8

Accountability and Algorithms

The "news about the news" is often bad, particularly for investigative reporting.[1] The journalism that holds institutions accountable involves original work, about substantive issues, that someone wants to keep secret. These three distinctive features cause it to be underprovided in the marketplace. It is costly to create, hard to translate its social benefits into revenues, and easy to stymy by keeping information hidden inside government agencies.

The research presented in previous chapters puts number values on the nature of investigative reporting. The type of work submitted in IRE's annual prize competitions often takes six months to do. When the results are published, debates ensue, individuals may be fired, and, in some instances, new legislation gets passed. The case studies I developed show that each dollar invested by a media outlet in investigative work can yield hundreds of dollars in benefits to society when public policies change. Investigative reporting costing thousands of dollars can generate millions in benefits spread throughout a community. Pat Stith's career shows how, at its best, investigative reporting can consistently change laws and lives.

The results presented here suggest that the type of journalism that holds institutions accountable is increasingly at risk, particularly at the local level. Prize-winning work is more concentrated among a small number of outlets. In an era when prospective journalists face uncertain incomes and little time for training in the workplace, the average age of reporters winning Pulitzers for investigative work has increased by nearly a decade since the 1980s. Local newspapers are less likely to be home for reporters who share innovations and best practices in investigative work at conferences. Freedom of Information Act requests at a subset of federal agencies by local newspapers dropped by nearly a half between 2005 and 2010. The drop in IRE submissions during

the 2008/2009 recession came with a shift toward tales of individual wrongs, which are cheaper to produce than investigations into institutional patterns, and a drop in work about social justice stories like neglect.[2]

Newspapers accounted for more than half of the story submissions in IRE's prize competitions from 1979 to 2010, consistent with the role attributed to print outlets as a large source for investigative work. By almost any metric, however, the set of resources that newspapers can deploy has radically shrunk since the financial collapse of 2008. Between 2007 and 2014, full-time newsroom employment at daily newspapers dropped by almost 40%.[3] When looking at an overall decline in newsroom staffing of about 18,000 journalists, Ken Doctor noted that the disappearance of these staffers from local daily newspapers was equivalent to a "cumulative loss of about 216,000 years of local knowledge."[4] Newspaper ad revenue went from $37.8 billion ($3.1 billion digital advertising, $34.7 print advertising) in 2008 to $19.9 billion ($3.5 billion digital, $16.4 billion print) in 2014.[5] The dismal economic prospects for newspapers brought large declines in the market value of papers. Translating sale prices for newspapers into 2014 dollars, the Pew Research Center calculated the change in valuation of major daily newspapers that were sold across two decades: the *Boston Globe/Worcester Telegram & Gazette* dropped 96% in value between 1993 and 2013; the *Philadelphia Inquirer/Philadelphia Daily News* declined 91% from 2006 to 2012; the *Chicago Sun-Times* valuation dropped 93% between 1994 and 2011; and the *Minneapolis Star Tribune* valuation dropped 95% between 1998 and 2012.[6]

The same technologies and trends that caused revenue prospects for original content creation by print newspapers to plummet also created an abundance of opportunities in other parts of the information environment. Once digital content was created, tech firms and social media networks made it easier than ever to search, aggregate, curate, and share data. In 2014, about two-thirds of Twitter users and a similar percentage of Facebook users indicated that they got news from these social media platforms.[7] Tweets and Facebook posts became a way to share stories, and a way for reporters to discover new ideas, events, and sources. For those setting out to cover public affairs, the Web offered cheaper and quicker ways to access data, video, and sentiments. As in print and broadcast markets, however, the demand for information of interest to voters tended to be overwhelmed by consumer, worker, or entertainment demands. In a study of news shared on Facebook,

researchers found that, for news story URLs shared by Facebook users, 13% were "'hard' (such as national news, politics, or world affairs)." The rest were classified by the researchers as "'soft' content (such as sports, entertainment, or travel)."[8]

Assessing overall whether a given type of accountability coverage is more likely now than in 2007 would depend on factors such as the type of institutional breakdown to be discovered, the geographic scope of the community affected, the demographics of the potential audience for the tale, availability of and access to data, and the incentives for public or private actors to shield their actions from view. The relevant question, though, is not how to preserve a specific number of investigative reporting positions, but rather how to support the accountability function traditionally performed through investigative journalism. In this chapter, I explore two ways to support the watchdog function—changes in government policy, and expansion of the field of computational journalism.

Three changes in government policy would readily reduce the hassle costs involved with investigative work: (1) real reform of FOIA laws; (2) changes in federal research and development policies and IRS nonprofit rules, both of which would flow from recognition of the market failures involved with public affairs reporting; and (3) truer implementation of open government and transparency policies. FOIA reform at the federal level would include new legislation to codify a presumption that documents and data are open, a central portal to make requests across agencies easier, and support for agencies to share previously released documents. At the state level, reforms would include expanding parts of government covered by FOI laws, speeding up reply times, and reducing fees placed on journalists requesting information.

If policymakers acknowledged that reporting about public affairs is underprovided by the market, decisions across many different agencies could support investigative reporting. Recognition of national defense, education, and environmental protection as public goods means that federal programs supporting research and development are aimed at furthering these goals. Recognition of journalism as a public good worthy of support could increase the likelihood that the National Science Foundation would support research on open-source software to aid reporters. Journalists face multiple hurdles in turning unstructured information into structured data for analysis.

Adding journalism as a field to support in government research competitions would spur advances in algorithms to discover and tell stories. Recognizing the educational value of public affairs reporting would also lead the IRS to make it easier for newly formed online sites that provide hard news and investigative work to gain nonprofit tax status.

While open government and transparency policies offer rhetorical support for accountability, the reality of their implementation often tips more toward provision of information that builds businesses or solicitation of suggestions via the wisdom of crowds. Making data available to increase scrutiny of government decisions is not often a priority. Truly open government would entail the proactive release of the artifacts of governing—calendars, contracts, e-mails, and records. Designing government information systems to make public access to data easier up front would also cut down on later costs of redaction and retrieval. Changes to open government policies, real FOIA reform, and alteration of research and development policies have many advantages as media policies. They do not favor any particular medium or direct support to any specific topic in public affairs coverage. They hold the prospect of lowering the costs of discovering stories by making it harder for agencies to hide the raw data of governance. Their impact would also be multiplied by advancements in computational journalism.

Computational journalism, like the emerging field of data science, does not have a readily agreed-upon definition. It is likely to be defined by the set of practices of reporters using increasingly larger data sets and more sophisticated algorithms to find patterns. One working definition for computational journalism is the combination of data, algorithms, and knowledge from social science to supplement the accountability function of journalism. A wider take involves the interaction of computation and journalism to change how stories are discovered, presented, aggregated, monetized, and archived. Computational journalism may entail stories by algorithms (e.g., content generated by code), through algorithms (e.g., patterns and breaks surfaced through software), and about algorithms (e.g., investigations of how algorithms work and impact decisions).

Advances in computational journalism could improve the economics of accountability journalism in two ways. On the supply side, new ways of combining data and algorithms could lower the costs of discovering stories. Investigative work is akin to drilling for oil—the prospects of whether you find

what you're looking for are uncertain, and the setup costs can be large. Progress in making sense of new and larger data sets may allow reporters to spot outliers, find miscreants, and connect patterns to identify which institutions are breaking down. On the demand side, telling stories in more personalized and engaging ways for readers or viewers can increase the likelihood of revenues through subscription or advertising. Computation may give you an edge in originality, either in finding stories or conveying them. Offering a distinctive take brings the benefits of product differentiation in a market often swamped with commodity-like facts.

Policy Portfolio

The First Amendment offers a clear but impossible charge to U.S. legislators: "Congress shall make no law . . . abridging the freedom of speech, or of the press."[9] Congress in fact makes many laws that affect freedom of speech and the press, as an inevitable part of defining property rights to information. Laws about copyright, government secrecy, transparency, advertising, campaign finance, and nonprofit activity all can involve restrictions that change the calculus of media outlets of what they include in their output. Budget decisions about funding government information officers, public media, information infrastructure, and even access to phone lines and broadband all implicate what types of speech will be subsidized in content creation and distribution. In a world where government actors make decisions that inevitably affect media operation, and where investigative reporters hope to hold public officials accountable, a major challenge involves how to get government officials to pass and implement policies that ultimately increase scrutiny of their actions and thereby reduce their decision-making discretion.

The Freedom of Information Act passed Congress in 1966 not because of a groundswell of support from voters concerned about transparency, but rather from a desire from legislators to open up the actions of an expanding executive branch to more examination.[10] Journalists supported and benefitted from the legislation, which helped lower the costs of discovering and pursuing stories by changing the property rights to many types of documents held by federal agencies and departments. The right to get copies of paper, and now electronic documents, however, does not always translate

into the reality of receiving the data, since the transfer depends on the willingness of the government agency receiving the FOIA request to comply readily with the law. On his first day in office, President Obama issued a Memorandum on Transparency and Open Government that stressed federal agencies should be transparent, participatory, and collaborative.[11] Yet, in 2015, three analyses of implementation of FOIA by the federal government still found striking levels of noncompliance. Assessing the fifteen federal agencies that account for 90% of FOIA requests, the Center for Effective Government gave ten of the agencies a grade of D or F for timely and consistent responses to FOIA.[12] A National Security Archive analysis found that, "Nearly 20 years after Congress passed the Electronic Freedom of Information Act (E-FOIA), only 40 percent of agencies have followed the law's instruction for systematic posting of records released through FOIA in their electronic reading rooms."[13] A March 2015 article entitled "Obama Administration Sets New Record for Withholding FOIA Requests" began with these assessments:

> The Obama administration set a record again for censoring government files or outright denying access to them last year under the U.S. Freedom of Information Act, according to a new analysis of federal data by The Associated Press. The government took longer to turn over files when it provided any, said more regularly that it couldn't find documents and refused a record number of times to turn over files quickly that might be especially noteworthy. It also acknowledged in nearly 1 in 3 cases that its initial decisions to withhold or censor records were improper under the law—but only when it was challenged. Its backlog of unanswered requests at year's end grew remarkably by 55 percent to more than 200,000.[14]

Real FOIA reform would be an effective policy change to bolster monitoring of government activity, which could be performed by journalists, NGOs, or individuals requesting and using government records.[15] The problems investigative reporters currently face in pulling data from federal agencies are evident in the suggested policy changes set out in FOIA reform bills in 2015. Both House and Senate FOIA measures would codify a presumption of openness for federal documents and data. This would re-

quire agencies that wanted to deny requests to cite specific, foreseeable harms from the release or the exact legal requirements that prevent their release. The codification of the openness presumption would also make it difficult for future presidents to relax compliance with FOIA. One review of the reform legislation notes: "The bills would also put a 25-year sunset on agencies' ability to withhold records reflecting internal deliberations or other privileged communications. Openness advocates complain that the 'deliberative process privilege' is often interpreted so broadly that it obscures virtually all of the inner workings of government that FOIA [covers]."[16] Proposed reforms would also create a central FOIA Internet portal where filers could submit a FOIA to many agencies at once, and would change fee use by departments. While the prospects for FOIA reform are often uncertain, particularly given concern from financial regulators that banks would be less willing to share information if it were easily discoverable via FOIA, improvements are feasible even absent legislative action. In 2015, for example, seven federal agencies began a pilot program of putting online the documents and data they release under FOIA.[17] This "release to one is release to all" endeavor, which included the EPA and parts of the Department of Homeland Security and Department of Defense, would make it easier to access information already released (though there would be an exemption for individuals requesting their own files).

Freedom of information laws at the state level vary in terms of what parts of government are covered, deadlines for timely provision, and the degree that requesters have to pay fees related to search costs and document or data copying. Often, state government agencies appear to use the prospect of high fees to discourage journalists from requesting data. Each March during Sunshine Week, news outlets share examples of successes and failures in freedom of information requests at the federal, state, and local levels. The Associated Press's (AP) recounting in 2015 of its experience with state fees shows the hurdles its reporters face in pursuing government investigations.[18] In California, the Department of Motor Vehicles quoted a minimum fee of $19,950 to respond to a request for data that would allow a reporter "to determine whether poor people had their driver's licenses suspended at a disproportionate rate"; the AP dropped the request. In Colorado, the Department of Law quoted the AP a fee of $350 to estimate how much it would cost the agency to provide documents about state communications with

the federal government about marijuana. In Kansas, the Department of Aging and Disabilities Services estimated it would cost $600 to generate copies of e-mails from the governor and two others about Medicaid. In Oregon, the state police requested $4,000 to cover twenty-five hours of search and review time for a request related to the director of a commission that regulated boxing and martial arts (in whose office $22,000 had been discovered). The AP dropped the request. In Virginia, Governor Terry McAuliffe's office wanted $500 for copying ten months of his daily calendar, a document type routinely provided for free to the AP by California's Governor Jerry Brown.

The documents and data provided under state FOI laws allow reporters to cover a wide range of government actions. Consider the experience of *The State,* a local newspaper published in South Carolina's capital, Columbia. Like many news outlets located in a state capital, *The State* treats coverage of state government as local news. The headlines of stories made possible by the paper's use of South Carolina's FOI law show the type of scrutiny made possible by government documents and data: "Expenses Quietly Inflate Lawmakers' Salaries"; "S. C. Trooper Has Record of Complaints; Driver Has Driving, Marijuana Convictions"; "EXCLUSIVE: Accused Schemers Courted Columbia City Council Members"; "40% of SC Child-Welfare Workers Bear Heavy Caseloads"; and "How DHEC's [Department of Health and Environmental Control's] Oversight Fell Short."[19] As is true in other states, South Carolina's original Freedom of Information Act (passed in 1974) has been amended over time to make clearer what information is required to be turned over. In the case of South Carolina, this meant changes to the law to clarify that data on the salaries of public officials and crime reports created by police are all public records required to be released. There are still gaps in coverage, with *The State* noting in 2015: "There is no law that requires top local and state agency officials to use state email accounts that could be subject to Freedom of Information requests. And some officials still stonewall legitimate requests for public information, forcing time-consuming and expensive lawsuits."[20]

Accountability stories about state and local government would be easier if the search costs and copying fees charged were lower and if the set of documents clearly covered were expanded. State press associations, newspaper editorial page writers, and NGOs support improving the implementation of

state FOI laws. But the experience of the South Carolina Press Association's Freedom of Information Committee is typical of reform efforts. When the press association put together a task force on how to make FOI enforcement more likely (in a state where there is not a central agency with authority to act on complaints about FOI implementation), attempts to strengthen the law's enforcement failed.[21]

Beyond freedom of information laws, there is a large set of government policies affecting the accountability function of journalism that could change if the market failures inherent in investigative reporting were acknowledged. At the Federal Communications Commission (FCC), there is a long history of commissioners denying that the public goods and positive externalities problems in information markets exist. FCC chairman Mark Fowler in the 1980s called television "a toaster with pictures," and famously said of people's preferences in the marketplace: "The public's interest . . . defines the public interest." When asked in 2001 about the digital divide, FCC chairman Michael Powell declared, "I think there's a Mercedes divide. I'd like one, but I can't afford it."[22] These assessments imply that there is nothing special about the operation of media markets that requires interventions based on content deficiencies. The Waldman Report released by the FCC in 2011 did focus on journalism as a public good and effectively outlined the content gaps arising from difficulties in public affairs coverage. Yet even in this comprehensive report of 468 pages, the word "market failure" was only used once, in a footnote.[23]

Part of the tumult in news markets is clearly evidence of "creative destruction," the phrase economist Joseph Schumpeter used to describe the changes in economic structures that can come from industrial innovation, technical change, and the opening of new markets. The drop in advertising revenues at legacy media outlets; rapid expansion of social media networks; birth of online news sites staffed by digital natives; explosion in search, aggregation, curation, and expression; triumph of mobile devices as a way to navigate and fill daily life; and drop in the ranks of print reporters can all rapidly decide the winners and losers in the news world. Policymakers may understandably wish to avoid subsidizing firms and formats being replaced by more efficient approaches to information delivery, and in a world of quickly evolving technology, may doubt their ability to intervene in a timely and effective manner.

Against the backdrop of technical advances and cyclical booms and busts, there remains the real phenomenon that news markets for public affairs

continue to involve the market failures associated with public goods and positive externalities. The three essential elements of investigative reporting—original content, positive spillovers on society, a desire by institutional actors to impose transaction costs on discovery—all particularly discourage the provision of accountability reporting in the market. Important stories that would readily pass a social cost-benefit test go undiscovered and untold because they are expensive to produce and because media outlets are not adequately rewarded via advertising or subscriptions for the changes brought about by these stories. If policymakers acknowledged the existence of market failures in investigative reporting, this would open up a wide range of government actions that could support the accountability function of journalism. These would include changes in policy affecting federal support for research and development, nonprofit tax policies, and open government operations.

A journalist with a cache of e-mails, stack of forms, set of audio and video files, and collection of electronic documents faces a challenge familiar to many academic researchers and government analysts: How do you transform unstructured information into structured data to analyze? Many parts of the federal government have funded research and development projects to aid the analysis of data in particular policy areas. The Defense Advanced Research Projects Agency, Department of Defense, and National Science Foundation (NSF) have supported the development of analytical software aimed at providing one type of public good: national defense.[24] The Library of Congress (LOC) and National Archives and Records Administration (NARA) have funded innovative work to further education about the operation of government. The National Endowment for the Humanities and the NSF have provided financial support to jump-start the development of fields such as digital humanities and e-government. If the software supported by this federal funding were released in open-source formats, it could be modified to help investigative reporters. Yet much of the research ends up supporting algorithms that remain inside national security agencies, becomes the basis for commercial software with prices beyond that affordable by journalists, or never progresses beyond academic prototypes.

Reporters face a set of readily described problems in monitoring government actions that could be aided by inventive algorithms: deciphering scanned documents with frequent redactions or scrawled handwriting (a

challenge for Optical Character Recognition software); transcribing, indexing, or searching audio or video recordings of events such as city council meetings, state legislative committee hearings, or court proceedings (tasks made more difficult by the challenge of multivoice recognition); pulling data into a spreadsheet from forms provided as pdfs; mining documents to recognize entities, topics, and sentiments and how these cluster (a challenge in examining debates, rule makings, press releases, and news coverage); and combining information from multiple digital streams to track an entity or event over time and across sources to find new actions or decisions.[25] Solving these problems for journalists would increase government accountability, but the agencies with research and development grants and contracts rarely see aiding reporters as part of their research mission. If organizations such as NSF, NARA, and the LOC added public affairs reporting as a public good to be supported through their work, then the computational challenges faced by journalists would more readily attract the attention of academic and NGO researchers.[26]

Recognition that investigative reporting involves the market failures of public goods and positive externalities could also change how the Internal Revenue Service (IRS) treats nonprofit media. A long list of media outlets operate as nonprofits under 501(c)(3) of the Internal Revenue Code, including *Consumer Reports, Mother Jones, National Geographic,* and local public radio stations. In the wake of the financial collapse of 2008, the IRS started to receive more applications from organizations focused on local public affairs coverage and hoping to operate as nonprofit media (most often online). A 2013 report from the Council on Foundations Nonprofit Media Working Group found evidence that the IRS had significantly delayed granting nonprofit status to these local news outlets. Those experiencing delays in securing 501(c)(3) status included the *San Francisco Public Press, Chicago News Cooperative* (which ceased existence before approval was ever received), El Paso's *Newspaper Tree,* and *The Lens* of New Orleans. Difficulties in receiving the nonprofit status in turn hindered their ability to secure donors and foundation funding. In communications with applicants, the IRS officials often expressed a reluctance to declare nonprofit news outlets as educational, though tax regulations define educational as "the instruction of the public on subjects useful to the individual and beneficial to the community."[27] In order to secure its nonprofit status, the Investigative News

Network (now called the Institute for Nonprofit News) was required by the IRS to "remove the word 'journalism' from the 'purpose' clause in its articles of incorporation."[28] After the IRS informed the editor of a local public affairs newspaper and Web site in Rhode Island, "While most of your articles may be of interest to individuals residing in your community, they are not educational," the organization suspended publication in the absence of a 501(c)(3) designation.[29]

The Council on Foundations report argued that investigative reporting deserves support because of the public goods and positive spillovers generated by this type of journalism. In describing why public affairs coverage is sometimes underprovided in the market, the report noted that high-cost accountability journalism includes "Investigative pieces that require lengthy documents and records searches. Stories where government officials actively resist the disclosure of records and information. Stories that involve knowledge best gained through beat reporting, since understanding some policy areas involves spending time observing a set of institutions and issues. . . . Local accountability journalism is of great civic importance and value, but does not generate significant consumer demand to fuel healthy media business models."[30] The report urged that, in evaluating future applications for nonprofit media status, "the IRS should evaluate whether the media organization is engaged primarily in educational activities that provide a community benefit, as opposed to advancing private interest, and whether it is organized and managed as a nonprofit, tax-exempt organization."[31] Despite the attention focused on this issue by the Council on Foundation report, Waldman report, and the Federal Trade Commission staff proposal draft on policies to support the reinvention of journalism, the IRS to date has not changed its rules to reflect that nonprofit media focused on public affairs reporting can clearly be educational and eligible for 501(c)(3) status.[32]

At first glance, open government and transparency policies appear to foster the ability of many different actors—voters, NGOs, reporters—to hold public officials accountable. President Obama's open government policies are explicitly based on making government more transparent, collaborative, and participatory. The implementation of open government policies at multiple levels of government, however, often focuses on crowdsourcing of expertise and suggestions to aid government decision making, pushing government data out that can be used to build businesses (so-called DC to VC [Venture

Capital] approaches), or making data available that facilitates delivery of government services. These uses of transparency generate significant gains for society.[33] Making information available that would generate government accountability as a public good, however, is less of a priority. In an essay entitled "Transparency and Public Policy: Open Government Fails Accountability," Pulitzer Prize–winning reporter Sarah Cohen noted that, after President Obama's first term, much remained opaque in an avowedly transparent government:

> certain records that are widely acknowledged as public in the spirit of FOIA remain locked within virtual and physical cabinets. Few public officials make their desk calendars public, despite repeated rulings at the federal level that they are open for inspection. Obtaining basic spending documents, such as contracts, grants and purchase orders, is usually a two-year effort, thanks to provisions that the recipients can review the documents and censor information they consider sensitive. Even basic records that most cities have long ago released remain almost impossible to obtain elsewhere.[34]

Cohen observed that transparency policies sometimes resulted in the creation of two sets of books, one used internally by an agency and another created to share with the public.[35]

One way to see how the implementation of open data policies at the federal, state, and local level could be improved is to compare a given agency or area's approach with a template of best practices. The Sunlight Foundation's *Guidelines for Open Data Policies* provides an outline for how to determine which data to be made public, ways to share the information, and how to implement policies. Suggested open data policies that would particularly help reporters engaged in accountability work include:

> Proactively release government information online. . . . Create a public, comprehensive list of all information holdings. . . . Stipulate that provisions apply to contractors or quasi-governmental agencies. . . . Appropriately safeguard sensitive information. . . . Mandate data formats for maximal technical access. . . . Require publishing metadata. . . . Publish bulk data. . . . Create public APIs for accessing information. . . .

> Create or appoint oversight authority. . . . Create guidance or other binding regulations for implementation. . . . Ensure sufficient funding for implementation.[36]

In comments prepared for a Federal Trade Commission workshop on journalism's challenges, Sarah Cohen specified a number of ways transparency policy could change to facilitate public affairs reporting. She noted that, instead of waiting for FOIA requests, agencies could proactively release "records like correspondence logs, desk calendars of cabinet and sub-cabinet level officials, payroll records of political appointees and contracts, grants and their audits."[37] She also recommended that governments consider openness in the design of new or revised information systems so that data were easier to extract for public access. This would reduce the need for redactions and lower the costs of releasing many types of data, including "personnel records, calendar and email systems, spending records, client records, inspection and compliance records and benefit records."[38]

Open data help many different types of actors hold government accountable. Nonprofit Web sites built on open data include those that visualize Chicago zoning (secondcityzoning.org), report grants made in Detroit (detroitledger.org), and describe crime trends in San Francisco (sanfrancisco.crimespotting.org).[39] Tools for transparency from for-profit sites include data on spending by New York City (Checkbook NYC 2.0) and an app to report and track responses to public complaints about neighborhood problems (seeclickfix.com). Open data also give investigative reporters new ways to discover and tell stories. WAMU, a public radio station in Washington, DC, used open data to demonstrate that developers contributing in city council races were receiving significant subsidies from the government. A reporter from the *Chicago Sun-Times* used data from Chicago's Plow Tracker site to demonstrate the heavy attention that snowplows gave to the street where the city council's most senior member lived.[40] ProPublica used information from the Centers for Medicare and Medicaid Services to provide Yelp with government data on hospitals, dialysis centers, and nursing homes, which Yelp in turn added to its health care review pages so that it was accessible to the review site's millions of users.[41]

These proposed changes in public policy to support investigative reporting have the desirable property that they are content neutral and platform

agnostic. Policies to lower the cost of story discovery related to FOIA, open government, and support for computational research are neutral in the sense that they are not aimed at increasing coverage of a particular policy agenda issue or promoting a given political viewpoint. The tools and data can be used by journalists, NGOs, and individuals, who can communicate their findings via print, broadcast, or the Web. The public goods nature of a story, whose facts can circulate freely across many platforms once created, means that aiding story discovery can support accountability stories in any medium.

Reporters do change coverage depending on the costs of assembling a story. Consider the case of information about air and water releases of chemicals by manufacturing plants in the United States. Prior to the passage of the Emergency Planning and Community Right-to-Know Act of 1986 (EPCRA), firms could keep the nature and size of their chemical releases private. For a subset of chemicals, EPCRA required firms to report each year to the U.S. EPA their releases and transfers of these toxic substances. The data were made available to the public via the Toxics Release Inventory program (TRI), and for the first time in a regulatory program, Congress required that information be made available to the public via an electronic database. The result: reporters wrote stories about pollution revealed in the TRI, firm stock prices dropped in reaction to the information, and companies with greater stock reactions engaged in more reductions.[42] The scrutiny generated by the release of the data varied systematically. Reporters were more likely to write about a company's TRI releases in the first year of the program (1989) if air emissions were higher or more chemicals were involved, and less likely if the pollution was spread across many plants or if the company was already known to be in a pollution-intensive line of business.[43] For a given level of pollution, firms were more likely to reduce where emissions posed a greater health hazard, but less likely in communities where residents were not politically active.[44] Firms were less likely to report their releases accurately for heavily regulated substances.[45] The case of the TRI shows that reporters will use newly released government data to generate stories, though the impact of the scrutiny will depend on factors such as what prior information circulated about the topic, the concentration or dispersion of the activity to be written about, and the political and economic processes where the scrutiny is generated.

Many of the policy changes that would make government officials more accountable concentrate costs on those officials and widely disperse benefits across voters, which can make their passage or implementation less likely. Legislators may take credit for making their records available, but the actions can be symbolic rather than real if the data come wrapped in transaction costs. Records of contributions to Senate campaign committees illustrate symbolic transparency. While candidates for the House and the presidency file contribution records electronically with the Federal Election Commission (FEC), the upper chamber takes a different approach: "the Senate for years has chosen to maintain its archaic system, which works like this: Candidates send paper copies of their campaign reports to the Secretary of the Senate, who then submits them to the FEC, which has to pay a contractor to key the data into its electronic systems. This process is estimated to cost up to half-a-million dollars each year, and it delays availability of the public records by weeks, or sometimes months."[46] In the fall of 2014, this meant that candidates in eleven competitive Senate races gave the Senate Office of Public Records more than 80,000 pages of documents (many of which were printouts from electronic databases) by an October 15 filing deadline. The Senate Office then scanned those documents into electronic form and shared them with the FEC, which thirteen days later posted images of the records online for the public to view.[47] Information in the form of an electronic database about these races would not be available until months after the election. Transaction costs similarly limit scrutiny of other congressional actions, with both the Office of the House Clerk and Office of the Secretary of the Senate maintaining documents or data that are only available to people willing and able to travel to their offices to make hard copies of the records.[48]

At the agency level, officials may resist policies relating to FOIA, open government, or transparency either because scrutiny limits their discretion or because the costs of providing documents and data are an underfunded mandate from Congress. Bureaucrats can place the anticipated costs of transparency back on the public by failing to comply with FOIA provisions and daring requestors to engage in litigation to force compliance. Despite this political calculus, which means that those in power have concentrated incentives to resist accountability, progress on policies to support the watchdog function is feasible and evident. In battles between Congress and the executive branch or Democrats versus Republicans, making data more

accessible to the public and reporters may at times provide an electoral advantage. When a media outlet does take up the challenge and go to court to exercise the public right to know, the resulting court decision may set a precedent that makes similar data accessible to others in the future. There are a set of NGOs whose mission is to make government data and actions more accessible, including the Center for Effective Government, Center for Responsive Politics, Project on Government Oversight, and the Sunlight Foundation. The philanthropy that supports these NGOs effectively overcomes free-rider problems, since these groups often bear the costs of making data available and comprehensible or producing analyses and studies about transparency. Civic hackers can also support the accountability function by taking on programming projects, devoting the time to solve coordination or collection problems, and producing information that is then freely available to voters and reporters.[49]

Changes to public policies to support the accountability function of journalism involve group decision making and collective action about property rights to data and the flow of government funding. A different and parallel track would be to take current policies as given and ask how advances in computing might enhance our ability to use data to hold institutions accountable. Developing this approach does not require changing the minds of legislators or regulators. Advances in this area, increasingly called computational journalism, rather involve changes in the decisions of a diverse set of private actors: computer scientists, philanthropists, entrepreneurs, academics, and reporters.

Pathways to Computational Journalism

Sigma Delta Chi, the journalism organization that eventually became the Society of Professional Journalists, began publishing *Quill* in 1912. The magazine focused on trends and challenges in journalism. Despite the many investigations, exposés, and crusades undertaken by daily newspapers each year, the *Quill* appears not to have used the term *investigative reporting* until 1948.[50] While many other terms, including *muckraking,* were used to describe accountability work, the term *investigative reporting* only came to be applied after the approach had been on the scene for many decades.

The term *data journalism* had an even longer gestation period. The first issue of the *Guardian* (called the *Manchester Guardian* at its founding), contained a table listing the schools in Manchester with their total pupils (broken down by boys and girls) and total annual expenditures. The issue date was May 21, 1821.[51] In 1848, Horace Greeley used his access to travel reimbursement records as a Congress member to generate a story in the *New-York Tribune* that laid out, in table form, "each congressman by name with the mileage he received, the mileage the postal route [an alternative measure of distance from the district to the Capitol] would have granted him, and the difference in cost between them."[52] This set off a debate in the House about proper routes and rates for reimbursements. It would be approximately 160 years later before *data journalism* became a term applied to this type of analysis. In contrast to terms like *computer-assisted reporting* (CAR) and *data journalism*, the phrase *computational journalism* has emerged in anticipation of likely types of future work rather than a descriptor of current activities. Understanding the evolution of terms from *CAR* to *data journalism* to *computational journalism* allows you to see how innovative uses of data and algorithms could eventually help sustain investigative work.

By 1989, the use of databases and computers by a small but growing number of reporters led to the establishment at the Missouri School of Journalism of the IRE program that became known as the National Institute for Computer-Assisted Reporting (NICAR). In 1995, Brant Houston wrote one of the first CAR textbooks. In its fourth edition, published in 2015, Houston described CAR in this way: "Over time, three skills . . . for computer-assisted reporting have emerged: online resources (primarily finding and downloading databases), spreadsheets, and database management. As journalists have become more technologically sophisticated, other tools have joined these three, including statistical software, geographical information systems (GIS), or mapping software, and social network analysis."[53] In 1998, Bruce Garrison noted a long list of reasons why CAR (which he defined simply as "the application of computers to gather information for a news presentation") was being adopted in newsrooms: "Increased productivity by journalists. . . . Cost savings in information gathering. . . . Increased quality of local reporting. . . . Increasing meaning in analysis of information and less dependence on sources for interpretation of that information. . . . Keeping up with the competition. . . . Increased access to information. . . . Technical

reliability and greater accuracy of information. . . . Better storage and faster retrieval for follow-up uses and other needs."[54]

For the most part, CAR practitioners pursued stories in data by using four techniques that Elliott Jaspin, the journalist who pioneered the transfer of mainframe data to PCs via the Nine Track Express software, advised: "searching, counting, sorting, and cross-indexing."[55] In analyzing a subset of CAR investigative stories, Margaret DeFleur showed that the predominant analytical techniques were very basic: counts, percent, rank order, or before-after comparisons. Only a third of the stories went as far as calculating a mean or trend, and less than 5% attempted sampling, hypothesis tests, or probability estimations.[56] CAR stories often had great impacts through relatively simple methodologies, because sometimes putting a frequency or percentage or category on the results of public policy was enough to spur change. The *Sun Sentinel* won the 2013 Pulitzer Prize for Public Service, for example, by calculating the speed of police cars from toll data and showing how frequently high rates of speed were observed.[57]

By 2012, the year when *The Data Journalism Handbook* was first published, journalists were more likely to use the term *data journalism* than *computer-assisted reporting* when talking about innovations in reporting.[58] NICAR's Web site treated the changing taxonomy with this explanation: "The term 'computer-assisted reporting' was used widely in the last two decades to describe what many now call 'data journalism.'"[59] However, definitions of data journalism usually described work that could include CAR but also could include a wider array of reporting. Jonathan Stray defined it as "obtaining, reporting on, curating and publishing data in the public interest."[60] Alex Howard described the evolving use of data by reporters as:

> work [that] is data journalism, or gathering, cleaning, organizing, analyzing, visualizing, and publishing data to support the creation of acts of journalism. A more succinct definition might be simply the application of data science to journalism, where data science is defined as the study of the extraction of knowledge from data. . . . In its most elemental forms, data journalism combines: (1) the treatment of data as a source to be gathered and validated, (2) the application of statistics to interrogate it, (3) and visualizations to present it, as in a comparison of batting averages or stock prices.[61]

Mark Coddington analyzed CAR, data journalism, and computational journalism by describing how these approaches differed on four dimensions: professional vs. network orientation; opacity vs. transparency; sampling vs. big data; and passive vs. active visions of the public. He found that the development of CAR was heavily influenced by Philip Meyer, whose 1973 text on "Precision Journalism" advocated the use of social science methods (including sampling and surveys) in reporting on topics often involving investigative or public affairs journalism. Data journalism is less linked to accountability work, and much more likely to provide the underlying data in a story for readers to explore. This provides a greater role for audience members to make sense of what the data describe and makes the work more transparent. As Liliana Bounegru explains:

> By enabling anyone to drill down into data sources and find information that is relevant to them, as well as to verify assertions and challenge commonly received assumptions, data journalism effectively represents the mass democratization of resources, tools, techniques, and methodologies that were previously used by specialists; whether investigative reporters, social scientists, statisticians, analysts or other experts. While currently quoting and linking to data sources is particular to data journalism, we are moving towards a world in which data is seamlessly integrated into the fabric of media.[62]

Coddington correctly notes that these approaches often overlap, with pieces of each evident within the same series or story. Summarizing the divergences, he concluded, "The three practices . . . are distinct quantitatively oriented journalistic forms: CAR is rooted in social science methods and the deliberate style and public-affairs orientation of investigative journalism, data journalism is characterized by its participatory openness and cross-field hybridity, and computational journalism is focused on the application of the processes of abstraction and automation to information."[63]

The subheadings of the section in *The Data Journalism Handbook* on the importance of this evolving methodology capture its functions and value: filtering the flow of data; new approaches to storytelling; like photo journalism with a laptop; number-crunching meets word-smithing; a remedy for information asymmetry; an answer to data-driven PR; providing independent

interpretations of official information; dealing with the data deluge; a way to save time; a way to see things you might not otherwise see; and a way to tell richer stories.[64] These benefits come with costs, including time to train, time to explore, and funds for data and software. These costs translate into stories not published or pursued because of resources devoted to data journalism. Like CAR or any other approach that involves original content creation about complex topics, there may not always be a story that results from digging with data journalism. As one practitioner noted: "We can spend days on it, without results. . . . By cross-tabulating different databases, one might come up with a scoop. But it happens one time out of ten. It requires a lot of time, without immediate or systematic results."[65]

There can be different types of investigative insights from reporters approaching a story with CAR versus other data journalism backgrounds. Sylvain Parasie and Eric Dagiral found that in looking at reporting in Chicago, "programmer-journalists" brought different assumptions to story construction: news should be viewed as structured data of a type that could be captured in a database; data sets should be shared with the public so that audience members can delve deeply on their own; and people can use the data to explore how government works and, through this data use, hold government accountable.[66] CAR enthusiasts sometimes questioned the degree that readers could filter data and provide context to arrive at an understanding of the operation of a policy, and believed that expecting individuals to dive into data to find principal-agent breakdowns on their own may be expecting too much.

In a case study of the Center for Investigative Reporting's (CIR) 2011 series ("On Shaky Ground") on seismic safety and California public schools, Parasie found that reporters with CAR roots were more likely to follow a hypothesis-driven path, while those rooted in tech were more data driven. The project, which cost approximately $550,000 to report, initially involved cleaning databases that contained dirty data and (in some cases) contained biases introduced by pressure from affected stakeholders. Describing his desire to avoid incorrect assertions about the safety of a particular school, one of the journalists noted: "No information is ever clean. No data is ever perfect. I'm willing to accept that. But given that limitation, it's really important and really necessary and really hard to figure out what we can responsibly say with this information."[67] Reporters iterated between the field and the

data. Visits to individual schools helped identify problems with the database. Eventually the data were cleaned enough so that the database helped journalists to identify schools illustrating particular seismic risks and to interview experts about additional problems. While the written reporting provided a context for overall policy failures, CIR's publicly accessible database allowed granular analysis so that parents could look up seismic risk information about their children's school. The reporting resulted in significant debates, audits, and investigations in California about school seismic safety, earned a 2011 IRE Medal (the highest IRE honor for investigative work), and made CIR's California Watch a 2012 finalist for the Pulitzer Prize for Local Reporting.[68]

Professor Irfan Essa organized the first conference on computation and journalism at Georgia Tech in 2008, which featured Google News inventor Krishna Bharat as a keynote speaker.[69] Since then, debate has proceeded about the definition of computational journalism, even as early manifestations of this approach rapidly evolve. In 2009, Fred Turner and I wrote: "What is computational journalism? Ultimately, interactions among journalists, software developers, computer scientists and other scholars over the next few years will have to answer that question. For now though, we define computational journalism as the combination of algorithms, data, and knowledge from the social sciences to supplement the accountability function of journalism."[70] Working with Sarah Cohen in 2011, we noted that computational journalism, "Broadly defined . . . can involve changing how stories are discovered, presented, aggregated, monetized, and archived."[71]

Discussions with investigative reporters laid out challenges that computation could help solve, including aggregation of content across many sources, finding entities within and across records, clustering and classifying documents, indexing audio and video files, and pulling data more easily out of forms. We noted that innovations in these areas would help investigative journalists and consumers overcome a key problem: "too much material too difficult to obtain containing too little information."[72] In a 2011 article entitled "The Promise of Computational Journalism," Terry Flew and coauthors described the dimensions of computation that could advance sense making in journalism:

> *Automation* alleviates activities such as data gathering and interpretation, number crunching, network analysis, sorting, and processing that would otherwise need to be done manually; *algorithms* allow operators to follow predefined steps needed to accomplish certain goals, identify problems, find suitable solutions in a large set of alternatives, and verify information in a reliable, consistent and efficient manner; and *abstraction* enables the qualification of different levels or perspectives from which an idea may be presented or new directions that may be explored.[73]

Though computational journalism builds on and incorporates elements of CAR and data journalism, the new approach can sometimes involve larger data sets and more sophisticated algorithms. Data sets in computational journalism may often be larger than spreadsheets such as Excel can incorporate. The use of "Big Data" by corporate and government actors may eventually mean that reporters will use data of similar size in their attempts to scrutinize public and private actors. While the term "Big Data" does not have a widely accepted definition, Seth Lewis and Oscar Westlund note that it often refers to data whose "volume, variety, and velocity" are not easily handled by standard computer storage and processing.[74] Advances in the combination of computation and journalism are uneven across types of reporting functions and computational processes. Assessing the research literature in computer science in 2012, Nicholas Diakopoulos found: "concepts such as natural language processing, data mining, social computing, and information visualization have garnered the most amount of attention in terms of their application to news and journalism. Topics such as machine learning, knowledge representation, information retrieval, and computer vision have also gotten some attention. But . . . very little research has looked at how machine translation, tangible user interfaces, agents, or virtual reality can be applied to news information or journalism."[75] In analyzing recent developments in computational journalism, I see three distinct types of advances: reporting by algorithms, about algorithms, and through algorithms.

In automated journalism, the combination of data and code yields a story written via algorithm. This happens now in areas where events follow easily predictable patterns and involve quantitative measures. The software firm

Narrative Science has analyzed typical story frames and language in sports stories and, using repetitively structured data from box scores, can craft news stories about games. The firm notes its software "generates real-time communications, such as Tweets, game updates, game recaps and game analysis, customized to the audience and the level of play, whether that's youth, college or professional level sports."[76] In 2014, Automated Insights' software generated more than one billion stories via algorithm, including quarterly earnings report stories written for release by the AP for over 3,000 companies.[77] On March 17, 2014, the *Los Angeles Times* carried a short article about a local earthquake that was posted to the Web three minutes after the vibrations ceased. The article was generated by Quakebot, an algorithm designed by *LA Times* reporter Ken Schwenke, so that: "Whenever an alert comes in from the U.S. Geological Survey about an earthquake above a certain size threshold, Quakebot is programmed to extract the relevant data from the USGS report and plug it into a pre-written template. The story goes into the *LAT*'s content management system, where it awaits review and publication by a human editor."[78]

To date, stories by algorithm have not covered investigative topics, in part because of the relatively high fixed costs of discovering new examples of institutional breakdowns and the transaction costs often associated with getting the data involved in accountability work. The closest automated journalism relating to accountability work involves Narrative Science partnering with ProPublica to produce more than 52,000 stories about the availability of advanced classes at public schools. ProPublica used federal education data from 2009 to 2010 to develop a series entitled "Opportunity Gap," which explored the degree students from poor families had access to AP or advanced coursework. The Narrative Science algorithms generated a story for each school in the database, so that when parents searched for data from their particular public school, the results returned that story.[79]

Investigative reporting about how algorithms work is a growing though technically challenging field. Diakopoulos points out the need to "assess algorithmic power by analyzing the atomic decisions that algorithms make, including prioritization, classification, association, and filtering."[80] Companies and campaigns may be unlikely to reveal willingly how economic, political, social, and demographic factors influence their interactions with individuals. Reporters, though, have begun to reverse engineer the rules that

appear to govern how people are treated by interactions with institutions using big data and complex algorithms to guide their decisions. Diakopoulos describes such work by *Wall Street Journal* reporters writing about potential online price discrimination, that is, charging different prices to consumers for the same good online because of different consumer characteristics. The *Journal* team built software that allowed them to simulate Internet searching that appeared to come from different types of computers and browsers, from different parts of the country, from users with different browsing profiles. This allowed the *Journal* reporters to simulate, for example, visiting the Staples Web site seeking price quotes for a Swingline stapler as if they were users from each of the more than 42,000 zip codes in the United States.[81] Their finding was that "the Staples Inc. website displays different prices to people after estimating their locations. More than that, Staples appeared to consider the person's distance from a rival brick-and-mortar store, either OfficeMax Inc. or Office Depot Inc. If rival stores were within 20 miles or so, Staples .com usually showed a discounted price."[82] An apparent by-product of taking into account the ability of online shoppers to visit competitors' nearby stores is that people in areas with higher mean incomes were more likely to get the discounted price for the stapler. Diakopoulos describes similar reverse engineering challenges where reporters investigated autocompletions on Google and Bing, autocorrections with the iPhone, and stock-trading plans used by executives. He notes that one obstacle to the spread of this type of accountability work is that "the number of computational journalists with the technical skills to do a deep investigation of algorithms is still limited."[83]

The operation of accountability reporting through the use of algorithms illustrates the distinctive characteristics of computational journalism. Case studies and reflections on the sociology of reporting emphasize the effects of increasing amounts of data and public participation. Seth Lewis and Nikki Usher note that the rise of programmer-journalists has brought an open-source approach to reporting, which means "values of iteration, tinkering, transparency, and participation, each embedded in the open-source ethic, can be brought into the newsroom as architecture *and* culture—as a structural retooling of news technologies and user interfaces, and as a normative re-articulation of what journalism means in a networked setting."[84] The *Los Angeles Times* Homicide Report provides systematic coverage of murders in the area through a public facing database, map, and

blog, which includes initial posts written by algorithm when data are received from the coroner's office. Ben Welsh, editor of the paper's Data Desk, points out that this comprehensive, computational approach allows a reader to see beyond high-profile crimes: "Mr. and Mrs. Outlier get covered really well in crime news. And they drive tons of traffic, and we have full-time reporters who do that. But as you know that's an incredibly small fraction of the amount of crime that happens. And they probably reflect certain cultural predispositions. . . . But what data can bring us . . . is to try and give some fuller sense of crime as a phenomenon in the city."[85]

Analyzing the 2009 "Toxic Water" series from the *New York Times,* Astrid Gynnild determined that this data-intensive reporting about drinking-water safety involved more than 500 FOI requests spread across every state and multiple federal agencies.[86] The resulting databases and articles led to changes in environmental regulations and more funding for water projects. Gynnild stressed the scale of this project, which involved envisioning interactions across multiple levels of government, scraping and mining data across many sources, and coordinating a reporting team of ten across months of work. Yet computational projects can also be developed quickly and with fewer resources. Anna Daniel and Terry Flew note that the *Guardian* crowdsourced initial scrutiny of Member of Parliament expense reports by posting claim documents on the Web and creating a gamelike setting where readers could suggest claims for journalists to investigate. This resulted in 170,000 MP documents being examined in the initial eighty hours the data were available.[87]

Advances in computational journalism could improve the economic prospects for investigative reporting in several ways. On the supply side, research that lowers the costs of discovering accountability stories through better use of data and algorithms could make investigations more likely. On the demand side, research that allows outlets to tell investigative stories in more engaging or personalized ways raises the probability that media could attract more readers or viewers and monetize their attention through advertising or subscription. The original creation of facts about substantive policy issues can be expensive, and once the facts are known they can be copied and retold by those who did not bear the original costs of discovery. If engagement and personalization create product differentiation, however, a reader or

viewer may consistently seek out content from a particular outlet because the experience of learning facts there is distinct and more highly valued.

Discovering and Telling Campaign Finance Stories

To see how computational journalism might evolve, consider how journalists currently cover federal campaign finance. Reporting on this topic explores whose contributions are fueling campaigns, what types of access contributions produce, and how financial support in politics can translate into quid pro quo exchanges that influence policy and engender corruption.

Stories about financial support for candidates and their expenditures often start with the data filed by candidates and committees with the Federal Election Commission (FEC). Since the FEC Web site is not very user friendly, media outlets and NGOs have created algorithms to make the data and underlying patterns more accessible and transparent.[88] ProPublica's Campaign Finance API (application programming interface) looks at FEC data in fifteen-minute intervals to check for new filings, which means reporters can devise alerts for campaign contributions of a particular size. ProPublica's FEC Itemizer allows you to search individual records of contributions and expenditures involving federal political committees right after they appear in the FEC data. To explore the long-term relationships between donors and recipients and the similarity of giving patterns across political committees, Derek Willis created a software program called Bedfellows at the *New York Times* Upshot blog to allow users to explore PAC data from 1980 to 2014. CIR and IRE collaborated on a competition hosted by Kaggle that yielded algorithms that spot out-of-the-ordinary donations by a PAC, identify relations among committees supported by a relatively limited set of donors, and tap Natural Language Processing techniques to use employer and occupation data listed for FEC contributions to see trends in support. The Center for Responsive Politics created Anomaly Tracker, which allows users to explore FEC data by looking for instances of six different patterns, for example, "lawmakers sponsoring legislation that was lobbied by only one company or other organization whose employees or PAC also donated to the

sponsoring lawmakers" or "lawmakers receiving more than 50 percent of their itemized contributions from out of state."[89]

The crowd also plays a role in campaign finance analysis. In 2012, ProPublica's Message Machine project solicited fund-raising e-mails from readers, who eventually provided the nonprofit reporting outlet with the text of over 30,000 political e-mails.[90] ProPublica used Natural Language Processing and machine-learning techniques to reverse engineer the e-mail texts and explore how different demographic factors (including donation history and state location) affected the pitch campaigns used to solicit support. In the Free the Files project, ProPublica took on the problem that the documents on political ad spending at local television stations posted by the FCC did not contain a standard format. Involving almost 1,000 volunteers to log data on who spent what dollars on political ads in thirty-three swing markets, ProPublica was able to document about $1 billion in ad spending at local television stations in the 2012 election. The Sunlight Foundation's Party Time site allows individuals to upload invitations to and information about fund-raising events, which Sunlight supplements with data gathered from media accounts of political events.[91] This has generated data on thousands of political fund-raisers held since 2008, and allows the public to see details such as giving levels required for access at particular events. The most sophisticated extraction of information from the crowd comes through Crowdpac, a commercial venture that takes candidate voting records; candidate statements on the floor of Congress, on Facebook, and on Twitter; and donations to candidates at the federal and state levels from individuals and political committees to derive an ideological score for candidates that ranges from –10 for liberals to 10 for conservatives. The scoring system has enabled journalists in the 2016 presidential race, for example, to use contributions data to derive ideological rankings of the contenders.[92]

Describing what is sought through donations, including the *quo* in influence exchanges, can involve the question of how to describe policy decisions that range from contract decisions, regulatory actions, and legislative outcomes. Computational approaches here are more nascent, in part because outcome variables are harder to describe than the "input" variables of campaign contributions. Sunlight Foundation's Influence Explorer allows searches of lobbying registrations, which links lobbyists with clients and political issues. GovLab's Legisletters captures letters to agencies that Congress

members post on their Web sites, which provides a window into legislative oversight and interventions in regulatory decision making. Muckrock is a collaborative news site that allows free search of thousands of pages of government documents released under FOI, provides readers with low-cost options to submit records requests through Muckrock's software, and offers stories on topics such as private prisons and drone use generated through FOI releases.[93] ProPublica used sampling and advanced regression analyses to explore the factors that influence a powerful but rare decision, presidential pardons. Reuters used machine learning to classify types of petitions to the Supreme Court into different categories so the reporters could associate particular lawyers with types of cases (e.g., those where companies are sued by workers). This approach is useful in classifying bill texts and exploring the ability of interest groups to secure the introduction of similar types of legislation across states. ProPublica has pioneered the approach of Reporting Recipes, where this national nonprofit media outlet creates a database involving policy outcomes, creates a national story, and then offers advice on how local journalists can take the information and localize the policy story (and sometimes even provides matching funding for reporters who use crowdsourcing to fund their local investigations).[94]

As many journalists have noted, establishing the *pro* part of an exchange may often be the hardest link, since correlation between campaign donations and policy actions does not prove causation. Petty corruption is sometimes the easiest to spot. Representative Aaron Schock resigned when *Politico* established that he had billed his campaign and office accounts for reimbursement for driving more than 170,000 miles, though public records indicated the Chevrolet Tahoe he drove had only traveled about 80,000 miles when he sold it, suggesting that "he was reimbursed for 90,000 miles more than his car was driven."[95] Liz Whyte at the Center for Public Integrity suggests that, in looking for quid pro quos, part of the challenge is to "look in all the 'buckets' where politicians can receive benefits—campaign contributions (to everything from individual campaigns to leadership PACs to party committees), gifts, the promise of jobs for themselves or family members, and contributions to charities or foundations with ties to the politician."[96]

At their best, algorithms may suggest the most likely relationships to investigate for solid evidence of malfeasance. In a state where "one out of every eleven legislators to leave office since 1999 has done so under the cloud of

ethical or criminal violations," the Moreland Commission in New York hired a consulting firm with access to the Palantir software, technology used by government agencies in counterterrorism investigations, to spot relationships within a mosaic of connections.[97] In its preliminary report, the commission noted that it had used an

> analytics platform to ingest and analyze Board of Elections campaign finance information; elected officials' financial disclosure statements; lobbyist and client disclosures; legislative election results; legislative initiatives; publicly available biographical and professional data about elected officials mined from media, social media sites, and other databases; and proprietary research meticulously gathered by the Commission's investigative staff. To date, we have used this analytics tool to focus in on and uncover connections and relationships that otherwise would have been difficult or impossible to discern, thereby allowing Commission staff to create dossiers on companies, organizations, and persons of interest to "connect the dots" and construct timelines and relationships maps.[98]

As the commission began to narrow in on targets for investigation, Governor Cuomo abruptly disbanded the group. While the analytic platform used by the commission is beyond the tools available to most journalists, the case also illustrated the need for work by investigative reporters, who are independent of control by the public actors whose actions are under scrutiny.

Personalization of news about legislators has centered on tracking votes rather than donations. Part of the process of tailoring the information can start with voters' geographic coordinates. The Voter Information Project, a partnership involving the Pew Charitable Trusts and Google, created software that allows voters to text the word VOTE to a designated number or search an address on an app to receive information on polling place locations and ballot information. The *New York Times* created Represent, an app that took an address in New York City, determined the relevant districts of the over 150 legislators from the area associated with it, and tracked data such as votes and activities described in media articles. The *Times* later released the Districts API, which would take the latitude and longitude for any spot in New York City and list the political districts it belonged to, including "City

Council, the State Assembly, State Senate, and finally, the U.S. House of Representatives." Sunlight's Congress app allows you to select representatives to follow, so that you can get real-time notifications of votes and follow legislator Tweets, YouTube posts, and media coverage. Users of GovTrack.us can similarly follow votes, bills, and committee meetings. In 2014, the site sent out 4 million e-mail alerts about legislation and was used by 7 million people.[99]

True personalization in campaign finance coverage would involve an outlet that knew many things about your news interests: the candidates and policy outcomes you follow; your preferences about narrative, visualization, video, data, and documents; the coverage you've consumed to date and what would next be news for you; the context surrounding a story and the degree you know these details. An outlet that knew these preferences and histories would be able to offer you a distinct way of engaging with a story, and the more you visited and interacted, the more would be learned and the more distinct the experience. This would accentuate a lock-in effect, similar to that enjoyed by Amazon, since a media outlet would know you better than others and offer a differentiated product. The more readers who shared this experience, the greater the site's learning and the stronger the network effect would be. The more differentiated the product, the higher the chance that the outlet could monetize attention through advertising or subscription.

Parts of this personalization and engagement pathway exist, although no outlet has succeeded in creating this experience around public affairs reporting in general or campaign finance stories in particular. In 2009–2010, Google partnered with the *New York Times* and *Washington Post* to create Living Stories, a format designed to put coverage of a story on a single Web page that would contain an updated summary, provide new developments, and provide ways to go deeper for context and explore elements of multimedia.[100] The experiment stopped in part because of lack of resources and limits on technology at that time. Since then, research on story presentation has advanced, so that by 2013 the Metro Map project could take a collection of news articles about a topic, organize them by algorithm into a flow of events, and allow readers to explore the "subway map" of article clusters depending on what they knew about a topic and the degree they wanted to zoom in and learn finer details.[101] The *New York Times* had created an experimental text editor that suggested tags and possible annotations to reporters as they were writing articles.[102]

In the future this might allow a journalist to pull contextual information from a database, link to prior coverage, and insert quotes from other events while composing a story. Research at the Northwestern InfoLab News Context Project aims to "create an open source framework and toolkit with methods for automatically identifying, selecting, and presenting the broad range of contextual information that users need in order to gain a more nuanced understanding of news stories and other information."[103] This platform would involve recognizing entities in a story and posting information from other articles about these people or organizations, offering new quotations involving sources mentioned in the original article, generating coverage examples about a story from outlets from specific locations, and offering relevant tweets about the story or associated with stakeholders involved with the issue. The hypothetical platform would make explicit trade-offs that readers might make on settings they would prefer in editorial algorithms, including preferences about data versus human interest, context versus brevity, and timeliness versus analysis. Campaign finance stories might also lend themselves to "structured journalism" approaches, where events are described via database schemas so that stories can reuse and build upon prior observations.[104]

In an era of precarious economic support for accountability reporting, computational advances offer at least two avenues to support investigative journalism about the connections between money and politics. One is lowering the cost of discovering and telling stories, which can involve stories told by, about, and through algorithms. In the world of campaign finance, this would entail reports of financial contributions that could be written by algorithms similar to those used in financial reporting. Investigations about algorithms could focus on campaign targeting of voters and contributors. Work done through algorithms would essentially be electronic tip generation, with data suggesting correlations among donations and policy outcomes serving as the basis for more investigations.[105] These probabilistic assessments would not be public facing, for example, you would not say there is a .7 probability a legislator is corrupt. Rather the algorithm would narrow the set of legislative acts for a reporter to examine. With sufficient advances in personalization and engagement, the campaign stories discovered could be told in ways that lead readers and viewers to seek out the outlets that originally devoted resources to the creation of new, important content.[106]

Advancing Computational Journalism

Whose decisions would need to change to spur the development of this type of computational journalism? If computer science researchers and programmers came to view the accountability function of investigative reporting as at risk and worth sustaining, then the field would advance more rapidly. Social gains from computing are reflected in many forms: the high returns in IPOs for firms that disrupt and scale; long-term funding for advances that support national security and defense; and the freedom and self-expression that come from newly envisioned social media networks. If the positive spillovers from holding institutions accountable became better recognized and supported, more time and attention would flow toward helping solve the computational problems involved with accountability journalism. Many of the National Academy of Engineering's fourteen Grand Challenges involve policy areas that are the topic of investigative work (e.g., environment, health, energy, defense, urban infrastructure).[107] Computational advances in holding public and private institutions accountable would involve work in the policy areas at the heart of engineering's Grand Challenges. If research challenges such as the ACM's Data Mining and Knowledge Discovery competition and the Text Retrieval Conference contests used journalism as a context to explore innovative techniques, this would rapidly increase the set of people working in the field. A focus on data and algorithms meant to promote social good would also likely bring in underrepresented constituencies into computing.[108]

Philanthropists concerned about the media have made many investments in online public affairs sites, in part to determine how these operations might evolve to sustainability. Given the public goods and positive externalities inherent in high-quality local affairs coverage, the gap between the costs of investigative work and the ability of providers to monetize their social good may mean that some types of reporting will always need subsidy (just as some types of local public goods, such as symphonies and art galleries, merit continual support). Donors could focus attention on research questions that would help nonprofit media as a sector, including what types of incentives would increase individuals' willingness to support nonprofit media. There have been relatively few attempts to combine insights from behavioral economics, crowdsourcing, and philanthropy to run field experiments on

generating donations of money, time, or attention for accountability work. There is also a portfolio of research questions in computational journalism that have relatively high up-front costs but very high potential returns. Moonshots here would include truly personalized news, accountability stories generated primarily by algorithm, transcription with an accuracy and price that made video and audio text mining seamless, and analysis of corporate and open government data to reduce substantially the costs of story discovery.

Entrepreneurs may elect to invest in the field to do well and good at the same time. The dual stock structure adopted by some tech and media companies allows individuals with higher voting power, usually the founders or their descendants, to pursue objectives beyond profit maximization in their selection of what types of content to produce and pursue. Those involved with the producer demand for information about government, such as firms making government documents or data more transparent or accessible, have the opportunity to think about how the information they generate could serve as inputs into accountability reporting. Companies involved with immersive journalism (including virtual reality) may help media outlets produce distinctive content at low costs. The resulting product differentiation would bring viewers to the media source that created the information, and thus provide higher returns for original digging into stories. Finally, entrepreneurs in the consumer, producer, and entertainment areas may choose to reconstitute the bundle, combining public affairs reporting related to their primary product as another inducement for consumers. A private social network like Nextdoor, for example, might choose to add algorithmic accountability stories about neighborhood institutions.

Journalism educators can contribute to the development of computational journalism through the transformation of their teaching, research, and engagement. Many departments on campuses are working on turning unstructured information into structured data for analysis. Digital humanists are analyzing texts, computer scientists are mining social media, and social scientists are delving into documents, all via algorithms that could be used or modified by reporters. Journalism classes could involve joint teaching and joint research projects with professors from other disciplines. Given the difficulties involved in monetizing the full returns to public affairs reporting, journalism educators hoping to expand and deepen the ability of their graduates to

do data journalism should recognize that graduates with these skills will face many opportunities more financially rewarding than investigative reporting. This means that administrators may need to engage in extra fund-raising for journalism scholarships, so that students who do gain the requisite skills to approach text as data do not feel deterred by student debt from going into accountability work. Communications scholars may also need to engage with computer scientists to speed progress in algorithmic reporting, which may eliminate jobs dominated by repetitive tasks but free skilled reporters to deliver insights harder to replicate via computation.

Reporters will push the field of computational journalism forward in investigative areas for the same mix of reasons that have propelled past innovations in accountability. Some will focus on developing advances that raise the probability of laws and lives changing through the exposure of institutional breakdowns. The drive for social impact and the improving methods of measuring this will attract reporters to the challenge of finding stories amid increasing amounts of data. Journalists working in fields with currently high returns, such as business coverage, sports, or entertainment, may have the leeway and resources to experiment more with investigative puzzles. New ways of telling stories, such as structured journalism or virtual reality, will require investment in new skills. New ways of discovering stories, such as machine learning or sentiment analysis, will require a willingness to seek out new sources of data. The use (and misuse) of algorithms in markets and government will require a counter set of algorithms to detect and describe abuse and deception. Investigative work often revolves around the power of hidden action and hidden information. The evolution of computational journalism offers reporters new ways to find and tell stories with very familiar outlines of individuals and institutions going astray.

Conclusions

As a journalist at the *New York World,* Herbert Swope in 1917 won the first Pulitzer Prize given for reporting. He earned that distinction for his coverage from inside the German empire during World War I. As executive editor, Swope later led the *World* to two Pulitzer Prizes for Public Service in a three-year span. The *World* won in 1922 for its investigation into the revived and

growing Ku Klux Klan, and again in 1924 for its coverage of abuses in the Florida system of leasing out prisoners to work for private employers. Describing his approach to journalism, Swope declared, "What I try to do in my paper . . . is to give the public part of what it wants to have and part of what it ought to have, whether it wants it or not."[109] Fastforward about ninety years, and a very different ethos reigned in New York City. After Rupert Murdoch bought the *Wall Street Journal* in 2007, he criticized the investigative reporting approach that had long drawn and retained reporters at the paper. Murdoch declared at a conference, "Stop having people write articles to win Pulitzer Prizes. . . . Give people what they want to read and make it interesting."[110]

The gap between what people need to know as citizens and want to know as audience members persists because of rational ignorance, positive spillovers, and public goods. Why invest the time to learn about public policy when the statistical probability of your individual vote influencing the outcome of an election is vanishingly small? Why would an editor give a reporter free rein to spend six months on an investigative story that would likely change laws and lives, but not yield many additional financial returns to the company? Who in the media ecosystem gets directly rewarded as a reader or reporter for taking actions that create the public good of scrutiny and accountability? The logic is much different for other information demands. Consumers in search of products, workers in search of helpful data, and audiences in search of diversion seek out stories to read or programs to watch because they personally gain from the information. They make better purchases, find smarter workplace solutions, and see more enjoyable shows. This link between search, costs, and rewards at the individual level means that, for many types of information wants, the current digital age is a golden one. The drop in costs of creation, access, and distribution means that information abounds that matches consumers with services, producers with investments, and viewers with entertainment. Entrepreneurs and engineers in the San Francisco Bay Area have created companies, including Google, Facebook, Apple, Twitter, Snapchat, and Instagram, that are unrivaled in their ability to connect ideas, images, and updates with individuals.

Today, institutions private and public have unparalleled information about individuals, including their locations, expressions, pictures, networks, purchases, pasts, and likely futures. Advances in artificial intelligence and

virtual reality will even increase the data held about your daily decisions and desires. The sharing of information with companies and agencies often comes through transactions willingly made and often eagerly sought because of their benefits. You accept cookies because of content, you trade your location for speed, or share your past "likes" to get a chance at a better match or lower price in the future. In the midst of an era of big data and digital exhaust, it is hard to believe that important information or insights could go undiscovered. Yet the focus of accountability journalism is the very search for stories not told. Reflecting on the legacy of Watergate, *Washington Post* investigative reporter Bob Woodward said, "The central dilemma of journalism is that you don't know what you don't know."[111] Costly but influential investigations by reporters were once made possible by frictions in information markets. The difficulties of matching buyers with sellers gave rise to classified newspaper ads, which once made up 40% of daily newspaper revenues but today have been supplanted by craigslist and Zillow. Limited viewing options once kept audiences on a couch as consumers of television ads, while today spilt-screen viewing, split-second searching, and millions of video options make eyeballs elusive. Costly delivery and distribution methods once meant consumer, producer, entertainment, and voter information came bundled in a dominant local newspaper or widely viewed national broadcast. Cable, the Internet, and social media broke the bundle, making a wider variety of entertainment and expression possible. This also reduced bundling's support for information with relatively higher costs, lower expressed demand, but often greater social impact—namely accountability journalism.

Public ignorance means power for institutions, because it lowers the probability of accountability. Institutions are built on delegation of decision making, which allows division of labor and development of expertise. Yet delegating choice also creates the opportunity for abuse, for an agent to take hidden actions against the principal's interest or to use limited knowledge about options to pursue an agenda the principal would not favor if fully informed. Information costs involving uncertainty about quality or trust often drive choices away from markets and into the hierarchal structure of firms. Information costs contribute to imperfectly functioning markets when there are negative spillovers like pollution, anticompetitive acts in cartels, or asymmetries in knowledge about goods and services such as with credit, insurance, and health care industries. Government policies aim to improve on

these market failures, but imperfections in agency regulations or bureaucratic provision of services can also arise because of delegation within institutions.

What's new in investigative news is the discovery of which individuals in what institutions are violating laws, expectations, or norms. Stories from the muckraking era emphasized the power of individuals to do wrong and to remedy inequities. In the opening of his famous *McClure's Magazine* article in January 1903, entitled "The Shame of Minneapolis," Lincoln Steffens said: "Whenever anything extraordinary is done in American municipal politics, whether for good or for evil, you can trace it almost invariably to one man. The people do not do it. Neither do the 'gangs,' 'combines,' or political parties."[112] Digital economics, characterized by low-cost access, rapid and cheap distribution, millions of sources, and propagation through networks, has put expensive though impactful investigative reporting at risk. At a time when 1 billion people access Facebook in a single day, the actions of one person as an audience member or voter are negligible. The same economics that undercuts financial incentives to invest in accountability reporting means that once information and ideas are created, they can quickly circulate to great effect. This means that the work of a single individual can still strongly contribute to the accountability function of journalism. These individual roles are varied: a coder who spots new ways to discover or tell stories, a donor willing to bet on experiments in accountability by algorithm, an educator willing to cross disciplines to do research to turn unstructured information into structured data, a reporter eager to invest time and training to produce stories of moderate financial but high social returns, an editor able to make bets and solve puzzles with positive spillovers on government or corporate actions, a source choosing to deliver evidence of malfeasance or misdirection, or a witness with a smartphone willing to spread breaking news about wrongdoing. Breakdowns in delegated decision making in institutions are both enduring and inevitable. The combination of computation and journalism offers an expanded set of people new ways to hold those in power accountable, and allows them to serve as democracy's detectives.