Inter-American Development Bank

Innovations in Public Service Delivery

Issue No. 4

Predictive Analytics: Driving Improvements Using Data

Stephen Goldsmith, Susan Crawford, and Benjamin Weinryb Grohsgal

Institutions for Development Sector

Institutional Capacity of the State Division

DISCUSSION PAPER
IDB-DP-440

July 2016

Innovations in Public Service Delivery

Issue No. 4

Predictive Analytics: Driving Improvements Using Data

Stephen Goldsmith, Susan Crawford, and Benjamin Weinryb Grohsgal

July 2016

IDB
Inter-American Development Bank

Innovations in Public Service Delivery
Issue No 04
Predictive Analytics:
Driving Improvements using Data

Authors: Stephen Goldsmith, Susan Crawford,
and Benjamin Weinryb Grohsgal

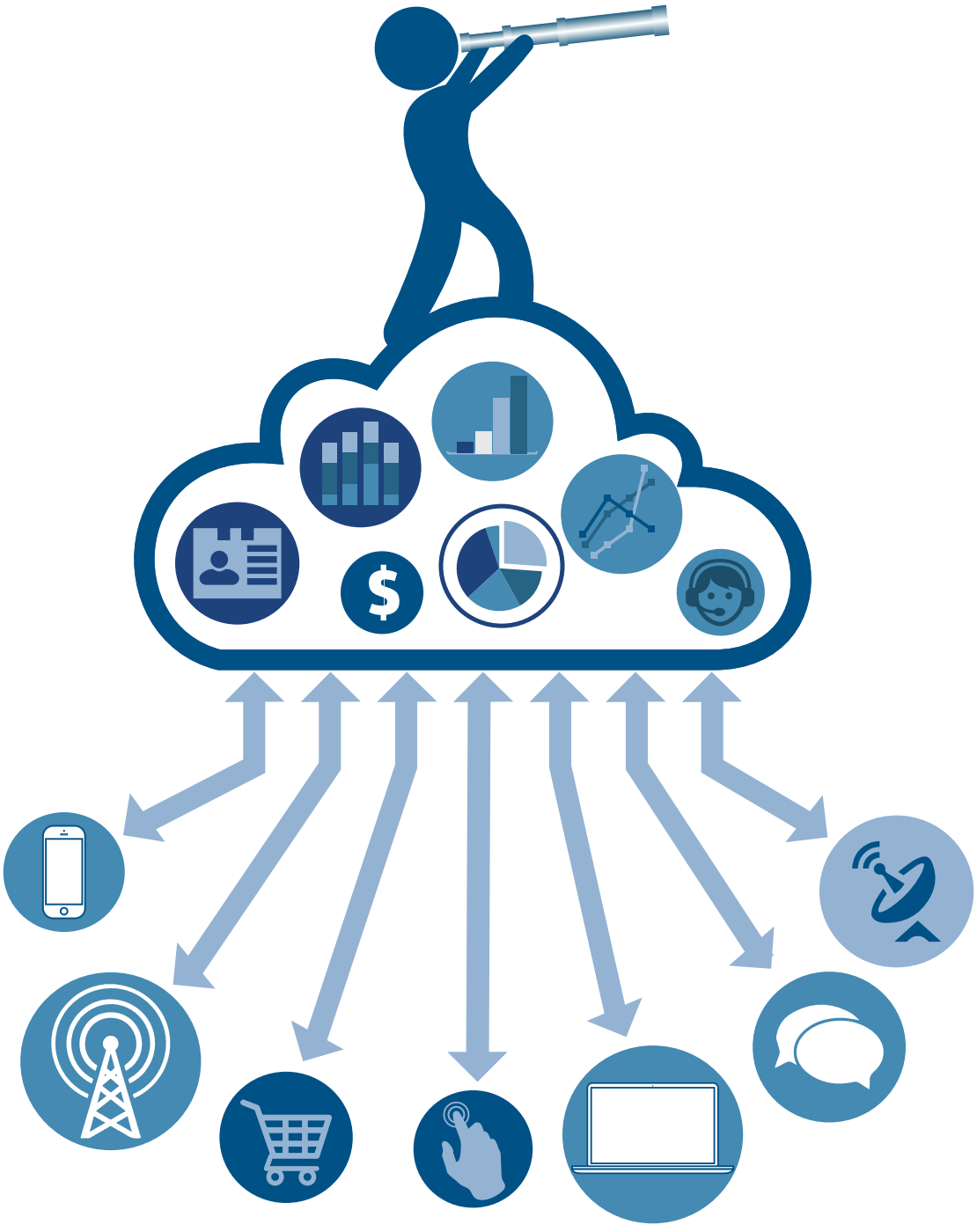

TABLE OF CONTENTS

Complexity is the defining feature of an increasingly interconnected world. It underlies the greatest challenges facing the global community in the twenty-first century, including climate change, the stability of markets, the availability of energy and resources, poverty, and conflict.[1] It is also the new reality in which governments must operate: more demanding citizens with greater awareness of their rights, better access to information through technology, and expectations of high-quality services; growing budgetary restrictions at a time when the demand for public services is on the rise; and a changing demographic composition, with important implications for service provision. In this context, government officials are pressed to devise innovative solutions to address these challenges.

This complexity is reflected in the massive volume of data produced by citizens and agencies that governments must harness using predictive analytics to better understand citizens' needs and allocate public resources efficiently.

What is predictive analytics? In a nutshell, it is the use of historical data to identify trends, which can be used to anticipate future needs. Some Latin American countries are already using this innovative practice to support more effective decision making through real-time situational awareness: Montevideo is following in the footsteps of Los Angeles, Santa Cruz, and Atlanta, using basic crime data (type, time, and place) to analyze patterns of criminal behavior and help the city's police force improve neighborhood patrolling schedules based on the identification of critical points of crime.[2] The Government of Mexico has used data on electricity consumption to get near real-time forecasts (nowcast) of quarterly GDP, addressing the need for up-to-date estimates of economic activity to formulate and assess policies;[3] and Colombia is using data from Google web searches to infer economic activity at the sectoral level, providing alternative indicators to traditional statistics in a much timelier manner.[4]

Despite these success stories, predictive analytics has had a slow uptake in the Latin America and Caribbean (LAC) public sector. Putting data to work for governments in the region requires a profound process and cultural transformation within agencies. Although many LAC governments are modernizing public management, institutional fragmentation and complicated administrative procedures are still a major hindrance to effective, efficient, and open government. For years, the Inter-American Development Bank (IDB) has been working with LAC governments to strengthen their management capacities and improve the quality of the services they deliver. This support includes technical and financial assistance and the generation and exchange of policy-relevant knowledge to better understand the drivers of institutional change and government modernization that will improve public service delivery and foster public sector innovation.

As part of these efforts, the IDB has created the "Innovations in Public Service Delivery" discussion paper series. Previous cases and lessons learned documented,

[1] http://www3.weforum.org/docs/WEF_GAC_PerspectivesHyperconnectedWorld_ExecutiveSummary_2013.pdf

[2] http://www.elpais.com.uy/informacion/predpol-software-policia-predecir-delitos.html

[3] Nowcasting Mexican GDP Based on Electricity Consumption (2015).

[4] http://cea.cepal.org/sites/default/files/lacbigdatansopapernov10finaldraft.pdf p.25

analyzed, and shared within the series include "Can 311 Call Centers Improve Service Delivery? Lessons from New York and Chicago;" "Los servicios en línea como derecho ciudadano: El caso de España;" and "Descomplicar para avançar: O caso Minas Fácil." The present issue examines how government officials can use properly analyzed historical data to look for patterns and trends to reorganize the way they deliver services, anticipate future events, and often even prevent potential problems. It outlines examples of the use of predictive analytics and their implications for future government action, beginning with two detailed case studies: Chicago's current predictive analytics pilots in the service of enhancing operational outcomes within particular city departments, and the process followed by the state of Indiana in launching a Management and Performance Hub that will permit analytics to be used across state databases. The paper provides a catalog of uses of predictive analytics by governments around the globe, organized by policy area, which can provide important lessons to LAC countries seeking to adopt similar solutions. Key areas of public policy, such as health, public safety, social services, taxation, and others are presented as potential beneficiaries of the use of predictive analytics. Finally, the authors present a summarized roadmap to start implementing predictive analytics in public administration.

Professor Stephen Goldsmith, one of the three authors of the paper, is the Daniel Paul Professor of the Practice of Government and the Director of the Innovations in American Government Program at Harvard University's Kennedy School of Government. He currently directs Data-Smart City Solutions, a project to highlight local government efforts to use new technologies that connect breakthroughs in the use of big data analytics with community input to reshape the relationship between government and citizen. Professor Susan Crawford is the John A. Reilly Visiting Professor in Intellectual Property at Harvard Law School. She served as Special Assistant to the President for Science, Technology, and Innovation Policy in 2009, co-led the FCC transition team between the Bush and Obama administrations, and was a member of Mayor Michael Bloomberg's Advisory Council on Technology and Innovation. Ben Weinryb Grohsgal is a software engineer at BuzzFeed. He previously worked as a senior consultant at Booz Allen Hamilton, the Data Smart City Solutions Initiative, the Mayor's Office of Civic Innovation in the City of San Francisco, and the Office of Economic Development in Littleton, Colorado.

This document is product of the collaboration between the IDB and the Ash Center for Democratic Governance and Innovation at the Kennedy School, under the coordination of Prof. Goldsmith. This work was important in laying the foundations for the discussion paper series coordinated by Pedro Farias, Principal Modernization of the State Specialist in the IDB's Institutional Capacity of the State (ICS) Division.

The authors would like to acknowledge the technical and financial support of the IDB's Institutional Capacity Strengthening Fund (ICSF), with funding from the Government of the People's Republic of China and the Republic of Korea through the Korean Capacity Building Fund, which made this publication possible. They are grateful for the valuable comments made by Pablo Valenti and Alejandro Pareja during the review process, and they wish to thank Ana Catalina García de Alba and Florencia Cabral for their contributions, which helped bring this document to fruition.

Miguel Porrúa
e-Government Lead Specialist
Institutional Capacity of the State Division
Institutions for Development Sector
Inter-American Development Bank

Although governments have reams of historic and current electronic data at their disposal, the insights and patterns that systematic inspection of that data might reveal are often hidden from civil servants. For the most part, government information technology systems are developed as a means to process isolated transactions. Data stored in incompatible formats and on outdated equipment was used to service those siloed transactions—allowing the government to keep doing what it was doing—but could not be linked to other data. As a result, insights drawn from the information generated by these disparate systems were simply not available to government; in an important sense, governments did not know what they knew.

Today, hardware is getting cheaper, data processing time is rapidly diminishing, data mining techniques can circumvent legacy system interoperability issues, and increasingly professionally-managed governments are hiring personnel comfortable with using data to manage performance and predict trends. Leaders are emerging who have the vision and forcefulness of personality to require agencies to work together to harness data in the service of better outcomes for citizens. As a result, a new era of predictive analytics is dawning across the world. This paper provides a set of tools and frameworks that will be useful to public sector employees considering the use of predictive analytics to improve the effectiveness of government operations.

What is Predictive Analytics?

Predictive analytics is the use of historical data to identify patterns and trends that can help anticipate the future. Businesses use these practices all the time: for example, Amazon predicts which books its customers want to read next and makes it easy to buy them; CVS, a large U.S. pharmacy chain, knows when its customers need to replenish their multivitamins before they do, and sends coupons to encourage repurchasing based on past sales data; and Walmart stores stock extra bottled water, duct tape, and snacks in preparation for major weather events. The Obama campaign used predictive analytics to identify swing voters in need of persuasion. The U.S. Department of Homeland Security is developing software that can predict which individuals are security risks based on their behavior while standing in the security line, including data such as fidgeting, sweating, eye movement, and heart rates.

Predicting the future is not a new trend. For decades, television weathermen have forecasted whether it will be sunny or raining, and pollsters have predicted who will win an election. What is new today is the sheer volume and ubiquity of data that can be used to make predictions. It is now possible to harness the computing power needed to analyze this massive amount of data and to combine and compare complex datasets—all in real-time and using increasingly smaller devices.

Just as commercial online companies such as Google and Facebook have been doing for some time, governments can now use predictive analytics to better target their activities. The idea is simple: using historic data about characteristics of a given subject (e.g., a transport network, a restaurant, obese children, buildings that might experience fires), statisticians isolate those characteristics that appear to correlate strongly with particular outcomes. Alternatively, they look for anomalies or outliers in data about these characteristics. They then rigorously test—first through use of data alone and later in the field—whether prioritizing activities concerning those characteristics in some way (changing the timing of trains, classifying restaurants differently, providing better exercise facilities, inspecting particular buildings first for fire risks) would lead to an improvement over the status quo with respect to a desired outcome. Once these predictions are validated, operational and policy engines result that prevent the identified undesirable outcomes. Then the process is repeated. Not all problems, however, fit this method; sometimes all that is needed to solve a problem is visibility of data across different databases rather than a prediction. Problems that involve a need to prioritize a list of government tasks are the most susceptible to predictive analytics.

Businesses and governments have similar problems to solve that can be aided by predictive analytics. Tasks that might have seemed impossible become possible, as illustrated in Table 1.

This description of predictive analytics is both under- and over-inclusive. It does not capture the dynamism of systems or the complexity of a world in which billions of pieces of data are collected every day. Today's society lives in a hyper-connected environment; thus predictions of later outcomes—and the enforcement mechanisms (or "treatments") that can be proven through testing to address them—may have to change on the fly in reaction to small changes in the world that end up having nonlinear impacts. At the same time, basic shortfalls in government resources make even the simplest of predictive analytics experiments extraordinarily difficult to implement.

Table 1. Predictive Analytics in Public and Private Sectors

Business application	Public sector application
A process involves a large number of similar decisions, but each decision requires considering a wide range of variables.	Inspections
The problem and the outcome involve a system and not just one agency (i.e., where treating the manifestation of the problem may miss much of the underlying cause).	Truancy (public safety and education)
There is abundant information in electronic data form available on which to base decisions and measure outcomes	Fraud detection
It is possible to insert a model calculation into the actual business process, either to automate decisions or to support human decision makers.	Welfare casework
Conditions of privacy, standardization, data security, and transparency have been explicitly considered.	Tax collection

It is important that predictive analytics be used as a tool to enhance the discretion and effectiveness of public employees, and not as a substitute for their intuition. As IBM puts it, the best professionals working with young people "make shrewd judgments about which of their charges is at risk," and "predictive analytics simply provide robust, empirical evidence to support their professional findings" (IBM, 2011a). At the same time, these techniques can be useful in spotting outliers that might escape the notice of even those professionals. The CGI Group (2013) has found that predictive models can help produce value by allowing workers to reach the right conclusion more quickly, cheaply, and effectively; in some cases, entire legacy processes can be automated with the aid of analytics. At bottom, however, predictive analytics should not replace local knowledge and expertise; instead, these techniques are best employed as adjuncts or complements to existing practices.

In the private sector, shareholders of companies are the beneficiaries of predictive analytics driving improved business practices and increased market share. When government uses these same techniques, citizens benefit.

Why Now?

Governments at all levels face daily demands for services that far outpace the resources available to produce them, requiring officials to produce better, cheaper, faster government. Predictive analytics will allow public sector employees to target resources to areas where they are most needed, improving the decision-making capacity of public employees while unlocking business process reengineering opportunities. Using data, governments can understand and take precautions against risk much more precisely by directing resources to areas where risk mitigation is most important. In a nutshell, predictive analytics allow governments to be more responsive to the needs of citizens in a time of limited government resources.

It is not technology that prevents us from achieving this new normal of responsive government. Instead, the problem is often outmoded government structures. Countries with advanced civil service systems have made major improvements in reducing corruption by requiring employees to operate inside a narrow set of rules limiting discretion and by auditing, inspecting, and second-guessing their compliance with these rules. This made sense in a world without predictive analytics; governments were unable to distinguish between the good exercise of discretion and its corrupt use, and sought to limit both.

Today's digital tools, however, provide governments with the opportunity to grant more discretion while at the same time tracking its exercise much more closely. The same tools that will help government find water leaks before they occur and identify taxpayers underpaying what they owe can also identify public employee outliers. Data about exactly where an employee is working during a day, how many infractions an inspector issues, to whom, when (or not), and which public officials or agencies use private service providers will likewise help in identifying employee outliers. We believe that governments cannot afford not to take advantage of these opportunities.

What Obstacles Need to be Overcome?

The building blocks of using data successfully to predict and solve problems require overcoming a number of non-technical obstacles. The first of these challenges involves a lack of clarity concerning the public values government seeks to serve by using data. The most exciting opportunities are focused on improving outcomes rather than outputs: improving public health rather than increasing the number of public hospitals; reducing homelessness rather than increasing the number of shelter beds; making the country safer rather than arresting more criminals. But defining the problem to be solved is a difficult step.

The second organizational challenge is the structure of government. Well-intentioned officials trying to use data to solve a systemic problem stretching across multiple departments and nongovernmental organizations will have problems if they operate in narrow activity-driven verticals.[1] Sharing data and implementing solutions often require government-wide actions.

When New York City data scientists discovered how to predict which buildings presented the highest risk of a deadly fire, the City's chief operating officer needed the authority to create multidisciplinary teams of fire, health, and building officials to leverage their collective know-how and enforcement authority to solve the problem. It turned out that slumlords and residents in those risky buildings more frequently responded to a fire official in uniform than to a building inspector, even when the latter had the true solution. Operationalizing the answer predicted by the data often requires new cross-agency structures to produce real value.

The third challenge is overcoming the status quo—whether management, culture, or perceived legal barriers among government agencies. Predictive analytics unlocks answers to problems when it operates on data across agencies, not just on data inside a single agency. Agencies, however, do not naturally cooperate with one another. Without a strong push from a top elected or appointed official, the power of analytics will never reach its potential. The examples that follow demonstrate that anyone leading a predictive effort needs to have the actual and apparent authority to create an appetite within his or her enterprise for the solutions that derive from analytics.[2]

---

[1] According to Unified Incident Command and Decision Support (UICDS, undated):

> In an emergency operation today—whether responding to an actual event or to indications and warnings coming from an intelligence agency—each person, team, and organization knows about their own information. Police knows police, fire knows fire, counter-terrorism knows counterterrorism. But, these response forces operate virtually in isolated information silos ... until each shares their information as they are able and enabled. This results in gaps, overlaps, and inconsistencies in who knows what, when, yielding isolated information and clouded decision making.

[2] According to LaVall et al. (2011):

> The leading obstacle to widespread analytics adoption is lack of understanding of how to use analytics to improve the business, according to almost four of 10 respondents.

Fourth, investments in people and technology will be necessary in order to drive future savings using predictive analytics. (Appendix A provides a step-by-step set of recommendations along these lines.) As a recent study observed: "Data are plentiful and typically easy to extract, but the resources (e.g., human) needed to transform the data into useful information are often scarce" (Van Barneveld, Arnold, and Campbell, 2012).

The fifth structural obstacle involves both acquiring usable data and overcoming the notion that the data being used needs to be perfect. Problems chosen early for solution, whether they are large or small, should relate to areas in which a reasonable amount of easily extracted data is available. By combining open data, transparency, and use of the best existing datasets, it is possible to achieve breakthroughs. Interested public officials need not mine every conceivably applicable dataset to find an answer. (Additionally, they will need to have in place the business intelligence and visualization tools necessary to translate their findings in a way that motivates and assists users. Appendix A provides more details.)

These five structural obstacles, taken together with inadequate hardware for processing data, turf battles among agencies unwilling to provide relevant data, inadequate skills in personnel, and destructive procurement and hiring requirements, have meant that governments around the world are still at the beginning of the story when it comes to the use of predictive analytics. Increasingly, however, public leaders are producing many strong examples of using these techniques effectively.

For example, Brazil's highest court, the Supremo Tribunal Federal (STF), carried out a reform that allowed the STF to drive performance through data by predicting future needs while simultaneously reallocating existing resources more effectively. Confronted with a major problem in resource planning, Almir Antonio da Costa, Head of Data Administration in the Secretary of Information Technology, focused on data quality and standardization as his first step. Before he started, there was neither unified reporting about the court's practices nor standardized data about its caseload. In order to produce the cultural change needed, he appointed an internal committee to standardize definitions and procedures. He then adopted business intelligence tools that could use that data to identify cases raising similar issues. As a result, the court can now make better decisions about its operations, using analytics to measure key indicators, and assess trends in the data that signal needs for future resources.

This paper outlines other examples of the use of predictive analytics and their implications for future government action, beginning with two detailed case studies: the City of Chicago's current pilots of predictive analytics in the service of enhancing operational outcomes within particular city departments, and the process followed by the State of Indiana in launching a Management and Performance Hub that will permit analytics to be used across state databases. Finally, the paper provides a catalog of uses of predictive analytics by governments around the globe, organized by policy area.

Chicago: Outcomes Driven by Predictive Analytics

In Chicago, Mayor Rahm Emanuel's leadership has been essential to the city's efforts to operationalize predictive analytics. As Brenna Berman, the city's Chief Information Officer, puts it: "He's been talking about being data-driven from day one, because that's how he makes his decisions—that's how he talks about everything. He always wants the numbers and he believes in predictive analytics." The mayor's determination, in turn, drives his staff: "We want to be the best at this," says Berman.

As of summer 2014, Chicago's deliberate and thoughtful planning for predictive analytics is beginning to bear fruit. The city's Department of Innovation and Technology (DoIT) is looking to expand predictive analytics pilot projects to more city departments in 2015. Every one of these pilots is aimed at a well-defined operational problem (e.g., How can one better target rat baiting?) and every experiment is subject to rigorous testing and validation. The results of all this steady work, city employees hope, will be measurably improved outcomes driven by predictive analytics.

These pilots, joined by maintenance of an ambitious open data portal, extensive work with city departments aimed at encouraging them to share additional data with DoIT, and the development, spurred by a US$1 million grant from Bloomberg Philanthropies, of an open-source SmartData platform that will be shareable with other cities, make for a long list of projects for DoIT. DoIT, however, is not alone in this; partnerships, formal and informal, with local universities, the local civic community, foundations in Chicago like MacArthur and the Chicago Community Trust, and many other people and groups, have both increased the city's capacity to take on difficult projects and challenged the city to constantly examine its use of data.

Spearheaded by the mayor, the City of Chicago sees itself as leading other cities around the world in the use of technology—solving the problems that big cities have and will continue to have as more people live in cities in the future. Chicago's experience in using predictive analytics to drive towards improved outcomes, in particular, is informative. The desired outcome for predictive analytics, Berman says, is to "make [it] the standard way of life" in city government operations. As Tom Schenk, Director of Analytics for the City of Chicago since late 2012, puts it, "Our projects are very focused on operational need. We won't take a project unless we can implement it on an operational basis."

To get there, Berman and her predecessor Brett Goldstein (the country's first municipal Chief Data Officer) needed to take some basic steps. First, Goldstein had the city invest in what Schenk calls "fantastic hardware"—machines that can process data extremely quickly. The city sorely needed this capacity when Goldstein arrived in the mayor's office; as a result of this investment, reports can be run in five percent of the time they used to take—saving untold numbers of hours of city employee time and increasing the city's ability to employ advanced analytics.

Then, Berman and her colleagues worked through business cases with city departments to ensure that they understood it would be worth it for the departments to migrate their data into DoIT's environment. One of DoIT's persuasive arguments in this regard was speed: according to Schenk, "people will migrate" when they understand they can save 95 percent of their processing time. Inquiring public employees will pursue answers to "what if" questions more frequently in a fast and responsive IT environment.

First Operational Pilot

DoIT took on its first operational pilot of predictive analytics: better-targeted rat baiting. This pilot was born of data mining across all of the city's 311 data—which includes 500 call types, from complaints about unclean restaurants to concerns about potholes. The team of engineers at DoIT noticed a relationship among 311 call types that correlated with rat infestation problems, and took that information to the Department of Streets and Sanitation (DSS).

Soon after, Schenk assembled a team to collaborate with Josie Cruz, DSS Director of Rodent Control. On hearing the proposal from Schenk, Cruz was hesitant to believe the pilot would lead to success. "At first, I found myself wondering if this was really possible, " says Cruz. "Our operations are 311 complaint-driven. Will using data like this really help get things done?"

To prove its effectiveness, Berman and Schenk presented DoIT's predictive model to Cruz. Based on thirty-one types of 311 calls that served as leading indicators for rat infestations, the system developed a map of the city that targeted specific areas for intervention. Overlaying these points with Cruz's schedules, Berman demonstrated that 80 percent of locations that DoIT identified were places that Cruz's team was already planning to go. According to Berman, "Then there was 20 percent that she had never been to, so she was completely unaware that there might have been an issue. That really intrigued her because we were aligned for a large majority, which gives you a lot of confidence. That 20 percent piqued her interest—and that really got us on the same page to then build a pilot and explore where this was coming from."

Berman and Charlie Williams, the Commissioner of the Department of Streets and Sanitation, arranged a pilot project with the assistance of computer scientists from Carnegie Mellon University. It was critical that the pilot be minimally invasive, and that it offer improvement without reinventing the wheel. DSS and DoIT were open to collaboration from the start and held multiple meetings to hash out the pilot's details, gaining a mutual understanding of the benefits their cooperation would produce for the city.

The only change to DSS operations was the pilot's minimally-invasive experimental design: baiters in a control group would receive work-order lists of trapping locations as usual, whereas experimental-group baiters would receive data-optimized lists. Fortunately, the model worked as intended: predictive analytics were not replacing DSS's traditional methods, but enhancing them. During the trial run, for instance, the DoIT system sent a rodent control team to a house where there had not been any 311 calls about rats. What they discovered, upon arriving, was the largest infestation the Chicago Department of Streets and Sanitation had ever seen. It turned out that even though there was no historical data indicating that rats were actually present in

that location—calls complaining about rats—there were other leading indicators present at that location—calls complaining about garbage, for example—that had historically correlated with the presence of rats and, when used to target inspection efforts, yielded results.

As a pilot, rodents were a perfect choice. Rats have no constituency, and so there was no political pressure not to study them. And the success of the program portends well for the future of the use of analytics in other agencies. "I've been very surprised and impressed with the pilot's findings," Cruz reflects. "I've become a believer. This can create a tool that will optimize our department's operations well into the future."

Berman says DoIT can run multiple operational predictive analytics pilots each year; DoIT now has several more pilots ready to go that are being vetted with the mayor's office and other stakeholders. Berman notes that working with stakeholders on these pilots is crucial: "The mayor's office is aware of what we're doing very early on. The more that they are aware and understand, the better, because sometimes there are implications to what we're doing that we're not even necessarily aware of. We keep the law department pretty close to what we're doing, as well."

Eventually, all of these individual operational pilots will be smoothed together into an overall set of business issues to which the city can apply predictive analytics. "What if" questions will someday be able to be asked, according to Berman, who has future plans to test the ability to query the system for problems that are not specifically defined.

The platform will not be capable of responding to those hypothetical questions for a couple of years; it needs data to be able to answer, and there are plenty of "unknown unknowns" out there, according to Danielle DuMerer, Director of Planning, Policy & Management, who has led the charge on a myriad of open data, business intelligence, and predictive analytics projects at DoIT. She believes that every time DoIT works on a process it should try to bring as much data as possible ("without being burdensome") into its environment.

Cross-department analytics are not yet in the works in Chicago. There the problems are still business intelligence. In sum, agencies simply need to understand better what other agencies are doing. "It's usually not analytics," says Schenk. "It's more about transparency about what's happening in between those two agencies."

The SmartData Platform

Chicago is working towards producing an open-source SmartData platform for predictive analytics, supported by a US$1 million Bloomberg Philanthropies grant. The plan is that this platform will be adopted by other cities. According to Danielle DuMerer, cities seeking to replicate Chicago's platform would get the cleaned-up code that drives DoIT's information environment, "scrubbed of our specific information." The package made ready for adoption for other cities would include advice as well as algorithms: "It's not just about putting up the code on Github and saying 'have at it,'" says DuMerer. "The chances of someone adopting that are going to be low." Instead, cities will be handed implementation documentation addressing everything from engagement with stakeholders to choices of questions. "Here are the best practices. Here is how you go about this as part of the plan," she says. "It will not be plug-and-play: each city will have to build its own hooks into its transactional systems," notes Brett Goldstein. "That's hard."

One of Berman's primary goals for the SmartData platform is that it be easy to use. "The overarching point of the SmartData platform is to build an application that allows the savvy business manager—that person who today can run that amazing report that tells the commissioner exactly what they need to do—to query the platform and leverage analytics without the involvement of a data engineer," she says. Not all cities have multiple data scientists on staff. To that end, the code for the SmartData platform will include the user-friendly WindyGrid interface—which allows representation of historical and real-time data mapped according to location.

The platform does not do everything. There are systems that prioritize work orders, for example, that have already been built; Chicago does not intend to build those itself as part of the SmartData Platform. As a basic tool, however, making progress toward replicable, scalable, and sustainable predictive analytics possible for smaller cities that do not have Chicago's resources, the SmartData Platform will be a major contribution to civic life. Ultimately, cities will be able to load their data in and compare their indicators—of wellbeing, pollution, and anything that can be built based on city data—to those of other cities.

Chicago wants predictive analytics to be the norm. The city's process of evangelism and business case discussion with its own agencies continues. "We're not completely done yet" with migrating data, according to Schenk; it may take until the end of 2014 for all city operational data to be under DoIT's central control and thus available for speedy processing. The fact that some agencies outsource the hosting of their data to private vendors—for whom helping DoIT add more data to its environment is not necessarily a high priority—has been a challenge; Schenk hopes to overcome this obstacle soon.

Another helpful lever used by DoIT with city departments is what Schenk calls Rapid Application Development (RAD). For example, a DoIT database analyst is skilled at using Oracle's APEX, a lightweight web application development tool for use with Oracle databases. If a department has a simple web-interface need—say it wants to be able to display data for citizens—DoIT will build an APEX app for that department. The condition? The department has to migrate its data. Finally, DoIT is having its predictive engine incorporate non-city open data, like weather and geo-located public tweets. Chicago has been collecting located tweets since early 2012 to take advantage of insights this data may provide.

As DuMerer puts it, analytics should be "just a part of how we function in the city. This is how we work to improve services. Analytics help us focus our energies on not just reacting to problems but actually getting ahead of them." Putting the important in front of the urgent, for once, can greatly assist government operations and improve civic outcomes.

Indiana: Building a Hub for Analytics

In Indiana, the state's Chief Information Officer under Governor Mike Pence, Paul Baltzell, oversees databases that store all of the state's information. Where other CIOs might have to persuade agencies or departments to allow the CIO to host their data, Baltzell, who took office in January 2013, is ahead of the game: he already has all the data due to the state's centralized model.

Since he took office, and with Governor Mike Pence's leadership and encouragement, the Office of Management and Budget (OMB) in partnership with the Indiana Office of technology (IOT) have been working closely with Chris Atkins, the director of Indiana's Office of Management and Budget, toward setting up a dedicated analytics center located in the Indiana statehouse and expected to open for business in the fall of 2014. The state-of-the-art center, called the Management and Performance Hub (MPH), will aim to address some of Indiana's most pressing problems by pulling data stored in Baltzell's databases into an analytics engine and by employing data scientists and analysts to assess patterns in that data that could help target state interventions. The MPH tool will be a resource for all state agencies. Baltzell's and Atkins' recent experience in getting to this point can, we believe, be helpful to other governments that are considering making predictive analytics a central tool of governance.

When a private company wants to instruct its units to cooperate in using data to drive results, it is relatively simple: all the management of Amazon or Priceline.com has to do is issue an edict. In government, it is different: there are laws, bureaucratic turf battles among agencies, sharp limits on resources, and built-in incompatibilities that raise barriers to widespread collaboration with data. To overcome the inertia of the status quo, it helps to have an initial shared problem—a "use case" for data analytics that forces cooperation—and a strong signal from top-level executive leadership. Baltzell has both.

Indiana has one of the highest infant mortality rates in the nation, as well as some of the highest levels of obesity and smoking of any state in the country: 7.7 infant deaths per 1,000 births, compared with the national rate of 7 deaths per 1,000 births. Indiana has chosen infant mortality as the first use case for the Management and Performance Hub. And the MPH Project has leadership from the top: the governor has tasked OMB and IOT to make this initiative a success. "Governor [Mike] Pence is in on this," Baltzell observes. "He gets where data analytics is going."

The correlation between infant mortality and such indicators as poor nutrition, unsuitable housing, and insufficient access to healthcare is well known and easily understood. It has been historically difficult, however, to operate on that correlation by increasing funding in a targeted way to support those programs and caseworker efforts that could most effectively address infant mortality. The MPH's analytics platform, however, allows the hub to integrate government data from family, health, finance, business, and employment agencies into its analysis, revealing less obvious connections.

The project's goal is to use these results to empower caseworkers to make data-driven decisions. With access to the tools developed by MPH, a caseworker will be able to compare a given family's information against data on all past and present at-risk families in the database. The result would be a clear risk estimate that allows the caseworker to determine the probability that a child will be at risk of future harm. For example, the algorithm may determine that "this child has an 80 percent chance of something really bad happening," Baltzell explains. "You need to remove him right now." Alternatively, a risk estimate of 10 percent may indicate a slight but significant chance of future problems. The platform, in turn, will enable the worker to more directly mitigate

the risk, whether that means contacting the right parole officer, counselor, or any other specialist in another government agency.

In order to focus on infant mortality, IOT had to do more than just host, or store, information from state agency databases. It had to build links between databases in ninety-two Executive Branch agencies and an analytics engine that would be able to ingest data, chew through it, and generate reports in as little time as possible. (The MPH Steering Committee chose to use SAP's HANA data platform as the analytics engine.) Baltzell decided not to build those links simply by fiat; instead, he worked closely with the agencies whose data he was already storing to build the case for analytics.

Baltzell encountered some pushback from midlevel civil servants in the agencies who were used to doing things in a particular way. These concerns were not just cultural—they were also about potentially violating federal regulations and about security. The MPH Steering Committee worked carefully through all of these concerns with the help of outside legal counsel. There is a healthy respect for the agencies' substantial legal concerns. An executive order issued March 16 by Governor Pence provides an avenue for agencies to raise legal concerns with OMG and IOT.

Security proved to be the biggest pre-launch task for Baltzell—which did not surprise him. Although his centralized data center manages the operating systems and databases used by state agencies, the applications used to extract the data are often under the control of individual agencies (as is usually the case when a specialized application is used by just one agency). The agencies felt there was a risk that the infant mortality project would somehow put their data into the wrong hands.

Baltzell built security in from the ground up. Indiana had not had a statewide data analytics operation in the past, so in an important sense, Baltzell was operating in a greenfield area—which made things easier. Baltzell could plan carefully for security. As he puts it, "We're doing everything we can."

First, his team physically and logically walled off the data analytics operation—the Management and Performance Hub—from the environment in which agency data is stored. Agency data was already protected by firewalls keeping out the rest of the world; now there is "another layer of the onion" keeping "big data away from everybody else here." Additionally, the Hub is not connected to the Internet. The room housing the data scientists is inaccessible to unauthorized employees and is a controlled and monitored environment.

Data will not be brought into the Hub (or, more accurately, be made part of the Hub) unless it is tagged with metadata showing where it came from, when it was created, and what legal requirements are attached to it—such as federal health, tax, or education laws limiting its use. This will allow Baltzell's team to track which data is being used and combined. The requirement of metadata insertion means that legal restrictions will travel with the data wherever it goes inside the Hub.

Employees from the Office of Technology will be responsible for the packages of mathematical algorithms that make sense of the Hub's data. After moving through an elaborate approval process, these individuals will gain access to a secure room that runs algorithms with SAP HANA's

software tools. The programs sort through data temporarily stored on the computer and allow immediate processing—and as a result, enable real-time decision making. In order to get into that secure room, employees will have to obtain an appropriate level of clearance and make a series of enforceable contractual promises. Baltzell's architects working on the Hub have elevated background security checks, which allow for administrative access to the data; his "database guys" have a lower level of access and do not have administrative access. The data scientists who will be working on predictive analytics algorithms are strictly monitored and will not be permitted to work remotely or bring USB keys or their own devices into the secure room. They must physically work in the controlled environment.

Data brought out of the Hub as reports for display will not include personally identifiable information. Case workers dealing with infant mortality risks in the field will not learn details beyond what is necessary for them to do their jobs—they will have access to private information, such as social security numbers. They will, however, have access to contact information for programs and agencies that may be helpful to the families they serve.

In order to ensure that the Hub meets the highest security standards, Baltzell has gone through several security reviews by outside vendors before launching Hub operations. He is in the process of making changes in response to their recommendations, and anticipates going through more "penetration testing"—attempts by "white hat" security experts operating from the outside (as well as pretending to be disgruntled inside employees) to defeat his electronic walls—before beginning work. All of the security practices he is adopting are being recorded in a (very long) document that he will have at hand should there ever be a question; he has an audit trail ready to go that, among other things, meets the NIST cyber security standards that have been adopted at the federal level.

Baltzell is proud of what he's done about security for the Hub and he wants to extend this protection more broadly: "I think we've wrapped more security around this than anything to protect people's private data." He acknowledges that he has gone beyond what others might have done with respect to security, even though it makes working on the data less convenient: "Some of the steps we've taken, we've taken to the extreme, but we've done it because we don't want there to be any risk of anything bad happening. It's actually made it more challenging for those mathematicians and those data scientists to work on the data because they can't work remotely," he says.

Space and Employees

As of fall 2014, a dingy office in the basement of the State House has been transformed into the Management and Performance Hub with a look and "feel like Silicon Valley"—it is aimed at engaging people and fostering collaboration, with interactive displays, an open physical layout for employees, meeting rooms, and whiteboard walls. The space will be useful for training, education, and outreach, as well as the day-to-day work of data scientists and analysts.

About 10 to 15 employees will work in the space full-time. Analysts will be paired with data scientists to work in the space. Those pairs of workers will be able to bring agency heads or

other agency actors already doing business intelligence work into the space; together, says Baltzell, they will work on statewide problems. The data scientists will come in and out as needed.

One of Baltzell's key challenges in getting the Hub going has been working with highly intelligent data scientists who are less experienced in communicating the results of their research. As he puts it, "I'm like, 'Hey, I actually need to pick out some key points to save some kids.'" It is very difficult to find data scientists who can communicate: "It's hard enough to find somebody with the skill set just to do the mathematical portion of it, the algorithmic portion. Then if you could find somebody who has both skill sets, you probably couldn't afford them." His hope is that he will be able to use the Hub's relationships with Purdue and IU to build a pipeline of data scientists who will work in the Hub. (Baltzell is involving researchers from IU, which runs the state's back-up disaster recovery data center, as well as academics and data scientists from Purdue.) "I'm willing to take a young guy and work with him for a couple of years, fresh out of college, even if he leaves to go make twice as much out on one of the coasts." So far, however, he has had to contract out all of this work, and he knows that hiring will continue to be difficult: "I already know that those guys can make considerably more in the private sector."

Future Use Cases for the Hub

Baltzell plans to hold brainstorming sessions about use cases. He says, "The governor has a roadmap of things he wants to particularly go after" that are problems for Indiana. Baltzell will bring in groups that might have relevant data about particular problems; he's also interested in identifying other problems that might be responsive to predictive analytics. Next up: recidivism. Baltzell thinks that data might reveal which areas of particular facilities or which offenders (or groups of offenders) are correlated with higher rates of return to prison.

Criminal justice, in general, may be a ripe area for the use of analytics in Indiana: Baltzell believes analytics might be useful in determining which offenders have lower associated risks of violence—and therefore can be supervised with GPS ankle bracelets or other monitoring rather than being sent back to jail. Indiana recently changed its criminal justice code by revising the way felonies are classified, which means that more prisoners will be assigned to county rather than state facilities. Predictive analytics may help assess how many beds and staff members a particular county facility will need in the future.

Another key issue for Indiana is economic development. Baltzell hopes to be able to identify drivers for business formation in Indiana. "If there's something we're not doing from a legislative or policy standpoint that can help increase business revenues or bring new businesses to Indiana, let's do it," says Baltzell.

One of the Hub's most promising implications may be in increasing the effectiveness of Indiana's government itself. In his March 2014 Executive Order directing the state Office of Management and Budget and Office of Technology to create the Management and Performance Hub, Governor Mike Pence affirmed that the Hub be a "tool for continuous process improvement for the State of Indiana," and asked that OMB provide him with recommendations about "opportunities to use data collected by state agencies to drive innovation and efficiency across state agencies."

To this end, Baltzell's staff has been converting existing state agency key performance indicators (KPIs) into forms that the data analytics engine can work with. This has been a challenge; agencies in the past calculated their own indicators using in-house statisticians, and much of this historical work lives only in Excel spreadsheets and homegrown visualizations. Baltzell's staff has been automating KPIs. Once the KPIs have been worked through by the Hub's analysts, they will be rolled out to agency heads and advanced using high-quality visualization tools combined with the speed of in-memory computing.

An important advantage of centralizing KPIs and automating them is that this will give the Hub the ability, over time, to drive better data practices into agencies through the program measures that underlie overall agency KPIs. Data quality, to date, has been a major issue for agency data; using analytics, Baltzell's team will, in the long term, be able to do what Baltzell calls "master data management." As he puts it, "We can correlate the data to make the quality better and then feed [the cleaned data] out to those [old] systems to correct their faults," in an ever-reinforcing loop. At this point, the Hub is still several steps away from this, but Baltzell can see the possibilities.

Building Support

Announcements about the Hub will come in concert with the launch of Indiana's public-facing open data portal. Using Socrata, and following the lead of many cities and countries around the world, Indiana will be publishing data and an API that allows the public to create applications using that data. Indiana and Indianapolis have already held public hackathons that Baltzell thinks have been fruitful. He is planning short-term fellowships that will involve working with this public data.

As Paul Baltzell and Chris Atkins move steadily toward making the Management and Performance Hub called for by Governor Pence a reality, they are cautiously optimistic. After months of coordination and steady attention to security, privacy, and the use case of lowering infant mortality rates, the Hub is almost ready for business.

The State of Indiana locates its predictive analytics unit inside its performance and innovation group. The insights gleaned from this data will give this highly interactive team both leads on where changes will produce the most value and the confidence they need to take on bold change. Armed with data, the team can more easily move reluctant bureaucrats, as well as defend an effort that might not lead to the predicted successes. Using data allows the Indiana governor's team to both reduce risk and generate a favorable environment for innovation.

Predictive analytics was first used by commercial enterprises for which slight improvements in performance were worth large investments in technology and personnel. For example, since the business models of credit card companies and insurers depend on detecting fraud and scoring customers based on predictions of future ability to pay stemming from past activities, those companies saw early on the benefit of automating statistical work that used to be done (laboriously) by hand. As hardware and data management tools have become cheaper and more accessible, the use of predictive technologies has ballooned. Today, reports that might have taken hours to produce or were not produced at all—because computer processors were too slow and the processors themselves were so expensive—can take minutes if the right hardware is in place. As a result, nearly every corner of private industry takes advantage of predictive analytics.

Governments have been slower to adopt these techniques. Again, a particular sector has led the charge: The public safety community's "business model" depends on quick detection and prediction of crime in order to protect the public, and so it has been willing to invest in easily deployable software packages and faster hardware. As a result, "hot-spotting," predicting where the next crime is likely to happen, is widespread, particularly in the United States.

This section lays out examples of the use of predictive analytics by governments, organized by policy area. These examples range from potential to mature uses of technology. Some of the techniques we describe were developed in-house by government agencies. Others have been sold to governments by outside vendors. We hope this catalog will inspire additional experiments in predictive analytics. The use of data will not only improve the results of government programs, but will also reduce risks to government innovators. Risk is often the most important stumbling block within governments that have historically been averse to trying new things, but data-driven predictions will enhance confidence in change and thus, create conditions more conducive to innovation. In other words, data-driven innovations will assist the innovator in arriving at the right choices and also provide support (in the form of analysis) in the event of failure.

Public Safety

For years, prisons have used data and detailed statistics in their predictions of which prisoners should be allowed out on parole. Many jurisdictions use predictive models to determine who should be detained before trial and who may be safely allowed out in the community while awaiting trial. Yet these assessment tools do not generally access the wide variety of available data; they too often rely on data generated by the offender himself. They have no insight into which factors affect the conduct of a particular person.

Predictive analytics is a natural next step for police departments engaged in intelligence-led policing or data-driven policing. The advent decades ago of computerized records, management systems, and automated call dispatch systems created large stores of data about basic policing operations. In the 1990s, the development of CompStat applied computer statistics as a management tool to allocate resources to police beats with specific problems and to serve as a management and accountability tool. Simultaneous developments in geographic information systems and automated data presentation tools have allowed departments to identify and map geographic "hot spots" of crime and to devote resources to those areas. These innovations have driven a performance measurement-based philosophy that provides a strong foundation for predictive analytics.

There are a number of ways police departments can use predictions: to predict patterns of crime; to predict which offenders may commit further crimes and merit more careful surveillance; and to predict which individuals, groups, or neighborhoods are at risk of becoming victims of crime. Again, predictive policing is not intended to replace existing policing techniques, but rather to complement them. Predictive analytics builds on problem-oriented policing, community policing, evidence-based policing, intelligence-led policing, hot spot policing, and other proven strategies. Predictive policing is most effective when it is part of larger proactive strategies that build strong relationships between police departments and their communities to solve crime problems.

One thing that predictive analyses cannot replace is solid data infrastructure. Quality data is an important input for predictive analytics, and predictions cannot replace basic data management. Police leaders still need their management reports—dashboards, crime maps, key performance indicators, and so on.

Predictive policing uses statistical algorithms to anticipate likely crime events, allowing departments to prepare for and prevent such crimes. Predictions can focus on variables such as places, people, groups or incidents. Demographic trends, parolee populations, and economic conditions may all affect crime rates in particular areas. The goal of predictive analytics is similar to the goals of many prior technology innovations in policing: to reduce administrative work and to allow officers and investigators more time for proactive patrol, engagement with the community, response to incidents, and active investigation of crimes.

Large and small public safety agencies are using predictive analytics to address a variety of challenges. Departments develop predictive algorithms to look at relationships among various elements, such as the relationship between school truancy and a rise in neighborhood burglaries, abandoned buildings and illegal drug markets, and so on. Other possible factors to examine are public parks, eviction data, school data, zoning information, and mental health data.

- Los Angeles was among the first cities to experiment with predictive analytics, leveraging the work of academics at the nearby University of California at Los Angeles. Professor Jeff Brantingham of the Anthropology Department developed a mathematical model, based on predictions of the aftershocks from earthquakes that allowed for predictions of future crimes. The system calculates and creates predictions based on simple, constantly updated data such as the location, time and type of crime. The system then creates

prediction boxes of 500 square feet and displays them, color-coded, on a patrol map. In one of the first areas to implement predictive policing, the Foothill Division, crimes were down 13 percent in the four months following the rollout compared to an increase of 0.4 percent in the rest of the city where the rollout had not happened. In the words of the LAPD leader for the predictive analytics project, "We have prevented hundreds and hundreds of people coming home and seeing their homes robbed" (CBS Los Angeles, 2012).

- The Chicago Police Department (CPD) has a long track record of leadership in technology innovation. CPD was selected by the U.S. Department of Justice for implementation and research funding to advance predictive analytics. The pilot project under way at CPD will evaluate the ability of modified pattern-matching software currently used for medical diagnostic purposes to predict crime patterns. It will also evaluate the efficacy of a software tool that quantifies and maps gang activity to predict emerging areas of gang conflict, leveraging Yale sociologist Andrew Papachristos' research on social networks. Dr. Papachristos has discovered that violent activity is limited to a small number of individuals concentrated in one particular neighborhood. By identifying not just "hot spots" of crime, but also "hot" individuals and their networks, police officers can proactively approach members of the network. When contacting members of the network, police and social workers encourage them to leave behind their violent gang lifestyles by helping them to find jobs. As a result of this work, crime statistics in the area are showing early decreases. As Commander Jonathan Lewin, told The Verge: "This [program] will become a national best practice. This will inform police departments around the country and around the world on how best to utilize predictive policing to solve problems. This is about saving lives" (Stroud, 2014).
- NYPD Commissioner Bill Bratton has been a proponent of data as a management tool for decades and is widely credited with developing and advancing CompStat as a police management tool. He implemented predictive policing in Los Angeles. Now in his second tour of duty as NYPD Commissioner he plans to do the same, so that his officers can predict and prevent crimes. "Already, the department is looking back at the hours before every shooting to see which 311 complaints and 911 calls for minor events—loud music playing, crowds gathering—preceded the violence and, if addressed in the future, could help head it off."
- Miami-Dade County Police Department's predictive analytics tool is focused on solving cold cases and catching repeat offenders. When a crime is committed, an officer can quickly access a list of potential suspects based on match probability from existing data in the system. The list allows investigators to narrow their focus from thousands of known offenders to those with the highest probability of having committed the crime. The system leverages crime patterns and offender modus operandi information from huge volumes of historical data.
- The Shreveport Police Department conducted a pilot project in six of its highest-crime districts to see if a predictive model could reduce crime. The model used data on juvenile

complaints, loud music, disorderly persons, suspicious activity, loitering, disputes, and prowlers. Crime analysts developed maps of likely locations of property crime. An experiment to determine whether the predictions could reduce crime was inconclusive due to a lack of statistical power and inconsistent fidelity to the program model. However, the areas where the predictive model was used spent less (six to ten percent less in labor costs) for the same results. For the first four months of the experimental period, the areas using predictive models reduced property crime by 35 percent.

- The Charlotte-Mecklenberg Police Department (CMPD) has a predictive analytics system that uses operational, tactical, and predictive dashboards to visualize on a map the areas that have the highest probability of a crime occurring during any four-hour window. Data elements in the predictive model include historical crime patterns, recent crimes (in the last 24 and 48 hours), recent call for service activity, motives, physical location, day of the week, time of day, weather, political factors, economic factors, school calendars, pay periods, special events, and more. The insight gained from the solution helps command staff to more knowledgeably deploy resources and allows officers to more effectively manage their areas of responsibility to deter crime.
- The Memphis Police Department (MPD) reduced crime by more than 30 percent, including a 15 percent reduction in violent crimes over a four-year period, by using predictive analytics. MPD is now able to evaluate incident patterns throughout the city and forecast criminal "hot spots" to proactively allocate resources and deploy personnel, including directed patrol, targeted traffic enforcement, task forces, operations, high-visibility patrol, and targeted investigations. The system compiles large volumes of crime data, including incoming data sources from patrols pertaining to type of criminal offense, time of day, day of week, or various victim/offender characteristics. MPD can now better guide daily decisions that address criminal activity and place officers in a better strategic position to respond to an unfolding crime.
- The Arlington, Texas Police Department used data on residential burglaries to identify hot spots and then compared these locations to areas with code violations. According to Chief Theron Bowman, officers found a direct connection between physical decay and residential burglaries. Arlington then developed a formula to help identify characteristics of "fragile neighborhoods" so that they could work proactively with other city agencies to help prevent crime (Pearsall, 2010).
- Lancaster, CA successfully deployed a predictive model to reduce crime by 37 percent over a three-year period. It began with hiring a data analyst. That one staff person was able to build a predictive data model, and then was able to show crime patterns on heat maps and color codes. The maps allowed the data to be used proactively to allocate police resources across the city while achieving a dramatic decrease in crime and a US$1 million annual benefit in productivity (IBM, 2011b).

Social media has provided police departments with a new way of sharing information with the public in real-time. As was demonstrated by the Boston Police Department in its response

to the Boston Marathon bombing, using Twitter to update the public can be an effective way to directly provide important safety information, as well as to dispel rumors and share accurate information.

Social media data can also help predict crime trends. Public Twitter data provides a vast trove of information. A recent study in Chicago used 1.5 million geo-coded tweets over a three-month period. This study showed that for 19 of the 25 crime types studied, adding Twitter data to existing crime prediction models improved the accuracy of predictions (Gerber, 2014).

Public Safety

Limited public funds plus the constant impulse to prioritize today's urgent problems over important long-run issues pose a great danger to bridges, tunnels, and roads in need of inspection and maintenance. Knowing which segments of physical infrastructure are likely to be deteriorating would be useful to government managers anxious about maximizing the benefits of expensive repairs and upgrades. Perfect physical inspection, however, would require huge numbers of inspectors who themselves might be imperfect reporters.

As sensors of all kinds become cheaper and easier to deploy, they can assist infrastructure managers in targeting their upgrade and repair efforts. For example, sensors that "phone home" with the data they automatically collect can provide information about humidity, temperature, jostling, pollution, and other environmental issues that cause infrastructure to degrade. "Predictive repair" is now a possibility—using data to predict the likelihood of degradation, and then setting a threshold that triggers repairs before rather than after a system is damaged. Scheduling predictive maintenance lowers long-run costs and is less disruptive of citizens' lives than reacting to crises.

Public transportation systems can be kept up using preventative maintenance. For instance, in the London Underground, a thorough collection of data revealed key indicators that were strongly correlated with equipment failures. This work also identified mechanical parts in the system's escalators that needed to be replaced in order to avoid future failures (Shueh, 2014). As a result of this work, repairs went from 50 percent to 70 percent efficacy on the first visit, which will translate to considerable savings over time and easier trips for riders.

Networks of sensors can become knowledge producing infrastructure that can prompt far more efficient deployment of government services. In Rotterdam, Paris, Leuven, and London, for example, the RainGain project is covering these cities with rainfall sensors (Rain Gain, 2012). Catchment data drawn from these sensors will be combined with radar mapping data to help the cities dynamically predict the effects of floods on a street-by-street basis. (Satellite imagery, taken alone without the sensor data, would not provide these predictions at such a fine-grained level.)

In Leuven, sensors detect the river's water level as well as the velocity of water movement within the sewer and height of overflows. The system is still being built, but the idea is that forecasters will be able to more intelligently design city infrastructure to manage periodic rainfall and to reduce the risks of future disasters. Rotterdam, as a city mostly below sea level, relies

on a network of mobile water pumps and storage basins to keep its streets and basements dry. Real-time predictions from the radar system will help the city deploy pumps where they are most needed and reduce the risk of flooding.

In the United States, the New Jersey Turnpike serves 200 million vehicles a year in one of the country's densest regions. Traffic continues to grow, causing congestion at peak travel times. The Turnpike Authority has responded by building more lanes to ease traffic flow. At the same time, however, the Authority has begun installing sensors at regular intervals that gather information on traffic volume, lane occupancy, and speed. Based on this information and the predictions it facilitates, traffic managers can reroute cars along the roadway and signal to trucks to avoid accidents (Goldsmith, 2014a). Networks of sensors and the predictive analytics their data enables can be used to make expensive physical infrastructure more useful, producing more intelligent management of these resources.

Fraud Detection

Large entitlement programs and social services attract fraud and abuse, putting a strain on treasuries and jeopardizing funding intended for those in need. However, the prohibitively high cost of audits and human-led fraud detection is a drain on state resources. Entities like Service Canada's Integrity Service Branch—responsible for disbursing employment insurance funds—are using predictive risk analysis to detect fraud and abuse Office of the Privacy Commissioner of Canada, 2012). Los Angeles County's Department of Public Social Services does the same, analyzing its data to find anomalies that can then be used to prioritize the workload of investigators responsible for detecting potential fraud within the state's childcare program (Goldsmith, 2014b). Patterns in this data reveal leading indicators that have historically correlated with the presence of fraud. By focusing prospectively on those indicators rather than using some other method of prioritizing their workload, auditors can increase the likelihood that their work will pay off in the form of a higher success rate. The presence of these indicators becomes a red flag—a sign that abuse may have occurred, and the resulting audits conserve resources. In 2012 this approach yielded substantial savings: 200 cases were flagged with 85 percent accuracy and saved the department US$6.8 million in eliminated fraudulent cases (Heaton, 2012). Agencies using data can be proactive: no longer dependent on tips from hotlines or hunches to discover fraud, but rather knowing exactly what easily detectable patterns are likely to be associated with fraud and investigating these occurrences directly.

Human Resources

As budgeting practices have improved, revealing the true sources of cost in the operation of large entities, those in charge of finances have come to realize how costly attracting and sustaining human capital is. Predictive analytics can help by improving the success of hiring practices (Fitz-Enz, 2009). A good potential application can be found in the military and other sectors that invest

heavily in their human resources yet risk losing that investment if new hires do not stay in place. Predictive analytics can be used to improve recruitment in the first place for the sake of retention. Leading indicators that predict success and long-term job satisfaction can be used to focus scarce government recruiting and mentoring resources on those soldiers who are most likely to remain in the ranks over the long haul (Web Builders, 2011).

Several leading-edge companies, including Google, Cisco, and GE, are using predictive analytics to guide their hiring rather than relying on their intuitions about candidates (Human Capital Media, 2014). The United States Government Office of Personnel Management is pursuing predictive analytics to help make decisions about what positions to fill and which individuals will be the best fit for the necessary work. OPM is using workforce data that includes skills assessment, workforce diversity, retirement eligibility, location of employees, hiring and retirement patterns, and turnover statistics to guide it in its decision making (Ward, Tripp, and Maki, 2012).

Social Services

When it comes to providing services like child welfare, the diversity of existing clients as well as the difficulty and importance of working with at-risk populations, make it difficult to achieve positive outcomes. Even worse, while experienced providers develop a knack over the years for connecting the right children to the right services to improve outcomes, that knowledge is difficult to transmit between workers in a formal way. It is lost and must be rebuilt whenever a social worker or manager stops working.

Indiana (as we discussed) is planning to use analytics to help reduce infant mortality (Howard, 2013). While officials picked this particular goal because of its political feasibility—a cause everyone can get behind—the effort is not overly expensive. The state is receiving a grant of US$500,000 from a private philanthropy to pursue this project as well as other technology improvements; most of the money is coming from the regular IT budget (Hughes, 2014).

The Medway Youth Trust has begun using predictive analytics instead of its laborious prior process to find and identify high-risk young adults. By analyzing its digital records and finding historical correlations between particular characteristics of youths and outcomes, the Trust has begun to be able to predict which youths will have trouble in the future. It has replaced a manual process involving review of thousands of records with a much faster and more accurate way to link high-risk youth with the services and opportunities they need to be successful. This substitution of analytics for manual processes can help all parts of the system: it can reorient youths before they are pushed down paths that turn out to be worse for them and cost the state more (IBM, 2011a).

New York City's Department of Homeless Services is working with the SumAll Foundation to comb through its shelter check-in data. Knowing that a family's recent address has become a hotspot for a large number of problems—such as evictions—the city can focus its limited resources on that address and reduce the risk that future families will need to enter the homelessness system. By turning its attention to causes of homelessness rather than merely

reacting to its consequences, the city can move toward better outcomes for parents and children (Mascarenhas, 2014). This means lower costs for the city and better-quality and better-targeted social services for at-risk populations.

Tax Collection

Tax agencies have begun using predictive analytics to perform a first pass of their own records, flagging likely cases of unpaid taxes or misleading accounting practices. Based on these flags, particular files can be prioritized for audits and legal reviews (SPSS, 2010). Predictive analytics can also be used to complement existing economic forecasting techniques, allowing agencies to better understand what effects new policies may have on tax revenues in the future (Desigan, 2011).

For example, New York State uses predictive analytics to preemptively flag questionable tax returns in its work stream to collect its revenues, while also saving the cost of pursuing these problematic returns later. Since its implementation, the program has helped recapture US$400 million that was being lost in wrongful returns annually, in addition to US$100 million of increased revenue availability. The department's leadership is pleased that this kind of investment was made during better economic times when they were flush with funds and could take such preventative steps. Now these efforts are paying off (IBM, 2011c).

Disaster Management

In times of disaster, quick access to relevant predictions of impact and estimates of needed resources can make the difference between life and death. Predictive techniques can assess what responses may be needed to incoming information more quickly than a human may be able to realize what is happening, balance all the variables in his or her mind, and make a conscious decision to intervene.

The United States is continuing to develop its National Response Framework (NRF) and National Incident Management System (NIMS)—both aimed at performing analytical information management and getting needed real-time information to those who are in a position to reduce the risks of disasters (UICDS, undated). These two national frameworks and support structures, part of the country's homeland security planning, are intended to add to the capacity of disaster management agencies across the country—both big and small[3]—to respond effectively to disasters. The availability of ever-improving weather modeling supports these efforts: data scientists can now predict more accurately where severe weather might be headed, which, when combined with GIS data, can be used to aid in all stages of a disaster. These models will help predict impact before a storm hits to assist in evacuation planning or during a disaster so that local agencies will

[3] For example, the Los Angeles Operational Area Critical Incident Planning and Training Alliance (http://www.catastrophicplanning.org/alliance.html) and Cowell County Emergency Management (see http://www.cowleycounty.org/wp-content/uploads/ESF-5-Emergency-Management.pdf).

know where aid will be needed. Their ability to plan ahead of time to target disaster assistance efforts and post-disaster recovery will both save money and keep more Americans safe.

Public Health

When the United States saw the growth of the H1N1 pandemic in 2009, the Centers for Disease Control and Prevention and the National Institutes of Health used predictive analytics to guide their activities and health warnings based on the predicted spread and strength of the virus (Desigan, 2011). Tools like Google's Flu Trends and Mapmyhealth may add new data sources to help target where diseases like flu may be spreading—long before traditional metrics like hospital emergency room visits or laboratory tests begin to spike. Google's Flu Trends gathers and geographically analyzes users searching for flu related resources, while Mapmyhealth and SickWeather mine Tweets and Facebook postings in real-time, searching for people who report feeling sick with the flu (Kongel, 2013). These steps may help save lives. By the time a patient goes to the hospital, the opportunity to prevent further outbreaks of disease may be lost.

Experimental uses of mass social media data may enrich our ability to predict and track many public health problems. For example, researchers at the University of Rochester have shown that geo-located Tweets can be used to monitor food poisoning stemming from restaurants in New York City, prioritizing the strained workload of the City's food inspectors by uncovering insights that allow them to focus on restaurants with spikes in social media "reporting" (Kautz, 2013). The app has not yet been launched, however, and remains a research product, largely due to entrenched interests. Since the app uses crowd-sourced data, groups like the New York State Restaurant Association have opposed its use. They worry that competitors could manipulate source data to "frame restaurants" (Taney, 2013). Researchers are working to improve the algorithms to be immune to such manipulation, but its non-implementation points out two key issues with predictive analytics: (1) if using crowd-sourced data, which is common, efforts must be made to ensure it has not been tampered with; and (2) even then, existing political groups and interests may fiercely protect the status quo.

Analytical techniques have allowed researchers at Microsoft to work toward identifying women at risk of severe post-partum depression by way of changes in the patterns of their online habits, as well as the content of their postings (Kautz, 2013). Other researchers at Microsoft have demonstrated that they can uncover previously unknown problematic drug interactions through an analysis of users' anonymous web searches (Greenemeier, 2013). Thinking through the privacy and consumer implications of these approaches will be crucial.

The predictive approach is also used with more traditional medical data sources. IBM and the University of the Ontario Institute of Technology are using biomedical readings from monitored premature babies to predict potentially fatal infections nearly a day before they may be observed otherwise (IBM, 2010). That single day can give doctors an important head start in initiating treatment and can prevent infections from developing into more serious complications.

The aging population is growing worldwide, which is consuming more resources and putting additional stress on the system for routine distribution of healthcare services. The United States Affordable Care Act created financial incentives for hospitals to reduce their readmissions rate patients being hospitalized again within 30 days of their previous hospitalization—because repeated visits generate US$30 billion in costs each year (Health Affairs, 2013). To find ways to reduce these costs, the Heritage Health Prize, a contest put on by the Heritage Provider Network, uses historical claims data to predict which patients are most likely to be readmitted. Flagged patients receive extra attention to their discharge and post-hospitalization treatment plans, with the goal of reducing or preventing readmission. Similarly, the King's Fund has been working in the UK to build a software system that will identify patients with a high risk of hospital readmission, so that local primary care trusts can intervene and reduce the risk of patients returning to the hospital (The Kings Fund, 2014).

Predictive analytics and machine learning can also be used to complement standard medical methodological studies. Using medical and pharmacy claims data, data scientists can begin to test the leading indicators for multiple combinations of therapies that may be correlated with desired outcomes. This complex approach—something few human minds would be capable of—will help providers develop solutions to large-scale health issues, like type 2 diabetes. Addressing these mass problems will have a major impact on public health budgets; the cost to society of type 2 diabetes alone is expected to balloon over the coming years (Maguire and Dhar, 2013). This same data-driven and predictive approach can be used to refocus the healthcare provision on prevention rather than just reimbursement for visits and services—bringing attention and resources to public health (the desired outcome) rather than merely assessing the outputs of public hospitals and other institutions.

In the United States, the country's health welfare systems for the elderly and the poor, Medicare and Medicaid, have adopted predictive analytic technologies in order to detect fraud The enormous amount of data collected by these programs could be used to better understand patterns of activities for all users, not just those involved in fraud. In other words, the aggregated characteristics and actions of beneficiaries could be used to predictively tailor eligibility and coverage plans. As a result, populations can be insured more efficiently, putting affordable healthcare in the hands of more segments of the population and creating a better standard of health for the citizenry at large.

Education

A major implementation of predictive analytics in education that has emerged recently involves flagging students who appear at risk of dropping out. Educational institutions can provide support for that student even before the student him or herself realizes that he or she may be at risk (Luan, 2004). This support has been made possible through improvements in institutions' ability to track students from institution to institution; at-risk students tend to jump among schools. In the past, legal barriers made such cross-institutional tracking difficult. Today, the National Student

Clearing House allows for matching of data across universities and community colleges, allowing for a more complete understanding of students' characteristics and propensities. Rich data is becoming available for use as indicators; systems can incorporate grade data, financial aid data, and student data to build a predictive model to allow the prioritization of interventions and the provision of additional resources to high-risk students (Barber and Sharkey, 2012). While these efforts are just beginning, it is clear that predictive analytics will hold an important place in education, particularly in ensuring that public programs aiding and supporting disadvantaged students are as effective as possible and generate the highest possible graduation rates. Programs can do this by aiding students as their circumstances change rather than waiting for problems to arise to take action. By then, it may be too late.

Conclusions

Better, faster, and cheaper government is within our grasp, in developing nations as well as developed countries. An increasing number of national and subnational governments are taking advantage of open source tools to launch uses of data that will help make their government more effective and transparent. Indeed, the high penetration of smartphones in developing nations can be a built-in advantage for government actors in those countries: government employees providing services can inexpensively inform central databases about what they are doing, via GPS technologies, and open data efforts can be immediately visible to those with hand-held devices.

Technology has provided a spectacular set of tools that can help public servants solve problems and produce better outcomes with the same amount of effort, but more intelligently applied. If our goal is to increase the connection between effort and results, predictive analytics can improve the ratio dramatically. We have cloud computing, data mining and sentiment analysis tools, nearly universal connectivity with smartphones and other mobile devices in the hands of workers and citizens alike, and processing tools unheard of or unaffordable just a few years ago.

Standing in the way is mostly us—the way we run government. Operating with executive leadership, cross-agency tools, transparency and open data, and a concentration on results (instead of activities) can, in fact, produce a more responsive government. Data will drive better services, more effective ways to manage and evaluate employees, and through transparency and social media, more pervasive citizen involvement.

Steps Toward Success: Where To Begin

A new Chief Data Officer (CDO) arriving in City Hall who wants to use predictive analytics will need to consider a wide range of issues: the availability and nature of public data, the skills of his or her employees, the political heft of the office, and the nature of the problems that can and should be addressed. The following section lists a set of basic questions any Chief Information Officer or CDO will need to answer in setting up a predictive analytics function.

For purposes of this discussion, the person in charge of the predictive analytics agenda is not directly responsible for the demanding and time consuming responsibilities of securing data and running the day-to-day information technology needs of his city or state (such as ensuring that the email servers work or addressing employees' requests for technical help). The CDO we have in mind will have responsibility for database management, open data, advanced analytics, business intelligence, and data warehousing.

What is the existing data infrastructure? The first layer of any predictive analytics capacity is data. That data may be (and probably is) in distributed databases controlled by different municipal or state agencies. The agencies and relevant cities and states may not yet be collecting data that will be crucial, such as information flowing from GPS devices on buses and trains.

The CDO should carry out a thorough inventory of existing datasets and any description of the fields within those datasets—often called "documentation"—that has been maintained. If such documentation does not exist, the CDO will need to ensure that it is created. Often, particular individuals within agencies maintain specialized datasets and dedicated enterprise systems. These individuals, who themselves know what is in that system and what it is useful for, have had no real incentive to make that information available to anyone else. The CDO's initial investigation will be necessary to make knowledge about that data or existing system available to his operation.

The City of Chicago recently released a first-of-its-kind "Data Dictionary,"[4] which describes the fields contained in all its datasets. While Chicago is not releasing all of its data to the public, it is cataloging all the different kinds of data it has. This makes it possible for city staff, residents, and others to understand Chicago's data resources, how and if those resources can be accessed, and in what formats they can be accessed. Other cities are adopting similar, interoperable data dictionaries. If the CDO from the outset has aggregated (or commissioned) documentation of the data in use, participation in cross-city data analytics efforts will someday be possible—which will be a powerful tool.

[4] See http://datadictionary.cityofchicago.org/.

Understanding the gaps between the data the CDO's unit of government is collecting and data collected by other cities or states will also be useful at this point. Down the line, the CDO may the government to increase its data collection efforts; having a complete inventory and gap analysis on hand will be essential to that step.

In addition to finding out who is responsible for each kind of data that is collected, the CDO will need to determine what transparency, quality, and retention requirements apply to the data with which he will be working. Close and friendly coordination with the legal department (and the IT security officer charged with these responsibilities) will be useful.

Finally, this inventory stage will need to include an assessment of how much data (and which data) is under the direct physical control of the CDO's function. Different data systems strike different balances between centralization and decentralization: the State of Indiana's Management and Performance Hub, described above, was relatively easy to create because the state CIO was already hosting the data of all state departments.

Chicago's data analytics function, by contrast, has been built across distributed databases (meaning that many datasets are hosted by city departments and not by the central Department of Innovation & Technology) and has worked by building automated processes (ETL processes, meaning "Extract, Transform, Load") to those distributed databases. Over time, Chicago has persuaded agencies to make their data available for analytics by demonstrating how much more quickly and efficiently that data can be worked with when it is hosted on centralized, powerful servers.

In general, existing transactional databases (crime, city services, permitting) should not themselves be used for predictive analytics; they are busy serving the city or state and should not be interfered with. Instead, the job of the CDO is to build an analytics engine or layer that is fed by constantly, automatically refreshed data drawn from those transactional databases.

Insight: The more the CDO knows about his or her city's data resources, the more effective he or she will be.

What are your analytic needs? Not all analytics are predictive. Sometimes what is needed is simply evaluation: what percentage of graduates is a school system producing compared to its peers? Has a street camera system been rigged by a shady vendor to produce more traffic violation tickets than it should? How much money has the city spent on a particular cross-agency sector?

If predictive analytics are already a strategic need of the city or state, the CDO will need to establish the principles of his office: will he prioritize analytics that have immediate operational

impact (resulting, for example, in someone in a truck with a flashing yellow light on it being sent out to a particular street to make a particular repair), or that are useful for long-term economic planning (resulting in a twenty-year plan for a park)? The City of Chicago chose immediate operational impact as its priority, building use cases for predictive analytics in close coordination with city departments. Architects in Chicago are also using predictive analytics based on historical environmental and transportation data to plan the building of a new section of the city; that building project will take at least another decade. Most city or state CDOs need to operate using limited financial resources, and setting priorities—creating a decision rule that aids in accepting or turning down proposed projects will help in the allocation of those scarce funds.

Overpromising the benefits of predictive analytics—raising the expectations of government chief executives, so that they believe data will give them a crystal ball in which they can see the future of their city or state—can be destructive. Those CEOs will be prone to ask broad, vague questions like "How can we make life better for citizens?" No CDO can answer that question, because its answer requires an enormous number of policy choices and assumptions (what does "better" mean?) that are inherently subjective. The CDO who initially takes on small problems whose solution will be a minor but visible improvement in city services will probably be more successful. Particularly when the entire idea of predictive analytics is new, progress will be smoother when results are visible: it is difficult for humans to imagine the usefulness of something they cannot see. Focusing on early, visible wins will help create additional champions and enthusiasm for larger projects and more staff.

Here is an example of a very simple and small problem: payment systems sometimes generate errors, and even very small errors can add up to real money. If the dataset containing payment information meets statistical requirements (including if it is sufficiently large and has values that are widely distributed), running a simple "expected first-digit distribution" algorithm across that data will reveal whether there are outlying payments that should be examined.

Another example of a category of problem that may be susceptible to visible, straightforward analytics involves asking the question "where are city problems [crime, trash] actually happening, and what have been the drivers for these problems?" Most cities have a license to Esri's ArcGIS, software that facilitates spatial presentations of data. Disparate transactional databases (such as ones holding crime, 311 data, building inspection data) can be loaded into that software, and then small areas can be compared and analyzed to reveal leading indicators that may have relationships to the dependent variable of interest.

Some CDOs decide to start with the creation of performance metrics for city departments and then use analytics to assess whether those metrics are being met. Others see this process as a distraction: the scope of questions that can be answered by key performance indicators will be limited by the data already being collected by the agency in question ("we don't have data on this, so we can't measure our performance there"), which makes the process circular.

Some CDOs want to work primarily on inductive research, plowing through enormous amounts of disparate data to find unexpected patterns and then exploring them ("what is in our 311 plus 911 data that is interesting and unexpected?"). Others are more focused on deductive

work, looking for discrete problems, finding drivers, and creating proposed test solutions to improve the status quo. Whatever prioritization scheme the CDO decides to follow should be informed by and closely aligned with the strategic priorities of the chief executive.

Insight: Predictive analytics successes will bring increased demand for data analytics projects. The CDO should be clear about his or her priorities, refining those priorities in light of what he or she learns over time.

Who are your champions? Buy-in from the chief executive will be essential for any predictive analytics function to be successful. Chicago, again, is an example of a city in which a data-savvy leader—Mayor Rahm Emanuel—has given his employees running room to carry out extensive predictive analytics work. Mayor Emanuel's analytics staff, in turn, has been careful to work closely with city departments in eliciting key business problems they want to solve (including the deployment of city resources).

Having champions within the agencies, in the form of people who are knowledgeable about the data their department controls and enthusiastic about the power of data-driven government, is extremely valuable, but will not be sufficient in the absence of political authority stemming from the highest level of government. Effective data analytics requires political cover, which carries with it increased resources and conferred credibility—both attributes that are useful when trying to persuade disparate elements of any government to work with you. Effective CDOs will need to be both skilled in analytics and effective in working easily with entrenched bureaucracies.

One former CDO told us that he wished he had known about politics and power plays before he entered office. He had not understood the significance of the difference between career civil servants and political appointees, and he was unaccustomed to having to "tell a story" about why analytics is important. He is a "data guy," impatient with the slow pace of government functions. He did not know about procurement rules, buzzwords, or lobbying pressures. In his words: "Data is easy; but everyone will be your obstacle."

Insight: It is important that the CDO role be seen as directly supportive of the city's (or state's) overall goals and that the chief executive visibly support the CDO.

What skills do your employees have? Just as important as an inventory of datasets will be an inventory of people. Talent identification and expansion will be essential to any successful predictive analytics function.

The CDO will need database administrators who are curious, energetic, and knowledgeable about statistical analysis—and can get things done. Even if their past experience has been limited to working with enterprise software (Oracle, for example), they can be retrained to work with open source analytic packages and will likely be thrilled to learn new things.

The CDO will also need a full-time ETL administrator—someone whose job it is to build and maintain ETL processes that extract stable information from existing databases and automatically load it into the predictive analytics engine. This is the root canal of predictive analytics work; there is no substitute for having these ETL processes in place, and it is painstaking work to create and scale them.

Project managers who are skilled at working with city departments will be essential, both for their political skills and their specific domain expertise; these can be less technical people who have backgrounds in policy issues on which the chief executive is focused (transportation, climate, etc.), have experience in running cross-agency processes, and are excellent communicators. These managers can stay in constant touch with departments; managers will thus understand the problems the departments face and have opportunities to suggest how working with the CDO can make the agency's job easier.

Jacks-of-all-trades are important; these are people who are flexible, have some technical skills (most importantly, visualization expertise), and are focused on client service. Lightweight development work will be essential for making the story of any predictive analytics experiment accessible to laypeople, and simple web tools that are easy for departments to use to present their own narratives will help bring them onboard with the CDO's predictive analytics agenda. These generalists can both serve more specialized, highly technical employees working on analytics or databases and support work on presentations and web interfaces.

Data scientists who have a deep, fundamental comfort with statistics, including the necessary elements of any valid statistical statement, and who are skilled at working with data, are central. These people should have experience with forecasting and prediction, know how to design and analyze the experimentation (AB testing and other techniques) that validates differential treatments, be creative in coming up with test design, sampling, and measurement solutions in non-standard arenas, be able to write scripts, and have experience with R or other statistical analysis packages. Often, partnering with local universities will be needed in order to bring these people inside local or state government.

If the CDO elects to create an open data portal, someone who is knowledgeable about open data practices and can provide insight about the attributes of databases that are being made available by way of the portal will be valuable. Open data can be an important driver for predictive analytics functions; the same scripts used to automate the availability of refreshed data for analytics purposes will be needed to populate a data portal, meaning that programming code can be used for dual purposes. Practitioners have found that open data portals are often even more useful inside

government than they are outside the walls of city hall, because one agency can use the portal to see what data other agencies have made public without having to pick up the phone. A CDO with authority over both open data and advanced analytics can ensure that these two functions serve one another; he can also make investments in staff and resources that are useful to both.

Insight: The CDO will need to combine understanding of government processes with technical facility. Putting in place a team with expertise across these areas, and encouraging ongoing communication among the members of that team, should be a top priority.

What vendors should you work with? Government CDOs are targets for extensive vendor attention, and any CDO will need to get on top of existing procurement processes. The business of predictive analytics is booming, and vendors know that few longtime government employees have experience in the field; often, existing IT employees have gotten in the habit of carrying out autonomous IT procurement processes. These days, however, a knowledgeable CDO has leverage: vendors will often share information with government actors about their private-sector implementations, because the local government is not competing with them and they want to make the sale. So ask questions.

Also, a CDO can now avail himself of many free, open source tools that can provide powerful analytics support. Even with limited resources and disparate existing transactional databases, a CDO can use multiple if not state-of-the-art computers and servers, use an open source ETL software (Pentaho) to ingest distributed data, use an open source software project that enables the distributed processing of large datasets across clusters of commodity servers (Hadoop) on top of all of those machines, use open source analytics platforms or layers (PostgreSQL or MongoDB) to grapple with that data, use an open source statistical program (R) to do regression tables, use an open source scripting language (Python) to run processes against the data, use an inexpensive visualization package (Tableau), and emerge with useful results.

Employees who understand statistics (e.g., Does the data have a normal distribution? Are there enough data? When are regressions allowed?) can learn to carry out these functions and talk about the results knowledgeably and clearly. Fortunately for government employees, information technology is becoming "consumerized;" thus, reducing reliance on private vendors can not only save the city or state a great deal of money but can also free resources for long-term investment in people. Moreover, as more cities and states use these platforms and create reusable packages, relatively inexpensive analytics will become ever more widely available.

Insight: The successful CDO will ask vendors well-informed questions and study his or her options carefully.

CDOs need to be thoughtful about "make or buy" decisions. There are many feature-rich products from companies like Oracle, IBM, and SAP that may be useful for some functions but not necessary for others. For example, private cloud platforms may be enticing to a CDO because they appear to solve a host of problems at once. (Someone else, after all, has bought the hardware and licensed the software used by the cloud, and the employees used by the cloud company are not line items on the CDO's budget.) If a CDO is working extensively with the data for 24 hours a day, renting a cloud facility can quickly become prohibitively expensive; every time the CDO asks a question, another charge is incurred.

Insight: Evaluation of data analytics projects should be built into those projects from the outset. The successful CDO will look for early wins.

On the other hand, renting cloud resources can be helpful when the CDO needs to have access to expanded processing capabilities for a short burst of time. The CDO may want to scale up—to use a great deal of computing horsepower—to carry out a particular task, and then bring the results back to his or her own operation; in this context, renting may make sense. Similarly, Oracle services can be extremely useful, but their expenses mount up quickly.

We are not advocating that CDOs operate free of support from vendors. CDOs should, however, be cautious about locking themselves into highly architected enterprise packages, particularly now that alternatives are widely available.

How will you show the effectiveness of data-related projects? A CDO will always be on the lookout for new opportunities to integrate and leverage the information the city has already gathered or will gather under his or her direction. Constant, visible evaluation of the impact of these projects will help bolster support for the activities and enlist other employees to the cause. Successful predictive analytics practices frequently demonstrate value, so that their champions can see examples of the usefulness of these systems.

How helpful are the peers of the CDO? At the moment, many city and state CDOs are not aware of what their colleagues in other cities are doing. This means that master techniques are not being handed off and today's pioneers may be feeling unduly isolated and embattled. Just as a successful predictive analytics function is built on collaboration across many business units of government, a CDO's functions will be enriched by collaboration with other jurisdictions; eventually, cross-jurisdiction analytics efforts will become routine. In creating a data strategy, hiring employees, finding champions, setting priorities, and evaluating personal successes, the CDO should reach out to his or her peers to learn about what works (and what does not) in other locations.

Insight: These days, a CDO charged with carrying out predictive analytics projects does not need to work alone. Networks of other CDOs should be leveraged.

REFERENCES

Barber, R. and M. Sharkey. 2012. "Course correction: Using analytics to predict course success." In S. B. Shum, D. Gasevic, and R. Ferguson (Eds.), In Proceedings of the 2nd international conference on learning analytics and knowledge (LAK '12). New York, NY: ACM.

Capt. Sean Malinowski as quoted in AP. 2012. "LAPD The largest agency to embrace 'predictive policing.'" CBS Los Angeles. Retrieved from http://losangeles.cbslocal.com/2012/07/01/lapd-the-largest-agency-to-embrace-predictive-policing/.

CGI Group. 2013. "Predictive analytics: The rise and value of predictive analytics in enterprise decision making." Retrieved from http://www.cgi.com/sites/default/files/white-papers/Predictive-analytics-white-paper.pdf.

Desigan, S. 2011. "Predictive analytics: A case for nonprofits and government contracting agencies." AnalyticBridge. Retrieved from http://www.analyticbridge.com/profiles/blogs/predictive-analytics-a-case-for-nonprofits-and-government.

Fitz-Enz, J. 2009. "Predicting people: From metrics to analytics." Employer Relations Today 36, 1–11.

Gerber, M. S. 2014 "Predicting crime using Twitter and kernel density estimation." Decision Support Systems 61. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167923614000268.

Goldsmith, S. 2014a. "Digital transformation: Wiring the responsive city." Civic Report 87. New York, NY: Manhattan Institute for Policy Research. http://www.manhattan-institute.org/html/cr_87.htm.

———. 2014b. "Big data gives a boost to health and human services." Data-Smart City Solutions. Retrieved from http://datasmart.ash.harvard.edu/news/article/big-data-gives-a-boost-to-health-and-human-services-380.

Greenemeier, L. 2013. "Your smartphone just diagnosed you with postpartum depression." Scientific American. Retrieved from http://blogs.scientificamerican.com/observations/2013/05/03/your-smartphone-just-diagnosed-you-with-postpartum-depression/.

Health Affairs. 2013. "Medicare hospital readmissions reduction program." Retrieved from http://www.healthaffairs.org/healthpolicybriefs/brief.php?brief_id=102.

Heaton, B. 2012. "Los Angeles County uses analytics to stop child-care fraud." Retrieved from http://www.govtech.com/health/Los-Angeles-County-Uses-Analytics-to-Stop-Child-Care-Fraud.html.

Howard. A. 2013. "On the power and perils of 'preemptive government'." O'Reilly Radar. Retrieved from http://radar.oreilly.com/2013/02/preemptive-government-predictive-data.html.

Hughes, J. 2014. "Indiana uses data analytics to lower infant mortality, child fatality." Government Technology. Retrieved from http://www.govtech.com/health/Indiana-Uses-Data-Analytics-to-Lower-Infant-Mortality-Child-Fatality.html.

Human Capital Media. 2014. "Predictive hiring: Find candidates who will succeed in your organization." http://www.slideshare.net/humancapitalmedia/predictive-hiring

IBM. 2010. "University of Ontario Insitute of Technology: Leveraging key data to provide proactive patient care." Retrieved from http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=AB&infotype=PM&appname=SNDE_OD_OD_USEN&htmlfid=ODC03157USEN&attachment=ODC03157USEN.PDF.

———. 2011a. "Simple solutions for local government in an era of austerity." Retrieved from http://public.dhe.ibm.com/common/ssi/ecm/en/yts03031gben/YTS03031GBEN.PDF.

———. 2011b. "City of Lancaster takes a predictive approach to policing." Retrieved from http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=PM&subtype=AB&htmlfid=YTC03339USEN

———. 2011c. "New York state tax: How predictive modeling improves tax revenues and citizen equity." Retrieved from http://www.ibm.com/smarterplanet/us/en/leadership/nystax/assets/pdf/0623-NYS-Tax_Paper.pdf.

Kautz, H. 2013. "There's a fly in my tweets." The New York Times. Retrieved from http://www.nytimes.com/2013/06/23/opinion/sunday/theres-a-fly-in-my-tweets.html.

Konkel, F. 2013. "Predictive analytics allows feds to track outbreaks in real time." Federal Computer Weekly. Retrieved from http://fcw.com/articles/2013/01/25/flu-social-media.aspx.

LaVall, S., E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, N. 2011. "Big data, analytics and the path from insights to value." MIT Sloan Management Review 52 (2).

Luan, J. 2004. "Data mining applications in higher education." SPSS. Retrieved from http://www.spss.ch/upload/1122641492_Data%20mining%20applications%20in%20higher%20education.pdf.

Maguire, J. and V. Dhar. 2013. "Comparative effectiveness for oral anti-diabetic treatments among newly diagnosed type 2 diabetics: Data-driven predictive analytics in healthcare." Health Systems 2: 73–92.

Mascarenhas, R. 2014. "Illuminating housing challenges with data." Data Smart City Solutions. Retrieved from http://datasmart.ash.harvard.edu/news/article/illuminating-housing-challenges-with-data-389.

Office of the Privacy Commissioner of Canada. 2012. "The age of predictive analytics: From patterns to predictions." Privacy Research Papers. Quebec, Canada: Office of the Privacy Commissioner of Canada. Retrieved from http://www.priv.gc.ca/information/research-recherche/2012/pa_201208_e.asp.

Pearsall, B. 2010. "Predictive policing: The future of law enforcement?" NIJ Journal 266. Retrieved from http://www.nij.gov/journals/266/Pages/predictive.aspx.

Rain Gain. 2012. "Work Package 2: Fine-scale rainfall data acquisition and prediction." Retrieved from http://www.raingain.eu/en/fine-scale-rainfall-data-acquisition-and-prediction.

Shueh, J. 2014. "Predictive analytics aboard the London underground." Government Technology. Retrieved from http://www.govtech.com/transportation/Predictive-Analytics-Aboard-the-London-Underground.html.

SPSS. 2010. "Making critical connections: predictive analytics in government." Chicago, IL: SPSS Inc. Retrieved from ftp://ftp.boulder.ibm.com/software/data/sw-library/spss/IMW14284USEN-00.pdf

Stroud, M. 2014. "The minority report: Chicago's new police computer predicts crimes, but is it racist? The Verge." Retrieved from http://www.theverge.com/2014/2/19/5419854/the-minority-report-this-computer-predicts-crime-but-is-it-racist.

Taney, P. 2013. "New food illness tracking app getting mixed reviews." WHEC-TV NBC. Retrieved from http://www.clipsyndicate.com/video/play/4225546/new_food_illness_tracking_app_getting_mixed_reviews?wpid=5435.

The Kings Fund. 2014. "Predicting and reducing re-admission to hospital." Retrieved from http://www.kingsfund.org.uk/projects/predicting-and-reducing-re-admission-hospital.

UICDS (Unified Incident Command and Decision Support). Undated. "Unified Incident Command and Decision Support (UICDS): A Department of Homeland Security Initiative in Information Sharing." McLean, VA: UICDS.

Van Barneveld, A., K. E. Arnold, and J. P. Campbell. 2012. "Analytics in higher education: Establishing a common language." Educause Learning Initiative. ELI Paper 1. Retrieved from http://net.educause.edu/ir/library/pdf/ELI3026.pdf.

Ward, D. L., R. Tripp, and B. Maki. 2012. Positioned: Strategic workforce planning that gets the right person in the right job. Saranac Lake, NY: Amacom.

Web Builders. 2011. "Predictive analytics for federal government." Retrieved from http://www.informationbuilders.com/pdf/factsheets/FS_Solution_Rstat_GovFed_2011.pdf.

IDB

Inter-American
Development Bank

2016