Deloitte. Insights

beeckcenter
social impact + innovation

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.

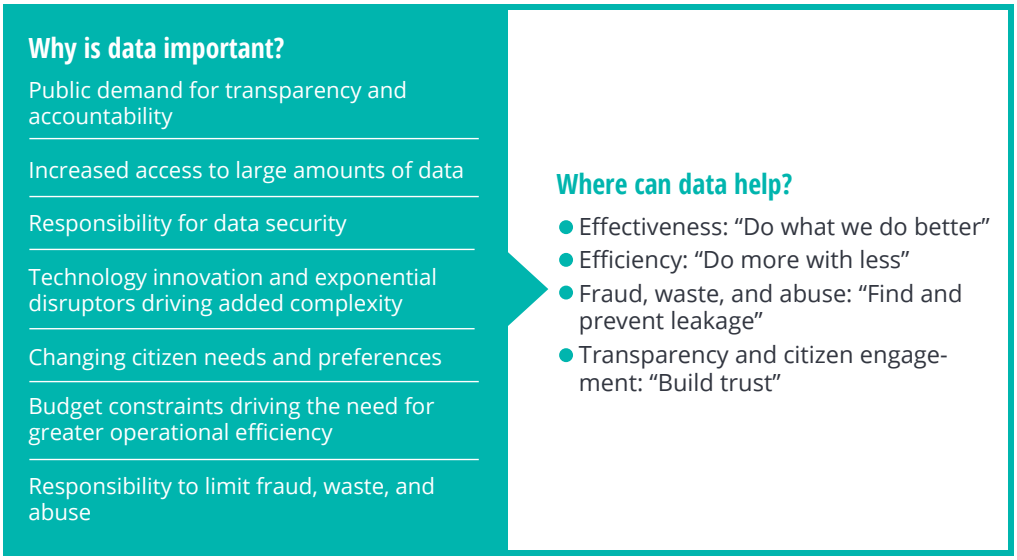
As digital tools and technologies continue to rapidly evolve, the role of data in government and the roles of those who oversee it—chief data officers (CDOs), chief information officers (CIOs), and chief technology officers (CTOs)—will require more clarity and definition if governments are to put data 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,

FIGURE 1

Why data is important to government



Source: Deloitte analysis.

as well as to operate more efficiently and get more done with less.

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.

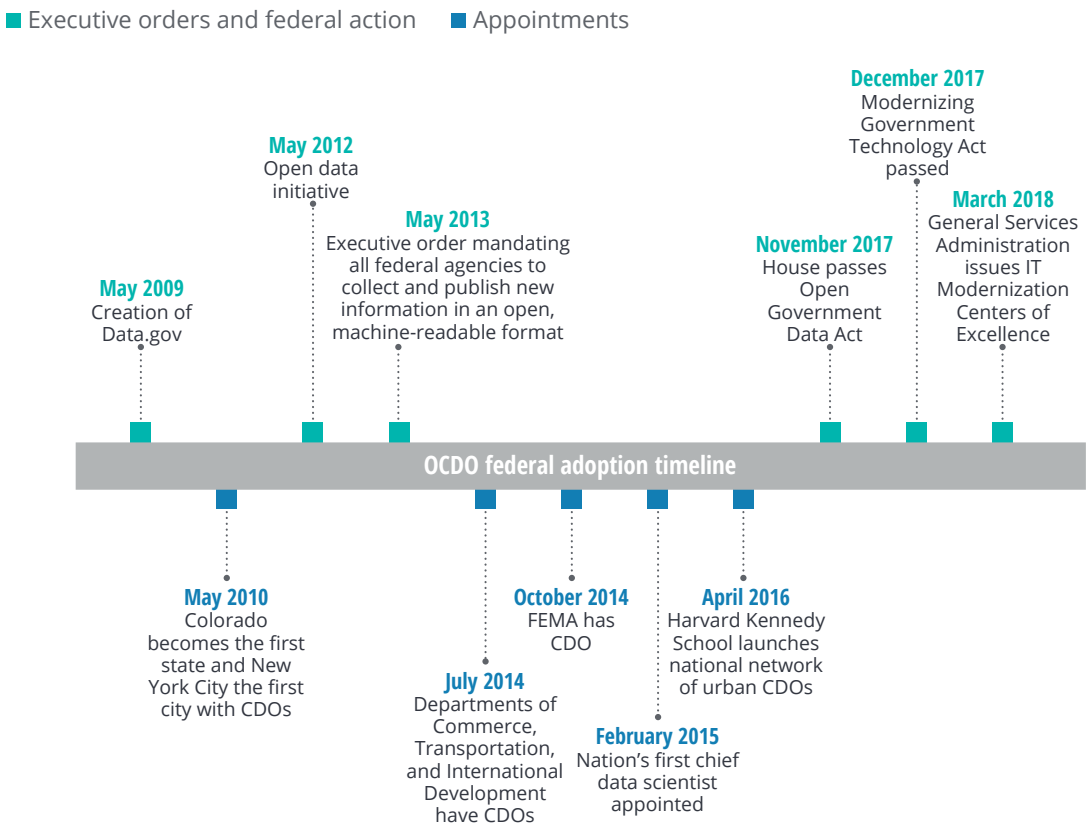
To do this, CDOs will likely need to coordinate with CIOs, CTOs, and chief information security officers across agencies to build a team, structure, and budget that can support and appropriately manage data assets. The time is ripe for CDOs to take a leadership role in organizing these key decision-makers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

FIGURE 2

Government CDOs are becoming more common



Source: Jack Moore, "Rise of the data chiefs," *NextGov,* March 18, 2015.

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.

Endnotes

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

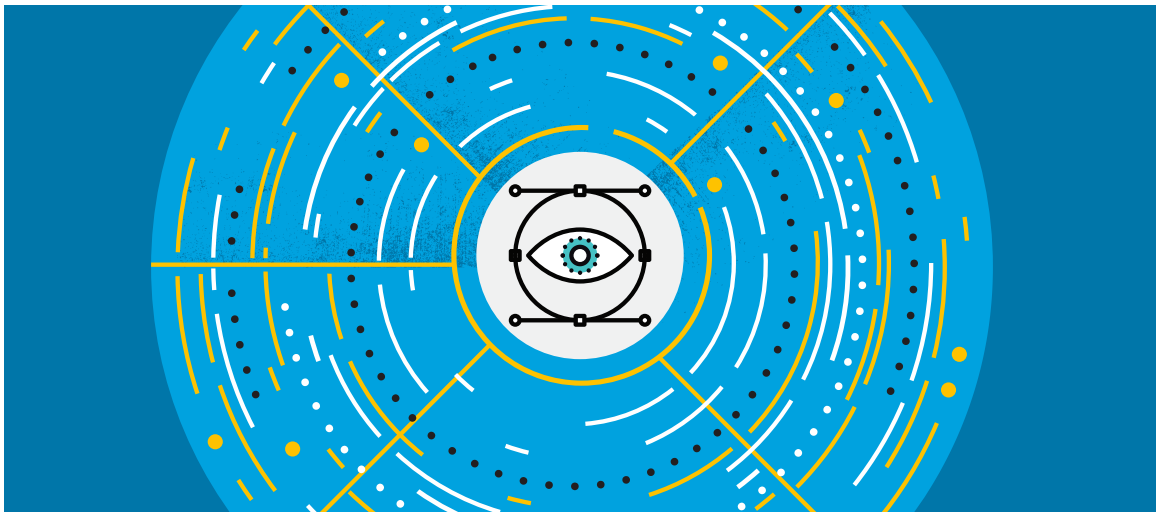
About the authors

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.

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 multimodal 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]

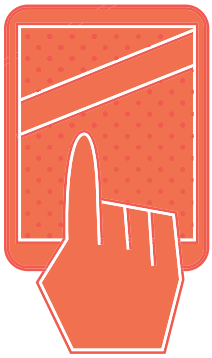
Similarly, DataUSA combines publicly accessible US government data from a variety of agencies and brings it to life in more than 2 million visualizations. In addition to allowing users to search, map, compare, and download data sets, the site also shows them what kinds of insights the data can reveal 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.

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]

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.



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 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.

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.

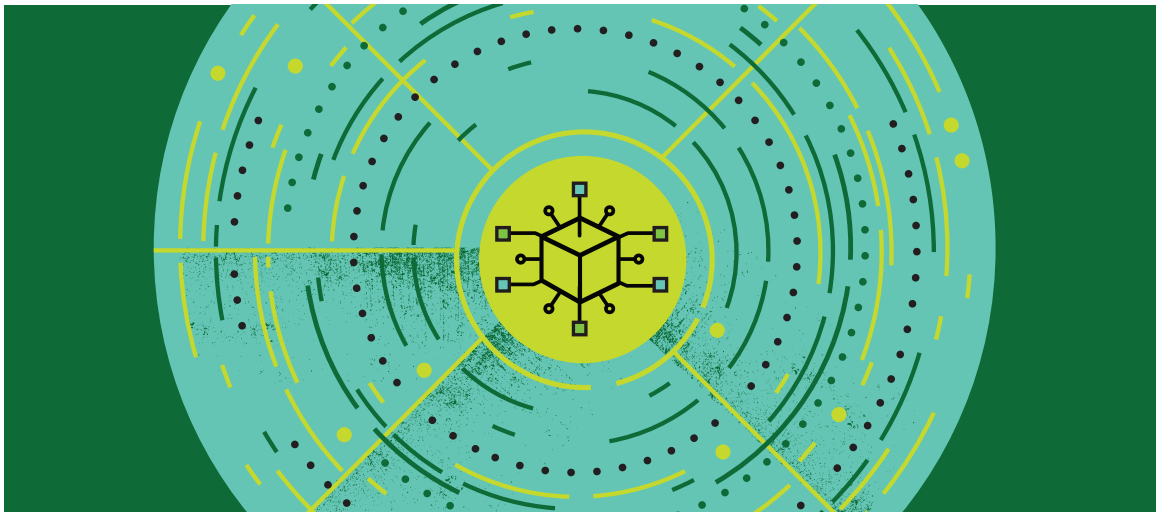
About the authors

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]

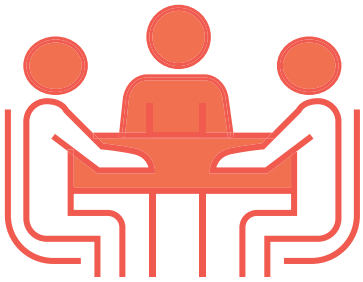
It is time for the open data community to pivot based on the lessons learned over the past decade, and governmental chief data officers (CDOs) can lead the way.

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

Intermediate approaches could allow even some sensitive data to be shared.



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, executive 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

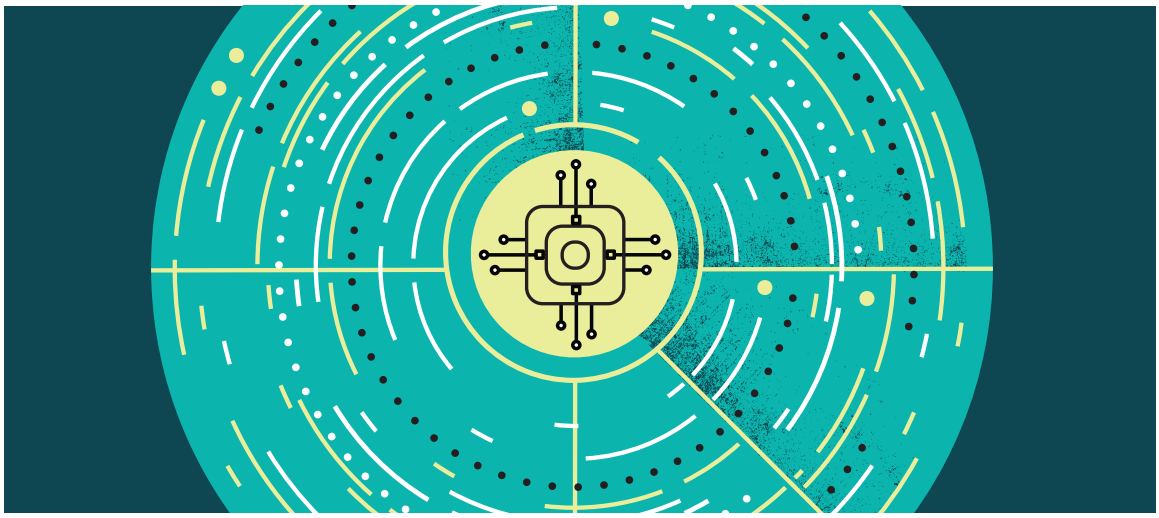
1. Josh Hicks, "This bill to track every federal dollar somehow united Cummings and Issa," *Washington Post*, April 11, 2014.
2. US government, "Digital Accountability and Transparency Act of 2014," accessed May 3, 2018.
3. Data.gov, "The home of the US government's open data," accessed May 3, 2018.
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).

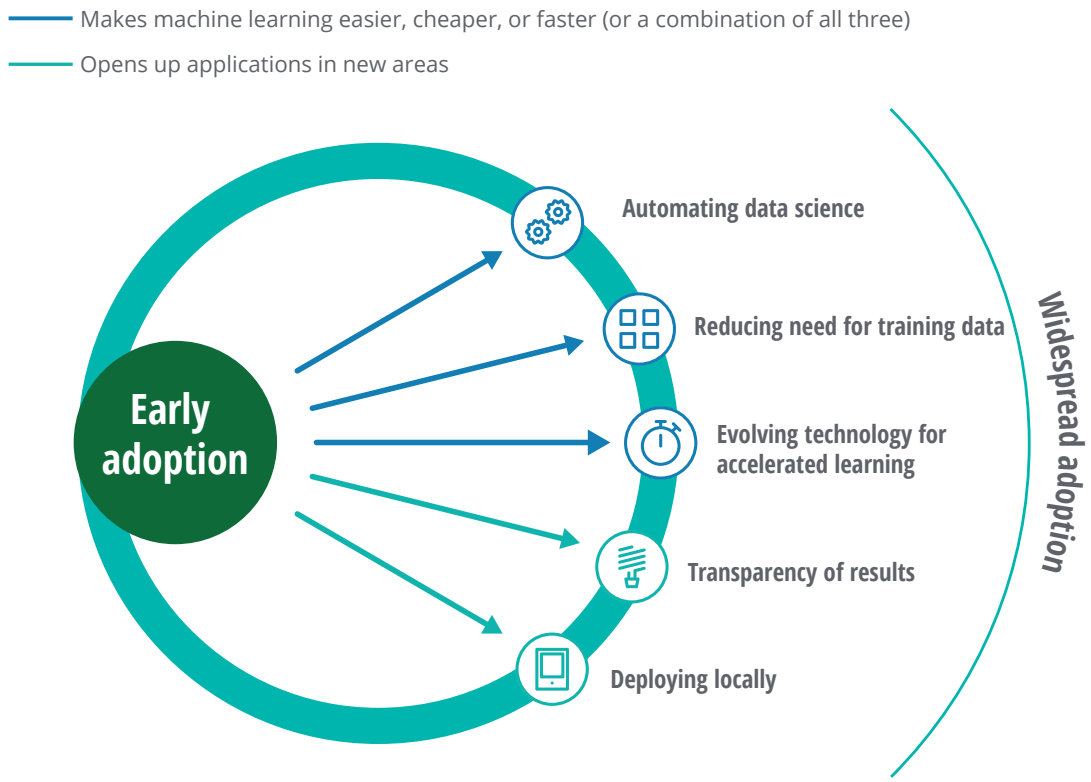
Three of these vectors include automation, data reduction, and training acceleration, which make machine learning easier, cheaper, and/or faster. The other two are model interpretability and local machine learning, both of which can open up applications in new areas.

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

FIGURE 1

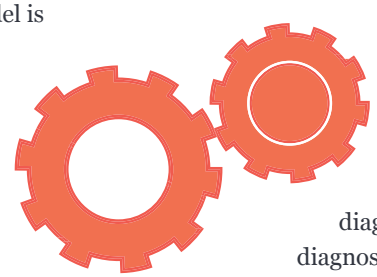
The five vectors of progress



Source: Deloitte analysis.

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 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.

A number of potentially promising techniques for reducing the amount of training data required for machine learning are emerging.



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, 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.

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]

- **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. 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

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.

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.

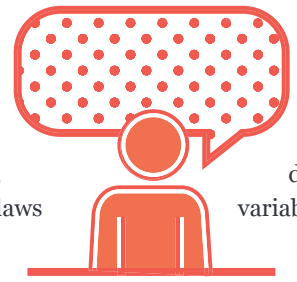
From a policy perspective, the idea that automated decisions should be "explainable" to those affected has recently gained prominence, although this is still a technically challenging proposition. In May 2018, the European Union began enforcing laws that require companies to be able to explain how their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

Researchers have developed a number of techniques to construct algorithmic models in ways in which they can better explain themselves.



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.

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:

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.

- 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 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 Journa*l, 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.

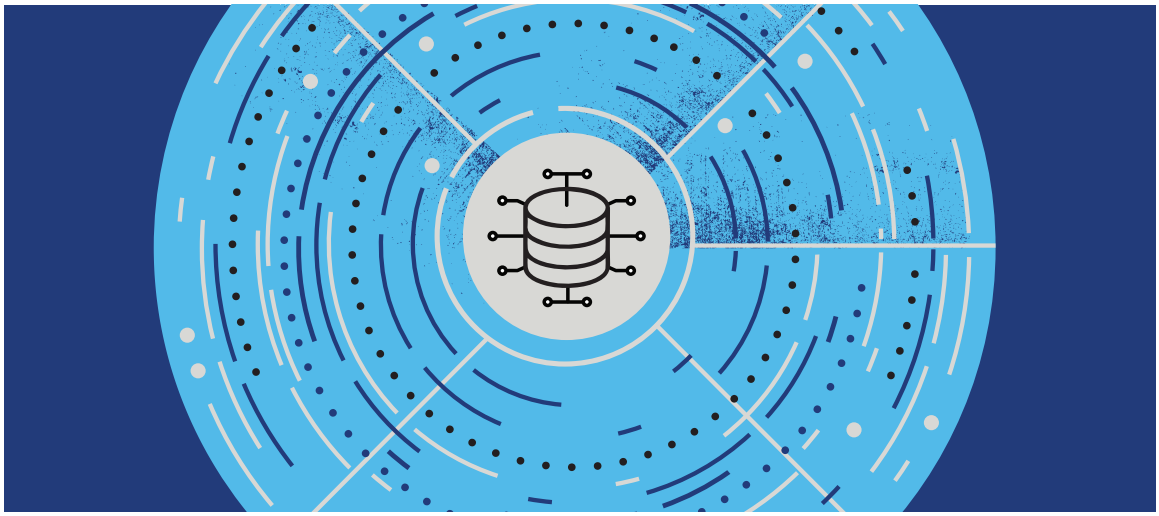
Yang Chu is a senior manager at Deloitte & Touche LLP. She is based in San Francisco, CA.

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



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.

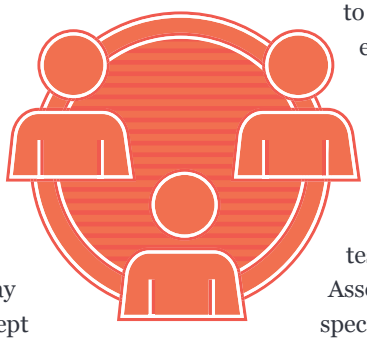
While the DATA Act deals with federal government data, it can indirectly affect how state and local governments manage their data as well. Data officers from state and local governments will likely need to be familiar with DAIMS and the Data Broker if they hope to collect grants from the federal government. 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

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



effective data analytics. Likewise, users unfamiliar with the DAIMS system may never bother to become adept with it.

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 efficient 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 important 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

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

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.

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.

Managing data ethics: A process-based approach for CDOs

Christopher Wilson

THE PROLIFERATION OF STRATEGIES for leveraging data in government is driven in large part by a desire to enhance efficiency, build civic trust, and create social value. Increasingly, however, this potential is shadowed by a recognition that new technologies and data strategies also imply a novel set of risks. If not approached carefully, innovative approaches to leveraging data can easily cause harm to individuals and communities and undermine trust in public institutions. Many of these challenges are framed as familiar ethical concepts, but the novel dynamics through which these challenges manifest themselves are much less familiar. They demand a deep and constant ethical engagement that will be challenging for many chief data officers (CDOs).

To manage these risks and the ethical obligations they imply, CDOs should work on developing institutional practices for continual learning and interaction with external experts. A process-oriented approach toward data ethics is well suited for data enthusiasts with limited resources in the fast-changing world of new technologies. Prioritizing flexibility over fixed solutions and collaboration over closed processes could lower the risk of ethical guidelines and safeguards missing their mark by providing false confidence or going out-of-date.

The old, the new, and the ugly

To a casual observer, many ethical debates about data might sound familiar. One doesn't have to engage deeply, however, before it becomes clear that contemporary informatics have radically reshaped the way we think about traditional concepts (table 1).

TABLE 1

Ethics then and now

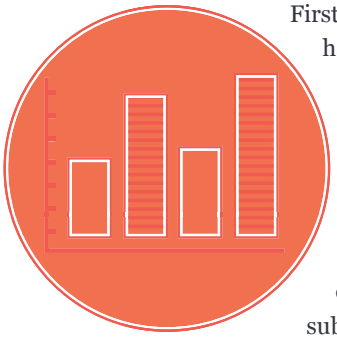
Traditional concepts	Ethics in a data context
Privacy has traditionally been understood as a *possessive concept*, in which information is intentionally shared or withheld.	Privacy is increasingly seen as *transactional and collective,* whereby access to personal information is granted on a differential basis in exchange for access to services and interactions online.
Prior and informed consent is the presumed condition for all use of an individual's personal data that is secured through research or medical processes. In effect, people must sign a form.	In the world of digital services, consent is often granted by individuals checking boxes on a Terms of Service form that they do not read. When large-scale data is collected from the internet or from administrative processes, without direct interaction with individuals, consent is rarely pursued. Informed consent seems, moreover, increasingly implausible, given an unpredictable digital context.
Sensitive data and personally identifiable information (PII) have traditionally been considered as discrete bits of information about individuals.	Sensitive data is increasingly understood to include novel types of data,[1] including data about interactions between individuals,[2] and some practitioners have also begun discussing *community identifiable information.*[3]
Data has traditionally been described as *anonymous* when all PII has been removed, and was thus understood as safe for public release.	Novel techniques for *reidentifying individuals* in anonymized data raise the question as to whether there is such a thing as anonymity in data.
There are several examples in the analog era of *repurposing of data* for uses contrary to the intended ones, but this was largely a function of reinterpretation of findings.[4]	The advent of digital data sharing and duplication opens new opportunities for possible manipulation, repurposing, and reuse of data in ways that can directly contradict the actual original data and the individual's consent.

Privacy is an excellent example. Following revelations on the controversial use of personal social media data for political campaign efforts, privacy has come to dominate popular debate about social and political communication online. This has highlighted the ease with which personal data can be collected, used, shared, and even sold without individuals' knowledge. It also has reinforced the popular claim that digital privacy applies not only to the content of messages and personal information, but to metadata on when, how, and with whom individuals interact.

This kind of data is common to all kinds of digital interactions. *Digital traces* are created any time a person logs into a government website, receives a digital service, or answers a survey on their phone. These interactions need not involve users explicitly supplying information, but data about the interaction itself is logged and can be traced back to users. Often these interactions are premised with some kind of agreement to provide that data, but recent controversies illustrate just how tenuous that permission often is, and just how important people feel it is to exercise control over any type of data in which they are reflected.

These dynamics recall the concept of *consent*. Classically understood in terms of academic and scientific research on human subjects, the idea of consent has taken a distinct turn in the context of social media and interactions online. Not only is consent often more implied than given, the potential for informed consent is complicated by the fact that it has become virtually impossible to anticipate all the ways in which personal

Digital traces are created any time a person logs into a government website, receives a digital service, or answers a survey on their phone.



information can be used, shared, and compromised. Instant duplication and sharing are some of the greatest strengths data offers to government innovation, but these advantages also completely undermine the presumption that it is possible to control how data might be used, or by whom. The internet never forgets, yet from a data protection perspective, it is also wildly unpredictable.

We might put trust in government security protocols to protect sensitive data, but even in the most stable of democratic contexts, we can never be entirely certain of what political agendas will look like a decade from now. Even when they do remain stable, however, technology can throw a wrench into future-proofing data. Consider the *mosaic effect*, the phenomenon whereby it is possible to identify individuals in an anonymized data set by combining that data with external data sets. The trouble with this phenomenon is that it is never possible to know how advanced technologies for re-identification will become—they are consistently surprising experts with their effectiveness.[5] Thus, it is never possible to determine how much deidentification is sufficient to protect data subjects. Even without such capacities or access to multiple data sets, recent events highlight how easy it is to identify individuals in deidentified data sets on the basis of public information about singular events.[6] There really no longer is any such thing as completely anonymous data.

These are profound complications to familiar ethical challenges, but digital data poses entirely new challenges as well, and at least two types of potential harm deserve mention for being central to democratic processes.

Firstly, datafied government processes have the potential to cause *procedural harm*. Data-driven and evidence-based policy is often heralded as an inherent good by many commentators. But data-driven policy is only as good as the data that drives it, and if technical capacities in government agencies are sub-optimal, poor data accuracy and reliability can produce worse policy than would have been created otherwise. The ideologies that underpin data-driven processes can also end up privileging certain social groups over others. For example, some commentators have noted how an economic rationale for innovation could inevitably privilege those who contribute most to economic activity, at the expense of the poor or economically marginalized.[7]

The collection of data can itself also have a deleterious effect on communities and be perceived as exploitative when data collection is not accompanied by visible benefits to those communities. This has led several groups, and indigenous communities in particular, to pursue guidelines for ensuring responsible data collection processes and interactions.[8] There are common threads to these guidelines, having to do with meaningful engagement and participation, but they also display a profound diversity, suggesting that any government effort to collect data on vulnerable groups will require a thoughtful allocation of time and resources to avoid getting it wrong.

Secondly, and closely related to procedural harms, data-driven government processes can lead to *preferential harms* by over- or underrepresenting specific groups. This is most easily considered in terms of the "digital divide." Individuals with

TABLE 2

Concepts and terms of art

Data creep	The tendency of data to assume an increasingly important role in strategic planning and project development, without a clear need or demand for data. Closely related to the hype surrounding data in contemporary governance discourse.
Data subjects	The people who are reflected in data, whether it is voluntarily provided or collected without their knowledge.
Digital exhaust/ digital traces	Data that is automatically created about individuals' interactions and activities online, often without their knowledge.
Mosaic effect	The phenomenon whereby anonymous data sets can be combined with other publicly available data sets to reidentify the individuals in the presumed anonymous data.
Digital divide	Refers to inequalities in access to digital media and digital literacy, and directly impacts issues of representation and voice in government data and digital engagement activities.
Data deserts/ data invisibles	Refers to the lack of representation of individuals and communities in data that is used for policymaking and service delivery. When individuals or communities do not generate relevant data, they are invisible and excluded from such processes.

access to the internet and with technological literacy will be most likely to participate in participatory processes or to be represented in digitally curated data, which can result in prioritizing the voice and representation of groups that are already well represented.[9] "Data deserts" and "data invisibles" are terms that have been coined to understand how some groups are not represented in the data used to develop government policy and services.[10] In extreme cases, institutional constraints and limited information mean that this can effectively exclude consideration of the interests of vulnerable groups from consideration in government processes.

Procedural and preferential harms are especially difficult to anticipate given the institutional tendency toward *data creep*, whereby an interest in technology's potential may drive the adoption of data and technology to be pursued as an end in itself. When data is itself presumed to be a selling point, or when projects aspire to collect and manage maximum amounts of data without clear use cases, it can be hard to spot and mitigate the kinds of ethical risks described above.

Much has been written about the novel ethical challenges and risks posed by contemporary technology and data strategies. A number of taxonomies for harm have been created in the contexts of government work, private sector activities, smart cities, and international development.[11] At the end of the day, however, the list of things that can go wrong is as long and diverse as the contexts in which data can be leveraged for social good. Context is indeed king. And for data enthusiasts in government, no context is typically more important or unique than the institutional context in which these strategies are developed.

Ethics in an institutional context

Government use of data and technology is often divided into "front office" and "back office"

activities. It can be tempting to consider ethics as most relevant to the former, where there are direct interactions with citizens and an opportunity to proactively engage on ethical issues. This would be a mistake, however. Even when data is collected without any direct interaction with citizens, there are important questions to be asked about consent and representation. Apparently anodyne methodological issues having to do with data validity and harmonization can have ethical consequences just as profound as processes related to data collection, data security, or open data publication. Perhaps most importantly, it is worth recalling that unforeseen challenges, whether related to anonymity, reuse, or perceptions of representation, can impact the public's trust in government, and aggrieved citizens are unlikely to make nuanced distinctions between back- and front-office processes.

Data ethics should be considered across governmental institutional processes and across different types of data, whether they target government efficiency or civic engagement. The most useful heuristic may be that ethical questions should be considered for every process in which data plays a role. And data plays a role in almost all contemporary projects.

Once it is clear whether or not an activity or policy development process has a data component, it is important to ask questions about what ethical risks might be present, and how to address them. In particular, there are several inflection points at which ethical vulnerabilities are most profound. These are listed in the sidebar "Common points of vulnerability in project cycles," together with examples of the types of questions that can be asked to identify ethical risks at various stages.

These are some key points in projects and processes where asking questions about ethics can be most effective. These questions should focus on potential risks and are likely most effective when pursued by groups and in conversation. When potential risks have been identified, there are many types of potential responses, some of which are listed in table 3.

It should be noted that any response or ethics strategy could further exacerbate ethical challenges by installing a false sense of security. It is very easy for civil servants to overestimate the security measures taken to protect personal data when their technical capacities are limited or when they do not have a full overview of the vulnerabilities and

TABLE 3

Potential responses and safeguards

Relational	• Providing data subjects with ongoing information about how data is used and opportunities to withdraw their consent • Designing user-friendly privacy setting notices • Engaging data subjects in collaborative processes of data collection and analysis • Different approaches to data ownership and licensing
Technical	• Data anonymization or deidentification • Data security and pseudonymization (e.g. encryption, masking) • Specific technical solutions such as differential privacy, virtual private networks (VPNs), and onion routing
Design	• Providing data collection materials and analysis in multiple languages and for user accessibility • UX design sensitive to particular literacies, digital divides, or political contexts
Regulatory	• Ethical policies and frameworks, including guidelines • Oversight bodies • Adherence to the sector or issue-specific standards

COMMON POINTS OF VULNERABILITY IN PROJECT CYCLES AND QUESTIONS TO CONSIDER

- **Project planning and design**
  Does the project collect the right data and the right amount of data? What are the ethical implications? What are the most immediate risks and who are the most important stakeholders? What resources are needed to manage data ethics? What are the opportunities for engaging with experts and the public along the way? How can the data ethics strategy be documented? What are the most important points of vulnerability? What data protection strategies and controls to deploy in case of a breach?

  Note that organizations can employ a "privacy-by-design" strategy to comprehensively address vulnerabilities across the project life cycle. This approach aims to protect privacy by incorporating it upfront in the design of technologies, processes, and infrastructure. It can help restrict the collection of personal data, enable stricter data encryption processes, anonymize personal data, and address data expiry.

- **Data collection and consent**
  Who owns data? Who should give permissions and consent? How will the act of data collection affect the people it is collected from and their relationship to the government? Is more data being collected than necessary? Is the data secure during collection? Are there any methodological issues that affect the reliability of the data? Would the data collected be seen as valid by the people it is collected from? How to procure consent for alternate use of the same data?

- **Data maintenance**
  Who has access to the data? Does the level of data security match the potential threats to the data? Is there a timeline for how long data will be maintained and when it will be destroyed? What are the specific guidelines around data portability? Which formats and structures are used to share data with the data subjects, other data controllers, and trusted third-party vendors?

- **Data analysis**
  Does the data have any biases or gaps? Does other information exist that would contradict the data or the conclusions being drawn? Are there opportunities to involve data subjects in the analysis? How can you avoid inherent algorithmic bias in the data analysis?

- **Data sharing and publication**
  Is an appropriate license selected for open data? Does the data contain sensitive information? Have all potential threats and harms from releasing or publishing the data been identified? Are there explicit ethical standards in data-sharing agreements?

- **Data use, reuse, and destruction**
  What were the ethical issues surrounding how the data was originally collected? Has the context changed since then in a way that requires regaining consent of data subjects? Are there data ownership or licensing issues to be aware of? What methods are used for secure data destruction?

threats that might be posed to that data or the individuals reflected in it.

Chief data officers and information officers likely have an advantage in this regard but will likely also struggle to keep up to date on all cutting-edge data ethics issues. This is an inevitable challenge for people working to advance innovative use of data inside of government, where demands are often high and resources low.

It is also worth noting that the danger of false security can be just as important for policy responses as it is for technical responses. It might be easy to create a consent policy for digital service delivery that would check internal boxes and satisfy internal project managers, but eventually could lead to resentment and anger from communities that did not understand how the data would be used or shared. The danger that ethical regulatory measures will produce a false sense of security has on occasion been blamed on poor "hard" technical capacities.[12] However, the "soft" capacities required to assess ethical risks, conduct threat assessments, and anticipate how specific constituencies will experience data-driven processes are typically just as important, and can be just as challenging for civil servants to secure.

In addition to capacity constraints, the use of data in government is often subjected to a host of other limitations, including resource constraints, institutional cultures, and regulatory frameworks. Each of these poses unique challenges to managing ethical risks. Resource constraints are perhaps the most obvious, as community consultations, developing tiered permission structures for data, and even SSL certificates all cost money and time. Some of the more novel and aggressive approaches to managing ethical risks might clash with political priorities or cultures for short-term outputs. Regulations such as the Paperwork Reduction Act are notorious for the impediments they pose for proactive engagement with the public.[13]

These challenges will manifest differently for different types of projects and in different contexts. Some will require deep familiarity of differential privacy models or the UX implications of 4G penetration in specific geographic areas. Others will require deep expertise in survey design or facilitation in multiple languages. Nearly all will likely require close and thoughtful deliberation to determine what the ethical risks are and how best to manage them. The ethical challenges surrounding innovative data use are generally never as straightforward as they first appear to be. It's simply not possible for any one person or team to be an expert in all of the areas demanded by responsible data management. Developing cultures and processes for continual learning and adaptation is key.

Recommendations for CDOs: A process-focused response

Ethically motivated CDOs could find themselves in a uniquely challenging situation. The dynamic nature of data and technology means that it is nearly impossible to anticipate what kinds of resources and expertise will be needed to meet the ethical challenges posed by data-driven projects before one actually engages deeply with them. Even if it were possible to anticipate this, however, the limitations imposed by most government institutions would make it difficult to secure all the resources and expertise necessary, and the fundamentally ambiguous nature of ethical dilemmas makes it difficult to prioritize data ethics management over daily work.

Progressively assessing and meeting these challenges requires a degree of flexibility that might not come naturally to all institutional contexts. But there are a few strategies that can help.

PRIORITIZE PROCESSES, NOT SOLUTIONS

Whenever possible, CDOs should establish flexible systems for assessing and engaging with the ethical challenges that surround data-driven projects. Identifying a group of people within and across teams that are ready to reflect on these issues and

NETWORKS AND RESOURCES FOR MANAGING DATA ETHICS

There are several nonprofit, private-sector, and research-focused *communities and events* that can be useful for government CDOs. The Responsible Data website curates an active discussion list on a broad range of technology ethics issues.[14] The International Association of Privacy Professionals (IAPP) manages a community list serve,[15] and the conference on Fairness, Accountability, and Transparency in Machine Learning convenes academics and practitioners annually.[16]

Past activities and consultations like the UK government's public dialogue on the ethics of data in government can also provide useful information,[17] and has resulted in the adoption of a governmentwide data ethics framework, which includes a workbook and guiding questions for addressing ethical issues.[18] Responding to the EU General Data Protection Regulation (GDPR) regulations, the International Organization for Standardization (ISO) has set up a new project committee to develop guidelines to embed privacy into the design stages of a product or service.[19]

Several other *useful tools and frameworks* have been produced. The Center for Democracy and Technology (CDT) has developed a Digital Decisions Tool to help ethical decision-making into the design of algorithms.[20] The Utrecht Data School has developed a tool, data ethics decision aid (DEDA) that is currently being implemented by various municipalities in the Netherlands.[21] The Michigan Department of Transportation has produced a decision-support tool dealing with privacy concerns surrounding intelligent transportation systems,[22] and the IAPP provides a platform for managing digital consent processes.[23] The Sunlight Foundation has developed a set of tools to help city governments ensure that open data projects map community data needs.[24]

Many organizations also offer *trainings and capacity development*, including the IAPP,[25] journalism and nonprofit groups like the O'Reilly Group,[26] and the National Institute of Standards and Technology, which offers trainings on specific activities such as conducting privacy threat assessments.[27]

Several *white papers and reports* also offer a general overview of issues and approaches, including the European Public Sector Information Platform's report on ethical and responsible use of open government data[28] and Tilburg University's report on operationalizing public sector data ethics.[29]

This list is not exhaustive, but it does illustrate the breadth of available resources, and might provide a useful starting point for learning more. Participants in the MERL Tech Conference on technology for monitoring, evaluation, research, and learning also maintain a hackpad with comparable networks and resources for managing data ethics in the international development sector.[30]

are willing to be on standby for discussions can greatly enhance the efficiency of discussions. Setting up open invitations at the milestones and inflection points for every project or activity that has a data component (see sidebar, "Networks and resources for managing data ethics") can facilitate constant attention. Also, it allows the project team to step back and explore ways to embed privacy principles in the early design stages. Keeping these discussions open and informal can help create the sense of dedication and flexibility often necessary to tackle complex challenges in contexts with limited resources. Keeping them regular can help instill an institutional culture of being thoughtful about data ethics.

In some contexts, it might make sense to formalize processes, creating bodies similar to the NYC task force mandated to assess equity, fairness, and accountability in how the city deploys algorithms. In other contexts, it may make more sense to consider alternative formats like data ethics lunches or short 30-minute brainstorming sessions immediately following standing meetings and try

to get everybody on the same page about this being an effort to build and sustain meaningful trust between constituents and government.

Flexibility can be key to making this kind of engagement effective, but it's also important to be prepared. For each project, consider identifying a key set of issues or groups that are worth extra attention, and prioritize getting more than two people into discussion regularly. Group conversations can help surface creative solutions and different points of view, and having them early can help prevent unpleasant surprises.

ENGAGE WITH EXPERTS OUTSIDE THE BOX

A process-based approach to managing data ethics will only be effective if teams have the capacity to address the risks that are identified, and this will rarely be the case in resource-strapped government institutions. CDOs should invest in cultivating a broad familiarity with discourses on data ethics and responsible data and the experts and communities that drive those discourses. Doing so can help build the capacity of the teams and stakeholders inside government and also support innovative approaches to solving specific ethical challenges through collaboration.

Many sources of information and expertise are available for managing data ethics. Research communities regularly publish relevant reports and white papers. Government networks discuss the pros and cons of different policy options. Civil society networks advance cutting-edge thinking around data ethics and sometimes provide direct support to government actors. Increasingly, private sector organizations, funders, consultants, and technology-driven companies are also offering resources.

Becoming familiar with these communities is a first step; just subscribing to a few RSS feeds can provide prompts every day, flagging issues that need attention and honing it to keep ethical challenges from slipping through the cracks. Cultivating relationships with experts and advocates can provide important resources during crises. Attending conferences and events can provide a host of insights and contacts in this area.

OPEN UP INTERNAL ETHICAL PROCESSES

Perhaps most importantly, process-focused approaches to managing data ethics should be open about their processes. Though some government information will need to be kept private for security reasons, CDOs should encourage discussions about keeping the management of ethics open and transparent whenever possible. This adheres to an important emerging norm regarding open government, but it's also critical for making data ethics strategies effective.

Open source digital security systems provide an illustrative example. A number of services are available for encrypting communications, but digital security experts recommend using open source digital security software because its source code is consistently audited and reviewed by an army of passionate technologists who are vigilant to vulnerabilities or flaws. As it is not possible to audit closed source encryption tools in the same way, it is not possible to know when and to what degree the security of those tools has been compromised.

In much the same way, government data programs may be working to keep information or personal data secure and private, but by having open discussions about how to do so, they typically build trust with the communities they are trying to serve. They also open up the possibility of input and corrections that can improve data ethics strategies in the long and short run.

This kind of openness could involve describing processes in op-eds, blog posts, or event presentations or inviting the occasional expert to data ethics lunches or the other flexible activities described above. Or it might involve the publication of documents, regular interaction with the press, or a more structured way of engaging with the communities that are likely to be affected by data ethics. Whatever the mechanism or the

ENABLING STRONG DATA GOVERNANCE AND ETHICAL USE OF DATA

This article focuses on improving data ethics through process changes and improvements. However, it needs to be acknowledged that although the process-based approach is important and sometimes a critical first step, it's not the only way to achieve privacy protection outcomes. There are also a host of technologies, strategies, and solutions that can enable strong data governance and ethical use of data.

- **Data discovery, mapping, and inventorying solutions:** These solutions can help you to understand data across organizational silos, map the data journey from the point of collection to other systems (both internal and external), and determine the data format and validity in the system.[31]
- **Consent and cookie management solutions:** There are many software-as-a-service (SaaS) solutions that can manage your website's user consents. These services typically manage your website's cookies, automatically detect all tracking technologies on your website, and provide full transparency to both the organization and the user.[32]
- **Individual rights management solutions:** This is a subset of digital rights management technologies that provide users the right to access, use, reuse, move, and destroy their data.[33]
- **Data protection impact assessment (DPIA):** An organization can conduct an extensive impact assessment of its data, systems, and other digital assets to identify, assess, and mitigate future privacy risks. These assessments can help organizations understand their technical capacities and evolve appropriate organizational measures around privacy.[34]
- **Incident management solutions:** These solutions help an organization to identify, analyze, quarantine, and mitigate threats to avoid future recurrences. Such software solutions can help restrict threats from becoming full-blown attacks or breaches.

particular constraints on CDOs, a default inclination toward open processes will contribute toward building trust and creating effective data ethics strategies.

Navigating the trade-off between rigor and efficiency

Data is hard. So are ethics. There is nothing easy about their interface either, and CDOs operate in a particularly challenging environment. This article has not provided any out-of-the-box answers to those challenges, because there are none. A proactive and meaningful approach to data ethics will typically involve compromises in efficiency and effectiveness. Ethical behavior isn't easy and civic trust in a datafied society isn't free. Being attuned to the ethical challenges surrounding government data means that CDOs and other government data enthusiasts will necessarily be faced with trade-offs between data ethics rigor and efficiency—between the perfect and the good. When that time comes, it's important to have realistic expectations about the cost of strong ethics, the risks of getting it wrong, and the wealth of options and resources for trying hard to get it right. A careful and informed balancing of these trade-offs can increase CDOs' chances of managing data ethics in a way that helps build trust in government and hopefully avoids disaster following the most innovative applications of data in governance.

Toward that end, this article aims to provide a brief discussion on why data ethics is such a challenging and important phenomenon, and to offer some entry points for informed, critical, and consistent ways of addressing them. The hard work of actually pursuing that challenge falls on the CDO.

Endnotes

1. Acxiom, "Defining 'sensitive' in world of consumer data," July 27, 2015; Safecredit, "Home," accessed July 30, 2018.
2. See Jonathan Mayer, Patrick Mutchler, and John C. Mitchell, "Evaluating the privacy properties of telephone metadata," PNAS, May 16, 2016.
3. See UN OCHA, "Humanitarianism in the age of cyber-warfare," p.6, October 2014.
4. For example, research on vulnerable populations (on substance abuse and violence among Australian aboriginals in the '80s and gay men in the United States in the '90s) has been leveraged to promote discriminatory legislation.
5. ASPE, "Minimizing disclosure risk in HHS open data initiatives," accessed January 8, 2019; Rich Murnane's Blog, "The 'mosaic effect," January 8, 2011.
6. Olivia Solon, "'Data is a fingerprint': Why you aren't as anonymous as you think online," *Guardian*, July 13, 2018.
7. Abhishek Gupta, "AI in smart cities: Privacy, trust and ethics," New Cities, May 7, 2018.
8. Responsible Data, "Indigenous peoples and responsible data: An introductory reading list," September 22, 2017.
9. This phenomenon, has been well documented in research on government crowdsourcing and participation initiatives, for example. See Helen K. Liu, "Exploring online engagement in public policy consultation: The crowd or the few?," *Australian Journal of Public Administration* 76, no. 1 (2017): pp. 1–15; and Benjamin Y. Clark and Jeffrey Brudney, "Too much of a good thing? Frequent flyers and the implications for the coproduction of public service delivery," SSRN, March 28, 2017.
10. Daniel Castro, "The rise of data poverty in America," Center for Data Innovation, September 10, 2014; Justin Longo et al., "Being digitally invisible: Policy analytics and a new type of digital divide," *Policy & Internet* 9999, no. 9999 (2017): DOI:10.1002/poi3.144.
11. Daniel J. Solove, Marc Rotenberg, and Paul M. Schwartz, *Privacy, Information, and Technology* (Aspen Publishers Online, 2006); Responsible Data, "Responsible Data reflection stories," accessed November 26, 2018.
12. Ben Rossi, "Why the government's Data Science Ethical Framework is a recipe for disaster," Information Age, June 2, 2016.
13. Eric Keller and Kathy Conrad, "Overcoming a big obstacle to citizen feedback," Government Executive, accessed June 8, 2018.
14. Responsible Data, "Home," accessed July 31.
15. IAPP, "Resource center," accessed July 31, 2018.
16. Fairness, Accountability, and Transparency in Machine Learning, "Home," accessed July 31, 2018.
17. Ipsos MORI Social Research Institute, *Public dialogue on the ethics of data science in government*, May 2016.
18. Gov.uk, "Data ethics framework," August 30, 2018.
19. Clare Naden, "Data privacy by design: A new standard ensures consumer privacy at every step," ISO, May 11, 2018.
20. Natasha Duarte, "Digital decisions tool," CDT, August 8, 2017.
21. Utrecht Data School, "Data ethics decision aid (DEDA)," accessed July 31, 2018.

22. Michigan Department of Transportation and Center for Automotive Research, "ITS data ethics in the public sector," June 2014.

23. IAPP.

24. Greg Jordan-Detamore, "Use our new tools to learn about community data needs," Sunlight Foundation, July 9, 2018.

25. Ibid.

26. O'Reilly, "Data ethics: Designing for fairness in the age of algorithms," accessed July 31 2018.

27. NIST, "Privacy risk assessment: A prerequisite for privacy risk management," accessed July 31, 2018.

28. Karolis Granickas "Ethical and responsible use of open government data," EPSI Platform, February 2015.

29. L. E .M. Taylor, R. E. Leenes, and S. van Schendel, "Public sector data ethics: From principles to practice," Universiteit Van Tilberg, April 2017.

30. MERL Tech., "Responsible data hackpad," accessed December 19, 2018.

31. Rita Heimes, "Top 10 operational responses to GDPR: Data inventory and mapping," IAPP, February 1, 2018.

32. Cookiebot, "How to manage consents in compliance with the GDPR," accessed November 22, 2018.

33. PRNewswire, "TrustArc launches 3-in-1 GDPR Individual Rights Management solution," January 24, 2018.

34. IT Governance, "Data Protection Impact Assessment under the GDPR," accessed November 22, 2018.

About the author

Christopher Wilson is a research fellow at the University of Oslo and the Beeck Center for Social Impact + Innovation at Georgetown University. He is based in Oslo, Norway.

Pump your own data: Maximizing the data lake investment

Paul Needleman, Eric Rothschild, and Stephen Schiavone

ORGANIZATIONS ARE CONSTANTLY LOOKING for better ways to turn data into insights, which is why many government agencies are now exploring the concept of *data lakes*.

Data lakes combine distributed storage with rapid access to data, which can allow for faster analysis than more traditional methods such as enterprise data warehouses. Data lakes are special, in part, because they provide business users with direct access to raw data without significant IT involvement. This "self-service" access lets users quickly analyze data for insights. Because they store the full spectrum of an enterprise's data, data lakes can break down the challenge of data silos that often bedevil data users. Implemented correctly, data lakes provide insight at the point of action, and give users the ability to draw on any data at any time to inform decision-making.

Data lakes store information in its raw and unfiltered form—whether it is structured, semi-structured, or unstructured. A data lake performs little automated data cleansing or transformation. Instead, data lakes shift the responsibility of data preparation to the business.

Providing broad access to raw data presents both a challenge and an opportunity for CDOs. By enabling easy access to enterprise data, data lakes allow subject matter experts to perform data analytics without going through an IT "middleman." At the same time, however, these data lakes must provide users with enough context for the data to be usable—and useful.

CDOs can play a major role in the development of a data lake, providing a strategic vision that encourages usability, security, and operational impact.

RAPID INSIGHTS AT A FEDERAL MANUFACTURING FACILITY

A federal manufacturing facility's CIO wanted faster access to large volumes of data in its native format to scale and adapt to the changing needs of the business. To accomplish this, the facility implemented a data lake, which stores distributed servers to efficiently process and store nonrelational data. This platform complements the organization's existing data warehouse to support self-service and open-ended data discovery. Users now have on-demand access to business-created data sets from raw data, thereby reducing the time to access data from 16 to three weeks.

Establishing the data lake platform: Avoiding a data swamp

A poorly executed data lake is known as a *data swamp*: a place where data goes in, but does not come out. To ensure that a data lake provides value to an organization, a CDO should take some important steps.

METADATA: HELP USERS MAKE SENSE OF THE DATA

Imagine being turned loose in a library without a catalog or the Dewey Decimal System, and where the books are untitled. All the information in the books is there, but good luck turning it into useful insight. The same goes for data lakes: To reap the data's value, users need a metadata "map" to locate, make sense of, and draw relationships among the raw data stored within. This metadata layer provides additional context for data that flows through to the data lake, tagging information for ease of use later on.

Too often, raw data is stored with insufficient metadata to give the user enough context to make gainful use of it. CDOs can help combat this situation by acting as a metadata champion. In this capacity, the CDO should make certain that the metadata in the data lakes he or she oversees is well understood and documented, and that the appropriate business users are aware of how to use it.

SECURITY: CONTROL ACCESS FOR INDIVIDUAL ROLES

By putting appropriate security and controls in place, CDOs will be better positioned to meet increasingly stringent compliance requirements. Given the vast amount of information data lakes typically contain, CDOs need to control which users have access to which parts of the data.

Role-based access control (RBAC) is a control mechanism defined around roles and privileges through security groups. The components of RBAC—such as role permissions, user roles, and role-to-role relationships—make it simple to grant individuals specific access and use rights, minimizing the risk of noncleared users accessing sensitive data. Within most data lake environments, security typically can be controlled with great precision, at a file, table, column, row, or search level.

Besides improving security, role-based access simplifies the user experience because it provides users with only the data they need. It also enhances consistency, which can build users' trust in the accessed data; this, in turn, can increase user adoption.

DATA PREPARATION: EQUIP USERS TO CLEANSE DATA SETS

Preparing a new data set can be an extremely time-consuming activity that can stymie data analysis before it begins. To obtain a reliable analytic output, it's usually necessary to cleanse, consolidate, and standardize the data going

in—and with a data lake, the responsibility of preparing the data falls largely into the hands of the business users. This means the CDO must work with business users to give them tools for data prep.

Thankfully, software is emerging to help with the work of data preparation. The IT organization should work collaboratively with the data lake's business users to create tools and processes that allow them to prepare and customize data sets without needing to know technical code—and without the IT department's assistance. Equipped with the right tools and know-how, business data users can prepare data efficiently, allowing them to focus the bulk of their efforts on data analysis.

ENABLEMENT: ALLOW USERS TO USE FAMILIAR TOOLS

Self-service data analysis will go more smoothly if users can use familiar tools rather than having to learn new technologies. CDOs should strive to ensure that the business's data lake(s) will be compatible with the tools the business currently uses. This will greatly enhance the data lake platform's effectiveness and adoption. Fortunately, data lakes support many varieties of third-party software that leverage SQL-like commands, as well as open source languages such as Python and R.

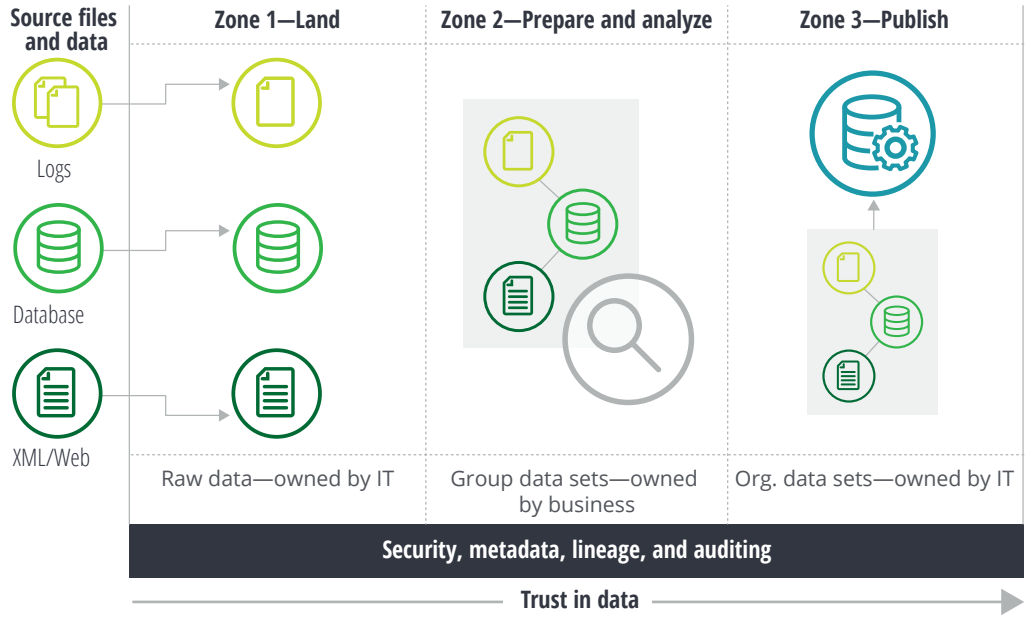
GOVERNANCE: MAINTAIN CONTROLS FOR NON-IT RESOURCES

Once users have access to data, they will use it—which is the whole point of self-service. But what if a user makes an error in data extraction, leading to an inaccurate result? Self-service is fine for exploring data, but for mission-critical decisions or widespread dissemination, the analytical outcomes must be governed in a way to guarantee trust.

One approach to maintaining appropriate governance controls is to use "zones" for data

FIGURE 1

Creating data "zones" with different verification requirements can enhance analytical reliability and accuracy



Source: Deloitte analysis.

ENABLING ANALYTICS AT A FEDERAL AGENCY

A federal agency CIO team built and deployed analytics tools to support operations to influence an insight-driven organization. The goal was to create an environment where stakeholders were consistently incorporating analysis, data, and reasoning into the decision-making process across the organization, such as enhancing data infrastructure. To give users the ability to utilize these tools to their full potential, an "Analytics University" was implemented. This was well-received; more than 20,000 field employees completed level 1 courses, with 90 percent saying they would recommend them to a colleague. The support by upper management to invest in the use and understanding of data analytics across the organization encouraged a data-driven culture, and this culture shift continues to enable business adoption of big data technologies.

access and sharing, with different zones allowing for different levels of review and scrutiny (figure 1). This allows users to explore data for inquiry without exhaustive review while simultaneously requiring that data that will be broadly shared or used in critical decisions will be appropriately vetted. With such controls in place, a data lake's ecosystem can perform nimbly while limiting the impact of mistakes in extraction or interpretation.

Figure 1 illustrates one possible governance structure for a data lake ecosystem in which different zones offer appropriate governance controls:

- **Zone 1** is owned by IT and stores copies of the raw data through the ingestion process. This zone contains the least trustworthy data and requires the most vetting.
- **Zone 2** is where business users can create their own data sets based on raw data from zone 1 as well as external data sources. Zone 2 would be trusted for group (i.e., office or division) use, and could be controlled by group-sharing settings.
- **Zone 3** data sets, maintained by IT, are vetted and stored in optimal formats before being shared with the broader organization. Only data in zone 3 would be trusted for broad organizational uses.

Adopting a culture of self-service analytics: Empowering people to use data lakes

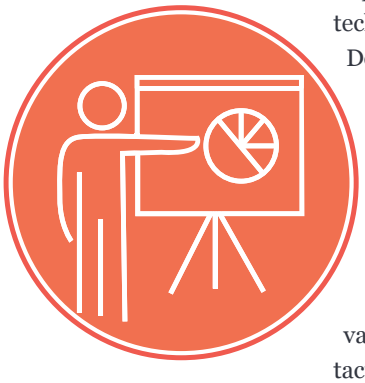
Implementing a data lake is more than a technical endeavor. Ideally, the establishment of a data lake will be accompanied by a culture shift that embeds data-driven thinking across the enterprise, fostering collaboration and openness among various stakeholders. The CDO's leadership through this transition is critical in order to give employees the resources and knowledge needed to turn data into action.

INVEST IN DATA LEADERS

CDOs are responsible for more than just the data in the data lake; they are also responsible for helping to equip the workforce with the data skills they need to effectively use the data lake. One way to help achieve this is for CDOs to advocate for and invest in employees that have the necessary skills, attitude, and enthusiasm. Specialized trainings, town halls, data boot camps—a variety of approaches may be needed to foster not only the technical skills, but the courage to change outdated approaches that trap data in impenetrable silos. The best CDOs will create an organization of data leaders.

CDOs may need to work with senior business leaders and HR in the drive for change. They should strive to overcome barriers, highlight data champions throughout the organization, and lead by example.

Work collaboratively with the data lake's business users to create tools and processes that allow them to prepare and customize data sets without needing to know technical code.



PRACTICE NIMBLE GOVERNANCE

Governance over data lakes needs to walk a very fine line to support effective gatekeeping for the data lake while not impeding users' speed or success in using it. Traditionally, governance bodies for data defined terms, established calculations, and presented a single voice for data. While this is still necessary for data lakes, governance bodies for a data lake also should establish best practices for working with the data. This includes activities such as working with business users to review data outputs and prioritizing ingestion within the data lake environment. Organizations should establish thorough policy-based governance to control who loads which data into the data lake and when or how it is loaded.

KEEP CURRENT ON THE TECHNOLOGY

Technology is never static; it will always evolve, improve, and disrupt at a dizzying speed. The technology surrounding data lakes is no exception. Thus, CDOs must continue to make strategic investments in their data lake platforms to update them with new technologies.

To do this effectively, CDOs must educate themselves about current opportunities for improving the data lake and about new technologies that will reduce users' burden. Doing so will open up the ability for more users to use data in their everyday work and decisions. Keeping oneself up to date is straightforward: Read journals and trades, attend conferences and meetups, talk to the users, and be critical of easy-sounding solutions. This will empower a CDO to sift through the vaporware, buzzwords, and flash to identify tactical, practical, and necessary improvements.

To invest or not to invest?

The challenges associated with traditional data storage platforms have led today's business leaders to look for modern, forward-looking, flexible solutions. Data lakes are one such solution that can help government agencies utilize information in ways vastly different than was previously possible.

It would be easy to say that there is a one-size-fits-all approach and that every organization should have a data lake, but this is not true. A data lake is not a silver bullet, and it is important for CDOs to evaluate their organization's specific needs before making that investment. By planning properly, understanding user needs, educating themselves on the potential pitfalls, and fostering collaboration, a CDO can gain a solid foundation for making the decision.

About the authors

Paul Needleman is a specialist master with Deloitte Consulting LLP. He is based in Rosslyn, Va.

Eric Rothschild is a senior consultant with Deloitte Consulting LLP. He is based in Rosslyn, Va.

Stephen Schiavone is a business technology analyst with Deloitte Consulting LLP. He is based in Rosslyn, Va.

Data as an asset: Defining and implementing a data strategy

Max Duhe, Matt Gracie, Chris Maroon, and Tess Webre

HOW DO YOU GET your organization to value data? Accurate data is the fuel that propels government organizations toward achieving their mission. Increasingly, government organizations agree it is time to view data as a critical strategic asset and treat it accordingly.[1]

Managing and leveraging data typically falls to the chief data officer (CDO). In the Big Data Executive Survey of 2017, 41.4 percent of the executives surveyed believed that the CDO's primary role should be to manage and leverage data as an enterprise business asset.[2] However, many government organizations fail to invest in the resources necessary to realize the data's inherent value.

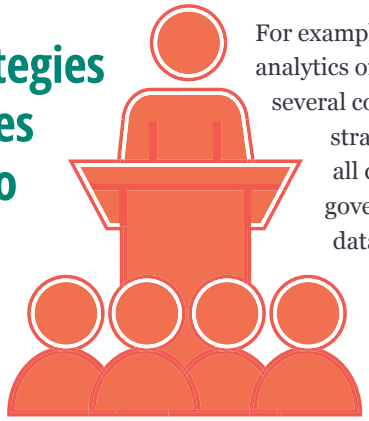
It is easy to become overwhelmed by the challenge of turning data from an afterthought into a core facet of business operations. Organizations can become paralyzed because they don't know where to begin. But CDOs can take comfort in knowing that change doesn't happen overnight.

To unlock the value of an organization's data, a CDO should develop and implement a clear data strategy. The data strategy can help organizations take a strategic view of data and use it more effectively to drive results. The best data strategies are generally tailored to the organization's needs and help the CDO engage necessary stakeholders, plan for the future, implement strategic projects, develop partnerships across the organization, and emphasize successes to drive a strategic mindset.

Where to begin: Defining a data strategy with success in mind

A data strategy provides an organization with direction. CDOs can use the data strategy to organize disparate activities, consolidate siloed data, and orient the organization toward a cohesive and unified goal. The aim is to set the stage for treating data as an asset, resulting in improved decision-making, enhanced user insights, and greater mission effectiveness.

Successful data strategies come in many shapes and sizes, tailored to each organization's strengths and weaknesses.



TAILORING A DATA STRATEGY TO AN ORGANIZATION'S UNIQUE NEEDS

Every organization is different; there is no definitive checklist for a data strategy. Successful data strategies come in many shapes and sizes, tailored to each organization's strengths and weaknesses.

CDOs who are unsure of their organization's strengths and weaknesses benefit from an assessment of their data maturity. Assessments are intended to provide a pulse check that CDOs can use to prioritize goals and initiatives within the data strategy to meet the organization's unique needs. With this understanding of strengths and weaknesses, CDOs can tailor their strategy to build upon organizational data opportunities while being cognizant of limitations.

The aims of an organization's data strategy should align with the overall mission and goals. For instance, the US Navy CIO's data strategy emphasizes data analytics and data management to enhance combat capabilities.[3] Similarly, the Department of Health and Human Services' data strategy focuses on consolidating data repositories to create a shareable data environment for all relevant stakeholders.[4]

IT'S ALL ABOUT THE PEOPLE

To be effective, a data strategy should also consider the human side: owners, stakeholders, analysts, and other users. Organizations that encourage staff to think about information and data as a strategic asset can extract more value from their systems.

For example, New York City's first chief analytics officer, Michael Flowers, addressed several complex problems through a data strategy that emphasized engagement of all data owners across the local government. "Our big insight was that data trapped in individual agencies should be liberated and used as an enterprise asset," he said.[5] Flowers' efforts led to the development of New York City's data integration platform, which now allows different parts of the local government to share data with each other to collaborate and solve problems.[6]

Gaining buy-in across the organization is instrumental in developing a successful data strategy, as is understanding all relevant organizational needs. The CDO should engage all parts of the organization from day one. Without input from key stakeholders, the CDO may fail to incorporate critical organizational considerations into the data strategy.

PLANNING FOR THE FUTURE

Nothing is stationary. CDOs should recognize that not only will their organization change, but so will various industry tools and technologies, as well as broader government policies and practices. It is imperative to plan and establish a data strategy that accounts for future changes. A flexible data

strategy can open up the ability for the organization to continue to use data as an asset for the long term.

As an example, the City of San Francisco addresses this challenge by regularly revisiting its data strategy. The City reviews and revises its data strategy each year—including refining its mission, vision, and approach—all while adhering to a set list of core goals. This periodic review of its data strategy keeps the City's approach to data use up to date while sustaining accountability for pursuing the City's overall strategy. [7]

How to implement a data strategy: Turning a document into a movement

Implementing a data strategy is a daunting task. One common difficulty is that many organizations are hesitant to change legacy IT operations—especially for government, whose budgeting process can make even small changes difficult to implement. However, difficult does not equate to impossible. CDOs can nudge their organization toward alignment with the data strategy's principles and goals.

TRANSFORMING STRATEGY INTO ACTION

Even the most brilliant strategy will not improve an organization's use of its data assets if it sits on a shelf. Converting a data strategy from a piece of paper into a state of mind can be messy, and CDOs should be realistic about the pace of change, especially early on. The most effective approach in the face of organizational inertia can be to set realistic expectations and identify opportunities to show value early on.

Once the data strategy is developed, CDOs should identify and list key business issues that the data strategy is designed to address or solve. For example, will the data strategy enable the organization to meet upcoming regulatory or legislative deadlines? Are there existing modernization efforts underway that require a data conversion? Is there a particular weakness from the assessment that can be addressed by implementing a data governance council? CDOs can develop a list of projects by identifying specific ways the data strategy can address these issues.

It is important to prioritize issues that will add the most value to the organization. To establish the data strategy's credibility and utility, it's helpful to start with high-visibility projects that draw on key components of the data strategy and that support the CDO's own key goals.

All parties should have skin in the game; this way, once the solution is deployed, everyone can declare victory.

What defines a good opportunity will be different for each organization. Finding the right project requires the CDO to have a clear understanding of the organization's wants and needs. For example, while developing the US Air Force's data strategy, the CDO identified manpower shortages as a critical issue. The CDO prioritized this limitation early on in the implementation of the data strategy and developed a proof of concept to address it.[8]

TURNING ACTION INTO VICTORY

A CDO can improve the chances of a project's success by developing partnerships across the organization. The best partnerships are those that are mutually beneficial, where all parties are invested in the effort's outcome. All parties should have skin in the game; this way, once the solution is deployed, everyone can declare victory.

An effective partnership can be maintained by simple, frequent, prioritized, and actionable communication. Simple and frequent communication keeps all parties informed about progress and minimizes negative impressions from minor setbacks. Delivering prioritized and actionable information enables all parties to act efficiently, and fosters a shared sense of ownership.

One effective partnership model can be for the CDO to share resources with partners across the organization. For example, a member of the CDO's team could work with a partner for a limited time to implement a specific project. This could benefit both teams: The project team gets an additional resource, and the CDO can be confident that the work aligns with the overall strategy. If a team member cannot be spared, the CDO can provide the project team with tools, subject-matter expertise, or other assets. Not only does this increase the odds that the project will meet its objectives, but it also affords the CDO greater control over the project's alignment with the data strategy.

An example of such a partnership is the US Department of Transportation's (DOT's) development of a national transit map in collaboration with several state and local transportation organizations.[9] In this effort, the DOT provided technical assistance to local transit agencies, who also benefited from having their data made publicly available.

TURNING A VICTORY INTO A STRATEGIC MINDSET

Success breeds success, and CDOs should capitalize on every victory. For example, enhanced data availability and a strategic analysis enabled the CDO of the US Department of Health and Human Services' Office of the Inspector General to detect hundreds of millions of dollars in fraud in 2017.[10] CDOs can point to victories like these when making a budgetary case for additional resources, technologies, or capabilities. Additionally, a CDO can use victories to foster excitement and buy-in from the organization.

Every victory counts. After one victory, find the next. Find another project, turn it into a success, and publicize it appropriately. Engage the energetic participants, and continue to work on the less-than-enthusiastic ones.

THE FUTURE STARTS NOW

The years to come will present many unique opportunities and challenges for data management. CDOs are uniquely positioned to guide organizations through the process of managing data and unlocking its value.

To be successful in this effort, a CDO must have a nuanced understanding of the organization's current data culture, resources, and opportunities for improvement. Through this understanding, CDOs can develop and implement an actionable data strategy to achieve the desired future state.

Understanding how to create and tailor a data strategy will be a critical skill for CDOs. By carefully selecting and implementing a data strategy and capitalizing on victories, CDOs can position their organizations for success in making use of data as a valuable strategic asset.

Endnotes

1. Executive Office of the President of the United States, "Leveraging data as a strategic asset," accessed February 5, 2019.
2. NewVantage Partners, "Big data business impact: Achieving business results through innovation and disruption," 2017.
3. Department of the Navy, United States of America, "Strategy for data and analytics optimization," September 15, 2017.
4. Office of the Assistant Secretary for Planning and Evaluation, US Department of Health & Human Services, "Improving data for decision making: HHS data collection strategies for a transformed health system," December 21, 2011.
5. William D. Eggers, *Delivering on Digital: The Innovators and Technologies That Are Transforming Government* (Deloitte University Press, 2016).
6. Ibid.
7. Edwin M. Lee and Joy Bonaguro, "Data in San Francisco: Meeting supply, spurring demand," City and County of San Francisco, July 31, 2015.
8. Rusty Frank, "AF chief data officer: Data is the future of the force," US Air Force, February 23, 2018.
9. Ben Miller, "US DOT calls on state, local agencies to help build national transit map," *Government Technology*, April 20, 2016.
10. Frank Konkel, "How federal agents took down a $1.3 billion health care fraud scheme," Nextgov, July 20, 2017.

About the authors

Max Duhe is a consultant with Deloitte Consulting LLP. He is based in Arlington, Va.

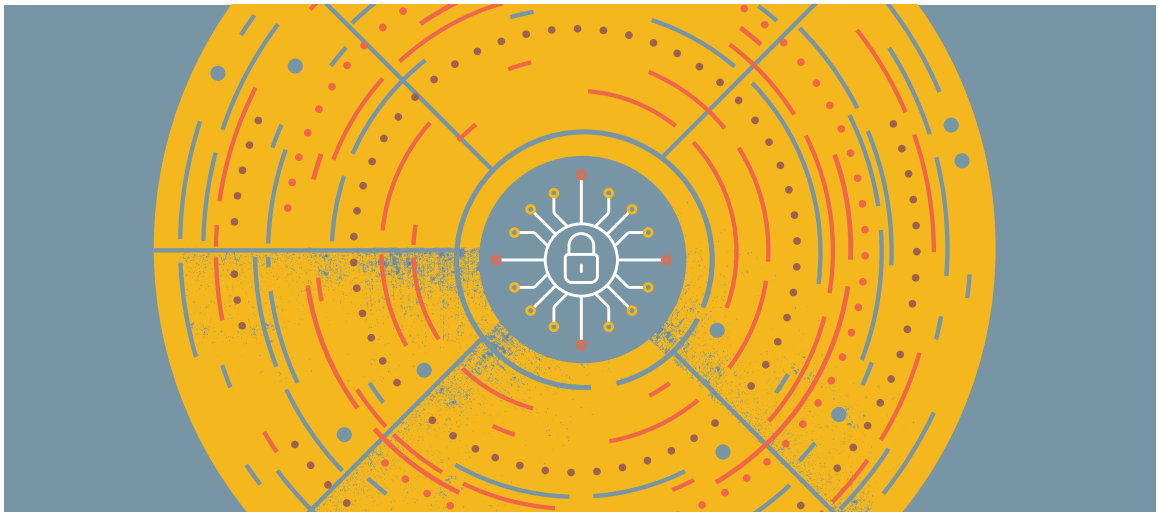
Matt Gracie is a managing director with Deloitte Consulting LP's strategy and analytics team. He is based in Arlington, Va.

Chris Maroon is a senior consultant with Deloitte Consulting LLP. He is based in Arlington, Va.

Tess Webre is a senior consultant with Deloitte Consulting LLP. She is based in Rosslyn, Va.

Data tokenization for government: Enabling data-sharing without compromising privacy

Tab Warlitner, John O'Leary, and Sushumna Agarwal



WE'VE ALL HEARD THE stories: If only information housed in one part of government had been available to another, tragedy might have been averted. From the 9/11 terrorist attacks to fatal failures of child protective services, we are often left to wonder: What if government's left hand knew everything it had in its right hand?

That's a tough ideal to attain for many government agencies, both in the United States and around the world. Today, much of the information held in government programs is isolated in siloed databases, limiting the ability to mine the data for insights. Attempts to share this data through interagency agreements tend to be clunky at best and nightmarish at worst, with lawyers from multiple agencies often disagreeing over the meaning of obscure privacy provisions written by disparate legislative bodies. No fun at all.

This isn't because agencies are being obstructionist. Rather, they're acting with the best of intentions: to protect privacy. Most government programs—such as Supplemental Nutrition Assistance Program (SNAP), Medicare, Unemployment Insurance, and others—have privacy protections baked into their enabling legislation. Limiting data-sharing among agencies is one way to safeguard citizens' sensitive data against exposure or misuse. The fewer people have access to the data, after all, the less likely it is to be abused.

The flip side, though, is that keeping the data separate can compromise agencies' ability to extract insights from that data. Whether one is applying modern data analytics techniques or just eyeballing the numbers, it's usually best to work with a complete view of the data, or at least the most complete view available. That can be hard when rules governing data-sharing prevent agencies from combining their individual data points into a complete picture.

What if data could be shared across agencies, *without* compromising privacy, in a way that could enable the sorts of insights now possible through data analytics?

That's the promise—or the potential—of data tokenization.

How data tokenization works

Data tokenization replaces sensitive data with substitute characters in a manner similar to data masking or redaction. Unlike with the latter two approaches, however, the sender of tokenized data retains a file that matches the real and tokenized data. This "token," or key file, does two things. First, it makes data tokenization reversible, so that any analysis conducted on the tokenized data can be reconstituted by the original agencies—with additional insights gleaned from other data sources. Second, it makes the tokenized data that leaves the agency virtually worthless to hackers, since it is devoid of identifiable information.

A simplified example can help illustrate how tokenized data can allow for personalized insights without compromising privacy. Imagine that you are a child support agency with the following information about an individual:

Name: Marvin Beals
Date of birth: September 20, 1965
Street address: 23 Airway Drive
City, state, zip: Lewiston, Idaho, 83501
Gender: Male
Highest education level: Four-year college

You might tokenize this data in a way that keeps certain elements "real" (gender and education level, for example), broadens others (such as by tokenizing the day and month of birth but keeping the real year, or tokenizing the street address but keeping the actual city and state), and fully tokenizes still other elements (such as the individual's name). The result might look something like this:

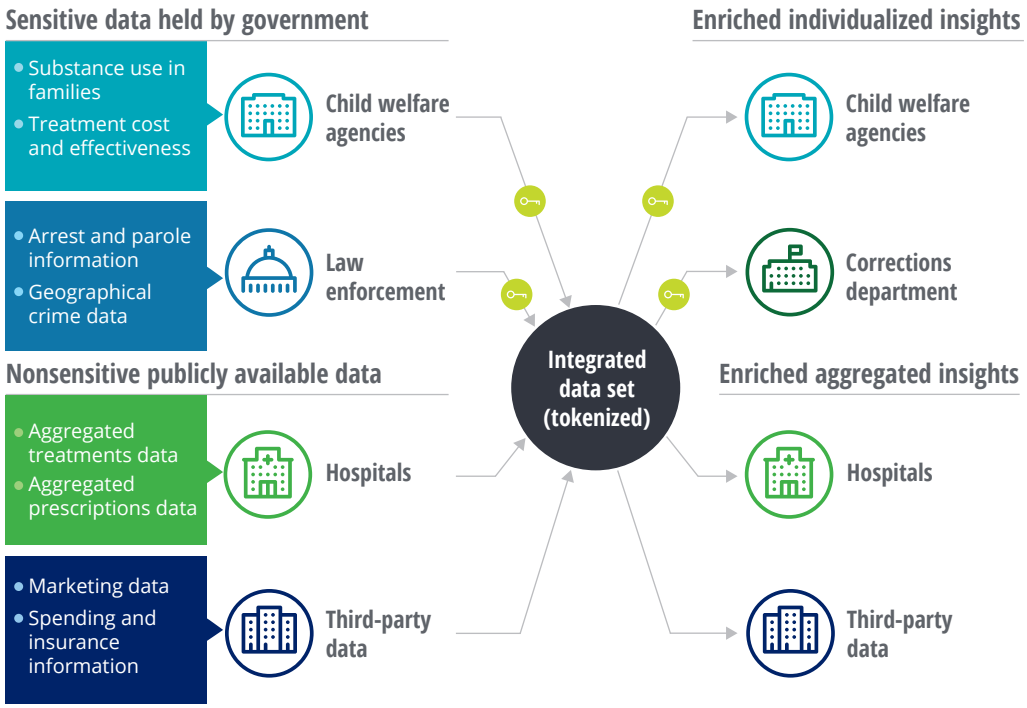
Name: Joe Proust
Date of birth: May 1, 1965
Street address: 4 Linden Street
City, state, zip: Lewiston, Idaho, 83501
Gender: Male
Highest education level: Four-year college

You could readily share this tokenized data with a third party, as it isn't personally identifiable. But you could also combine this tokenized data with, for example, bank data that can predict what a 54-year-old male living in that zip code is likely to earn, how likely he is to repay loans, and so forth. Or you could combine it with similarly tokenized data from a public assistance agency to learn how likely a male of that age in that geographical area is to be on public assistance. After analysis, you (and you alone!) could reverse the tokenization process to estimate—with much greater accuracy—how likely Marvin is to be able to pay his child support. Going deeper, you could work with other government agencies to tokenize some of the other personally identifiable information such as yearly income, social security number etc. in the same way, allowing you to connect the data more precisely with additional data.

Data tokenization's most powerful application is likely this mingling of tokenized government data with other data sources to generate powerful insights—securely and with little risk to privacy (figure 1). Apart from the ability to deidentify structured data, tokenization can even be used to deidentify and share unstructured data. As governments increasingly use such data, tokenization offers many new use cases for sharing data that resides in emails, images, text files, and other such files.

FIGURE 1

Data tokenization can allow data to be shared across both private and public sector organizations without compromising privacy



Note: Combining data from different sources, as illustrated in the figure above, requires either a unique identifier common to all data sets to tie the data together, or the use of an identical tokenization process across all the data sets.
Source: Deloitte analysis.

Data tokenization already helps some companies keep financial data safe

Data tokenization is already considered a proven tool by many. It is widely used in the financial services industry, particularly for credit card processing. One research firm estimates that the data tokenization market will grow from US$983 million in 2018 to US$2.6 billion by 2023, representing a compound annual growth rate of 22 percent.[1]

It's not hard to understand why data tokenization appeals to those who deal with financial information. Online businesses, for instance, want to store payment card information to analyze customer purchasing patterns, develop marketing strategies, and for other purposes. To meet the Payment Card Industry Data Security Standard (PCI DSS) for storing this information securely, a company needs to put it on a system with strong data protection. This, however, can be expensive—especially if the company must maintain multiple systems to hold all the information it collects.

Storing the data in tokenized form can allow companies to meet the PCI DSS requirements at a lower cost compared to data encryption.[2] (See the sidebar "The difference between encryption and tokenization" for a comparison of the two methods.)

Instead of saving the real card data, businesses send it to a tokenization server that replaces the

actual card data with a tokenized version, saving the key file to a secure data vault. A company can then use the tokenized card information for a variety of purposes without needing to protect it to PCI DSS standards. All that needs this level of protection is the data vault containing the key file, which would be less expensive than working to secure multiple systems housing copies of real credit card numbers.[3]

How data tokenization could help government address the opioid crisis

Policymakers often describe the US opioid crisis as an "ecosystem" challenge because it involves so many disparate players: doctors, hospitals, insurers, law enforcement, treatment centers, and more. As a result of this proliferation of players, information that could help tackle the problem—much of it of a sensitive nature—is held in many different places.

Government health data is difficult to share—as it should be. Various agencies house large amounts of sensitive data, including both personally identifiable information (PII) and personal health information (PHI). Given government's significant role in public health through programs such as Medicare, Medicaid, and the Affordable Care Act, US government agencies must expend considerable resources in adhering to Health Insurance Portability and Accountability Act (HIPAA) regulations. HIPAA alone specifies 18 different

THE DIFFERENCE BETWEEN ENCRYPTION AND TOKENIZATION[4]

Encryption is the process of transforming sensitive data into an unreadable form using an algorithm. A password, also known as a "key," is generally needed to decrypt the data. Encryption is useful when sensitive information needs to be exchanged securely—although both parties will need to hold an encryption key (either a symmetric or asymmetric key). However, encrypted data can be reversed into its original form if the key is compromised. Many attackers resort to what is known as a dictionary attack—trying millions of likely passwords—in attempts to hack encrypted data.

Tokenization, on the other hand, does not use a traditional algorithm to drive the masking process. Rather, it replaces sensitive data with random data, maintaining a one-to-one mapping between each sensitive data point and its corresponding random data point. This mapping is stored securely in a "token store," and only individuals with access to the token store can reverse the tokenization process.

Even after encryption, the sensitive data is, in its essence, still there, vulnerable to sophisticated cybercriminals who can crack the encryption algorithm. But even if bad actors were to steal tokenized data, it would be worthless to them without the corresponding token information. This is because tokenized data does not, in itself, contain any useful information, since the random replacement data points have no inherent value. And because tokenized data cannot be understood without the key, the tokenization process allows the original data to be analyzed while completely preserving the anonymity of the sensitive aspects of that information.

Another feature of tokenization not available with encryption is the ability to "forget." It is possible to delete the token so the true values can never be reidentified, which may be useful when individuals ask for their info to be erased and "forgotten" as under the European Union privacy regulations.

Other approaches to merging data without compromising security are under development. For example, it may be possible for a number of government agencies to use the same "public" key to tokenize data while making detokenization possible only with a "private" key held by a single high-level privacy office. This would do away with the need to use a common identifier across data sets.

types of PHI, including social security numbers, names, addresses, mental and physical health treatment history, and more.[5]

However, the US Department of Health and Human Services (HHS) guidelines for HIPAA note that these restrictions do not apply to *de-identified* health information: "There are no restrictions on the use or disclosure of de-identified health information. De-identified health information neither identifies nor provides a reasonable basis to identify an individual."[6] By tokenizing data, states may be able to share opioid-related data outside the agency or organization that collected it and combine it with other external data—either from other public agencies or third-party data sets—to gain mission-critical insights.

At no point in the entire cycle should any other agency or third party be able to see real data.

Data tokenization might enable states to bring together opioid-related data from various government sources—including health care, child welfare, and law enforcement agencies—and combine this data with publicly available data related to the social determinants of health and health behaviors. The goal would be to gain insights into the causes and remedies of opioid abuse disorder. By tokenizing the data's personal information, including PII and PHI, government agencies can share sensitive but critical data on opioid use and abuse without compromising privacy. Moreover, only the government agency that owns the sensitive data in the first place would be able to reidentify (detokenize) that data, assuming that the matching key file never leaves that agency's secure control. At no point in the entire cycle should any other agency or third party be able to see real data.

Why not simply use completely anonymized data to investigate sensitive topics like opioid use? One reason is that tokenized data, but not anonymized data, can provide insights at the individual level as well as at the aggregate level—but only to those who have access to the key file. For example, tokenization can turn the real Jane Jones into "Sally Smith," allowing an agency to collect additional data about "Sally." If we know that "Sally Smith" is a 45-year-old female with diabetes from a certain zip code, the agency can merge that with information from hospital records about the likelihood of middle-aged females requiring readmission, or about the likelihood of a person failing to follow his or her medication regimen. An analysis of this combined information can allow the agency to come up with a predictive score—and the agency can then detokenize "Sally Smith" to deliver customized insights for the very real Jane Jones. This ability to gain individual-level insights could be helpful both in delivering targeted services and in reducing improperly awarded benefits through fraud and abuse.

The mechanics of securely combining different data sets after tokenization can be complicated, but the potential benefits are immense.

The opioid crisis is not government's only ecosystem challenge. As we've seen in the private sector, in industries from retail to auto insurance, more information generally means better predictions. (That's why your auto insurer wants to know about your driving patterns, and why they might offer a discount if you ride around with an app that captures telemetrics.) Many companies' websites require an opt-in agreement to allow them to use an individual's data. This approach is more challenging in government, however, due to the sensitive nature of data that is collected and the fact that citizens must be served whether they opt in or not. Where circumstances make it impractical to obtain consent, data tokenization can make it possible for governments to use data in ways other than the originally intended use without violating individuals' privacy.

PARTIAL TOKENIZATION CAN ENABLE COMPLETE INSIGHTS

Tokenizing data can make more data available for analysis—but what if the data points that are swapped out for random information are precisely what you're interested in analyzing? Some attributes, such as age or address, that may be valuable to the analysis could be lost if the tokenization process doesn't take this into account. You certainly wouldn't want to analyze data for insights on the opioid epidemic with anything other than real health information.

"Partial tokenization" offers a way around this problem. With partial tokenization, government agencies can obscure the *personally identifiable* attributes of the data without losing the ability to detect critical insights. This is done by tokenizing only those parts of the data that can pinpoint an individual. For example, a tokenized date of birth might tokenize the day and month but leave the year intact—which provides sufficient information to analyze patterns of behavior that may vary with age.

Tokenization can be combined with redaction or truncation. For example, fields that can identify individuals (such as name or social security number) can be completely tokenized, fields like address and date of birth partially tokenized, and other fields such as health information untokenized. Such techniques can help government detect fraud, identify patterns of use, and provide predictive insights to better serve constituents.

Taking data tokenization beyond data-sharing

Beyond allowing agencies to share data without compromising privacy, data tokenization can help governments in other ways as well. Three potential uses include developing and testing new software, supporting user training and demos, and securing archived data.

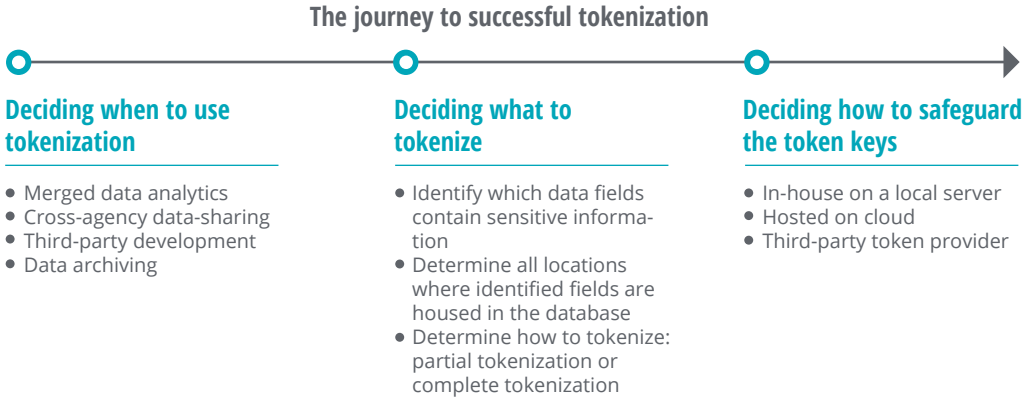
Developing and testing new software. Developers need data to build and rigorously test new applications. Government frequently relies on third-party vendors to build such systems, because it can be prohibitively expensive to require all development and testing be done in-house on government premises. But what about systems such as those relating to Unemployment Insurance or Medicaid, which contain a great deal of PII and PHI? By using format-preserving tokens to tokenize actual data while maintaining the original format—where a tokenized birth date, for example, looks like 03/11/1960 and not XY987ABC—third-party developers can work with data that "feels" real to the system, reliably mimicking the actual data without requiring all the security that would be needed if actual data was being shared. Some US states, including Colorado, have used data tokenization in this manner. Data tokenization apps are often a cost-effective way to give developers tokenized data.

User training and demos. When new employees join government agencies, they often undergo a probationary period during which they need to be trained on various applications and evaluated on their performance.[7] During this time, government agencies can create training environments using tokenized data, enabling new hires to work and interact with data that looks real but does not compromise security.

Securing archived data. Data tokenization can also allow governments to archive sensitive data offsite. For US government agencies, securing sensitive data not in active use in a production

FIGURE 2

How can data tokenization work for government?



Source: Deloitte analysis.

environment has been a challenge due to costs and competing priorities.[8] A 2018 report by the US Office of Management and Budget found that, while 73 percent of US agencies have prioritized and secured their data in transit, less than 16 percent of them were able to secure their data at rest[9]—an alarming statistic, considering that governments are often a prime target for cyberattacks and have been experiencing increasingly complex and sophisticated data breaches in the last few years.[10]

How can governments get started?

Figure 2 depicts three important decisions government departments should carefully consider to successfully implement data tokenization. First, when should they use tokenization? Second, what data should they tokenize? And third, how will they safeguard the token keys? Once the decision to use tokenization has been made, there is still much important work to be done. The team tokenizing the data must work closely with data experts to ensure that tokenization is done in a way that allows the end users to meet their intended objectives yet ensures privacy.

Public officials are often frustrated by their lack of ability to share data across organizational boundaries, even in situations where sharing the data would have a clear benefit. This lack of cross-agency sharing can mean that agencies don't make the most of data analytics that could improve public health, limit fraud, and make better decisions. Data tokenization can be one way for government agencies to share information without compromising privacy. Though it is no magic bullet, the better insights that can come from sharing tokenized data can, in many circumstances, help governments achieve better outcomes.

Endnotes

11. Research and Markets, *Tokenization market by component, application area (payment security, application area, and compliance management), tokenization technique (API-based and gateway-based), deployment mode, organization size, vertical, region—global forecast to 2023,* December 2018.
12. John Zyskowski, "Tokenization: A privacy solution worth watching," *Federal Computer Week*, March 10, 2011.
13. Ibid.
14. Interviews with Deloitte subject matter experts.
15. United States Department of Health and Human Services, "Summary of the HIPAA privacy rule," May 2003; Martin Sizemore, "Is tokenization the solution for protected healthcare information (PHI)?," Perficient, February 22, 2011.
16. United States Department of Health and Human Services, "Summary of the HIPAA privacy rule."
17. FEDweek, "Probationary period," accessed October 28, 2019.
18. The White House, *Federal cybersecurity risk determination report and action plan*, May 2018.
19. Ibid.
20. US Government Accountability Office, "Cybersecurity challenges facing the nation—high risk issue," accessed August 27, 2019.

About the authors

Tab Warlitner, a principal with Deloitte Consulting LLP and a member of its board of directors, is the lead client service partner supporting the state of New York and New York City. He is based in Arlington, VA.

John O'Leary, a senior manager with Deloitte Services LP, is the state and local government research leader for the Deloitte Center for Government Insights. He is based in Boston.

Sushumna Agarwal is a senior analyst with the Deloitte Center for Government Insights, Deloitte Services LP. She is based in Mumbai.

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

Deloitte Insights contributors

Editorial: Junko Kaji, Abrar Khan, Nairita Gangopadhyay, and Preetha Devan

Creative: Anoop K R

Promotion: Alex Kawecki

Cover artwork: Lucie Rice

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

About Deloitte