BIG DATA: THE ROLE OF EDUCATION AND TRAINING

Julia Lane

The Jarmin and O'Hara piece describes a turning point in statistical data collection and dissemination. The federal agencies can no longer bear the full burden of producing and disseminating data for public policy research. Public policy researchers and schools should seize the new opportunity to complement the Census Bureau initiative—which also represents a turning point for both public policy education and research. We should adapt our educational programs so that we can produce the workforce capable of working with new data. We should adapt our research programs to move from largely artisanal individual efforts to large-scale "big science"; I have seen other scientific areas do so with great success.

WHAT DOES THIS MEAN FOR EDUCATION?

Public policy schools must both train a new generation of data scientists skilled in public policy and also increase the skill and competency level of the existing workforce. Data science is of great value to students and employers—not only is it the "sexiest job" in the 21st century, but employers pay a substantial premium for data scientists (Davenport & Patil, 2012).

Why? The new focus on evidence-based policy now draws heavily on data science (Barbosa, Pham, Silva, Vieira, & Freire, 2014; Catlett et al., 2014), yet data science training has too often been confined to technical training in computer science departments. As a result, while there are many courses in machine learning, database management or text analysis, the content of those curricula do not focus on the scientific analysis of social problems. In addition, the target audience should go beyond traditional students to include nontraditional students in government agencies and the private sector—so they can use data science tools as part of their regular employment (Holdren, Marrett, & Suresh, 2013).

Why? We have a lot to contribute. There is a lack of capacity to deal with big data as evidenced by the results of the open data movement. The barriers to value creation from open data platforms are well documented—including problems of diverse user needs and capabilities, the limitations of internally oriented data management techniques, untested assumptions about information content and accuracy, and issues associated with information quality and fitness for use (Dawes & Helbig, 2010). Having capable, in-house data scientists who can demonstrate to their fellow civil servants the value of big data to solve practical problems may be one of the most significant steps any government can take in breaking down the barriers to value creation (Jarmin, Marco, Lane, & Foster, 2014). Cities are a good place to start (Pardo, n.d.).

How? The challenge for public policy is that the newness of big data means that there is not a long history of knowledge of how to teach the right skills, but in a very influential series of papers, Handelsman and others argue that the new types of science need to adopt active learning techniques (Handelsman, Ebert-May,

Journal of Policy Analysis and Management, Vol. , No. , 1–3 (2016)

© 2016 by the Association for Public Policy Analysis and Management

Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
DOI:10.1002/pam.21922

Beichner, & Bruns, 2004). By this they mean changing teaching from a lecture-based format to one which is *inquiry based and modular learning* and which *treats students as scientists* who "develop hypotheses, design and conduct experiments, collect and interpret data, and write about their results." The approach appears to be effective: a recent meta-analysis of 225 studies of the effectiveness of "learning by telling" vs. "learning by doing," albeit in the undergraduate context, suggests that the latter increases examination performance while the former increases failure rates. The positive effects are particularly pronounced for students from disadvantaged backgrounds and for women in male-dominated fields (Freeman et al., 2014).

What? We can design experiential courses around policy issues—such as inequality, health, education, or transportation. Much can be learned from the experience in other fields in moving from a curriculum based on providing content to one that is both interdisciplinary and driven by concepts. In the biological sciences, Gutlerner and Van Vactor argue forcefully for the development of modular classes—what they call "nanocourses"[1] (Gutlerner & Van Vactor, 2016). Their approach brings together students from multiple backgrounds, engages faculty from a variety of disciplines, and creates "small discussion group activities that allow students to practice framing experiments into larger scientific contexts and disciplines."

How to evaluate? Of course, new educational techniques must be scientifically evaluated. Some of these measure the set of skills acquired. For example, Hoey suggests that the key measures include knowledge of concepts in the discipline; ability to conduct independent research; ability to use appropriate technologies; ability to work with others, especially in teams; and ability to teach others (Hoey, 2008).

WHAT DOES THIS MEAN FOR RESEARCH?

We need to move toward big data as big science. Universities are now both collecting and disseminating new types of data—for example, NYU's Center for Urban Science and Progress (CUSP) hosts both city agency data and data based on satellite and ground-based remote sensing (https://datahub.cusp.nyu.edu/); the University of Chicago has developed Plenario (Catlett et al., 2014). This is insufficient, however. For data to have value, they must be *used* by researchers, comparable across cities, and statistically reliable. The new Census initiative provides such an opportunity. The community can build a network of sound, accessible, and useful university-based data facilities that features common capacity and connectivity.

How? Each facility could be the equivalent of an urban statistical agency based on sound science: a trusted independent resource that can be used to draw valid inference for decisionmaking based on new types of data. The facility could work with state and local agencies, drawing on Census expertise, to enable the curation, management, and linkage of disparate data sources; facilitate analysis, collaboration, and replication of results; and apply modern statistical techniques to limit the potential for reidentification of human subjects (National Research Council, 2014)—and work with the federal statistical system to benchmark and validate results. Each facility could adopt the approach of such private sector initiatives as Amazon.com to build referral capabilities, of TripAdvisor and Yelp to promote metadata documentation and automated tools for project management and reporting. And each facility could be the locus of data-based experiential learning modules—described above—that bring together agency staff, urban researchers, and subject matter experts to engage with and validate the quality of different data sources, in the context of key policy issues.

[1] See, for example, https://nanosandothercourses.hms.harvard.edu/node/8.

The results of such a fully formed network will be the development of "big science": data sets and data collection methodologies that are comparable across geographies; common standards, methodologies, and tools, as well as the establishment of a large team-focused community of research and practice that improves decisionmaking within and across geographies.

JULIA LANE is a Professor at the Center for Urban Science and Progress and the Wagner Graduate School of Public Affairs at New York University, 295 Lafayette Street, New York, NY 10012; also a Co-Founder of the Longitudinal Employer-Household Dynamic program at the U.S. Census Bureau, and the new Institute for Research on Innovation and Science (e-mail: Julia.lane@nyu.edu).

REFERENCES

Barbosa, L., Pham, K., Silva, S., Vieira, M., & Freire, J. (2014). Structured open urban data: Understanding the landscape. Big Data Journal, 2, 144–154.

Catlett, C., Malik, T., Goldstein, B., & Giuffrida, J. (2014, December). Bulletin of the Technical Committee on Data Engineering. Vol. 37 (No. 4), IEEE Computer Society Special Issue on Urban Informatics.

Davenport, T. H., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century—A new breed of professional holds the key to capitalizing on big data opportunities. But these specialists aren't easy to find—And the competition for them is fierce. Harvard Business Review, 70, 70.

Dawes, S., & Helbig, N. (2010). Information strategies for open government: Challenges and prospects for deriving public value from government transparency. In M. Wimmer (Ed.), Electronic government: Lecture notes in computer science, In Electronic government, pp. 50–60. Springer Berlin Heidelberg, 2010. LNCS 6228.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. Proceedings of the National Academy of Sciences, 111, 8410–8415.

Gutlerner, J. L., & Van Vactor, D. (2016). Catalyzing curriculum evolution in graduate science education. Cell, 153, 731–736 (http://doi.org/10.1016/j.cell.2013.04.027).

Handelsman, J., Ebert-May, D., Beichner, R., & Bruns, P. (2004). Scientific teaching. Science, 304, 521.

Hoey, J. J. (2008). Tools and assessment methods specific to graduate education. In J. E. Spurlin, A. R. Sarah, P. L. Jerome (Eds.), Designing better engineering education through assessment: a practical resource for faculty and department chairs on using assessment and ABET criteria to improve student learning, pp. 149–167. VA: Stylus Publishing, LLC.

Holdren, J. P., Marrett, C., & Suresh, S. (2013). Federal science, technology, engineering, and mathematics (STEM) education 5-year strategic plan: A report from the committee on STEM Education National Science and Technology Council. Washington, DC: The National Academies Press.

Jarmin, R., Marco, A., Lane, J., & Foster, I. (2014). Using the Classroom to Bring Big Data to Statistical Agencies. AMSTAT news: the membership magazine of the American Statistical Association, 499, 12–13.

National Research Council. (2014). Proposed Revisions to the Common Rule for the Protection of Human Subjects in the Behavioral and Social Sciences. Committee on Revisions to the Common Rule for the Protection of Human Subjects in Research in the Behavioral and Social Sciences. Board on Behavioral, Cognitive, and Sensory Sciences, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Pardo, T. (n.d.). Making data more available and usable: A getting started guide for public officials. Retrieved from http://cusp.nyu.edu/data-privacy-book/.