Deloitte. Insights

The Chief Data Officer in Government

A CDO Playbook

A report from the Deloitte Center for Government Insights

About the Deloitte Center for Government Insights

The Deloitte Center for Government Insights shares inspiring stories of government innovation, looking at what's behind the adoption of new technologies and management practices. We produce cutting-edge research that guides public officials without burying them in jargon and minutiae, crystalizing essential insights in an easy-to-absorb format. Through research, forums, and immersive workshops, our goal is to provide public officials, policy professionals, and members of the media with fresh insights that advance an understanding of what is possible in government transformation.

About the Beeck Center for Social Impact + Innovation

The Beeck Center for Social Impact + Innovation at Georgetown University engages global leaders to drive social change at scale. Through our research, education, and convenings, we provide innovative tools that leverage the power of capital, data, technology, and policy to improve lives. We embrace a cross-disciplinary approach to building solutions at scale.

Deloitte Consulting LLP's Technology Consulting practice is dedicated to helping our clients build tomorrow by solving today's complex business problems involving strategy, procurement, design, delivery, and assurance of technology solutions. Our service areas include analytics and information management, delivery, cyber risk services, and technical strategy and architecture, as well as the spectrum of digital strategy, design, and development services offered by Deloitte Digital. Learn more about our Technology Consulting practice on www.deloitte.com.

Contents

Introduction

The government CDO: Turning public data to the public good

Sonal Shah and William D. Eggers

A DECADE AGO, A key focus of government pertaining to data was how to make it more open and easily accessible to the public. Ten years later, with thousands of open government data sets worldwide, the discussion has evolved and become more nuanced. Governments are considering their role in overseeing the types and validity of the data they make available, seeking ways to create greater public value from data, and debating how best to protect privacy and govern data use. The rise of government APIs—of which about 700 exist in the United States alone[1]—and developments such as machine learning, the Internet of Things, smart transportation, and blended data make the role of data management in government even more critical.

As digital tools and technologies continue to rapidly evolve, the role of data in government and the roles of those who oversee it—chief data officers (CDOs), chief information officers (CIOs), and chief technology officers (CTOs)—will require more clar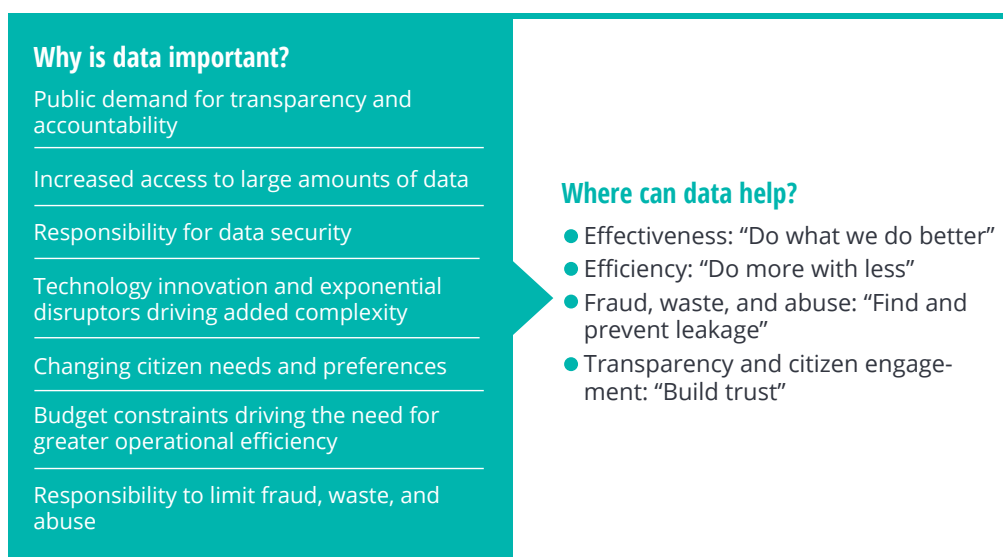ity and definition if governments are to put data to use in governing more effectively. In particular, as data becomes more important in finding solutions to public problems (see figure 1), these government technology leaders will play an increasingly important part in delivering better public outcomes at the city, state, and national levels.

New challenges call for expanded CDO responsibilities

Public sector data is becoming more important for myriad reasons. Public pressure for transparency and accountability is mounting. Many companies, social sector organizations, and others are calling on governments to leverage data to gain greater insights and formulate better policies. And data can offer new ways to curb waste, fraud, and abuse, as well as to operate more efficiently and get more done with less.

FIGURE 1

Why data is important to government

Why is data important?

Public demand for transparency and accountability

Increased access to large amounts of data

Responsibility for data security

Technology innovation and exponential disruptors driving added complexity

Changing citizen needs and preferences

Budget constraints driving the need for greater operational efficiency

Responsibility to limit fraud, waste, and abuse

Where can data help?
- Effectiveness: "Do what we do better"
- Efficiency: "Do more with less"
- Fraud, waste, and abuse: "Find and prevent leakage"
- Transparency and citizen engagement: "Build trust"

Source: Deloitte analysis.

Governments collect vast amounts of data on everything from health care, housing, and education to domestic and national security—both directly and through nonprofits that they support. Governments also produce data, such as census data, labor information, financial market information, weather data, and global positioning system (GPS) data.

This data can be a valuable core asset with the potential to influence program outcomes and public policy. For instance, government data from Medicare and Medicaid can help doctors and hospitals better understand how to reduce the cost of treatment, and help insurance companies provide greater incentives to motivate people to take care of their health. Timely data can also illuminate faster transportation routes in real time, better measure the impact of government programs, and spur new investment opportunities. And in terms of guiding policy, data can help inform decisions on multiple fronts: infrastructure, small business investment, housing, education, health, energy, and many other areas.

Given the immense quantities of data government holds, the governance structures for public data are important and need to be addressed. For example, who gives permission for data use? How will permissions be designed? What is the best way to share data sets between agencies while maintaining privacy? Should there be a standard reporting format across multiple levels of government? When can data collected for one purpose be used for other purposes? What are the legal guidelines around data-sharing?

The increased use of data in policymaking and operations also raises many questions about data provenance, integration with private data sets, individual privacy, and data ethics. Hence, as government CDOs become more prevalent across cities, states, and counties (figure 2), it is important for these CDOs to understand the role's multiple responsibilities and its scope. Yes, CDOs are responsible for safeguarding government data, but they should also help agencies better use their data, and connect citizens with government data to make it more actionable. At the same time, they should provide oversight in managing privacy and protecting citizens' information, especially as digital technologies become more ubiquitous within society.

To do this, CDOs will likely need to coordinate with CIOs, CTOs, and chief information security officers across agencies to build a team, structure, and budget that can support and appropriately manage data assets. The time is ripe for CDOs to take a leadership role in organizing these key decision-makers around using public data for the public good.

The CDO Playbook: Exploring the CDO's toughest challenges

The CDO Playbook, produced by Georgetown University's Beeck Center and Deloitte's Center for Government Insights, explores some of the hardest questions facing CDOs today. The playbook draws on conversations we've had over the past year with CDOs from multiple levels of government as well as in the private, nonprofit, and social sectors. Insights from these leaders shed light on opportunities and potential growth areas for the use of data and the role of CDOs within government.

The playbook is written for government executives as well as for government CDOs. For executives, it provides an overview of the types of functions that CDOs across the country are performing. For CDOs, it offers a guide to understanding the trends affecting public sector data, and provides practical guidance on strategies they can pursue for effective results.

We hope this playbook will help catalyze the further evolution of CDOs within government and provide an accessible guide for executives who are still evaluating the creation of these positions.

FIGURE 2

Government CDOs are becoming more common
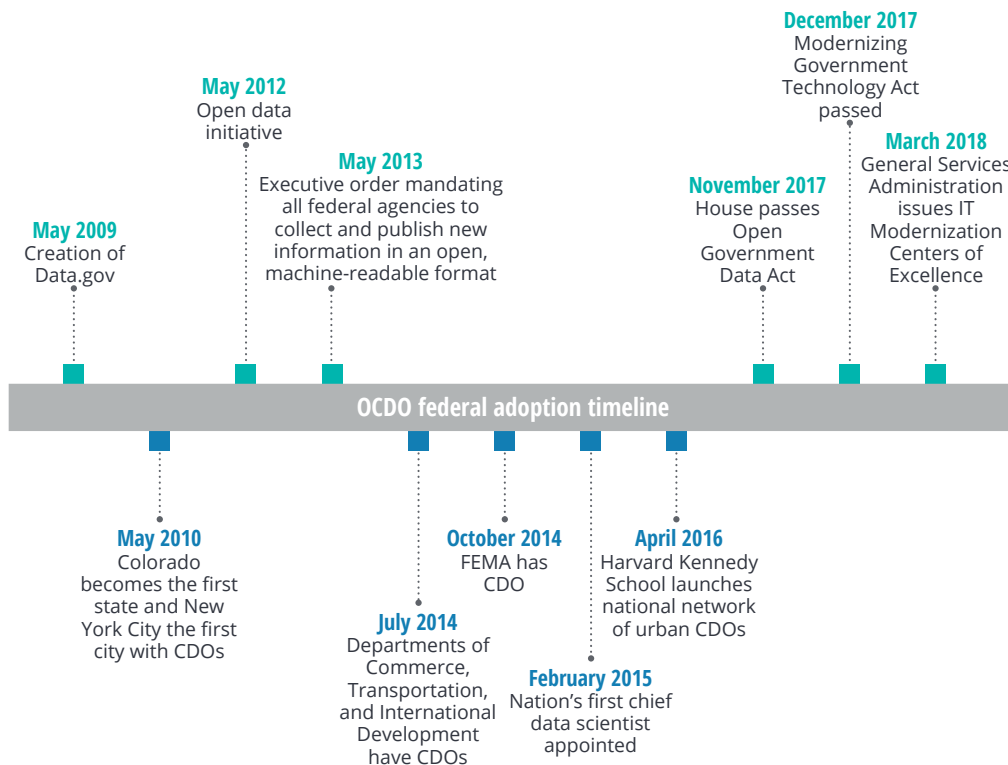
Source: Jack Moore, "Rise of the data chiefs," *NextGov,* March 18, 2015.

Endnotes

1. Wendell Santos, "Most popular government API categories," ProgrammableWeb, January 2, 2018.

About the authors

Sonal Shah is executive director, professor of practice, at the Beeck Center for Social Impact + Innovation. She is based in Washington, DC.

William D. Eggers is the executive director of Deloitte's Center for Government Insights, where he is responsible for the firm's public sector thought leadership. He is based in Arlington, VA.

Deloitte Consulting LLP's Technology Consulting practice is dedicated to helping our clients build tomorrow by solving today's complex business problems. Our service areas include analytics and information management, delivery, cyber risk services, and technical strategy and architecture, as well as the spectrum of digital strategy, design, and development services offered by Deloitte Digital. Learn more about our Technology Consulting practice on www.deloitte.com.

Connecting data to residents through data storytelling

William D. Eggers and Amrita Datar
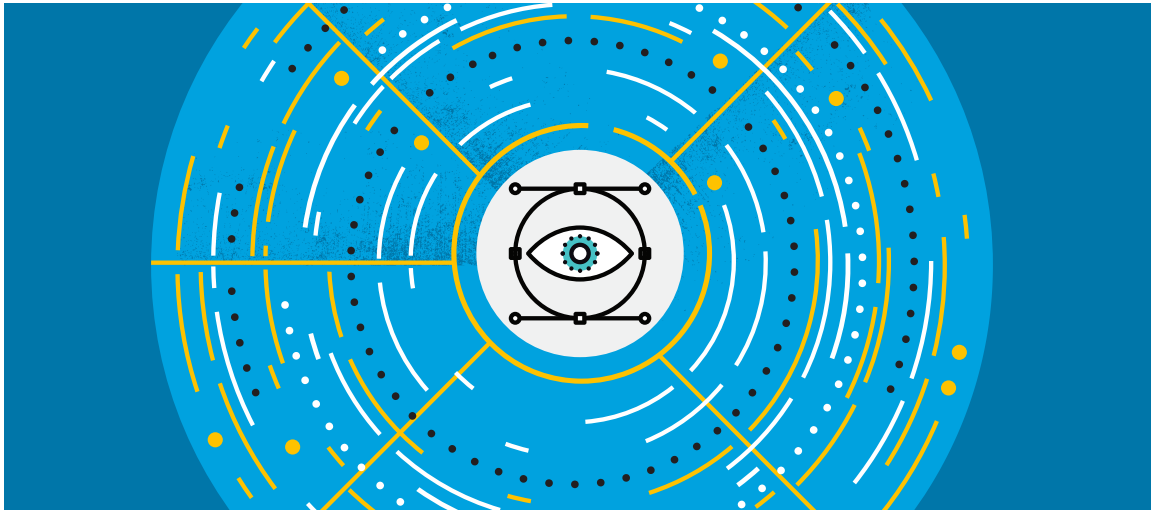
The story is mightier than the spreadsheet

In the past decade, many governments have taken significant strides in the open data movement by making thousands of data sets available to the general public. But simply publishing a data set on an open data portal or website is not enough. For data to have the most impact, it's essential to turn those lines and dots on a chart or numbers in a table into something that everyone can understand and act on.

Data itself is often disconnected from the shared experiences of the American people. An agency might collect and publish data on a variety of areas, but without the context of how it impacts citizens, it might not be as valuable. So how do we connect data to the citizenry's shared everyday lives? Through a language that is deeply tied to our human nature— stories.

Four ways to harness the power of data stories

SHOW, DON'T JUST TELL

As human beings, our brains are wired to process visual data better than other forms of data. In fact, the human brain processes images 60,000 times faster than text.[1] For example, public health data shown on a map might be infinitely more meaningful and accessible to citizens than a heavy table with the same information. Increasingly, governments are tapping into the power of data visualization to connect with citizens.

In Washington, DC, the interactive website District Mobility turns data on the DC area's multi-modal transportation system into map-based visual stories. Which bus routes serve the most riders? How do auto travel speeds vary by day of week and time of day on different routes? How punctual is

the bus service in different areas of town? These are just some of the questions that residents and city planners can find answers to through the site.[2]

Similarly, DataUSA combines publicly accessible US government data from a variety of agencies and brings it to life in more than 2 million visualizations. In addition to allowing users to search, map, compare, and download data sets, the site also shows them what kinds of insights the data can reveal through "DataUSA stories." "People do not understand the world by looking at numbers; they understand it by looking at stories," says Cesar Hidalgo, director of the Massachusetts Institute of Technology Media Lab's Macro Connections group, and one of DataUSA's creators.[3] DataUSA's stories combine maps, charts, and other visualizations with narratives around a range of topics—from the damage done by opioid addiction to real estate in the rust belt to income inequality in America—that might pique citizens' interest. Some states, such as Virginia, have also embedded interactive charts from DataUSA into their economic development portals.

As human beings, our brains are wired to process visual data better than other forms of data. In fact, the human brain processes images 60,000 times faster than text.

reports, inspections, hearings, and scheduled demolitions. The app connected both citizens and city building inspectors to the data and presented it in an easily accessible map-based format along with the context needed to make it actionable.[5]

Data-driven stories can also reveal hidden truths about institutionalized biases. Across America, social justice movements are highlighting citizen disparities. While protests grab the attention of some and repel others, telling compelling stories supported by data can spur meaningful shifts in thinking and outcomes. For example, in the United States, the data shows that black women are 243 percent more likely to die than white women from birth-related complications.[6] This disparity persists for black women who outpace white women in education level, income, and access to health care. The data challenges an industry to address the quality of care provided to this population of Americans.

PICK A HIGH-IMPACT PROBLEM

To connect with citizens across groups, focus data and storytelling efforts around issues that have a far-reaching impact on their lives.

In the aftermath of hurricane Katrina, many neighborhoods across New Orleans were full of blighted and abandoned buildings—more than 40,000 of them. Residents and city staff couldn't easily get information on the status of blighted properties—data that was necessary for communities to come together and make decisions around rebuilding their neighborhoods.[4]

New Orleans city staff worked with a team of Code for America fellows to build an open data-powered web application called Blight Status, which enabled anyone to look up an address and see what reports had been made on the property—blight

SHARE HOW DATA DRIVES DECISION-MAKING

Another way to bring citizens closer to data that matters to them is by telling the story of how that data can shape government decisions that impact their lives. This can be accomplished through a blog, a talk, a case study, or simply in the way public officials communicate successes to their constituents.

Consider the example of Kansas City's KCStat program. KCStat meetings are held each month to track the city's progress toward its goals. Data is used to drive the conversation around a host of issues, from public safety and community health to economic development and housing. Citizens are invited to the meetings, and stats and highlights from meetings are even shared on Twitter (#KCStat) to encourage participation and build awareness.[7]

"As data becomes ingrained systemically in your operation, you can use facts and data to create, tweak, sustain, and perfect programs that will provide a real benefit to people, and it's verifiable by the numbers," said Kansas City mayor Sly James in an interview for Bloomberg's What Works Cities.[8]

The city also publishes a blog called Chartland that tells stories drawn from the city's data. Some focus on themes from KCStat meetings, while others, often written by the city's chief data officer (CDO) and the office of the city manager, explore pertinent city issues such as the risk of lead poisoning in older homes, patterns in 311 data, or how results from a citizen satisfaction survey helped drive an infrastructure repair plan.[9] These blogs are conversational and easy to understand, helping to humanize data that can seem intimidating to many.

MAKE STORYTELLING A TWO-WAY STREET

Hackathons and open data-themed events give citizens a way to engage with data sets in guided settings and learn to tell their own stories with the data. To celebrate the five-year anniversary of the NYC Open Data Law, for instance, New York City's Open Data team organized its first-ever Open Data Week in 2017. The week's activities included 12 events revolving around open data, which attracted over 900 participants. The city's director of open data also convened "Open Data 101," a training session designed specifically to teach nontechnical users how to work with open data sets.[10]

It's important for event organizers to be cognizant that, for data storytelling to bring citizens closer to data, activities should be designed to enable participation for all—not just those who are already skilled with technology and data. For example, when Pittsburgh hosted its own Open Data Day—an all-day drop-in event for citizens to engage in activities around data—the event included a low-tech "Dear Data" project in which participants could hand-draw a postcard to tell a data-based story. Organizers also stipulated that activity facilitators should adopt a "show and play" format—a demo followed by a hands-on activity instead of a static presentation—to encourage open conversation and participation.[11]

Citizens telling their *own* stories with data can shed light on previously unknown challenges and opportunities, giving them a voice to drive change. For example, Ben Wellington, a data enthusiast looking through parking violation data in New York City, discovered millions of dollars' worth of erroneous tickets issued for legally parked cars. Some patrol officers were unfamiliar with a recent parking law change and continued to issue tickets—a problem that the city has since corrected, thanks to Wellington's analysis.[12]

LOOKING AHEAD

The value of data is determined not by the data itself, but by the story it tells and the actions it empowers us to take. But for citizens to truly feel connected to data, they need to see more than just numbers on a page; they need to understand what those numbers really mean for them. To make this connection, CDOs and data teams will need to invest in how they present data and think creatively about new formats and platforms.

Endnotes

2. Rachel Gillet, "Why we're more likely to remember content with images and video," *Fast Company*, September 18, 2014.

3. Government of the District of Columbia, "District mobility: Multimodal transportation in the district," accessed February 12, 2018.

4. Tanvi Misra, "The one-stop digital shop for digestible data on your city," *Citylab*, April 4, 2016.

5. Code for America blog, "New Orleans fights blight with data," accessed February 12, 2018.

6. Ibid.

7. Renee Montagne, "Black mothers keep dying after giving birth. Shalon Irving's story explains why," NPR, December 7, 2017.

8. Sharman Stein, "Kansas City: Winning residents' trust by asking, listening, and acting," Medium, January 25, 2018.

9. Ibid.

10. City of Kansas City, MO, "Chartland," accessed March 29, 2018.

11. Colin Wood, "As public outreach matures, NYC renews Open Data Week for 2018," StateScoop, October 26, 2017.

12. Bob Gradek and Eleanor Tutt, "Pittsburgh's Data Day: Using civic data to spark hands-on community engagement," Living Cities, October 19, 2017.

13. Max Galka, "How an open data blogger proved the NYPD issued parking tickets in error," *Guardian*, July 26, 2016.
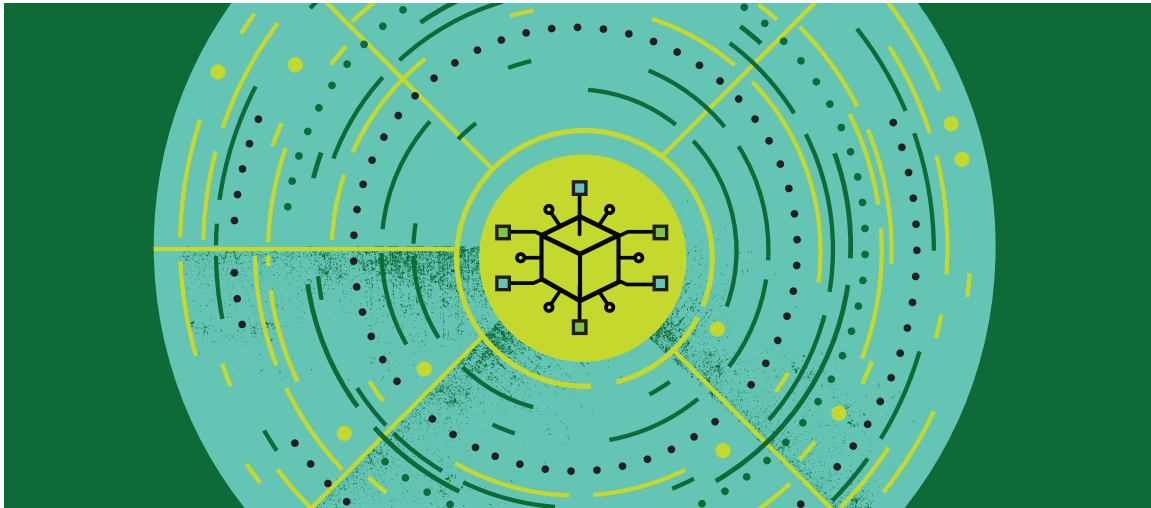
About the authors

William D. Eggers is the executive director of Deloitte's Center for Government Insights, where he is responsible for the firm's public sector thought leadership.

Amrita Datar is an assistant manager with the Deloitte Center for Government Insights.

How CDOs can overcome obstacles to open data-sharing

Adam Neufeld

OPEN DATA HAS been a hot topic in government for the past decade. Various politicians from across the spectrum have extolled the benefits of increasing access to and use of government data, citing everything from enhanced transparency to greater operating efficiency.[1]

While the open data movement seems to have achieved some successes, including the DATA Act[2] and data.gov,[3] we have yet to achieve the full potential of open data. The McKinsey Global Institute, for example, estimates that opening up more data could result in more than $3 trillion in economic benefits.[4]

It is time for the open data community to pivot based on the lessons learned over the past decade, and governmental chief data officers (CDOs) can lead the way.

Much valuable government data remains inaccessible to the public. In some cases, this is because the data includes personally identifiable information. But in other situations, data remains unshared because government has procured a proprietary system that prevents sharing. Moreover, when government does share data, it sometimes does so in spreadsheets or in other formats that can limit its usefulness, rather than in a format such as an application programming interface (API) that would allow for easier use. In fact, some of the potentially most valuable public information, such as financial regulatory filings, is typically not machine-readable.

CDOs looking to achieve greater benefits through open data should devise a plan that addresses both the technical and administrative challenges of data-sharing, including:

- **Mismatched incentives between political leaders and their staff:** Not all data can or should be shared publicly. Agencies are prohibited from sharing personally identifiable data, medical data, and certain other information. There are, however, many gray areas regarding what can or cannot be disclosed. In these instances, the decision on whether and how to standardize or publish a government data set has all the ingredients of a standard principal-agent problem in economics.[5] The principals (here, the public, legislators, and, to some extent, ex-

Intermediate approaches could allow even some sensitive data to be shared.
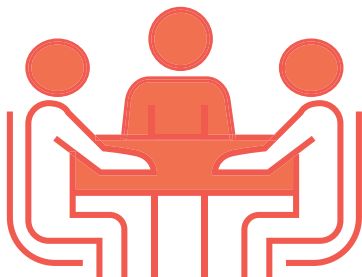
ecutive branch leaders) generally want data to be open because they stand to reap the societal and/or reputational benefits of whatever comes from releasing it. However, the decision of whether to standardize or release data is made by an agent (here, usually some combination of program managers, information technology professionals, and lawyers). The agent tends to gain little direct benefit from releasing the data—but they could face substantial costs in doing so. Not only would they need to do the hard work of standardization, but they would incur the risk of reputational damage, stress, or termination if the data they release turns out to be inaccurate, creates embarrassment for the program, or compromises privacy, national security, or business interests. As a result, even if a political leader wants to share data, there may still be obstacles to doing so.

- **An "all or nothing" approach to data-sharing:** The discussion of open data is often presented in binary terms: Either data is open, meaning that it is publicly available in a standardized format for download on a website, or it is not accessible to outsiders at all. This type of thinking takes intermediate options off the table that could provide much of the benefit of full disclosure, but at less cost and/or lower risk. The experience of federal statistical agencies suggests that intermediate approaches could allow even some sensitive data to be shared on a limited basis.[6] For example, the Center for Medicare and Medicaid Services allows companies to apply for limited, secure access to transaction data to help them develop products

that aim to improve health outcomes or reduce health spending.[7]

- **Lack of technical expertise:** Releasing a data set is generally time-consuming technical work that may require cleaning the data and deciding on privacy protections. Some governments may have limited in-house technological expertise, however, and these technical experts are often needed for other competing priorities. The skills needed to appropriately release data sets that contain sensitive information are even more technical, requiring people with an understanding of advanced cryptographic and technical approaches such as synthetic data[8] and secure multiparty computation.[9] Usually, the subject-matter experts who control whether a given data set will be opened do not have this expertise. This is understandable, as such skills were not historically necessary or even useful, but the skill set gap can prevent governments from sharing data even when all stakeholders agree that it should be shared.

- **Difficulty in prioritizing data sets:** Just as releasing data typically requires a rare combination of subject matter and technical expertise, so can figuring out which data sets to prioritize. How government data might be put to beneficial use requires imagination from people with varied perspectives. Government officials cannot always predict what data sets, especially when used in concert with other data sets, might prove transformative. This is even more true when considering the details of how data should be shared.

CDOs looking to unleash the potential of open data should consider ways that they could address these obstacles. One potential approach is to centralize decision-making authority and technical capabilities rather than having these distributed among the numerous offices and departments that "own" the data. The General Services Administration, for example, created a chief data officer position to act in this capacity. Several other agencies have done the same, and Congress is currently considering legislation to require every agency to do so.[10]

The open data community, for its part, can play an important part in encouraging data-sharing by helping agencies understand what data would be

most useful under what conditions. CDOs sometimes do not have the political strength or the management or technical bandwidth to release all of their agencies' data, even if this were always desirable, so prioritization is key. Regulated entities and beneficiaries should also help the government determine what the next-best alternative is if full openness is not possible. A few agencies, such as the US Department of Health and Human Services with its Demand-Driven Open Data effort, have invited the public to engage in prioritization. To promote greater openness, however, such efforts should be spread across more agencies and involve more levels at those agencies. Understanding the perspectives of those outside government can help officials balance the trade-off between releasing data and controlling the risks and costs.

CDOs' leadership will be important in encouraging government to move swiftly to release all appropriate data that could benefit our society, democracy, and economy. To be most effective, they may need private-sector input and policy guidance that can help them and support them on the open data journey.

Endnotes

1. Josh Hicks, "This bill to track every federal dollar somehow united Cummings and Issa," *Washington Post*, April 11, 2014.

2. US government, "Digital Accountability and Transparency Act of 2014," accessed May 3, 2018.

3. Data.gov, "The home of the US government's open data," accessed May 3, 2018.

4. James Manyika et al., "Open data: Unlocking innovation and performance with liquid information," McKinsey Global Institute, October 2013.

5. Wikipedia, "Principal-agent problem," accessed May 3, 2018.

6. Office of Management and Budget, "Federal Register," December 2, 2014.

7. Research Data Assistance Center, "Innovator Research," accessed May 3, 2018.

8. Wikipedia, "Synthetic data," accessed May 3, 2018.

9. Wikipedia, "Secure multi-party computation," accessed May 3, 2018.

10. US government, "H.R.4174 - Foundations for Evidence-Based Policymaking Act of 2017," accessed May 3, 2018.

About the authors

Adam Neufeld is a senior fellow at the Beeck Center for Social Impact + Innovation. He is based in Washington, DC.

How CDOs can promote machine learning in government

David Schatsky and Rameeta Chauhan
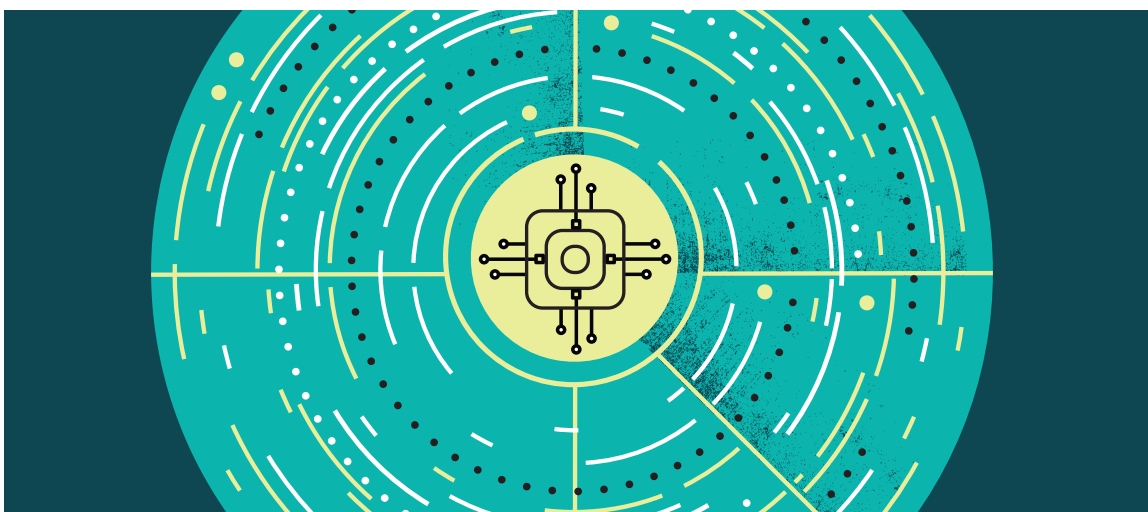
ARTIFICIAL INTELLIGENCE (AI) holds tremendous potential for governments, especially machine learning technology, which can help discover patterns and anomalies and make predictions. There are five vectors of progress that can make it easier, faster, and cheaper to deploy machine learning and bring the technology into the mainstream in the public sector. As the barriers continue to fall, chief data officers (CDOs) have increasing opportunities to begin exploring applications of this transformative technology.

Current obstacles

Machine learning is one of the most powerful and versatile information technologies available today.[1] But most organizations, even in the private sector, have not begun to use its potential. One recent survey of 3,100 executives from small, medium, and large companies across 17 countries found that fewer than 10 percent of companies were investing in machine learning.[2]

A number of factors are restraining the adoption of machine learning in government and the private sector. Qualified practitioners are in short supply.[3] Tools and frameworks for doing machine learning work are still evolving.[4] It can be difficult, time-consuming, and costly to obtain large datasets that some machine learning model-development techniques require.[5]

Then there is the black box problem. Even when machine learning models can generate valuable information, many government executives seem reluctant to deploy them in production. Why? In part, possibly because the inner workings of machine learning models are inscrutable, and some people are uncomfortable with the idea of running their operations or making policy decisions based on logic they don't understand and can't clearly describe.[6]

Other government officials may be constrained by an inability to prove that decisions do not discriminate against protected classes of people.[7] Using AI generally requires understanding all requirements of government, and it requires making the black boxes more transparent.

Progress in these five areas can help overcome barriers to adoption

There are five vectors of progress in machine learning that could help foster greater adoption of machine learning in government (see figure 1). Three of these vectors include automation, data reduction, and training acceleration, which make machine learning easier, cheaper, and/or faster. The other two are model interpretability and local

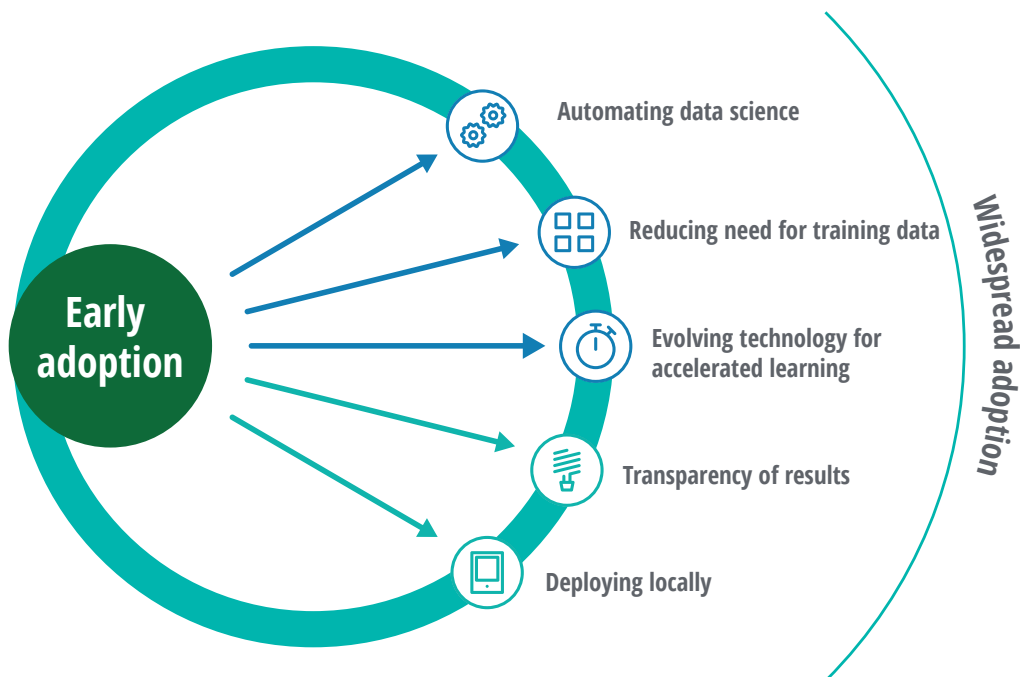machine learning, both of which can open up applications in new areas.

AUTOMATING DATA SCIENCE

Developing machine learning solutions requires skills primarily from the discipline of data science, an often-misunderstood field. Data science can be considered a mix of art and science—and digital grunt work. Almost 80 percent of the work that data scientists spend their time on can be fully or partially automated, giving them time to spend on higher-value issues.[8] This includes data wrangling—preprocessing and normalizing data, filling in missing values, or determining whether to interpret the data in a column as a number or a date; exploratory data analysis—seeking to understand the broad characteristics of the data to help formulate hypotheses about it; feature engineering and selection—selecting the variables in the data that are

FIGURE 1

The five vectors of progress

———— Makes machine learning easier, cheaper, or faster (or a combination of all three)

———— Opens up applications in new areas

Early adoption

- Automating data science
- Reducing need for training data
- Evolving technology for accelerated learning
- Transparency of results
- Deploying locally

Widespread adoption

Source: Deloitte analysis.

most likely correlated with what the model is supposed to predict; and algorithm selection and evaluation—testing potentially thousands of algorithms to assess which ones produce the most accurate results.

Automating these tasks can make data scientists in government more productive and more effective. For instance, while building customer lifetime value models for guests and hosts, data scientists at Airbnb used an automation platform to test multiple algorithms and design approaches, which they would not likely have otherwise had the time to do. This enabled Airbnb to discover changes it could make to its algorithm that increased the algorithm's accuracy by more than 5 percent, resulting in the ability to improve decision-making and interactions with the Airbnb community at very granular levels.[9]

A growing number of tools and techniques for data science automation, some offered by established companies and others by venture-backed startups, can help reduce the time required to execute a machine learning proof of concept from months to days.[10] And automating data science can mean augmenting data scientists' productivity, especially given frequent talent shortages. As the example above illustrates, agencies can use data science automation technologies to expand their machine learning activities.

REDUCING THE NEED FOR TRAINING DATA

Developing machine learning models typically requires millions of data elements. This can be a major barrier, as acquiring and labeling data can be time-consuming and costly. For example, a medical diagnosis project that requires MRI images labeled with a diagnosis requires a lot of images and diagnoses to create predictive algorithms. It can cost more than $30,000 to hire a radiologist to review and label 1,000 images at six images an hour. Additionally, privacy and confidentiality concerns,

A number of potentially promising techniques for reducing the amount of training data required for machine learning are emerging.
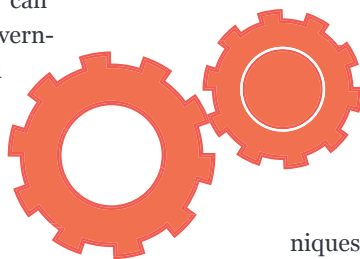
particularly for protected data types, can make working with data more time-consuming or difficult.

A number of potentially promising techniques for reducing the amount of training data required for machine learning are emerging. One involves the use of synthetic data, generated algorithmically to create a synthetic alternative to mimic the characteristics of real data.[11] This technique has shown promising results.

A Deloitte LLP team tested a tool that made it possible to build an accurate machine learning model with only 20 percent of the training data previously required by synthesizing the remaining 80 percent. The model's task was to analyze job titles and job descriptions—which are often highly inconsistent in large organizations, especially those that have grown by acquisition—and then categorize them into a more consistent, standard set of job classifications. To learn how to do this, the model needed to be trained through exposure to a few thousand accurately classified examples. Instead of requiring analysts to laboriously classify ("label") these thousands of examples by hand, the tool made it possible to take a set of labeled data just 20 percent as large and automatically generate a fuller training dataset. And the resulting dataset, composed of 80 percent synthetic data, trained the model just as effectively as a hand-labeled real dataset would have.

Synthetic data can not only make it easier to get training data, but also make it easier for organizations to tap into outside data science talent. A number of organizations have successfully engaged third parties or used crowdsourcing to devise machine learning models, posting their datasets online for outside data scientists to work with.[12] This can be difficult, however, if the datasets are

proprietary. To address this challenge, researchers at MIT created a synthetic dataset that they then shared with an extensive data science community. Data scientists within the community built machine learning models using the synthetic data. In 11 out of 15 tests, the models developed from the synthetic data performed as well as those trained on real data.[13]

Another technique that could reduce the need for extensive training data is transfer learning. With this approach, a machine learning model is pre-trained on one dataset as a shortcut to learning a new dataset in a similar domain such as language translation or image recognition. Some vendors offering machine learning tools claim their use of transfer learning has the potential to cut the number of training examples that customers need to provide by several orders of magnitude.[14]

EVOLVING TECHNOLOGY FOR ACCELERATED LEARNING

Because of the large volumes of data and complex algorithms involved, the computational process of training a machine learning model can take a long time: hours, days, even weeks.[15] Only then can the model be tested and refined. Now, some semiconductor and computer manufacturers—both established companies and startups—are developing specialized processors such as graphics processing units (GPUs), field-programmable gate arrays, and application-specific integrated circuits to slash the time required to train machine learning models by accelerating the calculations and by speeding up the transfer of data within the chip.

These dedicated processors can help organizations significantly speed up machine learning training and execution, which in turn could bring down the associated costs. For instance, a Microsoft research team, using GPUs, completed a system that could recognize conversational speech as capably as humans in just one year. Had the team used only CPUs, according to one of the researchers, the same task would have taken five years.[16] Google has stated that its own AI chip, the Tensor Processing Unit (TPU), when incorporated into a computing system that also includes CPUs and GPUs, provided such a performance boost that it helped the company avoid the cost of building a dozen extra data centers.[17] The possibility of reducing the cost and time involved in machine learning training could have big implications for government agencies, many of which have a limited number of data scientists.

Early adopters of these specialized AI chips include some major technology vendors and research institutions in data science and machine learning, but adoption also seems to be spreading to sectors such as retail, financial services, and telecom. With every major cloud provider—including IBM, Microsoft, Google, and Amazon Web Services—offering GPU cloud computing, accelerated training will likely soon become available to public sector data science teams, making it possible for them to be fast followers. This would increase these teams' productivity and allow them to multiply the number of machine learning applications they undertake.[18]

TRANSPARENCY OF RESULTS

Machine learning models often suffer from the black-box problem: It is impossible to explain with confidence how they make their decisions. This can make them unsuitable or unpalatable for many applications. Physicians and business leaders, for instance, may not accept a medical diagnosis or investment decision without a credible explanation for the decision. In some cases, regulations mandate such explanations.

Techniques are emerging that can help shine light inside the black boxes of certain machine learning models, making them more interpretable and accurate. MIT researchers, for instance, have demonstrated a method of training a neural network that delivers both accurate predictions *and* rationales for those predictions.[19] Some of these techniques are already appearing in commercial data science products.[20]

As it becomes possible to build interpretable machine learning models, government agencies could find attractive opportunities to use machine learning. Some of the potential application areas include child welfare, fraud detection, and disease diagnosis and treatment.[21]

DEPLOYING LOCALLY

The emergence of mobile devices as a machine learning platform is expanding the number of potential applications of the technology and inducing organizations to develop applications in areas such as smart homes and cities, autonomous vehicles, wearable technology, and the industrial Internet of Things.

The adoption of machine learning will grow along with the ability to deploy the technology where it can improve efficiency and outcomes. Advances in both software and hardware are making it increasingly viable to use the technology on mobile devices and smart sensors.[22] On the software side, several technology vendors are creating compact machine learning models that often require relatively little memory but can still handle tasks such as image recognition and language translation on mobile devices.[23] Microsoft Research Lab's compression efforts resulted in models that were 10 to 100 times smaller than earlier models.[24] On the hardware end, various semiconductor vendors have developed or are developing their own power-efficient AI chips to bring machine learning to mobile devices.[25]

Prepare for the mainstreaming of machine learning

Collectively, the five vectors of machine learning progress can help reduce the challenges government agencies may face in investing in machine learning. They can also help agencies already using machine learning to intensify their use of the technology. The advancements can enable new applications across governments and help overcome the constraints of limited resources, including talent, infrastructure, and data to train the models.

CDOs have the opportunity to automate some of the work of often oversubscribed data scientists and help them add even more value. A few key things agencies should consider are:

- Ask vendors and consultants how *they* use data science automation.
- Keep track of emerging techniques such as data synthesis and transfer learning to ease the challenge of acquiring training data.
- Investigate whether the agency's cloud providers offer computing resources that are optimized for machine learning.

Endnotes

1. For an introduction to cognitive technologies, including machine learning, see David Schatsky, Craig Muraskin, and Ragu Gurumurthy, *Demystifying artificial intelligence: What business leaders need to know about cognitive technologies*, Deloitte University Press, November 4, 2014.

2. SAP Center for Business Insight and Oxford Economics, *SAP Digital Transformation Executive Study,* 2017.

3. For a discussion of supply and demand of data science skills, see IBM, "The quant crunch: How the demand for data science skills is disrupting the job market," accessed April 9, 2018.

4. For insights on the early stage of development of machine learning tools, see Catherine Dong, "The evolution of machine learning," TechCrunch, August 8, 2017.

5. For a discussion of the challenge and some ways around it, see Alex Ratner, Stephen Bach, Paroma Varma, and Chris Ré, "Weak supervision: The new programming paradigm for machine learning," Stanford DAWN, July 16, 2018.

6. See, for example, Cliff Kuang, "Can A.I. be taught to explain itself?," *New York Times Magazine*, November 21, 2017.

7. For a high-level discussion of this challenge, see Manny Moss, "The business case for machine learning interpretability," Cloudera Fast Forward Labs, August 2, 2017. For a discussion of how the European Union's new General Data Protection Regulation effectively creates a "right to explanation" that will increase demand for interpretable algorithms and models, see Bryce Goodman and Seth Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation,'" *AI Magazine* 38, no. 3 (2016).

8. For a discussion of machine learning automation, see Jakub Hava, "Sparkling Water 2.2.10 is now available!," H2O.ai blog, March 22, 2018.

9. Hamel Husain and Nick Handel, "Automated machine learning—a paradigm shift that accelerates data scientist productivity @ Airbnb," Medium, May 10, 2017.

10. For a partial list, see Gregory Piatetsky, "Automated data science and data mining," KDnuggets, March 2016. As of October 2017, one startup in this area, DataRobot, had raised $125 million from venture investors. Google has introduced machine learning modeling techniques called AutoML. See Quoc Le and Barret Zoph, "Using machine learning to explore neural network architecture," Google Research Blog, May 17, 2017.

11. Sergey Nikolenko, "New resources for deep learning with the Neuromation platform," Medium, October 9, 2017.

12. Lisa Morgan, "9 reasons to crowdsource data science projects," *InformationWeek*, February 2, 2016.

13. Stefanie Koperniak, "Artificial data give the same results as real data—without compromising privacy," MIT News, March 3, 2017.

14. See, for instance, Indico, "Extract insight from data with Indico's API," accessed April 9, 2018.

15. Khari Johnson, "Google unveils second-generation TPU chips to accelerate machine learning," VentureBeat, May 17, 2017.

16. Cade Metz, "How AI is shaking up the chip market," *Wired*, October 28, 2016.

17. Cade Metz, "Building an AI chip saved Google from building a dozen new data centers," *Wired*, April 5, 2017.

18. See, for instance, IBM, "IBM Cloud first to offer latest NVIDIA GRID with Tesla M60 GPU, speeding up virtual desktop applications," press release, May 19, 2016; Tiffany Trader, "Microsoft Azure will debut Pascal GPU instances this year," HPCwire, May 8, 2017.

19. Larry Hardesty, "Making computers explain themselves," MIT News, October 27, 2016.

20. For examples, see the data science automation platform H2O Driverless AI (Patrick Hall et al., "Machine learning interpretability with H2O Driverless AI," H20.ai, April 2018); DataScience.com's new Python library, Skater (DataScience.com, "DataScience.com releases Python package for interpreting the decision-making processes of predictive models," press release, May 23, 2017); and DataRobot's ML-powered predictive modeling for insurance pricing (DataRobot, "New DataRobot release extends enterprise readiness capabilities and automates machine learning in insurance industry pricing models," press release, July 24, 2017).

21. Fast Forward Labs, "Interpretability," July 2017.

22. David Schatsky, *Machine learning is going mobile*, Deloitte University Press, April 1, 2016.

23. John Mannes, "Google's TensorFlow Lite brings machine learning to Android devices," TechCrunch, May 18, 2017; Liam Tung, "Microsoft wants to bring AI to Raspberry Pi and other tiny devices," ZDNet, June 30, 2017; John Mannes, "Facebook open sources Caffe2, its flexible deep learning framework of choice," TechCrunch, April 19, 2017; James Vincent, "Apple announces new machine learning API to make mobile AI faster," Verge, June 5, 2017.

24. Tung, "Microsoft wants to bring AI to Raspberry Pi and other tiny devices."

25. Tom Simonite, "The rise of AI is forcing Google and Microsoft to become chipmakers," *Wired*, July 25, 2017; Mark Gurman, "Apple is working on a dedicated chip to power AI on devices," Bloomberg, May 27, 2017.

About the authors

David Schatsky, a director with Deloitte Services LP, analyzes emerging technology and business trends for Deloitte's leaders and clients. He is based in New York City.

Rameeta Chauhan, of Deloitte Services India Pvt. Ltd., tracks and analyzes emerging technology and business trends, with a primary focus on cognitive technologies, for Deloitte's leaders and clients. She is based in Mumbai, India.

How CDOs can manage algorithmic risks

Nancy Albinson, Dilip Krishna, Yang Chu, William D. Eggers, and Adira Levine

THE RISE OF advanced data analytics and cognitive technologies has led to an explosion in the use of complex algorithms across a wide range of industries and business functions, as well as in government. Whether deployed to predict potential crime hotspots or detect fraud and abuse in entitlement programs, these continually evolving sets of rules for automated or semi-automated decision-making can give government agencies new ways to achieve goals, accelerate performance, and increase effectiveness.

However, algorithm-based tools—such as machine learning applications of artificial intelligence (AI)—also carry a potential downside. Even as many decisions enabled by algorithms have an increasingly profound impact, growing complexity can turn those algorithms into inscrutable black boxes. Although often enshrouded in an aura of objectivity and infallibility, algorithms can be vulnerable to a wide variety of risks, including accidental or intentional biases, errors, and fraud.

Chief data officers (CDOs), as the leaders of their organization's data function, have an important role to play in helping governments harness this new capability while keeping the accompanying risks at bay.

Understanding the risks

Governments increasingly rely on data-driven insights powered by algorithms. Federal, state, and local governments are harnessing AI to solve challenges and expedite processes—ranging from answering citizenship questions through virtual assistants at the Department of Homeland Security to, in other instances, evaluating battlefield wounds with machine learning-based monitors.[1] In the

coming years, machine learning algorithms will also likely power countless new Internet of Things (IoT) applications in smart cities and smart military bases.

While such change can be considered transformative and impressive, instances of algorithms going wrong have also increased, typically stemming from human biases, technical flaws, usage errors, or security vulnerabilities. For instance:

- Social media algorithms have come under scrutiny for the way they may influence public opinion.[2]
- During the 2016 Brexit referendum, algorithms received blame for the flash-crash of the British pound by six percent in two minutes.[3]
- Investigations have found that an algorithm used by criminal justice systems across the United States to predict recidivism rates is biased against certain racial groups.[4]

Typically, machine learning algorithms are first programmed and then trained using existing sample data. Once training concludes, algorithms can analyze new data, providing outputs based on what they learned during training and potentially any other data they've analyzed since. When it comes to algorithmic risks, three stages of that process can be especially vulnerable:

- **Data input:** Problems can include biases in the data used for training the algorithm (see sidebar "The problem of algorithmic bias"). Other problems can arise from incomplete, outdated, or irrelevant input data; insufficiently large and diverse sample sizes; inappropriate data collection techniques; or a mismatch between training data and actual input.
- **Algorithm design:** Algorithms can incorporate biased logic, flawed assumptions or judgments, structural inequities, inappropriate modeling techniques, or coding errors.
- **Output decisions:** Users can interpret algorithmic output incorrectly, apply it inappropriately, or disregard its underlying assumptions. The immediate fallout from algorithmic risks can include inappropriate or even illegal decisions.

THE PROBLEM OF ALGORITHMIC BIAS

Governments have used algorithms to make various decisions in criminal justice, human services, health care, and other fields. In theory, this should lead to unbiased and fair decisions. However, algorithms have at times been found to contain inherent biases, often as a result of the data used to train the algorithmic model. For government agencies, the problem of biased input data constitutes one of the biggest risks they face when using machine learning.

While algorithmic bias can involve a number of factors other than race, allegations of racial bias have raised concerns about certain government applications of AI, particularly in the realm of criminal justice. Some court systems across the country have begun using algorithms to perform criminal risk assessments, an evaluation of the future criminal risk potential of criminal defendants. In nine US states, judges use the risk scores produced in these assessments as a factor in criminal sentencing. However, criminal risk scores have raised concerns over potential algorithmic bias and led to calls for greater examination.[5]

In 2016, ProPublica conducted a statistical analysis of algorithm-based criminal risk assessments in Broward County, Florida. Controlling for defendant criminal history, gender, and age, the researchers concluded that black defendants were 77 percent more likely than others to be labeled at higher risk of committing a violent crime in the future.[6] While the company that developed the tool denied the presence of bias, few of the criminal risk assessment tools used across the United States have undergone extensive, independent study and review.[7]

And due to the speed at which algorithms operate, the consequences can quickly get out of hand. The potential long-term implications for government agencies include reputational, operational, technological, policy, and legal risks.

Taking the reins

To effectively manage algorithmic risks, traditional risk management frameworks should be modernized. Government CDOs should develop and adopt new approaches that are built on strong foundations of enterprise risk management and aligned with leading practices and regulatory requirements. Figure 1 depicts such an approach and its specific elements.

STRATEGY, POLICY, AND GOVERNANCE

Create an algorithmic risk management strategy and governance structure to manage technical and cultural risks. This should include principles, ethics, policies, and standards; roles and responsibilities; control processes and procedures; and appropriate personnel selection and training. Providing transparency and processes to handle inquiries can also help organizations use algorithms responsibly.

From a policy perspective, the idea that automated decisions should be "explainable" to those affected has recently gained prominence, although this is still a technically challenging proposition. In May 2018, the European Union began enforcing laws that require companies to be able to explain how their algorithms operate and reach decisions.[8] Meanwhile, in December 2017, the New York City

FIGURE 1

A framework for algorithmic risk management

Strategy, policy, and governance		Design, development, deployment, and use	Monitoring and testing
Goals and strategy	Principles, ethics, policies, standards, and guidelines	Algorithm design process	Algorithm testing
Accountability and responsibilities	Life cycle and change management	Data assessment	Output logging and analysis
Regulatory compliance	Hiring and training of personnel	Assumptions and limitations	Sensitivity analysis
Disclosure to user and stakeholder	Inquiry and complaint procedures	Embedding security and operations controls	Ongoing monitoring
Inventory and risk classifications		Deployment process	Continuous improvement
		Algorithm use	Independent validation
Enterprise risk management			

Source: Deloitte analysis.

Council passed a law establishing an Automated Decision Systems Task Force to study the city's use of algorithmic systems and provide recommendations. The body aims to provide guidance on increasing the transparency of algorithms affecting citizens and addressing suspected algorithmic bias.[9]

DESIGN, DEVELOPMENT, DEPLOYMENT, AND USE

Develop processes and approaches aligned with the organization's algorithmic risk management governance structure to address potential issues in the algorithmic life cycle from data selection, to algorithm design, to integration, to actual live use in production.

This stage offers opportunities to build algorithms in a way that satisfies the growing emphasis on "explainability" mentioned earlier. Researchers have developed a number of techniques to construct algorithmic models in ways in which they can better explain themselves. One method involves creating generative adversarial networks (GANs), which set up a competing relationship between two algorithms within a machine learning model. In such models, one algorithm develops new data and the other assesses it, helping to determine whether the former operates as it should.[10]

Another technique incorporates more direct relationships between certain variables into the algorithmic model to help avoid the emergence of a black box problem. Adding a monotonic layer to a model—in which changing one variable produces a predictable, quantifiable change in another—can increase clarity into the inner workings of complex algorithms.[11]

MONITORING AND TESTING

Establish processes for assessing and overseeing algorithm data inputs, workings, and outputs, leveraging state-of-the-art tools as they become available. Seek objective reviews of algorithms by internal and external parties.

Researchers have developed a number of techniques to construct algorithmic models in ways in which they can better explain themselves.
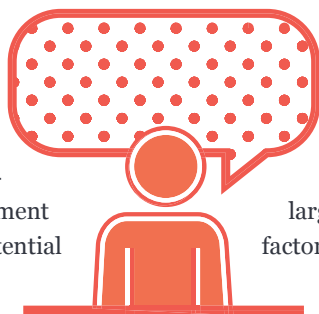
Evaluators can not only assess model outcomes and impacts on a large scale, but also probe how specific factors affect a model's individual outputs. For instance, researchers can examine specific areas of a model, methodically and automatically testing different combinations of inputs—such as by inserting or removing different parts of a phrase in turn—to help identify how various factors in the model affect outputs.[12]

Are you ready to manage algorithmic risks?

A good starting point for implementing an algorithmic risk management framework is to ask important questions about your agency's preparedness to manage algorithmic risks. For example:

- Where are algorithms deployed in your government organization or body, and how are they used?
- What is the potential impact should those algorithms function improperly?
- How well does senior management within your organization understand the need to manage algorithmic risks?
- What is the governance structure for overseeing the risks emanating from algorithms?

Adopting effective algorithmic risk management practices is not a journey that government agencies need to take alone. The growing awareness of algorithmic risks among researchers, consumer advocacy groups, lawmakers, regulators, and other stakeholders should contribute to a growing body

. . . .

THE ALLEGHENY COUNTY APPROACH

Some governments have begun building transparency considerations into their use of algorithms and machine learning. Allegheny County, Pennsylvania provides one such example. In August 2016, the county implemented an algorithm-based tool—the Allegheny Family Screening Tool—to assess risks to children in suspected abuse or endangerment cases.[13] The tool conducts a statistical analysis of more than 100 variables in order to assign a risk score of 1 to 20 to each incoming call reporting suspected child mistreatment.[14] Call screeners at the Office of Children, Youth, and Families consult the algorithm's risk assessment to help determine which cases to investigate. Studies suggest that the tool has enabled a double-digit reduction in the percentage of low-risk cases proposed for review as well as a smaller increase in the percentage of high-risk calls marked for investigation.[15]

Like other risk assessment tools, the Allegheny Family Screening Tool has received criticism for potential inaccuracies or bias stemming from its underlying data and proxies. These concerns underscore the importance of the continued evolution of these tools. Yet the Allegheny County case also exemplifies potential practices to increase transparency. Developed by academics in the fields of social welfare and data analytics, the tool is county-owned and was implemented following an independent ethics review.[16] County administrators discuss the tool in public sessions, and call screeners use it only to decide which calls to investigate rather than as a basis for more drastic measures. The county's steps demonstrate one way that government agencies can help increase accountability around their use of algorithms.

. . . .

of knowledge about algorithmic risks and, over time, risk management standards. In the meantime, it's important for CDOs to evaluate their use of algorithms in high-risk and high-impact situations and implement leading practices to manage those risks intelligently so that their organizations can harness algorithms to enhance public value.

The rapid proliferation of powerful algorithms in many facets of government operations is in full swing and will likely continue unabated for years to come. The use of intelligent algorithms offers a wide range of potential benefits to governments, including improved decision-making, strategic planning, operational efficiency, and even risk management. But in order to realize these benefits, organizations will likely need to recognize and manage the inherent risks associated with the design, implementation, and use of algorithms—risks that could increase unless governments invest thoughtfully in algorithmic risk management capabilities.

Endnotes

1. William D. Eggers, David Schatsky, and Peter Viechnicki, *AI-augmented government: Using cognitive technologies to redesign public sector work*, Deloitte University Press, April 26, 2017.

2. Dilip Krishna, Nancy Albinson, and Yang Chu, "Managing algorithmic risks," *CIO Journal, Wall Street Journal*, October 25, 2017.

3. Jamie Condliffe, "Algorithms probably caused a flash crash of the British pound," *MIT Technology Review*, October 7, 2016.

4. Issie Lapowsky, "Crime-predicting algorithms may not fare much better than untrained humans," *Wired*, January 17, 2018.

5. Julia Angwin et al., "Machine bias," ProPublica, May 23, 2016.

6. Ibid.

7. Ibid.

8. Bahar Gholipour, "We need to open the AI black box before it's too late," Futurism, January 18, 2018.

9. Julia Powles, "New York City's bold, flawed attempt to make algorithms accountable," *New Yorker*, December 20, 2017.

10. Deep Learning for Java, "GAN: A beginner's guide to generative adversarial networks," accessed May 3, 2018.

11. Paul Voosen, "How AI detectives are cracking open the black box of deep learning," *Science*, July 6, 2017.

12. The Local Interpretable Model-Agnostic Explanations (LIME) and other techniques modifying and building upon this help look at local instances of a model prediction. (Patrick Hall et al., "Machine learning interpretability with H2O driverless AI," H2O.ai, April 2018; Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Why should I trust you? Explaining the predictions of any classifier," arXiv.org, February 16, 2016; Voosen, "How AI detectives are cracking open the black box of deep learning.")

13. Dan Hurley, "Can an algorithm tell when kids are in danger?," *New York Times Magazine*, January 2, 2018.

14. Virginia Eubanks, "A child abuse prediction model fails poor families," *Wired*, January 15, 2018.

15. Hurley, "Can an algorithm tell when kids are in danger?"

16. Ibid.

About the authors

Nancy Albinson is a managing director with Deloitte & Touche LLP and leader of Deloitte Risk & Financial Advisory's innovation program. She is based in Parsippany, NJ.

Dilip Krishna is the chief technology officer and a managing director with the Regulatory & Operational Risk practice at Deloitte & Touche LLP. He is based in New York City.

Yang Chu is a senior manager at Deloitte & Touche LLP. She is based in San Francisco, CA.

William D. Eggers is the executive director of Deloitte's Center for Government Insights, where he is responsible for the firm's public sector thought leadership. He is based in Arlington, VA.

Adira Levine is a strategy consultant at Deloitte Consulting LLP, where her work is primarily aligned to the public sector. She is based in Arlington, VA.

Implementing the DATA Act for greater transparency and accessibility

Dave Mader, Tasha Austin, and Christina Canavan

With data an often-underutilized asset in the public sector, enhancing availability and transparency can make a big difference in enabling agencies to use data analytics to their advantage—and the public's.

THOUGHTFUL USE OF data-driven insights can help agencies monitor performance, evaluate results, and make evidence-based decisions. Having access to key facts can drive impressive improvements: When the United States Postal Service compiled and standardized a number of its data sets, the office of the USPS Inspector General's data-modeling team was able to use them to identify about $100 million in savings opportunities, as well as recover more than $20 million in funds lost to possible fraud.[1]

For government chief data officers (CDOs), one of the key drivers for data transparency is the federal government's effort to implement wide-scale data interoperability through the Data Accountability and Transparency Act of 2014 (DATA Act), which seeks to create an open data set for *all* federal spending. If successful, the DATA Act could dramatically increase internal efficiency and external transparency.[2] However, our interviews with more than 20 DATA Act stakeholders revealed some potential challenges to its implementation that could be important to address.

The DATA Act's intent

Before addressing these implementation challenges, it may help to know how the DATA Act sets out to make information on federal expenditures more easily accessible and transparent.

Implementation of the DATA Act is still in its early stages; the first open-spending data set went live in May 2017.[3] If the act is successfully implemented, by 2022, spending data will flow automatically from agency originators to interested government officials and private citizens through publicly available websites. This could save time and increase efficiency across the federal government in several ways, possibly including the following:

Spending reports would populate automatically. Agency leaders wouldn't need to request distinct spending reports from different units of their agencies—the information would compile automatically. For example, a user could see the Department of Homeland Security's spending at a summary level or review spending at the component level.

Congress could make appropriations more transparent. When crafting legislation, Congress could evaluate the impact of spending bills with greater ease. Shifting a few sliders on a dashboard could show the impact of proposed changes to each agency's budget. Negotiations could be conducted using easy-to-digest pie charts reflecting each proposal's impact.

Auditors would need to do less detective work. Auditors would have direct access to data describing spending at a granular level. Rather than often digging through disparate records and unconnected systems, auditors could see an integrated money flow. Using data analytics, auditors could gauge the cost-effectiveness of spending decisions or compare similar endeavors in different agencies or regions. These efforts could help root out fraud.

Citizens could see where the money goes. With greater spending transparency, citizens could have real-time clarity into how government decisions might influence local grant recipients, nonprofits, and infrastructure. It could be as easy for a citizen to see the path of every penny as it would for an agency head.

OMB's data schema: The foundation for change

The DATA Act has the potential to transform various federal management practices. While much work remains to be done, the technology to support the DATA Act has already been developed, giving the act a strong foundation.[4]

The DATA Act mandates that the White House Office of Management and Budget (OMB) maintain a unified data format, or "schema," to organize all federal spending reports. This schema, known as DAIMS (DATA Act Information Model Schema), represents an agreement on how OMB and the Department of the Treasury want to categorize federal spending.[5] It's a common taxonomy that all agencies can use to organize information, and it could shape how the federal government approaches budgeting for years to come. To allow other agencies to connect to DAIMS, OMB has built open-source software—the "Data Broker"—to help agencies report their data.

While the DATA Act deals with federal government data, it can indirectly affect how state and local governments manage their data as well. Data officers from state and local governments will likely need to be familiar with DAIMS and the Data Broker if they hope to collect grants from the federal government. And when contractors adopt federal protocols, they'll likely prefer to report to states in a similar format.

Implementation challenges and approaches

As federal CDOs transform their organizations to meet the DATA Act's new transparency standards, they could face a number of challenges, both cultural and technical.

If users see the DATA Act as a reporting requirement rather than as a tool, they are unlikely to unlock its full potential. Bare minimum data sets, lacking in detail, might satisfy reporting requirements, but they would fail to support effective data analytics. Likewise, users unfamiliar with the DAIMS system may never bother to become adept with it.

The current DAIMS schema fails to account for the full federal budgeting life cycle.

Technical challenges also threaten DATA Act implementation. Legacy reporting systems may not be compatible with DAIMS. The federal government currently identifies grant recipients and contractors using DUNS, the Data Universal Numbering System, a proprietary system of identification numbers with numerous licensing restrictions. A transparent federal data set won't be able to incorporate new data sets from state and local partners unless those partners also spend scarce resources on the DUNS system to achieve compatibility. Lastly, the DAIMS schema, while a monumental achievement, will continue to need improvement. The current DAIMS schema fails to account for the full federal budgeting life cycle. Therefore, the ability to use the data to organize operations is incomplete at best.[6]

With care and commitment, however, these problems can be surmountable. Two steps CDOs can take are:

Convince managers to see the DATA Act as a tool, not a chore. To truly fulfill the DATA Act's promise, workplaces should approach it as a managerial tool, not merely a reporting requirement. If managers use the DAIMS system to run their own organizations, the data they provide would be granular and more accurate. That said, one of the best ways to convince managers to adopt DAIMS for daily use will likely be through active congressional buy-in. If congressional budgeters and appropriators begin relying on DAIMS-powered dashboards to allocate funds, agency managers could naturally gravitate to the same data for budget submissions—and, eventually, for other management activities.
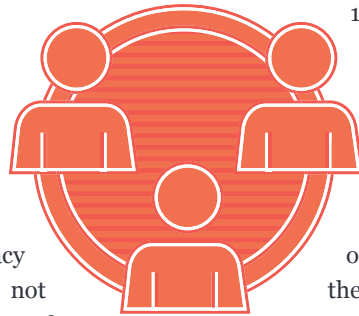
Educate users and managers to show them the benefits. Education can encourage agencies to incorporate DAIMS data into their own operations. One of the test cases for Data Broker, the Small Business Association (SBA), worked with technology specialists on the federal government's 18F team to find uses for the new data system. In the process, they found mislabeled data, made several data quality improvements, and even discovered discretionary funds that they had thought were already committed.[7] Agencies like the SBA, which experienced significant improvements, could evangelize the benefits of clean, transparent data for decision-making to the larger public sector community. Further, more can be done to invest in the upskilling of managers. This could help managers to develop a vision for how data can be used and begin to provide the resources needed to get there.

Improving execution

For all its laudable intent, the DATA Act may fail to deliver its full potential unless it is effectively executed. Some steps for the federal government to consider include:

Establish a permanent governance structure. Currently, OMB and Treasury are responsible for managing data standards for spending data. While this fulfills the basic mandates of the DATA Act, experts acknowledge that, with their current resources, these two agencies can't do the work indefinitely.[8] To ensure DAIMS's flexibility and stability, a permanent management structure should oversee it for the long term.

Extract information directly from source systems. Currently, when a government agency awards a contract, it reports the contract data using several old reporting systems, many of which have well-documented accuracy problems.[9] Currently, DAIMS extracts financial information from these inconsistent sources. The first major revision to DAIMS should require agencies to extract contract information directly from their source award systems. Going straight to the source for both financial and award data should lead to more effi-

cient processing, boost data quality, and could save agencies time and effort.

Adopt a numbering system that anyone can use. Everyone, from local governments to American businesses, should be encouraged to integrate their own budgeting data with the federal government's. Instead of using a proprietary numbering system that excludes participants, the government could consider adopting an open-source or freely available numbering system.

Expand the DAIMS to reflect the full budget life cycle. The federal budget follows a life cycle, from the president's proposed budget to congressional appropriations to payments. To properly track the flow of funds through this life cycle, the spending data in DAIMS should reflect the budget as something that evolves over time from the beginning, with the receipt of tax revenues to final payments to grantees and contractors.

CDOs will likely recognize both the potential benefits of enhancing an organization's ability to leverage data, and the challenges of changing the way public organizations manage data. CDOs would have to thoughtfully manage through the barriers to realize the potential benefits of readily available, transparent data. Leaders would be wise to prepare their own organizations for change even as the DATA Act takes hold at the federal level.

Endnotes

1. Jessica Yabsley, "What the DATA Act means for anti-fraud analytics," Data Coalition, January 26, 2017.

2. David Mader et al. "DATA Act 2022: Changing technology, changing culture," Deloitte and Data Foundation, May 2017.

3. Office of Management and Budget, *Report to Congress: DATA Act pilot program*, August 10, 2017.

4. Mader et al., "DATA Act 2022."

5. Ibid.

6. Ibid.

7. Ibid.

8. Ibid.

9. United States Government Accountability Office, "Data transparency: Oversight needed to address underreporting and inconsistencies on federal award website," June 2014.

About the authors

Dave Mader is the chief strategy officer for the civilian sector within Deloitte Consulting LLP's Federal practice. He is based in Arlington, VA.

Tasha Austin, a senior manager in Deloitte & Touche LLP's Federal practice, leads the DATA Act offering for Deloitte's Risk and Financial Advisory practice. She is based in Arlington, VA.

Christina Canavan, a managing director in Deloitte & Touche LLP's Federal practice, leads the Advisory Analytics practice for Financial Risk Transactions and Restructuring. She is based in Arlington, VA.

CDOs, health data, and the Open Science movement

Juergen Klenk and Melissa Majerol

Open Science: In need of champions

The health care sector is teeming with data. Electronic health records, technologies such as smart watches and mobile apps, and major advances in scientific research—especially in the areas of imaging and genomic sequencing—have given us volumes of medical and biological data over the last decade. One might assume that such a data-rich landscape inherently accelerates scientific discoveries. However, reams of data alone cannot generate new insights, especially when they exist in silos, as is often the case today.

Open Science—the notion that scientific research, including data and research methodologies, should be open and accessible—can offer a solution. Without powerful champions, however, such openness may remain the exception rather than the rule. Practicing Open Science inherently requires cross-sector collaboration as well as buy-in from the public. This is where government chief data officers (CDOs) could play a key role.

Now is the time for Open Science

The early stages of the Open Science movement can be traced back to the 17th century, when the idea arose that knowledge must flow freely across the scientific community to enable and accelerate scientific breakthroughs that can benefit all of society.[1] Four centuries later, Open Science remains an idea that has yet to be fully realized. However, collaborative tools and digital technologies are making the endeavor more achievable than ever before. Rather than simply sharing knowledge in scientific journals, we now have the ability to share electronic health records, patient-generated data, insurance claims data—even genomic data—in standardized, interoperable formats through

web-based tools and the cloud. Moreover, with advanced analytics and cognitive technologies, we can process large volumes of data to identify complex patterns that can lead to new discoveries in ways that were almost unimaginable until recently. Using these data and tools is essential to achieving Open Science's so-called FAIR principles—that data should be findable, accessible, interoperable, and reusable[2] (see the sidebar, "What is FAIR?").

WHAT IS FAIR?[3]

The FAIR principles are a set of guiding principles for scientific data management and stewardship to support innovation and discovery. Distinct from peer initiatives that focus on the human scholar, the FAIR principles put specific emphasis on enhancing the ability of machines to automatically find and use data—in other words, making data "machine-actionable"—in addition to supporting its reuse by individuals. Widely recognized and supported in the scientific community, the principles posit that data should be:

- **Findable.** Data must have unique identifiers that effectively label it within searchable resources.

- **Accessible.** Data must be easily retrievable via open systems that have effective and secure authentication and authorization procedures.

- **Interoperable.** Data should "use and speak the same language" by using standardized vocabularies.

- **Reusable.** Data must be adequately described to a new user, include clear information about data usage licenses, and have a traceable "owner's manual" or provenance.

Consider cancer research. Dr. Jay Bradner, a doctor at a small Harvard-sponsored cancer lab, created a molecule called JQ1—a prototype for a drug to target a rare type of cancer. Rather than keeping the prototype a secret until it was turned into an active pharmaceutical substance and patented, the lab made the drug's chemical identity available on its website for "open source drug discovery." The concept of open source drug discovery borrows two principles from open source computing—collaboration and open access—and applied them to pharmaceutical innovation. Scientists from around the world were able to learn about the drug's chemical identity so that they could experiment with it on various cancer cells. These scientists, in turn, have created new molecules to treat cancer that are being tested in clinical trials.[4] Collaborations like these allow hundreds of minds to study the individual pieces of a complex problem, multiplying the usual pace of discovery.

Government CDOs can help accelerate Open Science

Federal and state governments—and their CDOs—have two unique levers that they can apply to encourage greater openness and collaboration: They hold enormous quantities of health data, and they have the ability to influence policy and practice.

US government health data derives from public programs like Medicare and Medicaid, which collectively cover one in three people in the United States;[5] government-sponsored disease registries; the Million Veteran Program (MVP), one of the world's largest medical databases, which has collected blood samples and health information from a million veteran volunteers; and the National Institutes of Health's (NIH's) recent All of Us initiative, a historic effort to gather data from 1 million or more US residents to accelerate research and improve health.[6] In addition, federal agencies such as the Department of Health and Human Services (HHS), as well as a handful of states, cities, and counties around the country, have begun hiring CDOs to help determine how data is collected, organized, accessed, and analyzed. According to Project Open Data, an online public repository created by President Barack Obama's Open Data Policy and Executive Order,[7] the CDO's role is "part data strategist and adviser, part steward for improving data

quality, part evangelist for data sharing, part technologist, and part developer of new data products."[8]

CDOs looking to advance Open Science should consider ways to meaningfully share more government health data and to encourage nongovernment stakeholders, including academic researchers, health providers, and ordinary citizens, to participate in Open Science data platforms and share their own data. To do so, they will need to address the various technological, policy, and cultural challenges.

Overcoming the barriers: Technology, policy, and culture

TECHNOLOGY: MOVING GOVERNMENT HEALTH DATA TO THE CLOUD

Open Science requires a technological infrastructure that allows data to be securely shared, stored, and analyzed. In an effort to develop this infrastructure, the NIH has begun piloting a "Data Commons," a virtual space where scientists can store, access, and share biomedical data and tools. Here, researchers can utilize "digital objects of biomedical research" to solve difficult problems together and apply cognitive computing capabilities in a single cloud-based environment.[9] This platform embraces the FAIR principles, including the need to safeguard the data it contains with secure authentication and authorization procedures. The pilot is due to be completed in 2020,[10] after which lessons learned are expected to be incorporated into a number of permanent, interoperable, sustainably operated Data Commons spaces.

A Data Commons, however, is only as good as the quality and quantity of the health data it contains. Government health agency CDOs can play an important role in increasing participation in Data Commons by moving their agency's data from on-premise storage units to large-scale cloud platforms that are interoperable with the NIH's Data Commons, making it more *accessible*. Equally im-

portant is to improve the quality of the shared data, which means putting it in formats that are *findable*, *interoperable*, and *reusable*—that is to say, making it machine-actionable.

POLICY: EDUCATING STAKEHOLDERS AND IMPLEMENTING DATA-SHARING REGULATIONS

The legal and regulatory landscape surrounding what data can be shared, with whom, and for what purpose can be a source of confusion and caution among health care providers and institutions that collect or generate health data. The real and/or perceived ethical, civil, privacy, or criminal risks associated with data-sharing have led many researchers and health care stakeholders to avoid doing so entirely unless they feel it is essential. This "better safe than sorry" approach can impede high-impact, timely, and resource-efficient discovery science. Furthermore, in academia, a researcher's career advancement can depend on his or her ability to attract grant funding, which in turn depends on his or her ability to generate peer-reviewed publications. In this competitive environment, researchers have little incentive to collaborate with and share their valuable data with their peers. On top of these barriers, the effort and cost associated with making data FAIR are significant.

> This "better safe than sorry" approach can impede high-impact, timely, and resource-efficient discovery science.

Government CDOs have an opportunity to overcome such barriers to data-sharing through a combination of education, support structures, and appropriate policies and governance principles. CDOs could conduct educational outreach to academics, health care providers, and other stakeholders to clarify data privacy laws such as the Health Insurance Portability and Accountability Act (HIPAA) and the Health Information Technology for Economic and Clinical Health (HITECH) Act.

The goal would be to help these stakeholders understand that, rather than prohibiting data-sharing, these laws merely define parameters around when and how to share data. Through written materials, videos, and live workshops, CDOs can clarify regulatory requirements to encourage data-sharing among health care stakeholders and individuals who are being asked to share their personal health information.

In addition to educating stakeholders, CDOs can prompt agencies to take advantage of certain policies that allow government agencies to require data-sharing. The 21st-Century Cures Act, for instance, gives the director of the NIH the authority to require that data from NIH-supported research be openly shared to accelerate the pace of biomedical research and discovery.[11] Such policies must be complemented with appropriate benefits for researchers who share their data—for instance, giving such researchers appropriate consideration for additional grants and/or naming them as co-authors on publications that use their data.

CULTURE: ENGAGE THE BROADER COMMUNITY

Open Science requires cross-sector participation and engagement from government entities, health care stakeholders, researchers, and the public. As part of their efforts to evangelize data-sharing, CDOs should consider engaging the broader community by stoking genuine interest and appreciation of the crucial role data-sharing plays in science and innovation and the benefits every player can gain from it.

One way of engaging health care stakeholders and scientists is by giving them access to appropriate government data and tools so that they can begin using shared data and seeing its value for themselves. Another way is to seek innovative solutions to health and scientific challenges using community engagement models such as code-a-thons, contests,

and crowdsourcing.[12] CDOs can also encourage the general public to ensure that their data contributes to Open Science by educating them on how they can—directly or through patient advocacy organizations—encourage researchers and clinicians to share the data they collect. Lastly, with private individuals increasingly generating large volumes of valuable health data through wearables and mobile devices, CDOs can help such individuals understand how they could best share this data with researchers.

Looking ahead

The proliferation of digital health data, coupled with advanced computational capacity and interoperable platforms such as Data Commons, gives society the basic tools to practice Open Science in health care research. However, making Open Science a reality will require all health care stakeholders, including ordinary citizens, to participate.[13]

Government CDOs can accelerate the spread of Open Science in several ways. They can establish policies and governance principles that encourage data-sharing. They can conduct education, outreach, and community engagement efforts to help stakeholders understand why and how to share data and to encourage them to do so. And they can serve as role models by making their own agencies' data available for appropriate public use.

Like all important movements, Open Science will likely face ongoing challenges. Those at the helm will need to balance the opportunities it provides with the inherent risks, including those related to data privacy and security. Of all the stakeholders in scientific discovery, government CDOs may be among the best placed to help society sort through these opportunities and risks. As public servants, they have every incentive to embrace a leadership role in promoting Open Science for the common good.

Endnotes

1. Foster, "Open Science definition," accessed August 1, 2018; Sarita Albagli, Maria Lucia Maciel, and Alexandre Hannud Abdo, *Open Science, open issues*, Brazilian Institute for Information in Science and Technology (IBICT) and Federal University of the State of Rio de Janeiro (UNIRIO), accessed August 1, 2018.

2. Mark D. Wilkinson et al., "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data* 15, no. 3 (2016): DOI: 10.1038/sdata.2016.18.

3. Mark D. Wilkinson et al., "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data* 15, no. 3 (2016).

4. National Public Radio (NPR), Ted Radio Hour, "How will open-source research help cure cancer?," transcript, July 31, 2015.

5. Center for Medicare and Medicaid Services (CMS), "Medicare and Medicaid: Keeping us healthy for 50 Years," accessed August 22, 2018.

6. US Department of Veterans Affairs, "Million Veteran Program," accessed August 1, 2018; National Institutes of Health, "The All of Us research program," accessed August 1, 2018.

7. Todd Park and Steven Vanroekel, "Introducing: Project Open Data," The White House, President Barack Obama, May 16, 2013.

8. Project Open Data, "CDO position description," accessed August 22, 2018.

9. National Institutes of Health, Office of Strategic Coordination—The Common Fund, "New models of data stewardship—Data Commons pilot," July 24, 2018.

10. Ibid.

11. Kathy L. Hudson and Francis S. Collins, "The 21st century Cures Act—A view from NIH," *New England Journal of Medicine*, 376 (2017): pp. 111–3, DOI: 10.1056/NEJMp1615745.

12. National Institutes of Health (NIH), *NIH strategic plan for data science*, accessed August 22, 2018.

13. J. Klenk et al., "Open science and the future of data analytics," in M. Edmunds, C. Hass, and E. Holve (Eds.), *Consumer Informatics and Digital Health* (Springer, 2018).

About the authors

Dr. Juergen Klenk is a principal with Deloitte Consulting LLP's Monitor Strategy practice, focusing on advancing precision medicine and data science in health care and biomedical research. He is based in Arlington, VA.

Melissa Majerol is a health care research manager with the Deloitte Center for Government Insights. She is based in Washington, DC.

Contacts

William D. Eggers
Executive director, Deloitte Center
for Government Insights
Deloitte Services LP
+1 571 882 6585
weggers@deloitte.com

Vishal Kapur
Principal
Deloitte Consulting LLP
+1 571 814 7510
vkapur@deloitte.com

Derick Masengale
Managing director
Deloitte Consulting LLP
+1 571 814 7580
dmasengale@deloitte.com

Deloitte.
Insights

Sign up for Deloitte Insights updates at www.deloitte.com/insights.

Follow @DeloitteInsight

About Deloitte Insights

Deloitte Insights publishes original articles, reports and periodicals that provide insights for businesses, the public sector and NGOs. Our goal is to draw upon research and experience from throughout our professional services organization, and that of coauthors in academia and business, to advance the conversation on a broad spectrum of topics of interest to executives and government leaders.

Deloitte Insights is an imprint of Deloitte Development LLC.

About this publication

This publication contains general information only, and none of Deloitte Touche Tohmatsu Limited, its member firms, or its and their affiliates are, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your finances or your business. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser.

None of Deloitte Touche Tohmatsu Limited, its member firms, or its and their respective affiliates shall be responsible for any loss whatsoever sustained by any person who relies on this publication.

About Deloitte

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee ("DTTL"), its network of member firms, and their related entities. DTTL and each of its member firms are legally separate and independent entities. DTTL (also referred to as "Deloitte Global") does not provide services to clients. In the United States, Deloitte refers to one or more of the US member firms of DTTL, their related entities that operate using the "Deloitte" name in the United States and their respective affiliates. Certain services may not be available to attest clients under the rules and regulations of public accounting. Please see www.deloitte.com/about to learn more about our global network of member firms.