BIG DATA FOR PUBLIC POLICY: THE QUADRUPLE HELIX

Julia Lane

INTRODUCTION

Data from the federal statistical system, particularly the Census Bureau, have long been a key resource for public policy. Although most of those data have been collected through purposive surveys, there have been enormous strides in the use of administrative records on business (Jarmin & Miranda, 2002), jobs (Abowd, Haltiwanger, & Lane, 2004), and individuals (Wagner & Layne, 2014). Those strides are now becoming institutionalized. The President has allocated $10 million to an Administrative Records Clearing House in his FY2016 budget. Congress is considering a bill to use administrative records, entitled the Evidence-Based Policymaking Commission Act, sponsored by Patty Murray and Paul Ryan. In addition, the Census Bureau has established a Center for "Big Data." In my view, these steps represent important strides for public policy, but they are only part of the story. Public policy researchers must look beyond the federal statistical system and make use of the vast resources now available for research and evaluation.

All politics is local; "Big Data" now mean that policy analysis can increasingly be local. Modern empirical policy should be grounded in data provided by a network of city/university data centers. Public policy schools should partner with scholars in the emerging field of data science to train the next generation of policy researchers in the thoughtful use of the new types of data; the apparent secular decline in the applications to public policy schools is coincident with the emergence of data science as a field of study in its own right. The role of national statistical agencies should be fundamentally rethought—and reformulated to one of four necessary strands in the data infrastructure; that of providing benchmarks, confidentiality protections, and national statistics.

The federal statistical data infrastructure was built in response to post World War II and the Cold War policy imperatives. It is deeply grounded in collecting data on manufacturing processes, building national income and product accounts, and describing national trends in population, employment, and social dynamics. Also in the 1950s and 1960s, the federal research agencies provided resources for major social science data collections—notably the Panel Survey of Income Dynamics, the General Social Survey, and the American National Election Survey. Despite the proposed moves to incorporate administrative data, that infrastructure still exists, largely unchanged. Yet there has been a fundamental transformation in the way in which data are now being used for both policy and research as Figure 1 demonstrates. The transformation has resulted from a tsunami of change—and social scientists stand in danger of being supplanted by data scientists.

The first of these is the change in the role of policymakers. Policy paralysis at the national level has been matched by a surge on the part of cities and states creating and implementing policies at the local level. The resultant interest in evaluating the impact of those policies has created enormous demand for local data, but limited willingness on the part of federal government to pay for it (the American Community

Journal of Policy Analysis and Management, Vol. , No. , 1–8 (2016)

Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
DOI:10.1002/pam.21921

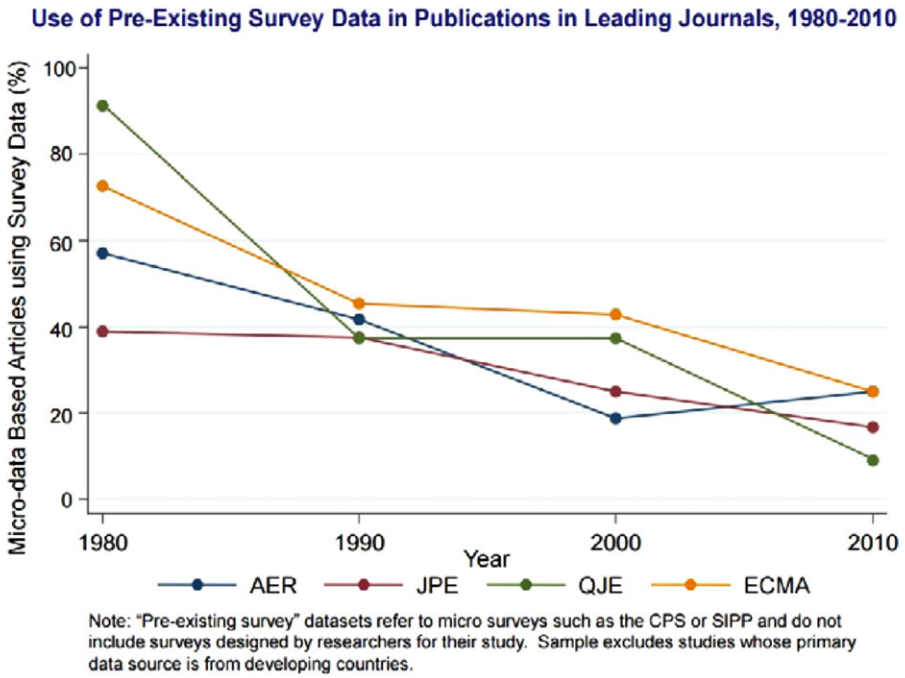

Figure 1. The decreasing use of survey data (Chetty, 2012).

Survey is the poster child for that tension). States and cities have responded by establishing their own institutions, such as the Offices for Data Analytics, Open Data Initiatives, and Chief Data Officers (Pardo, 2014).

The second has been a change in the type of data that are available. The change in the nature of the new type of data is transformative. Its characteristics—its velocity, volume, and variety—and the way in which it is collected, mean that a new analytical paradigm is open to statisticians and social scientists (Hey, Tansley, & Tolle, 2009). The classic statistical paradigm was one in which researchers formulated a hypothesis, identified a population frame, designed a survey and a sampling technique, and then analyzed the results (Groves, 2011). The new paradigm means that it is now possible to digitally capture, semantically reconcile, aggregate, and correlate. Those correlations might be effective or suspect (Couper, 2013). They certainly enable completely new analysis to be undertaken—none of which can be done using survey data alone. For example, the new type of analysis might be one that captures rich environmental detail on individuals (from sensors, google earth, videos, photos, or financial transactions). Or the analysis might include rich and detailed information on unique and quite small subsets of the population (from microbiome data, or websearch logs). Or the analysis might be on completely new units of analysis, like networks of individuals or businesses whose connections can only be captured by new types of data (like tweets, cell phone conversations, and administrative records). As Kahneman points out, the new measurement can change the paradigm in its own right (Kahneman, 2011); and there has indeed been fundamental change in empirical social science, which is increasingly focused on issues of measurement (Warsh, 2015). Researchers are collecting and linking their own data (Einav & Levin, 2013), whether it be generating new measures of unemployment (Antenucci, Cafarella, Levenstein, Ré, & Shapiro, 2014), or creating new data sets for cities (Glaeser, Kominers, Luca, & Naik, 2015).

The third has been a change in the way in which data can be processed and used for analysis. The data are not provided by federal statistical agencies to social scientists ready-made for analysis, with full documentation. Documentation, structure, coverage, and provenance—as well as understanding—must be developed de novo. New techniques, such as text, graph and network analysis, and machine learning have been developed by data scientists to structure, link, manage, and analyze the new types of data.

The opportunities for changing policy are breathtaking. For example, it is now possible to use graph and network analysis to rethink the fundamental social and economic units of analysis that go beyond a basis grounded in a program or tax identity, but to move to one grounded on transactions—like cell phone or media communications or financial transactions. It should be possible to use text analysis to create classification systems that are not based on manual allocation of predetermined categories but that are based on the words used by economic and social entities to describe their activities. Sensor data can be a catalyst for not only collecting data in real time but for capturing the dynamics of use at extremely granular temporal and spatial levels—for example, recording activities in parks, types of sound in neighborhoods, or particulate pollution (Holland & Koonin, 2014). Surveys can be then rethought to purposefully collect information that is not captured in other ways (the Kavli Human Project; Azmak et al., 2015).

THE FUTURE: A QUADRUPLE DATA HELIX

The data helix of the future will have four strands—state and city agencies, universities, private data providers, and federal agencies. Each contributes key features; each reinforces the weaknesses of the other.

The role of State and City agencies is clearly critical in terms of framing key policy questions that are necessary to drive data collection and organization (Holland & Koonin, 2014). They are generators of data in their own right and are, of course, not only the ultimate users of policy analysis, but employers of policy analysts. Despite the enormous value provided by the creation of the state and city data offices, their role, almost by definition, must be to address immediate policy issues, rather than to build long-term data infrastructures; their mandate is to work with city data, rather than the full spectrum of available data; their viability is subject to political whims, and their external credibility can also be affected by political variation.

Universities can play a separate role. First and foremost, they can act as a trusted independent third party to process, store, analyze, and disseminate data. They have a long history of doing so—state data centers exist in all 50 states, as do bureaus of economic research and often specialized centers in such areas as housing, workforce, and education policy. Second, policy schools, partnering with the new field of data science, can provide new sources of trained human capital to both the city agencies and the private sector. Third, social scientists, again acting in partnership with the data science community, can act as the R&D lab for purposive future data collection. And as a network, they can host data from a variety of different sources and provide integration and benchmarking across regions. However, they, too, cannot stand alone. The incentive structure for University researchers is based on publications and grants—namely, producing frontier research—not producing sustainable data infrastructures. Their skill set typically does not include sophisticated confidentiality protection techniques, and their focus is not necessarily cross-disciplinary.

Companies in the private sector host vast treasures of proprietary data (Einav & Levin, 2013). But, as Bob Groves has pointed out, they have four common salient attributes that require the input of social scientists: they measure behavior, are highly granular, are wide but not deep, and do not cover the full population of interest. To this, we would add that the data are often ephemeral, unstructured, and undocumented. Universities and statistical agencies will be necessary to develop statistical and metadata documentation techniques that exploit the richness of the data but preserve inference (Varian, 2014); city agencies will be necessary to guide the strategic collection of the data to answer policy questions.

The role of federal statistical agencies should be structured to play to their strengths, rather than in a vain attempt to restructure and address their core weaknesses. They are the only source of consistent long-term national frames against which the variety of different sources can be benchmarked. They are the trusted source and provide the long-term expertise in confidentiality protections that has been deemed so critical by the National Academies of Sciences (National Academies, 2014). But they should draw on the strengths of the other strands of the helix. Federal agencies are unlikely to be able to respond quickly to change—either in terms of data processing and collection or in terms of creating new series. Surveys, which have to go through a long bureaucratic approval and congressional funding process, followed by stakeholder consultation and Office of Management and Budget (OMB) approval of the data collection instrument, before field collection even starts, can take 15 years to move from conception to realization.

What Must Be Done

A new set of institutions—city/university data facilities—must be established. These institutions should form the backbone of the quadruple helix, with direct connections to the private sector and to the federal statistical agencies. There are four core sets of issues—technical, legal, privacy, and training—that the institutions must address as a network.[1] But, most urgently, the framework must prove that it both has value and is sustainable. We address each in turn.

Value

Our experience is that there must be some short-term demonstration of value. The most obvious approach is to build on existing city capacity and existing partnerships. There are nascent existing successful models, such as the new Institute for Research on Innovation and Science (IRIS) at the University of Michigan (Zolas et al., 2015) and the data facility at the Center for Urban Science and Progress at NYU. Municipalities are building their own capacity for data science, with groups established in multiple cities. Prominent examples include the Mayor's Office of Data Analytics (MODA) in New York City and a Mayor's Office of New Urban Mechanics in both Boston, MA and Philadelphia, PA, while Chief Information Officers in cities as large as Chicago, IL and as small as Asheville, NC are taking steps to develop data science capacities as part of smart cities initiatives (Gil-Garcia, Pardo, & Nam, 2015).

[1] There is a substantial literature on building infrastructures for such combined social science and physical measurement data. The Research Data Alliance is establishing procedures in the broader research data arena (Berman, Wilkinson, & Wood, 2014). The Berkeley Initiative has developed a manual for best practices in transparency (Christensen & Soderberg, 2015). Harvard's Institute for Quantitative Social Science has produced important research tools (King, 2014). There are practical examples of efforts to build successful data infrastructures, including the Administrative Data Research Network in the United Kingdom (http://www.adrn.ac.uk/), the U.S. Federal Statistical Research Data Network (https://www.census.gov/ces/rdcresearch/), and the German RDC-in-RDC-network of the German Research Data Center (Bender, Burghardt, & Schiller, 2014; Bender & Heining, 2011).

Technical Issues

The new infrastructure must ensure that data from disparate sources are collected managed and used in a manner that is informed by end users. There are many technical challenges: disparate data sets must be ingested, their provenance determined, and metadata documented. Researchers must be able to query data sets to know what data are available and how they can be used. And if data sets are to be joined, they must be joined in a scientific manner, which means that workflows need to be traced and managed in such a way that the research can be replicated. In addition, the legal issues associated with combining information on individuals, businesses, the environment, and the physical facilities in which economic and social activity occurs must be addressed, together with the associated bureaucratic barriers.

Legal Issues

Data sharing is often "governed by a hodge-podge of contractual instruments" (Wilbanks, 2014). The new institutions must develop standardized data-sharing agreements appropriate to the laws and regulations of city jurisdictions in consultation with interested individuals, civic organizations, and businesses (Pentland, Greenwood, Sweatt, Stopczynski, & de Montjoye, 2014).

Privacy and Confidentiality

The historical approaches to protecting privacy and confidentiality—informed consent and anonymity—no longer hold (Baracas & Nissenbaum, 2014). The lack of guidance is serious, since absent acceptable protocols, Institutional Review Boards will limit access (National Academies, 2014). Systematic approach is needed, with the development of standards that evolve as data types evolve. Some have suggested using Big Data itself to keep track of user permissions for each piece of data and act as a legal contract (Pentland et al., 2014). Others have suggested using analysis on encrypted files or building systems in which information flow, rather than access control, is used to enforce policies (Landwehr, 2014).

Training

While Big Data has enormous appeal in its potential to describe economic and social activity, the structure of data science training has often been confined to technical training in computer science departments. As a result, while there are many classes in machine learning, database management, or text analysis, the syllabi are typically not focused on how to use these skills to contribute to the scientific analysis of social problems. In addition, most are targeted at traditional graduate students, and are part of a continuous course of study. Yet there is a great need for graduate training for nontraditional students, who need to understand how to use data science tools as part of their regular employment. The network of institutes, situated in universities, should both train a new generation of data scientists and increase the skill and competency level of the existing workforce—by training students to identify and capture the appropriate data, understand how data science models and tools can be applied, and determine how associated errors and limitations can be identified from a social science perspective.

Sustainability

The institutions could have four separate sources of funding for what is essentially a major research infrastructure (MRI). One is to provide data as a service to schools of public policy and social science—and institute a modest lab fee. The second is to provide certificate classes in urban informatics to city and state agencies; having capable, in-house data scientists who can demonstrate to their fellow civil servants the value Big Data has for solving 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). The third is to support the R&D efforts of faculty in the same way that MRIs do in the natural sciences, again through an infrastructure fee. The fourth is through the production of reports and analysis. Each of these would sustain a professional technical staff in the same manner as the University of Michigan's ICPSR or the University of Chicago's NORC data enclave.

CONCLUDING COMMENTS

It is exciting to see that change is coming to the federal statistical system. The establishment of a new administrative data clearinghouse and the Census Bureau's Big Data Center are welcome additions to the new world of data. But the federal statistical system is no longer the center of the universe for the provision of the data upon which public policy is based. It is time for new institutions to arise that can make intelligent, informed, and ethical use of all new sources of data, not just those provided by the federal government.

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

Abowd, J. M., Haltiwanger, J., & Lane, J. (2004). Integrated longitudinal employer-employee data for the United States. American Economic Review, 94, 224–229.

Antenucci, D., Cafarella, M., Levenstein, M., Ré, C., & Shapiro, M. D. (2014). Using social media to measure labor market flows. Working paper No. w20010. Cambridge, Massachusetts: National Bureau of Economic Research.

Azmak, O., Bayer, H., Caplin, A., Chun, M., Glimcher, P., Koonin, S., & Patrinos, A. (2015). Using Big Data to understand the human condition: The Kavli HUMAN project. Big Data, 3, 173–188.

Baracas, S., & Nissenbaum, H. (2014). The limits of anonymity and consent in the Big Data Age. In J. Lane, V. Stodden, S. Bender, & H. Nissenbaum (Eds.), Privacy, Big Data, and the public good: Frameworks for engagement. New York, New York: Cambridge University Press.

Bender, S., & Heining, J. (2011). The research-data-centre in research-data-centre approach: A first step towards decentralised international data sharing. IASSIST Quarterly, 35, 10–16.

Bender, S., Burghardt, A., & Schiller, D. (2014). International access to administrative data for Germany and Europe, Vancouver, Canada. Available at SSRN 2393357.

Berman, F., Wilkinson, R., & Wood, J. (2014). Building global infrastructure for data sharing and exchange through the research data alliance. D-Lib Magazine. Reston, VA: Corporation for National Research Initiatives.

Chetty, R. (2012). Time trends in the use of administrative data for empirical research. Presentation Slides. Retrieved April 23, 2016, from http://www.rajchetty.com/chettyfiles/admin_data_trends.pdf.

Christensen, G., & Soderberg, C. (2015). Manual of best practices in transparent social science research. Berkeley, CA: University of California. Retrieved April 23, 2016. from http://www.bitss.org/education/manual-of-best-practices/.

Couper, M. (2013). Is the sky falling? New technology, changing media, and the future of surveys. Survey Research Methods, 7, 145–156.

Einav, L., & Levin, J. D. (2013). The data revolution and economic analysis. Working Paper Series, No. 19035. Cambridge, Massachusetts: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w19035.

Gil-Garcia, J. R., Pardo, T. A., & Nam, T. (2015). What makes a city smart? Identifying core components and proposing an integrative and comprehensive conceptualization. Information Polity, 20, 61–87.

Glaeser, E. L., Kominers, S., Luca, M., & Naik, N. (2015). Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life. No. w21778. Cambridge Massachusetts: National Bureau of Economic Research.

Groves, R. M. (2011). Three eras of survey research. Public Opinion Quarterly, 75, 861–871.

Hey, T., Tansley, S., & Tolle, K. (2009). The fourth paradigm: Data intensive scientific discovery. Redmond, WA: Microsoft Research.

Holland, M., & Koonin, S. E. (2014). The value of Big Data for urban science. In J. Lane, V. Stodden, S. Bender, & H. Nissenbaum (Eds.), Privacy, Big Data, and the public good: Frameworks for engagement. Cambridge, UK: Cambridge University Press.

Jarmin, R., & Miranda, J. (2002). The longitudinal business database (CES Working Paper No. 02-17). Suitland, Maryland: Census Bureau working paper.

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 449, 12–13. Alexandria, VA.

Kahneman, D. (2011). Thinking fast and slow. Farrar, Straus and Giroux, New York, New York.

King, G. (2014). Restructuring the social sciences: Reflections from Harvard's Institute for quantitative social science. PS: Political Science and Politics, 47, 165–172. Retrieved from http://journals.cambridge.org/repo_A9100Nlq.

Landwehr, C. (2014). The operational framework: Engineered controls. In J. Lane, V. Stodden, H. Nissenbaum, & S. Bender (Eds.), Big Data and privacy. Cambridge, UK: Cambridge University Press.

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. (2014). Making data more available and usable: A getting started guide for public officials. Albany, NY: Center for Technology in Government, State University of New York. Retrieved from http://cusp.nyu.edu/data-privacy-book/.

Pentland, A., Greenwood, D., Sweatt, B., Stopczynski, A., & de Montjoye, Y.-A. (2014). Institutional controls: The new deal on data. In J. Lane, V. Stodden, H. Nissenbaum, & S. Bender (Eds.), Privacy, Big Data, and the public good: Frameworks for engagement. Cambridge, UK: Cambridge University Press.

Varian, H. R. (2014). Big Data: New tricks for econometrics. Journal of Economic Perspectives, 28, 3–27.

Wagner, D., & Layne, M. (2014). The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software. Suitland, MD: Census Bureau. Retrieved from https://www.census.gov/srd/carra/CARRA_PVS_Record_Linkage.pdf.

Warsh, D. (2015). On taking things apart. Economic Principals. Retrieved April 23, 2016, from http://www.economicprincipals.com/issues/2015.12.13/1837.html

Wilbanks, J. (2014). Portable approaches to informed consent and open data. In J. Lane, V. Stodden, S. Bender, & H. Nissenbaum (Eds.), Privacy, Big Data, and the public good: Frameworks for engagement. Cambridge, UK: Cambridge University Press.

Zolas, N., Goldschlag, N., Jarmin, R., Stephan, P., Owen-Smith, J., Rosen, R. F., Allen, B. McFadden, Weinberg, B. A. & Lane, J. I. (2015). Wrapping it up in a person: Examining employment and earnings outcomes for PhD recipients. Science, 360, 1367–1371.