DEMOCRATIZING OUR DATA

A MANIFESTO

JULIA LANE

First MIT Press paperback edition, 2021

This book was set in Adobe Garamond and Berthold Akzidenz Grotesk by Jen Jackowitz. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Names: Lane, Julia I., author.
Title: Democratizing our data : a manifesto / Julia Lane.
Description: Cambridge, Massachusetts : The MIT Press, [2020] | Includes bibliographical references and index.
Identifiers: LCCN 2019057265 | ISBN 9780262044325 (hardcover)
ISBN 9780262543521 (paperback)
Subjects: LCSH: United States--Statistical services--Standards. | Statistical decision--Standards. | Quantitative research--Standards.
Classification: LCC HA37.U55 L34 2020 | DDC 352.7/50973--dc23
LC record available at https://lccn.loc.gov/2019057265

10 9 8 7 6 5 4 3 2

Contents

Preface

Not long ago I had lunch with one of the top statisticians in the country, who looked across the table and said, "The information needs of people making important decisions are changing so rapidly. It's difficult for the federal government to meet those needs—the system is not designed for rapid change. We need to rethink how people can get consistently high-quality information when they need it from a trustworthy source. Our future depends on it." Another former government statistician put it more succinctly: "The federal system is broken, and I don't know if anything can fix it. So many good people, and nothing ever changes." A colleague who has spent over thirty years working within the system believes the problem is that even when senior management knows what to do, there is no incentive to take big risks and rock the boat, so the safest thing for both managers and frontline staff is to continue doing what they have always done and make very small incremental changes. And that does not meet today's needs.

Think of the massive impact of the coronavirus on jobs and society in 2020. Governments at all levels urgently needed numbers that could tell them how many jobs were lost and how the

most vulnerable in society were affected; the infrastructure was not there.

This book is intended to drive change in the system that the United States uses to produce public statistics. It is not a call for more funding, although lack of funding has contributed to the current crisis. It is a call to fundamentally reorganize data production.

The reasons are clear. Access to high-quality information to make good decisions is necessary for society to function. Reliable, accurate, timely information levels the playing field for businesses and individuals alike. It is necessary at every level of our economy and society—to help small businesses succeed, schools serve parents and students, central banks make sound policy, and people make major life decisions.

The need for change is also clear. Costs to collect information through traditional means continue to increase, and response rates to government surveys continue to decrease. Government-produced data should accurately represent American economic and social activity, and our current system is badly strained and at high risk of future failure. Our democracy is threatened without timely, relevant data and evidence that reflects our economy and society, and we should be outraged that our system has fallen behind. This book is a wake-up call for Americans to understand how our data are used to create important information underlying major decisions that affect our lives every day, why the current system is breaking, and the immediate steps necessary to fix it.

The path outlined in the book is the result of over twenty-five years of experience working with data in academia, many levels of government, and the private sector. Of course, no one

can possibly understand the workings of the entire system, but I stand on the shoulders of the writings of many giants, not least of whom is Janet Norwood, a noted economist and statistician who headed the US Bureau of Labor Statistics for thirteen years.

This is the time to effect change. There are new data, new tools, and new technologies that can be combined in new ways to create new evidence. There are enough people of good will with enough determination to get things done. Recent legislation has created an opportunity to rethink the organizational data infrastructure. New legislation could take advantage of this golden moment and truly democratize our data.

The time is now.

Like any author, I owe enormous debts to the many colleagues with whom I have worked over the years. I have been privileged to work with incredibly dedicated and visionary colleagues at New York University's Coleridge Initiative, and our associated university and agency partners. I am incredibly grateful to all the staff in the federal statistical system and in programmatic agencies at the federal, state, and local government levels who work hard to make a difference, often against daunting odds. I name just a few in the book, but literally hundreds of people contributed to the LEHD (Longitudinal Employer–Household Dynamics) program, to IRIS (Institute for Research on Innovation and Science) at the University of Michigan, and to the Coleridge Initiative. I am indebted to all of them.

The philanthropic foundations, particularly Schmidt Futures, the Alfred P. Sloan Foundation, the Overdeck Family Foundation, and the Bill and Melinda Gates Foundation, have been game changers in placing their trust in the work that we

have been doing and enabling us to demonstrate what can be done.

My developmental editor David Weinberger helped shape and reshape this book with constant good humor and wise suggestions. Ian Glennon provided the initial research, particularly for chapter 2. Jason Owen Smith and Nancy Calvin-Naylor provided very useful comments on chapter 4. Paco Nathan, Jonathan Morgan, Ian Mulvany, and Drew Gordon were extremely influential in the discussion of automation in chapter 5. Stefan Bender, Nick Greenia, Frauke Kreuter, Nancy Potok, Bryant Renaud, and Brock Webb provided valuable suggestions and input at all stages of the manuscript. Mike Holland provided great assistance in reviewing the entire book, particularly in the discussion of federally funded research and development centers in chapter 7. All remaining errors are, of course, my own.

My MIT Press editor, Emily Taber, has provided an unbelievable amount of support and guidance—well beyond anything I deserved.

Finally, I owe my greatest debt to my husband, Dennis Glennon, who has put up with me during this process—and for many years before that!

There are massive challenges to be addressed. The national statistical system—our national system of measurement—has ossified. Public agencies struggle to change the approach to collecting the statistics that they have produced for decades—in some cases, as we shall see, since the Great Depression. Hamstrung by excessive legislative control, inertia, lack of incentives, ill-advised budget cuts, and the "tyranny of the established," they have largely lost the ability to innovate or respond to quickly changing user needs.[3] Despite massive increases in the availability of new types of data, such as administrative records (data produced through the administration of government programs, such as tax records) or by digital activities (such as social media or cell phone calls), the US statistical agencies struggle to operationalize their use.[4] Worse still, the government agencies that produce public data are at the bottom of the funding chain—staffing is being cut, funding is stagnant if not being outright slashed, and entire agencies are being decimated.[5]

If we don't move quickly, the cuts that have already affected physical, research, and education infrastructures[6] will also eventually destroy our public data infrastructure and threaten our democracy. Trust in government institutions will be eroded if government actions are based on political preference rather than grounded in statistics. The fairness of legislation will be questioned if there is not impartial data whereby the public can examine the impact of legislative changes in, for example, the provision of health care and the imposition of taxes. National problems, like the opioid crisis, will not be addressed, because governments won't know where or how to allocate resources. Lack of access to public data will increase the power of big busi-

nesses, which can pay for data to make better decisions, and reduce the power of small businesses, which can't. The list is endless because the needs are endless.

This book provides a solution to the impending critical failure in public data. Our current approach and the current budget realities mean that we cannot produce all the statistics needed to meet today's expectations for informing increasingly complex public decisions. We must design a new statistical system that will produce public data that are useful at all levels of government—and make scientific, careful, and responsible use of many newly available data, such as administrative records from agencies that administer government programs, data generated from the digital lives of citizens, and even data generated within the private sector.

This book will paint a picture of what this new system could look like, focusing on the innovations necessary to disrupt the existing federal statistical system, with the goal of providing useful and timely data from trusted sources so that we, the people, have the information necessary to make better decisions.

WHY IT MATTERS

Measurement is at the core of democracy, as Simon Winchester points out: "All life depends to some extent on measurement, and in the very earliest days of social organization a clear indication of advancement and sophistication was the degree to which systems of measurement had been established, codified, agreed to and employed."[7] Yet public data and measurement have to be paid for out of the public purse, so there is great scrutiny of costs and quality. The challenge public agencies face is that, as

Erik Brynjolffson, the director of MIT's Initiative on the Digital Economy, points out, we have become used to getting digital goods that are free . . . and instant and useful. Yet in a world where private data are getting cheaper, the current system of producing public data costs a lot of money—and costs are going up, not down. One standard is how much it costs the Census Bureau to count the US population. In 2018 dollars, the 1960 Census cost about $1 billion, or about $5.50 per person. The 1990 Census cost about $20 per head.[8] The 2020 Census is projected to cost about $16 billion, or about $48 per head.[9] And the process is far from instant: Census Day is April 1, 2020, but the results won't be delivered until December.

Another standard is the quality of data that are collected. Take a look, for example, at the National Center for Health Statistics report to the Council of Professional Associations on Federal Statistics.[10] Response rates on the National Health Interview Survey have dropped by over 20 percentage points, increasing the risk of nonresponse bias, and the rate at which respondents "break off" or fail to complete the survey has almost tripled over a twenty-year period.

As a result, communities are not getting all the information they need from government for decision-making. If we made a checklist of features of data systems that have made private sector businesses like Amazon and Google successful, it might include producing data that are: (1) real-time so customers can make quick decisions; (2) accurate so customers aren't misled; (3) complete so there is enough information for the customer to make a decision; (4) relevant to the customer; (5) accessible so the customer can easily get to information and use it; (6) interpretable so everyone can understand what the data mean;

(7) innovative so customers have access to new products; and (8) granular enough so each customer has customized information.

If we were to look at the flagship programs of the federal system, they don't have those traits. Take, for example, the national government's largest survey—the Census Bureau's *American Community Survey* (ACS). It was originally designed to consistently measure the entire country so that national programs that allocated dollars to communities based on various characteristics were comparing the whole country on the same basis. It is an enormous and expensive household survey. It asks questions of 295,000 households every month—3.5 million individuals a year. The cost to the Census Bureau is about $220 million[11] and another $64 million can be attributed to the respondents in the value of the time taken to answer the questions.[12] Because there is no high-quality alternative, it is used in hundreds if not thousands of local decisions—as the ACS website says, it "helps local officials, community leaders, and businesses understand the changes taking place in their communities."[13] In New York alone, the police department must report on priority areas that are determined, in part, using ACS poverty measures,[14] pharmacies must provide translations for top languages as defined by the ACS,[15] and the New York Department of Education took 2008 ACS population estimates[16] into account when it decided to make Diwali a school holiday.

Yet while reliable local data are desperately needed, the very expensive ACS data are too error prone for reliable local decision-making. The reasons for this include the survey design, sample sizes that are too small, public interpretation of margins of error when sample sizes are small, and lack of timely dissemination of data.

I'll discuss some of the details of these reasons in chapter 2—but one core problem is the reliance on old technology. The data are collected by means of mailing a survey to a random set of households (one out of 480 households in any given month). One person is asked to fill out the survey on behalf of everyone else in the household, as well as to answer questions about the housing unit itself. To give you a sense of the issues with this approach: there is no complete national list of households (the Census Bureau's list misses about 6 percent of households), about a third of recipients refuse to respond, and of those who respond, many do not fill out all parts of the survey.[17] There is follow-up of a subset of nonresponders by phone, internet, and in-person interviews, but each one of these introduces different sources of bias in terms of who responds and how they respond. Because response rates vary by geography and demography, those biases can be very difficult to adjust for.[18] Such problems are not unique to the ACS; surveys in general are less and less likely to be truly representative of the people in the United States and the mismatch between intentions and reality can result in the systematic erasure of millions of Americans from governmental decision-making.

Statistical agencies face major privacy challenges as well. The increased availability of data on the internet means that it is much easier to reidentify survey respondents, so more and more noise has to be introduced into the data in order to protect respondent privacy. This noise results in reduced data reliability, particularly for small populations.[19] For example, the Census Bureau is systematically making data worse to protect privacy.

> Census data from 2010 showed that a single Asian couple—a 63-year-old man and a 58-year-old woman—lived on Liberty Island, at the base of the Statue of Liberty. That was news to

mation to travelers about the best way to get from A to B. Their business, and others like them, replaced the business of producing physical maps that were difficult to use and often out of date.

Since that solution doesn't work for governments, we need to identify what parts of the federal statistical system should be retained and what parts should be reallocated. The challenge is identifying an alternative. An important argument in this book is that the Data Revolution makes it possible.

Changing the workforce is critical. For data to have value, the employees in an organization have to have the skills necessary to translate that data into information. The entire structure of the private sector has been transformed in the past twenty years to reflect the need for such skills. In 2018, one of the biggest US companies, Facebook, grounded in data, had a market value per employee of about $20.5 million, with very little physical capital and a workforce skilled in manipulating data. Twenty years ago, one of the biggest US companies, General Motors, grounded in manufacturing, had a market value of $230,000 per employee, with a great deal of physical capital and a skilled manufacturing workforce.[24] Such change is difficult to effect in the public sector. Government salary structures make it difficult to hire and retain enough in-house data analysts, let alone respond quickly to reward employees for acquiring new skills. The government is competing against Facebook and Google not only for salaries but also prestige. The occupational classification of "data scientist" didn't even exist in the federal government until June of 2019.[25] Open source tools, like Python, which are commonly used in private-sector data analysis, are regarded with suspicion by many government IT organizations. The pressures to meet existing program needs make it difficult for agency

staff to try something new, and while failure is celebrated in the private sector, it can be career ending in the public sector. These combined challenges have led to the current situation—agencies cannot get the significant resources necessary to make use of new data, and because they don't use new data, they don't get new resources.

New products that respond to community needs must be developed. There is a huge opportunity to do so. The amount of new data available is overwhelming.[26] Real-time data can be collected on cell phones, from social media sites, as a result of retail transactions, and by sensors or simply driving your car. Turning the data into useable information requires a very different set of skills than the ones deployed in the survey world. Data need to be gathered, prepared, transformed, cleaned, and explored, using different tools. The results need to be stored using new database tools, and analyzed using new techniques like machine learning and network analysis. Visualization and computational techniques are fundamentally different with data on a massive scale, rather than simply tens of thousands of survey answers. The privacy issues are different, as are the requirements for data search and discovery and reproducibility.[27]

While today's data world is, in many ways, a Wild West, data being produced for the public sector need to be designed carefully. The key elements of the federal statistical infrastructure are too important to lose: we need to expand the current statistical system to think about how public data should be produced, and how they must be trustworthy and measured well and consistently over time, and how confidential information should be protected. A world in which all data are produced by a market-driven private sector could be a dangerous one—where there are

many unidentified or unreported biases; where privacy is not protected; where national statistics could be altered for the right price; where if a business changes its data collection approach, the unemployment numbers could skyrocket (or drop); where respondents' information could be sold to the highest bidder.[28]

Action is required because the way governments produce statistics won't change by itself. In the private sector, market forces create the impetus for change, because organizations that don't adapt are driven out of business. There's no similar force driving government change. Over the past thirty years, I've worked with people at all levels of government—federal, state, county, and city—in the United States and throughout the world. I've developed tremendous respect and admiration for the highly skilled and dedicated workforce that brings us the information driving our economy. These professionals know what needs to be done to make change happen. Hundreds of studies have provided useful recommendations. But when, in the course of thirty years, hundreds of good people try to change the system and the system doesn't change, it's clear nothing is going to happen without disruption.

This book proposes a new and, yes, disruptive approach that spells out what to do. It keeps the best elements of the current model—the trust, professionalism, and continuity—while taking away the worst elements—the bureaucracy and rigidity. It proposes a restructuring to create a system that will:

1. Produce public statistics that are useful at all levels—federal, state, and local.
2. Empower a government workforce to innovate in response to new needs.

3. Create a trusted organization that is incentivized to respond to community demand.

This is a golden moment to rethink data use by establishing, codifying, agreeing to, and employing new systems of measurement. Governments at the state and local levels are upping their investment in developing analytics teams to support better management. At the federal level, Congress passed the Foundations for Evidence-Based Policymaking Act of 2018 and the White House published the first Federal Data Strategy. Both efforts require agencies to invest in data, analytical approaches, and more thorough evaluation activities to get rid of programs that don't work and expand programs that do. Many state and local governments are turning to data- and evidence-driven decision-making and forming new partnerships with universities, with the private sector, and with each other to do so.

The challenge is making sure that the focus is on creating new value rather than creating new processes. In the private sector, thousands of firms get started; only the successful ones survive. Federal, state, and local governments don't have the pressure of failure, so their response is to establish new positions. The federal government's response has been to require each of the twenty-four major US government agencies to have a chief data officer (CDO), a chief evaluation officer, and a senior statistical officer; at last count, nearly fifty states, counties, and cities had also hired CDOs. Ensuring that the people in these positions have the support or control that they need to succeed is essential: if an ineffective system is introduced in government, it can be hard to course-correct. Governments at all levels are investing in training their staff to acquire data skills; it will be

similarly critical to ensure those investments are substantive rather than perfunctory.

Chapter 2 goes into the details of how key indicators of factors such as economic activity are measured, and why measurement is so difficult. The current system was designed to be great at counting guns and butter for World War II supplies. But although manufacturing and agriculture are much less important now than they were a century ago, the government continues to be much better at counting manufacturing output (648 industry categories) than finance and insurance output (89 industry categories).[29] Why? It's like the old story of a drunk looking for his lost keys under a streetlight. A policeman stops to help, and after a fruitless search, the policeman asks if the keys were really lost there. The drunk says, no—he lost them in the alley. The policeman, of course, asks the drunk why he isn't looking in the alley, and the drunk answers, "This is where the light is." That's government. The public sector continues to look under the streetlight because it's so difficult to change.

We'll highlight the issues by discussing how people work and generate products and why governments need to rethink how and what they measure and why they measure it.[30] We'll start by talking about one of the most important measures that government produces—gross domestic product, or GDP, which is the international measure of economic activity in each country. GDP is sometimes also used to measure economic well-being, but as we'll see later on, it is not designed for that, so it doesn't measure it very well. Digital technologies have fundamentally changed the way in which business is done and people interact.[31] This leads to huge and important questions about how to measure twenty-first-century activity.[32] How should

Of course, a key feature of this new system is an approach to inspiring civil servants and creating and empowering an engaged and innovative workforce. That is the focus of chapter 6, where I draw on my experiences working with federal, state, and local agencies.

The last chapter proposes a new organizational model that is inspired by institutional success in other areas. This new model has the potential to transform the world of measurement and statistics in a way that democratizes both access and use.

Failure to act could threaten our democratic infrastructure. The slow decline in data quality combined with increasing costs could well lead budget-conscious legislatures to simply save money by shutting down parts of the statistical system. We must begin the process of careful restructuring now—while we can—so that all Americans will be fairly represented when our democratic institutions are charged with making evidence-based decisions. As Janet Norwood said, this is essential so that the people making these decisions "can rely on accurate and objective statistical information to inform them of the choices they face and the results of choices they make."[34]

We need to build a new public data infrastructure that democratizes data.

2 THE CURRENT STATE OF PLAY

The problems of our public measurement system are perhaps best illustrated by detailing the problems of one of our best-known national statistics—gross domestic product—which counts the market value of all final goods (in other words, the finished products: paper, not timber) and services produced within a country in a given period of time. When analysts talk about the strength of the American economy, or national income accounts, they are often talking about our GDP. However, the imperfections of GDP are well known, and have been well known for a very long time. It has been referred to as a "Frankenstein's monster" and an "arbitrary, oversimplified human invention that we slavishly follow" . . . a modern "cult."[1] Sadly, this is not a new revelation. Paraphrasing an eloquent speech by presidential candidate Robert F. Kennedy in 1968: "We judge the United States by production—we count air pollution, cigarette advertising, locks for our doors and jails for people who break them. We count the destruction of the redwood, the production of nuclear warheads, but not the health of our children, the beauty of poetry, the intelligence of our public debate or the integrity of our officials. . . . It measures everything in short, except that which makes life worthwhile."[2]

Ouch.

But the GDP measure is not the only flawed measure—it is only illustrative. Another well-known set of statistics used to describe the strength of the economy is for the levels of unemployment and employment, and there is a parallel set of problems associated with these as well. The most well-known measure is drawn from the Current Population Survey that surveys about 60,000 households every month to find out people's employment status. The reports of the results, which are issued by the Bureau of Labor Statistics on the first Friday of every month with great fanfare, can change government policies and the direction of markets. Election campaigns are often won or lost on the basis of employment statistics. The problem is that the jobs numbers are increasingly incomplete, both because of the development of the gig economy[3] and because the survey is missing more and more people at both ends (or margins) of the income distribution.[4] In the careful language of economists, this leads to undercounting the employment status of such marginal individuals "even though marginal workers and marginal jobs are the most sensitive to changing labor market conditions."[5]

So the official data don't fully measure either the degree to which important swaths of the workforce are marginalized or their vulnerability to economic shocks. It has been argued that the 2016 presidential election results were directly a result of Democrats not fully understanding the disproportionately negative effect of shocks like the North American Free Trade Agreement and immigration on income inequality and working-class voters.

The problem is not that the measures have serious flaws. That is old news to economists. The economy is complex, and