JULIA LANE

A Vision for Democratizing Government Data

Building an information marketplace about how government data are used can enable new types of informed governance, strengthen science, and engage the public.

During the earliest weeks of the COVID-19 pandemic, there was a desperate need for data to respond to the worst labor market crisis in almost a century. State-administered unemployment claims skyrocketed. Jobseekers, employers, state governors, and state legislators required practical information to address rapid and repeated shocks that put as many as 20 million people out of work. Some midwestern states, in need of quick ways of understanding what was happening, where, and to whom, found that their administrative data on certified unemployment claims could be restructured to get a much better idea of the impact the crisis was having on individuals. This gave decisionmakers access to information such as how long different groups of people were unemployed and how length of unemployment varied by gender, race, education, industry, geography, and the timing of individual layoffs in the pandemic. The restructured data were timely, relevant, and actionable in a rapidly changing environment where evidence was desperately needed, and provided local information that could be used to allocate resources accordingly.

As it turns out, these important data were available because of the Midwest Collaborative, a multistate activity that began in 2018 and was subsequently funded by philanthropic foundations and state and federal agencies. Although US states have a history of leading the way in the use of data, they have often been hobbled by the reality that each state's data end at its borders, as residents cross state lines to attend school, go to work, or change jobs. To enable cross-state data collaborations, the Midwest Collaborative made use of a secure data-sharing platform combined with a hands-on training program—providing the core infrastructure necessary to create better understanding of how to use the data to create evidence.

The Midwest Collaborative is just one of many projects inspired by recommendations from the federal Commission on Evidence-Based Policymaking, a bipartisan effort established by Senator Patty Murray (D-WA) and Speaker Paul Ryan (R-WI) in 2016 to marshal data and evidence to guide and improve the effectiveness of government investments. A subset of the commission's recommendations was incorporated into the Foundations for Evidence-Based Policymaking Act of 2018 (the Evidence Act), which requires all federal agencies to submit yearly systemic plans for the collection, storage, and analysis of data. The law established the basis for a national approach to evidence-building, set ground rules for privacy and statistical efficiency, and complemented the 10-year Federal Data Strategy. When the act was passed, Senator Murray noted, "Whether you think we need more government or less government—you should agree that we should at least have better government."

Labor market outcomes are but one of a myriad of issues where better evidence, if unlocked and democratized, could help inform and enhance policymaking. Health care, education, social services, and infrastructure planning—

not to mention investments in science and technology and workforce development—could also benefit from such an informed approach. But the challenge of gathering and analyzing data to plan for the future is a longstanding issue for government agencies as well as for industry. In the 1990s, Lew Platt, then chief executive office of Hewlett-Packard, famously said, "If HP knew what HP knows, it would be three times more profitable."

Although the challenges of developing a system that can take full advantage of existing data and evidence are significant, such a system can realize three goals common to many policy areas and challenges. First, evidence can ultimately reveal which strategies work, how they work, and what their outcomes are. When thinking about investments in science and technology (S&T), for example, evidence can open the "black box" between research funding and societal benefits, demonstrating what actually happens when money is spent on particular programs and fields. This knowledge can enable the second goal: strategic planning and investment in programs, processes, places, and people to increase the likelihood of achieving targeted outcomes. And finally, by "knowing what is known," policymakers will become better equipped to make timely and effective decisions grounded in granular, useful, linked data—as the midwestern states did with labor force data during the pandemic. The result will be more thoughtful, productive, and transparent policymaking that is more likely to accomplish public goals.

Already, the Evidence Act has inspired widespread action across the federal government. Federal agencies have appointed chief data officers and chief evaluation officers and established interagency councils for both groups. The White House Office of Management and Budget (OMB) has directed each agency to develop learning agendas and evaluation plans. The Interagency Council on Statistical Policy, a group of federal officials who advise OMB in coordinating the federal statistical system and setting statistical policy, has established a Standard Application Process for outside researchers to gain data access. Meanwhile, the advisory committee established by the law will release its final report in October 2022. Now that the scaffolding is in place, the next step is to build a community of practice around the data, so that knowledge about data and measurement can be shared.

To be clear, in harnessing evidence for policymaking, the problem is rarely a lack of data. Government data are everywhere: generated by federal and state programs administering tax, labor, justice, welfare, and education policies, for example, and from comprehensive surveys—such as the Decennial Census and the Survey of Earned Doctorates—that have been run by federal statistical agencies for decades. The problem is that analysts often don't know how to use the data once they get access. One reason is that the data are often poorly documented. Another is that many data sets have been siloed from one another from conceptualization to collection and use. The silos result from legal, regulatory, and other hurdles to sharing—including valid concerns about privacy and confidentiality. Even when data sets are brought together, they are often tough to accurately match against one another or against outside data, biasing analysis in arbitrary and potentially harmful ways. Making the problem worse, analysts frequently don't—or can't—know how these data sets have been previously used. Each new analysis starts without building on established knowledge, which wastes time and potentially introduces error. In sum, harnessing data for evidence requires discovering information about how relevant data sets are used, identifying the experts, and sharing community knowledge so that governments can be much more productive. But because governments don't know what governments know, there are massive challenges associated with sharing knowledge, which hinder the deep assessments necessary to provide truly powerful evidence for policymaking.

What a data information marketplace could do

The private sector can provide inspiration for indexing and making information available about how data are used: look at Amazon.com. Before the company's arrival, people seeking information about books either went to libraries or to bookstores, relying on book reviews and informal recommendations from friends. Jeff Bezos changed that by giving people the information they wanted: which books addressed topics they were most interested in, how similar they were to other books, and which ones were highly rated by the community. In other words, the breakthrough was that Amazon provided customized, useful information in a way that was easy for people to find and understand. In this particular sense, Amazon democratized access to books—and later to many other products—by lowering search costs and creating an information marketplace of people contributing knowledge through user reviews and purchases of related books on similar topics.

The government needs to do something similar by building an information marketplace for evidence and data. The impact could be transformational. Amazon sells retail goods, which tend to be bought and sold once or just a few times, but ideas and knowledge about data can be reused over and over again, with benefits continuing to accrue as the community learns and shares more knowledge about how to create evidence from data.

At the moment, this effort is in its infancy, but a cluster of initiatives are starting to coalesce. The Evidence Act and the Federal Data Strategy have created incentives for agencies to provide more transparency—a more public information marketplace—about their funding, data investments, and how their data are used. One effort I have been involved in seeks to automate the process of understanding how publicly funded data sets are being used. We set up a competition, Show US

the Data, in which more than 1,600 data science teams competed to develop the best machine learning approaches to understand how government data are being used, by whom, and for what purposes. The competition was hosted on kaggle.com, an online community of data scientists that is a subsidiary of Google. The results of the competition, which were highlighted in a 2021 conference, showed the power of artificial intelligence to search and discover how data sets are being used in scientific publications—successfully identifying topics and the experts utilizing the data, and even pointing to the documents containing published research that used federal data sets.

These search and discovery tools are intended to be public. A pilot project including a dashboard and a prototype interface that provides key information about how the data are used (called an application programming interface) is being sponsored by agencies including the National Center for Science and Engineering Statistics at the National Science Foundation (NSF) and the National Center for Education Statistics at the Department of Education. Making these tools a community asset is a critical part of increasing the capacity to use data across government, academia, and private industry, with the goal of creating a broader cultural shift toward action grounded in data.

Much in the same way that Amazon's infrastructure to collect and interpret user preferences changed the way retail works, agencies will soon be able to see which data sets are most in demand and which are underutilized, discovering new areas and topics for which their data provide insights. This can help inform investment decisions for future data collection and quality improvements, as well as inspire collaborations with other federal agencies that have complementary data. And then there are the second-order effects, as researchers discover more about other scientists with whom they share common interests by viewing authors, articles and papers, and related data sets in an easily arrayed and searchable format. This knowledge about data use will be available to a diverse spectrum of researchers from many types of institutions across the country. As more researchers combine efforts, their work may move faster, include more diverse insights, and become more replicable. As policymakers learn to understand how data are used to anticipate and measure outcomes, policies can become more effective and more targeted. And the public can begin to ask for policies that more reliably and effectively deliver the desired results at national and local levels.

A key part of this process will be building trust with the public, which provides data to the government through surveys, censuses, and by enrolling or participating in government programs. Citizens must trust that the access to data will generate evidence that improves policies—and also trust that their privacy will be respected and confidentiality will be protected.

A vision for democratizing data

One aspect of earning public trust is demonstrating that data are being used to improve the lives of citizens and taxpayers. A data-driven approach could, for example, assist in the knotty challenge of planning investments in research and development and the scientific workforce to foster economic growth. An extensive literature ties regional economic development with the presence of research universities, and innovative new businesses are often located near universities precisely because it's easier to hire researchers. It's long been known that ideas travel through interpersonal interactions. Certainly, when the White House science advisor John Marburger III called for a "science of science policy" in 2005, he was thinking of tracing the flows of people and ideas from the bench to the workplace.

Almost 20 years after Marburger's call, there are now better ways to ensure that investments, such as those in the CHIPS and Science Act of 2022, have the best chance of leading to the desired outcomes. It's possible to measure the links between S&T investments—including in critical technologies—and resulting economic activity by looking at the career outcomes of graduate students and postdoctoral researchers, and at the growth of companies that provide goods and services related to research grants. A pathbreaking initiative at the Institute for Research on Innovation & Science (IRIS) involves secure automated approaches to pull deidentified information on the workforce from university human resources departments, sponsored projects, and finance systems on all individuals engaged in grant-supported research. IRIS's data now cover more than 40% of all academic R&D, including monthly transaction information on more than 535,000 sponsored projects. What this means is that researchers and policymakers can directly "see" how investments impact individuals, businesses, universities, and regions. These data complement existing statistical surveys and make tracing the impact of S&T investments more scientific.

Rather than relying on anecdote and supposition, IRIS data can show how students and postdocs employed by universities subsequently move into industry, positioning them to transmit the new scientific knowledge they helped create. Likewise, businesses that sell high-tech equipment to power these projects may be well positioned to develop their own innovations. The resulting knowledge gains can accrue to the employing firms, their workers, and ultimately to society. These data on S&T investments can be further connected with state and federal education and workforce data, enabling state agencies to ensure that firms can hire workers with the appropriate credentials, thereby matching workers to the resulting high-wage jobs. This vision of matching labor demand and workforce skills has been a goal of agencies for decades, but it has been stymied by the lack of sufficient data. Now the data and evidence can be linked to make sure that S&T stimulates economic growth in places that need it.

In keeping with the vision of building a community of knowledge about how data are being used, IRIS supports a large and growing network dedicated to using and enhancing these data. Including nearly 370 researchers from more than 80 institutions, the community has added valuable new data assets to its infrastructure, such as federal survey and transaction data and privately held and collected resources. To sustain this community, IRIS data are made available to approved and vetted researchers through well documented annual research releases with multiple portals for research access, including the potential to partner with state collaboratives like the Midwest Collaborative. Within the next few years, this virtuous circle of more data and more analysis will have made many things that are now buried—such as the expected impact of research spending on local skills, local businesses, and local economies—measurable and actionable. But for this to work, researchers must be able to connect with a community of practice so that they can share ideas and build on each other's research, as well as communicate the value of data use to the public.

edge inputs to research projects and on scientists' career trajectories in research-intensive firms. While the scientific enterprise has long struggled to spell out its benefits to society in a way that resonates with the public, IRIS data show a way to move beyond proxy indicators to measure directly what policymakers—and society—care about.

Much more can be done. IRIS could be expanded to all universities, particularly minority-serving institutions, so that more ideas could be included. Existing partnerships with statistical agencies could be strengthened so that new measures of critical technologies could be tied to standard economic statistics and industry classifications. Federal and state education and labor agencies—now forewarned and informed by evidence—could work with science funding agencies to proactively invest at all levels in the necessary workforce training and skills so that cutting-edge ideas can be effectively adopted and deployed.

Just as Amazon changed how Americans buy goods and services, a public-sector information marketplace could change how government and industry make decisions. Such a change may well disrupt the way that

A data-driven approach could assist in the knotty challenge of planning investments in research and development and the scientific workforce to foster economic growth.

And this is where the new infrastructure for evidence could change and increase the impact of investments in science and technology. Indicators used by policymakers have often relied, of necessity, on indirect proxy measures. One proxy for effectiveness has been measuring the number of publications that resulted from investments, but this reveals little about the processes or people involved, may not correlate with economic or societal impact, and for some types of research publications may not even be the main product.

IRIS data, which is more complete and includes key features missing from bibliometric data, can now trace how research and training influence career trajectories. This allows policymakers, funders, and institutional administrators to fine-tune programs and approaches to meet specific objectives. For example, the data include information on full teams: faculty, staff, and students, including those who may not appear on author lists. This provides a unique lens on diversity, equity, and inclusion issues, as seeing all people employed on grants allows for analyses of who gets what kinds of credit and what the implications are for careers. It also enables a whole series of reports about economic impact not possible before, such as on research-intensive companies that supply cutting-

government works. The increasing influence of Amazon's interface empowered many new businesses to serve specific communities—but it also helped send many brick-and-mortar stores into bankruptcy. Likewise, while the focused use of evidence in policymaking may empower governments to produce targeted information for local decisionmaking and open the door for local experimentation and course correction, it may also reduce funding to long-established institutions that fail to produce equitable social improvements. Agencies newly empowered with the ability to measure specific impacts will also have a greater responsibility to spell out a theory of change: exactly how those expected impacts from scientific research will be achieved. To take just one example, the broader impacts criterion, which has long required NSF proposals to consider the societal ramifications of research, has been difficult to measure and evaluate. But with an evidence-based approach, it is possible to specify broader impacts as an outcome, measure them, and fine-tune investments to obtain those impacts more effectively.

A key tenet of the marketplace describing how data are used is that the information is owned by everyone, and everyone has a part to play in contributing knowledge.

Already, the prototype for the data usage dashboard contains information about how 51 different data sets—from NASA, the National Oceanic and Atmospheric Administration, NSF, and the US Department of Agriculture, among others—are being used. In building this wider system, there is a role for many stakeholders to play in validating the information, which will provide a mechanism for continuous improvement and added value. And, as the IRIS example illustrates, meaningful data sets can also come from sources outside federal and state governments, including from universities and companies. The very process of making available more information about how data are used could galvanize a community of researchers, analysts, and agencies; inspire new uses; and create new evidence.

This move toward evidence-driven policy will also transform civil society in important ways. Consider what is possible when data sets, and information about their use, become widely shared and democratized. In 2013, Johns Hopkins University developed SciServer, a platform that allows large groups of citizen scientists to collaborate on categorizing data. In particular, the project made finding massive amounts of astronomical data easy and intuitive for researchers, students, and the public. Galaxy Zoo, one citizen science project on SciServer, has resulted in reliable classifications for hundreds of thousands of galaxy images, assembled by more than 100,000 volunteers. But the community goes beyond astronomy fans: the SciServer system now includes oceanography, mechanical engineering, social sciences, and finance. In addition, SciServer features a learning environment that is being used in K–12 and university education. The more that people learn about data—and the more they can contribute new ideas about how data can be used—the greater the potential for data and evidence to transform society.

A public marketplace of ideas will radically change policymaking in ways that are impossible to foresee. The effects may be profound. Consider all the change that Amazon's interface wrought. By offering a new way to shop, it gave people more tools to save time and money. As consumers revealed their preferences, industries were reshuffled, upending the business models of publishers, manufacturers of household goods, and the entertainment industry, particularly with the rise of streaming services. The effects of the interface have cascaded across the economy—for better and worse—affecting infrastructure, logistics, shipping, labor relations, and the way that data are stored in the cloud. These economic changes were accompanied by shifts in norms and expectations, so that by the time the pandemic hit, it was possible for segments of society to shift to remote working and shopping. So too, a marketplace of data use is a force powerful enough to disrupt the status quo. Today's decisionmaking processes—highly centralized and based on anecdotes, rough proxies, and years-old data—may one day be seen as a relic of the past.

As efforts to create a marketplace providing more information about how data are used move forward, there are three potential pitfalls of special concern. The first is what's known as Campbell's law: "The more any quantitative social indicator is used for social decisionmaking, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor." As systems develop, vigilant community-driven monitoring will be needed to limit abuse, ensure inclusivity, and protect valuable features that are difficult to measure from being left aside. Second, insufficient protection of privacy and confidentiality, including disclosure of intellectual property and national security risks, could lead to a loss of trust and a consequent disincentive to participating in sharing knowledge. It will be vital to institute rules for sharing information about data use that make sense. Finally, it is important to recognize that investments in social goods, including education and S&T, take a long time to bear fruit. The timeframes for assessment should be realistic and calibrated to the scale and complexity of the effort.

With these risks in mind, there remain many good reasons for a wider number of players to participate in an expanded data use infrastructure. Researchers can achieve more visibility and wider recognition, find collaborators more easily, and connect with a broader research community. Ultimately, these incentives can lead to sharing code and insights about data quality that will improve the replicability, efficiency, and integrity of science. Agencies could better connect with each other and prioritize high value areas as they discover common topics of interest. They will also be better able to communicate the utility of their investments in data and generate resources to support investments in data quality, all of which will provide more value to taxpayers. Additional transparency may help improve the public's trust in government over time. And, of course, implementing a government marketplace with better information about data will be key in realizing the goals of the Evidence Act—leading to, as the Commission on Evidence-Based Policymaking put it in their final report, "a future in which rigorous evidence is created efficiently, as a routine part of government operations, and used to construct effective public policy."

Julia Lane *is a professor at New York University's Wagner Graduate School of Public Service and has been involved in founding many data initiatives to serve the public good, including the Longitudinal-Employer Household Dynamics Program, the STAR METRICS/UMETRICS program, and the Coleridge Initiative. She currently serves on the Advisory Committee on Data for Evidence Building. She is the author of* Democratizing Our Data: A Manifesto *(The MIT Press, 2021).*