe Data

Robert GOULD

This is one of the most exciting papers I have read in a while. George Cobb has, as he has before, clearly identified a challenge to our statistics community that many of us have been aware was lurking somewhere on the margins, but have not seen so clearly until now. There's much to comment on here, and I expect there will be years of discussion, but I'd like to emphasize two topics mentioned in this paper – data and curricula.

Here was my very first introduction to data as an undergraduate math major: "Let $X_1, X_2, \ldots, X_n$ denote n random variables that have the joint p.d.f. $f(x_1, x_2, \ldots, x_n)$." (Hogg and Craig, 1978, p. 122). Some of you who enjoyed a similar introduction to data are probably marveling that Hogg and Craig were so far-sighted as to introduce the topic as early as page 122. Today, most data used in examples and homework problems in introductory courses—while represented less abstractly than in my course and possessing, thank goodness, some level of "realness"—are derived from random samples or studies that applied random assignment. When they are not, homework questions often begin "Assume that these are from a random sample" because without that assumption there's not much we can ask students to do. These probabilistic-culture data, I suggest, represent a very small fraction of data that our students encounter in life and maybe in their careers. This attention to only one type of data in our classrooms risks making our profession insignificant.

Despite being the "science of data," the statistics classroom has a narrow vocabulary for describing data. Let me expand this vocabulary by two terms. The first, "opportunistic data," was coined, to the best of my knowledge, by Amy Braverman, a statistician at Jet Propulsion Labs. The second is my own: "algorithmic data." Algorithmic data are data collected through an algorithm. Sensors collect data algorithmically, for example. The algorithmic trigger might be an occasional event, such as when a sensor detects motion, or might be a semi-continuous event, such as when sensors on a satellite are programmed to collect a stream of measurements. Opportunistic data are often collected by sensors, but more generally are data sets that are collected and await an opportunity for analysis. This category includes large, national databases which continue to foster research for purposes not originally foreseen by those who collected the data. (Think of the NHANES dataset.)

Opportunistic and algorithmic data challenge educators because they do not fit into the inference box; these approaches usually do not produce random samples, and a naive approach can lead to philosophical and scientific mistakes. (For example, see "The Parable of Google Flu: Traps in Big Data Analysis," Lazer. D et . al. 2014). And yet these data provide a pivotal role in students' lives and so provide a platform in which the science of data analysis can be introduced to a very wide audience.

When designing curricula, we should keep in mind this motto: Data First. We should design curricula that help *all* students understand *all* data, including algorithmic and opportunistic data. As George recommends, we should order topics in the order that best helps them understand data and not because we are supporting a "beautiful structure." We should exclude topics that do not help students understand data.

I would like to depart from George's recommendations, though, and urge us to think, when designing curricula, not in terms of semesters, but years. How should students learn about data from Kindergarten through retirement? This question had an easy answer when learning statistics meant learning mathematics. (Answer: wait until they've learned calculus.) However, as George points out, many useful and important tools can be understood through algorithms, and are more accessible at younger ages. In addition, educational technology can provide students with experiential access to abstractions such as random samples or repeated sampling, and so many topics now taught in graduate school or in the last months of a bachelor of science program can be introduced much earlier.

I have some experience with this first-hand. As the principal investigator of Mobilize, an NSF-funded project dedicated to bringing a "data science" curriculum to high schools, I've been struggling with the challenges of helping high school teachers teach their students to find meaning in data that do not belong to the probability culture. Our students use their cell-phones to engage in "participatory sensing campaigns," a form of algorithmic data collection in which they strive to gain insight into their lives and their communities. The data they collect are rich. They include geocoded locations, dates, photos, text, as well as answers to survey questions that fall into the more mundane categories of categorical and numerical.

From the Mobilize project, I've learned a few lessons about designing curricula. The most important: emphasize the statistical investigation process, as outlined in the GAISE K-12 report (Franklin et al. 2007). This consists of four stages: Ask Questions, Examine/Collect data, Analyze, Interpret. Most statistics curricula I've seen emphasize only the last two stages. Most high school science curricula emphasize the first two stages. Future citizens need all four stages. This investigation process works well in either of Breiman's two cultures and it keeps us focused on what matters: understanding of our lives, community, world.

The second important lesson I've learned is that engaging students in this cycle is not easy and requires considerable professional development for teachers. Our community needs to engage seriously in the preparation of teachers, not just through hosting workshops, but through changing teacher preparation at the undergraduate, graduate, and credentialing levels. Both science and math teachers are, with some exceptions, frightfully

Online discussion of "Mere Renovation is Too Little Too Late: We Need to Rethink Our Undergraduate Curriculum From the Ground Up," by George Cobb, *The American Statistician*, 69. Robert Gould, UCLA, Department of Statistics, Math Sciences Building, Los Angeles, CA 90095-1554 (Email: rgould@stat.ucla.edu).

unprepared to teach students to engage meaningfully with data.

"Big Data" and the algorithmic data culture provide a way for us to move forward to reach more students and to reach them through engagement in authentic analysis of data. George is to be applauded for shoving us in the right direction.

References

Cobb, G. W. (2015), "Mere Renovation is Too Little Too Late: We Need to Rethink Our Undergraduate Curriculum from the Ground Up," *The American Statistician*, 69(4), doi:10.1080/00031305.2015.1093029.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., and Scheaffer, R., (2007), "Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework," [online] The American Statistical Association. Available at *www.amstat.org/education/gaise/*

Hogg, R. and Craig, A., (1978), *Introduction to Mathematical Statistics* (4th ed.), London: Macmillan.

Lazer, D., Kennedy, R., King, G., Vespignani, A. (2014), "Big Data: The Parable of Google Flu: Traps in Big Data Analysis," *Science*, 343(6176), pp. 1203–1205. DOI: 10.1126/science.1248506.