CENTER FOR THE ADVANCEMENT OF PUBLIC INTEGRITY

Taking a Byte Out of Corruption

A Data Analytic Framework for Cities to Fight Fraud, Cut Costs, and Promote Integrity

The Results of the Center for the Advancement of Public Integrity Data Analytics Working Group
February 28, 2017

COLUMBIA LAW SCHOOL

TABLE OF CONTENTS

EXECUTIVE SUMMARY

In recent years, the emerging science of data analytics has equipped law enforcement agencies and urban policymakers with game-changing tools. Many leaders and thinkers in the public integrity community believe such innovations could prove equally transformational for the fight against public corruption. However, corruption control presents unique challenges that must be addressed before city watchdog agencies can harness the power of big data. City governments need to improve data collection and management practices and develop new models to leverage available data to better monitor corruption risks.

To bridge this gap and pave the way for a potential data breakthrough in anti-corruption oversight, the Center for the Advancement of Public Integrity (CAPI), with the support of the Laura and John Arnold Foundation, convened an expert working group of leading practitioners, scholars, engineers, and civil society members to identify key issues, obstacles, and knowledge gaps, and map a path forward in this promising area. CAPI supplemented the deliberations of this working group with further research and more than forty field interviews in New York and Chicago. (See Appendix A for more detail on our sources.)

This yearlong study came to the following conclusions:

- Data-driven investigative approaches require robust, consistent, and reliable data. City governments should implement policies to improve data collection and management in order to undertake pilot projects testing the usefulness of data analytics to monitor and detect potential fraud, waste, and corruption.

- Efforts to improve data collection and pioneer data-driven approaches should focus on high-priority **areas of corruption risk**—situations or circumstances common to city government that are particularly prone to fraud, waste, abuse, or corruption. These areas, outlined in the report, are where cities can get the most "bang for their buck" by using known fraud indicators to detect red flags among existing data sources. While data-driven approaches will not magically identify corruption, they can help pinpoint potential corruption risks that enable investigators to more efficiently target their efforts.

- What cities can do now is to identify starting points to test data-driven efforts, using existing data sources and known indicators. And as practitioners in various cities make headway in this fast-changing field, CAPI can facilitate the sharing of knowledge and best practices to accelerate progress against mutual challenges.

The following framework explores these findings in greater detail, to provide independent guidance for cities seeking to develop data-driven investigative approaches. Neither a technical manual nor a theoretical study, the framework aims to be practical, actionable, adaptable, and sensitive to the needs and interests of city watchdogs.

The framework has several parts:

- **Part I** gives background information about project goals and context and delineates key concepts.
- **Part II** details key challenges facing cities in terms of fraud, waste, abuse, and corruption for data-driven approaches, and discusses available data sources and potential indicators that may present pathways to address such challenges.
- **Part III** presents best practices concerning data collection and maintenance, indispensable to any city or agency intending to use data-driven investigative methods.
- **Part IV** notes relevant cautions and considerations, such as foreseeable obstacles and issues of performance measurement.
- **Part V** concerns next steps for development—a few illustrative examples of data-driven monitoring and investigative methods that may be relatively easy to implement, as well as areas for future research.
- **Part VI** concludes the framework with a final summary of key points.

- Finally, there are three appendices:
  - Appendix A summarizes the research methodology and lists the working group members who contributed to this report and the interviewees who helped inform it.
  - Appendix B discusses relevant examples of applications of data analytics in related fields.
  - Appendix C is a "Top Ten" list of starting points for data-driven approaches to fraud and corruption risks using readily available resources.

I. BACKGROUND

20th Century Tools for a 21st Century Challenge

Corruption, the abuse of public authority for private gain, poses a tremendous challenge for communities across the country, costing U.S. taxpayers tens of billions of dollars each year and spurring a vicious cycle of citizen distrust and disengagement.[1] Recently, state and local corruption schemes have grown more sophisticated, involving manipulation of public resources, complex bid rigging, and money laundering conspiracies. For example, New York City's CityTime scandal involved a massive $600 million procurement fraud related to a city-wide payroll computerization project.[2] New Jersey's Operation Bid Rig resulted in the convictions of dozens of people for a conspiracy involving money laundering, political corruption, and even interstate organ trafficking.[3] The unprecedented scale of corruption and fraud in the Hurricane Katrina recovery efforts have resulted in more than 1,300 indictments.[4] Additional examples of data analytics are described in Appendix B.

Currently, corruption investigations predominantly rely upon random spot-checks and tips from whistleblowers. Both strategies are time-tested but limited. Random spot-checks can be scattershot and laborious. Tip-driven investigations are reactive and often dependent upon whistleblowers or other witness statements, which may or may not be reliable. In theory, data-driven approaches could be more **efficient**, directing limited resources to the most probable problem areas. They could be more **effective**, uncovering evidence of systemic corruption risks. And they could be **proactive**, identifying risks before they metastasize into entrenched problems.

Unfortunately, the application of data analytics to corruption control is complex. Unlike typical uses of urban informatics, corruption has no straightforward inputs, outputs, or metrics of success. Which data would be relevant? What red flags should investigators look for? What benchmarks can agencies use to measure gains and fine-tune approaches? Such thorny questions have delayed the deployment of data-driven approaches to the fight against municipal corruption. Data-driven investigative methods used in other fields of law enforcement have yet to be translated into effective and scalable corruption control.

Data-Driven Approaches

Information is at the heart of any investigation. Any investigation involves obtaining information and then sorting through it to uncover facts that are relevant, significant, and credible, and to marshal those facts into an accurate narrative. Throughout this process, leads from a variety of sources can help an investigator find, interpret, and evaluate information.

A data-driven investigation might use data to gather evidence to supplement leads from other sources such as complaints, witness interviews, or evidence from other investigations. **Data Analytics** is the process of examining large data sets in order to generate useful leads, patterns or relationships directly from the data itself.

Types of data analytics particularly relevant to data-driven corruption investigations include:

- **Exploratory**: used to uncover new patterns within data
- **Confirmatory**: used to support or weaken existing hypotheses about data
- **Predictive**: uses current and historical facts to inform hypotheses about future or unknown incidences
- **Quantitative**: uses numerical data
- **Qualitative**: uses non-numerical data like natural language, images, names, or addresses
- **Data mining**: used to discover hidden patterns and relationships within large data sets
- **Outlier analysis**: used to detect anomalies within data, such as incidences that deviate from an expected pattern

In theory, data-driven corruption monitoring could initially involve exploratory analysis of relevant city-specific data sets to find outliers for further evaluation. If this process were effective, the anomalies detected by the process would provide enough useful leads to justify further investigation and evaluation.

II. COMMON RISKS AND POTENTIAL APPROACHES

All cities face common challenges in addressing waste, fraud, abuse, and corruption. This section highlights nine areas of corruption risk that are recommended as starting points in the use of data-driven methods because of their prominence in municipal corruption investigations and the availability of usable data and time-tested indicators in these areas. For our recommendations for ten approaches across the below topics that are relatively easy to implement with existing data sources please see appendix C.

A. Fraud by Inspectors

What are the corruption risks?

Many regulated industries in cities rely upon periodic inspections, including daycare centers, restaurants, healthcare providers, senior care facilities, for-hire vehicles, and waste removal services. Inspections of buildings and construction sites represent one particular area of fraud in many cities, involving many actors—developers, contractors, expeditors, and inspectors—who interact with one another and the city over the lifespan of a construction project, exposing multiple corruption risks. Bribery schemes involving inspectors can compromise public safety and cost city taxpayers. For example, building developers might offer bribes to building inspectors in exchange for helping to accelerate projects to completion notwithstanding the failure to complete critical steps in the process. Fire inspectors might be offered bribes in exchange for overlooking expensive code violations.

How can cities use existing data sources to better target corruption risks?

Inspection records can be a rich source of information to help anti-corruption officials identify incidences of bribery or fraud. Records documenting inspection activities, typically maintained by municipal building departments in the construction arena, for example, may contain common indicators of potential corruption.

We note that the ability to access the information contained in inspection records, like all of the other data sources identified in this framework, will vary greatly depending on whether the information is available electronically or on paper. While all cities should move towards electronic records, some will be further along than others in that process, and the ability of a city to easily implement some of these suggestions will be based on large part on how its information is kept.

A careful review of inspection records may reveal the following indicators of bribery or fraud:

- Outliers in the number of reported code violations by inspectors, for example an inspector whose reports of violations seems exceptionally high or low without an obvious explanation;
- Missing inspection sheets and/or route sheets that may indicate skipped visits;
- Inspection sheets with irregularities, for example being signed twice by the same inspector, rather than by both an inspector and a supervisor;
- A short amount of time between a failed inspection and a passing inspection of the same property, which may indicate a suspicious acceleration of the inspection process.
- In cases where inspectors work in pairs, especially when pairings are supposed to be randomly assigned, look for recurring pairings of the same inspector teams;
- Look for evidence of inspectors frequently being (allegedly) denied access to a site, which may be serving as an excuse to close a complaint when the actual reason is a bribe;
- Look for relationships between particular licensees or contractors or site owners/operators and the individual inspector who works with them—consistent sign-offs from the same inspector across different, ostensibly unrelated, projects could signal fraud; and

- Examine how often a licensed private inspector is being used to determine if the frequency of jobs is realistic, or whether it suggests exaggerations in the number of reported activities.

Hotline complaints can also point toward potential corruption hot spots. Cities that have complaint hotlines (e.g. 311 in New York) or smartphone applications can leverage these robust data sources to identify incoming complaints that are tied to particular inspectors.

Important strategies to consider include:

- Look for evidence of complaints that were changed or downgraded, or where the language of a complaint itself was downgraded, following its initial logging;
- Identify cases in which complaints were negotiated to lower levels; and
- Identify occurrences of unusually swift site visits or resolutions of complaints.

If a particular inspector, or group of inspectors, is suspected of engaging in corrupt practices, **GPS data** from government issued cell phones and cars can indicate potential red flags for follow-up investigation.

During the investigation, cities should consider using the following strategies:

- Use GPS data to track movement to ensure that inspectors are at reported locations at expected times—deviations from assigned routes may indicate suspicious behavior; and
- Map data to find geographic trends—if one or more individuals are engaging in potentially fraudulent behavior, track their movements to uncover other partners in crime or other projects.

Spotlight on: Fraud by Inspectors

In 2015, after a two-year investigation by the New York City Department of Investigation and the New York County District Attorney's Office, over 50 defendants, including 11 building inspectors and five Department of Housing Preservation and Development employees, were charged with multiple counts of bribery and fraud. The building inspectors allegedly cleared complaints and "stop work" orders, and expedited inspections, allowing for construction to continue even if the site was unsafe or in violation of building and construction codes. In exchange, the inspectors received sports vehicles, a luxury cruise, and other payments. In one case, a construction worker was injured on the job site, and an inspector instructed the contractors not to report the case because it would raise questions about the site's safety. The investigation started when the Department of Housing Preservation and Development noticed discrepancies in an internal audit.[5]

B. Human Resources-Related Fraud by Public Employees

What are the corruption risks?

Like any large employer, the city oversees massive systems of human resources to govern hiring, payment, and benefits provision. Within this structure, cities run the risk of being defrauded by public employees who attempt to exploit payment mechanisms and benefits for personal gain.

Examples of human resources-related fraud include:

- Fraudulently submitted timesheets to obtain overtime payment;
- Abuse or overuse of vacation time;
- Falsifying mileage on cars for increased reimbursement;
- Recruitment schemes that involve nepotism and conflicts of interest; and
- City employees fraudulently receiving state benefits, or vice versa.

If an employee is found to be committing one of these types of fraud, they are likely to be engaging in other forms of human resources-related fraud and should be reviewed accordingly.

How can cities use existing data sources to better target corruption risks?

Timekeeping data is the most important source of information to examine to identify human resources-related fraud. Many cities have electronic systems for employees to log hours, or require employees to swipe in and out of their work places. These data can be analyzed to spot unusual trends, particularly regarding the payment of overtime.

Proven indicators of fraud to look for in timekeeping data include:

- Excessive logging of overtime hours, including unusually high weekend overtime;
- Overly consistent employee overtime (*e.g.,* always the same day of the week, or the exact same number of hours every time);
- Individual employee timekeeping records that deviate from departmental norms—these can often be identified by comparing time swipes among similarly situated employees;
- Overtime taken without a formal request or supervisor approval;
- Time sheets that have been revised after, or upon, supervisor approval;
- Missing approval signatures from supervisors on time sheets;[6]
- Vacation hours that do not line up with employee time swipes;
- Increased labor hours with no corresponding increases in materials used or units produced; and
- Relationships between specific employees and their work patterns, such as one being sick and the other working overtime, or clusters of individuals with comparable titles across departments earning the same overtime payments.

Cities that allow employees to submit **reimbursement requests** for miles driven on personal vehicles run the risk that employees will exaggerate mileage accrued to obtain reimbursement.

Strategies for identifying fraudulent claims regarding mileage reimbursement include:

- Look for individuals, agencies, or teams that are routinely taking maximum or close to maximum mileage claims;

- Compare mileage claims for personal cars with mileage logs of government-issued cars to identify irregularities;
- Verify claimed miles by double-checking the standard mileage calculation between given points (such as the employee's home address and work address).

GPS data on government-issued cell phones and vehicles can also indicate time-stealing and/or employee fraud. More specifically, these datasets can help public integrity officials assess whether an employee is where he or she should be to perform expected duties, or whether he or she might be absconding from official responsibilities to conduct personal business on government time.

- Use **geofencing** technology to establish a pre-set grid where employees are expected to work, in order to receive notifications when someone has deviated from the area;
- Investigate cases in which employees report a high number of broken GPS systems;
- Use mobile phone metadata, including the times of calls, how long calls lasted, the numbers dialed, and the approximate locations of calls, to confirm locations of employees; and
- Identify remote logins to the department's network and the timing of such logins to determine employee location.

Finally, **onboarding paperwork** can be analyzed to identify potential cases of nepotism—the preferential treatment of relatives and friends. Using records of new hires, promotions, appointments, and contractors, anti-corruption agents can flag instances that warrant further investigation. Strategies to consider include:

- Look for an unusually high number of employees appointed by the same person;
- Identify instances of newly created positions that originate from a single individual; and
- Review employee names to look for commonly occurring family names, indicating that employees may be related, and review employee addresses to look for recurring addresses, indicating that employees live in the same household.

Spotlight on: Human Resources-Related Fraud by Public Employees

A common problem for many cities is the fraudulent use of overtime by city employees. For example, in 2014, a former Baltimore Department of Transportation employee was indicated on theft charges for allegedly forging signatures of his supervisor on overtime authorization forms for up to 2,227 hours, for which he received nearly $73,000 in overtime pay, nearly double his salary.[7]

In another example, a 2015 audit found that employees in Los Angeles' Department of Transportation collected overtime pay totaling $3.3 million in a year. It amounted to a 263% increase in overtime pay in that division between 2000 and 2015. The audit found that employees in the traffic paint and sign division overbilled an average of $48,100 between 2013 and 2014. Several supervisors received as much as $70,000 in overtime pay while one superintendent's salary tripled after he claimed $155,319 in overtime.[8]

C. Benefits Fraud

What are the corruption risks?

Many cities manage the distribution of federal, state, and local benefits to eligible residents, including those receiving Medicaid, cash assistance, food stamps, or other types of public assistance. In many cases, residents receive payment cards to access their food and cash benefits. This card may be used to purchase authorized food items at participating retailers or to make cash withdrawals from participating banks. Unfortunately, common risks are that city employees illegally create false benefits for personal gain, or conspire with beneficiaries committing fraud, often in collaboration with certain retailers. (Note that while fraud by beneficiaries alone is another huge concern for cities that will benefit from data analytics tools, fraud against the city that does not involve public officials—and therefore does not implicate corruption—is beyond the scope of this framework.)

How can cities use existing data sources to better target corruption risks?

Information about how **electronic benefit transfer (EBT) cards** are used and how employees are creating benefits cases may lead cities to identify fraudulent employee behavior. Although the program is managed by the United States Department of Agriculture, city and state social services offices typically maintain benefits case records that may provide insights into this type of fraudulent behavior.

To identify benefits fraud, cities can helpfully examine when and how benefits are redeemed:

- Check to see if transactions from EBT cards are occurring outside the zip code, or proximate zip codes, of the beneficiary's reported residence;
- Examine EBT card transactions—card information that is manually inputted may indicate that the individual does not physically have the EBT card and that a fraudulent purchase has occurred;
- If a "natural disaster" code was used to justify the benefit, ensure that the disbursal of benefits related to an actual natural disaster, particularly in timing and location;
- Review the timing within which benefits are redeemed—fraudulent benefits may be immediately redeemed (*e.g.*, at midnight on the first day benefits are active);
- Search death records to identify potential usage of beneficiary activities on behalf of deceased clients;
- Ensure individuals receiving disability benefits are documented to have disabilities by checking employee rolls, business licenses with the Secretary of State, status with the Department of Health and the Social Security Administration, and the State Department of Professional Regulation;
- Map business processes to track the data flow of benefits claims and distribution, which can often involve multiple levels of government that may each introduce potential risk; and
- Monitor inventory and distribution to look for outliers such as single retailers or processing centers that record disproportionate usage. Relatedly, check for higher than normal incidences of certain usages at single retailers/processing centers, such as manual entry or "keying-in" of codes as opposed to card swipes.

Cities can also adopt the following strategies to identify suspect behavior of city employees:

- Look for employees who issue emergency benefits at an unusually high rate;
- Review timestamps when a new benefits case is opened and closed—if new cases are quickly closed, a case worker may be disbursing benefits and closing the case before it is properly reviewed;
- Observe whether certain benefits disbursal codes that require less documentation are used more frequently than others;
- Look at computer login data to ensure case workers are taking action on assigned cases;

- Ensure the case worker and person auditing the case are not same person by cross-checking both log-in and IP address information; and
- Search for the use of expired benefit reason codes.

Spotlight on: Benefits Fraud

In 2015, multiple city employees in the New York City Human Resources Administration were charged with exploiting benefits, such as rental assistance and nutrition programs typically reserved for the city's neediest residents. One such scheme involved using fraudulent EBT cards to buy food and beverage items for resale. Another scheme involved falsely registering individuals as landlords and giving them fraudulent rental subsidy payments. To avoid detection, the employees used paper records to bypass computerized systems. In the end, the investigation by local, state, and federal authorities found that these employees generated over $2.1 million through their fraudulent schemes.[9]

D. Campaign Finance Violations and Theft of Public Funds

What are the corruption risks?

Campaign donations are subject to strict regulation, and must be monitored by city agencies to ensure compliance with local donation limits. For example, in New York City, the contribution limit to a mayoral campaign for the entire election cycle is $4,950. If there is a business relationship between a contributor and the City, the limit is lowered to $400.[10]

Some cities also have **public matching programs**, in which public funds are given to campaigns to match private contributions. As of 2015, at least seventeen local jurisdictions and thirteen states in the United States made some form of public funding available for political campaigns.[11] For example, the Campaign Finance Board in New York City matches small campaign contributions at a six-to-one ratio. Recently, the city of Seattle implemented a new program in which, instead of providing matching campaign contributions, four $25 vouchers are mailed to each voter. The voter is then free to donate the vouchers to his or her preferred candidates.[12] Regardless of the local system, public campaign funds must be spent appropriately, and unused funds are required to be returned to the city at the conclusion of a campaign. Cities that choose to operate public matching programs must be supported by a robust enforcement mechanism that is both tough and fair. Actually enforcing the law in this area is crucial to incentivizing compliance and giving the public confidence that their funds are not being wasted.

Campaign finance fraud can be committed by campaigns, donors, or both. Common types of campaign finance fraud include:

- "Straw donors"—the making of surreptitious donations by illegally using another person's (the "straw donor's" name;
- Donations made by non-residents, in jurisdictions where there are residency requirements for donors;
- Reporting of fictitious contributions to defraud public matching programs; and
- Improper splitting of large contributions from a single source into small donations to obtain increased public matching funds.

How can cities use existing data sources to better target corruption risks?

For cities that have donation limits, publicly available information about **campaign donations** can be analyzed to detect fraudulent activity. Primary data sources include lists of individual, corporate, and intermediary campaign donors, which can be cross-referenced with detailed administrative datasets that contain information about:

- The addresses of city residents;
- Registered lobbyists;
- Vendors that do business with the city;
- Tax documents showing income or financial losses; and
- Discretionary spending by City Council members.

Approaches for reviewing campaign donations and associated datasets for corruption risks include:

- Review donations from repeating out-of-district donor zip codes;
- Determine whether there are multiple donations from the same address;
- Determine if a campaign has received a high percentage of cash donations;
- Flag and investigate same-day donations and contributions from first-time donors;

- Review donor history to see if the amount given is in line with prior donations;
- Check previous donations by the same donor to compare handwriting against the current donation;
- Look for sequential money orders, which may suggest straw donors;
- Verify that donations from corporations do not exceed the legal limit; and
- Use a manual inspection to identify a high number of donor cards where the handwriting looks similar or the cards look altered.

In cities that have public matching programs, or that otherwise provide public funding, it is also important to examine **public fund reporting**. A city's campaign finance board will require supporting materials to accompany reports regarding the use of public funds, following the issuance of matching dollars. Review of the supporting materials—including leases, contracts, consulting agreements, and receipts—can reveal irregularities indicative of corruption or fraud.

When considering documents submitted in conjunction with the use of public or matching funds, it is important to:

- Determine if the documents are complete and contain sufficient information;
- Assess the timing of expenses, given knowledge about the sequence of campaign events;
- Ensure each expense has a corresponding receipt and is related to an eligible campaign purpose; and
- Confirm employment payments match up with the timing of the campaign.

Spotlight on: Campaign Finance Violations and Theft of Public Funds

The New York City Campaign Finance Board manages a public program that matches small campaign donations at a six-to-one rate. In the 2013 New York City mayoral race, two members of candidate John Liu's campaign were convicted for violating campaign finance laws. To take advantage of the public matching program, they created "straw donors" whom they reimbursed in order to enable the campaign to access matching funds. Following the investigation, Liu was denied up to $3.5 million in public matching funds.[13]

E. Petty Theft of Public Resources and Inventory

What are the corruption risks?

On a day-to-day basis, city agencies are responsible for managing thousands of assets and transactions, all of which are overseen by city employees with direct access to items themselves, the financial statements underlying transactions, and other records documenting inventory. Fraud may occur when city employees steal inventory for their personal use or resale, or collude with outside contractors to help them obtain supplies or materials at a discount.

How can cities use existing data sources to better target corruption risks?

Inventory records—including information about the acquisition, storage, and distribution of goods—can be used to identify fraudulent activity. Per the Government Accountability Office (GAO), cities should ensure they have detailed, accurate electronic inventory systems in place, along with proper accountability mechanisms for recording and updating inventory data on a routine basis.[14] When maintained properly, these automated inventory management systems position city officials to leverage data analytics to identify fraud. Cities will greatly benefit in trying to identify and prevent inventory fraud by moving to an electronic system; the need to wade through paper documentation of inventory will make the suggestions below significantly more challenging.

When examining inventory records, the following strategies may help public integrity officials identify incidences of corruption:

- Look for the continued purchase of items that have been declared in excess and/or unnecessary;
- Look for a significant uptick in ordering, potentially indicating that employees are selling excess inventory;
- Look for abnormalities in inventory requirements, including high stock levels, and incomplete, inaccurate, or outdated inventory records;
- Examine cases in which inventory is declared lost, damaged, or broken, to verify the accuracy of such reports;
- Compare freight expenses—the cost of delivering goods—to inventory purchased to identify suspicious patterns; excessive freight expenses relative to inventory purchased may be a sign that assets are being diverted for personal use;
- Look for suspicious patterns in goods being sold on secondary markets, such as eBay—an unusually high number of sales of a particular item the city uses may suggest fraud;
- Examine sales documents to ensure that all funds are transferred appropriately and match the inventory that is sold; and
- Compare requirements for the production of a certain material to the actual number of materials ordered—for example, assume one item X is purchased for each item Y to make product Z. If records reflect the purchase of l000X and 500Y to produce 500Z, examine the excess purchases of item X.

In addition to reviewing data, there are certain management techniques that are in wide use in the private sector that cities might employ to limit the possibility of theft:

- Implement an automated tracking system, such as an RFID (radio frequency identification) system, to track inventory;
- Track inventory at multiple points in the supply chain—each time that an item is moved, ensure that its location is known; and

- Limit the number of employees who have direct access to the inventory, including the number of employees who have direct access to editing inventory data.

Spotlight on: Petty Theft of Public Resources and Inventory

In 2015, the Atlanta Police Department arrested a number of Watershed Management employees for stealing city inventory. The Department of Watershed Management is a city agency responsible for providing drinking water and wastewater services to city residents and businesses. The employees were accused of selling Watershed equipment, such as copper, brass, and water meters, to recycling centers over a period of eight years. The investigation into the thefts began when the City Auditor's Office failed to find receipts for the missing equipment. In addition to tightening management and security measures, including limiting the number of individuals who can order, receive, and distribute equipment and requiring manager sign-off, Watershed Management planned to implement a $20 million barcode system to track when equipment is moved.[15]

F. Procurement Fraud

What are the corruption risks?

Cities procure myriad goods and services from private contractors and nonprofit providers, ranging from construction, to health and human services, to waste removal, to the purchasing of items like textbooks. The vendor selection process is a particularly high-risk area for fraud. Cities should pay careful attention to how contracts are awarded and to whom. Examples of procurement and contractor fraud include:

- Collusive bidding for city contracts, which involves agreements between multiple prospective contractors to undermine the competitive bidding process and inflate prices;[16]
- Overcharging the city for projects via payroll manipulation and/or change orders;
- Fraud related to the status and use of Disadvantaged Business Enterprises; and
- Nepotism or conflicts of interest in the selection of vendors or contractors—where decision-makers favor certain vendors or contractors because of family connections, friendships, or, for elected officials, because of campaign contributions previously made by the businessperson or his company.

How can cities use existing data sources to better target corruption risks?

Information about the vendors that compete for city contracts is a rich source of data for combatting procurement fraud. Many cities have centralized databases with information about the vendors that have held city contracts, which can be cross-referenced with other publicly available data—such as the federal "no contract list," city and state records of sanctions against corporations and nonprofits, records of political donors, and federal tax forms (e.g., Form 990 for nonprofit organizations)—to identify corruption red flags.

To minimize the risk of fraud at the time of vendor selection, consider the following strategies:

- When conducting background checks, search databases or lists of debarred vendors, such as the federal "no contract" list or city or state records, to identify vendors with a suspect history;
- Flag vendors that received poor performance ratings by city agencies for previous services provided;
- Look for instances of fraudulent use of **Disadvantaged Business Enterprises** (DBE)—businesses that claim to be women, minority, disabled, or veteran-owned—by the DBEs themselves or by prime contractors by comparing the date the company was DBE certified to the start date of the contract, and reviewing prior contracts to verify their DBE status;
- Look at the number of active contracts that a DBE has—a high number of contracts may mean they are not completing the work themselves and passing contracts through to other organizations or contractors;
- Compare stated capacity to contract utilization rates in contract performance;
- Review federal tax documents to identify organizations that are financially insolvent—this may indicate a higher risk of fraud; and
- For nonprofits, review Form 990s for further information about:
    - Prior government grants received, key officers, board members, and stated purpose;
    - Unexplained financial losses;
    - Problems with third-party audit reports;
    - Reported instances of fraud; and
    - Any lawsuits involving the organization (e.g., employees bringing wage and hours lawsuits).

Bids and proposals that corporations and nonprofit organizations submit to acquire contracts with city agencies can signal potential fraud—particularly collusive bidding. Cities that have electronic procurement systems, such as Health and Human Services (HHS) Accelerator in New York City, contain detailed and

historical information about bidders and awards, which, when properly analyzed, can lead corruption officials to identify suspicious patterns in procurement.

Proven strategies for leveraging proposal data to combat **collusive bidding** include:

- Implement software programs called "cartel screens"[17] that help analyze bids to identify patterns indicative of collusion, such as South Korea's Bid Rigging Indicator Analysis System.[18]
- Look for unusual bidding patterns (e.g., bids that are identical, very close to each other, or suspiciously far apart compared to previous tenders; or several bids that are the same percentage difference apart from each other)—these tactics are known as **complementary bidding**,[19] and are used by firms to ensure a particular bidder wins at an inflated price;
- Identify instances where multiple bidders adjust their prices at the same time and to the same extent, or bidders that submit identical bids for specific line items;
- Look for patterns in who bids for certain contracts based on location or type of work—if the same group of firms always bids, rotating who wins, this might be a sign of collusion;
- Look for identical errors in calculations or typos across multiple vendors' proposals;
- Flag cases where one bidder withdraws and becomes a subcontractor, or where a losing bidder becomes a subcontractor—this could mean the losing or withdrawing bidder was promised subcontractor status in exchange for colluding with the winning bidder;
- Look for frequent change orders that increase overall contract value;
- Look for unbalanced bidding, especially from incumbent bidders, where items not likely be needed are underestimated and items likely to be used are price-inflated; and
- Look for patterns of "lowballing"—e.g., one vendor is consistently the low bidder in a certain geographical area, or in a fixed rotation with other bidders.

Strategies for identifying **nepotism or conflicts of interest** in the procurement process include:

- Look for connections between bidders and the city employees making procurement decisions by identifying matching addresses, phone numbers, and zip codes, or connections on social media;
- Look for connections between bidders and estimators in government who decide how much a project or an item should cost—if a bidder appears to be consistently close to the independent government estimate, the bidder may be receiving inside information;
- Identify cases where a vendor's Employer Identification Number matches an employee's Social Security Number;
- Flag cases where numerous "sole source" contracts have been awarded to the same contractor, or cases where an open bid receives only one proposal; and
- Look for potential conflicts of interest, such as government employees with family members that work for bidding companies, or city employees that serve on the board of a nonprofit that does business with the city.

Spotlight on: Procurement Fraud

In 2016, John Bills, a former assistant transportation commissioner for the City of Chicago, was sentenced to ten years in federal prison for his role in a scheme involving the City's red-light camera contracts. In 2003, Redflex Traffic Systems Inc. was awarded a contract to install red-light cameras in Chicago's intersections. In the approximately ten years following this initial award, Bills used his influence to continuously expand Redflex's business with the City, resulting in millions of dollars for the company's installation of additional red-light cameras. In exchange, Redflex provided Bills with cash and other personal benefits, including meals, golf outings, rental cars, airline tickets, hotel rooms, and other entertainment. The case was investigated by the Inspector General for the City of Chicago with municipal, state, and federal partners.[20]

G. Vendor and Contractor Fraud

What are the corruption risks?

After a city contract has been awarded to a vendor, corruption risks remain, and vendor performance and compensation should be monitored throughout the duration of the agreement. Cities run the risk of being defrauded by vendors who submit inaccurate or misleading invoices for payment, or who collude with city employees to receive fraudulent payments.

How can cities use existing data sources to better target corruption risks?

Invoices submitted by vendors, and subsequent payments made to these vendors, provide information that can be analyzed to detect potential fraud. Many cities keep electronic records of invoices, and have a centralized system to track payments issued. Following the award of a city contract, pubic officials should remain on the lookout for the falsification or manipulation of invoices submitted for payment.

Cities can effectively review invoices with the following strategies in mind:

- Flag cases of unusually high hours, especially when reported to be logged over the weekend, or contracts that are significantly over-budget;
- Verify that the requested payment is consistent with how much work the contractor is doing on an hourly basis, and the progress being made on the project, to protect against intentional over-billing;
- Guard against payroll fraud by cross-referencing the names and addresses of all contracted employees using available information from the Department of Motor Vehicles (DMV) or Social Security Administration (SSA)[21]—each employee should have a unique taxpayer identification and address;
- Compare requests for reimbursement for goods and materials against the standard market price;
- For construction projects, look for frequently submitted change orders, especially those without the architect's approval;
- For recurring material contracts (*e.g.,* tiles), check the vendor's prior contracts to confirm consistent pricing;
- Flag increases in the volume of contractor claims for reimbursement, or changes in the appearance of claims forms;
- Confirm requests for reimbursement are accompanied by supporting receipts;
- Ensure payments are not being made to inactive or terminated city vendors;
- Confirm segregation of duties for approval, recording and custody of assets;
- Identify cases where the vendor's purchase order is later than the date of the corresponding invoice; and
- Investigate unusual patterns in charges for "overhead" or for general administrative expenses.

Spotlight on: Vendor and Contractor Fraud

In 2012, the URS Corporation, an engineering, design, and construction firm, settled with the Massachusetts Port Authority ("Massport") to resolve fraudulent invoicing during the renovation of Logan International Airport. Lawsuits filed against URS Corporation allege that a former employee submitted hundreds of false and inflated invoices, resulting in overpayment of more than $1.3 million. The employee in question, George Papadopoulos, the URS project manager for the Logan renovation, had submitted monthly claims to URS for thousands of dollars for personal reimbursements. Papadopoulos would inflate labor and other charges in his invoices, and URS would overcharge Massport and then reimburse Papadopoulos. To settle the lawsuit, URS Corporation paid over $3.3 million, which includes repayment of overcharges. Papadopoulos was sentenced to state prison.[22]

H. Fraudulent Legal Claims against the City

What are the corruption risks?

Fraudulent legal claims occur when individuals knowingly submit a false claim for payment or compensation, or knowingly make or use a false record or statement in pursuit of government payment. Fraudulent legal claims typically take the form of personal injury or property damage incidents,[23] in which the claimant seeks liability compensation from the city government. Individuals may submit fraudulent claims or collude with city claims officers to receive undeserved settlement payments. Claims officers are often overly-specialized in terms of the subject matter of the claims they handle, which may lead to collusion between the officers and the lawyers who file claims.

How can cities use existing data sources to better target corruption risks?

Officials seeking to identify this type of fraud should begin by examining the **claims** submitted and approved for payment by the city. The office of the city auditor or comptroller often oversees the review and settlement of these claims, and holds either electronic or paper records of each claim and the result.

Strategies for identifying fraudulent legal claims include:

- Look for approval of claims at amounts right below the city's cap—this could mean claims officers are knowingly approving false claims as part of a scheme with the claimants and/or attorneys;
- Compare the address and any other identifying information of the claimants and attorneys to the claims officer, to identify matches or trends that suggest a personal relationship;
- Use mapping techniques to identify the geographic areas where people most frequently file complaints—an unusually high number of complaints in a particular region or at a specific site could indicate fraud; and
- Examine suspicious trends in casework with certain attorneys or law firms.

In addition to reviewing data, there are certain management techniques that cities might employ to limit the possibility of fraudulent legal claims:

- Rotate claims officers so they have to move around to different substantive areas, forcing them to diversify their expertise; and
- Ensure random assignment of claims to claims officers.

Spotlight on: Fraudulent Legal Claims against the City

In 2014, the District Attorney of Philadelphia announced charges against 24 individuals who allegedly submitted false insurance claims totaling over $400,000. Over the course of seven years, the defendants staged slip and fall accidents and claimed to have suffered injuries. Claimants were encouraged to select a location where the defect in the sidewalk was not too noticeable. They were then instructed to go to the hospital or call an ambulance after a staged fall, make follow-up appointments with doctors, and engage in physical therapy to bolster their claims. [24]

I. Fraud and Corruption by Elected Officials or High-Level Officials

What are the corruption risks?

Public functionaries and contractors are not the only participants in public fraud and corruption. Fraud by elected or appointed officials can be difficult to detect and prosecute due to the status and prominent positions of these individuals. If left undetected, however, a culture that tacitly accedes to corruption squanders public resources and erodes community trust.

How can cities use existing data sources to better target corruption risks?

Information about an elected official's **donors, discretionary spending, and personal contacts** may lead to the identification of corruption. By identifying who has donated to an elected official's campaign, examining where the official chooses to allocate discretionary funding (in the case of city councilmembers), and determining who has frequent contact with the official—particularly lobbyists—cities can identify cases of potentially illicit activities, such as bribery. Note, however, that data work in this area can raise significant data privacy and even First Amendment concerns as it relates to campaign donations, so watchdogs must know and consider the relevant laws and regulations as they proceed.

In particular, cities can use the following strategies to identify corruption:

- Review an elected official's use of discretionary funds, such as city council member items—to identify suspicious patterns;
- Match campaign contributions data with vendor and lobbyists data;
- Cross-reference individual and corporate donations to elected officials with specific success later on (e.g., a businessman who has contributed to an official has obtained a sought-after contract or license, potentially benefiting from the relationship); and
- During a follow-up investigation, look at the data contained in the envelope information (including timestamps, length of call, number dialed, or email recipient) of an official's phone and email contacts during the application period for a city contract—look especially for suspicious contact between contractors and the officials overseeing selection and procurement.

Public officials must file **financial disclosure forms** on a regular basis to publicly disclose personal financial holdings, assets, and transactions, as well as information on income, gifts, and reimbursements. Evaluating these disclosure filings may reveal fraudulent activity. In particular, anti-corruption officials may look for improper disclosure filings, including a lack of specificity on matters subject to lobbying. In addition, "lifestyle checks" or unexpectedly high income reported on financial disclosure forms may indicate illegal activity. In reviewing disclosure forms, look specifically at information about the individual's positions and organizations, outside employment, and income.

Spotlight on: Fraud and Corruption by High-Level Officials

New York City Councilman Ruben Wills, was arrested and charged in 2014 with multiple counts of fraud and grand larceny. Wills was accused of stealing over $30,000 in public funds from the New York City Campaign Finance Board and the New York State Office of Children and Family Services. In both instances, Wills used a non-profit organization that he founded, NY 4 Life, to hide and redirect funds. In one scheme, when running for his seat on the City Council, Wills received $11,500 in matching funds. Wills used these funds to pay a shell company that supposedly translated and distributed campaign literature, while actually redirecting the funds to NY 4 Life. In another scheme, NY 4 Life was awarded a contract to conduct four public service projects, but only implemented one program, while Wills pocketed the remaining $19,000 in grant funds. Through a series of cash withdrawals, Wills used the money from NY 4 Life to pay for extravagant shopping expenses. The Wills case was led by the Attorney General's Office and the Comptroller's Office of New York State.[25]

III. BEST PRACTICES IN DATA MANAGEMENT

In order to capitalize on the power of data-driven approaches to revolutionize corruption investigations—or even to test their feasibility through pilot projects—city governments need meaningful, manageable data. A truism in computer science states: "Garbage in, garbage out." Data is only useful for computer-assisted analysis if it is reliable, consistent, and machine-readable.

An important step in this process is to clearly define the purpose of data collection. Knowing the intended use of a data source will help shape each stage of the process of data collection and management. And it will also help guide the decisions of **data reporters**, the people who input the data.

Data Structuring

For data to be digitized and made useful for data analytics, it must be consistently structured. In the case of qualitative data, this may require conscientious tradeoffs. Data that is more rigidly categorized is easier to input and compare, but may oversimplify situations or fail to capture significant details or distinctions.

One important trade-off is between standardization and flexibility. Counterintuitively, even simple sets of qualitative data can contain ambiguities that are hard to reconcile. For example, many individuals may reasonably type in the same address in varying ways:

- 15 Lovers' Lane, Apt. 1 A
- 15 Lovers Ln. #1A
- 15 Lovers Ln, 1a
- *Etc.*

Discrepancies can be minimized through internal transmission standards, the use of clear guidelines and standardized templates. Collection specifications should include data nomenclature, definitions, acceptable enumerations, and rules on **referential integrity** to ensure that changes to one data set do not unintentionally affect related sets of data. The specifications should also include rules for formatting common values such as dates, names, addresses, and organizational names. It is crucial that reporters, custodians, and managers of data are trained to follow these rules consistently and are given reference materials and channels to seek help or further guidance.

End-user software design may help reduce discrepancies further. A drop-down menu of options or series of binary "yes/no" questions may force users to choose from an array of acceptable inputs. An "auto-complete" feature or auto-check feature may flag improperly entered data. Users may be prompted to confirm entered data after it is translated into a standardized format, like the way that many e-commerce retailers prompt shoppers to confirm their address. The Python programming language (discussed below) converts addresses into machine-readable quantitative information as geocoordinates.

Data is useless without context. Thus, it is imperative to develop a **data dictionary**, a comprehensive directory of information about the structure and format of data in a data management system, using industry standards whenever possible. A data dictionary may include a catalog of codes that represent relevant information values. For example, the medical field has long used volumes of diagnostic codes, billing codes, and other classification systems to ensure that all medical professionals can communicate potentially critical information with the same vocabulary. If all participants input data using a shared vocabulary and format, the data is more consistent and machine-readable.

However, in some situations involving qualitative data, open fields are preferable for data entry. Forcing users to pigeonhole data into preconceived typologies may introduce distortions or biases by shoehorning complex information into predetermined value sets. It's often useful to include at least one open field for notes or comments so that uncategorizable information is not lost entirely.

For some purposes, it might be useful to develop a set of standardized tags to index data. Tags are non-hierarchical keywords or terms that serve as **metadata**, or information about data, making related data points easier to index for browsing or searching. Data sets like case records, incident reports, or networks of related individuals have trends that are easier to analyze if occurrences of select attributes or circumstances can be easily tracked. In a large set of tagged data, machine learning may be applied predictively to find latent relationships or trends. Quantitative tags, such as geotags (which convey geographical information) are designed for machine-readability.

When data collection is outside of an agency's control, in the case of data supplied by outside departments, it is especially important to track the source and provenance (or chain of custody) of each data set, to follow up on any inconsistencies or issues that emerge in later analysis. The data management team should keep a record of names of individuals from outside departments who supply information, as well as a master list of all computer programs or other data sources used in cleaning or otherwise manipulating the data. This list should be regularly updated, as data protocols may change rapidly as government operations are reorganized. Some questions to ask in assessing data received from an outside source include:

- Where did this data come from?
- How was is collected?
- Why was it originally collected?
- What kinds of biases might affect this data?
- Who is the **Point of Contact** for this data?

Perhaps the surest and most cost-effective way to ensure data accuracy is to make data sets public whenever possible. Watchdogs in the media, civil society, and general public can then do their own checks and data analysis to add extra sets of eyes to existing oversight.

Spotlight on: Open Data Standards

New York City's Department of Information Technology and Telecommunications maintains the Open Data Policy and Technical Standards Manual, an updated manual of technical standards and best practices to harmonize data collection and dissemination efforts citywide. The manual is a great resource for other cities seeking advice on issues raised above, such as reconciling addresses and building data dictionaries.

Data Cleaning

For government entities working to combat corruption and fraud through data analytics, it is crucial to ensure the data being analyzed is both accurate and usable. Data cleaning is a process used to correct and remove irrelevant and inaccurate records from a database. Common errors in a database include spelling errors, missing information, invalid entries, redundant entries, and inconsistent formatting.[26] Most data sets require some form of cleaning, from reordering columns or references, to removing duplicates, to combining related sets.

The process of data cleaning can be carried out manually or automatically. Manual data cleaning is generally used in simple cases and involves individual, line-by-line review of data entry to check for irregularities, duplications, and omissions. In more complicated, larger data sets, automatic data cleaning can be carried out with the help of computer software. Data cleaning software performs a variety of functions, including:

- Standardize information within a data set;
- Remove inconsistencies in data;
- Identify duplicate entries;
- Identify mismatched data entries;
- Transform data formatting;
- Identify and exclude missing fields;
- Maintain archives of raw data; and
- Provide a collaborative platform for users to interact with data.

Three commonly used software languages for data cleaning are **Python**, **R, and SQL**. Python and R are open-source, which makes them especially appealing to cost-sensitive users, due to the widespread availability of free resources such as online tutorials and documentation, adaptable modules, and developer communities. SQL, which can be integrated with Python and R, is optimal for relational databases (such as an organizational chart or a social network or any data stored in XML format), since it stores data relationally using metadata rather than in fixed columns. Those languages all allow for the application of interactive commands to process data, generating automated reports and graphics.

Below are some examples of more commonly used and well-reviewed data cleaning software programs. Before introducing particular software options, officials should understand the specific needs of their agency and define a clear set of goals for using the software:

- Ab Initio is used for both data integration and data management. A unique feature of this software is its Graphical Development Environment (GDE). The GDE components can be used to access data from a variety of sources, for example, web services. This extracted data can then be transformed and loaded to any given destination.
- Datapreparator provides a variety of tools to clean and transform data sets using a graphical user interface. The software does not store data sets on the computer hard disk and, therefore, can handle very large data sets.
- Datamartist is a tool for data examination. It can be used to combine and clean multiple large data sources into one data set and provide instant analysis of key indicators to assess data quality.
- OpenRefine is useful for transforming data formats. A significant feature of this program is that it allows users to link and extend data sets with other web services.
- Syncsort provides data cleaners with fast, high volume sorting. It can collect, integrate, and distribute big data efficiently.
- Pentaho Data Integration uses a consistent data cleansing approach that is repeatable and automated.

Regardless of the selected software, it is important to protect against the unnecessary loss of information. Prior to deletion of duplicate or invalid data, as a practical matter, the data cleaning process should include an opportunity to review and correct flagged entries. In addition, once a dataset has been cleaned, it is important to preserve its integrity by preventing the introduction of new "contaminated" data. Original data may need to be carefully preserved "as is" for legal purposes.

Finally, it should be noted that technological gains are helping to reduce the burden of data cleaning. For example, probabilistic data matching techniques can help to link similar entities in data sets that contain errors and inconsistencies, without undertaking more comprehensive data cleaning efforts.

Creating a Data-Oriented Organizational Culture

A government agency must integrate data collection, management, and usage into its organizational culture, to ensure that decision-makers have access to rich, reliable, and usable data.

Many municipal offices still use paper-based data records. Such records are easy to misplace or misread and costly to digitize. There are three basic options for digitization of paper records. The cheapest option is to scan them, which produces files in formats that are not machine readable, such as portable document format (.pdf). A more costly option is to use document reading software, which may be costly and error-prone, especially if the records are handwritten or otherwise difficult to interpret. Finally, records may be re-entered into a digital format, which can be labor-intensive.

Thus, it can be worth the upfront cost to create a digital system for data entry. Even better is a secure web-based system that allows data to be shared, updated in real time, stored securely (such as in a cloud-based system), and accessed remotely.

But even if an office adopts better data management practices for the future, *legacy databases* may still be in suboptimal formats. Organizations may need to carefully consider how to integrate legacy databases with more current sources, and how to ensure all data is accessible and properly contextualized.

Data reporters need to be included as an integral part of the data management process. They are the ones making frontline decisions about data input that may significantly impact results downstream. It is important that they be properly trained and informed about the purpose of data collection and given a channel to ask questions and offer feedback, since they may be best positioned to know if questions or categories don't accurately reflect the facts on the ground. Furthermore, their morale may suffer if they feel alienated from the process, which can lead to inconsistent or inadequate data entry. Anecdotal evidence suggests that city workers who understand that the data they input into a system will be used to help make decisions about targeting corruption investigations to find misuses of taxpayer funds may be more conscientious and resourceful than those who believe the data they input will just sit in an overlooked file, gathering real or virtual dust. Data reporters who feel alienated from the process or its outcomes may enter data with more inconsistencies.

End-users, such as key decision-makers, must be adequately trained and committed to data usage. Veteran investigators or policymakers who are unfamiliar with data-driven methods may be skeptical of their reliability and reluctant to integrate new tools into their workflow. Since data analytics work best in tandem with traditional investigative methods, these new tools won't produce results unless continuously tested and honed. Users need to know the purpose, advantages, and limitations of these methods, and their feedback must be incorporated into new iterations.

Because people tend to be best at interpreting visual information, data visualization is worth investment in design, engineering, and user experience. Without a clear, intuitive visual interface, users may overlook important results and trends.

Technological adoption often follows a steep learning curve, with corresponding upfront costs. Data reporters and end users need retraining, so that they are both capable and willing to deploy new methods. Engineers and designers need to field test and refine the technology and address unanticipated problems. Data structures need to be adapted to be more useful for data analytics, which may require re-entry, re-formatting, reinterpretation, and new checks for accuracy and consistency. All these changes may have second-order repercussions that require time and effort to address.

Thus, it may be easier to implement data-driven approaches as part of a broader revamp of technology, processes, and organizational culture. New agencies or agencies that are currently in transition, or

contemplating overhauling internal procedures, may be better positioned to adopt new methods than more established ones.

At its best, the usage of data analytics is an **iterative process**, a cycle of trial and error that leads to continuous improvement of methods and greater accuracy and reliability of results. Hypotheses should be tested and confirmed as new data is received. Core indicators must be routinely reviewed and reassessed. Employees and partners need to periodically query data sources and use curiosity and creativity to expose potential risks, biases, latent assumptions, or inefficiencies in data-driven processes.

Data reporters should be included in cycles of improvement as well, since they may be best positioned to know if the data captured accurately reflect facts on the ground. They should be given channels to provide feedback through an online form, e-mail address, a phone number, or smartphone application.

Increasingly, government organizations are creating data officer positions or teams to coordinate and improve data management policies and facilitate data sharing across offices, such as New York City's Mayor's Office of Technology and Innovation and Mayor's Office of Data Analytics, or Seattle's Chief Technology Officer, or Philadelphia's Chief Data Officer. A city's chief data officer, chief technology officer, or open data coordinator can play a crucial role in harmonizing data collection and implementing best practices. Such offices can be important partners for watchdog agencies.

IV. CAUTIONS AND CONSIDERATIONS

Data analytics can yield unclear and unreliable results when critical variables can't be accurately or consistently measured. Unfortunately, these are common challenges in monitoring for corruption. Agencies implementing data-driven approaches must be careful to consider relevant benefits and risks.

Where to House a Data Analytics Program

Careful consideration should be given to what component of city government will be responsible for conducting corruption-related data analytics and resulting investigations. If possible, cities should consider building internal capacity for this purpose, either within an existing city watchdog agency or otherwise, as this will be preferable in the long term to outsourcing. The least desirable option may be contracting with a private sector company to provide data analytics services of this sort to the city as this is likely cost-prohibitive in the long term and raises privacy concerns, discussed further below. If data capacity is built internally, cities must consider how success should be measured and put into place mechanisms for ensuring that goals are being met and that there is someone watching the watchdog, so to speak.

Data Access and Privacy and Bias Issues

Unlike many other areas of urban informatics, corruption investigations often involve data that is legally restricted due to concerns of privacy and confidentiality. Such concerns can be addressed through the conscientious stewardship of data, attention to cyber security, and careful data aggregation and disaggregation. After all, in tracing patterns indicative of corruption risks, individual identity may be less relevant than the various values or relationships attached to each data point. In the private sector, data aggregation to preserve privacy is a frequent strategy employed in areas with sensitive data such as healthcare, consumer behavior, and metering of utilities or usage of wireless networks.

Law enforcement agencies must be careful to safeguard the data they share with partners and set boundaries on data access or usage as necessary. Relatedly, the design of a process for data-driven investigative methods should account for differing levels of data access among users, and the extent to which certain data sets are access restricted. Is the data publicly available? Is it available only to certain users? Does it require a specific request, or even a court-ordered warrant? Can the data be accessed only on a case-by-case basis or is it available in aggregate? Legal restrictions may slow down, if not necessarily bar, usage of certain data sources, limiting their usefulness, portability, or scalability. Arguably, caution about such restrictions contributes to information "siloing" in city government, as agencies are reluctant to share data that may result in legal risks if inadvertently exposed or deliberately leaked.

Of course, law enforcement agencies tend to have measures in place to manage confidential data and the findings of their investigations are not intended for public access. But agencies must be careful to maintain confidentiality of data and comply with legal standards and requirements and local ethics codes. Ultimately, an agency that seeks to use restricted data in novel ways should seek advice and approval from relevant legal authorities, to avoid legal risks down the road. For example, the use of warrantless geolocation information or the monitoring of public employees may raise considerations untested in court, depending on the jurisdiction in question and specific circumstances at issue.

Agencies will also want to be sensitive to the concerns of protected classes and groups when conducting data analysis. Whenever historical data is used to make predictions, there is the possibility that the data encodes past biases that can be exacerbated through analysis. Even perceptions of such bias can be problematic, so cities should be careful to consider the lingering effects of structural discrimination when developing their programs.

The Reliability of Results

Outlier analysis is only effective if there is enough reliable data to establish a baseline or control of what values or data patterns would constitute "normal." In some data sources, this can be difficult as well, due to the lack of established benchmarks or known correlations of conditions and results. When it comes to restaurant inspections, for example, there may a wealth of historical data. But when it comes to big-ticket procurement contracts, the data points available may be too limited and hard to compare.

Even in systems that feature large-volume, multi-variable data sets, investigators must be skeptical and measured and wary of over-reliance on data. False positives and misleading trends can emerge from large data sets.

One of the greatest challenges in using data analytics is avoiding **confirmation bias**, the selective interpretation of data to reinforce preexisting beliefs or support preset theories. Data can be used to test hypotheses, or explored for new hypotheses, but interpretation must be carefully controlled to allow for alternate explanations or additional analysis. Chains of causation can be difficult to infer with confidence, without relying on other investigative steps or methods, and misunderstood correlations are a constant risk. Ultimately, agencies must ensure that the right expertise is deployed to evaluate the data and approaches used, without being biased by other objectives.

One important check used in data analytics is to ensure that results are not only verifiable, but **falsifiable**. In other words, elements should be included to flag results that differ from expectations. Disconfirmation can serve as a brake on the system, to indicate to an analyst that there may be issues with data collection, analysis, or interpretation. Results that are consistently either confirmatory or inconclusive are a signal that analysts' hypotheses are not being properly formulated or tested to allow for rival explanations.

Over-Reliance on Data

The flipside of the previous risk is that organizations may become too reliant on data. They may set their expectations too high for data-driven efforts, may use data as a crutch, and may allow their work to be driven by data collection—rather than the other way around.

Like the man in the joke looking for his lost car keys under a streetlight because that's where the lighting is best, investigative processes can err by focusing exclusively on areas with strong data, rather than the areas of most need or with the potential for greatest impact.

For example, imagine a data-driven approach that produces strong results for detecting theft of time by city clerical workers, because of the availability of rich time data about when such workers enter and leave restricted facilities or access online systems. This approach could be a great program to cut down on fraud and save the city money. However, the success of the program can lull investigators into a false sense of complacency. Are investigators ignoring similar fraud by non-clerical workers, whose time management data is less accessible? Are investigators cognizant of how employees might have strategically responded to crackdowns by developing crafty new workarounds that render data misleading? It is important for those responsible for anti-corruption monitoring to constantly evaluate and re-evaluate approaches, especially in light of information gathered through more traditional investigative approaches.

To avoid such risks, data-driven approaches are strongest when combined with human intelligence and shoe-leather investigations. Especially in cases in which data analytic projects are used to explore or test hypotheses, intelligence gathered through other methods are critical checks. For example, a program that checks for favoritism in hiring solely by looking at whether a given set of workers are neighbors or family

members may miss workers who were otherwise closely related, merely because such data was unavailable or veteran investigators weren't consulted about real-world cases or known risks.

Gauging Effectiveness

Data-driven approaches can entail considerable upfront costs. As a new tool, the cost-effectiveness of analytics can be difficult to measure, especially given the difficulty of measuring or assessing many indicators of corruption. Furthermore, since data-driven methods are designed to work in tandem with more traditional methods, their specific impact may not always be obvious or traceable. Ideally, those methods would provide useful leads to investigators, but if those leads include false positives, then are they truly time-savers? And if data-driven methods are meant to improve over time through a process of iteration, how long should it take for a pilot project to achieve cost-effectiveness?

Questions of cost-effectiveness are especially important given the limited resources available to watchdog agencies and municipal government generally, and the frequent need for such agencies to justify their budgets in a political climate that is often wary of bloated public-sector budgets, alarmed by escalating information technology costs, and skeptical of the value of oversight itself. Thus, performance assessment must be a priority for data-driven pilot projects, and project leaders and technical staff should work together to choose specific, appropriate metrics for periodic monitoring. Such metrics might include closed cases, the duration of open investigations, assets recovered, and qualitative assessments from data reporters and end users.

V. POSSIBLE NEXT STEPS

The proposals below are intended to be immediately actionable, leveraging existing knowledge and resources, to illuminate various potential pathways for future development.

Compile a database of corruption cases

Veteran practitioners often tap their knowledge of previous cases and successful investigations to spot trends and useful commonalities. Some even believe in a periodic cycle of corruption waves, as criminals are emboldened, develop workarounds to existing systems, trigger crackdowns, and then lie low. However, historical records of past cases handled by anti-corruption authorities are often ad hoc and uncatalogued.

A comprehensive database of known municipal corruption cases within one city, region, or the whole country would provide a rich trove of useful information that could be mined to track emerging trends and guide future efforts. Using a data dictionary, each case could be tagged with significant indicators about the circumstances of the case, sources of evidence and investigative methods used, or other relevant factors. Although such a database would require an upfront investment of time and resources, it could eventually provide the bedrock for novel applications of data analytics and testing of hypotheses about corruption trends.

Revamp vendor tracking

Currently, only a few major cities maintain comprehensive databases about vendors, with information about existing contracts, historical performance, corporate management and ownership, and other information relevant to corruption cases. One example is New York City's "VENDEX" database, a restricted-access manual-input system that can be unwieldly to search and to cross-reference with other data sources. The Mayor's Office of Contract Services (MOCS), which maintains VENDEX, is currently implementing a top-to-bottom overhaul of VENDEX, to make the system web-based, real-time, fully paperless, electronically-inputted, standardized, and fully searchable.

Fraud detection is not a primary purpose of VENDEX, but the database includes "cautions" noted by oversight agencies, which routinely search VENDEX entries in their investigations into waste, fraud, and corruption. With technical assistance and input from anti-corruption practitioners, a revamped VENDEX could pioneer a powerful engine for procurement fraud detection. The system could potentially be made more accessible, more reliable, easier to search, and better integrated with related data sources. Other cities could consider developing similar systems, especially those that have already begun to move procurement processes online.

Cross-reference data on nonprofit contractors

Nonprofit vendors already disclose valuable information through their tax forms, including information on budgeting, loans, management composition, lawsuits, and instances of fraud. In New York City, information is also collected on such vendors through VENDEX, including performance assessment and cautions logged by law enforcement agencies. Potentially, a program able to systematically cross-reference those databases could help identify nonprofits that are relatively risk-prone—due to red flags such as lawsuits, operational deficits, management turnover, records of underperformance, and other issues. Data could also be cross-referenced with databases of political donors or discretionary spending by public officials to pinpoint nonprofits to whom they are or were formally connected. Of course, such indicators are not dispositive, and may reveal many instances of organizational turbulence unrelated to integrity violations. Nevertheless, a data-driven approach could help investigators focus their efforts on higher-priority investigative targets.

Make public disclosure statements searchable

States and cities are making progress in moving towards paperless, web-based databases of public financial disclosure statements by political candidates, elected and non-elected public officials, and registered lobbyists. Such disclosures present crucial information that could be cross-referenced with other data sources to find connections between political leaders, lobbyists, and vendors. An investment of resources and technical expertise could allow oversight offices that collect this information to work together with local watchdog agencies to share this information more efficiently. For example, the Joint Commission on Public Ethics in New York State is moving towards a paperless system, yet currently provides financial disclosure forms as scanned pdf documents which are not machine-readable. A consistent online data entry system would make the data it collects much more usable, streamlining corruption investigations and allowing watchdogs to better track entries and relationships that merit further investigation.

Encourage data disclosure and sharing

The more data accumulated, the more opportunities there are for watchdogs to make actionable connections. Cities should consider establishing open data policies and pressure private companies to disclose useful data (aggregated to protect consumer privacy). The more data is open, the more that "civic hackers," civil society activists, journalists, and concerned citizens can participate in data collection and analysis. Cities might consider supporting "hackathons" or collaborating with academic institutions, private start-ups, and civil society groups to brainstorm and test new ideas to use data analytics to fight fraud and save money. Other ideas include holding conferences to more widely brainstorm and distribute best practices in this area, and developing an online repository of successful examples where corruption has been detected using data analytics.

VI. CONCLUSION

Data analytics is a promising new field with the potential to kick-start a transformation in how cities identify waste, fraud, and corruption. Data-driven efforts to monitor and investigate abuses of public trust can help watchdog agencies perform their duties more efficiently, more effectively, and more proactively. Particularly in the nine areas of common municipal corruption risks cited in this framework, cities already have some available data sources and indicators to begin to implement data-driven approaches. However, data collection and management processes need significant improvement and harmonization before the potential for those methods can be fully tested. Furthermore, cities seeking to make headway in this area need to consider potential pitfalls, such as privacy concerns and the dangers of confirmation biases and an over-reliance on data.

Cities interested in pursuing the potential gains of a data revolution in anti-corruption enforcement have a great incentive to collaborate. Despite differences in specific data sets and circumstances, problems of waste, fraud, and corruption echo across jurisdictions. Ultimately, cities should share data with each other and work together on honing indicators and developing stronger and more innovative data-driven methods. CAPI aims to serve as a knowledge hub for the public integrity community, for cities to share resources and lessons learned in the fight against corruption, including the new battleground of data analytics. To find out more about CAPI or to join our public integrity community, please visit our website at http://www.law.columbia.edu/public-integrity or contact us at capi@law.columbia.edu or 212-854-6186. We welcome any comments and feedback about how we can assist practitioners, scholars, and policymakers to promote public integrity and turn the tide in the joint struggle against corruption in communities across the country and around the globe.

APPENDICES

Appendix A: Research Methodology

To investigate the applicability of data analytics to the problems of municipal corruption control, CAPI assembled a Data Analytics Working Group (the "Working Group"). The Working Group met four times from March 2016 to February 2017 at Columbia Law School in New York City, to discuss questions such as:

1. What are feasible applications of data analytics to targeting corruption?
2. What data sources would be relevant, and how can they be accessed?
3. What are useful indicators of fraud or corruption among those data?
4. What tools could be developed, purchased, or adapted to glean actionable intelligence and useful information from those data?
5. What would be the specifications for designing such tools to cost-effectively meet agency needs?
6. What technical, logistical, or political challenges must be addressed to make those tools operational?
7. How can a pilot be developed and tested within particular jurisdictions in order to test feasibility?
8. What are realistic expectations for those tools, and by which metrics should they be evaluated?
9. How can such tools be designed to comply with applicable laws and legal standards and to address any concerns of citizens, public servants, and policymakers?
10. How can tools be shared across jurisdictions to minimize unnecessary duplication of efforts?

Work Group Members:

Jennifer Rodgers – Center for the Advancement of Public Integrity (chair)
Brian Browne - Ernst & Young
Aaron Chalfin - University of Pennsylvania and University of Chicago Crime Lab in New York City
John Curran – Walden Macht & Haran LLP (formerly Stroz Friedberg)
Jacqueline Eppolito - New York City Department of Investigation
Joe Ferguson – Inspector General of the City of Chicago
Stephanie Glive – New York City Department of Investigation
John Kaehny - Reinvent Albany
Andrew Kalloch – formerly Office of the New York City Comptroller
Gabriel Kuris - Center for the Advancement of Public Integrity
Calvin Lam - New York City Department of Investigation
Charlie Linehan – formerly the New York County District Attorney's Office
Elizabeth Marcello - Reinvent Albany
Jeff Merritt – New York City Mayor's Office of Tech and Innovation
Pasqualino Russo – Windels, Marx, Lane, and Mittendorf, LLP
Ravi Shroff – New York University Center for Urban Science and Progress
Marcos Soler – New York City Mayor's Office for Criminal Justice
Peter Weitzman - Stroz Friedberg
Milton Yu - New York City Department of Investigation
Ryan Zirngibl - New York City Mayor's Office of Data Analytics

At the suggestion of the Working Group, CAPI conducted additional semi-structured research interviews with municipal leaders, current and former law enforcement officials, data science scholars, civil society representatives, and private sector experts, including:

Stacy Aronowitz, Office of the New York State Attorney General
Salvador Arrona, NYC Business Integrity Commission, New York City
Patrick Blanchard, Inspector General of Cook County, Illinois
Andrew Brunsden, New York City Department of Investigation

Thomas Caulfield, Procurement Integrity Consulting Services
Pei Pei Cheng de Castro, New York State Joint Commission on Public Ethics
Gregory Cho, New York City Department of Investigation
Lisa Flores, Office of the Comptroller, New York City
Anthony Florio, Office of the Inspector General of the City of Chicago
Michael Flowers, Enigma.io
Ronald Goldstock, Pugh, Jones, & Johnson, P.C.
Stephen Hamilton, Inspector General for the Office of the New York State Comptroller
Noel Hidalgo, BetaNYC
Ahmed Jir, Office of the Inspector General of the City of Chicago
Dmitri Jones, Office of the Inspector General of the Metropolitan Transit Authority
Jamie Kalven, Invisible Institute
Lacey Keller, Office of the New York State Attorney General
Barry Kluger, Inspector General of the Metropolitan Transit Authority
Marjorie Landa, Office of the Comptroller, New York City
William Marback, Office of the Inspector General of the City of Chicago
Michele Mark-Levine, Office of the Comptroller, New York City
James McIsaacs, Department of Procurement Services, City of Chicago
Michael Owh, Mayor's Office of Contract Services, New York City
Steven Pasichow, Deputy Inspector General/Director of Investigations for the Port Authority of New York and New Jersey
Brian Peete, Office of the Inspector General of the City of Chicago
James Perazzo, Mayor's Office of Operations, New York City
Roy Pollitt, Exiger
Richard Ponce, Department of Finance, City of Chicago
Deirdre Power, Office of the Inspector General of the Metropolitan Transit Authority
Ashley Przestrzelki, Business Integrity Commission, New York City
Jesse Schaffer, Campaign Finance Board, New York City
Nicholas Schuler, Inspector General of Chicago Public Schools
Andrew Sein, New York City Department of Investigation
Matthew Serio, Office of the Inspector General of the City of Chicago
Andrew Smyth, Data Science Institute, Columbia University
Sherrill Spatz, Inspector General for New York State Courts
Sheryl Steckler, Procurement Integrity Consulting Services
Mindy Tarlow, Mayor's Office of Operations, New York City
Lise Valentine, Office of the Inspector General of the City of Chicago
Melissa Villa, Office of the Inspector General of the City of Chicago
Jennifer Way, Mayor's Office of Contract Services, New York City
Eugene Wu, Data Science Institute, Columbia University

In addition, CAPI extends its appreciation to **Bennett Midland, LLC**, and to interns and pro bono volunteers from Columbia Law School, for assistance with research and drafting of the framework.

Appendix B: Relevant Applications of Data Analytics

Data-driven approaches have already led to game-changing applications in municipal government, due to the density of information, manageable scale, concentration of resources, range of service delivery needs, and proximity between government and citizens. Data-driven policing has contributed to the rapid decline of crime in most major cities. For example, the New York Police Department's "CompStat" program to identify and map crime-prone hot spots and emerging trends has been adopted by dozens of cities nationwide, from Baltimore to Los Angeles. The Manhattan District Attorney's Crimes Strategies Unit has pioneered intelligence-driven prosecution strategies, such as sophisticated mapping of gang activity and weapons usage. The University of Chicago's Crime Lab, which recently expanded to New York City with support from the Arnold Foundation, uses data-driven insights to determine cost-effective interventions against violent crime. Private-sector firms are exploring the uses of data analytics to detect latent crimes such as human trafficking, insider trading, and insurance fraud.

In the field of procurement fraud, investigators and researchers abroad and in multilateral agencies have applied data-driven tactics to procurement contracts over the last few years:

- In 2013, the European Commission, together with Transparency International, developed the ARACHNE data analytics software to cross-check data from various public and private institutions and identify corruption risks, conflicts of interest, and irregularities. Data is collected from publicly available sources, including government reports and media.
- Two European anti-corruption research centers, Corruption Research Center Budapest and U4, built a python-based database of Hungarian public procurement records to study relationships between contracts, corporate ownership and management, and political officers. The project findings included potential indicators of corruption risk and indicators of undue political influence. The related Government Transparency Institute conducts innovative data-driven research into procurement and corruption risks within the European Union.
- The World Bank Group's Integrity Vice Presidency (INT) has partnered with the Data Science for Social Good (DSSG) program at the University of Chicago to preemptively avoid procurement fraud by analyzing bids for Bank-financed projects for suspicious patterns and behavior among competing firms.
- Brazil's Office of the Comptroller General (CGU) recently established a permanent office, the Public Spending Observatory (ODP), dedicated to using emerging data analytics techniques to prevent waste and abuse in public expenditures.

Domestically, data analytics has been applied by public and private-sector organizations to detect tax fraud and public benefits fraud. For example:

- At least 20 states use, or are developing, data-driven predictive approaches to target fraud in unemployment insurance programs. New Mexico, for example, used data analytics to cut unemployment fraud by 60% and is now testing a program that combines data analytics with behavioral economics.[27]
- Various states and municipalities are employing emergent data analytics techniques to catch tax cheats.[28]
- Data analytic techniques have been applied to review large healthcare claims and billing information to target fraud indicators and catch billing errors, "up-coding," and ghost patients.[29]
- Private companies like Paypal, FICO, and IBM have applied machine learning to fight fraud by identifying anomalous transactions using a nonlinear network of algorithms.[30]
- Palantir Technologies, Inc. has applied data analytics to forms of fraud including insider trading, fraud, tax fraud, tax evasion, and synthetic identity fraud, including for clients in the law enforcement field.
- Apache Hadoop produces machine learning software used to detect fraud in banking and healthcare.

Appendix C: Starting Points – Ten Ways to Use Existing Data to Fight Fraud

Of the many indicators of fraud discussed in this report, we recommend considering the following as a starting point for data-driven approaches. The below "Top Ten" approaches are relatively easy to implement and likely to unearth corruption more readily than some of the more complicated methods described in the body of the report.

Fraud by Inspectors

- Perform an outlier analysis to identify inspectors with the highest and lowest tallies of reported code violations.
- Search for code violations that were changed or downgraded following their initial logging.

Human Resources-Related Fraud by Public Employees

- Perform an outlier analysis to identify the employees who report the most overtime hours, including unusually high weekend overtime.
- Search for suspiciously consistent employee overtime claims (*e.g.,* always the same day of the week, or the exact same number of hours every time).

Benefits Fraud

- Search for transactions from Electronic Benefit Transfer (EBT) cards outside the zip code of the beneficiary's reported residence.
- Perform an outlier analysis to identify those employees who issue emergency benefits at the highest rates.

Campaign Finance Violations and Theft of Public Funds

- Perform an outlier analysis to identify campaign donations from out-of-district donor zip codes.

Petty Theft of Public Resources and Inventory

- Search for significant upticks in ordering, potentially indicating employees are selling excess inventory.

Procurement Fraud

- Implement cartel screen software programs to help identify collusive bidding.

Fraudulent Legal Claims against the City

- Search for approved claims at amounts right below the city's cap to identify possible knowing approval of false claims.

Endnotes

[1] Mikesell, John. *Study: Corruption Increases and Distorts Spending by the U.S. State.* Indiana University Bloomington, 2014. http://news.indiana.edu/releases/iu/2014/06/cost-of-corruption-paper.shtml.
[2] Weiser, Benjamin. *Three Contractors Sentenced to 20 Years in CityTime Corruption Case.* New York Times, 2014. http://www.nytimes.com/2014/04/29/nyregion/three-men-sentenced-to-20-years-in-citytime-scheme.html.
[3]Sherman, Ted. *NJ Corruption Case Finally Ends.* NJ.com, 2014. http://www.nj.com/news/index.ssf/2014/10/landmark_nj_corruption_case_ends_5_years_after_the_fbi_sting_operation_came_to_light.html.
[4] *Katrina: Four Years Later.* The FBI, 2009. https://www.fbi.gov/news/stories/2009/september/katrina_090109.
[5] *Manhattan DA's Office, DOI, NYPD Announce Arrests and Criminal Charges in Widespread Bribery Schemes Involving DOB and HPD Employees.* District Attorney's Office, 2015. http://manhattanda.org/press-release/manhattan-da%E2%80%99s-office-doi-nypd-announce-arrests-and-criminal-charges-widespread-briber.
[6] *Fraud Case Study- Timesheet Fraud.* http://www.dodig.mil/resources/fraud/pdfs/FraudCaseStudy_TimesheetFraud.pdf.
[7] Wenger, Yvonne. *Former City DOT Worker Indicted in Alleged Overtime Scheme.* The Baltimore Sun, 2014. http://www.baltimoresun.com/news/maryland/bs-md-ci-alleged-overtime-fraud-20141106-story.html; *City Employee Indicted For Fraudulent Overtime Scheme.* Office of Inspector General, 2014. http://inspector-general.baltimorecity.gov/news/inspector-general-press-releases/2014-11-16-city-employee-indicted-fraudulent-overtime-scheme.
[8] *Department of Transportation Audit on Overtime.* Los Angeles Times, 2015. http://documents.latimes.com/department-transportation-audit-overtime/.
[9] *New York City Human Resources Administration Supervisor Pleads Guilty to Defrauding Two Public Assistance Programs of More than $1.8 Million.* The United States Department of Justice, 2016. https://www.justice.gov/usao-sdny/pr/new-york-city-human-resources-administration-supervisor-pleads-guilty-defrauding-two; Muller, Benjamin and Nikita Stewart. *Two New York City Workers Charged With Stealing $2.1 Million from Benefits Programs.* The New York Times, 2015. https://www.nytimes.com/2015/12/02/nyregion/employees-arrested-in-scheme-to-defraud-human-resources-administration.html.
[10] *Limits and Thresholds 2017 Citywide Elections.* New York City Campaign Finance Board. http://www.nyccfb.info/candidate-services/limits-thresholds/2017.
[11] Malbin, Michael. *Citizen Funding for Elections.* The Campaign Finance Institute, 2015. http://www.cfinst.org/pdf/books-reports/CFI_CitizenFundingforElections.pdf.
[12] Berman, Russell. *Seattle's Experiment With Campaign Funding.* The Atlantic, 2015. https://www.theatlantic.com/politics/archive/2015/11/seattle-experiments-with-campaign-funding/415026/.
[13]*Former Campaign Treasurer and Fundraiser Found Guilty in Manhattan Federal Court of Campaign Finance Fraud.* The United States Department of Justice, 2013. https://www.justice.gov/usao-sdny/pr/former-campaign-treasurer-and-fundraiser-found-guilty-manhattan-federal-court-campaign.
[14] *Best Practices in Achieving Consistent, Accurate Physical Counts of Inventory and Related Property.* United States General Accounting Office, 2002 http://www.gao.gov/new.items/d02447g.pdf.
[15] Leslie, Katie. *Four More Arrests at Watershed.* AJC.com, 2015. http://www.ajc.com/news/local-govt--politics/four-more-arrests-watershed/3jdMEXbDiABNycKF6mfKqJ/; *Performance Audit: Department of Watershed Management Inventory Management.* City of Atlanta Auditor's Office, 2014. http://www.atlaudit.org/uploads/3/9/5/8/39584481/dwm_inventory_management.pdf.
[16] *Guide to Combating Corruption and Fraud in Development Projects*, 2017 http://guide.iacrc.org/potential-scheme-collusive-bidding/.
[17] Messick, Rick. *Procurement Agencies: Put in Screens!.* The Global Anticorruption Blog, 2017. https://globalanticorruptionblog.com/2017/01/11/procurement-agencies-put-in-screens/.
[18] *Fighting Bid Rigging in Public Procurement.* OECD, 2017. https://www.oecd.org/daf/competition/Fighting-Bid-Rigging-in-Public-Procurement-%202016-report.pdf.
[19] *Guide to Combating Corruption and Fraud in Development Projects*, 2017. http://guide.iacrc.org/potential-scheme-collusive-bidding/.
[20] *Former City of Chicago Transportation Official Sentenced to Ten Years for Corruption in Awarding of Red-Light Camera Contracts.* The United States Department of Justice, 2016. https://www.justice.gov/usao-

ndil/pr/former-city-chicago-transportation-official-sentenced-ten-years-corruption-awarding-red; *Examples of Public Corruption Investigations- Fiscal Year 2016.* IRS, 2016. https://www.irs.gov/uac/examples-of-public-corruption-investigations-fiscal-year-2016.

[21] Marasco, Jim. *Payroll Fraud: How It's Done, How to Prevent It.* Stonebridge, 2007. https://stonebridgebp.com/library/uncategorized/payroll-fraud-how-its-done-how-to-prevent-it/.

[22] *Construction Company Pays $2.3 Million to Resolve Fraudulent Overbilling in the Renovation of Logan Airport.* Mass.gov, 2012. http://www.mass.gov/ago/news-and-updates/press-releases/2012/2012-01-19-urs-settlement.html.

[23] http://www1.nyc.gov/nyc-resources/service/1654/file-claim-against-the-city.

[24] *46 People Charged in Elaborate Slip and Fall Scheme.* Philadelphia Office of the District Attorney, 2014. https://phillyda.wordpress.com/2014/04/24/46-people-charged-in-elaborate-slip-and-fall-scheme/.

[25] *A.G. Schneiderman and Comptroller DiNapoli Announce Indictment of NYC Councilman Ruben Wills in Public Corruption Scheme.* New York State Office of the Attorney General, 2014. https://ag.ny.gov/press-release/ag-schneiderman-comptroller-dinapoli-announce-indictment-nyc-councilman-ruben-wills.

[26] Rahm, Erhard and Hong Hai Do. *Data Cleaning: Problems and Current Approaches.* University of Leipzig, Germany. http://betterevaluation.org/sites/default/files/data_cleaning.pdf.

[27] *Construction Company Pays $2.3 Million to Resolve Fraudulent Overbilling in the Renovation of Logan Airport.* Mass.gov, 2012. http://www.mass.gov/ago/news-and-updates/press-releases/2012/2012-01-19-urs-settlement.html.

[28] Newcombe, Tod. States Use Big Data to Nab Tax Fraudsters. Governing, 2015. http://www.governing.com/columns/tech-talk/gov-states-big-data-tax-fraud.html.

[29] http://www.govhealthit.com/news/part-3-9-fraud-and-abuse-areas-big-data-can-target.

[30] Knorr, Eric. *How Paypall beats the bad guys with machine learning.* InfoWorld, 2015. http://www.infoworld.com/article/2907877/machine-learning/how-paypal-reduces-fraud-with-machine-learning.html; Nash, Kim. *FICO Takes Fraud- Detection Techniques to Cybersecurity.* The Wall Street Journal, 2016. http://blogs.wsj.com/cio/2016/02/17/fico-takes-fraud-detection-techniques-to-cybersecurity/; Goldschmidt, Yaara. *Using Machine Learning to Stream Computing to Detect Financial Fraud.* IBM Research. https://www.research.ibm.com/foiling-financial-fraud.shtml.

laura and john arnold foundation

Cover Design by Freepik