Building an Infrastructure to Support the Use of Government Administrative Data for Program Performance and Social Science Research

Julia Lane

Julia.lane@nyu.edu

NOTE: Many thanks for very useful comments from Erica Groshen, Danny Goroff, David Ellwood, Mark Lowenstein, Sarah Dale, Kristen Monaco, Scott Fricker, and Andrew Reamer.

This article provides an overview of the elements necessary to build a sustainable research data infrastructure. I argue that it needs the financial and intellectual engagement of a community of practice. Most attention has been paid to researchers and policy-makers, but a third group—government programmatic agencies—must be a focal point since they act as both data producers and as policy implementers. I also discuss possible business models that are both consistent with serving the needs of multiple stakeholders and that are not completely dependent on the largesse of the public purse

Keywords: data infrastructure, linked data, evidence based policy

There is a new opportunity to link administrative data across agencies at all levels of government—federal, state and local areas of all sizes. There is clear interest in data-driven research and policy as evidenced by the push toward open data (Catlett 2014) and the proliferation of government chief data officers (Pardo 2014), the establishment of the Evidence Based Policymaking Commission, and the active engagement of many important private foundations in supporting linked data systems. It is fair to say that there is now the potential for the evolution of a new research infrastructure that joins datasets across federal and local agencies and enhances decision-making. Building such an infrastructure will require a thoughtful balancing of costs and benefits, documented value, and the engagement of the full community.

The potential value is immense. If the infrastructure is well designed, agencies can manage their programs better. In terms of programmatic operations, respondent burden could be reduced if statistical agencies used administrative records to replace or enhance survey question. Operational costs could also be reduced to save taxpayer dollars – for example, the current cost estimate for the 2020 Decennial Census of $15.5 billion, or approximately $107 per man, woman and child in the United States, could be reduced by about $1.5 billion1. In terms of creating more programmatic value, better policy interventions could be designed – just as early uses have led to the development of permanent supportive housing for the homeless, the design of effective training programs for dislocated workers, or the implementation of school curricula that really change the way in which children view drug use2. In the most forward looking sense, social science research could be galvanized by access to new sources of high quality data. For

[1] http://www.gao.gov/highrisk/2020_decennial_census/why_did_study; http://www.washingtonexaminer.com/2020-census-to-cost-107-per-household-156-billion-most-ever/article/2639061

[2] Commission on Evidence Based Policy making Final report pp 9-10; https://www.cep.gov/content/dam/cep/report/cep-final-report.pdf

example, work by Raj Chetty using tax records has documented that the intergenerational mobility is declining, and the American Dream is less robust than previously thought3

The potential risks of massive integrated data systems are high as well, though. Poorly designed infrastructure could lead to security breaches that could harm vulnerable populations – imagine the effect of an Equifax style breach on data that pertained to immigrant populations. Poorly documented or intermittently available datasets could result in higher, not lower, costs to statistical agencies as staff struggle to reconcile changing classification systems, or fill in gaps that occurr because data providers went out of business. Imagine, for example, if the Bureau of Labor Statistics came to rely on ADP employment data to measure employment, and ADP[4] simply stopped providing information or decided to triple prices.

In order for the opportunity to be realized, it is critical to move from hypothetical to actual value and to demonstrate that confidentiality can be protected. (Abowd, Haltiwanger, and Lane 2004, 224-29). Many agencies are not legally permitted to share data for research unless the work that is to be done is consistent with the agency mission. For example, the success of the Longitudinal Employer Household Dynamics (LEHD) program is at least partly due to the development of the Quarterly Workforce Indicators and the "On the Map" program, which generated value for both statistical and programmatic agencies (see Figure 1) while protecting confidentiality.

Figure 1

Example of Data Infrastructure Producing Value to Data Providers

[3] http://www.mobilitypartnership.org/blog/intergenerational-mobilitys-downward-trend
[4] https://www.adpemploymentreport.com/2017/June/NER/NER-June-2017.aspx

SOURCE: https://lehd.ces.census.gov/. US Census Bureau

The success of LEHD (and other similar efforts) was also made possible because it could be embedded in a strong research infrastructure: the Center for Economic Studies at the US Census Bureau. That enabled the linkage activities to be institutionalized and professionalized – in essence, creating a 'coral reef" of data.

What precisely is a research infrastructure? Funding agencies tend to define it by its operational characteristics. The European Commission says "The term 'research infrastructure' refers to facilities, resources and related services used by the scientific community to conduct top-level research in their respective fields, ranging from social sciences to astronomy, genomics to nanotechnologies."[5] In the United States, the National Science Foundation's Major Research Instrumentation Program (MRI) says it "serves to increase access to shared scientific and engineering instruments for research and research training in our Nation's institutions of higher education, not-for-profit museums, science centers and scientific/engineering research organizations."[6] The National Institutes of Health does not precisely define research infrastructures, but "includes the physical, intellectual and human resources that advance biomedical research at the NIH."[7]

In this article I argue that the real mark of a good data infrastructure is that it is sustainable, and that its results are valued by the broad community that it serves. Without that value, there will be no long-term sustainability. In other words, an infrastructure needs the

[5] https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=what accessed January 15, 2017; EU support for RIs in the context of its Framework Programmes (FPs) began with FP2 (1987-1991) when it had a budget of about €30 million. From these relatively humble beginnings, FP7 (2007-2013) earmarked €1.85 billion for RIs between 2007 and 2013 and the Horizon 2020 Programme will support research infrastructures with about €2.5 billions between 2014 and 2020.

[6] https://nsf.gov/funding/pgm_summ.jsp?pims_id=5260&org=OIA&from=home

[7] https://dpcpsi.nih.gov/sites/default/files/ORIP%20Strategic%20Plan%20Final-%20April%202016.pdf

financial and intellectual engagement of a community of practice. In terms of developing a new infrastructure for the use of administrative data, most attention has been paid to researchers and policy-makers, but a third group—government programmatic agencies—must be a focal point since they act as both data producers and as policy implementers. 8I argue that the provision of technical solutions to housing safe data is a fundamental and necessary step in order to get data providers to contribute data. However, technical solutions, while necessary, are not sufficient to ensure that data is made available for operational and research purposes. I thus also describe the core organizational features necessary for all stakeholders to participate in the establishment of a new intra-agency data infrastructure: reputation, reciprocity, and trust. (Ostrom 1998, 1-22) I describe the complex motivations of the different stakeholder groups and argue that there should not be a "one size fits all" approach to designing a successful institutional framework (Ostrom 2010, 641-72) though ultimately some core elements of technology, law, and privacy protection can move toward a much greater degree of standardization. I conclude with a discussion of possible business models that are both consistent with serving the needs of multiple stakeholders and that are not completely dependent on the largesse of the public purse.

Infrastructure

Merriam-Webster defines an infrastructure as the basic equipment and structures that are needed for an economy or an organization to function properly. [9] This volume documents in extensive detail the extent to which an administrative data infrastructure will make the country function better through accessing and using data. The operational aspects are obvious: there are many data uses such as being able to track program participants across programmatic agencies, across time

[8] Private data providers should, of course, be part of a broader conversation.

[9] https://www.merriam-webster.com/dictionary/infrastructure

and space that simply make programs more efficient. Research is equally important: determining what works and what does not enables better utilization of taxpayer dollars. Researchers, policy-makers, and government agencies (in their twin roles as data producers and policy implementers) need to work together for the country to be served well by a data infrastructure. For example, institutions like the University of Chicago's Chapin Hall10 have worked with all three stakeholders to combine administrative data on children from many sources and have been a remarkable resource for evidence to Illinois policy makers and communities for over 50 years The challenge before us is to determine the institutional characteristics necessary to engage stakeholders at a national scale.

One of the challenges of building a sustainable model is dealing with mission divergence among those stakeholders. University scholars are motivated by research that takes years and aimed at peer-reviewed journals. Policy-makers have short time frames to make a difference. Agencies, in their role of providing support to policy-makers, are focused on concrete service improvements for citizens. Agencies in their role of data provider must both protect data and ensure that data access is granted for mission specific purposes.

The work of Nobel Laureate Elinor Ostrom is extremely instructive in designing a sustainable infrastructure. She has described in detail how her empirical study of different institutional structures led her to identify three key features of these structures: reputation, reciprocity, and trust. She emphasized the importance of small groups, face-to-face communications, and the development of shared norms in developing those three features. Ostrom also made it clear that there are multiple (seven) types of rules that govern behavior but "gave up on the idea that [there were] specific rules that were associated with [a] successful

[10] http://www.chapinhall.org/

case"; in other words, she was against one monolithic approach (Ostrom 2010, 641-72) She argues for the facilitation of institutional development that brings out the best in humans and deals with the inherent complexity of human dynamics, what she calls a polycentric approach. Such a polycentric approach should sit at the core of the development of a data infrastructure. The human dimension, which brings together the three groups of stakeholders, should be as carefully thought through and made as central to the design as the technical dimension. The following describes each of their interests in turn.

Stakeholders

Policy-makers

The professional practice of policy-makers stands to gain in terms of designing effective operations and policy - administrative data are often of higher quality than survey data in key dimensions. In particular, administrative data are, by definition, better at capturing the population of participants in government programs than are surveys. This is important, since if participation is underestimated, particularly for particular groups, the structure of the interventions will be biased. In a series of high impact and thoughtful papers, Bruce Meyer and coauthors have shown that survey data suffer from nonresponse both at the unit and item level, as well as response measurement error, and documented that using administrative data can significantly increase quality. (Meyer, Mok, and Sullivan 2015, 199-226) (Meyer and George 2011) (Meyer, Mok, and Sullivan 2009)

Administrative data can be structured to be longitudinal, with typically less attrition than that which is exhibited by surveys. Such data can permit the examination of the effect of interventions. In the educational arena, for example, longitudinal administrative data make it

possible to examine the returns to pre-school education or No Child Left Behind . As pointed out by Figlio, Karbownik, and Salvanes 2015

> "Registry data have been used to study the introduction of new technologies to schools in England (Machin, McNally and Silva 2007), experimental evidence on schools' influence on parents' involvement in education in France (Avvisati et al. 2014), experimental evidence on gender differences in competitiveness and its consequences for educational choices in the Netherlands (Buser, Niederle, and Oosterbeek 2014), the role of school quality in Romania (Pop-Elches and Urquiola 2013), experimental evidence on learning incentives in Mexico (Behrman et al. 2015), perceived effects of school quality on the housing market (Figlio and Lucas 2004), the role of peer effects utilizing student reshuffling due to extreme events (Imberman, Kugler, and Sacerdote 2012), and the ability of principals to recognize effective teachers (Jacob and Lefgren 2008). " Educational Research and Administrative Data",(Figlio, Karbownik, and Salvanes 2015)

Of course, there are negatives to linking data that policy-makers also need to recognize. Administrative records are collected for the purposes of administering programs, so their coverage and applicability depends on the nature of the program for which the data were collected. It may be difficult to generalize conclusions from one population to another.

Government agencies

Government programmatic agencies are being forced by Congress and budgetary expediency to use new technology and approaches to make better use of existing data. At the state and local levels, agencies have created a new job title, "chief data officer" and built dashboards, initiated predictive analytic and smart sensor projects in the name of better

efficiency, accountability and improved community engagement. (President's Council of Advisors on Science and Technology 2016). Local government organizations (e.g., U.S. Conference of Mayors, National League of Cities and the International City/County Management Association) have launched data initiatives.

There are similar pressures on government statistical agencies. The declining response rate and quality of surveys has caused widespread concern (Meyer, Mok, and Sullivan 2015, 199–226) (Meyer and George 2011). In response, the Office of Management and Budget's Interagency Committee on Statistical Policy has encouraged a set of system-wide pilot projects to advance the statistical uses of administrative data (Smith 2014).

There are many reasons to link datasets. By linking to an existing source of data instead of implementing a new survey, there is a cost savings (and almost certainly a time savings as well). For some research questions (e.g., a survey of the reasons for death of a longitudinal cohort of individuals), a new survey may not be possible. In the case of administrative data or other automatically generated data, the sample size is much greater than would be possible from a survey. And once links are made to one dataset, there may be an opportunity to vastly expand links in the future.

There are also, however, multiple challenges for agencies to provide and link data—there are many legal and technical hurdles to clear and few prototype successes to point to. As some white papers have pointed out, the legal framework varies from agency to agency, and negotiations between lawyers and analysts can take many months of staff time. In addition, the pressures to meet existing program needs make it difficult for agencies to allocate staff time to try something new and create pipelines of new products. There can also be serious downsides to allowing access to data, either because of the potential for breaches or poor quality analysis that

results in agency embarrassment. In addition, government salary structures often make it difficult to hire and retain enough in-house data analysts, so agencies do not have the capacity to work with new linked data. The lack of workforce capacity is a binding constraint, because agencies often do not have the mechanisms and resources to share data and build the linked datasets. They also do not have the resources to develop new products in their own right. And while some federal agencies have professional development funds for staff, some do not.

The research community

The value to researchers of using administrative data has been well established (Card, Chetty, Feldstein and Saez 2011) (Jarmin and O'Hara 2016, 715-21) indeed, the use of administrative data by researchers has increased substantially over the past 30 years (see Figure 2). There are many reasons for this. The coverage is broad and the sample size large. This means that it is possible to study rare events, and small segments of the population. Since much of the economic and social activity of interest is concentrated in a small segment of the distribution—such as health care costs or job creation—it is critical to have sufficient data to study outliers. So, for example, business dynamics, which are heavily skewed can only be studied using administrative data (Decker, Haltiwanger, Jarmin, and Miranda 2016). One important finding has been the decline of transformational entrepreneurial firms—those that introduce major innovations and make substantial contributions to growth—particularly in the high-tech sector. On the human dimension, a small number of individuals disproportionately contribute to crime incidents; a small number disproportionately contribute to health care costs; a small number disproportionately contribute to welfare costs. The study of the behavior of such groups, and designing appropriate interventions, is made possible by large-scale administrative data.

Figure 2

The Use of Administrative Data

SOURCE: Raj Chetty.

There are a number of barriers to accessing administrative data—many of which can substantially reduce the willingness of researchers, particularly junior researchers, to work with administrative records. A major issue is simply getting access to data. There are substantial legal restrictions on data use and access. Researchers need to first identify champions within an agency who are willing to take the time to work with them. They must then identify what data are available, develop projects that are consistent with the agency mission and develop detailed data management and security plans. If a researcher wants to link data across agencies, she/he must work with the agencies to develop memoranda of understanding or interagency agreements, common access protocols, and research review requirements. This process can take years, which a junior researcher, wanting tenure, simply does not have. However, these barriers should not be completely eliminated. Access to confidential micro data is a privilege, not a right. And some difficulty is necessary so that the researcher is constantly reminded that she/he is working with data that deal with human subjects. (Ohm 2014)

The incentive structure for the hard work of building widely useable data infrastructures is also lacking. There are many academic rewards for publishing. There are few rewards for building and documenting data infrastructure efforts. It is also much easier to get funding for

projects that promise publications; there are few initiatives that fund infrastructure development. [11]

General barriers for all stakeholders

A major issue with linked data, although difficult to quantify, is dealing with turf battles, both internal and external to the pertinent agency. I experienced a turf battle when I worked with an internal champion in a major governmental agency to develop a researcher access program that provided access to extremely sensitive microdata. That program was established and is now extremely successful. However, the internal champion had to be creative within his agency to move things forward; once he was uncovered he was sidelined and subsequently retired. There were similarly many external turf battles in the establishment of the Longitudinal Employer Household Dynamics (LEHD) program; dealing with those challenges took more time than dealing with the technical issues.

Building a Business Model

Infrastructure development should explicitly address how to develop reputation, reciprocity, and trust between agencies and the research community. Given the complexities of the different agencies and policy issues, the infrastructure should draw on common standards, while customizing some elements—as Ostrom pointed out one size should not fit all. Indeed, Ostrom identified some key mechanisms in building successful institutions: reliance on small to medium-size organizations to ensure flexibility, voluntary participation to ensure that the needs of stakeholders are met on an ongoing basis, and multiple service providers to ensure responsiveness. (Ostrom 1990)

An example from personal experience – building a training program for government agencies - might serve to fix ideas and show how such centers can work with sufficient initial investments. New York University was tasked by the Census Bureau to build an Administrative Records Research Facility to inform the decision making of the Commission on Evidence Based Policy. The goal of that facility was to build and support an infrastructure that will expedite the acquisition of federal and federally sponsored administrative data,

[11] There are notable exceptions.

sources for program evaluation, improve data documentation and linkage techniques, and leverage and extend existing systems for governance, privacy protection, and secure access to these data. That facility thus was designed to address the technical problem of security and access. However, because of the ADRN UK experience, which struggled to gain agency acceptance and data, it was clear that it was critical to also build an agency engagement strategy. I worked with Rayid Ghani, a computer scientist at the University of Chicago, Bob Goerge at Chapin Hall at the University of Chicago and Frauke Kreuter, a statistician at the University of Maryland, to develop training classes centered around empirical policy issues (Foster et al. 2016). These training classes have been a huge success. They address the key agency challenges: workforce capacity and the need to serve agency missions. The classes are designed to build new products through linking data, using modern technology and applying active learning techniques in a sandbox environment. It (i) creates a pipeline of new prototype products central to an agency's mission as defined by senior management, (ii) develops teams of practitioners who can demonstrate the value of the new types of data for solving real world practical problems and who become embedded in their organizations, and (iii) makes new linked data available on an ongoing basis. This framework builds trust, reputation, and as it is repeated, reciprocity—the features that Ostrom identified as critical.

Our initial experience has been extremely positive. We have engagement in two key policy areas of interest: the labor market outcomes of ex-offenders and the labor market outcomes of welfare recipients. By providing incentives to provide data, resources to work on high priority problems and demonstrated value in sharing data across agency lines, we demonstrate the value of linked data by engaging staff from multiple agencies to work on cross-agency problems. The classes are not structured as lectures but rather are inquiry based and modular. We engage agency staff in a lab in which participants implement skills to produce actionable analyses (Handelsman, Ebert-May, Beichner and Bruns 2004, 521-22). Our strategy has been to build ongoing agency use of linked administrative records and other sources of data for program evaluation by demonstrating value to the agency.

For each project, the NYU/Chicago/UMD teams build the core linked dataset that can be modified and expanded as engagement increases (see Figure 3). That infrastructure makes use of new technology (JupyterHub (21)), (Perez and Granger 2007, 21-9), with specific examples and code developed through the notebooks and the companion book (Foster, Ghani, Jarmin, Kreuter, and Lane 2016) so that participants can have direct, replicable, and high-value interaction with the data and with each other. The expectation is that networks will be formed, new data assets will be created, and useful reports and analyses will be generated (Figure 3). We have well over 175 students signed up from almost 60 government agencies. Agencies are providing data to be linked—across state, local and federal agencies—since the activity serves their missions.

The role of external funding was critical. The Census Bureau and the Laura and John Arnold Foundation were critical to the success of our initiative. The Census Bureau provided initial funding for a remote access data facility; the Arnold Foundation provided scholarships for the test cases.

FIGURE 3

Canonical Example of Linked Data for Policy

This approach is one example of how financial sustainability can be developed, once initial startup funding is made available. There are multiple options to revenue generation. Class tuition is certainly one, but it is also possible to build a membership model jointly with city agencies to study topics of cross-cutting interest. If the linked data are built on—using a "coral reef" approach—the university centers could partner with government agencies to develop customized reports; such an approach has been successful in other contexts.[12] There is already some interest in the classes being used as the basis for developing long-term state and local government and federal statistical agency data exchange partnerships.

Of course, there have been many other models in the past where universities have partnered with key communities of interest; the agricultural extension program is one example (McDowell 2003a, 31-50) (McDowell 2003b, 116-18) (Cash 2001, 431-45). There are certainly a number of centers that have been successful in their own right, as Dennis Culhane and Bob Goerge have demonstrated in their contributions to this volume. The question is how to scale to a national model. The role of private foundations could be critical here. The Alfred P. Sloan Foundation's call for a network of research facilities, and the Laura and John Arnold Foundation's substantial investments in research centers have led the way in building that mass. Scaling and building on the existing polycentric communities would seem natural, given that universities have both research and education missions, and agencies need both research and workforce training. It is important to note the importance of entry and exit in shaping the infrastructure (also identified by Ostrom (Ostrom 1990). In our model, successful activities (such as classes), aimed at serving the interests of researchers, policy-makers and government agencies, would grow and expand; those that were not as focused on stakeholder interests should

[12] See, for example, the model developed by the Institute for Research on Innovation and Science (iris.isr.umich.edu)

be allowed to die.

Infrastructure is long term in nature. Thus, attention must be paid to financial sustainability and to building a vibrant community of practice. There are some major investments that must be made to lower fixed costs so that more researchers become involved and more agencies provide data. For example, developing common agreed-upon legal and technical standards would likely create both a safer and more interoperable environment, while maintaining some flexibility. The white papers by Petrila, Culhane and Goerge have addressed the importance of establishing a coherent legal framework with standards for privacy protection, including Memoranda of Understanding and Nondisclosure Agreements; conflicting and confusing rules add risk and cost to data providers and cost and burden to researchers. The development of such frameworks can draw on the expertise of legal scholars, and ideally be the basis for standard-setting legislation. Other white papers by Foster and Culhane have addressed the importance of providing a template for technically safe environments, so that resources are not wasted by duplicating technical efforts. The federal investment in the Federal Risk and Authorization Management Program, or FedRAMP, standards is illustrative of the value of established standards in promoting access. [13] The development of standardized environments can draw on the considerable expertise of computer scientists.

The startup costs for building infrastructure are high. Private foundations could play a fundamental initial role. Private foundations could invest in providing support for developing and disseminating best practices for the technical security of data facilities, as well as disseminating information about disclosure limitation techniques. They could identify key policy foci—whether they be pathways out of poverty, improving community services, or stimulating

[13] www.fedramp.gov

entrepreneurship—and pay for scholarships or initial projects for the development of classes in these areas. New ideas could be generated by annual workshops that bring together experts on the many facets of the community. They would include data providers, policy-makers, and researchers (social scientists and computer scientists).

Where to go from here

This article has argued that the development of a new administrative data infrastructure has to be informed by thoughtful and deliberate engagement with appropriate stakeholders. Because of the legal and technical barriers associated with providing data access, serious resources must be devoted to reducing those barriers. Building an operational business model requires that stakeholders be engaged in a mechanism that builds reciprocity, reputation, and trust and that there are sufficient resources—financial, technical, and human—to make it all happen(Pardo 2014). The current person-by-person and ad hoc links between individual data producers and a few scholars is not sustainable over the long term.

A scalable and sustainable approach could involve developing a network of city/organizational partnerships where data producers, scholars, and policy-makers come together to work. One possible starting point is to build those partnerships around executive education–style classes. In those classes, agency staff could work together in secure state of the art neutral data repositories, working to produce useful projects and research that are mutually beneficial. If all projects made use of a common interoperable data infrastructure that served as the backbone for the safe use of linked administrative data across governmental units, it would both assure data providers that their data are being accessed in a secure and safe environment and reduce onerous, time-consuming and costly startup time.

A natural place for this to occur is at universities. If a group of universities in collaboration with governments built safe, credible high-service data centers that specialized in data linkage, usage, and dissemination and demonstrated their operational and policy value through training programs targeted at agencies' workforces, the world might beat a path to their door. We have seen how university based centers can support themselves and provide good operational policy analysis. Two illustrative examples are the Institute for Research on Innovation and Science at the University of Michigan, which is almost completely funded by member universities, andTulsa's MHealth network which now includes more than 4,000 providers and their patients, and demonstrated significant cost reductions and shown improvement in health outcomes

While each center is different, some core lessons are surprisingly similar. Each group started by focusing on a core problem. Integrated data and technology was treated as a tool to help solve the problem, not the primary goal in and of itself. The effort had strong and visionary leadership which convened and engaged key stakeholders around the problem and created an effective

governance system that ensured each had voice and a strong stake in the work. That leadership did see a sustainable system for data integration as an essential goal of the work, but it was always described as serving the large goals of health and efficiency. And the project started with demonstration support, while ultimately creating a system that was self-sustaining because of the demonstrated benefits to the producers themselves.

A massive step forward could be made if a consortium of foundations made major similar investments that would be competitively allocated across multiple centers. Such a consortium could create a mechanism for local governments to connect to national efforts to link data for the purposes of designing cutting edge policy/program pilots combined with proven and developing technologies for data access. This would both assure data providers that their data are being accessed in a secure and safe environment and reduce onerous, time-consuming and costly startup time. Since technical solutions are necessary, but not sufficient, to ensure sustainability: the approach should include the building of community data skills and associated human capital, which has the added benefit of ensuring sustainability if the project focus is in demand.

References

Catlett, C. et al. 2014. Plenario: An open data discovery and exploration platform for urban science. Bulletin of the IEEE Computer Science Technical Committee on Data Engineering. Available from http://sites.computer.org/debull/A14dec/p27.pdf.

Pardo, Theresa A. 2014. Making data more available and usable: A getting started guide for public officials. Presentation at the Privacy, Big Data, and the Public Good Book Launch. Available from http://cusp.nyu.edu/wp-content/uploads/2014/07/Pardo.pdf.

Ostrom, E. 1998. A behavioral approach to the rational choice theory of collective action: Presidential address, American Political Science Association, 1997. *American Political Science Review* 92:1–22.

Ostrom, E. 2010. Beyond markets and states: Polycentric governance of complex economic systems. *American Economic Review* 100 (3): 641–72.

Kleiman, N. 2015. *Striking a (local) grand bargain: How cities and anchor institutions can work together to drive growth and prosperity*. New York, NY: The National Resource Network.

President's Council of Advisors on Science and Technology. 2016. *Technology and the future of cities*. Washington DC: President's Council of Advisors on Science and Technology

Meyer, B. D., W. K. C. Mok, and J. X. Sullivan. 2015. Household surveys in crisis. *Journal of Economic Perspectives* 29:199–226.

Meyer, B. D., and R. M. Goerge. 2011. Errors in survey reporting and imputation and their effects on estimates of food stamp program participation. Discussion Paper. Washington, DC: U.S. Census Bureau, Center for Economic Studies.

Smith, K. 2014. Increasing researcher access to federal administrative data: A project of the Council of Professional Associations on Federal Statistics. Available from http://www.copafs.org/UserFiles/file/ IncreasingAccesstoFederalAdministrativeDataProject.pdf.

Abowd, J. M., J. Haltiwanger, and Julia Lane. 2004. Integrated longitudinal employer-employee data for the United States. *American Economic Review* 94:224–29.

Card, D., R. Chetty, M. Feldstein, and E. Saez. 2011. Expanding access to administrative data for research in the United States. White paper. Washington, DC: National Science Foundation. Available from https://eml.berkeley.edu/~saez/card-chetty-feldstein-saezNSF10dataaccess.pdf.

Jarmin, R. S., and A. O'Hara. 2016. Big data and the transformation of public policy analysis. *Journal of Policy Analysis and Management* 35 (3): 715–21.

Decker, R. A., J. Haltiwanger, R. Jarmin, and J. Miranda. 2016. The decline of high-growth entrepreneurship. VoxEU.org.

Ohm, P. 2014. In Privacy, big data, and the public good: Frameworks for engagement, eds. Julia Lane, V. Stodden, H. Nissenbaum, S. Bender, New York, NY: Cambridge University Press.

Meyer, B. D., W. K. C. Mok, and J. X. Sullivan. 2009. *The under-reporting of transfers in household surveys: Its nature and consequences*. Cambridge, MA: National Bureau of Economic Research.

Figlio, D. N., K. Karbownik, and K. G. Salvanes. 2015. *Education research and administrative data*. Cambridge, MA: National Bureau of Economic Research.

Tirole, J. 2014. Market failures and public policy. Nobel Prize Lecture.

Rochet, J-C., and J. Tirole. 2004. Two-sided markets: An overview. Institut d'Economie Industrielle Working Paper, Alle de Brienne.

Foster, Ian, R. Ghani, R. S. Jarmin, F. Kreuter, and Julia Lane. 2016. *Big data and social science: A practical guide to methods and tools*. New York, NY: Taylor & Francis Group.

Handelsman, J., D. Ebert-May, R. Beichner, and P. Bruns. 23 April 2004. Scientific teaching. *Science* 304 (5670): 521–22.

Granger, B. E., Jupyter, and JupyterHub. 2015 available at https://jupyter.org/

Pérez, F., and B. E. Granger. 2007. IPython: A system for interactive scientific computing. *Journal of Computing in Science and Engingeering* 9:21–29.

McDowell, G. R. 2003a. Engaged universities: Lessons from the land-grant universities and extension. *The ANNALS of the American Academy of Political and Social Science* 585:31–50.

McDowell, G. R. 2003b. Land-grant universities and extension into the 21st century: Renegotiating or abandoning a social contract. *Journal of Higher Education* 74:116–18.

Cash, D. W. 2001. "In order to aid in diffusing useful and practical information": Agricultural extension and boundary organizations. *Science, Technology, & Human Values* 26:431–45.

E. Ostrom, E. 1990. *Governing the commons: The evolution of institutions for collective action*. New York, NY: Cambridge University Press.

Notes