Accountability Through Algorithm:

Developing the Field of Computational Journalism

James T. Hamilton

Charles S. Sydnor Professor of Public Policy, Duke University

Fred Turner

Assistant Professor, Department of Communication, Stanford University

A report from *Developing the Field of Computational Journalism*, a Center For Advanced Study in the Behavioral Sciences Summer Workshop, July 27-31, 2009. Comments are welcome at jayth@duke.edu and fturner@stanford.edu

Table of Contents

Acknowledgements

The Center for Advanced Study in the Behavioral Sciences has a well-earned reputation for encouraging innovation, reflection, and interdisciplinary discovery. We appreciate the leadership of Claude Steele (Director) and Anne Petersen (Associate Director) in creating the Summer Workshop Program, the decision by the CASBS Board to select computational journalism as one of the inaugural workshops, and the work of Susan Beach, Christy Duigan, Cynthia Pilch, Elisabeth Ponsot, and Ravi Shivanna that made it work so well.

Writing a report about potential advances in reporting is an iterative process. We thank all the workshop participants and presenters for contributing to the creation of ideas about how the field of computational journalism may evolve. As the authors of the report, we remain responsible for errors that remain.

Overview

In recent years, ubiquitous computation has transformed the landscape of journalism. It has undermined business models, rebalanced the relative power of reporters and audiences, and accelerated the delivery of information worldwide. Even as older modes of journalism are beginning to slip from view however, we believe that computation has also begun to present reporters with a series of new techniques with which to pursue journalism's long-standing public interest mission. Public and private data sources are expanding exponentially and transparency advocates are pushing for more. Computer scientists are rapidly creating algorithms to make sense of large-scale data sets. Social scientists too are working with data in real time, publishing quickly, and thus finding themselves pursuing social problems alongside reporters. We believe that the convergence of these trends holds the promise of the development of a new field: computational journalism (CJ).

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. In some ways computational journalism builds on two familiar approaches, computer-assisted reporting (CAR) and the use of social science tools in journalism championed by Phil Meyer in *Precision Journalism: A Reporter's Introduction to Social Science Methods* (Rowman and Littlefield, 2002). Like these models, computational journalism aims to enable reporters to explore increasingly large amounts of structured and unstructured information as they search for stories.

At the same time though, computational journalism offers a new way to help sustain the watchdog reporting on which democratic citizenship so clearly depends. The term "watchdog reporting" encompasses both traditional investigative or enterprise reporting and the everyday beat coverage of key institutions. Together these kinds of reporting hold leaders accountable, unmask malfeasance, and make visible critical social trends. Without them, citizens would have little of the information they need to make many important choices. Yet, as the traditional business models of journalism have collapsed, so too have incentives to take on the high fixed costs of watchdog work. Computational journalism cannot transform the business situation of contemporary journalism. But it can create new tools that may reduce the cost of watchdog reporting in certain circumstances, take better advantage of the new information environment, and ultimately help sustain watchdog work during the technological sea change now under way.

What Could Computational Journalism Look Like?

During the last week of July 2009, the Center for Advanced Study in Behavioral Sciences (CASBS) hosted a summer workshop on *Developing the Field of Computational Journalism*. This workshop, directed by James T. Hamilton (Duke University) and Fred Turner (Stanford University), brought together journalists, researchers, and NGO participants (see Appendix) to discuss the evolution of this nascent field. This report describes how the working group participants viewed the likely growth and contributions of computational journalism in the months and years ahead.

The discussions among the CASBS Summer Workshop participants identified at least four distinct areas for innovation in computational journalism: techniques for data transformation and pattern discovery in investigative reporting; a digital "dashboard" for journalists; new social and technical structures for interactions among readers and reporters; and sense-making advances from other disciplines. Participants noted that innovations in each of these areas might come from a wide variety of communities and might involve repurposing existing tools as often as inventing new ones. In the following sections we provide concrete examples of the tools that may emerge from these areas and the processes by which they may be developed. We describe the envisioned innovations in detail, both to convey how the advances would help journalists and to stimulate discussion among readers of this report about how they themselves might spur the development of these tools.

1. Information Extraction, Integration, and Visualization

Ever-expanding sources of public and private data are dramatically increasing the range of opportunities for watchdog reporting. In order to take advantage of these, reporters need two kinds of assistance: first, they need ways to extract and integrate structured information from a variety of data sources such as text, video and the Web; second, they need tools with which to make visible and exploit patterns in the data.

The Challenge of Data

Investigative reporters often face the challenge of working with records developed for other purposes, with highly unstructured data, or with information riddled with inconsistencies or uncertainties. In a working paper entitled *Collecting, Mining, Visualizing, and Analyzing: A Menu of Projects to Advance Reporting Tools*, Duke University Knight Professor Sarah Cohen developed a comprehensive list of tools that would help reporters in pattern discovery. The excerpts from her paper below provide more details on the similar challenges faced by many reporters pursuing investigative stories:

Text extraction from web-based collection of documents

Collections of government documents are often still provided on paper (often with large sections blacked out because of privacy or security concerns). Others are a collection of searchable documents and images of pages. Reporters have no way to index or search these documents. Few have access to sophisticated optical character recognition (OCR) software and fewer have the necessary computing power available. An ideal tool would allow cash-strapped reporters to feed pdf documents into a web service and return a searchable version that can be indexed by tools like Google Desktop. The software would also make it easy to tag the documents with peoples' names, places and dates.

Examples of where this tool would have helped reporters include:

1. The transition team of the Obama administration posted letters from dozens of interest groups advising the new president on issues ranging from nursing homes to farm policy. However, the actual letters are not searchable or downloadable making it difficult to do serious analysis of policy once the administration began acting: http://change.gov/open_government/yourseatatthetable
2. Some enforcement agencies post original documents on their websites. An example of one that was used in a story this summer is at the National Transportation Safety Board: http://www.ntsb.gov/Dockets/Aviation/DCA09MA027/default.htm . Although "download-them-all" can be used to put all of these into a searchable folder, some of the documents are images of forms or scanned letters. Reporters have no way to distinguish among those formats and can't tell whether they've done a complete search of the documents.

Transcription of audio or video files

Congressional committees, state legislatures, county councils and some courts only make video or audio files available. Reporters may know approximately what was said, but have to watch or listen to hours of

meetings to find pieces they want to use in a story or want to ask a politician about. They also use these hearings as ways to find new sources, but they often aren't indexed well enough to know who was present. Current transcription software is too expensive or too difficult to use. Researchers could evaluate existing voice and video (speech) recognition software with an eye toward the level of accuracy and ease that would be needed in a news environment. Among the considerations is how well the software recognizes changes in the person speaking, indexes the video or audio file and can summarize terms found. Note that C-SPAN has started looking into this technology and may be a good partner for additional research in this area.

The lack of transcripts hampers reporting on local government, courts, state government and federal government. Effective tools to reduce the time needed – and improve the span of videos reviewed – would benefit reporters in almost every news organization. An example at the state level of video recordings that could be transcribed and made searchable is WisconsinEye, the state version of C-SPAN, at http://www.wiseye.org/legislature0910.html

Handwriting transcription

Investigative reporters often receive hand-written or typed forms in response to public records requests. Analyzing these often involves retyping the responses into a database before any analysis can be run. These documents are often basic accountability reports: politicians' financial disclosures or inspection results, such as restaurant inspections. Researchers could help evaluate the possible technologies for use of handwritten forms, particularly those used in medical form extraction technology and management of document-intensive legal cases.

This is one of the most time-consuming data collection tasks undertaken in many investigations across all size of news organizations. Despite the proliferation of electronic forms, many are available to the public only in an image form. Examples of forms include financial disclosure forms at http://clerk.house.gov/public_disc/financial-pdfs/ . To see a sample form, go to http://clerk.house.gov/public_disc/financial-pdfs/2009/8142761.pdf.

Seeing Patterns in the Data

Visualization templates

Visualizations of data collected for stories come too late for most reporters to use them effectively before publication. Researchers could create several basic Flash or lightweight Web templates for visualizing a combination of time and place. Some of the places to look for tools include:

Google Charts / Google Maps: http://code.google.com/apis/chart/

MIT's Simile project: http://simile-widgets.org/exhibit/

IBM's ManyEyes: http://manyeyes.alphaworks.ibm.com/

The first two are too difficult for most reporters, and the last is too limited and requires publishing your un-confirmed data to the world. The templates – in order to be widely adapted – should allow for many data formats, ranging from original XML or CSV files, to copying and pasting from a spreadsheet, to directly entering and editing the data into a form. They should be easy to update, easy to customize, and they should show a sense of design. One reason to use Flash and Flex is that most news

organizations still do their published visualizations in that form. This would allow their work to move seamlessly from analysis to publication with just a little editing.

The ability to quickly understand a complex dataset – large or small – is becoming more important in accountability reporting. Improving the coordination among pieces of a complex story is also important for building credibility with audiences. For current examples and discussion, see:

http://oakland.crimespotting.org/

http://projects.nytimes.com/crime/homicides/map

http://www.nytimes.com/ref/us/20061228_3000FACES_TAB2.html

http://www.poynter.org/column.asp?id=101&aid=161675.)

Flexible timeline/chronology tools

Any long-running story cries out for a chronology and timeline as a reporting and writing tool. Reporters usually create them out of Excel spreadsheets or retype notes into a 40-page Word document in chronological order. But they can't zoom in, tag events for publication, turn on and off players or events and otherwise use them effectively. The stories can range from a long court case to a police investigation or even a narrative reconstruction of events.

Advances in programming and design could create a flexible, easy and aesthetically decent chronology and timeline maker. It should allow reporters to enter the data in several different ways, let them edit it and export the information in many forms, let them turn on and off people or types of events, zoom in and out on some time periods, and allow for thousands of records if necessary. Currently, Simile's Exhibit and Timeline is the closest open source available, but there may be commercial products that are used in litigation support or law enforcement.

2. The Journalist's Dashboard

Reporters on a beat face a constant challenge of patrolling familiar information sources in search of new information. As the number and range of digital data sources increases, journalists increasingly need a tool with which to spot what's new and what's important in the flow of daily information. At the same time, thanks to the efforts of both open source and corporate software developers, the number and diversity of tools available to do parts of that work is also growing rapidly. We propose the creation of a Journalist's Dashboard that could host those tools and aid the reporter in her daily work. Elements of the Dashboard discussed at the CASBS conference include:

Customized Google News for Reporters

In her description of potential tools that would help journalists discover new stories, Sarah Cohen notes:

> Beat reporters – the core of accountability journalism – have to keep up with local blogs, online news organizations, emailed press releases and government websites. They attempt to monitor the sources, but many just repeat the same stories as aggregators. They need a fast way to organize the new discoveries on their beats and a way to find the original source. Software developers could create something like a customizable Google News for beat reporters. It would scan the sites selected by the reporter and list only distinct stories. It would list only the original source and show which of the sites had linked to it, as a way of determining importance. It might be sensitive to timing – the last time that the reporter looked at the tool.
>
> Constructing this tool could involve a building a set of sites and feeds for testing that are a realistic set of sources for beat or local reporters. Minneapolis, San Diego, Chicago or New York City are good test grounds since they feature many independent news organizations and have fuzzy geographic lines. The effort might build on an open source project from Phase 2 Technology called Tattler and might start with only RSS feeds before expanding to more content.

Keeping Track of Sources

Reporters today often keep source lists in notes in a variety of formats: Excel spreadsheets, Word documents, and address books. A tool to keep track of sources for a beat reporter would have contact information recorded in a database. The tracker would also capture news from the web, however, so that when a source was mentioned in a news report or blog the tracker would note this and could alert the reporter. The tracker could also mine a reporter's own archived stories, so that context and history were readily apparent to the journalist as she considered a source's contribution to a story.

Data Alerts on Trends and Outliers

As the Everyblock website illustrates (http://www.everyblock.com/), it is now possible to get daily updated data feeds about many aspects of city life. Spotting the story in these data points, however, may be very difficult. By setting up the bounds of what changes in data values might indicate something is amiss or afoot, a reporter could be alerted when a trend appears, an anomaly arises, or when a specific individual or entity or location is referred to in the data stream.

Timeline Generator

A timeline generator on a journalist's dashboard could mine stories and information on the web to show at least two chronologies. For a given story, an events timeline would chart the specific incidents mentioned in coverage of a particular story. A coverage timeline would show how coverage of the story unfolded in specific news outlets and blogs. An overlay function would allow you to see when and where particular incidents in the Events timeline were first reported.

Annotator

The annotator on a dashboard would allow a reporter to see past stories, images, references, and contextual information as she wrote a story. As she drafted a story and mentioned an entity such as a politician, contextual information would appear on a side rail that a reporter could choose to click into the story. This function would allow a journalist to provide more easily depth and context, and bring out into the open the source data that often form the basis of statements in a story's text. The dashboard tool would help move much more information from the reporter's laptop to the consumer's screen or paper.

3. Interactions among Readers and Reporters

The text mining tools involved in pattern discovery and the journalist dashboard share a similar goal, to help a reporter discover new accountability stories. Yet, at a broader level, digital technologies have already dramatically reshaped the organizational environment in which reporters work and the forms of the stories they tell. Research in computational journalism should enable reporters to develop new social roles and to sustain watchdog work under a variety of institutional conditions. It should also lead to new ways of telling stories, which ultimately may leave readers better informed and may help news organizations gain more revenue to sustain watchdog coverage.

Phil Bennett, the former managing editor of the *Washington Post*, offered this example of how CJ could have changed the presentation of the 2007 Pulitzer Prize winning investigative series on Walter Reed Army Medical Center (see http://www.washingtonpost.com/wp-dyn/content/article/2007/02/17/AR2007021701172.html). Imagine that as a person read the story about the military's disastrous handling of veterans' health care at Walter Reed, there could be many different layers of information presented. A person could simply read the original text. Or she could also follow links to read original documents, listen to interviews, and follow trails to other veterans' healthcare information news sources. As what she read revealed more of her interests, different types of information could be offered. This would allow a reader to have a differentiated news experience depending on her interests, in a way that used her prior reading and interests as an indicator of what content she might be most interested in (akin to Amazon's reader recommendations). By offering different readers deeply differentiated content, a news outlet would create an experience that would be more distinct and hence less subject to competition from news aggregators.

Bennett pointed out that long-term investigative projects today are like marathons, with publication day being the finish of the race and the start (if the piece is notable) of a short victory lap. He points out, however, that publication online could correctly be seen as the midpoint of the process. A series like the Walter Reed investigation, for example, could become a focal point online for readers interested in veterans' affairs and veterans' healthcare. If the paper could nurture a community of interest around the story, readers might use the site as a discussion place for the action that follows from investigation. The core group interested in the topic might be an ideal target audience for advertisers involved in health care, national defense, and military affairs.

Advances in CJ could thus alter how a story is told by offering, through algorithm, different layers of stories depending on a reader's interest and choices. This could make the originating site more distinctive, attract the subset of readers highly interested in a topic, and help monetize their attention through targeted advertising and sustained attention to the site. It could also help transform journalism from a deadline-driven practice of news reporting into a more blended practice of reporting and social organizing. That is, by taking advantage of the ways in which digital technologies facilitate both speaking with and actually gathering audiences, CJ might help create new blendings of audience, reporter, and commentator. This in turn might help grow the audience for watchdog journalism and enhance the involvement of citizens in the democratic watchdog process.

4. Sensemaking Advances in Other Disciplines

Another source of innovation in CJ is advances made in other disciplines that could be readily transferred to the problems that investigative reporters face. Researchers working on homeland security, digital humanities, political science, and medical research face quandaries about how to draw information from many disparate pieces of information. Consider how each of the following projects could be the basis for tools helpful to reporters:

Homeland Security Research Example: http://www.cc.gatech.edu/gvu/ii/jigsaw/

Jigsaw: Visualization for Investigative Analysis is software developed by a team of researchers at Georgia Tech that offers a visual representation of the connections among individuals and entities that may be mentioned across many different sets of documents. This software helps you identify unlikely associations and determine which documents you may want to read if you are interested in particular connections among specific individuals or groups.

Digital Humanities Research Example: http://www.muninn-project.org/index.html

The Muninn World War I project is a multidisciplinary international research project that aims to translate the information from World War I military forms into searchable databases. The research question for computer scientists in this project is how to use statistical methods to decipher hand-written responses to forms used in this time period. Breakthroughs from this project would be very helpful in analyzing financial disclosures forms. Andy Hall of the Wisconsin Center for Investigative Journalism points out that in Wisconsin, for example, nearly 2,700 officials annually file financial disclosure forms that provide data on investments, directorships, and income sources. The filings are in hard copy and often contain handwritten responses. The state's Government Accountability Board has not taken action to put the forms online or make the information easily translatable into a database. Advances in form processing in Digital Humanities research, however, could easily transfer into tools helpful to journalists trying to analyze disclosure forms.

Political Science Research Example: http://www.comparativeagendas.org/

Professors Frank Baumgartner, Bryan Jones, and John Wilkerson have developed methods to analyze changes in the US public policy agenda since WW II (http://www.policyagendas.org/). A challenge in studying the ways different issues surface in legislation and media stories is creating a way to classify the many topics addressed in millions of different bills and stories. While hand-coding a sample of laws and stories was once the most feasible way to proceed to study agenda change, recent advances in text mining are allowing political scientists to automate parts of the tasks of categorizing documents based on the political topics they discuss. The Comparative Agendas Project uses automatic text analysis tools developed by Professor Paul Wolfgang of Temple University to analyze changes in issue agendas across many different countries. This type of tool could be modified to enable reporters to classify how individuals or

organizations have changed their focus or priorities over time through analyzing documents and statements associated with the politician or group.

Medical and Health Informatics Research Example: http://ils.unc.edu/~cablake/evid/

The rate of article production in fields such as medicine and biology make it difficult for any one researcher to keep current with key findings. The vast amount of medical data available makes it hard to synthesize results and spot unexpected outcomes. Increasingly, content analysis tools are being developed in this area to do text mining that isolates fact claims, spots contradictions, and surfaces "secondary information" that may not be the focus of an article but may arise from its results. Catherine Blake's research, such as her NSF funded project on Evidence-Based Discovery in health sciences, tries to use content analysis to help researchers identify claims made in articles. This type of tool could be helpful once developed to reporters seeking to establish the claims made across different documents dealing with government activities.

Some Likely Effects of Computational Journalism

In sum, the CASBS Summer Workshop participants identified four areas as potential sources for advances in CJ: techniques for data transformation and pattern discovery; a digital dashboard for journalists; new watchdog roles for readers as well as reporters; and narratives and spillovers from cutting edge research in areas such as homeland security, digital humanities, and medical research. Key points to note are:

Computers Won't Replace People

In all these areas the algorithms will surface data and ideas to be further explored by reporters. These are tools to supplement rather than substitute for efforts by reporters. These tools will essentially involve data mining in the public interest. Though the phrase computational journalism carries for some the suggestion of robotic reporters, it is really through the interaction of practitioners in the fields of computer science and journalism that tools will be developed that will be used by reporters to discover new stories.

New Tools Will Engage New Players in Watchdog Journalism

The tools developed will also help change the journalism ecosystem. By lowering the cost of investigating government activity, computational journalism will help spread the watchdog function to a larger set of eyes. Small nonprofit reporting outlets, deeply engaged citizens, and NGOs involved in policy debates can all use the text mining tools and dashboard software envisioned in this report. While computer assisted reporting was often viewed as the province of a subset of investigative reporters, a goal of CJ algorithms is that the tools can be readily used by anyone interested in following the performance of public and private institutions.

New Tools Will Change the Data Too

The data and documents used may be imprecise, dirty, and misleading. Some hunches suggested will be investigated and proven incorrect. As raw data are shared with readers and the reporting process is made more transparent, indicators of data quality and provenance will need to be developed.

Tools Will Need to Be Open-Source and Easy to Use

The tools developed for reporters will likely need to be open-source or carry a very low cost of acquisition, since local papers and online news providers will be hard-pressed to make investments in accountability coverage. The tools will need to be easy to operate too, since journalists may not be given the time or training to use complex algorithms.

Funding Will Have to Come from Outside Professional Journalism

The funding for CJ innovations will need to come from academia, government, nonprofits, and foundations. Few media organizations are currently willing to invest substantial funds in areas that are not readily monetized.

How Is The Field of Computational Journalism Evolving? Why Does This Matter?

In February 2008 Prof. Irfan Essa of Georgia Tech convened *Journalism 3G: The Future of Technology in the Field, a symposium on Computation + Journalism*. This conference, described at http://www.computation-and-journalism.com/main/, brought together the key parties involved in the creation of the new field of computational journalists: computer scientists, engineers, journalists, and communication researchers. This meeting helped spark coverage of the new field, in articles such as "Deep Throat Meets Data Mining" (http://www.miller-mccune.com/media/deep-throat-meets-data-mining-875), "Tracking Toxics When Data Are Polluted" (http://www.nieman.harvard.edu/reportsitem.aspx?id=100933), and "Can Computer Nerds Save Journalism?(http://www.time.com/time/business/article/0,8599,1902202,00.html). Professors such as Rich Gordon at Northwestern, Brant Houston at the University of Illinois, and researchers at Stanford's Department of Communication and Harvard's Berkman Center for Internet and Society are pursuing teaching and research projects that are helping to bring together software developers and journalists. In July 2009 Sarah Cohen became the first chaired professor in the field of computational journalism when she became the Knight Professor at Duke University (http://news.duke.edu/2009/04/cohen.html). In February 2010 Georgia Tech will host a second conference on computation and journalism, aimed at advancing the research agenda in the field.

These developments suggest that academia may become a source for the innovations that will help advance the development of CJ tools. The participants at the CASBS workshop also noted that new organizations were moving to form a "middle layer" of public information providers. Working with an array of public data sources, organizations such as MAPLight.org, GovTrack.us, OpenSecrets.org, ProPublica.org, EveryBlock.com, and others have rapidly arisen to mediate between public databases and audiences of laymen and reporters. Most of these organizations are non-profit, though some are for-profit entities. Taken individually, they represent a fascinating set of experiments in using computing to inform the public. Taken together however, we believe they constitute an emerging infrastructure for the provision of public information. As journalistic institutions disaggregate and journalists come to work in smaller units or independently, this new infrastructure will represent a key public resource. We urgently need to understand how that resource is coming into being, how different funding models affect its journalistic and pro-democratic potential, and how we can sustain those elements of it we value most.

It is too early to demonstrate the impact of computational journalism, since the tools for reporters and partnerships among academia and journalists remain to be developed. At its heart the field will involve using computing power to lower the cost of accountability coverage. Three recent examples that attempt to do this are:

Public Insight Journalism: http://americanpublicmedia.publicradio.org/publicinsightjournalism/

American Public Media has helped Minnesota Public Radio and others develop a Public Insight Network, a database which has detailed data on the demographics, interest, and expertise of more than 70,000 individuals. Radio journalists have used the database to query listeners about issues they care and know about. As the APM puts it, "Our network of more than 70,000 public sources

has helped us find and report stories on the growing obesity epidemic in rural areas, the decline of labor unions, and the impact of the Iraq War on families of soldiers."

Crowdsourcing at the Guardian: http://www.niemanlab.org/2009/06/four-crowdsourcing-lessons-from-the-guardians-spectacular-expenses-scandal-experiment/

Faced with looking through electronic copies of hundreds of thousands of newly released expense reports from Members of Parliament, *The Guardian* turned to crowd-sourcing. The paper created a site that allowed readers to look at a document and flag it as Not Interesting, Interesting but Known, Interesting, and Investigate This (http://mps-expenses.guardian.co.uk/). The game-like setup led more than 20,000 volunteers to review more than 170,000 documents over the first 80 hours of the operation of the site. These readers flagged expense reports that were then further investigated by *The Guardian*.

Stimulus Spending at the Local Level: http://www.recovery.gov/?q=content/recipient-reporting

On October 30, 2009 the federal government will make available the first detailed data on recipients of funding under the stimulus program. Reporters will be hard pressed to parse and analyze the data, which will not be provided in a user-friendly format that would be easily digestible by local reporters. Leading an effort that will draw upon student software developers, nonprofit reporting centers, and Investigative Reporters and Editors, Duke's Sarah Cohen will work to devise a setup ahead of the data release so that once the information is released by the government it can be quickly parsed and distributed to local reporters.

In an era when media organizations are shedding reporters and beats, the development of computational journalism offers a way to expand the reach of both professional and citizen journalists. The increasing amounts of data becoming available at the federal level mean that CJ tools will offer a way to capitalize on (anticipated) government transparency. Beth Noveck, Deputy Chief Technology Officer for Open Government in the Obama administration, described in her workshop interactions how federal sites such as www.data.gov and local sites such as www.datasf.org will provide increasingly large amounts of information for journalists. Civic hackers such as Josh Tauberer, the founder of GovTrack.us, will have increasing amounts of unstructured government data to organize and analyze. Workshop participant Jun Yang stressed that the new streams of data becoming available and the time-sensitive nature of reporting raise familiar research challenges for computer scientists dealing with massive and often uncertain datasets, including scalable continuous querying and data mining, data provenance, cost-benefit optimization of data acquisition, privacy-preserving data publishing, and reasoning with uncertainty.

Workshop participants Rayvon Fouché and Lucy Suchman stressed that the CJ tools will also have an influence on the types of organizations that evolve to perform the watchdog function of journalism. Algorithms will need to be easy to use and data will need to be widely available if reporters across beats are to adopt them. Moreover, they will need to be designed in ways that do

not entirely overthrow the traditional practices of watchdog reporting. In the past, techniques of computer-assisted reporting tended to be the province of a specialized subset of investigative reporters. Yet, we believe the tools of CJ will also be adopted by citizen journalists, nonprofit news outlets, and NGOs working on government accountability. In order to make that adoption work, however, Fouché and Suchman noted that future computational journalists will need to be as innovative in thinking about the organization of their work process as they are about the technologies they deploy.

Next Steps

Computational journalism offers the prospect of lowering the costs of accountability reporting, a genre increasingly at risk in local media markets given its high cost and difficulty in monetizing. The CASBS Summer Workshop outlined how the development of text mining tools and a journalist's dashboard could speed the discovery of investigative stories. The pursuit of configuring projects which draw together people from multiple disciplines will demonstrate the degree to which CJ can lead to the development of stories that would not otherwise be told.

The CASBS participants believed that if effective and low-cost reporting tools were built, both professional and citizen journalists would use them. As in many questions surrounding information markets, the prime question is who will pay for the creation of these tools. Many types of actors may have a role in building this field:

Foundations are a driving force today in media experimentation. The Knight Foundation's willingness to redefine the Knight Chair at Duke to be a position in computational journalism is an early sign of faith in this field. The recent round of winners in the Knight News Challenge included projects that may lead to the development of algorithms and datasharing for reporters. The Document Cloud project (http://www.newschallenge.org/winner/2009/document-cloud), funded by the Knight Foundation, will pioneer the process of transforming a large collection of documents into a searchable database for reporters. Foundations particularly interested in sustaining accountability reporting and public affairs coverage should find investments in CJ projects to yield high impacts. Since the reporting tools will be widely applicable across many geographic and topic areas, the algorithms developed will be scalable, have definable goals, and yield measurable impacts in terms of story production and (in some cases) policy impacts.

Government agencies such as the National Science Foundation and the National Endowment for the Humanities have funded successful cyber-research projects in areas as diverse as health care, terrorism, and literature. The development of computational journalism can be speeded by the transformation of products already developed by government support into tools that can be used by journalists. Agencies may also want to hold funding competitions explicitly for the development of accountability algorithms that could be used by reporters or citizen journalists. The text mining and dashboard tools would also be of use to government workers. People working in public health or environmental policy agencies, for example, could use these tools to examine topics that they follow as part of their interest in social and environmental outcomes. Government efforts to push out raw data in forms that can be used by reporters, including the development of APIs on government websites, can also lower the costs of producing stories.

Academic research centers can use their convening power to draw together university researchers to focus on CJ tools. The DeWitt Wallace Center for Media and Democracy at Duke, for example, is drawing together journalists, communications scholars, and computer scientists to work on a suite of reporting tools. Similar multidisciplinary efforts are underway at Georgia Tech, Harvard, Northwestern, Stanford, and the University of Illinois. Most of these efforts are

not currently centered at journalism schools. If journalism schools used discretionary endowments to fund research and development in this field, however, they could become true incubators for tools that would be readily taken up by reporters (and readers).

Nonprofits are often the favored organizational form of groups that help transform government data into useable formats. More reporting is being done at the local and state level by newly formed nonprofits, such as MinnPost and the Center for Investigative Reporting's California Watch program. Nonprofits can play a key role in the evolution of computational journalism, especially if they are willing to share the data they develop with other reporting outlets and are willing to partner with innovative academic programs. Investigative Reporters and Editors can serve as a partner for those hoping to transmit their new reporting tools to investigative journalists in the field.

Open source developers can help translate tools developed in other fields into reporting algorithms. Progressive programming firms often encourage their employees to work part-time on open source code. For many software developers, contributing to code offers a form of expression and a way to help others. The potential impact of the reporting made possible through computational journalism has already begun to attract developers to the field who are willing to push the growth of open source tools. The Drupal community at Stanford University has already begun to organize efforts to help create a journalist's dashboard. Note that the algorithms used by reporters will need to be both easy to use and cheap to procure, since media companies are unlikely to invest in software development.

Journalists and citizens interested in monitoring institutional performance will ultimately determine how quickly and how far the field of computational journalism develops. The best algorithms will essentially be public interest data mining. They will generate leads, hunches, and anomalies to investigate. It will remain for reporters and others interested in government performance to take the next step of tracking down the story behind the data pattern. With fewer reporters walking fewer beats, however, it will be up to foundations, government agencies, academic institutions, nonprofits, and software developers to summon the resources and creativity to develop the computational tools that will help preserve the watchdog function of journalism.

Readers of this report can also play a role in the evolution of this field. This is a snapshot of computational journalism from the vantage point of October 2009. To see how the field is continuing to change and explore the role you might play, we hope you'll go to www.sanford.duke.edu/centers/dewitt and to http://www.computation-and-journalism.com/main/.

Appendix: Developing the Field of Computational Journalism, a Center for Advanced Study in the Behavioral Sciences Summer Workshop, July 27-31, 2009

The CASBS summer workshop on computational journalism brought together journalists, researchers, and NGO representatives to focus on how to develop this nascent field. The titles for the daily agendas for the week evoke the wideranging nature of the conversations: the journalistic ecosystem; data; algorithms and interface design; thinking with the social sciences; and next steps. The working group members were:

Phil Bennett, Eugene C. Patterson Professor of the Practice of Journalism and Public Policy, Duke University

Catherine Blake, Associate Professor in the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

Sarah Cohen, Knight Professor of the Practice of Journalism and Public Policy, Duke University

Irfan Essa, Professor in the School of Interactive Computing of the College of Computing, and Adjunct Professor in the School of Electrical and Computer Engineering, Georgia Institute of Technology

Rayvon Fouché, Associate Professor of History, University of Illinois at Urbana-Champaign

James T. Hamilton, Charles S. Sydnor Professor of Public Policy, and Director of the DeWitt Wallace Center for Media and Democracy, Duke University

Phil Meyer, Professor Emeritus, School of Journalism and Mass Communication, University of North Carolina at Chapel Hill

Lucy Suchman, Professor, Sociology Department and the Centre for Science Studies, Lancaster University UK

Joshua Tauberer, Software Developer, Civic Hacker, Ph.D. in the linguistics (expected 2010) from University of Pennsylvania

Fred Turner, Assistant Professor and Director of Undergraduate Studies in the Department of Communication, Stanford University

Jun Yang, Associate Professor of Computer Science, Duke University

We also benefitted from the willingness of other scholars and journalists to make presentations to the working group and engage in discussions about how computational journalism might evolve through their work. These presenters included:

Jim Bettinger, Director of the John S. Knight Fellowships, Stanford University

Zach Chandler, Academic Technology Specialist**,** Stanford University

Louis Freedberg, California Watch Director for the Center for Investigative Reporting and the Founder and Director of California Media Collaborative

Andrew Haeg, John S. Knight Fellow**,** Stanford University

Barry Hayes, Google News

Jeff Heer, Assistant Professor of Computer Science, Stanford University

Carl Malamud, Technologist, Author, Public Domain Advocate, and Founder of Public.Resource.Org

Christopher Manning, Associate Professor of Computer Science and Linguistics and Sony Faculty Scholar, Stanford University

Daniel Newman, Co-Founder and Executive Director of MAPLight.org

Beth Noveck, Deputy Chief Technology Officer for Open Government, Office of Science and Technology Policy, Executive Office of the President