DOI:10.1145/2001269.2001288

How computer scientists can empower journalists, democracy's watchdogs, in the production of news in the public interest.

BY SARAH COHEN, JAMES T. HAMILTON, AND FRED TURNER

Computational Journalism

THANKS IN NO small part to the modern computer's ability to gather and disseminate seemingly limitless amounts and types of data, the institutions on which the public depends for information about government are melting away. To some this shift may look like a good deal: Why not trade a few newspapers for what appears to be infinite access to information? But as news staffs decline, so too does the public's ability to monitor power.

If there's a silver lining in this situation, it is the ability of computer scientists to strengthen the hands of the remaining professional reporters and engage new players in the watchdog process. Advances in analytic techniques, computing power, and the volume of digitally stored documents have prompted improvements in making sense of unstructured data. Much of the work to date has focused on the consumer arena: Web searches, blog discussions, tweets, and text messages that generate terabytes of information. Marketers, social scientists, information professionals, and governments have all invested heavily in innovative algorithms to analyze these sources.[12]

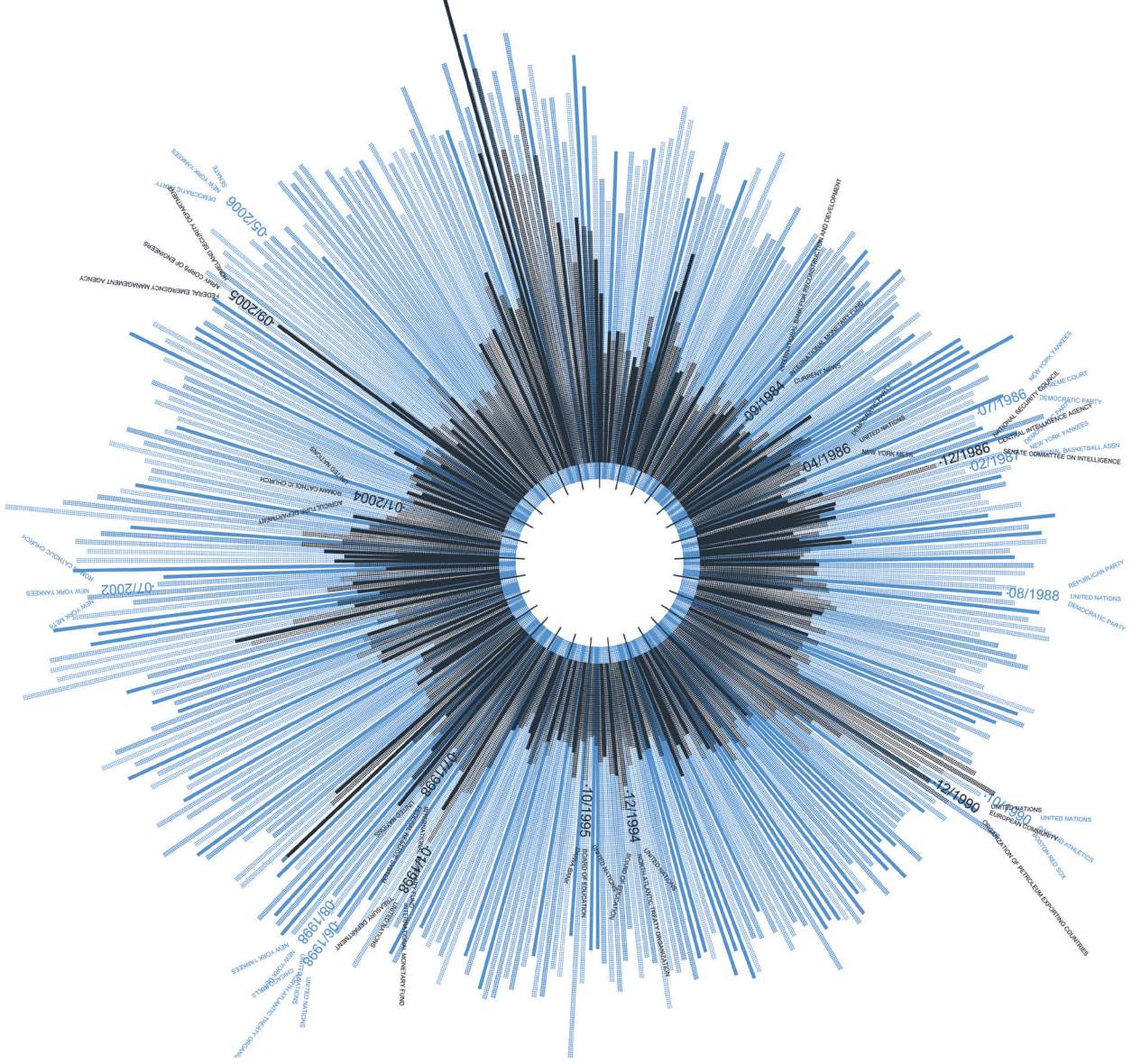
A similar push is also under way in investigative and public-affairs reporting. Researchers and journalists are exploring new methods, sources, and ways of linking communities to the information they need to govern themselves. A new field is emerging to promote the process: computational journalism. Broadly defined, it can involve changing how stories are discovered, presented, aggregated, monetized, and archived. Computation can advance journalism by drawing on innovations in topic detection, video analysis, personalization, aggregation, visualization, and sensemaking.[6–8,13]

Here, we focus on an aspect of computational journalism with a particularly powerful potential impact on the public good: tools to support accountability reporting. Building on the experience of an earlier generation of computer-assisted reporting, journalists and computer scientists are developing new ways to reduce the cost and difficulty of in-depth public-affairs reporting.

This aspect of computational journalism faces technical challenges ranging from how to transform paper-based documents into searchable repositories to how to transcribe collections of public video records. It also faces difficulty applying existing technology through user interfaces that accommodate the specific needs of journalists. Finally, it faces cultural challenges, as computer scientists trained in the ways of information meet journalists immersed in the production of news. If it is able to

» key insights

- **The public-interest journalism on which democracy depends is under enormous financial and technological pressure.**
- **Computer scientists help journalists cope with these pressures by developing new interfaces, indexing algorithms, and data-extraction techniques.**
- **For public-interest journalism to thrive, computer scientists and journalists must work together, with each learning elements of the other's trade.**



Visualization of the frequency of the words "hope" (blue) and "crisis" (graphite) published in the *New York Times*, 1981–2010.

overcome these hurdles, the field may sustain both public-interest reporting and government accountability.

IT and Watchdog Journalism

In the popular view, investigative reporting looks like the movie version of Watergate: secret meetings with sources, interviews, leaks, and cloak-and-dagger work. Due in large part to IT and statistical methods in the newsroom, it has long been more mundane and systematic.

A half-century ago, photocopying machines quietly revolutionized accountability journalism. The ability to copy documents worked in tandem with new freedom-of-information laws to make possible more sophisticated investigations. The machines let whistleblowers share agency records (such as correspondence and memoranda, inspection forms, and audits). Previously, investigations would often depend on undercover reporting, an ethically dicey practice. But the copy machine turned reporters' attention to documents and with them, to a new level of ethical clarity and accuracy.

In the late 1960s, a few reporters, led by Philip Meyer at Knight Newspapers, started using such research methods as sampling and correlation analysis to find and document stories.[14] Social-science tools have since helped establish notable patterns among variables, as in Bill Dedman's Pulitzer Prize-winning series in the *Atlanta Journal-Constitution* in 1988 on racial discrimination in mortgage lending in Atlanta.[5] Journalists have also used them to scout for outliers as news; for instance, reporters now make scatter plots to see the relationship between school test scores and student income, sometimes uncovering cheating by administrators and teachers in the odd, stray points.[9]

In the 1970s, reporters began to deploy the relatively novel methods of relational databases in their investigations. Using a portable nine-track tape reader and his newspaper's mainframe computer, Elliot Jaspin of the *Providence Journal* matched databases he acquired through government-in-sunshine laws. He found convicted drug dealers driving public school buses and local officials giving themselves discounts on their property-tax bills. By the late 1980s, he had shown that relational database technology and the Structured Query Language could be used to cross-reference systems to find news. He began traveling the U.S. showing reporters how to use these tools and founded an organization that later became the National Institute for Computer-Assisted Reporting (http://data.nicar.org/), an arm of the 4,500-member association Investigative Reporters and Editors (http://www.ire.org/).

VISUALIZAITON BY JER THORP

Today, matching data sets never intended to be matched is standard fare in newsrooms; stories highlighting child-care providers with felony records and voting rolls populated with the names of the dead are examples of the genre. The technique requires painstaking data cleansing and verification to deal with ambiguous identities and errors. Public records almost never include Social Security numbers, dates of birth, or other markers that would provide more accurate joins. But traditional news organizations have been willing to devote their time because they view documenting the failure of government regulation, unintended consequences of programs, and influence-peddling as core elements of their public-service mission.

However, such public-affairs reporting is increasingly at risk due to the decline in revenue and reporting staff in traditional news organizations—and is where the field of computational journalism can help the most. By developing techniques, methods, and user interfaces for exploring the new landscape of information, computer scientists can help discover, verify, and even publish new public-interest stories at lower cost. Some of this work requires developing brand-new technology, much of it involving work on new user interfaces for existing methods and some on simple repurposing. Technologies and algorithms already developed for informatics, medicine, law, security, and intelligence operations, the social and physical sciences, and the digital humanities all promise to be exceptionally useful in public-affairs and investigative reporting. At the same time, coupling the promised increased availability of government information with easy-to-use interfaces can aid nonprofessional citizen-journalists, non-governmental organizations, and public-interest groups in their own news gathering.

Understanding News Data

For computationalists and journalists to work together to create a new generation of reporting methods, each needs an understanding of how the other views "data." Like intelligence and law-enforcement analysts, reporters focus on administrative records and collections of far-flung original documents rather than anonymous or aggregated organized data sets. Structured databases of public records (such as campaign contributions, farm-subsidy payments, and housing inspections) generate leads and provide context, sometimes documenting wrongdoing or unintended consequences of government regulation or programs. But most news stories depend as much or more on collections of public and internal agency documents, audio and video recordings of government proceedings, handwritten forms, recorded interviews, and reporters' notes collected piece-by-piece from widely disparate sources. Some (such as press releases and published reports) are born digital; others are censored and scanned to images before being released to the public.

In many academic and commercial environments, researchers analyze comprehensive data sets of structured or unstructured records to tease out statistical trends and patterns that might lead to new policy recommendations or new marketing approaches. Journalists, however, look for the unusual handful of individual items that might point toward a news story or an emerging narrative thread. Journalists often collect records to address a specific question, which, when answered, marks the end of the analysis and the beginning of the story. This suggests a strict limit on the time and money invested in any document or data; it must be more effective or newsworthy than the alternative path of asking whistle-blowers or partisan insiders for the material.

On the flip side, investigative reporters have gigabytes of data on their hard drives and reams of documents in their file cabinets and are often willing to share them with researchers after a story is published. They are not bound by rules regarding human-subject testing or the research standards of peer-reviewed journals. Investigative reporters also expect plenty of false starts, tips that can't stand up to scrutiny, and stories that rarely hew to the route expected at the outset. In short, they write stories, not studies.

These characteristics suggest several areas of opportunity for collaboration among journalists, social scientists, humanists, and experts in computing. Over the past two years, we have conducted scores of interviews with reporters, editors, computer scientists, information experts, and other domain researchers to identify collaborations and projects that could help reduce the cost and difficulty of in-depth public-affairs reporting. In July 2009, we also brought together leaders in the field for a one-week workshop at Stanford University's Center for Advanced Study in the Behavioral Sciences (http://www.casbs.org).[10] Our conversations identified five areas of opportunity:

Combining information from varied digital sources. In our interviews, one reporter asked if it is possible to routinely combine all press releases from the 535 members of the U.S. Congress and their committees into a single searchable collection. Another wanted to search all 93 U.S. Attorney Web sites each week for news of indictments or settlements not likely to be shown in routine reviews of court records. A third dreamed of combing documents from the Web sites of the U.S. Securities and Exchange Commission, the Pentagon, and defense contractors to identify military officials who had moved into industry as consultants, board members, or executives.

What ties these ideas together is the ability to put into one repository material not easily recovered or searched through existing search engines. Each project features a limited number of discrete sources chosen by the reporter. Little of the material is available in RSS feeds. Each source has independent control of the form and format of its holdings; for instance, about one-quarter of the members of Congress keep press releases in databases accessed by Cold Fusion applications, while another significant portion post links to Adobe Acrobat files or organize HTML pages in subject-specific sections of their Web sites surrounded by member-specific templates. Each member's site is organized at least slightly differently.

This problem also arises for others, including public officials who want to monitor blogs, news sites, and neighborhood email lists, many not available in Google News alerts; corporations that want to monitor the public communications and regulatory filings of

their competitors and clients; and citizens who want to know everything their elected officials have done in the past week. Search and retrieval technology may be up this task, but the existing free and open source user interfaces to the technology remain crude and fail to address the variety of sources where finding answers to queries is less important than exploring what's new.

Information extraction. Most information collected by journalists arrives as unstructured text, but most of their work involves reporting on people and places. A beat reporter might cover one or more counties, a subject, an industry, or a group of agencies.

Most of the documents they obtain would benefit from entity extraction. Thomson Reuters allows the public to use its OpenCalais service (http://www.opencalais.com/), and at least a half-dozen open source and academic entity-extraction tools have been available for several years. The intelligence community and corporations depend on this basic but relatively new technique. But effective use of these tools requires computational knowledge beyond that of most reporters, documents already organized, recognized, and formatted, or an investment in commercial tools typically beyond the reach of news outlets in non-mission-critical functions.

Being able to analyze and visualize interactions among entities within and even outside a document collection—whether from online sources or boxes of scanned paper—would give stories more depth, reduce the cost of reporting, and expand the potential for new stories and new leads.

Document exploration and redundancy. There are two areas—finding what's new and mining accumulated documents—in which the ability to group documents in interesting ways would immediately reduce the time and effort of reporting.

Audiences, editors, and producers expect reporters to know what has been published on their beats in real time. Reporters need to notice information that is not commonly known but that could lead to news in interviews, documents, and other published sources. The recent explosion in blogs, aggregated news sites, and special-interest-group compilations of information makes distinguishing new stories time consuming and difficult. Collections of RSS feeds might comprise hundreds of stories with the same information.

Journalists look for the unusual handful of individual items that might point toward a news story or an emerging narrative thread.

In our interviews with journalists, we were told this challenge is more difficult than it seems for reporters lacking technical knowledge. But solving it would immediately reduce the amount of time spent distinguishing "commodity news," or news widely known and therefore uninteresting, from news their audience might not know or items that could prompt further reporting.

Another scenario arises in the collections of documents and data accumulated in a long investigative project. In some cases, existing search tools are not robust enough to find the patterns journalists might seek. For example, in 2006, reporters at the *New York Times* used more than 500 different queries to find earmarks for religious groups in federal legislation.[11]

In other cases, simply exploring a collection of documents might suggest further work if grouping them would help identify patterns. For example, in June 2010, the William J. Clinton Presidential Library released more then 75,000 pages of memoranda, email messages, and other documents related to Supreme Court nominee Elena Kagan. Grouping them in various ways might help better identify her interests, political leanings, and areas where she disagreed with others in the White House and suggest stories that could be missed simply by reading and searching the collection.

Combining these projects—content aggregation, entity extraction, and clustering of documents—could provide breakthrough innovation in investigative reporting. Together, they would directly address the key problem faced by most news consumers, as well as by producers: too much material too difficult to obtain containing too little information. These advances might allow for efficient, effective monitoring of powerful institutions and people and reduce the mind-numbing repetition and search in-depth reporting often requires.

Audio and video indexing. Public records increasingly consist of audio and video recordings, often presented as archived Webcasts, including government proceedings, testimony, hear-

ings, and civil- and criminal-court trials. Unless a third party has already transcribed, closed-captioned, or applied speech-recognition techniques on the record, most reporters have no way to search even a rough transcript. In addition, many reporters record many of their interviews digitally but rarely have useful speech-recognition software to index them. Basic consumer software products (such as Dragonspeech from Nuance) work on simple, short recordings or trained voices. Other promising projects (such as Google's Audio Indexing, or GAUDi) are not publicly available. GPS and voice recognition on mobile phones and voice mail could make reporters think solving their problem is simple.

Reporters could make near-daily use of technology and a user interface that would provide approximate indexing of a variety of voices and conditions, leading them to the portions they most want to review. They do not require the accuracy of, say, e-discovery by lawyers or official government records. Instead, they want a quick way to move to the portion of a recording that contains what may be of interest, then carefully review and transcribe it. Existing technology is probably adequate for reporters' immediate needs, but we are unable to find reasonably simple user interfaces to the technology that would allow unsophisticated users to test the technology on their own recordings.

Extracting data from forms and reports. Much of the information collected by reporters arrives in two genres: original forms submitted to or created by government agencies, often handwritten, and reports generated from larger systems, sometimes electronically and sometimes on paper. Examples include financial disclosure statements of elected officials, death certificates, safety inspections, sign-in sheets at government checkpoints and police incident reports. Journalists have few choices today: retype key documents into a database; attempt to search recognized images; or simply read them and take notes. An in-house programmer can occasionally find the pattern of digital reports intended for printing that can be leveraged to reverse them back into a structured database, but this time-consuming job requires skill well beyond nearly all reporters.

Extracting meaningful information from forms is among the most expensive and time-consuming large news investigations. Its cost sometimes results in abandoning promising stories. Reducing that cost could encourage substantially more important reporting on government, particularly at state and local levels where special-interest groups and NGOs are less likely to step in to help.

New Tools, New Organizations

A handful of new services have emerged to help address journalism's data challenges. Usually free for small-scale or non-commercial use, they facilitate analysis, visualization, and presentation of structured data: Google Refine promises to let reporters scrap their spreadsheets for filtering, viewing, and cleaning basic data sets; ManyEyes from IBM lets news organizations visualize and share data on their Web sites; Tableau Public from Tableau Software, Google Earth, and other such products are routinely used by news organizations to generate and publish visualizations. New tools (such as TimeFlow developed at Duke University as an investigative tool for temporal analysis) are being created to address some longstanding needs of reporters.[4]

Another set of tools created for other purposes, often experimental or academic, shows promise for the fast-paced, ad hoc nature of reporting challenges. Several political-science scholars have created tools for clustering legislation and other public documents; homeland-security developers have created tools (such as Georgia Tech's Jigsaw[15]) for visualizing the connections among documents; and the CMU Sphinx project[3] has created reasonably accurate open source speech-recognition technology. Applications developed for intelligence, law enforcement, and fraud investigations by such companies as Palantir Technologies and I2 are expensive and finely tuned to specific industries, though they address similar challenges on a different scale and with different requirements for speed and accuracy.

DocumentCloud (http://www.documentcloud.org), a nonprofit founded in 2009 by journalists at the *New York Times* and ProPublica, hopes to address one of the most vexing issues in documents reporting: scanned images files. With it, reporters can annotate their documents as they find interesting or questionable sections and see which entities appear in multiple documents. At this writing, most news organizations have used it most effectively to publish government documents. But the project, which includes information extraction as a standard feature, shows great promise helping address some of the problems of digesting large document collections.

If such new methods and tools could be more widely adopted in journalism, they could perhaps do for investigative reporting what the photocopier and relational database did in decades past.

Despite these possibilities, challenges persist in working with unstructured data for watchdog reporting. News organizations increasingly look toward their audiences to fill some of the gaps. Thanks to methods of collaboration pioneered in computer science, amateurs and professionals now find themselves reporting side-by-side.

In 2005, Josh Marshall of the online news site Talking Points Memo (http://talkingpointsmemo.com/) encouraged his readers to contribute local stories of politicization of the Justice Department under the George W. Bush administration. His work, and the work of his audience, won a prestigious George Polk award for investigative reporting in the first known use of crowdsourcing to uncover an important investigative story. The *Guardian* in London has enlisted its readers to help review payment records of members of Parliament on deadline.[1] American Public Media (http://americanpublicmedia.publicradio.org/) created possibly the largest crowdsourcing network, with more than 60,000 members in its Public Insight Network (http://www.publicinsightnetwork.org/). It now needs to find ways to better understand and mobilize them while encouraging local National Public Radio affiliates and other partners to use the network effectively. Leaders in social media and collaboration, including ProPublica, are helping reporters learn to motivate their audiences and organize ad hoc communities around projects. These models all

show promise enlisting more eyes and ears in accountability reporting.

Next Steps

For many aspects of accountability reporting, including beat and investigative work, a key question for the future is whether journalists' research problems are scientifically interesting, challenging, or even new enough. Some could be solved with new user interfaces that accommodate journalism's quirks. Others are bothersome impediments to more interesting work. For example, possibly the most intractable problem in investigative reporting remains the form in which the material arrives, often on paper, with large sections blacked out by censors, or in large files combining images of thousands of individual records (such as email messages, memos, forms, and handwritten notes). The investment required to address the problem is unlikely to be made in the news industry and may not interest software developers or scientists. Philanthropists, academic institutions, and spin-offs from government-funded research may ultimately provide the solution, not computer scientists.

Journalism and computer science schools have begun to address the questions at the intersections of their fields. Two top journalism programs, the Columbia University Graduate School of Journalism and Northwestern University's Medill School of Journalism, have initiated interdisciplinary programs with computer science or engineering departments. At the Georgia Institute of Technology, professor Irfan Essa teaches an influential course in computation and journalism, conducting research in the field, while videogame scholar Ian Bogost explores new ways to use games in journalism.[2,7]

Fortunately, funders have also stepped into the financial vacuum surrounding watchdog journalism. The largest is the Knight Foundation, which funds the annual $5 million Knight News Challenge contest and other grants and programs, along with university centers, startups, and nonprofit investigative news sites. Knight has funded digital innovators EveryBlock, DocumentCloud, and others. Projects funded through government programs (such as scanning and digitization projects at the National Archives and the Library of Congress) might further help address the challenges.

The Association for Computing Machinery's special interest groups, most notably Information Retrieval (http://www.sigir.org/) and Knowledge Discovery in Data (http://www.kdd.org/), could foster sessions at their own meetings and with journalism organizations, including Investigative Reporters and Editors. Annual research competitions (such as the Text Retrieval Competition, the IEEE Visual Analytics Symposium's Challenge, and the ACM's Data Mining and Knowledge Discovery competition) could each include as research topics the kind of data challenges facing journalists today.

Conclusion

How might the worlds of politics, governance, and social discourse change when computational journalism fulfills its promise? Not surprisingly, part of the answer is journalistic: Stories will emerge from stacks of financial disclosure forms, court records, legislative hearings, officials' calendars or meeting notes, and regulators' email messages that no one today has time or money to mine. With a suite of reporting tools, a journalist will be able to scan, transcribe, analyze, and visualize the patterns in these documents. Adaptation of algorithms and technology, rolled into free and open source tools, will level the playing field between powerful interests and the public by helping uncover leads and evidence that can trigger investigations by reporters. These same tools can also be used by public-interest groups and concerned citizens.

Much more of the answer, though, involves democracy itself. How can citizens govern themselves if they are unable to hold their governments accountable? This ancient question is often phrased as "Who guards the guardians?" A hundred years ago, or even 20, the answer might have been "full-time journalists." But today they can be only part of the answer. Journalists need to partner with computer scientists, application developers, and hardware engineers. For decades, the computing community has empowered individuals to seek information, improving their lives in the process. Few fields have done more to give citizens the tools they need to govern themselves. Few fields today need computer scientists more than public-interest journalism.

Acknowledgments

We would like to thank James Bettinger, Glenn Frankel, Ann Grimes, Jeffrey Heer, Michael Schudson, John Stasko, Fernanda Viegas, and Martin Wattenberg for generously reviewing and improving this article. C

References

1. Anderson, M. Four crowdsourcing lessons from the Guardian's (spectacular) expenses-scandal experiment. *Nieman Journalism Lab* (June 23, 2009); http://www.niemanlab.org/2009/06/four-crowdsourcing-lessons-from-the-guardians-spectacular-expenses-scandal-experiment/
2. Bogost, I., Ferrari, S., and Schweizer, B. *Newsgames: Journalism at Play.* MIT Press, Cambridge, MA, 2011.
3. Carnegie Mellon University Sphinx; http://cmusphinx.sourceforge.net/
4. Cohen, S., Viegas, F., and Wattenberg, M. TimeFlow Github repository http://wiki.github.com/FlowingMedia/TimeFlow/
5. Dedman, B. *The Color of Money: Text of the Pulitzer-Winning Articles*; http://powerreporting.com/color/
6. Diakopoulas, N. Research and projects; http://www.nickdiakopoulos.com/research-and-projects/
7. Essa, I. Computation and Journalism class at Georgia Tech; http://compjournalism.wordpress.com/
8. GVU Center at Georgia Tech. *Journalism 3G: The Future of Technology in the Field. A Symposium on Computation and Journalism* (Atlanta, Feb. 22–23, 2008); http://www.computational-journalism.com/symposium/index.php
9. Hacker, H. and Benton, J. Faking the grade. *Dallas Morning News* (June 3–5, 2007).
10. Hamilton, J.T. and Turner, F. *Accountability Through Algorithm: Developing the Field of Computational Journalism. Report from Developing the Field of Computational Journalism.* Center for Advanced Study in the Behavioral Sciences Summer Workshop (Stanford, CA, July 27–31, 2009); http://dewitt.sanford.duke.edu/images/uploads/About_3_Research_B_cj_1_finalreport.pdf; and workshop bibliography: http://dewitt.sanford.duke.edu/images/uploads/About_3_Research_B_cj_2_ReadingLinks.pdf
11. Henriques, D.B. and Lehren, A.W. In God's name. *The New York Times* continuing series (Oct. 2006–Nov. 2007); http://www.nytimes.com/ref/business/churchstate.html
12. Hey, T., Tansley, S., and Tolle, K., Eds. *The Fourth Paradigm: Data-intensive scientific discovery.* Microsoft Research, Redmond, WA, 2009; http://research.microsoft.com/en-us/collaboration/fourthparadigm/
13. Mecklin, J. Deep throat meets data mining. *Miller-McCune* (Jan./Feb. 2009); http://www.miller-mccune.com/politics/deep-throat-meets-data-mining-4015/
14. Meyer, P. *Precision Journalism: A Reporter's Introduction to Social Science Methods.* Rowman & Littlefield Publishers, Inc., New York, 2002.
15. Stasko, J., Görg, C., and Liu, Z. Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization* 7, 2 (Summer 2008), 118–132; http://www.cc.gatech.edu/gvu/ii/jigsaw/

Sarah Cohen (sarah.cohen@duke.edu) is Knight Professor of the Practice of Journalism and Public Policy at the Sanford School of Public Policy, Duke University, Durham, NC.

James T. Hamilton (jayth@duke.edu) is Director of the DeWitt Wallace Center for Media and Democracy and Charles Sydnor Professor of Public Policy at the Sanford School of Public Policy, Duke University, Durham, NC.

Fred Turner (fturner@stanford.edu) is Associate Professor of Communication and Director of the Program in Science, Technology and Society, Stanford University, Stanford, CA.