

THE ART AND SCIENCE OF DATA-DRIVEN JOURNALISM

**WHEN JOURNALISTS COMBINE
NEW TECHNOLOGY WITH
NARRATIVE SKILLS, THEY CAN
DELIVER CONTEXT, CLARITY,
AND A BETTER UNDERSTANDING
OF THE WORLD AROUND US.**

ALEXANDER
BENJAMIN
HOWARD

Executive Summary

Journalists have been using data in their stories for as long as the profession has existed. A revolution in computing in the 20th century created opportunities for data integration into investigations, as journalists began to bring technology into their work. In the 21st century, a revolution in connectivity is leading the media toward new horizons. The Internet, cloud computing, agile development, mobile devices, and open source software have transformed the practice of journalism, leading to the emergence of a new term: data journalism.

Although journalists have been using data in their stories for as long as they have been engaged in reporting, data journalism is more than traditional journalism with more data. Decades after early pioneers successfully applied computer-assisted reporting and social science to investigative journalism, journalists are creating news apps and interactive features that help people understand data, explore it, and act upon the insights derived from it. New business models are emerging in which data is a raw material for profit, impact, and insight, co-created with an audience that was formerly reduced to passive consumption. Journalists around the world are grappling with the excitement and the challenge of telling compelling stories by harnessing the vast quantity of data that our increasingly networked lives, devices, businesses, and governments produce every day.

While the potential of data journalism is immense, the pitfalls and challenges to its adoption throughout the media are similarly significant, from digital literacy to competition for scarce resources in newsrooms. Global threats to press freedom, digital security, and limited access to data create difficult working conditions for journalists in many countries. A combination of peer-to-peer learning, mentorship, online training, open data initiatives, and new programs at journalism schools rising to the challenge, however, offer reasons to be optimistic about more journalists learning to treat data as a source.

Recommendations and Predictions

- 1) Data will become even more of a strategic resource for media.
- 2) Better tools will emerge that democratize data skills.
- 3) News apps will explode as a primary way for people to consume data journalism.
- 4) Being digital first means being data-centric and mobile-friendly.
- 5) Expect more robo-journalism, but know that human relationships and storytelling still matter.
- 6) More journalists will need to study social science and statistics.
- 7) Data journalism will be held to higher standards for accuracy and corrections.
- 8) Competency in security and data protection will become more important.
- 9) Audiences will demand more transparency on reader data collection and use
- 10) Conflicts will emerge over public records, data scraping, and ethics.
- 11) Collaboration will arise with libraries and universities as archives, hosts, and educators.
- 12) Expect data-driven personalization and predictive news in wearable interfaces.
- 13) More diverse newsrooms will produce better data journalism.
- 14) Be mindful of data-ism and bad data. Embrace skepticism.

Table of Contents

I.	Introduction.....	4
II.	History.....	7
	a. Overview	
	b. Data in the Media	
	c. Rise of the Machines	
	d. Computer-assisted Reporting	
	e. The Internet Inflection Point	
III.	Why Data Journalism Matters.....	14
	a. Shifting Context	
	b. The Growth of the News App	
	c. On Empiricism, Skepticism, and Public Trust	
	d. Newsroom Analytics	
	e. Data-driven Business Models	
	f. New Nonprofit Revenues	
	g. Fuel for Robo-journalism	
IV.	Notable Examples.....	30
	a. National and International Data Journalism Awards	
	b. Data and Reporting Paired with Narrative	
	c. Crowdsourcing Data Creation and Analysis	
	d. Public Service	
	e. All Organizations—Great and Small	
	f. Embracing Data Transparency	
	g. Following the Money	
	h. Mapping Power and Influence	
	i. Geojournalism, Satellites, and the Ground Truth	
	j. Working Without Freedom of Information Laws	
	k. Data Journalism and Activism	

V. Pathway to the Profession.....	44
a. Mentorship, Numeracy, Competition, Recruiting	
b. Massive Open Online Courses (MOOC) to the Rescue?	
c. Hacks, Hackers, and Peer-to-peer Learning	
d. Journalism Schools Rise to the Challenge	
VI. Tools of the Trade.....	58
a. Digging into the CAR Toolbox	
b. New Tools to Wrangle Unstructured Data	
VII. Open Government.....	61
a. Open Data and Raw Data	
b. Data and Ethics	
c. Gun Data, Maps, and Radical Transparency	
d. A Global Lens on FIOA and Press Freedom	
VIII. On to the Future.....	69
Recommendations and Predictions	
IX. Appendices.....	81
a. Acknowledgments	
b. About the Author	
c. Interviews	
d. Endotes	

I. Introduction

Today, the world is awash in unprecedented amounts of data and an expanding network of sources for news. As of 2012, there were an estimated 2.5 quintillion bytes of data being created daily, or 2.5 exabytes, with that amount doubling every 40 months. (For the sake of reference, that's 115 million 16-gigabyte iPhones.)

It's an extraordinary moment in so many ways. All of that data generation and connectivity have created new opportunities and challenges for media organizations that have already been fundamentally disrupted by the Internet.

To paraphrase author William Gibson, in many ways the post-industrial future of journalism is already here—it's just not evenly distributed yet.ⁱ Stories are now shared by socially connected friends, family, and colleagues—and delivered by applications and streaming video accessed from mobile devices, apps, and tablets. Newsrooms are now just a component, albeit a crucial one, of a dramatically different environment for news. They are also not always the original source for it. News often breaks first on social networks, and is published by people closest to the event. From there, it's gathered, shared, and analyzed; then fact-checked and synthesized into contextualized journalism.

Media organizations today must be able to put data to work quickly.ⁱⁱ This need was amply demonstrated during Hurricane Sandy, when public, open government data feeds became critical infrastructure.ⁱⁱⁱ Given a 2014 Supreme Court decision in which Chief Justice John Roberts opined that disclosure through online databases would balance the effect of classifying political donations as protected by the First Amendment, it's worth emphasizing that much of the “modern technology” that is a “particularly effective means of arming the voting public with information” has been built and maintained by journalists and nonprofit organizations.^{iv}

The open question in 2014 is not whether data, computers, and algorithms can be used by journalists in the public interest, but rather how, when, where, why, and by whom.^v Today, journalists can treat all of that data as a source, interrogating it for answers as they would a human.

That work 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.^{vi}

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.

Some proponents of [open data journalism](#) hold that there should be four components, where data journalists archive and publish the underlying raw data behind their investigations, along with the methodology and code used in the analyses that led to their published conclusions.^{vii}

In a broad sense, data journalism is telling stories with numbers, or finding stories in them. It's treating data as a source to complement human witnesses, officials, and experts. Many different kinds of journalists use data to augment their reporting, even if they may not define themselves or their work in this way.

"A data journalist could be a police reporter who's managed to fit spreadsheet analysis into her daily routine, the computer-assisted reporting specialist for a metro newspaper, a producer with a TV station investigative unit, someone who builds analysis tools for journalists, or a news app developer," said [David Herzog](#), an associate professor at the Missouri School of Journalism.

Consider four examples: A financial journalist cites changes in price-to-earning ratios in stocks over time during a radio appearance. A sports journalist adds a table that illustrates the on-base percentages of this year's star rookie baseball players. A technology journalist creates a graph comparing how many units of competing smartphones have been sold in the last business quarter. A team of news developers builds an interactive [website that helps parents find nearby playgrounds](#) that are accessible to all children and adds data about it to a public data set.^{viii}

In each case, journalists working with data must be conscious about its source, the context for its creation, and its relationship to the stories they're telling.

"Data journalism is the practice of finding stories in numbers and using numbers to tell stories," said Meredith Broussard, an assistant professor of journalism at Temple University.

To become a good data journalist, it helps to begin by becoming a good journalist. Hone your storytelling skills, experiment with different ways to tell a story, and understand that data is created by people. We tend to think that data is this immutable, empirically true thing that exists independent of people. It's not, and it doesn't. Data is socially constructed. In order to understand a data set, it is helpful to start with understanding the people who created the data set—think about what they were trying to do, or what they were trying to discover. Once you think about those people, and their goals, you're already beginning to tell a story.

Data-driven reporting and analysis require more than providing context to readers and sorting fact from fictions and falsehoods in vast amounts of data. Achieving that goal will require media organizations that can think differently about how they work and whose contributions they value or honor. In 2014, technically gifted investigators in the corner of the newsroom may well be of more strategic value to a media company than a well-paid pundit in the corner office.

Publishers will need to continue to evolve toward a multidisciplinary approach to delivering the news, where reporters, developers, designers, editors, and community managers collaborate on storytelling, instead of being segregated by departments or buildings.

Many of the pioneers in this emerging practice of [data-driven journalism](#) won't be found on

broadcast television or in the lists of the top journalists over the past century. They're drawn from the pool of people who are building collaborative newsrooms and extending the theory and practice of data journalism. These people see the reporting that provisions their journalism *as* data, a body of work that itself can be collected, analyzed, shared, and used to create insights about how society, industry, or government are changing.^{ix}

In the following report, I look at what the media is doing, offer insights from data journalists, list the tools they're using, share notable projects, and look ahead at what's to come—and what's needed to get there. You'll also find more to read and consider in the *Data Journalism Handbook* that O'Reilly Media published in 2012.^x

II. History

a. Overview

As Liliana Bounegru highlighted in the introduction to the *Data Journalism Handbook*, this idea of treating data as a source for the news is far from novel: Journalists have been using data to improve or augment traditional reporting for centuries.^{xi}

In the [Guardian's ebook on data journalism](#), Simon Rogers (now [Twitter's first data editor](#)) said that the first example of data journalism at the *Guardian* newspaper was back in 1821, reporting student enrollment and associated costs.^{xii}

The data journalism construction, however, is very much of the 21st century, although its origin is murky. Rogers said he heard the term [data journalism used first by software developer Adrian Holovaty](#),^{xiii} but it's possible that it may have originated earlier [somewhere else in Europe](#),^{xiv} in [conversations about database journalism](#) of the kind Holovaty advocated.^{xv} The European journalists I asked about [the term's origins](#),^{xvi} however, pointed back to Holovaty as the [patron saint of data journalism](#).^{xvii}

Holovaty, a talented software developer at the *Washington Post* and founder of [EveryBlock](#), decried how data was organized and treated by media organizations in a 2006 post on [how newspaper websites needed to change](#).^{xviii} As Holovaty noted in a postscript, his essay [inspired the creation](#)^{xix} of [PolitiFact](#) by Bill Adair and Matt Waite. The fact-checking website subsequently won the Pulitzer Prize in 2009.^{xx}

The use of data journalism gained momentum around the world after Tim Berners-Lee called [analyzing data the future of journalism](#) in 2010, as part of a larger conversation around opening government data up to the public through publishing it online.^{xxi}

The year before, the *Guardian* had launched its Datablog. Using structured data extracted from the PDF that the [United Kingdom's Parliament published](#) online,^{xxii} the *Guardian* [visualized the expenses of Ministers of Parliament](#), launching a public row about their spending that has continued into the present day.^{xxiii} In July 2010, the [Guardian began publishing data journalism based on the War Logs](#),^{xxiv} a massive disclosure of thousands of Afghanistan war records leaked through [Wikileaks](#).

Over the following years, the use of the term [data journalism began to catch fire](#), at least within the media world.^{xxv} The usage was [adopted by David Kaplan](#), a pillar of the investigative journalism community, and used as self-identification by many attendees of the annual conference of the National Institute for Computer-Assisted Reporting (NICAR), where [nearly a thousand journalists from 20 countries](#) gathered in Baltimore to teach, learn, and connect.^{xxvi}

It was in 2014, however, that data journalism entered mainstream discourse, driven by the highly publicized relaunch of Nate Silver's FiveThirtyEight.com and Vox Media's April release of

general news site Vox.com, as well as new ventures from the *New York Times* and *Washington Post*.

On that count, it's worth noting a broader challenge that the data journalism mainstream presents: the novelty of the term has divorced it from the long history of computer-assisted reporting that came before in public discourse. Hopefully, this report will act as a corrective on that count.

Today, the context and scope of data-driven journalism have expanded considerably from its evolutionary antecedent, following the explosion of data generated in and about nearly every aspect of society, from government, to industry, to research, to social media. Data journalists can now use free, powerful online tools and open source software to rapidly collect, clean, and publish data in interactive features, mobile apps, and maps. As data journalists grow in skill and craft, they move from using basic statistics in their reporting to working in spreadsheets, to more complex data analysis and visualization, finally arriving at computational journalism, the command line, and programming. The most advanced practitioners are able to capitalize on algorithms and vast computing power to deliver new forms of reporting and analysis, from [document mining](#) applied to find misconduct,^{xxvii} to [reverse engineering](#) political campaigns,^{xxviii} price discrimination, executive stock trading plans, and autocompletions.

Data journalists are in demand today throughout the news industry and beyond. They can [get scoops](#), [draw large audiences](#), and augment the work of other journalists in a media organization or other collaboration. By automating common reporting tasks, for instance, or creating custom alerts, one data journalist can increase the capacity of the people with whom she works, building out databases that may be used for future reporting.

“On every desk in the newsroom, reporters are starting to understand that if you don’t know how to understand and manipulate data, someone who can will be faster than you,” said Scott Klein, a managing editor at ProPublica. He continued:

Can you imagine a sports reporter who doesn’t know what an on-base percentage is? Or doesn’t know how to calculate it himself? You can now ask a version of that question for almost every beat. There are more and more reporters who want to have their own data and to analyze it themselves. Take, for example, my colleague, Charlie Ornstein. In addition to being a Pulitzer Prize-winner, he’s one of the most sophisticated data reporters anywhere. He pores over new and insanely complex data sets himself. He has hit the edge of Access’ abilities and is switching to SQL Server. His being able to work and find stories inside data independently is hugely important for the work he does.

There will always be a place for great interviewers, or the eagle-eyed reporter who finds an amazing story in a footnote on page 412 of a regulatory disclosure. But, here comes another kind of journalist who has data skills that will sustain whole new branches of reporting.^{xxix}

b. On Data in the Media

In many ways, journalists have been engaged in gathering trustworthy data and publishing it for as long as journalism itself has been practiced. The need for reported accuracy about the world is part of the origin story of newspapers five centuries ago in Renaissance Europe.

These newsletters had historical antecedents in the *Acta Diurna* (daily gazette) of the Roman Empire and the *tipao* (literally, “reports from the official residences”) of the Han dynasty in China hundreds of years prior, where governments produced and circulated news of military campaigns, politics, trials, and executions.

Five centuries ago, Italian merchants commissioned and circulated handwritten newsletters that reported news of economic conditions, from the cost of commodities to the disruption of trade by revolutions, wars, disease, or severe weather. The printed versions that followed in the 17th century, once the cost of paper fell and printing presses proliferated, included these same basic lists of data, as did the printed newsbooks that circulated in the next century.

After Scottish engineer and political economist William Playfair invented graphical methods for displaying statistics in 1786, periodicals began to use line graphs, bar charts, pie charts, and circle graphs.

“As technology got better in the late 18th century and readers started demanding a different kind of information, the data that appeared in newspapers got more sophisticated and was used in new ways,” said Scott Klein. “Data became a tool for middle-class people to use to make decisions and not just as facts to deploy in an argument, or information useful to elite business people.”

By the end of the 19th century, statistics were a part of stories in many newspapers, whether they appeared as figures, lists, or raw data about commodities or athletics that readers could pore over and consult themselves.

Long before stock market data systems went electronic, newspapers published prices to investors. Dow Jones & Company began publishing stock market averages in 1884 and continues to do so today in both print and online via the *Wall Street Journal*.

c. Rise of the Newsroom Machines

By the middle of the 20th century, investigative journalism featured teams of professional reporters combing through government statistics, court records, and business reports acquired by visiting state houses, archives, and dusty courthouse basements; or obtaining official or leaked confidential documents. These lists of numbers and accounts in the ledgers and filing cabinets of the world’s bureaucracies have always been a rich source of data, long before data could be published in a digital form and shared instantaneously around the world.

Database-driven journalism arrived in most newsrooms in a real sense over three decades ago, when microcomputers became commonplace in the 1980s—although the first pioneers used punch cards. When computers became both accessible and affordable to newsrooms, however, the way data could be used changed how investigations were conducted, and much more. Before the first laptop entered the newsroom, technically inclined reporters and editors had found that crunching numbers on computers on mainframes, microcomputers, and servers could enable more powerful investigative journalism.

d. Computer-assisted Reporting

While the various [histories of the development of computer-assisted reporting](#) offer context for

the work of today, most historians place its start in the latter half of the 20th century.^{xxx} Casual observers may not realize that many aspects of what is now frequently called data journalism are the direct evolutionary descendants of decades of computer-assisted reporting (CAR) in the United States. In fact, computing pioneer Grace Hopper, a computing pioneer, professor, and U.S. Navy rear admiral during World War II, made prescient predictions long before Nate Silver's electoral prognostications made him a media star.

In 1952, CBS famously used a mainframe computer, a Remington Rand UNIVAC, and statistical models to predict the outcome of the presidential race.^{xxxi} Meanwhile, Grace Hopper worked with a team of programmers to input voting statistics from earlier elections into the ENIAC and wrote algorithms that enabled the computer to correctly predict the result.

The model she built not only accurately predicted the ultimate outcome—a landslide victory for Dwight D. Eisenhower—with just 5 percent of the total vote in, but did so to within one percent. (Their calculations predicted 83.2 percent of electoral votes for Eisenhower; in actuality he received 82.4 percent.) Grace Hopper and her team used the ENIAC to accomplish something quite similar to what Nate Silver does six decades later: defy the election predictions of political pundits by using statistical modeling.



[UNIVAC designer J. Presper (“Pres”) Eckert and operator Harold Sweeney demonstrate the computer to CBS anchor Walter Cronkite. Photo Credit: Wikimedia Commons.]

In the years that followed this signal media event, change was slow, marked by pioneers experimenting with computer-assisted reporting in investigations. It was almost two more decades before CAR pioneers like Meyer Elliot Jaspin and Philip Meyer began putting cheaper, faster computers to work, collecting and analyzing data for investigative journalism. After he was granted a Nieman Fellowship at Harvard University in the late 1960s to study the application of quantitative methods used in social science, Philip Meyer proposed applying these social science research methods to journalism using computers and programming. He called this “[precision journalism](#), which included sound practices for data collection and sampling, careful analysis and clear presentation of the results of the inquiry.”^{xxxii}

Meyer subsequently applied that methodology to investigating the [underlying causes of rioting in Detroit in 1967](#),^{xxxiii} a contribution that was cited when the *Detroit Free Press* won the Pulitzer Prize for Local General Reporting the next year. Meyer’s analysis showed that college graduates were as likely to have participated in the riots as high school dropouts, rebutting one popular theory correlating economic and educational status with a propensity to riot, and another regarding immigrants from the American South. Meyer’s investigations found that the primary drivers for the Detroit riots were lack of jobs, poor housing, crowded living conditions, and police brutality.

In the following decades, journalists around the country steadily explored and expanded how data and analysis could be used to inform reporting and readers. Microcomputers and personal computers changed the practice and forms of CAR significantly as the tools and environment available to journalists expanded. More people began waking up to “[newsmen enlisting the machine](#),” as *Time* magazine put it in 1996.^{xxxiv} By the early 1990s, [journalists were using CAR techniques and databases](#) in many major investigations in the United States and beyond.

Data-driven reporting increasingly became part of the work behind the winners of journalism’s most prestigious prize: From Eliot Jaspin’s Pulitzer at the *Providence Journal* in 1979, to the work of [Chris Hambly](#) at the Center for Public Integrity in 2014, CAR has mattered to important stories.^{xxxv}

Brant Houston, former executive director of Investigative Reporters and Editors (IRE), said in an interview:

The practice of CAR has changed over time as the tools and environment in the digital world has changed. So it began in the time of mainframes in the late 60s and then moved onto PCs (which increased speed and flexibility of analysis and presentation) and then moved onto the Web, which accelerated the ability to gather, analyze, and present data. The basic goals have remained the same. To sift through data and make sense of it, often with social science methods. CAR tends to be an umbrella term—one that includes precision journalism and data-driven journalism and any methodology that makes sense of data, such as visualization and effective presentations of data.

By 2013, CAR had been recognized as an important journalistic discipline, as the assistant director of the Tow Center, Susan McGregor, [explored](#) last year in a *Columbia Journalism Review* article.^{xxxvi} Data had become not only an integral part of many prize-winning investigations, but also the raw material for applications, visualizations, audience creation, revenue, and tantalizing scoops.

e. An Internet Inflection Point

At the start of the 21st century, a revolution in mobile computing; increases in online connectivity, access, and speed; and explosion in data creation fundamentally changed the landscape for computer-assisted reporting.

“It may seem obvious, but of course the Internet changed it all, and for a while it got smushed in with trying to learn how to navigate the Internet for stories, and how to download data,” said Sarah Cohen, a *New York Times* investigative journalist and a former Knight professor of the practice of journalism and public policy at Duke University. She added:

Then there was a stage when everyone was building internal intranets to deliver public records inside newsrooms to help find people on deadline, etc. So for much of the time, it was focused on reporting, not publishing or presentation. Now the data journalism folks have emerged from the other direction: People who are using data obtained through APIs often skip the reporting side, and use the same techniques to deliver unfiltered information to their readers in an easier format than the government is giving us. But I think it’s starting to come back together—the so-called data journalists are getting more interested in reporting, and the more traditional CAR reporters are interested in getting their stories on the Web in more interesting ways.

Given the universality of computer use today among the media, the term computer-assisted reporting now feels dated, itself inherited from a time when computers were still a novelty in newsrooms. There’s probably not a single reporter or editor working in a newsroom in the United States or Europe today, after all, who isn’t using a computer in the course of his or her journalism.

Many members of the media, in fact, may use several during the day, from the powerful handheld computers we call smartphones, to crunching away at analysis or transformations on laptops and desktops, to relying on servers and cloud storage for processing big data at Internet scale.

Much has changed since Philip Meyer’s pioneering days in the 1960s, offered Scott Klein:

One is that the amount of data available for us to work with has exploded. Part of this increase is because open government initiatives have caused a ton of great data to be released. Not just through portals like data.gov—getting big data sets via FOIA has become easier, even since ProPublica launched in 2008.

Another big change is that we’ve got the opportunity to present the data itself to readers—that is, not just summarized in a story but as data itself. In the early days of CAR, we gathered and analyzed information to support and guide a narrative story. Data was something to be summarized for the reader in the print story, with of course graphics and tables (some quite extensive), but the end goal was typically something recognizable as a words-and-pictures story.

What the Internet added is that it gave us the ability to show to people the actual data and let them look through it for themselves. It’s now possible, through interaction design, to help people navigate their way through a data set just as, through good narrative writing, we’ve always been able to guide people through a complex story.

The past decade has seen the most dynamic development in data journalism, driven by rapid technological changes. Ten years ago, “data journalism was mostly seen as doing analyses for

stories,” said Chase Davis, an assistant editor on the Interactive News Desk at the *New York Times*. He explained:

Great stories, for sure, but interactives and data visualizations were more rare. Now, data journalism is much more of a big tent speciality. Data journalists report and write, craft interactives and visualizations, develop storytelling platforms, run predictive models, build open source software, and much, much more. The pace has really picked up, which is why self-teaching is so important.

These are all still relatively new and powerful tools, which both justify excitement about their application and prompt understandable skepticism about what difference they will make if practicing journalists or their editors don’t support developing digital skills. Going digital first brings with it concerns about potential privacy, security, and sustainability relying upon third parties.

III. Why Data Journalism Matters

a. Shifting Context

While it's easy to get excited about gorgeous data visualizations or a national budget that's now more comprehensible to citizens, the use of data journalism in investigations that stretch over months or years is one of the most important trends in media today.

Powerful Web-based tools for scraping, cleaning, analyzing, storing, and visualizing data have transformed what small newsrooms can do with limited resources. The embrace of open source software and agile development practices, coupled with a growing open data movement, have breathed new life into traditional computer-assisted reporting. Collaboration across newsrooms and a focus on publishing data and code that show your work differentiate the best of today's data journalism from the CAR of decades ago.

By automating tasks, one data journalist can increase the capacity of those with whom she works in a newsroom and create databases that may be used for future reporting. That's one reason (among many) that [ProPublica can win Pulitzer prizes](#) without employing hundreds of staff. "We live in an age where information is plentiful," said Derek Willis, a journalist and developer at the *New York Times*. "Tools that can help distill and make sense of it are valuable. They save time and convey important insights. News organizations can't afford to cede that role."

Data journalism can be created quickly or slowly, over weeks, months, or years. Either way, journalists still have to confirm their sources, whether they're people or data sets, and present them in context. Using data as a source won't eliminate the need for fact-checking, adding context, or reporting that confirms the ground truth. Just the opposite, in fact.

Data journalism empowers watchdogs and journalists with new tools. It's integral to a [global strategy to support investigative journalism](#) that holds the most powerful institutions and entities in the world accountable, from the wealthiest people on Earth, to those involved in organized crime, multinational corporations, legislators, and presidents.^{xxxvii}

The explosion in data creation and the need to understand how governments and corporations wield power has put a premium upon the adoption of new digital technologies and development of related skills in the media. Data and journalism have become deeply intertwined, with increased prominence given to presentation, availability, and publishing. Unfortunately, during recent years, [attacks on the press](#) have also grown,^{xxxviii} while global [press freedoms](#)^{xxxix} have diminished to the [lowest levels in a decade](#).^{xl}

Around the world, a growing number of data journalists are doing much more than publishing data visualizations or interactive maps. They're using these tools to find corruption and hold the powerful to account. The most talented members of this journalism tribe are engaged in multi-year investigations that look for evidence that supports or disproves the most fundamental question journalists can ask: Why is something happening? What can data, married to narrative structure and expert human knowledge, tell us about the way the world is changing?

Along with delivering the accountability journalism that democracies need to provide checks and balances—speaking truth to and about the powerful—data journalists are also, in some cases, building the next generation of civic infrastructure out of public domain code and data. Such code might include [open source survey tools](#)^{xli}, an [open election database](#)^{xlii}, a [better interface for U.S. Census data](#)^{xliii}, or ways to [find accessible playgrounds](#)^{xliv}. Such projects are informed by the principles that [built the Internet](#) and World Wide Web^{xlv} and strengthened by [peer networks](#)^{xlii} across newsrooms and between data journalists and civil society. The data and code in these efforts—small pieces, loosely joined by the social Web and application programming interfaces—will extend the plumbing of digital democracy in the 21st century.

“I’m really hopeful that by making data about these facets of our communities more accessible to journalists, we’ll make it easier for them to report stories that help readers unpack the complexity,” said Ryan Pitts, a developer journalist at Census Reporter, in an interview. “Narrative along with this kind of data is a really powerful combination. I think it’s the kind of thing a community needs before it can get at the really important question: So what do we do about this?”^{xlvii}

In the hands of the most advanced practitioners, data journalism is a powerful tool that integrates computer science, statistics, and decades of learning from the social sciences in making sense of huge databases. At that level, data journalists write algorithms to look for trends and map the relationships of influence, power, or sources.

As they find patterns in the data, journalists can compare the signals and trends they discover to the shoe-leather reporting and expert sources that investigative journalists have been using for many decades, adding critical thinking and context as they go. In addition to asking hard questions of people, journalists can now interrogate data as a source.

“What’s different about practicing data journalism today, versus 10 or 20 years ago, was that from the early 1990s to mid 2000s, the tools didn’t really change all that much,” said Matt Waite, a journalism professor at the University of Nebraska who co-created Politifact.com, the Pulitzer Prize-winning website:

The big change was we switched from FoxPro to Access for databases. Around 2000, with the [U.S.] Census, more people got into GIS. But really, the tools and techniques were pretty confined to that tool chain: spreadsheet, database, GIS. Now you can do really, really sophisticated data journalism and never leave Python. There’s so many tools now to do the job that it’s really expanding the universe of ideas and possibilities in ways that just didn’t happen in the early days.

Newsrooms, nonprofits, and developers across the public and private sector are all grappling with managing and getting insight from the vast amounts of data generated daily. Notably, all of those parties are tapping into the same statistical software, Web-based applications, and open source tools and frameworks to tame, manage, and analyze this data.

“Five years ago, this kind of thing was still seen in a lot of places at best as a curiosity, and at worst as something threatening or frivolous,” said Chase Davis. He continued:

Some newsrooms got it, but most data journalists I knew still had to beg, borrow, and steal for

simple things like access to servers. Solid programming practices were unheard of—version control? What’s that? If newsroom developers today saw Matt Waite’s code when he first launched PolitiFact, their faces would melt like *Raiders of the Lost Ark*.

Now, our team at the *Times* runs dozens of servers. Being able to code is table stakes. Reporters are talking about machine-frickin’-learning, and newsroom devs are inventing pieces of software that power huge chunks of the Web. The game done changed.^{xlviii}

In 2014, data journalism is mainstream and the market for data journalists is booming. New media outlets like FiveThirtyEight.com and Vox.com are competing for eyeballs with Appp3d.com from the Mirror, QZ.com from the Atlantic Media Group, *The Economist*’s Data Blog, the *Guardian* Datablog, The Upshot from the *New York Times*, and a forthcoming data-driven site from the *Washington Post*.

A growing number of tools, online platforms, and development practices have transformed the field, from the use of Google and Amazon’s clouds, to the creation and maturation of open source software and the proliferation of open data resources around the globe.

b. The Growth of the News App

Traditionally, computer-assisted reporting focused on gathering and analyzing data as a means to support investigations. Where traditional CAR focused on analysis, the data-driven journalism of today includes data publishing, reuse, and usability.

“Increasingly, I think data journalists also think about how they can provide these data sets in an easy-to-use way for the public,” said Charles Ornstein, senior reporter at ProPublica. “I don’t think we’re in an era anymore in which journalists can say, ‘We’ve analyzed the data, trust us.’ ”

Today, many journalists devote attention not only to finding data for investigations, but to publishing it alongside living stories, or news apps. News applications are one of the most important new storytelling forms of this young millennium, native to digital media and, often, accessible across all browsers, devices, and operating systems on the open Web. [ProPublica’s news app style guide](#) lays out core principles for how they should be built and edited.^{xlix}

News applications and newsroom analytics will be a core element of the way media organizations deliver information to mobile consumers and understand who, where, how, when, and perhaps even why they’ve become readers. Both will be a component of successful digital businesses.

In this context, a news app primarily refers to an online application or interactive feature, as opposed to a mobile software application installed on a smartphone. At their best, [news applications don’t just tell a story, they tell your story](#), personalizing the data to the user.^l News apps can give mobile users better ways to understand the world they’re moving through, from general topics like news, weather, and traffic, down to little league baseball scores. “I think news apps demand that you don’t just build something because you like it,” said Derek Willis. “You build it so that others might find it useful.”

News apps help make sense of vast amounts of data for people who need to understand a

complex subject but lack digital literacy in manipulating the raw data itself. For instance, ProPublica launched [Treatment Tracker](#) in May 2014, a news app based on the Medicare data released by the Centers for Medicare and Medicaid Services earlier in the year.^{li} The investigation published with the news app [examined how doctors bill Medicare for office visits](#).^{liii} ProPublica's data-driven analysis found that while health care professionals classified only 4 percent of the 200 million office visits for established Medicare patients in 2012 as sufficiently complex to earn the most expensive rates, some 1,800 providers billed at the top rate 90 percent of the time.

Charles Ornstein wrote in an email:

This took some time. The data itself is big and complex. We interviewed experts to understand which comparisons would be most meaningful in the data. We looked for top-line numbers that could serve as easy benchmarks people could understand quickly. One was Medicare services per patient, another was payment per patient. We also took a careful look at intensity of established-patient office visits as a benchmark that would be interesting and easily understood by readers. Some specialties, like psychiatry and oncology, have, on average, much more intensive and costly office visits. But in many specialties where the typical such visit is less likely to be so intensive, doctors can vary widely from the mean. If you see that your doctor has a lot more or a lot fewer high-intensity visits than the average doctor like him/her, it doesn't automatically mean there's something wrong, but it's one of the things worth having a conversation about.

What sets our app apart is that it allows you to compare your doctor to others in the same specialty and state. While it may satisfy your curiosity to know how much money a doctor earns from Medicare, it tells you little. We think it's more useful to look at how a doctor practices medicine (the services they perform, the percentage of patients who got them, and how often those patients got them). Our app gives you that information in context.

You can easily spot which doctors appear way different using red notes and orange warning symbols. Again, it's worth asking questions if your doctor (or other health provider) looks different than his/her colleagues.

News apps can enable people to explore a data set in a way that a simple map, static infographic, or chart cannot.

"There are ways to design data so that more important numbers are bigger and more prominent than less important details," said Scott Klein. "People know to scroll down a Web page for more fine-grained details. At ProPublica, we design things to move readers through levels of abstraction from the most general, national case to the most local example."

Increasingly, the creators of news apps are focusing on user-centric design, a principle Brian Boyer, the editor of NPR's Visuals team, explained:

We don't start with the data, or the technology. Everything we make starts with a user-centered design process. We talk about the users we want to speak to and the needs they have. Only then do we talk about what to make, and then we figure out how we're going to do it. It's tempting to start with technical choices or shiny ideas, but we try to stop ourselves and focus on what will work best for a specific group of people, the people who would most benefit from the data.

It may be useful, therefore, to differentiate between the process and the product, as Susan McGregor has:

News apps and data visualization generally describe a class of publishing formats, usually a combination of graphics (interactive or otherwise) and reader-accessible databases. Because these end products are typically driven by relatively substantial data sets, their development often shares processes with CAR, data journalism, and computational journalism. In theory, at least, the latter group is format agnostic, more concerned with the mechanisms of reporting than the form of the output.

News apps “are great to tell stories, and localize your data, but we need more efforts to humanize data and explain data,” said [Momi Peralta](#), of *La Nación*. She noted:

[We should] make data sets famous, put them in the center of a conversation of experts first, and in the general public afterwards. If we report on data, and we open data while reporting, then others can reuse and build another layer of knowledge on top of it. There are risks, if you have the traditional business mindset, but in an open world there is more to win than to lose by opening up.

This is not only a data revolution. It is an open innovation revolution around knowledge. Media must help open data, especially in countries with difficult access to information.

This ethos, where both the data and the code behind a story are open to the public for examination, is one that I heard cited frequently from the foremost practitioners of data journalism around the world. In the same way that open source developers show their work when they push updated software to GitHub, data journalists are publishing updates to data sets that accompany narrative stories or news applications.

This capability to publish data doesn’t change the underlying ethics or responsibility that journalists uphold: Not all data can or should be published in such work, particularly personally identifiable information or details that would expose whistleblowers or put the lives of sources at risk.

Some of the data journalists interviewed expressed a clear preference for creating news apps that are Web-native, as opposed to an app developed for an iOS or Android device. If nonprofit or public media wish to serve all audiences, the thinking goes that means publishing in accessible ways that don’t require expensive, fast data plans or mobile devices. News apps based upon open source and open standards can be designed to work on multiple mobile platforms and are not subject to approval by a technology company to be listed on an app store.

c. On Empiricism, Skepticism, and Public Trust

While the tools and context may have evolved, the basic goals of data-driven journalism have remained the same over the decades, observed Brant Houston, former executive director of [Investigative Reporters and Editors](#). “Sift through data and make sense of it, often with social science methods,” he said.

Today, powerful open source frameworks for the collection, storage, analysis, and publication of immense amounts of data are integrated with rigorous thinking, sound design principles, powerful narratives, and creative storytelling techniques to produce acts of journalism.

Practiced at the highest level, data-driven journalism can be applied to [auditing algorithms](#) or testing whether predictive policing is delivering justice or further institutionalizing inequities in society.^{lvi} When an algorithm may be responsible for mistakenly targeting an innocent citizen or denying a loan to another, the skills required of watchdog journalism move well beyond the rapid production of infographics and maps.

“Data is at the heart of what journalism is,” said *New York Times* developer advocate Chrys Wu, [speaking at the White House](#) in 2012, “and the more substantive it is, the more organized it is, the more easily accessible it is, the better we all can understand the events that affect our world, our nation, our communities, and ourselves.”^{lv}

As with human sources, however, not all data sets are synonymous with facts. They must be treated with skepticism, from origin to quality to hidden biases. “The Latin etymology of ‘data’ means ‘something given,’ and though we’ve largely forgotten that original definition, it’s helpful to think about data not as facts per se, but as ‘givens’ that can be used to construct a variety of different arguments and conclusions; they act as a rhetorical basis, a premise,” [wrote Nick Diakopoulos](#), a Tow Fellow. “Data does not intrinsically imply truth. Yes, we can find truth in data, through a process of honest inference, but we can also find and argue multiple truths or even outright falsehoods from data.”^{lv}

If reporting does become more scientific over time, it could benefit readers and society as a whole. A managing editor might float an assertion or hypothesis about what lies behind news, and then assign an investigative journalist to go find out whether it’s true or not. That reporter (or data editor) then must go collect data, evidence, and knowledge about it. To prove to the managing editor—and skeptical readers—that whatever conclusions presented are sound, the journalist may need to show his or her work, from the sources of the data to the process used to transform and present them. That also means embracing skepticism, avoiding confirmation bias, and not jumping to conclusions about observed correlations.

“In a world awash with opinion there is an emerging premium on evidence-led journalism and the expertise required to properly gather, analyze, and present data that informs rather than simply offers a personal view,” [wrote](#) Cardiff University journalism professor Richard Sambrook. “The empirical approach of science offers a new grounding for journalism at a time when trust is at a premium.”^{vi}

Brian Keegan, a professor in the college of humanities and social sciences at Northeastern University, highlighted many of these issues in a long essay on the [need for openness in data journalism](#).^{vii} The pressures of deadlines and tight budgets are real:

Realistically, practices only change if there are incentives to do so. Academic scientists aren’t awarded tenure on the basis of writing well-trafficked blogs or high-quality Wikipedia articles, they are promoted for publishing rigorous research in competitive, peer-reviewed outlets. Likewise, journalists aren’t promoted for providing meticulously documented supplemental material or replicating other analyses instead of contributing to coverage of a major news event. Amidst contemporary anxieties about information overload as well as the weaponization of [fear, uncertainty, and doubt](#) tactics, data-driven journalism could serve a crucial role in empirically grounding our discussions of policies, economic trends, and social changes. But unless the new leaders set and enforce standards that emulate the scientific community’s norms, this data-driven

journalism risks falling into traps that can undermine the public's and scientific community's trust.

Keegan suggested several sound principles for data journalists to adopt: open data, open deliberation, open collaboration, and data ombudsmen:

Data-driven journalists could share their code and data on open source repositories like GitHub for others to inspect, replicate, and extend. [This is already happening at ProPublica and other outlets.] Journalists could collaborate with scientists and analysts to pose questions that they jointly analyze and then write up as articles or features as well as submitting for academic peer review. But peer review takes time and publishing results in advance of this review, even working with credentialed experts, doesn't imply their reliability.

Organizations that practice data-driven journalism (to the extent this is different from other flavors of journalism) should invite and provide empirical critiques of their analyses and findings. Making well-documented data available or finding the right experts to collaborate with is extremely time-intensive, but if you're going to publish original empirical research, you should accept and respond to legitimate critiques.

Data-driven news organizations might consider appointing independent advocates to represent public interests and promote scientific norms of communalism, skepticism, and empirical rigor. Such a position would serve as a check against authors making sloppy claims, using improper methods, analyzing proprietary data, or acting for their personal benefit.

It now feels clichéd to say it in 2014, but in this context transparency really may be the new objectivity. The latter concept is not one that has much traction in the sciences, where observer effects and experimenter bias are well-known phenomena. Studies and results that can't be reproduced are regarded with skepticism for a reason

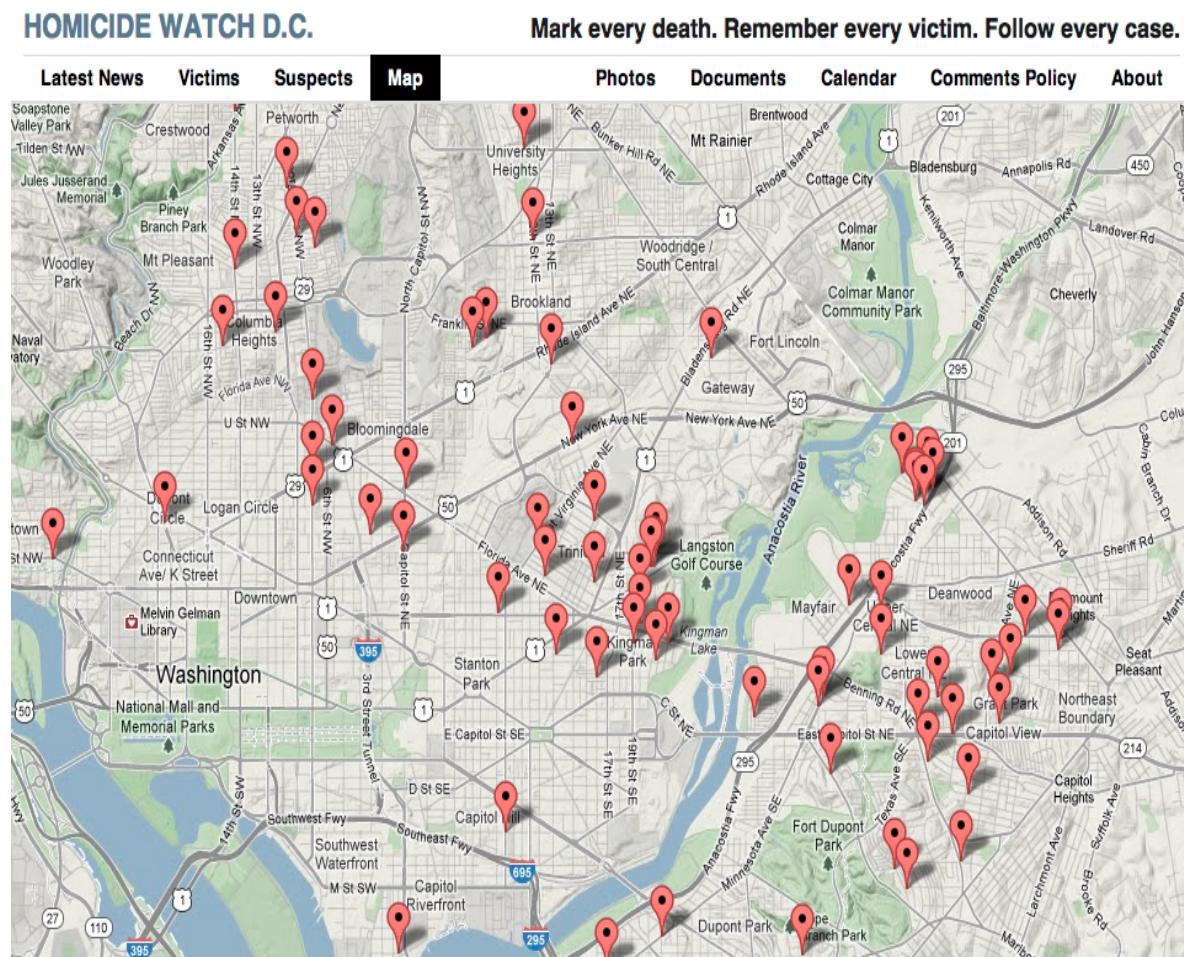
Such thinking about the scientific method and journalism isn't new, nor is its practice by journalists around the country who have pioneered the craft of data journalism with much less fanfare than FiveThirtyEight.com. Making sense of what sources mean, putting their perspective in context, and creating a narrative that enables people to understand a complex topic is what matters. The ultimate accomplishment for journalists may be to integrate data into stories in a way that not only conveys information, but imparts knowledge to the humans reading and sharing it.

To do this kind of work well, journalists need "a firm understanding of public records laws, a grasp of programs such as Excel or Access, contacts with statisticians, and a comfort level in creating data sets where none exist," said Charles Ornstein of ProPublica. "My colleagues and I put together a data set using Access when we were analyzing more than 2,000 disciplinary records from the California Board of Registered Nursing. It was the only way of analyzing real data and not piecing together anecdotes. It was very time consuming but very worthwhile."

Data-driven investigative techniques can substantially augment the ability of technically savvy journalists to master information and hold governments accountable. Applying data journalism enables investigative journalists to find trends, chase hunches, and explore hypotheses. It can enable beat reporters to look beyond anecdotes or a rotating cast of sources to find hidden trends or scoops. A body of empirical evidence, based upon rigorously vetted data, can also give editors and reporters the ability to move away from “he said, she said” journalism that leaves readers wondering where the truth lies.

d. Newsroom Analytics

While traffic data analytics and behavioral advertising aren’t directly involved in gathering data for investigations or publishing visualizations, they are now an integral part of digital journalism. Understanding who is interacting with a story, and how, informs the way future coverage can be extended and delivered.



Washington, D.C. is the epicenter for all kinds of data journalism these days, from politics to policy. Since Homicide Watch launched in 2009, it earned praise and interest from around the digital world, including a profile by the Nieman Lab at Harvard University that asked whether a local blog “could fill the gaps of D.C.’s homicide coverage.”^{lviii} Notably, Homicide Watch has turned up a number of unreported murders.

In the process, the site has also highlighted an important emerging set of data that other digital editors should consider: using inbound search-engine analytics for reporting.^{lxix} As Steve Myers reported for the Poynter Institute, Homicide Watch used clues in site search queries to ID a homicide victim.^{lx} The success of the husband and wife team behind Homicide Watch is an important case study into why organizing beats may well hold similar importance in investigative projects.^{lxii}

The use of data in editorial work at established media institutions like the Financial Times,^{lxii} the Guardian, Washington Post, or the New York Times, however, is still in its relatively early days. In an interview during the spring of 2014, Aron Pilhofer, associate managing editor for digital strategy at the New York Times, told me they had just launched a newsroom analytics team.

The kinds of projects we're doing there are entirely editorial. They are not tied to advertising at all. Right now, many newsrooms are stupid about the way they publish. They're tied to a legacy model, which means that some of the most impactful journalism will be published online on Saturday afternoon, to go into print on Sunday. You could not pick a time when your audience is less engaged. It will sit on the homepage, and then sit overnight, and then on Sunday a homepage editor will decide it's been there too long or decide to freshen the page, and move it lower.

I feel strongly, and now there is a growing consensus, that we should make decisions like that based upon data. Who will the audience be for a particular piece of content? Who are they? What do they read? That will lead to a very different approach to being a publishing enterprise.

Knowing our target audience will dictate an entirely differently rollout strategy. We will go from a “publish” to a “launch.” It will also lead us in a direction that is inevitable, where we decouple the legacy model from the digital. At what point do you decide that your digital audience is as important—or more important—than print?

As Pilhofer allowed, this is a lesson that online publishers started applying a decade ago. It’s time to catch up. “Listening to your readers is as old as publishing letters to the editor,” wrote Owen Thomas, editor-in-chief of ReadWrite. “What’s new is that Web analytics create an implicit conversation that is as interesting as the explicit one we’ve long been able to have.”^{lxiii}

e. Data-driven Business Models

Data journalism and the databases that drive it also offer dramatically improved means to organize and access source material over time. That’s not a minor issue: Newsrooms and media organizations are subject to the same challenges around knowledge management and collaborations that other organizations are in the 21st century. The McKinsey Global Institute estimates that knowledge workers spend 20 percent of their time trying to find information.^{lxiv}

Given that this activity is central to the work journalists do, improving collaboration through social software and digital source material condenses the time it takes to get a story researched, edited, and published. As editors in the business, tech, and finance world know well, that can mean real money—and stabilizing revenues and finding new sources of income is very much on the minds of publishers these days.

The 2013 State of the Media report from Pew Research Center’s Project for Excellence in

Journalism painted a picture of contraction, with newsroom closings and a digital advertising market dominated by technology giants like Facebook and Google.^{lxv} The disruption that the Internet has posed to the traditional business models of newspapers has been well-documented over the past decade. More than 166 U.S. newspapers of an [estimated 1,382](#) in total have stopped putting out a print edition or closed down altogether since 2008,^{lxvi} resulting in [more than 40,000 job losses](#) or buyouts in the newspaper industry since 2007.^{lxvii}

There's no going back to the days when newspapers enjoyed local advertising monopolies and 20 percent profit margins, either. Craigslist, eBay, and Monster.com have each become platforms for the classified revenue that once sustained local newspapers. No single, replicable business model for media in the information age has emerged since, although literally hundreds of panels, conferences, and colloquia have been held to debate the issue.

The economic pain remains most acute at the regional level, where daily newspapers face the difficult challenge of getting consumers to pay for yesterday's news. Just publishing or republishing rows of data alone will not come to the rescue. For instance, one of the canonical examples of data-driven news, [EveryBlock](#), never quite caught on. The site, reasonably described as the "[Xerox PARC of civic data](#),"^{lxviii} was acquired by MSNBC in 2009, [expanded](#) and [refocused on creating community features](#) on top of local data over the years.

In 2013, [NBC News shut down EveryBlock](#),^{lxix} citing issues with its [business model](#). EveryBlock faced other fundamental issues: Despite a [2011 redesign](#) that integrated more social features and topics, the local data that drove the service didn't prove compelling enough to attract consistent daily visitors to sustain the site. Pages of data weren't enough to engage the public on their own. EveryBlock needed more narratives and human interest pieces to keep people coming back for more, engaging them in participation and creating a community.

In 2014, [EveryBlock relaunched](#)^{lxx} in Chicago alone, continuing the experiment, however, hopes that the platform become the civic architecture to stitch together neighborhoods elsewhere are considerably dimmed. Instead, private social networks like Nextdoor or Facebook, bulletin boards like Craigslist, and mobile applications that follow will more likely help neighbors connect to one another or local services.

The [struggles that hyperlocal sites](#)^{lxxi} like [Patch.com](#) and [local news](#)^{lxxii} in general have faced in [searching for a business model](#) have left many observers wondering what will work.^{lxxiii} After a sale and new ownership that cut 85 percent of staff and shifted strategy from local advertising sales to national accounts, [Patch is on path to be profitable in 2014](#) with 17 million unique visitors across 906 sites in April of 2014.^{lxxiv}

In that context, publishers and editors have tough decisions to make about where to cut and where to invest. Despite the promise of data-driven journalism and its importance in the digital news environment, some are still choosing to close divisions dedicated to data.

Digital First Media, for instance, "[shuttered its Project Thunderdome](#)" in April of 2014.^{lxxv} As [Ken Doctor noted in an article for the Nieman Lab](#),^{lxxvi} Thunderdome was a promising digital startup within a much larger media company, producing solid videos and data-driven features like "[Firearms in the Family](#),"^{lxxvii} "[Decoding the Kennedy Assassination](#)"^{lxxviii} and a March

Madness bracket advisor.^{lxxix} Doctor wrote that the decision to shut Thunderdome down, however, was driven more by cost-cutting on the part of Digital First Media's majority owner, Alden Global Capital, than the success or failure of the unit.

Other media companies would be wise not to make the same decision, argued Scott Klein, who said that publishers can afford data journalism if they prioritize it:

News organizations are contracting and budgets are going down. Times are still very tough. That said, I suspect that some newsrooms say they can't afford to hire newsroom developers when they really mean that their budget priorities lie elsewhere—priorities that are set by a senior leadership whose definition of journalism is pretty traditional and often excludes digital-native forms. I also hear a lot from people trying to get data teams started in their own newsrooms that the advice that newsroom leaders get is that newsroom developers are unicorns, whom they can't afford. Big IT departments sometimes play a confounding role here.

I suspect many metro papers can actually afford one or two journalist/developers—and there's a ton of amazing projects a small team can do. For years, the *Los Angeles Times* ran one of the best news application shops in the country with only two dedicated staffers. (They still do great work, of course, and the team has grown.) If doing data journalism well is a priority of the organization, making it happen can fit into your budget.^{lxxx}

The data journalists that escaped alive from the Thunderdome, at least, will have many options in a booming market^{lxxxi} demanding their skills at digital startups^{lxxxii} like Vice, Politico, the Huffington Post, Vox Media, Buzzfeed, Gawker, Business Insider, and Mashable. According to Pew's 2014 State of the News Media, these relatively new entrants have created some 5,000 jobs.^{lxxxiii}

In 2014, data journalism went mainstream when Nate Silver's revamped FiveThirtyEight launched at ESPN and the *New York Times* started The Upshot. Whether some of the new entrants prove commercially successful is still in question, particularly for those pursuing explanatory journalism, seeking to help readers navigate the news.

"These publishers haven't talked much about their revenue strategy, but this is still publishing," noted Lucia Moses in an article on the ad model for explainer journalism in Digiday:

They're in it for the advertising. Online publishers can build an ad-based business one of two ways: Go the scale route by selling price-depressing ads programmatically, or focus on the long tail of lucrative, highly custom advertising (which presumably has a better shot at getting consumers' attention). These publishers are making a bet on the latter, and in the case of the startups, they have the benefit of having established backers—Vox is part of Vox Media; FiveThirtyEight has ESPN—to help with technology and ad sales.^{lxxxiv}

There are, however, more business models for data journalism than advertising, as Mirko Lorenz, a journalist and information architect at Deutsche Welle, highlighted in the *Data Journalism Handbook*:

The big, worldwide market that is currently opening up is all about transformation of publicly available data into something that we can process: making data visible and making it human. We want to be able to relate to the big numbers we hear every day in the news—what the millions and billions mean for each of us.

There are a number of very profitable data-driven media companies that have simply applied this principle earlier than others. They enjoy healthy growth rates and sometimes impressive profits. One example: Bloomberg. The company operates about 300,000 terminals and delivers financial data to its users. If you are in the money business this is a power tool. Each terminal comes with a color-coded keyboard and up to 30,000 options to look up, compare, analyze, and help you to decide what to do next. This core business generates an estimated U.S. \$6.3 billion per year, at least this what a piece by the *New York Times* estimated in 2008. As a result, Bloomberg has been hiring journalists left, right, and center; they bought the venerable, but loss-making *Business Week*, and so on.

Another example is the Canadian media conglomerate today known as Thomson Reuters. They started with one newspaper, bought up a number of well-known titles in the United Kingdom, and then decided two decades ago to leave the newspaper business. Instead, they have grown based on information services, aiming to provide a deeper perspective for clients in a number of industries. If you worry about how to make money with specialized information, the advice would be to just read about the company's history in Wikipedia.

And look at *The Economist*. The magazine has built an excellent, influential brand on its media side. At the same time the "Economist Intelligence Unit" is now more like a consultancy, reporting about relevant trends and forecasts for almost any country in the world. They are employing hundreds of journalists and claim to serve about 1.5 million customers worldwide.^{lxxxv}

Data is a strategic asset, given the insight that it can provide about the world. Proprietary data is a valuable resource that can and does drive the business models of giant companies. There's a reason data scientists are a hot commodity from Silicon Valley to Wall Street to intelligence agencies in Washington, D.C.: They can create valuable knowledge from vast amounts of data, both public and private. Similarly, there's a reason that [hedge funds use the Freedom of Information Act to buy government data](#):^{lxxxvi} It's useful business intelligence for investment management.

Outside of Western democracies with relatively well-established FOIA laws and governments that have been collecting and releasing data for decades, data stewardship may be even more strategic. Justin Arenstein, a Knight International Fellow embedded with the [African Media Initiative](#) (AMI) as a director for digital innovation, said in an [interview](#):

We've embedded open data strategists and evangelists into the newsrooms, backed up by an external development team at a civic tech lab. They're structuring the data that's available, such as turning old microfiche rolls into digital information, cleaning it up, and building a data disk. They're building news APIs and pushing the idea that rather than building websites, design an API specifically for third-party repurposing of your content. We're starting to see the first early successes. Four months in, some of the larger media groups in Kenya are now starting to have third-party entrepreneurs using their content and then doing revenue-share deals.

The only investment from the data holder, which is the media company, is to actually clean up the data and then make it available for development. Now, that's not a new concept. The *Guardian* in the United Kingdom has experimented with it. It's fairly exciting for these African companies because there's potentially—and arguably, larger—appetite for the content because there's not as much content available. Suddenly, the unit cost of value of that data is far higher than it might be in the United Kingdom or in the United States.

Media companies are seriously looking at it as one of many potential future revenue streams. It enables them to repurpose their own data, start producing books, and the rest of it. There isn't much book publishing in Africa, by Africans, for Africans. Suddenly, if the content is available in an accessible format, it gives them an opportunity to mash-up stuff and create new kinds of books.

They'll start seeing that content itself can be a business model. The impact that we're seeking there is to try and show media companies that investing in high-quality unique information actually gives you a long-term commodity that you can continue to reap benefits from over time. Whereas simply pulling stuff off the wire or, as many media do in Africa, simply lifting it off of the Web, from the BBC or elsewhere, and crediting it, is not a good business model.^{lxxxvii}

The [New York Times' syndication of Olympics data in 2012](#) is a useful example of an entrepreneurial business model of this sort,^{lxxxviii} as is [election data](#) from the Associated Press.^{lxxxix}

"Sports in general is big on stats, facts, and figures," wrote Jacqui Maher, assistant editor on the *New York Times* Interactive team, in a post on the "[data Olympics](#)." She said:

Just about any competition that tests the mettle of athletes can be broken down into data points, like personal-best times crossing the finish line of a 5k race, or top career home runs in Major League Baseball. Bringing a sport's national champions together in international competitions—for instance, soccer's World Cup—adds more layers of information. And then there's the Olympics. How much more data is that? Well, in two weeks of the Olympics over 204 gold, silver, and bronze medals were awarded after 7,000 competitions to the best of 32,000 athletes from around the world. It took us about 30,000 code commits to the main git repository to figure out how to show it.^{xc}

f. New Nonprofit Revenues

While revenue models for data-driven hyperlocal news or algorithmic reporting will continue to evolve, flourishing or withering on the vine, nonprofits like ProPublica or the *Texas Tribune* operate under different metrics than profit. The *Tribune*, which has emerged as a bright spot in the firmament of online media for state government, focuses on covering the Texas statehouse. It's now one of the most important examples of data journalism in the United States, given the success of its data visualizations and interactives.

"We turned [three-years-old](#) in November 2012, and we were profitable last year," said Rodney Gibbs, the *Texas Tribune*'s chief innovation officer. "A key to our sustainability is our diverse revenue stream: membership, events, earned income, corporate underwriting, and grants. In other words, we're not dependent on any one source of income. Plus, we've done a good job of keeping our expenses under budget while growing our reach and impact."^{xcii}

The Tribune now has over 200 different data tools and visualizations, including a Public Education Explorer and the [Higher Education Explorer](#), which collect and publish financial, demographic, and performance data for every Texas public school and college.^{xcii}

While the scope and granularity of the data that the *Texas Tribune* has amassed is impressive, it's the online traffic and interest that its work has received that make the case study important to the

future of news. Notably, all of that data has proven to be a hugely popular part of what the media organization publishes: Together, the *Texas Tribune*'s [data library^{xciii}](#) and [directory of public officials^{xciv}](#) account for a majority of its traffic. Such a data library is still a rarity in the media world.

In January of 2013, the *Texas Tribune* [launched^{xcv}](#) the [Lawmaker Explorer](#), the result of nine months of research by 20 different journalists. The news app draws from data on the Texas governor, lieutenant governor, and all members of the Texas House and Senate.

They have resources to apply to growing that success, as the Knight Foundation awarded the *Texas Tribune* a [grant](#) of nearly half a million dollars in 2011. Examples of the *Texas Tribune*'s data journalism include interactives on [Texas prisons^{xcvi}](#), [government employee salaries^{xcvii}](#) and [gubernatorial election results^{xcviii}](#).

"We think of ourselves as a tech startup that works in the news business, rather than a news organization that uses technology," said Gibbs. He elaborated:

I believe that's helped us stay nimble. While our tech group is small—four full-time developers plus one contractor—it's sufficient to not just support our primary site but also the data apps and visualizations we release each month. Moreover, our two data journalists work across the newsroom on a range of beats, so even reporters who aren't data nerds can leverage data and visualizations for their stories. In other words, no one here has to be sold on the value of data—the proof in the traffic and audience feedback has made believers of us all.

[ProPublica launched its own Data Store](#) in February of 2014,^{xcix} publishing raw data for free and [selling premium data](#) to those who would pay for the additional value that it's added.^c Wrote Scott Klein:

In the [Data Store](#) you'll find a growing collection of the data we've used in our reporting. For raw, as-is data sets we receive from government sources, you'll find a free download link that simply requires you agree to a simplified version of our Terms of Use. For data sets that are available as downloads from government websites, we've simply linked to the sites to ensure you can quickly get the most up-to-date data.

"It's a setup similar to [NICAR's Database Library](#), which offers journalists clean and formatted government data on things like plane accidents, federal contracts, and workplace safety records," wrote Justin Ellis for [the Nieman Lab](#):

For users wanting to get their hands on a state's worth of data from ProPublica's "[Dollars for Docs^{ci}](#)" series, for instance, the cost varies: \$200 for journalists, \$2,000 for academics. Companies looking to use the data for commercial purposes have to negotiate a (presumably higher) price with ProPublica. Like any good business, ProPublica offers potential customers free samples of the data before they make a purchase.

ProPublica has always encouraged a level of openness with its work, often making investigations [available to others by Creative Commons^{cii}](#) or building news apps that allow readers to play with data. The data store is an extension of that, but also a potential solution to a question many newsrooms face: how to extract additional value out of an investigation.

But don't expect the store to be a significant source of revenue, at least right away, according to [Richard Tofel](#), ProPublica's president. "It will take a while for us to see if that's a serious revenue source or not," Tofel told me.

In April of 2014, [ProPublica announced plans to grow its Data Store](#) to include almost every data set used in its reporting, citing strong interest. "If you look at newsrooms like the AP, Bloomberg, and Reuters, you'll see that at their core are data products, some of which are very profitable indeed," Scott Klein told the *Columbia Journalism Review*. "There's no question that selling data is a rich opportunity for many newsrooms."^{ciii}

g. Fuel for Robo-journalism

It's certain that data will also play a role in other kinds of ventures, perhaps underpinning "robo-journalism" from services like [Narrative Science](#).^{civ}

That future is now upon us: The [first news report on a March 2014 earthquake in Los Angeles was written by a robot](#)^{cv} created by Ken Schwencke, a journalist and programmer for the *Los Angeles Times*. It's not the first bot "roboporter" on staff; Schwencke and the *Times*' data desk modeled the "Quakebot" on a similar algorithm that creates [automatic reports about homicides](#) in the area.^{cvi}

Automated local news for traffic, weather, high school sports, and police blotters are inevitable, although a human editor may still play a role in publishing bot-reported stories. "Having spent some years as a local news reporter, I can attest that slapping together brief, factual accounts of things like homicides, earthquakes, and fires is essentially a game of Mad Libs that might as well be done by a machine," wrote Will Oremus at *Slate*. "... At the same time, Quakebot neatly illustrates the present limitations of automated journalism. It can't assess the damage on the ground, can't interview experts, and can't discern the relative newsworthiness of various aspects of the story."

In the near term, such newsbots may be most useful as early alert systems for beat reporters and editors, finding signal in the noise that journalists can then use as digital tips to assign, investigate, and confirm. This kind of data journalism—powered by alerts, scrapers, and algorithms—[created scoops](#), which should be catnip to city desk editors. Such automation has widespread applications, from government accountability to financial reporting.

"I would like to get more into monitoring and notification," said Aron Pilhofer. "We ingest millions of records of campaign finance contributions and expenditures every year. If for example, a member of Congress is at risk, you see a spike in 'legal services,' a standard variation above the mean. That should send a notification to congressional reporters. You'd be using tech to improve the reporter's ability to do their job."

What [Google Now](#),^{cvii} Narrative Science, and other algorithmic approaches to local information will all need is good data. Some data will come from municipalities, other data will come from the private sector, nonprofits, and academia, and some will be created by media organizations themselves using sensors and scrapers. (The *Omaha World-Herald*'s [Curbwise](#) is one such project, focused on real estate.)^{cviii}

Although it's not quite what Holovaty intended, it's in this area that EveryBlock may end up making the biggest contribution, in terms of making open government data more useful in an automated fashion. "The thesis of it was not take public records and make them usable," he told me in an interview in 2014. "It was to show what you need to know at the level of a block or neighborhood, but because we were doing it and no one else was, people focused on public records. People focused on things that were unique versus the purpose of what the site was." He added, "That's like focusing on the Beatles because of their use of a sitar—'I love the Beatles because they're such a great sitar band.' It's just something they used to the end of making great music."

The part of his vision for media organizations that is coming to pass may be expressed in news applications and other interactives that are born digital, divorced from the constraints of print and the daily front pages, personalized for individual users, and automatically updated with data as it becomes available. For some time to come, however, there will be a role and need for humans to fact-check the algorithms generating automated news from data, adding context, shaping visually compelling narratives, and conducting investigative journalism that algorithms alone cannot.

Someday, that may change, as Kristian Hammond, the CTO and cofounder of Narrative Science, suggested to Steven Levy:

Hammond believes that as Narrative Science grows, its stories will go higher up the journalism food chain—from commodity news to explanatory journalism and, ultimately, detailed long-form articles. Maybe at some point, humans and algorithms will collaborate, with each partner playing to its strength. Computers, with their flawless memories and ability to access data, might act as legmen to human writers. Or vice versa, human reporters might interview subjects and pick up stray details—and then send them to a computer that writes it all up. As the computers get more accomplished and have access to more and more data, their limitations as storytellers will fall away. It might take a while, but eventually even a story like this one could be produced without, well, me. "Humans are unbelievably rich and complex, but they are machines," Hammond says. "In 20 years, there will be no area in which Narrative Science doesn't write stories."

V. Notable Examples

a. National and International Data Journalism Awards

If you look around at the best data journalism in the world, you'll see a spectrum of achievement and sophistication. On one side, you'll find a lot of maps and data visualizations. These interactives may be the work of a few hours or a day. On the other side of the coin, you'll discover complex, multi-year investigations of education, health care, environment, crime, and government institutions.

The limits of both ends of the spectrum are important, with respect to the resources and time required. These efforts are being furthered by the efforts of many news organizations, including the *Washington Post*, the Center for Public Integrity, the Associated Press, Thomson Reuters, *USA Today*, NPR, the *Guardian*, and the *Chicago Tribune*.

"It's great to see journalists bravely jumping into complicated data sets, like [hospital billing](#) and [Medicare](#),^{cix} to find [investigative stories](#),"^{cxi} said David Herzog, who serves as the academic advisor for the National Institute of Computer-Assisted Reporting.

For example, he pointed to "[Million-dollar Hospital Bills in Northern California](#)" from *The Sacramento Bee*,^{cxi} "[Patient Safety at a Dallas Hospital](#) from *The Dallas Morning News*,^{cxii} "[The Epidemic in Deaths From Prescription Drugs](#)" from the *Los Angeles Times*,^{cxiii} "[Fake Medicare Providers](#)" from the *Atlanta Journal-Constitution*,^{cxiv} and "[Medical Helicopter Flights Mostly for Routine Transport](#)" from the *Argus Leader*^{cxxv} as some of the best data-driven investigative work in recent years.

"I'm really proud of the [elections work at the Times](#),^{cxvi} but can't take credit for how good it looks," said Derek Willis, who works at The Upshot. "A project called '[Toxic Waters](#)'^{cxxvii} also was incredibly challenging and rewarding to work on too. But my favorite might be the first one: the '[Congressional Votes Database](#)' that Adrian Holovaty, Alyson Hurt, and I created at the *Washington Post* in late 2005.^{c xviii} It was a milestone for me and for the *Post*, and helped set the bar for what news organizations could do with data on the Web."

There are an expanding number of notable data-driven journalism projects and sites around the world. The [Philip Meyer Awards](#)^{c xix} are a terrific place to find the best of each year's work, as are the Global Editors Network [Data Journalism Awards](#).^{c xx}

Given these indices, the following examples are chosen because they exemplify notable qualities in the evolving practice of data journalism.

b. Data and Reporting Paired with Narrative

"[The Prescribers](#)," ProPublica's series on fraud and influence in the Medicare drug system, is a masterful use of data in investigative reporting.^{cxxi} The project is an example of the explicit connection between data-driven investigative journalism and government or corporate

accountability. It's far from the first at that outlet.

"The project I'm most proud of is something I did before SOPA Opera, which was our 'Dollars for Docs' project in 2010," said Dan Nguyen, then a developer at ProPublica. "It started off with just a [blog post](#) I wrote to teach other journalists how Web scraping was useful.^{cxxii} In this case, I scraped a website Pfizer used to disclose what it paid doctors to do promotional and consulting work. My colleagues noticed and said that we could do that for every company that had been disclosing payments. Because each company disclosed these payments in a variety of formats, including Flash containers and PDFs, few people had tried to analyze these disclosures in bulk, to see nationwide trends in these financial relationships."

Nguyen explained that the ProPublica team wrote dozens of data scrapers to cross-reference their database of payments with state medical board and medical school listings. "For the initial story, we teamed up with five other newsrooms, including NPR and the *Boston Globe*, which required programmatically creating a system in which we could coordinate data and research," he said. "With all the data we had, and the number of reporters and editors working on this outside of our walls, this wasn't a project that would've succeeded by just sending Excel files back and forth."^{cxxiii}

The "Dollars for Docs" project shows how data-driven journalism can have an impact on patients, providers, private companies and universities; in short, an entire health care industry.

"The website we built from that data is our most visited project yet, as millions of people used it to look up their doctors," said Nguyen. "Afterwards, we shared our data with any news outlet that asked, so hundreds of independently reported stories came from our data. Among the results were that the drug companies and the med schools revisited their screening and conflict of interest policies."

c. Crowdsourcing Data Creation and Analysis

Today, journalists have many more options for sourcing. Newsroom reporters and developers can download, scrape, and digitize data from a wealth of sources, from websites to document dumps. In the future, more journalists will create data themselves using sensors, and engage their distributed audiences of readers, listeners, and watchers to help gather data with them.

In some forward-thinking media organizations, this is already happening. In 2011, ProPublica began an effort to "[Free the Files](#)," making the physical "public inspection documents" detailing political advertising spending at local stations open to the public.^{cxxiv}

In August of 2012, the Federal Communications Commission ordered TV stations in the top 50 markets in the United States to begin posting these documents online. The trouble is that the FCC didn't require that publishing be done in an open, standardized format. As a result, the stations submitted a [mass of unsearchable PDFs](#).^{cxxv} ProPublica created a software application to enable volunteers to translate the files into structured data and sort the files by market, amount, candidate, and political group. ProPublica later [open sourced the app as Transcribable](#).^{cxxvi}

"Our goal was to take thousands of hard-to-parse documents and make them useful, helping to reveal hidden spending in the election," senior engagement editor Amanda Zamora

explained^{cxxvii} “Nearly 1,000 people pored over the files, logging detailed ad spending data to create a [public database](#) that otherwise wouldn’t exist. We logged as much as \$1 billion in political ad buys, and a month after the election, people are still reviewing documents.”

In 2013, New York City’s public radio station asked its listeners to help [track the emergence of cicadas](#) with inexpensive sensors. WNYC’s “[Cicada Tracker](#)” project turned up some 8,000 cicada sightings, with 800 people making trackers.^{cxxviii} The project made [crowdsourcing data collection](#)^{cxxix} through [sensor journalism](#)^{cxxx} and a distributed listening audience a reality—not just a theoretical project—that resulted in 1,5000 collected temperature readings. The [lessons from the cicada tracker project](#) should inform future efforts at public media to engage in public engagement, citizen science, and data collection. For more of a deep dive into the topic, check out the proceedings of the [sensor journalism workshop](#) at the Tow Center last year.^{cxxxi}

In late 2013, the NPR Visuals team released a project around accessible playgrounds. NPR made a request of its community of listeners and readers: Help public media collect the data that drives it and make the resource better for everyone. The [NPR playgrounds app](#) enables parents and children to search for accessible playgrounds, taking commonly used consumer-recommendation engines and combining them with a strong public service element.^{cxxxii}

“This is sort of like Yelp, except for playgrounds for kids with special needs,” said Brian Boyer, the head of NPR’s Visuals team, in an interview. “It is the first of its kind—a nationwide database of playgrounds that are well suited to kids in wheelchairs, kids with autism, or kids with other special needs.”

NPR activated its audience to become participants in data collection, much in the same way that Audubon’s “Christmas Bird Count” and [eBird](#) are [crowdsourcing data collection](#) about bird species.^{cxxxiii} In the first 48 hours after the app was launched, data for 336 more playgrounds was added to the database, for a total of 1,293. In May of 2014, the playgrounds app had 1,907 entries, and counting.

The app is a notable case study for the power of public engagement and crowdsourcing-data creation. There are decades of precedent where a listening or viewing audience collaborate with a media organization in collecting images, videos, or stories.

What remains relatively new is the capacity for a networked populace to contribute data, whether it comes from [sensors](#) in [droughts](#)^{cxxxiv} or [geiger counters](#)^{cxxxv} near potential sources of radiation. If [turning data into stories](#) is now a core element of investigative journalism, WNYC and NPR’s Visuals team have showed how to do it best and serve the public in the process.^{cxxxvi}

d. Public Service

Internationally, the [Guardian Datablog](#) established itself early as one of the best sources of interesting, relevant data journalism, covering topics from sports, to popular culture, to government accountability.^{cxxxvii} The Datablog offers new posts daily, showing how free, open source tools, narrative skills, and online publishing can be used by a lean team to produce excellent journalism. Notably, its editors and contributors are not programmers: They leverage free online tools and open source software in their work.

Every Datablog post demonstrates an emerging practice wherein Datablog editors make it

possible for readers to download the data themselves. The Datablog is full of great examples, like mapping reactions to current events,^{cxxxviii} tracking government spending,^{cxxxix} evaluating government data quality,^{cxl} mapping gun crimes in the United States,^{cxli} and examining gun ownership and homicide rates.^{cxlii}

The *Guardian*'s data journalism is particularly important as the British government continues to invest in open data. In June of 2012, the United Kingdom's Cabinet Office relaunched Data.gov.uk and released a new open data white paper.^{cxlvi} The British government has doubled down again and again on the notion that open data can be a catalyst for increased government transparency, civic utility, and economic prosperity. While the evidence that has emerged in recent years strongly supports the connection of open data to economic activity, the role of data journalism in delivering accountability using the data released from these platforms and acquired by Freedom of Information Act requests is central.^{cxlvi}

e. All Organizations—Great and Small

When I interviewed the founding editor of the *Guardian* Datablog, Simon Rogers, he praised the work of institutions and the work of several practitioners. (While Rogers is now Twitter's first data editor,^{cxlv} he remains passionate about data journalism.^{cxlvi})

La Nación, in Argentina, “is doing an amazing thing, scraping all public data at the moment,” he said. “A lot of data journalism has to be about giving data to people and making it accessible.”

Rogers also pointed to the work of James Cheshire,^{cxlvii} the collaboration of government and academia on the Bombsight project, and the use of geodata by the Oxford Internet Institute.^{cxlviii} Those projects all show that acts of data journalism can be committed by much more than traditional media organizations.

It was particularly instructive to learn more about the work of large media organizations, like the *Los Angeles Times* and Canada's *Global News*, which have been building their capacity to practice data journalism.

“As part of large broadcast organizations, one thing that is very satisfying about data journalism is that it often puts our digital staff in the driver’s seat—what starts as an online investigation often becomes the basis for original and exclusive broadcast content,” wrote Keith Robinson, the senior producer for specials and interactive at *Global News* in Canada, in an email.

Robinson highlighted several examples of its Data Desk's work,^{cxlxi} from mapping and visualizing Canada's census data to investigating water main breaks in Toronto and the ways they're being addressed.

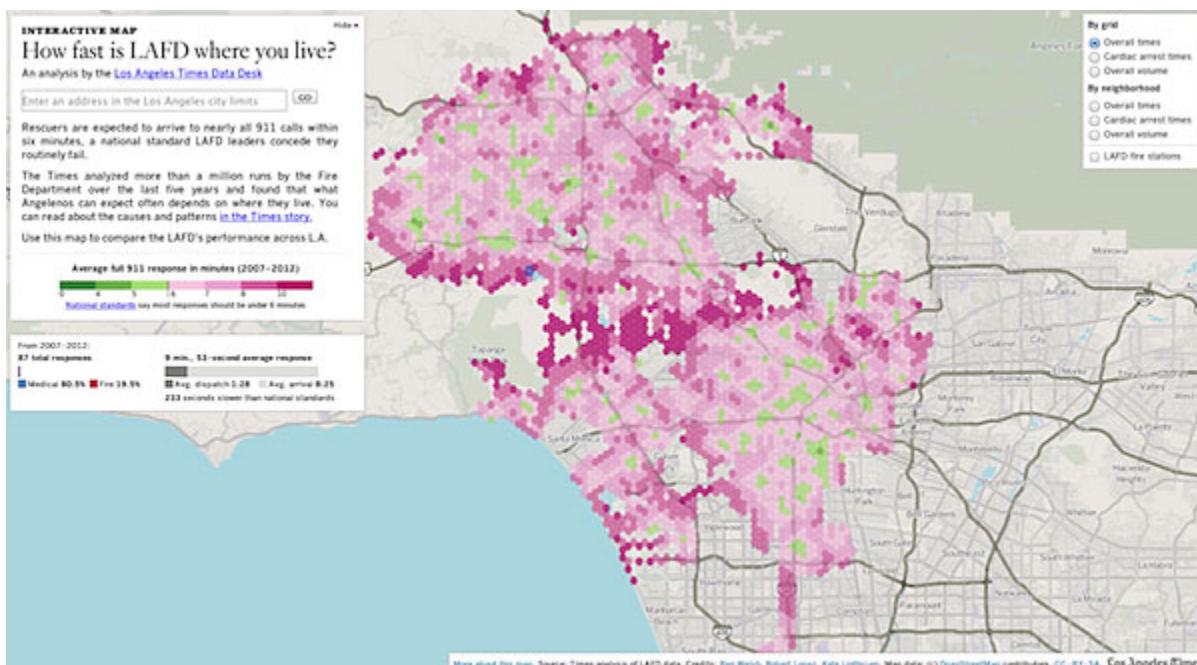
It's not just big-city newsrooms or stations that can afford teams of programmers and designers; these aren't the only players. Important, sophisticated data-driven journalism is also possible with smaller teams on tight deadlines. To put it another way, acts of data journalism by small teams or individuals aren't just plausible or possible, they're happening—from Italy to Chile to Brazil to Africa.

That doesn't mean that the news application teams at NPR or newspaper companies aren't setting the pace for data journalism when it comes to cutting-edge work—far from it, as this [news app of tornado damage in Moore, Oklahoma^{cl}](#) demonstrates—but the tools and techniques to make something worthwhile are available to smaller organizations, even with tightened spending.

f. Embracing Data Transparency

The Datablog and its editor set an important standard that many other data journalists continue to embrace: Show your work and share your data.

I profiled the [data journalism](#) work of the *Los Angeles Times* in early 2013, when I interviewed news developer Ben Welsh about the newspaper's Data Desk.^{cli} The relatively small team of reporters and Web developers specializes in maps, databases, analysis, and visualization. For instance, its interactive visualization mapped [how fast the Los Angeles Fire Department responds to calls](#).^{clii} The visualization was part of a longer [series on 911 breakdowns in the LAFD^{cliii}](#) that combined investigative journalism with data analysis to create important, compelling narratives that held the government accountable and demonstrated significant issues existed in the city's data-collection practices.



The investigation offered an ageless insight that will endure well beyond the “[era of big data](#)”: Poor collection practices and aging IT will derail any institutional efforts to use data analysis to improve performance.

The *Los Angeles Times* found that poor recordkeeping is holding back state government efforts to [upgrade California's 911 system](#). As with any database project, beware of “[garbage in, garbage out](#),” or “GIGO.”

As Ben Welsh and Robert J. Lopez reported for the *L.A. Times* in December of 2012, California's Emergency Medical Services Authority has been working to centralize performance data since 2009. Unfortunately, it's difficult to achieve data-driven improvements or manage against perceived issues by applying big data to the public sector if the data collection itself is flawed.^{cliv} The *L.A. Times* reported quality issues stemming from how response times were measured to record keeping on paper to a failure to keep records at all.

When I profiled [Ben Welsh's work](#) in 2012, he told me this kind of project was exactly the sort of work he's most proud of doing. "As we all know, there's a lot of data out there," said Welsh, "and, as anyone who works with it knows, most of it is crap. The projects I'm most proud of have taken large, ugly data sets and refined them into something worth knowing: a nut graf in an [investigative story](#) or a [data-driven app](#) that gives the reader some new insight into the world around them."^{clv}

Farther north in California, a new experiment is applying data journalism to local government accountability in Oakland, at a website called Oakland Police Beat that went live in the spring of 2014.^{clvi} The nonprofit site, which is part of Oakland Local and the Center for Media Change—and funded by the Ethics and Excellence in Journalism Foundation and the Fund for Investigative Journalism—was co-founded by Susan Mernit and Abraham Hyatt, the former managing editor of ReadWrite. (Disclosure: Hyatt edited my posts there.)

Oakland Police Beat is squarely aimed at shining sunlight on the practices of Oakland's law enforcement officers. Its first story out of the gate pulled no punches, finding that Oakland's most decorated officers were responsible for a high number of brutality lawsuits and shootings.

The site also demonstrated two important practices that deserve to become standard in data journalism: explaining the methodology behind its analysis, including source notes, and (eventually) publishing the data behind the investigation.

ProPublica does it, the Datablog does it, and so does the *Los Angeles Times*. The *Times* Data Desk set a high bar in its investigation of ambulance response times by not only making sense of the data, but also publishing the data behind the [open source maps of California's emergency medical agencies](#) as part of the series into the public domain.^{clvii}

This wasn't the first time the team made code available, nor the last. (Just visit the [Data Desk's Github account](#) for proof.)^{clviii} As Welsh noted in a [post](#) about the series,^{clix} the Data Desk has "previously written about the [technical methods](#)^{clx} used to conduct [the] investigation, [released the base layer](#) created for an [interactive map](#) of response times,^{clxi} and [contributed](#) the location of LAFD's 106 fire station to the Open Street Map.^{clxii}

Offered Scott Klein:

If it's done well, people have a really big appetite to see the data for themselves. Look how many people understand—and love—incredibly sophisticated and arcane sports statistics. We ought to be able to trust our readers to understand data in other contexts too. If we've done our jobs right, most people should be able to go to our [Prescriber Checkup](#) news application,^{clxiii} search for their doctors, and see how their prescribing patterns compare to their peers—and understand what's at play and what to do with the information they find.

Follow-through on this kind of thinking is what really made me sit up and take notice of [The Upshot](#)^{clxiv} in April of 2014, the *New York Times* new data-driven website. It made editorial decisions to share [how reporters found the income data](#)^{clxv} at [LIS](#), link to the [data set](#), and share both the [methodology](#)^{clxvi} behind the forecasting model and the [code](#) for it on Github.^{clxvii} That is precisely the model for [open data journalism](#)^{clxviii} that embodies the best of the craft as it is practiced in 2014, and sets a high standard right out of the gate for future interactives at The Upshot and for other sites that might seek to compete with its predictions.

I was not alone in my positive assessment of the content, presentation, and strategy of the *Times'* new site: Over at the *Guardian* Datablog, James Ball published an [interesting analysis of data journalism](#), as seen through the initial foray of The Upshot; FiveThirtyEight; and Vox, the “explanatory journalism” site Ezra Klein, Melissa Bell and Matt Yglesias, among others, launched in the spring of 2014.^{clxix}

Ball’s whole post is worth reading, particularly with respect to his points about audience, diversity, and personalization. The point that is particularly important is the one I’ve made repeatedly above, that data journalists should try to be open about the difficult, complicated process of reporting on data as a source:

Doing original research on data is hard: It’s the core of scientific analysis, and that’s why academics have to go through peer-review to get their figures, methods, and approaches double-checked. Journalism is meant to be about transparency, and so should hold itself to this standard—at the very least.

This standard is especially true for data-driven journalism, but, sadly, it’s not always lived up to: Nate Silver (for understandable reasons) won’t release how his model works, while FiveThirtyEight [hasn’t released the figures or work behind some of their most high-profile articles](#).

That’s a shame, and a missed opportunity: Sharing this stuff is good, accountable journalism, and gives the world a chance to find more stories or angles that a writer might have missed.

Counterintuitively, old media is doing better at this than the startups: The Upshot [has released the code driving its forecasting model](#), as well as the data on its launch inequality article. And the *Guardian* has at least tried to release the raw data behind its data-driven journalism since [our Datablog](#) launched five years ago.

In May of 2014, the backlash to data journalism is still growing, as more academics, economists, and statisticians read and react to the style and format of the pieces published at Vox, FiveThirtyEight, and The Upshot. The reaction, however, is to the brand and form of data journalism practiced there, in which data, available research, and charts are consulted by an author to examine a question or story, combined relatively rapidly, and presented in a series of charts or maps wrapped in narrative text. This form departs from the slower moving investigative features and news applications produced in proceeding years.

In a survey comparing the data publishing habits of these three sites, none is meeting the standard set by the *Guardian* Datablog or ProPublica.

Of the 290 items published in the catch-all [FiveThirtyEight RSS feed](#), available since the site launched in March of 2013, 114 are features.^{clxx} Only interactive .CSV files for data for 10 of these stories has been uploaded to its [data directory on Github](#); that's a transparency of only 3.4 percent.^{clxxi} The Upshot has demonstrated sound reporting and transparency regarding a story on inequality, publishing the data and the model used to analyze it to its [Github account](#). The *Times* has since published data for another story and [open sourced code for a Ruby gem](#) that extracts press releases and statements by members of Congress.^{clxxii} Vox.com does not publish or host any of the data used in its stories, although [Vox Media](#) has updated the code for Chorus, its content management system, over 62,000 times.^{clxxiii}

Not all data is equal, and not all data can be released in raw form, particularly if it contains personal or private details. Resource constraints may mean that scrubbing data properly isn't possible, which would argue against release. Practices can change too: The *Guardian* Datablog stopped publishing open data into its [data store](#) in 2014.^{clxxiv}

g. Following the Money

As David Kaplan, director of the Global Investigative Journalism Network, emphasized, huge databases, network analyses, and code [aren't a replacement](#) for investigative journalism.^{clxxv} These technologies augment and extend what's possible for media organizations, lone muckrakers, or even teams of journalists working collaboratively across borders and time zones. This kind of collaboration has moved from a potential to reality over the past three years, when a team of more than 80 journalists from 40 different countries worked together to map the world of secret trusts and offshore companies.

The [International Consortium of Investigative Journalists analyzed 260 gigabytes](#) of leaked corporate data, constituting text, PDF, images, spreadsheets, and images to reveal how government officials were offshoring money, how banks were involved in the practice, and how organized crime is using these same structures.^{clxxvi} “Offshore Leaks” is a [triumph for data journalism](#).^{clxxvii} Expect this model to be refined and applied again in the future.

h. Mapping Power and Influence

Mapping the hidden or tacit connections between powerful figures in business and government has long been a focus of investigative journalism. Today, data, software, and interactive visualizations can enable people to understand those networks of power and influence in unprecedented ways.

Reuters' [Connected China](#) is a brilliant example of how technology can give life and meaning to data, giving visitors the ability to explore relationships in a way that simply isn't possible on the printed page.^{clxxviii} According to Reuters investigative reporter Irene Jay Liu, Connected China was the outcome of 18 months of reporting, design, development, and research.^{clxxix} The database that resulted includes tens of thousands of people, organizations, and events, more than 30,000 relationships between them, and some 1.5 million words. The project is built on open standards, including HTML5, enabling people to use it on tablets, smartphones, laptops, or desktop computers alike.

The project “represents a new approach for Reuters News,” wrote Liu, “a model to take the reporting we do every day about people, institutions, power, and relationships and put it in a format that gives it sustained significance over time.” She added:

Adhering to Reuters’ high journalistic standards, we have structured inherently qualitative relationships—the connections between people (family, mentorship, rivalry, alliances), the importance of particular job roles, the power dynamic between the various institutions that govern China. By quantifying and categorizing these complex relationships, we break from the constraints of long-form text and allow new ways of communicating and interpreting this acquired knowledge.

By harnessing the collective intelligence gathered by a global team of reporters and editors, we can derive deeper insight to the political, societal, and economic implications of these connections.

Baseball has [sabermetrics](#); we’re trying to develop the field of sinometrics.

A few months before Connected China went live, Miguel Paz and his colleagues launched a similar data-driven approach to [mapping Chile’s elite](#).^{clxxx} In 2011, when Paz was still the managing editor of *El Mostrador*, he won a [Knight News Challenge](#) to create an interactive platform that would map the relationships between a database of entities from investigations and crowdsourcing.^{clxxxii}

In December of 2012, the beta version of the site went live and then grew, with support from [Startup Chile](#) and the [International Center for Journalists](#) (ICFJ). In the months since, Poderopedia (the platform’s name) has matured and grown beyond Chile, [powering a voting platform in Panama](#).^{clxxxii}

“Gabriel García Márquez said that to talk about ‘investigative journalism’ is redundant, because he assumes that any form of journalism should be an investigative one,” said Paz in an interview. “The purpose of journalism is to show you what the powerful want to hide. It’s the same with any form of journalism.”

Now the Poderopedia Foundation, in 2014 it announced [plans to expand to Venezuela and Colombia](#) in 2014.^{clxxxiii}

i. Geojournalism, Satellites, and the Ground Truth

[Open data from satellites can revolutionize environmental reporting](#), as InfoAmazonia has demonstrated over the past two years.^{clxxxiv} I first met [Gustavo Faleiros](#) in Brazil, where the Knight International Journalism Fellow was reporting on what was happening to the Amazon rainforest, in partnership with Washington-based organizations the [International Center for Journalists](#) and Internews. Faleiros is the project coordinator for [InfoAmazonia.org](#), a beautiful mash-up of open data, maps, and storytelling that enables people to explore how the rainforest of Brazil is changing.

Since its launch in 2012, InfoAmazonia has been training Brazilian journalists to use [satellite imagery](#) and collect data related to forest fires and carbon monoxide. It has now published 18 interactive maps online based upon gigabytes of geographical data that show [deforestation](#) over

time, among other subjects.^{clxxxv}

This approach to storytelling using maps has been dubbed “[geojournalism](#)” by its practitioners, or the practice of telling stories with geographic information systems’ (GIS) data generated by the earth sciences.

The [Environmental News Lab](#), a multidisciplinary team at Brazilian nonprofit media company [O ECO](#), has published a [Geojournalism Handbook](#)^{clxxxvi} in partnership with the ICFJ, Internews’ Earth Journalism Network, and the Flag It! Project. The online handbook explains how to use a series of open source and/or Web-based tools to collect, organize, visualize, and publish data, with a specific focus on contributing to and using the growing geocommons of Open Street Maps—the Wikipedia for maps.

Other [examples of geojournalism](#)^{clxxxvii} include the [Oxpeckers Center for Investigative Environmental Journalism](#) in South Africa, where journalists are tracking poaching of rhinoceroses in the country’s national parks.^{clxxxviii} [Internews Kenya](#) launched [Land Quest](#) in the country to increase the capacity of Kenyan journalists to report on international development and private financing.^{clxxxix}

In 2014, [InfoAmazonia plans to add ground reporting](#) in Brazil, creating applications that would enable nonprofits and residents to share data with O ECO.^{cxc}

j. Working Without Freedom of Information Laws

Four different perspectives that I heard from journalists in Spain, Italy, Argentina, and South Africa highlighted some of the challenges of practicing data-driven journalism in countries without strong right to information laws, noting it’s difficult but not impossible.

“Spain is a country lacking a Freedom of Information Act and an accountability culture,” wrote Javier de Vega, communications director for [Fundación Ciudadana Civio](#), a Spanish foundation that supports open data and data journalism in Spain, in an email. “We are the last big country in Europe to pass a freedom of information law, though a very unambitious text is being studied by the Congress.”

Long before data journalism entered the mainstream discourse, *La Nación* was pushing the boundaries of what was possible in Argentina, reporting on a country without a Freedom of Information Act Law. If you start exploring *La Nación*’s efforts to go online and treat data as a source, you’ll find [Angélica “Momi” Peralta Ramos](#), the multimedia development manager who originally launched LaNacion.com in the 1990s and now manages its [data journalism efforts](#).^{cxi}

Ramos contends that data-driven [innovation is an antidote](#) to budget crises in newsrooms.^{cxcii} Her perspective is grounded in experience: Peralta’s team at [La Nación is using data journalism to challenge a FOIA-free culture in Argentina, opening up data](#) for reporting and reuse to holding government accountable.^{cxciii}

Her team has published a notable array of data-driven stories to date, including:

- Argentina's Official Advertising Funds Distribution 2009–2013: Friends, Politicians, and a Stylist.^{cxciv}
- Public Officials' Salaries and Assets for Reporting and Accountability.^{cxcv}
- Monitoring the New Media Law in Argentina 2009–2013.^{cxcvi}
- VozData: The Senate Expenses (II).^{cxcvii}
- 2013: Legislative Elections in Argentina.^{cxcviii}
- Argentina's Senate Expenses 2004–2013.^{cxcix}

Peralta has seen the context for *La Nación*'s work change in recent years:

To take just one example, consider the inflation scandal in Argentina. Even *The Economist* removed our [national] figures from their indicators page. Media that reported private indicators were considered as opposition by the government, which took away most official advertising from these media, fined private consultants who calculate consumer price indices different than the official, pressed private associations of consumers to stop measuring price, and releasing price indexes, and so on.

Regarding official advertising, between 2009 and 2013, we managed to build a data set. We found out that 50 percent went to 10 media groups, the ones closer to the government. In the last period, a hairdresser (stylist) received more advertising money than the largest newspapers in Argentina. Last year, independent media suffered an ad ban, as reported in the *Wall Street Journal*: “Argentina imposes ad ban, businesses said.”^{cc}

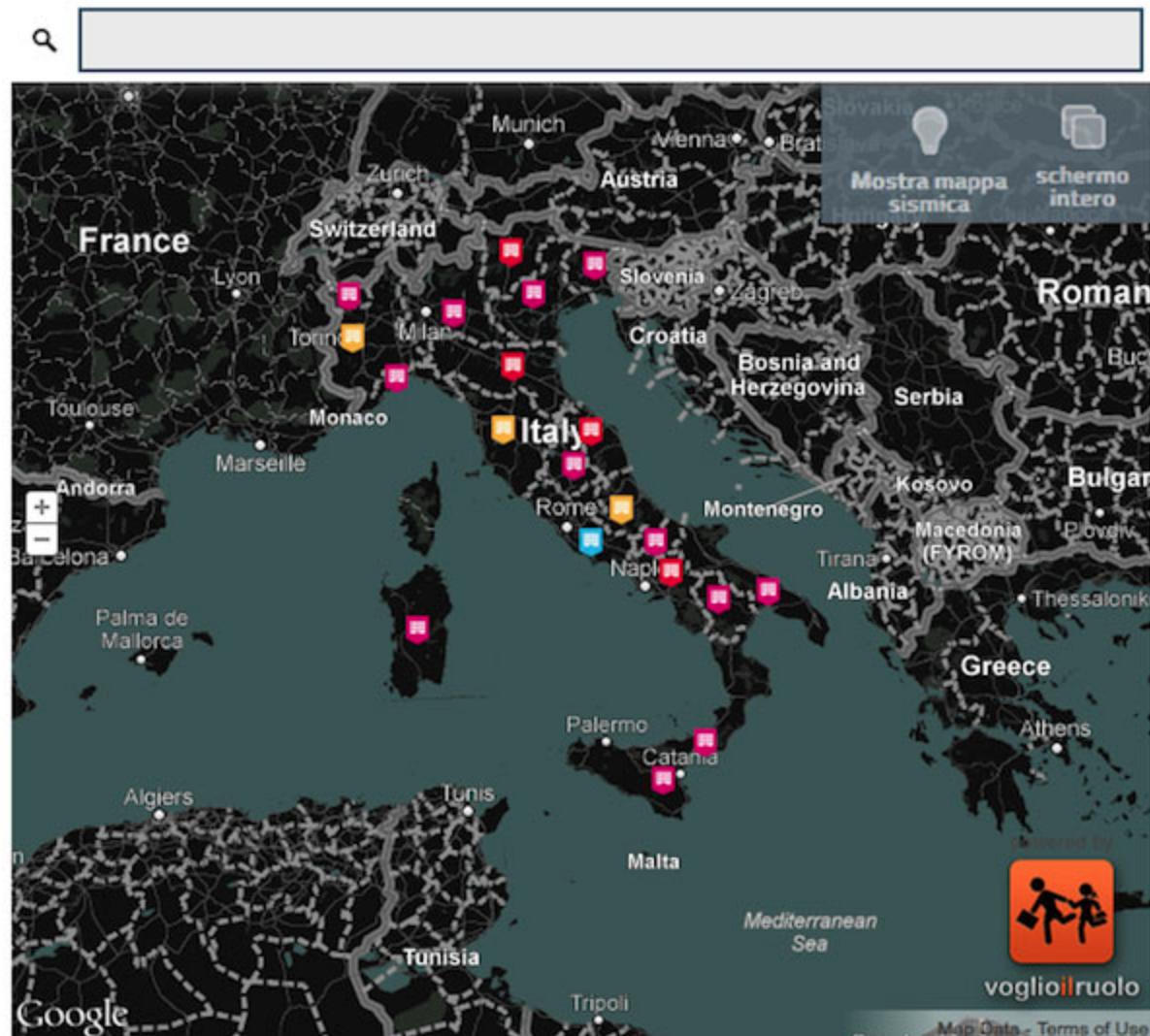
Argentina is ranked 106 / 177 in Transparency International Corruption Perceptions Index. We still are without a freedom of information law.

Journalists in Italy face a similar information landscape. Elisabetta Tola, an Italian data journalist, wrote in to share her work on a series of *Wired Italy* articles that featured data on seismic risk assessment^{cci} in Italian schools.^{ccii} The online interactive lets parents search for schools, a feature that embodies service journalism and offers more value than a static map.^{cciii}

La tua scuola è sicura? Cercala sulla mappa

L'inchiesta [#scuolesicure](#) di Wired arriva a conclusione. Ecco le informazioni su rischio sismico, verifiche e relativi costi di tutti gli edifici scolastici italiani

09 novembre 2012 di [Wired.it Staff](#)



La mappa aggiornata al 9 novembre 2012 comprende gli edifici scolastici di tutte le regioni italiane

[*Wired Italy Map of Earthquake Risk for Schools*]

Guido Romero, the science editor at *Wired Italy* who published the work, shared more of the backstory behind the project via email.

“In Italy there are some 50,000 school buildings,” said Romero. “Protezione Civile, the Italian FEMA, estimates about 22,500 need to be checked for earthquake security. The overall Italian school population is about eight million (students + teachers + personnel) so you can do the math of how relevant this problem is.”

The backstory behind the *Wired Italy* project highlighted a key challenge in Italy that exists in many other places around the world: How can data journalism be practiced in countries that do not have a Freedom of Information Act or a tradition of transparency on government actions and spending?

The Italian government, while well behind the pace set by the United Kingdom, has made [more open data available^{civ}](#) since it [launched](#) a national platform in 2011.^{ccv} In this case, however, the only data that the Italian Ministry of Education released was a list of school buildings published online.

As I recounted earlier, Tola and her team aggregated or created the rest of the data used in the project, from scraping and processing PDFs of spending data from regional government websites, then adding geolocation in cooperation with a local developer.

Romero said in our interview:

When we started looking into this last June [2012], the first door we knocked on was the Ministry of Education, notably their Office for School Buildings and Safety, as our sources inside the Ministry had told us they did have the data. Their non-response turned into a bitter attack to the magazine when we wrote that the very same ministry advertising itself as a groundbreaking pioneer of open data did not release information relevant for millions of families. Mario Di Costanzo, the Director of the Office for School Safety, did give us an [interview](#).^{ccvi} He said on the record they do have the data but would personally oppose any release of parts or all of them as “revealing which schools are at risk would be dangerous.”

As is the case around the world, culture and freedom of information laws matter, particularly with respect to access to data needed to hold governments accountable and audit their programs. Proactive, selective open data initiatives by government focused on services that are not balanced by support for press freedoms and improved access can fairly be criticized as “openwashing” or “fauxpen government.” Data journalists who are frequently faced with heavily redacted document releases or reams of blurry PDFs are particularly well placed to make those critiques. That currently appears to be the case in Italy.

Romero said, “Data journalism is not impossible over here—in fact, Elisabetta and myself believe there are great opportunities, but having a very poor access law and, even worse, a deep rooted culture of non-disclosure in the public administration makes data journalists’ work pretty hard.” He continued, “That said, there is a growing movement for reforming our access law (I’m personally engaged in that with www.dirittodisapere.it) but ‘open data’ is a word very much frowned upon by reporters, as it’s led to little relevant work.”

k. Data Journalism and Activism

Media throughout Africa face all of these challenges and more, fighting obstinate public officials, paper records, no access to information laws, and outright threats and physical violence directed at journalists. Building the capacity of African media to practice data-driven journalism has now taken on new prominence, as the digital disruption that has permanently altered the models of more developed countries bears down on countries in the continent.^{ccvii} The challenges that data journalism in West Africa faces are significant, though these are not unique from elsewhere on the continent.^{ccviii}

Justin Arenstein, a South African investigative journalist, is a fierce advocate for data-driven journalism that not only makes sense of the world for readers and viewers, but also provides them with tools to become more engaged in changing the conditions they learn about in the work. For instance, data journalism boosted voter registration in Kenya,^{ccix} creating a simple website using modern Web-based tools and technologies.

A “data boot camp” in Kenya in 2012 led to another excellent example of this dynamic. Arenstein explained:

NTV, the national free-to-air station, had been looking into why young girls in a rural area of Kenya did very well academically until the ages of 11 or 12—and then either dropped off the academic record completely or their academic performance plummeted. The explanation by the authorities and everyone else was that this was simply traditional; it’s tribal. Families are pulling them out of school to do chores and housework, and as a result, they can’t perform.

As it turned out, that was an incorrect conclusion. Irene Choge, a Kenyan journalist who attended the data journalism training, started mining the available data and public records. Choge first looked at medical records to see if cholera was involved. Then, she examined water records and physical infrastructure. It was there that she found a key correlation: The schools that saw the worst drop-offs in academic performance by teenage girls were the ones that didn’t have sanitation facilities.

Choge subsequently worked with developers to create a simple SMS-based phone application that enabled parents to determine how schools compared and, notably, to advocate for change. Her work in reporting school sanitation woes has led officials to shift resources to building sanitation facilities.^{ccx}

While such applications move further into the realms of political advocacy and citizen engagement than many journalists may find comfortable, the growth of services that span the intersection of open government and data journalism will continue to be an important, fertile ground in the years ahead.

V. Pathways to the Profession

a. Mentorship, Numeracy, Competition, Recruiting

While data journalism has gone mainstream in recent years, significant challenges lie ahead for traditional media and news organization to take full advantage of advances in technology.^{cxxi} Just as McKinsey identified a gap between available analytic talent and the demand created by big data, there is a data science skills gap in journalism.^{cxxii}

Rapidly expanding troves of data are useless without the skills to analyze them, whatever the context.^{cxxiii} Focusing too much on tech skills could exclude some of the best candidates for these jobs—but there will need to be training to bring them onboard.^{cxxiv} Foundations and universities have noticed the need to build capacity in these areas. In May of 2012, the Knight Foundation gave Columbia University \$2 million for research^{cxxv} to help close the data science skills gap.^{cxxvi}

“There’s such a demand and so little capacity,” said Emily Bell, director of the Tow Center for Digital Journalism at Columbia University, in 2012. “Lots of people teach at the very low level, very few at the elevated level. Nobody teaches the algorithmic, advanced courses that you’d see in computational journalism. There aren’t many people who can do the latter, either professionally or [on the] teaching side.”

In the United States, I’d estimate that the headcount of working data journalists numbers well under a thousand across all newsrooms and media organizations. Their ranks are growing, especially given clear demand in both traditional and startup media companies. Globally, there may be thousands of data specialists in the media, but not many more, unless we expand what practicing data journalism means.

If creating and generating charts or tables from financial or sport statistics qualifies as data journalism, there are many more people who could be fairly said to be practitioners. The number of people applying data science to journalism or practicing high-level computational journalism, however, is clearly far smaller.

The *New York Times*, for instance, has fewer than 10 staff working at that level in the entire organization, according to Aron Pilhofer, of which three are in editorial. “We have data scientists on the business side,” he said. “R&D has a couple, like Mike Dewar, who used to be at Bitly. These are people who are applying data science techniques to actual journalism, stories, infographics, and data visualizations.”

In the United States, much of the top talent in the field is split between the *New York Times*, ProPublica, NPR, the *Washington Post*, the *Chicago Tribune*, the *Wall Street Journal*, and the *Los Angeles Times*, although there are many smaller shops doing great work. In many ways, the *New York Times*’ growing data and interactive teams look a lot like the New York Yankees of data journalism—while they do grow their own players, they also find and acquire the best talent available.

Given the growth of news applications and online interactives in media, investment in this area is looking like a strategic imperative—and developing core competencies in creating them should be a preoccupation of university professors and journalism school deans around the world. When I interviewed academics and some of the leading practitioners of data journalism over the past two years, several obstacles to closing this gap emerged.

The first, mentorship, is common to any profession, but less of an issue in this field. Most data journalists had a mentor or two who guided them early in their development and helped them to get started. A capacity for self-motivation and self-guided learning is important: While mentors played an important role as data journalists developed, in many cases people have picked up the skills on the job or in their free time, learning online and in workshops, not in their undergraduate or graduate educations. In each of the profiles of data journalists that I've published over the years, mentors were an important part of development.

Said [John Keefe](#), the data editor at WNYC:

I could not have done so much so fast without kindness, encouragement, and inspiration from Aron Pilhofer, [Scott Klein](#), [Al Shaw](#), [Jennifer LaFleur](#), [Jeff Larson](#), [Chris Groskopf](#), [Joe Germuska](#), [Brian Boyer](#), and [Jenny 8. Lee](#). Each has unstuck me at various key moments and all have demonstrated in their own work what amazing things were possible. And they have put a premium on sharing what they know—something I try to carry forward. The moment I may remember most was at an afternoon [geek talk](#) aimed mainly at programmers.^{cxxvii} After seeing a demo of a phone app called [Twilio](#), I turned to Al Shaw, sitting next to me, and lamented that I had no idea how to play with such things.

“You absolutely can do this,” he said. He encouraged me to pick up [Sinatra](#),^{cxxviii} a surprisingly easy way to use the Ruby programming language. And I was off.

Sisi Wei, a [news application developer at Propublica](#),^{cxxix} similarly credits many different people for specific skills and ways of thinking:

[Tom Giratikanon](#) showed me that journalists could use programming to tell stories and exposed me to ActionScript and how programming works. [Kat Downs](#) taught me not to let the story be overshadowed by design or fancy interaction, and [Wilson Andrews](#) showed me how a pro handles making live interactive graphics for election night. [Todd Lindeman](#) taught me how to better visualize data and how to really take advantage of Adobe Illustrator. [Lakshmi Ketineni](#) and [Michelle Chen](#) honed my JavaScript and really taught me SQL and PHP.

Now at ProPublica, [my teammates](#) are my mentors.^{cxx} Here is where I learned Ruby on Rails, how news app development really works and how to handle large databases with first ActiveRecord and now ElasticSearch.

New York Times news developer Derek Willis started working with databases in graduate school at the University of Florida:

I had an assistantship at an environmental occupations training center and part of my responsibilities was to maintain the mailing list database. And I just took to it. I really enjoyed working with data, and once I found [Investigative Reporters and Editors](#), things just took off for me. A researcher [at the *Palm Beach Post*], Michelle Quigley, taught me how to find information online and how sometimes you might need to take an indirect route to locating the stuff you want. [Kinsey Wilson](#), now at NPR, hired me at Congressional Quarterly and constantly challenged me

to think bigger about data and the news.

Willis' experience was not unique: The trend that leapt out of the research was the degree to which peer-to-peer learning and peer networks are crucial in the practice and growth of data journalism. (He and other IRE members continue to pay it forward.)

The NICAR listserv is a busy, daily reminder of the generosity of the connected community of over 1700 subscribers. Given the reality of many working journalists who never attended journalism school, mentorship and networked learning will continue to be important factors in the development of more data journalists.

Second, there's improving the level of fundamental numeracy in the media, according to Pilhofer:

Journalism programs need to step up and understand that we live in a data-rich society, and math skills and basic data analysis skills are highly relevant to journalism. The 400+ journalists at NICAR still represent something of an outlier in the industry, and that has to change if journalism is going to remain relevant in an information-based culture. Journalism is one of the few professions that not only tolerates general innumeracy, but celebrates it.

I still hear journalists who are proud of it, even celebrating that they can't do math, even though programming is about logic. It's hard to get a journalist to open up a spreadsheet, much less open up a command line. It is just not something that they, in general, think is held to be an important skill.

It's baffling to me. Look at the *Sun-Sentinel*, which just won another Pulitzer for a story on speeding cops that you could only do with data analysis. You would think you wouldn't have to make the case that this is core to what journalists should know. It's a cultural problem. There is still far too much tolerance for anecdotal evidence as the foundation for news stories.

Like many data journalists I interviewed, Pilhofer originally learned to program because he needed to do something, in this case while he was at the Center for Public Integrity:

I can thank an IRS story on 527 committees, which were then the campaign finance loophole du jour. They were previously unregulated and Congress, in its wisdom, put the IRS in charge of regulating them. It was idiotic. The IRS is not a disclosure agency. They put together the world's worst disclosure website. There was basic data there, but you couldn't aggregate it or access it in a meaningful way. It would have taken thousands of mouse clicks to get all of it.

I talked to a public information officer, after they denied my FOIA request for the database underlying the site. He said it was all on the website. So, I created the world's worst Web scraper in PHP. It ran from the browser. I didn't know the command line well.

Cultural changes will need to start before journalists leave school. "I wish that no j-school ever reinforces or finds acceptable, actively or passively, the stereotype that journalists are bad at math," said Wei. "All it takes is one professor who shrugs off a math error to add to this stereotype, to have the idea pass onto one of his or her students. Let's be clear: Journalists do not come with a math disability." Chase David agreed, saying, "Most journalism students can't code

or do math, while most computer science students don't know storytelling. Hybrids on either side are rare, and we're scooping them up as fast as we can."

Third, students with the most aptitude for data journalism have data science skills that are in high demand in the private sector. In 2013, Indeed.com found that job postings for data scientists had jumped 15,000 percent between the summer of 2011 and 2012. McKinsey & Company [predicted](#)^{cxxi} in 2011 that there would be a 50 to 60 percent shortfall in data scientists by 2018.

The same data science skills that are useful in media, from programming, to statistics, to data cleaning, to analytical thinking are directly transferable to finance, business, or technology jobs, with a much bigger paycheck at the end of the week. In some cases, they're transferable into media as well. Many data journalists didn't go to journalism school, said Chris. W. Anderson, assistant professor of media culture at the City University of New York. For example, people like NPR's Brian Boyer or the AP's Jonathan Stray developed programming skills elsewhere and then entered journalism because of their interest in public interest work.

Organizations are now competing for talent with more than prestigious outlets or broadcast news. "The best people who could help media organizations are getting hired away by Silicon Valley—or 'Silicon Alley'—before they finish j-school," said Anderson. "The top-of-the-line programs at NYU and Columbia are beautiful recruiting grounds for Google or Facebook."

Finally, media companies and journalism schools need to value and fund training in digital skills, from teaching journalists how to use spreadsheets to thinking algorithmically. While not every journalist needs to code, everyone who works in the media does need to be digitally literate, numerate, and understand how technology relates to sourcing, storytelling, and audience development and relationships. The good news is that many journalists have been learning how to use these tools for decades, aided by the experiences and support of others.

"I discovered that I really enjoyed the coding part in addition to reporting," said Aron Pilhofer, associate managing editor for digital strategy at the *New York Times*. "The art of it. That's how I ended up shifting into my current job." Before, he reported about politics:

I was a political reporter, but always used data in my reporting. I just started doing it in college. I just started messing around. I had a history professor who was not well known then. Now, he's borderline famous from doing quantitative methods in history. He'd do statistical sampling of historical census data that had just been paper records before that. Suddenly, you could do queries on the 1930 Census. You were not just basing a historic analysis on papers or on interviews with people, or what you could glean from anecdotes. You were looking at data. It was incredible. That's not that different from a data journalist does, on the CAR side. Instead of a person, you're using data as a source.

Jeremy Bowers, a news developer at the *Wall Street Journal*, started on the tech side:

I started in data journalism at the *St. Petersburg Times*. I'd been working as the blog administrator for our online team and was informally recruited by Matt Waite to help out with a project that would turn into "MugShots."

I have no special degrees or certificates. I was a political science major and I had planned to go to law school before a mediocre LSAT performance made me rethink my priorities. I did have a background in server administration and was really familiar with Linux because of a few semesters spent hacking with a good friend in college, so that's been pretty helpful.

News app developer Dan Hill learned both journalism and computer science at Medill:

I've always wanted to be a reporter, but the work of Phillip Reese at *The Sacramento Bee* and [the Chicago Tribune's news apps team](#)^{cxxii} inspired me to enhance my storytelling with data.

I was a student fellow for the [Northwestern University Knight Lab](#)^{cxxiii} and studied journalism and computer science, but an internship with the *Washington Post* taught me how to apply what I was learning in a newsroom.

AP data journalist Serdar Tumgoren, co-creator of the [Knight News Challenge-funded OpenElections](#) project,^{cxxiv} began chasing documents as a print journalist and picked up new skills as he went:

The document chase quickly broadened to include data, and led me down a traditional “CAR path” of spreadsheets, to databases, to programming languages and Web development. When I first started programming around 2005, I took a Perl class at a community college.

...You don't need a computer science degree to master the various skills of data journalism. I learned how to apply technology to journalism through lots of late-night hacking, tons of programming books, and the limitless generosity of NICARians, who shared technical advice, provided moral support, and taught classes at NICAR conferences.

Mother Jones interactive editor Tasneem Raja also picked up data skills in journalism school:

I was a staff writer at the *Chicago Reader* in the mid-2000s, which was, of course, a scary time to be in news. When a bunch of my senior mentors there, all writers, got canned in 2007, I decided to reevaluate my career and went to j-school at Berkeley to learn new skills. I was lucky enough to be there while Josh Williams was teaching Web development (he left for the *NYT*, where he worked on “Snowfall” and tons of other big interactive pieces), and essentially attached myself at the hip. It turned into a year-long independent study, and got me a job on the launch team at The Bay Citizen, where I created a news apps team that made some really cool data projects for the Bay Area. (RIP, TBC.)

Culture really matters here, said Scott Klein:

People with the right mindset, who feel valued for their editorial judgment and creativity, and who are given real responsibility over their work, will learn whatever they need to learn in order to get a project done. The people on my team focus on telling great journalistic stories and don't let not knowing how to do something stop them from doing so. They learn whatever skills, techniques, and expertise they need to learn.

In terms of journalists learning how to program, I think there are some myths about what programming means. It doesn't have to mean a computer science degree and it doesn't have to mean what Google does. I know journalists who make incredibly complex scrapers for their reporting work who will tell you they don't know how to program. Really, making tools to automate tasks is what a programmer does. There's no magic threshold you have to pass between

programmer and not-programmer.

Of course, there is a difference between knowing how to code and being a computer scientist. If you've learned about algorithmic efficiency and can express it mathematically, and if you've studied how compilers work, all under the guidance of a person who knows the subject very well in an academic environment, you've got skills that will help you write better, faster, more efficient code. That's different than learning how to use a high-level programming language to get a task done.

Much of what we do in newsrooms is on deadline and meant to be put behind a caching system that makes efficient code much less important, so computer science is not a prerequisite for being a great newsroom coder. In newsrooms, most of us rely on frameworks like Rails or Django that already make great low-level programming decisions anyway.

"I suspect it is possible that a journalism degree will become a bolt-on for most of this kind of work," said David Johnson, a journalism professor at American University. "People will probably get their main degrees in hardcore fields, either doing minors in journalism or getting a degree like the [Columbia 2-year](#) or the [Medill program](#)."

Historically, only a few journalism schools have done a good job teaching data-driven journalism, said Anderson. (That's changing, as I explain later.) Much of what's cutting-edge today in data journalism, extending into data science, he suggested, goes well beyond traditional CAR and is being shared through peer-to-peer learning online and in person, at meetups, hackathons, and workshops.

One clear exception, however, lies along the Missouri River, in the center of North America. For decades, the National Institute for Computer-Assisted Reporting ([NICAR](#)) has been one of the most important institutions training journalists to use information technology. Founded in 1989, NICAR is a program of the Missouri School of Journalism and the [Investigative Researchers and Editors](#) (IRE).

Since NICAR was created, the use of data analysis and statistics has evolved into a core component of investigative reporting, augmenting and extending what journalists can do. If you want to find the people doing the best work, look to NICAR's extended community around the globe, subscribe to the email newsgroup, or attend its annual conference, which has become the [preeminent gathering of practicing data journalists in the world](#).^{ccxxv}

There is both a need and a demand for more tutorials and workshops on data-driven journalism tools and best practices beyond those offered in classrooms or at NICAR's annual conference. One of the most common questions I heard from members of the media over the past three years has been, "Where can I go to learn more?" As time has gone on, I've been able to point to more.

Interest in the industry as a whole is present: In the spring of 2013, the University of California at Berkeley's free [online data journalism training](#)^{ccxxvi} sold out and a [Kickstarter to create data journalism educational materials](#) was fully funded.^{ccxxvii} The 10 courses funded by the Kickstarter campaign were taught by leading practitioners in the field. "[For Journalism](#)" endures as a free online resource for anyone who wants to learn more, including webinars, ebooks, code repositories, and forums.^{ccxxviii}

b. Massive Open Online Courses (MOOC) to the Rescue?

In the spring of 2014, over 21,000 participants from more than 170 countries have registered for the online [data journalism course](#) offered by the European Journalism Centre, beginning on May 19.^{cxxxix}

“We’re very excited to be able to offer [this course](#) for free to anyone in the world who has an Internet connection,” wrote Liliana Bounegru, project manager on data journalism at the European Journalism Centre, in an email. “We’re proud to have been able to get support from Google, the Dutch Ministry of Education, and the African Media Initiative for this course and to get some of the best people in the business to teach and provide guidance to the course,” she said, including “the Walter Cronkite School of Journalism, the *New York Times*, ProPublica, *Wired*, Twitter, *La Nación*, the Chronicle of Higher Education, Zeit Online, and others.”

Will the thousands of participants get the skills that they need? A data journalism massive open online course (MOOC) that wrapped in 2013 suggests that many of them will find [the EJC’s course](#) valuable. Last fall, more than [3700 people](#) from 140 countries participated in a MOOC, hosted by the University of Texas, focused on building data journalism skills.^{cxxx} The reviews from participants and instructors were generally good.

Journalist Anna Li took the [online course on data-driven journalism](#) and really enjoyed it, save for some frustrations with the software design.^{cxxxii} Silvia Meave, a Mexico City-based journalist, heard about the MOOC through the Knight Center’s account on Twitter and enrolled. Meave, who has now participated in four MOOCs (two in English and two in Spanish), had not taken classes online previously.

“It’s great for me because I’m studying at my own pace at my computer,” she related in an email interview. Meave pursued and achieved certificates for all four of the MOOCs. “I wanted to get the certificates because these have been great courses,” she said. “I’ve learned too much and it’s good for my curriculum vitae.”

The data-driven journalism MOOC offered Meave an opportunity to obtain training that wasn’t otherwise as easily accessible. “The only way to take this kind of training in Mexico is to enroll in on-site seminars at the university, but I know these topics are not managed at my university yet,” she said.

The views I found on the other side of the screen were decidedly positive. “I thought it was pretty effective,” said Derek Willis in an interview.^{cxxxiii} Willis, who was one of the instructors, told me that this was his first experience teaching or participating in a MOOC:

We were able to put a lot of material in front of the students, with specific, concrete tasks for them to do. My week of it was particularly skills-based, in terms of spreadsheet skills. That can be tricky with two or three thousand students. Not everyone has the same background. The people who really want to stick with them are probably going to power through it, and that was our experience.

Willis’ success was aided by the fact that this wasn’t the first MOOC from the Knight Center for Journalism at the University of Texas. As I learned, it’s managed to conduct [5 MOOCs in 10 months.](#)^{cxxxiii}

The secret to the success, as Amy Schmitz Weiss [explained at PBS *MediaShift*](#), was engaging good instructors and doing a lot of planning.^{cxxxiv} Weiss, an associate professor in the School of Journalism & Media Studies at San Diego State University, offered practical advice for anyone who might try to follow in their footsteps:

Any online course, whether it is a MOOC or not, takes a lot of time to plan and maintain once it launches. Be prepared to spend twice or four times as much of your time on the course. Be available to the students. As the online medium creates an imaginary distance between people, students crave interaction with the instructor to know they are there. Try to be as present as possible in the course. Don't take on too much. Don't try to put everything into the MOOC about the given subject. I started out in the planning phase by including a lot of information and ended up reducing it by half to make sure it would be a manageable amount of course material for the students.

When I interviewed Weiss in December of 2013, she emphasized again how much of a difference thorough preparation made for making the MOOC work. "We planned for two and a half months to put it together," she said. "Planning and organization takes a big chunk of time. Once you have that set, you don't go on autopilot, but it makes the overall course flow much better allows the instructor to focus on teaching and the students."

Weiss shared Willis' positive assessment of the outcome:

Given the topic itself, the approach we took at addressing the basics went really well. Having the right people and right content makes a difference in any kind of course you do, whether face-to-face, online, or if it's a MOOC. For me, I consider it a success if students came away with skills they wouldn't have before and a learning community was formed. Knowing both were achieved makes me very happy.

She was also bullish about what she saw on social media. "The forums were fantastic, seeing the conversations, sharing ideas and examples of data-driven journalism around the world," said Weiss. "Seeing the challenges people had was quite eye-opening, as was seeing them help each other out was great. We have social media channels set up, so that those who wanted to go further could."

So are MOOCs the secret to unlocking [data journalism's secrets](#),^{cxxxv} meeting the demand for people who can build news apps and data visualizations, and crunch government data sets around the world? In a word, no. While the results and experiences of the students and instructors I interviewed are promising, these kinds of online courses will only be a part of the answer, not the singular solution to people's needs for skill development on their own.

"This kind of MOOC was focused on basics, for someone new to it or a citizen who wanted to know how data journalism worked," said Weiss. "More experienced people did sign up, and felt they weren't getting a lot of mileage. We assigned more reading to people who wanted to build those skills."

Given the amount of [vehement criticism](#) about the potential downsides of MOOCs for students and professors,^{cxxxvi} and [misunderstanding](#) of the dynamics of such instruction, any positive recommendation for these online courses should be leavened with a heavy dose of caution.

As Tamar Lewin reported at the *New York Times*, after setbacks, such [online courses are being rethought](#).^{cxxxvii} One of the [most publicized experiments](#),^{cxxxviii} a partnership between Udacity

and San Jose State University, has flopped: The students who participated in the MOOC in the spring of 2012 actually **fared worse** than students who took classes on campus.^{cxxxix}

Many students will benefit from in-person interaction with teachers and peers, from the ability to ask timely questions, and receive immediate follow-up on problems. As more schools experiment with **inverted classrooms**, more classroom time may be maximized in new, more effective ways.^{cxl}

Cheng, a journalism student who logged on from Vancouver, reflected on a challenge that online courses pose in general:

It's great to have direct contact with people. I'm finding that with just an online course, if there's no meeting with the teacher beforehand, even if there's just one day if I get to see the classmates, it's harder.

I'm doing that for Intro to Urban Politics. I have all these other journalism courses and deadlines and I'm constantly reminded of them. With this online course, it gets pushed off, because there's no reminder. I find it a bit of a struggle. I was on top of everything for data journalism because it was something I was personally invested in learning.

The remarks of Stanford professor Sebastian Thrun and challenges that Udacity and other early MOOCs have faced suggest that the current approaches to **distance learning are unsuccessful for most students**,^{cxli} with a **7 percent completion rate**.^{cxlii} In Tanzania, **MOOCs are seen as “too Western.”**^{cxliii} That grim attrition rate doesn't mean, however, that online courses for data-driven journalism or other subjects aren't worth experimenting with further, particularly as network capacity and video conferencing technology improve.

“I think it's having an open mind to new ways of incorporating digital technology to education,” said Rosental Alves, a journalism professor at the University of Texas at Austin and founder of the The Knight Center there, in an **interview with “MOOC News and Reviews”**.^{cxliv}

I think there is a lot of hype about MOOCs now, and there is a lot of negative energy and approach about the impact of the MOOCs, etc. And I think people should calm down and just be open minded to adopt technology in ways that break what we have been doing for centuries in one specific way and be open to check what is effective. It should be brought from the bottom up, not from us who teach and our own interests, but what is the interest of people who are the beneficiaries of the educational process. ... There are people out there who play an important role for democratic society, who are in trouble because the world is changing so rapidly in their area, and they need instruction. They need guidance. It's based on [their interests] that we are doing this.

When I called Alves, he told me that the University of Texas' College of Education has been looking at participants' evaluations and thinking through their approach to MOOCs. “The MOOCs that we do are different from the big movement,” he said. “We are not transforming college classes. In our case, it's more professional training, where we are creating a massive course out of what would be a workshop. It's short and very specific.”

The Knight Center launched this particular program in 2012 and is now working on more MOOCs. One course focused on entrepreneurial journalism is offered in Spanish and, according to Alves, has 5,015 people registered.

He described four broad types of registrants in their classes. The first includes people who register but don't do anything. The second defines people who log-in and take something away, either by watching some videos or downloading some material. The third category is the people who are getting something more from the MOOC. "It's hard to tell how many finish, because the concept of 'finishing' an open course is not very well defined," said Alves.

Cheng, who did not apply for a certificate, fell into this category. "By the third week, I just read everything," she told me. "I watched videos, but when it came to homework or forum questions, I didn't do them any more. It was free, I wanted the information, and I wanted structure."

The fourth category includes the people who pay for their certificates. According to the Knight Center, 278 people paid and earned certificates in University of Texas' data journalism MOOC, out of a total of 3,777 people who registered for the course.

That ratio is in line with other MOOCs: "It has been around 5–10 percent of people who get the certificates, which helps us financially," said Alves. "We charge \$30 if someone wants us to verify in the logs of the platform if they pass the course."

Research suggests those participation rates are roughly similar to other MOOCs. A study of one million MOOC users [published](#) by the University of Pennsylvania Graduate School of Education in December of 2013 found that about 4 percent of users completed the courses, with around half of people who registered for a given course never even signing in to view a lecture.^{cxxlv}

The upcoming European Journalism Center's MOOC is very similar to what the Knight Center offered, said Alves. "It's the same structure, same duration, same topic, and same style, with five different instructors with each one." Whether it gets the same results remains to be seen. A broader way to assess success, he suggested, is to think in terms of connecting peers and mentors, and increasing global access to specialized education:

The beauty of this program is to create a learning community. Each of the six MOOCs that we've done has created a sort of a virtual community with people all over the world interested in the same topic. People help each other a lot. I think people learn from each other as much as from the instructors. We think that's an exciting new opportunity and are very happy to offer these [online classes] to people who wouldn't have any other access to training. We don't want to put any barriers in front of them.

If MOOCs are compared to huge, in-person lectures, they fare better. When it comes to seminars or boot camps, however, their limitations become more apparent.

[Jonathan Groves](#), a journalism professor at Drury University, recommended the [boot camps](#) from Investigative Reporters and Editors for "great hands-on experience, which is the best way to dig into this stuff."^{cxxlvi} There are a couple of important differences between a MOOC and the IRE classes, noted Willis. "In person, you should get more individual instruction," he said. "They are broadly comparable. It can be a good experience. Get the right instructors, do a good job defining the scope of things..."

One concern Willis had turned out better than he expected. "I thought that it would be really tricky to do skills-based learning at scale," he said. "It turned out not to be as big or as bad. I think you can do it, but you're going to need to be very specific, very detailed, and there's only so much you can cover."

On the other hand, instructors can't freelance as they might in a live, in-person lecture. "It's tough to do, and somewhat risky," said Willis. "If someone asks a question in person, you can seize on it and use it to explain some concept to the class. Doing that in person is much easier than doing it in a classroom environment where people aren't all paying attention in the same time and place."

The feedback the University of Texas received from students indicated that the majority of the participants in the MOOC enjoyed it and liked having five instructors. Weiss told me that many students wanted to know when the next course would be offered and to retain access to the materials for a while. A few even wanted to use them to transfer these skills in their newsrooms or hometowns.

Weiss was adamant about continuing to improve the experience. "The MOOC model is something we can experiment with, and challenge ourselves to be better teachers and contribute to society," she said. "It's a way to scaffold learning, and to learn how to do it better. We can't afford not to take the risk as educators. We owe it to the students to keep experimenting and trying it out."

c. Hacks, Hackers, and Peer-to-peer Learning

MOOCs and online resources like "For Journalism" will offer those already in the industry a better, more flexible place to get started, along with those looking to break in a place to enter. Some journalists, however, won't be comfortable learning from a book or online alone: They need someone to answer questions and explain analogies. In other words, there's going to be continuing need for in-person, human-to-human interaction around learning.

"Journalism schools still teach journalism as a very hierarchical, often solitary pursuit," said Tasneem Raja. "That's not the way it works in data journalism, and the best learning is still gonna be on the job. That requires cross-pollination between folks with different skill sets. We need a pairing model across newsrooms, not just in the nerd corner."

People who want foundational skills need to get hands-on with dirty data and the tools needed to clean, organize, and present it. There are a number of non-governmental organizations that provide such forums, workshops, classes, and education, including [DataKind](#)^{cxxvii}, the [Sunlight Foundation](#)^{cxxviii}, the [World Bank](#)^{cxxix}, the [Open Knowledge Foundation](#)^{ccl}, [Code with Me](#)^{ccli}, and [Hacks and Hackers](#)^{cclii}, which now has dozens of chapters and thousands of members around the world.

Many "hacker journalist" projects and classes require collaboration with people outside of the journalism school, said Anderson, especially if professors don't have the needed skills to teach the students.

d. Journalism Schools Rise to the Challenge

While many data journalists enter the profession without a journalism degree, as is true for many people writing and reporting today, industry demand for data skills is leading to changes in the academy. In 2014, the University of Missouri is far from alone in teaching journalists how to treat data as a source.

For instance, if the [Knight Lab](#) at Northwestern University's Medill School can guide promising young [data journalists](#) like Dhruvil Mehta into journalism, they're doing something right.^{ccli} Along with the graduate school offering classes on [enterprise reporting with data](#)^{cclv} and [interactive storytelling with JavaScript](#),^{cclv} Medill is offering [scholarships to people with programming backgrounds](#).^{cclvi} The school is also [pairing journalism and computer science students together to develop interactive projects](#).^{cclvii}

Other universities are also building capacity to teach in these areas by collaborating with computer science departments. For instance, in England, [Cardiff University will introduce a masters in computational journalism](#).^{cclviii}

At the Columbia University Graduate School of Journalism, the [Lede Program](#)^{cclix} will seek to “[go beyond the data hype](#)” with a post-baccalaureate certification course that offers training in data, code, and algorithms to journalists.^{cclx} It will equip students with the technical skills required to enroll in the dual [Journalism/Computer Science master’s program](#) that Columbia began to offer in 2010.^{cclxi}

Journalism schools are also bringing practitioners with data journalism skills onto the faculty. At Missouri, Chase Davis teaches students [how to apply data science to all the news that’s fit to print](#).^{cclxii} While Davis said that journalism schools could be doing more to adjust to the changing needs of students, he emphasized that the current situation is not all educators’ fault:

It takes intellectual agility and natural curiosity to effectively develop hybrid skills. I don’t think that’s something we can teach solely through curriculum. That’s why I don’t think every journalism student should learn how to code. Being able to write a few lines of JavaScript is great, but if you let your skills dead end with that, you’re not going to be a great newsroom developer.

Folks on our interactive and graphics teams at the *Times* have remarkably diverse backgrounds: journalism and computer science, sure, but also cartography, art history, and no college degree at all. What makes them great is that they have an instinct to self-teach and explore.

That’s what journalism schools can encourage: Introduce data journalism with the curriculum, then provide a venue for students to tinker and explore. Ideally, someone on faculty should know enough to guide them. The school should show an interest in data journalism work on par with more traditional storytelling. Oh, and they should require more math classes.

In Philadelphia, [Temple is helping to ensure the future of data journalism](#)^{cclxiii} with a new course taught by assistant professor [Meredith Broussard](#), a computer scientist-turned-reporter. She starts her students by grounding them in the social sciences, a context that recalls Philip Meyer’s formative approach to precision journalism:

We read Joel Best’s *Damned Lies and Statistics* and talk about how data comes into being. Then, we go on to data analysis and we practice different ways of representing data. This might be infographics, or data visualization, or pivot tables in Excel. I focus on teaching the students how to use technological tools in the service of doing a story. We cover a variety of digital tools, and we analyze examples of journalists and scholars who are doing intellectually exciting work.

Broussard said that data journalism is now part of the curriculum at Temple at every level:

At Temple, our students have a well-rounded education that includes essential reporting skills, critical thinking, multimedia storytelling skills, visual analysis, and much more.

In our intro class this year, the students learned about Journalism++, Vox, FiveThirtyEight.com, and a handful of other exciting journalism projects. We had Aron Pilhofer from the *New York Times* as a guest speaker to talk with the students about what it's like to do data journalism in a world-class newsroom.

Students encounter data journalism again as curriculum units in mid-level classes: Our multimedia storytelling class does a unit about online data journalism, and our journalism research class introduces students to data analysis using Excel. Everyone has to fulfill a quantitative requirement, ensuring that all the students have basic statistical literacy.

I teach an upper-level class called "Data Journalism" in which the students do advanced data analysis with Excel, create data visualizations, work with databases, and create an original data journalism project. This semester, I had amazing student projects. Innovative news apps, visualizations I never imagined, infographics that were playful yet powerful. My students always impress me.

In Florida, the University of Miami is now deeply integrating data and visualization into its curriculum as well, explained Alberto Cairo, director of the visualization program at the Center for Computational Science at the university and professor of practice in the journalism department.

Visualization classes are now part of the core program for undergraduate journalism majors and in the university's master's degree program, along with mandatory introductions to design and Web design. He said in an interview:

We have hired two professors to teach data journalism and Web development classes. These classes are closely tied to the current Web design and visualization courses. Besides our journalism programs, we have an MFA in Interactive Media, and also a minor for undergrads. Journalism students can take classes in those programs as part of their electives (and vice versa). That is leading to strengthened ties with science departments across the university.

In California, veteran data editor [Cheryl Phillips was named a lecturer at Stanford](#),^{cclxiv} where she'll apply her experience as an award-winning investigative reporter to teaching classes on relational data, basic statistics, investigative reporting tools, and mapping at Stanford's Computational Journalism Lab. She spoke of evolution in education in an interview:

I think it's no secret that a lot of change is starting to take place in schools. Cindy Royal had an interesting piece about [platforms](#) just the other day.^{cclxv} We need to take a more integrated approach. Classrooms and their teachers should collaborate on work. For example, a multimedia class produces the visualizations and videos that go with the stories being written in another class. Stanford already does this.

Like Broussard and Davis, Phillips says that data journalism shouldn't be limited to just one class but infused into every part of the university curriculum:

Every type of journalist can learn data-related skills that will help them, whether they end up as a copyeditor, a reporter, a front-line editor, or a graphics artist. In general, I want to make sure the students are telling stories from data that they analyze. [They should be] not only learning the technical stack, but how to apply the technical knowledge to real-world journalism. I am hoping to create some partnerships with newsrooms as well.

VI. Tools of the Trade

a. Digging into the CAR Toolbox

As is true in the trades, the arts, and the sciences, the tools data journalists choose are driven by the needs of a given project, available resources, expertise, training, and time. These can be divided into five rough categories: data collection, cleaning, analysis, presentation, and publishing.

Cleaning data “is often the most time consuming part of the data journalism process,” said Jonathan Stray, an instructor at Columbia Journalism school, who has highlighted the widespread problem of governments publishing data locked up in the Portable Document Format (PDF) and the [heroic measures](#) needed to deal with the challenge.^{cclxvi}

A crucial part of what’s needed to practice data journalism, however, has little to do with tools and technology and everything to do with perspective and critical thinking. “You need a mindset which is about putting this in the context of the story and spotting stories, as well as having creative and interesting ideas about how you can actually collect this material for your own stories,” said Emily Bell. “It’s not a passive kind of processing function if you’re a data journalist: It’s an active speaking, inquiring, and discovery process. I think that that’s something which is actually available to all journalists.”

If you look at [data journalism and the big picture](#), more recent technologies are part of a continuum of technologically enhanced storytelling that traces back decades.^{cclxvii} The canonical suite of tools for computer-assisted reporting ran on desktops and servers, spreadsheets, databases, text editors, and statistics software.

Spreadsheets were the first “killer app” for data journalism, just as VisiCalc was the first killer app for the Apple computer. In many ways, they still are, even if the spreadsheets have become Web-based. Chris Amico and Laura Norton Amico’s work on Homicide Watch started as a spreadsheet and expanded over time. “No matter how advanced our tools get, I always find myself coming back to Excel first to do simple work,” said Minkoff, a data journalist at the Associated Press. “It helps us get an overall handle on a data set.”

After spreadsheets, the second most common tool applied in the field is database software, in particular, Microsoft Access, [MySQL](#), [PostgreSQL](#), or [SQLite](#). A text editor, like [TextMate](#) or [BBEdit](#), and statistics software, like [SPSS Statistics](#), round out the basic suite of tools that have been used for CAR for many years.

Today, data journalists leverage Web-based tools for data collection, manipulation, analysis, and visualization, like [Open Refine](#), [Google Fusion Tables](#), and [Tableau](#). They’re also working with modern programming languages, like [Python](#), [Ruby](#), and [Javascript](#), as well as [d3](#), a Javascript library.

“We love tools that don’t need a developer every time to create interactive content,” said Momi Peralta. “These are end user’s tools. Google Docs, spreadsheets, Open Refine, Junar’s open data

platform, Tableau Public for interactive graphs, and now Javascript or D3.js for reusable interactive graphs tied to updated data sets.”

Tool choice brings with it the thorny issue of newsroom culture, as previously referenced, right down to organizational DNA that venerates narrative writing and mistrusts the messy news environment online that is slow to adopt new technologies. It wasn’t so long ago that the people in charge of a newspaper’s website worked in different departments or even buildings than reporters working a story. (That’s still true in some media companies.)

The integration of the Internet into the collection and production of the news demonstrates that traditional media institutions can and will adapt and adopt new technologies and practices. That will continue to accelerate globally, once the advantages of data-driven storytelling become apparent.

b. New Tools to Wrangle Unstructured Data

The rapid expansion in the amount of unstructured data,^{cclxviii} however, has further changed the context and need for this kind of expertise in-house. When the *Guardian*’s data team was faced with making sense of the WikiLeaks cables, it took months to work through them.^{cclxix}

“For a long time, we’ve been hammering governments to give us data in columns and rows,” said Cohen. “I think we’re increasingly seeing that stories just as likely (if not more likely) come from the unstructured information that comes from documents, audio and video, tweets, other social media—from government and non-government sources.”

Making sense of all of that data is both a huge opportunity and an immense challenge for newsrooms. Once upon a time, it was difficult for investigators to find information relevant to answering a question. Today, in many (if not all) scenarios, the opposite is true, particularly in a world where readers have access to search engines. That has shifted the value that journalists can add—from finding information to making sense of what’s actually happening, processing, analyzing and vetting data, and finding signal in the digital noise.

That new landscape is precisely why the Knight News Challenge gave \$1.5 million to projects that filter and examine data.^{cclxx} Those include the PANDA Project,^{cclxxi} which tries to make research easier in the newsroom with a set of open source,^{cclxxii} Web-based tools oriented at making it easier for journalists to use and analyze data, and Overview,^{cclxxiii} which helps journalists find stories by cleaning, visualizing, and interactively exploring large documents and data sets, acting as a kind of “editorial search engine.^{cclxxiv}”

Associated Press data journalist Jonathan Stray, Overview’s project manager and a research fellow at the Tow Center, describes it as a organizational structure for data.^{cclxxv} Both PANDA and Overview are squarely aimed at bread-and-butter issues for newsrooms struggling to manage data. As of March 2014, PANDA has been installed in 25 newsrooms around the United States.

“It’s a pain to search across data sets, but we also have this general newsroom content management issue,” said Brian Boyer, the product manager for PANDA and head of NPR’s News Applications team. “The data stuck on your hard drive is sad data. Knowledge

management isn't a sexy problem to solve, but it's a real business problem. People could be doing better reporting if they knew what was available. Data should be visible internally."

Boyer thinks the trends toward big data in media are clear, and that he and other hacker journalists can help their colleagues to not only understand it, but to thrive. "There's a lot more of it, with government releasing its stuff more rapidly," he said in 2012. "This city of Chicago is releasing a lot of it. We're going for increased efficiency, to help people work faster and write better stories. Every major news org in the country is hiring a news app developer right now. Or two. For smaller news organizations, it really works for them. Their data apps account for the majority of their traffic." Once such databases are up and running, journalists can apply analytical tools to produce evidence-driven reporting.

The [difficulty](#) ProPublica had with building the "Dollars for Docs" project puts the scale of that work into perspective, from converting PDFs to dirty data, to fact-checking correlations within the massive databases.^{cclxxvi} Those who wish to follow their lead should read Dan Nguyen's [guide to scraping data](#),^{cclxxvii} Scott Klein's [style guide for news apps](#),^{cclxxviii} and Jacob Harris' exploration of "[how data sausage is made](#)".^{cclxxix}

As journalists start working more with data, they have more choices for tools than ever before. There is also powerful new data-journalism software coming online, from analysis to visualization tools. As [Eric Newton highlighted](#) at the Knight Foundation, many of these new tools help journalists gather, clean, analyze, and publish data and do not require sophisticated programming knowledge to use.^{cclxxx}

Open source tools are also popular. As [Dan Sinker](#), the head of the Knight-Mozilla News Technology Partnership for Mozilla, wrote last year, [journo-coders](#) are now taking social coding "to a whole new level."^{cclxxxi} Just as civic software^{cclxxxii} is increasingly baked into government, [open source](#) is playing a pivotal role in the practice of data journalism.^{cclxxxiii} While many news developers are agnostic with respect to which tools they use to get a job done, the people who are building and sharing tools for data journalism are often doing it with open source code.

While some of that open source development has been driven by the requirements of the Knight News Challenge, which funded the PANDA and Overview projects, there's a broader collaborative spirit evidenced in the interstitial communication on Twitter, GitHub, and mailing lists that connect the data-driven journalism community around the world.

Members of newsrooms that compete on beats are working together on code. For instance, [New York Times](#) and [Washington Post](#) developers are teaming up^{cclxxxiv} to create an [open election database](#).^{cclxxxv} Data journalists from WNYC, the *Chicago Tribune* and the *Spokesman-Review* are collaborating on building a [better interface for Census data](#).^{cclxxxvi} The same kinds of [peer networks](#) that helped build the Internet are building out civic infrastructure.^{cclxxxvii}

Beyond their contributions to the [newsroom stack](#),^{cclxxxviii} this is a group of people I found to be fiercely committed to "showing your work." For data journalists, that means sharing your source data, methodology, and code, not just a notebook. To put it another way, "[code, don't tell](#)".^{cclxxxix}

VII. Open Government

a. Open Data and Raw Data

Building capacity in data journalism is directly connected to the role the Fourth Estate plays in democracies around the world. There are important stories buried in that explosion of data from government, industry, media, universities, sensors, and devices that aren't being told because the perspective and skills required to do it properly aren't widespread in the journalism industry. The need for data-driven journalism comes at a time, unfortunately, when the news organizations that have housed them over the past centuries are contracting.

As that's happening, the demand for information about government is growing, in the areas of service, performance, and spending. Every day, more citizens turn to the Internet for government information,^{cxc} searching for more data, policy, and services. Research on community information systems from the Pew Internet and Life Project shows strong citizen interest in online resources for government and civic information.^{cxcii} When citizens are both aware of government information being released and can find it, open government policies can lead to higher levels of community satisfaction.^{cxciii} At the local level, however, limited budgets and technical ability will make opening data difficult.

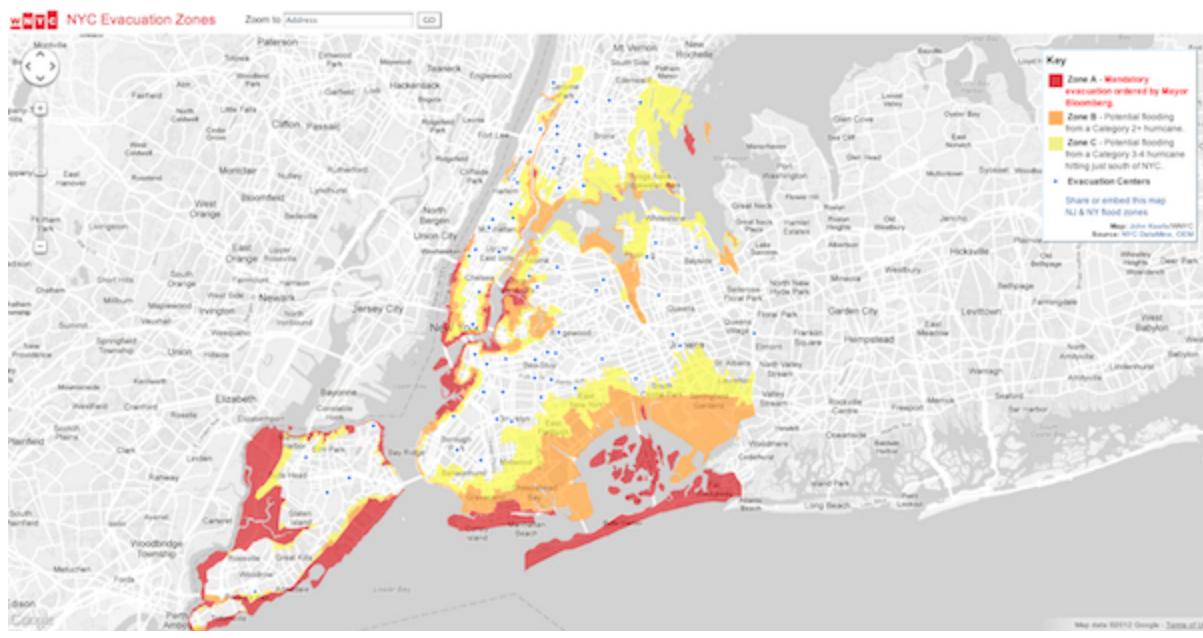
This situation may grow worse as more local newspapers close. That trend was one of the drivers for the landmark Knight Commission on the Information Needs of Communities in a Democracy.^{cxciii} One strategy for empowering citizens to be more informed about their communities includes a recommendation to create local online hubs based upon open government data.^{cxciv}

In the past several years, open government data platforms have grown around the globe, including in a number of big cities, providing more raw material for data journalism.^{cxcv} Said Momi Peralta of *La Nación*:

The open government movement is happening. We must be ready to receive and process open data, and then tell all the stories hidden in data sets that now may seem raw or distant. To begin with, it would be useful to have data on open contracts, statements of assets, and salaries of public officials, ways to follow the money and compare, so people can help monitor government accountability. Although we dream in open data formats, we love PDFs against receiving print copies.

The rewards that cities like New York have reaped from adopting a platform strategy are no longer theoretical, given that public open government data feeds become critical infrastructure during natural disasters.^{cxcvi} In New York City, as the city's websites faced heavy demand when residents went to its hurricane evacuation finder in advance of Hurricane Sandy, residents could also go and consult WYNC's lightweight, mobile device-friendly evacuation map.

WNYC data news editor John Keefe was responsible for the map, which put the city's open government data in action.^{cxcvii}



[WNYC Evacuation Map]

“We estimate that collectively we served and informed 10 times as many individuals by embracing an open strategy,” wrote Rachel Haot, then New York City’s chief digital officer, in a [blog post](#) for the Open Government Partnership.^{cxcviii} “That’s hundreds of thousands of people.”

If the evolutionary descendants of EveryBlock are ever going to be a meaningful replacement for local newspapers, however, they’ll need to be sustainable, independent from government’s influence, deliver a valuable information product and be interesting. They’ll have to feature compelling storytelling that’s citizen-centric, uses adaptive design, and provides information that’s relevant to what people need to know, now.

That’s a tall order but there’s hope: Hundreds of entrepreneurial journalists are working on creating versions of that future today, with more to come.

b. Data and Ethics

In recent years, more local, state, and national governments have begun proactively releasing public sector data in hopes of stimulated economic effects, improving services, or enhancing transparency and accountability. When these data sets detail performance, spending, budgeting, or services, if they do not include deliberations or policy decisions—which is to say how power or influence is exercised—journalists have to keep digging, scraping, and investigating.

There are good reasons for journalists to be careful about a complete embrace of open government data, at least with respect to the data’s relationship to government transparency. There’s now considerable [ambiguity regarding open government](#), as a 2012 paper on “[The New Ambiguity of ‘Open Government’](#)” by Princeton scholars David Robinson and Harlan Yu explored. From their abstract:

Open technologies involve sharing data over the Internet, and all kinds of governments can use them, for all kinds of reasons. Recent public policies have stretched the label “open government” to reach any public sector use of these technologies. Thus, “open government data” might refer to data that makes the government as a whole more open (that is, more transparent), but might equally well refer to politically neutral public sector disclosures that are easy to reuse, but that may have nothing to do with public accountability. Today a regime can call itself “open” if it builds the right kind of website—even if it does not become more accountable or transparent. This shift in vocabulary makes it harder for policymakers and activists to articulate clear priorities and make cogent demands.^{cxxix}

As skeptical data journalists know, there’s a difference between open data that’s proactively disclosed by governments and data buried in PDFs released in response to the Freedom of Information Act or lawsuits by media companies and advocates.

That said, there’s much to be gained by pitching a [big tent for open government](#), as Joshua Goldstein and Jeremy Weinstein argued in a response to Yu and Robinson in the *UCLA Law Review*, including benefits for data journalists. They wrote:

It is difficult to disagree with Yu and Robinson’s narrowest claim. Greater clarity about the complementary but distinct objectives of these different movements—and the likely impact of the specific governmental policies they advocate—is undoubtedly a good thing.

But saying that open data and open government can exist without the other, is not the same as saying that they should. Drawing on our respective experiences as a partner in Kenya’s Open Data effort and as a key architect of President Obama’s multilateral Open Government Partnership, we argue that the growing ties between the open data and open government movements, particularly in developing countries, can benefit *both* agendas.^{ccc}

Support for more open data releases was prevalent among most data journalists interviewed for this report, although it was coupled with ample caution and caveats.

“I can’t find any downsides of more data rather than less,” said Sarah Cohen, of the *New York Times*, “but I worry about a few things.” First, emphasized Cohen, there’s an issue of whether data is created open from the beginning—and the consequences of “sanitizing” it before release.

“The demand for structured, nicely scrubbed data for the purpose of building apps can result in fake records rather than real records being released,” Cohen said. “[USASpending.gov](#) is a good example of that—we don’t get access to the actual spending records like invoices and purchase orders that agencies use, or the systems they use to actually do their business. Instead we have a side system whose only purpose is to make it public, so it’s not a high priority inside agencies and there’s no natural audit trail on it. It’s not used to spend money, so mistakes aren’t likely to be caught.”

Second, there’s the question of whether information relevant to an investigation has been scrubbed for release, said Cohen:

We get the lowest common denominator of information. There are a lot of records used for accountability that depend on our ability to see personally identifiable information (as opposed to private or personal information, which isn’t the same thing). For instance, if you want to do

stories on how farm subsidies are paid, you kind of have to know who gets them. If you want to do something on fraud in Federal Emergency Management Agency claims, you have to be able to find the people and businesses who get the aid. But when it gets pushed out as open government data, it often gets scrubbed of important details and then we have a harder time getting them under the Freedom of Information Act because the agencies say the records are already public.

To address those issues, Cohen recommends getting more source documents, as a historian would. “I think what we can do is to push harder for actual records, and to not settle for what the White House wants to give us,” she said. “We also have to get better at using records that aren’t held in nice, neat forms—they’re not born that way, and we should get better at using records in whatever form they exist.”

Much of the time, government data is often “dirty,” with missing metadata, incorrect fields, or gaps in collection. Journalists have to extract data from PDFs, validate it, and clean up data sets^{ccci} to make them usable in applications, report it out, and then present it in context.

If the capacity to practice is there, data journalism can deliver notable results. For instance, ProPublica’s “Recovery Tracker”^{ccci} for stimulus data and projects is one of the best examples of the practice in action. Another gold standard for data journalism is the Pulitzer Prize-winning “Toxic Waters”^{ccci} project from the *New York Times*. The scale of that project makes it a difficult act to follow, though *Times* developers are working hard with projects like “Inside Congress.”^{ccci}

The parallels to what civic hackers are doing and what data journalists are working on is inescapable. Both are focused on putting data to work for the public good, whether it’s in the public interest, for profit, in the service of civic utility or, in the biggest crossover, government accountability.^{cccv} Said Peralta:

The open data movement and hacktivism can accelerate the application of technology to ingest large sets of documents, complex documents, or large volumes of structured data. This will accelerate and help journalism extract and tell better stories, but also bring tons of information to the light, so everyone can see, process, and keep governments accountable.

The way to go for us now is use data for journalism but then open that data. We are building blocks of knowledge and, at the same time, putting this data closer to the people, the experts and the ones who can do better work than ourselves to extract another story or detect spots of corruption.

It makes lots of sense for us to make the effort of typing, building data sets, cleaning, converting, and sharing data in open formats, even organizing our own “datafest” to expose data to experts. Open data will help in the fight against corruption. That is a real need, as here corruption is killing people.

To do so will require that data journalists and civic coders alike apply the powerful tools to the explosion of digital bits and bytes from government, business, and our fellow citizens. The need for data journalism, in the context of massive amounts of government data being released, could not be any more timely, particularly given persistent quality issues.

“Open government data means that more people can access and reuse official information

published by government bodies,” said Bounegru. “This in itself is not enough. It is increasingly important that journalists can keep up and are equipped with skills and resources to understand open government data. Journalists need to know what official data means, what it says, and what it leaves out.”

That requires journalists to possess both numeracy and digital literacy, if they’re going to interrogate the data. “Only by equipping journalists with the skills to use data more effectively can we break the current asymmetry, where our understanding of the information that matters is mediated by governments, companies, and other experts,” said Bounegru. “In a nutshell, open data advocates push for more data, and data journalists help the public to use, explore, and evaluate it.”

Open data needs to find people, not vice versa. For that to happen, supporting and extending the capacity of the media to practice data-driven journalism is a fundamental part of the equation. The role that the Fourth Estate plays in holding governments to account in the 21st century is no less pressing than in decades past. If anything, given how power is gathered and exercised in secret around the world, it’s more so.

There’s a long history of elected officials or government staff who want to prevent information that shows fraud, undue influence, embarrassing behavior, or outright criminality from coming to the public’s attention. That’s true today as well. To preserve such evidence, data journalists will also need to securely protect data, just as editors have historically protected human sources. When great investigative work is paired with data journalism, remarkable outcomes bloom.

“We took narrative reports from nursing home inspections and made them searchable^{ccvii} nationwide by keyword, something the government doesn’t allow,” said Ornstein, a senior reporter at ProPublica. The resulting data-driven tool, which enables people to shop for nursing homes online,^{ccviii} is a 21st-century version of service journalism, giving people a way to make more informed decisions and adding an accountability mechanism for businesses and government in the process.

At ProPublica, the data journalism team is conscious of deep linking into news applications, with the perspective that the visualizations produced from such apps are themselves a form of narrative journalism. With great data visualizations, readers can find their own way and interrogate the data themselves. Moreover, distinctions between a news story and a news app are dissolving as readers increasingly consume media on mobile devices and tablets.

One approach to providing useful context is the “Ion” format at ProPublica.org, where a project like “Eye on the Stimulus”^{ccviii} is a hybrid between a blog and an application. On one side of the Web page, there’s a news river. On the other, there are entry points into the data itself. The challenge to this approach is that a media outlet will need data specialists to work closely with the investigators—or that they become one and the same.

While that’s true regardless of the context, building data-driven capacity will necessarily start in different levels in different media cultures and climates. “Investigative journalism in Africa, like in many other places, tends to be scoop-driven, which means that someone has leaked you a set

of documents,” said Justin Arenstein, a Knight International fellow embedded with the African Media Initiative (AMI)^{cccxix} as a director for digital innovation.

“There are very few systematic, analytical approaches to analyzing broader societal trends,” he said. “You’re still getting a lot of hit-and-run reporting. That doesn’t help us analyze the societies we’re in, and it doesn’t help us, more importantly, build the tools to make decisions.”

The strategy that Arenstein and the AMI is pursuing diverges from the news applications and data visualizations that are common outcomes of data journalism in Europe and the United States. They don’t just tell a story but give people a tool to understand a specific area, make a decision, and then take action. Arenstein emphasized the need to think deeply about how journalists use data in investigations, as opposed to raw material for a visualization. The strongest commonalities between the work Code for Kenya is doing and ProPublica in the United States, in fact, lie in their use of data to support and augment investigative work, mapping the relationships of the powerful, and funding projects on extractive industries.

“We’re finding something that maybe you’re starting to see inklings of elsewhere as well: Data journalism doesn’t have to be the product,” he said. “Data journalism can also be the route that you follow to get to a final story. It doesn’t have to produce an infographic or a map.”

c. Gun Data, Maps, and Radical Transparency

The confluence of public data, digital media, and democratized publishing technology is going to lead media and advocacy organizations into challenging, uncomfortable places. Many of the issues data journalists face will be long-standing ones, like intransigent public officials or huge paper document dumps.

For instance, in the 1990s the District of Columbia water authority refused to publish the results of lead testing after it showed widespread contamination. “We got the survey from a source, but it was on paper,” said Cohen. “After scanning, parsing, and geocoding, we sent out a team of reporters to neighborhoods to spot check the data, and also do some reporting on the neighborhoods. We ended up with a story about people who didn’t know what was near them.”

In a harbinger of tensions to come, the *Washington Post* team chose not to publish the addresses of people identified in the data set. “The water authority called our editor to complain that we were going to put all of the addresses online—they felt that it was violating privacy, even though we weren’t identifying the owners or the residents,” said Cohen. “It was more important to them that we keep people in the dark about their blocks. Our editor at the time, Len Downie, said, ‘You’re right. We shouldn’t just put it on the Web.’ ”

At the end of 2012, similar questions arose when *The Journal News*, a newspaper in New York, displayed the names and addresses^{cccx} of handgun permit holders in an online map that was based upon the government’s regulatory data. The outrage^{ccxi} that resulted was instructive: This data was public and subject to a Freedom of Information law. Did that make it ethically sound to publish the names and addresses of permit holders?

The question of what to do about guns, maps, and disturbing data^{ccxii} in New York was

answered in part by the state legislature and senate, when it passed legislation that created an anonymity exemption^{cccxiii} for gun permit holders. The issues this situation raised, however, will be central to data journalism in every state and country around the world.

The conflict over guns and data showed how government data could be used by journalists in ways that could make many citizens quite uncomfortable.^{cccxiv} It also highlighted an issue with data quality and journalism: More than three quarters of the data in the gun map was inaccurate.^{cccxv} *The Journal News* took the map offline^{cccxvi} in January of 2013, although a version of it endures with zooming and data access disabled.

The reality is that government data is already consulted and used daily by media. Given the increased reach and velocity of digital media, data journalists must be more conscious of ethics than ever. “Journalists broadcast and publish criminal records, drunk driving records, arrest records, professional licenses, inspection records, and all sorts of private information,” wrote Al Tompkins,^{cccxvii} a senior faculty member at the Poynter Institute. “But when we publish private information we should weigh the public’s right to know against the potential harm publishing could cause.”

Journalists need to know how to turn data into journalism in a way that serves the public interest without harming it.^{cccxviii} If you’re applying a data-driven lens, as Jeff Sonderman highlighted at the Poynter Institute, you’ll need to ask a series of basic questions. He wrote:

In every situation you face, there will be unique considerations about whether and how to publish a set of data. Don’t assume data is inherently accurate, fair, and objective. Don’t mistake your access to data or your right to publish it as a legitimate rationale for doing so. Think critically about the public good and potential harm, the context surrounding the data, and its relevance to your other reporting. Then decide whether your data publishing is journalism.^{cccxix}

The issue of data journalism’s potential harms came up when Wikileaks released data from the U.S. Department of Defense and Department of State to multiple news organizations in 2010 and 2011. Every media organization that reviewed classified cables or logs from the Pentagon and State Department had to decide not only whether to publish them but how, balancing redacting the names of people who might be put at risk with the public’s right to know what was done on its behalf by government.

The technical capacity to move through millions of lines of messy data in proprietary formats, however, only rests with a limited number of news organizations. If the capacity to do data journalism at scale isn’t democratized, this dynamic could enshrine traditional media power structures.

“I helped out with the Wikileaks War Logs reporting,” said Jacob Harris, a data journalist at the *New York Times*. “We built an internal news app for the reporters to search the reports, see them on a map, and tag the most interesting ones. One of the unique things I figured out was how to extract MGRS [Military Grid References System] coordinates from within the reports to geocode the locations inside of them. From this, I was able to distinguish the locations of various homicides within Baghdad more finely than the geocoding for the reports. I built a demo, pitched it to graphics, and we built an effective and sobering look at the devastation on Baghdad from the

violence.”^{ccxx}

d. A Global Lens on FOIA and Press Freedom

In the United States, data journalists often run into bureaucracy, obfuscation, or years of drawn-out wrangling over Freedom of Information Act requests, fees, and redactions. Journalists trying to acquire or use data in countries without freedom of information laws or democratic institutions have an even harder time gaining the raw material for their stories.

Charles Andersen said that the issue of open government is hugely important to questions of data journalism’s future and relevance. Andersen, who co-authored a landmark [report on post-industrial journalism](#) with Emily Bell and Clay Shirky,^{ccxxi} said that a tradition of open government—which increasingly includes efforts to open data—is probably the biggest factor in the success of data journalism in developing countries. “Data journalists have a very hard time existing in countries where there isn’t open data,” he said. “For instance, there’s a huge difference between Germany and the United States. Germany has relevant laws but a culture of not sharing.”

The United States, at least by contrast, has a tradition of openness and government disclosure, said Andersen. Their research suggests that data journalism cannot exist in a given country without open government laws and policies. If elected officials, legislators, and staff want to see media using open data, they should also take substantive steps to ensure that policies, licenses, laws, and regulations are in place to permit that reuse. Similarly, if public services based upon open data feeds are performed by private parties, freedom of information laws in many countries may well need to be extended to the entities that deliver those services. Open data initiatives that aren’t accompanied by freedom of the press or freedom of information laws are unlikely to deliver on political rhetoric promising increased transparency or accountability.

VIII. On to the Future

Recommendations and Predictions

The world needs journalists with these skills more than ever. The same trends changing journalism and society^{ccxxii} have the potential to create significant social change throughout the world, as nation states move from conditions of information scarcity to abundance, causing vast disruptions to governance and governments.

Journalists have always needed to be able to write, interview, and fact-check their work. Today, photography, social media, video editing, and mobile devices have already become integral elements of the toolkits of many journalists. Whether news developers are rendering data in real time,^{ccxxiii} validating data in the real world,^{ccxxiv} or improving news coverage with data,^{ccxxv} good data journalism still must tell a story, solve a problem, or speak truth to power. Smartphones, notebooks, cameras, social media, and data sets can extend investigations in important ways.

In the near future, expect basic data-science skills to become baked into how investigative journalists gather sources, find evidence, and present their findings—from building databases, to creating visualization, to applying powerful analytical software. Along with those skills, journalists will still need to apply critical thinking and show how they reached conclusions.

While the need is acute and journalism schools are responding, significant cultural, fiscal, and technical barriers to the adoption of data journalism and digital skills remain. In May of 2014, a new report^{ccxxvi} by the Duke Reporters' Lab at the DeWitt Wallace Center for Media and Democracy in the Sanford School of Public Policy surveyed 20 newsrooms to find which digital tools are still missing. The top-line conclusions from Mark Stencel, Bill Adair, and Prashanth Kamalakanthan painted a sobering picture of an industry in flux.

The report found that many U.S. newsrooms aren't taking advantage of new, low-cost digital tools for reporting and presenting journalism, instead continuing to use familiar methods and practices. Its authors suggest that journalism awards and popular media conferences have created the perception that the adoption of digital tools and data journalism is more prevalent than it is.

While local newsroom leaders told the researchers that budget, time, and people were their primary constraints, deeper infrastructure and cultural issues are hindering adoption. The report describes an industry with a gap between "have and have-nots," with national organizations experimenting with data journalism and new digital tools while local newsrooms are not.

"The local newsrooms that have made smart use of digital tools have leaders who are willing to make difficult trade-offs in their coverage," write the authors. They prioritize stories that reveal the meaning and implications of the news over an overwhelming focus on chasing incremental developments. They also think of the work they can do with digital tools as ways to tell untold stories—not 'bells and whistles,' " wrote the authors.

Writing at Poynter.com,^{ccxxvii} Howard Finberg noted that the Duke report's conclusions support findings of Poynter's recent "Core Skills for the Future of Journalism" report, which was based on a broader sample of the industry—that is, more than 2,900 responses from media organization professionals, independent or freelance journalists, educators, and students. "Professional journalists in legacy media rated new digital skills as much less important than traditional skills," he wrote. "Educators, students, and independent journalists rated digital skills as much more important than the professionals."

Finberg's discussion of the report's finding and data journalism is a reality check on the challenges that remain for its adoption, revealing a schism between educators and professionals:

The ability to find and make sense of information is almost the definition of newsgathering, so it seems safe to call this an essential skill for the beginning journalist. We asked professionals and educators to rate the importance of two key aspects of newsgathering that require this ability. Both the ability to analyze and synthesize large amounts of data and the ability to interpret statistical data were rated as more important by educators than by professionals.

When it comes to the ability to analyze and synthesize large amounts of data, a little more than half (55 percent) of the professionals responded that this was important to very important. Almost three-fourths (73 percent) of the educators rated this skill as important to very important.

The response to the question about the ability to "interpret statistical data and graphics" was similar: 59 percent of professionals and 80 percent of educators called this skill important to very important.

Given the large amounts of data available on the Internet and the growing importance of presenting information in a pleasing and informative visual manner, the gap between educators and professionals is disturbing. The ability to make sense of our complex world by distilling meaningful information from the vast river of data is one of the great values professional journalists can offer their audience.

The third report, on innovation at the *New York Times*,^{ccxxviii} was prepared by a team inside the newspaper for an internal audience, not public consumption. After the document leaked online in May of 2014 to Buzzfeed and Mashable, however, it was hailed by Joshua Benton, the director of Harvard's Nieman Journalism Lab, as "one of the key documents of this media age."^{ccxxix}

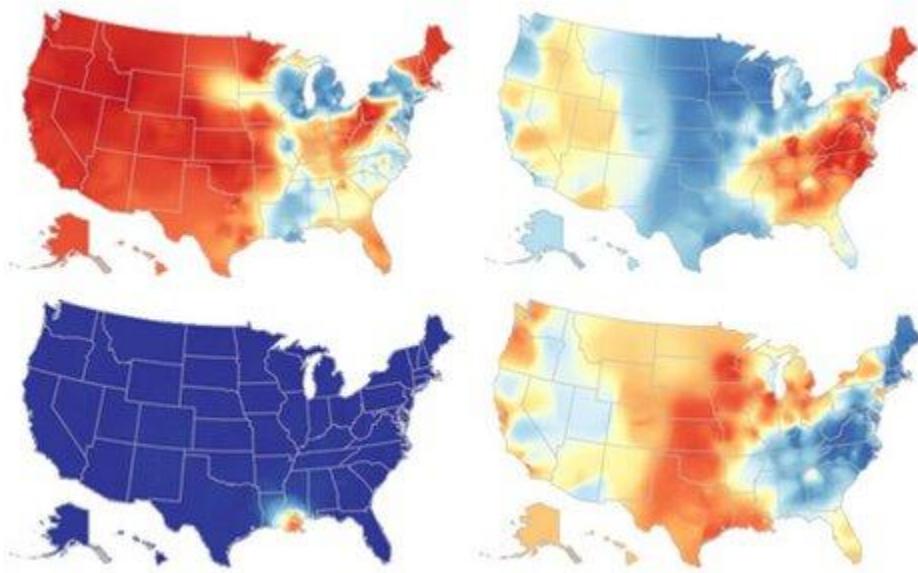
There's a tremendous amount of insight and introspection in the 97-page report, which surveyed the media landscape of today in depth, drawing on interviews with dozens of staff at the *New York Times* and dozens more with outside observers, including this author. I spoke with a researcher from the *Times'* team last year about the paper's approach to digital journalism, editorial analytics, social media and data, along with my own reading, sharing, and commenting habits.

The report paints a picture of an extraordinary organization housed within an institution and business grappling with the same fundamental shifts that broader society is enduring in the 21st century, struggling at times to escape a 20th century legacy of tools, infrastructure, and culture. Even though the digital audience of the *New York Times* is larger than its print readership (31

million unique visitors a month to nytimes.com versus 1.6 million total daily circulation), the daily editorial workflow described remains focused on the paper, not the pixel.

The report described the routine of a newsroom focused upon Page One and an incentive structure in which reporters are measured against their A1 stories. Instead of going “digital first” over the last decade, the publisher and leadership have continued to focus on the print edition. As the report notes, the paper currently derives three quarters of its revenues from print. That focus, however, cites a failure to convert the 14.7 million articles in the *Times*’ archive into structured data. Not doing so has meant that the newspaper is not capitalizing on one of its primary assets by making it more discoverable through search, sharing through social media, and data mining.

There are many reasons to think that “The Gray Lady” could become much more than she used to be in the years ahead. The first redesign of nytimes.com in eight years went live in January of 2014, optimized for mobile devices and integrating native advertising. The parent company was profitable in the first quarter of 2014. In March of 2014, the [Times expanded its digital offerings](#)^{ccxxx} to include [NYT Now](#), a lower-priced mobile app sold to iPhone users that summarizes the day’s top stories, and [Premier](#), which offers expanded access to behind-the-scenes stories, ebooks, videos, and crosswords. The *Times* may also explore events, a lucrative concern for other media companies. As noted earlier, The Upshot launched in April of 2014, to general acclaim.



The Upshot’s team includes the graduate student in statistics who helped to build the [news quiz on dialect](#) while he was an intern at the *Times*.^{ccxxxii} This interactive went on to be most read and shared content in the history of nytimes.com. In May of 2014, the [Times launched a lovely closed beta](#)^{ccxxxiii} of a new [cooking](#) Web application with more than 16,000 recipes. If the outlet can build a personalized recipe recommendation engine on top of its decades of dining and cooking archives, the platform could have tremendous potential.

The new executive editor of the *New York Times*, Dean Baquet, endorsed the report and the

digital-first strategy contained in it, both internally and publicly, once it leaked online. Whether he and his colleagues can execute against its recommendations remains to be seen.

The conclusions of these three reports, however, should still be sobering. The *Times* may be fine, but other papers will not be. Newsrooms face tight budgets, deep set cultural challenges, liabilities and debt, and historic lows in public trust. On the positive side, there is a tremendous upside for adoption and use of current tools and vast green fields for digitally native media organizations to experiment, create, and find audiences, as billions of people come online for the first time globally.

So what should we watch for next, and where? The following list of recommendations and predictions sketch out what to expect in the next decade and where publishers will need to adjust.

1) Data will become even more of a strategic resource for media.

If text is the next frontier in data journalism,^{ccxxxiii} it should be used in the service of telling stories more effectively, enabling digital journalism and digital humanities to merge in the service of a more informed society.^{ccxxxiv} Modern media organizations should become sources for trusted data.^{ccxxxv} Increasing amounts of data will be hosted by media organizations and leveraged as an asset. In some cases, media companies may be able to sell access to their archives and APIs.

Given the sensitivity of some data sets and the responsibility news organizations hold to confidential sources and whistleblowers, the media will need to improve its security practices. Recent widespread hacking incidents at major newspapers around the United States highlight the need for improvement.^{ccxxxvi}

2) Better tools will emerge that democratize data skills.

Even though the resources to learn data journalism are improving daily, there's still a high barrier to entry for people with no experience practicing it. That's changing as more powerful resources come online. Many of these tools for creating or presenting data-driven journalism will come from startups or nonprofits, like CartoDB, DocumentCloud, Timeline.js, Mapbox, Frontline SMS, Zeega, Kimono, Enigma.io, Amara, Plot.ly, DataWrapper, and Graf.ly.

Other tools will be provided by technology giants, like Google, Amazon, and Esri, as free Web services and open source code, or with enterprise licensing and API fees. Uncertainty about sustainability will drive foundations to fund tools and platforms, including pilot projects, entrepreneurial ventures, or components of open source civic infrastructure. The rest of the tools will be built by independent news hackers, university students, and data journalists as passion projects aimed at scratching someone's itch; these may well end up helping many other people solve similar problems as well.

Just as publishing text and editing photography or videos became accessible to hundreds of millions of people, analyzing and presenting data in maps, apps, and visualizations will become easier to do as well.

3) News apps will explode as a primary way for people to consume data journalism.

There have been hundreds of millions of iPhones, iPads, and Android devices sold in recent years, with billions of lower cost devices to follow as more of humanity goes online on mobile broadband networks. According to the [Pew Internet and Life Project](#), 42 percent of American adults over 18 years old owned a tablet in January of 2014.^{ccxxxvii}

Designing narrative stories, videos, and news applications for the growing number of readers using smartphones, phablets (a new class of mobile phones designed to straddle the functionality of phone and tablet), tablets, and laptops will only become more important to media organizations. That puts a premium on data journalists who can create apps, lightweight data visualizations, and story presentations that are optimized for mobile devices. Increasing demand for apps, quizzes, and interactive games will make news application developer a highly sought-after specialty at media companies.

Despite the growth in news apps, the narrative story format will endure as a complement to the news app, the summary for a blog, and access to the underlying data and model.

4) Being digital first means being data-centric and mobile-friendly.

As more and more people access the Internet and consume media on mobile devices, adopting a data-centric approach to collecting and publishing journalism will only grow in importance. The need to flexibly deliver content to multiple platforms and formats means that applications' programming interfaces that can supply data to any platform will continue to be a smart investment for organizations, particularly if they seek to be digital first. The *Washington Post*, NPR, and the *New York Times* have already moved in this direction. Others will follow, or lead.

Media companies will be competing for attention, and advertising and subscription dollars, with technology giants like [Google](#), Yahoo, Facebook, and startups that publish or curate user-generated content, along with vast amounts of data underpinning information services like mapping, shopping, or search. Facebook's Paper app, Google Play, Yahoo's News Digest, Narrative Science, Flipboard, and the automated information services yet to be created will be strong competition for media companies in the future.

5) Expect more robo-journalism, but know that human relationships and storytelling still matter.

We will see wunderkinds apply [computational journalism](#) to finding secrets and creating knowledge at vast scale, just as data scientists do in Silicon Valley, quants do on Wall Street, or spooks do at the National Security Agency.

"Robo-journalism" for commodity news from services like [Narrative Science](#) is already in the wild and will grow in use, particularly for areas that might have been previously uncovered by a beat reporter or for which a full time journalist is no longer economically viable. Wearable computers, drones, sensors, and algorithms are going to play a bigger role in the gathering of

data and consumption of media.

Despite changes in technology, humans will still matter in building relationships and making data into stories relatable to people. While the platforms and toolkits for journalism are evolving and the sources of data are expanding rapidly, many things haven't changed. The ethics that have long guided the choices of the profession remain central to the journalists working today, as NPR's new ethics guide makes clear.^{ccxxxviii} "Governments and others are going to learn how to hide their actions from open data," said Stray. "Personal relationships and skepticism will continue to be extremely important."

6) More journalists will need to study the social sciences and statistics.

"Philosophically, I think data journalism shares something with social science and also there's a real connection with the digital humanities," said Jonathan Stray, who teaches the subject at Columbia. "The emphasis is not just algorithms, but what do these algorithms tell us? How should we interpret all this fancy output?"

These questions have been integral to how sociologists, anthropologists, and ethnographers have conducted research for decades, particularly with respect to data collection and statistics. This means that if members of the media seek to practice data journalism, they'll need to be numerate, ethical, and thoughtful about the biases embedded in the data they're interrogating.

This is not a new idea, given how deeply Philip Meyer's "precision journalism" is grounded in applying social science to investigative reporting, but everyone who wishes to practice and publish sound data journalism is going to need to understand it.

Social scientists and biologists alike know that the sources for data and conditions under which it is collected will shape and bias any subsequent research conclusions made from it. To serve broad audiences, data journalists have to go beyond acquiring and cleaning data to understanding its provenance and source. Then, they'll need to make sure that its presentation doesn't tell a different story than the data itself allows.

None of that is easy for people trained as scientists, much less journalists. Some projects and analyses may exceed technical competence or subject matter expertise of select members of the media. Collaborating with academia and technologists will be preferable to flawed journalism, analyses, maps or visualizations that mislead readers, given the impact that inaccurate conclusions would have upon trust in the authors or publications.

7) There will be higher standards for accuracy and corrections.

Getting a fact wrong or screwing up a quote can sink a news story, leading to a correction or even retraction. Making a mistake in an algorithm or interpretation of data can similarly undermine the entire premise of an act of data journalism.

The mistakes and errors made in a post at FiveThirtyEight.com that sought to map kidnappings in Nigeria offer an instructive case study.^{ccxxxix} The post relied upon data sourced from the

Global Database of Events, Language and Tone (GDELT). As the correction to the story acknowledged, the journalism that was published was fundamentally flawed because the journalist failed to see that the data represented the rate of media stories as a proxy for the rate of kidnappings, did not account for duplicated reports, and used a default location if none was given. Decontextualizing the GDELT data led to a flawed post.^{ccxli}

Data journalism will attract numerate readers who are not only interested in the data behind stories and the analysis used to arrive at conclusions, but with the interest to try to reproduce them.

For instance, a FiveThirtyEight story on the Bechdel test in movies earned in-depth scrutiny from Brendan Keegan, who was able to replicate the findings. What that means in practice is that any media company that publishes such work should have a corrections policy in place for data journalism.^{ccxlii}

Andrew Whitby, an economist at Nesta, upon encountering examples of bad data journalism,^{ccxliii} proposed four principles of his own to improve upon the form:

1. **Choose the right stories:** In cases like this, a well-written review of the scholarly literature is likely to better inform public debate. Otherwise, stick to (a) lightweight but fun topics or (b) fast-moving topics yet to attract academic attention.
2. **Embrace complexity:** No interesting causal relationship involves only two variables.
3. **Use statistics intelligently:** A scatterplot of two variables with a least-squares regression line is not “doing statistics.” Bad statistics is worse than no statistics.
4. **Finally, be modest:** If you have so many caveats as to completely undermine any conclusion, then don’t offer a conclusion.

8) Competency in security and data protection will become more important.

In the United States, email hosted on private sector servers outside of a media company’s control does not have the same legal protections as email within an office. Until the Electronic Communications Privacy Act is reformed, journalists should be cautious about hosting sensitive email or data on other platforms.

People practicing data journalism or civic hacking need to know about the Computer Fraud and Abuse Act (CFAA),^{ccxliii} along with proposals for its reform.^{ccxliv} Journalists or members of the public who are unsure of the legality of data access or use, and don’t have the legal resources of major media organizations behind them, should think twice or thrice before clicking.

In general, journalists must consider when it’s appropriate to scrape data, access data, store it—or not. Does the story require storing personal information? If so, such sensitive data will need to be protected with the same vigor that journalists have protected confidential sources. Unfortunately, the information security practices of many media companies are not as robust as they will need to be to prevent determined intrusions by organized crime or nation states.

For more on data security, ethics, privacy, and journalism, consult the Tow Center’s white paper

on the subjects.

9) Audiences will demand more transparency on reader data collection and use.

Automated, personalized advertising or native advertising will be part of some living stories and news apps. The creators of these platforms it will have to carefully consider the context for matching ads with content.

Editorial and business departments are going to run up against difficult conversations about data access and sharing, with respect to audience analytics. Nonprofit organizations may not rely on advertising, instead taking underwriting or sponsorships, but they too will face pressure from funders and foundations to quantify their audiences and the impact of their journalism with data.

As editors, reporters, and publishers learn more about who is reading, sharing, and commenting on journalism through gathering data, they'll have to decide how transparent they'll be with readers about data collection and usage.

10) Conflicts over public records, data scraping, and ethics will surely arise.

For good or ill, we're likely to see more controversial online maps and interactive apps that show donations, votes, contributions, permits, convictions, and other public records. Along with voluntary disclosures, the data will be scraped, FOIA'ed or otherwise sourced from government publications, agencies, and websites.

Over time, much more of this data will end up in private hands, along with media, nonprofits, foundations, snarky online media outlets, and hacker collectives like Anonymous. Some of the resulting maps and charts will no doubt be found to be incorrect, made so by incompetence or malicious intent, resulting in misidentified people who will be subject to harassment or worse.

In turn, governments will try to deny access to data, heavily redacted documents, demand takedowns, and criminalize scraping or API calls. They will apply filtering, or extra-legal censorship through pressure on payment processors, seize servers or even direct denial of service attacks. Companies may deny access to their platforms for apps or services that use controversial data, similar to when Apple rejected an app showing drones strikes,^{ccxlv} or even accuse reporters of being hackers if they find data breaches or unprotected data online.^{ccxlvii}

Around the world, the conflict in societies with more closed governments and constricted information flows is likely to be explosive. Open data is not enough:^{ccxlviii} Investigative journalism will remain essential.

In the United States we'll run into more difficult First and Fourth Amendment issues as a result of all of this. It's going to be extremely messy. The chilling effects of mass surveillance on digital journalism will continue to be an issue for years to come. Just as sources may not trust the idea of a private conversation with a reporter, the provenance of data may be difficult to mask.

As a public comment^{ccxlviii} to the Review Group on Intelligence and Communication Technologies convened by President Barack Obama from Columbia Journalism School and the

MIT Center for Civic Media highlighted, mass surveillance makes investigative journalism much harder:

Put plainly, what the NSA is doing is incompatible with the existing law and policy protecting the confidentiality of journalist-source communications. This is not merely an incompatibility in spirit, but a series of specific and serious discrepancies between the activities of the intelligence community and existing law, policy, and practice in the rest of the government. Further, the climate of secrecy around mass surveillance activities is itself actively harmful to journalism, as sources cannot know when they might be monitored, or how intercepted information might be used against them.

11) Collaborate with libraries and universities as archives, hosts, and educators.

The government shutdown in the United States in the fall of 2013 demonstrated the need for media organizations and civil society to back up government data. At the time, many nonprofits, foundations, and individuals acted to preserve and mirror what they could. Around the rest of the globe, data sources may be even more tenuous.

In the years to come, journalists, universities, tech companies, businesses, and local governments will share a messy ecosystem of APIs, public, and private databases. There's already an emerging geocommons around OpenStreetMap, supported by rapidly improving open source tools and an emerging geojournalism speciality.

One strategy that may be fruitful is for city, county, and state governments to engage local media, universities, and libraries in public or civic data hosting and preservation.^{ccxlix} Librarians and academics have always been stewards of knowledge, in the forms of books and periodicals. As such, they and their institutions are well placed to host data for the public good, although legislators and executives will have to think through the economics of them doing so.

12) Expect data-driven personalization and predictive news in wearable interfaces..

In 2013, the most popular online content at the BBC was an economic class calculator. User-centric apps and services will enable people to understand how a given story or policy applies to them, their children, or their business. These kinds of news apps and data-driven platforms like Homicide Watch hint at what lies ahead. The current state of the art only scratches the surface of the ways that data will be personalized for individual readers as the use of analytics grows in media companies, helping editors get smarter.

As people express their interests through searches, clicks, saves, and shares, algorithms will use the data generated to suggest related editorial content and match advertising algorithms for relevant businesses or services with it. Recommendation engines will improve, across media companies, and be followed by predictive news that using social network analysis to suggest stories to users.

Over the next decade, a new wave of mobile computing will provide new platforms for nimble media companies to publish stories, from iWatches, to Google Glass, to smart appliances and wearable interfaces connected to an Internet of Things.

Some of these wearables won't just display data: They'll collect it. Such will include health data, geolocation, and air quality, which can then be used in citizen science and monitoring projects. They'll be part of a rich fabric of connected devices that, when combined with people, cellphones, and civic media, will enable [citizens to monitor infrastructure^{cccl}](#) or [water quality in China](#), extending into [networked civil society](#). The data generated from them will be rich source material for journalists to investigate and share.

Drones and sensors are both part of this picture and represent rich topics for more experimentation and inquiry, as explored by my colleague Fergus Pitt in his own research and workshops at the Tow Center.

13) More diverse newsrooms will produce better data journalism.

Diversity has been a challenge in the media for decades. Although far more minorities and women work in professional journalism than a century ago, a 2013 survey of [American Society of News Editors](#) (ASNE) found that of the 38,000 journalists currently working at 1,400 U.S. newspapers, 4,700 are minorities.^{cccli} A 2013 [ASNE survey of 68 online news organizations](#) found that 63 percent of them had no minorities at all.^{ccclii} The hiring choices at Vox Media, FiveThirtyEight, First Look Media, and other news startups garnered [criticism](#) in the spring of 2014,^{cccliii} including an [open letter](#) from the National Association of Black Journalists expressing concern regarding the lack of diversity.^{cccliv}

Diversity concerns are particularly relevant in the data journalism space, given the broader issues with women in technology that have become evident in recent years. Online and off, misogyny and discrimination endure in the industry, along with subtler sexism and racism. The challenge that editors face in hiring a diverse team of data journalists is structural, reflecting broader societal issues. As of 2010, 18 percent of undergraduates receiving degrees in computer science were women, [according](#) to the National Center for Women & Information Technology.^{ccclv} In 2013, just [0.4 percent of all female college freshmen said they intended to major in computer science](#).^{ccclvi} Given that context, perhaps it shouldn't have come as a surprise when Nate Silver said that 85 percent of the applicants to FiveThirtyEight were men.

There are reasons, however, to be cautiously optimistic about diversity in data journalism: Interviews with women and minorities in the United States suggest that the communities that have grown up around computer-assisted reporting over the decades may be more accepting of different faces than others in the technology world, perhaps because of the culture focused on peer-to-peer learning that celebrates mentorship.

"NICAR is a pretty healthy place to be a non-white, non-male person working in journalism," said Tasneem Raja. "I can't speak to issues of class, ability, gender identity, and other types of difference, other than to say we're almost definitely less good at them, and that needs to change." She went on:

I don't have experience with the way folks in this community handle issues of inclusion issues when they come up, but I have seen evidence of folks working preemptively to create environments that are less exclusionary than the norm in Web development, quantitative analysis, the visual arts, or journalism. Maybe it's because there haven't been that many of us webby data

journos till recently. Data journalists are pragmatic by nature, and maybe it just didn't make sense to alienate potential swaths of new recruits.

That's not to say everything is rainbows and sunshine, but I'm gonna take a rare moment of optimism here and say that I'm proud to represent this community, because in my experience, it's genuinely committed to inclusion.

No matter the country in which a media company operates, making an effort to include more women; minorities; gay, lesbian, bisexual, and transgender individuals; and people from multiple socioeconomic backgrounds will improve the work product and work environment. A diverse staff diminishes stereotypes and produces second-order reflection on unconscious biases, which in turn can lead to improved, more equitable evaluation of work, performance, promotion, and compensation. The absence of women, minorities, or GLBT persons in startups, media organizations, development teams, and in editorial or product leadership positions can signal to others that they aren't welcome.

Recruiting and hiring differently pays off: Media organizations that have diverse staffs are likely to produce better journalism, from story choice to source selection. Research suggests that teams with both men and women on them are more profitable and innovative. According to the National Center for Women and Information Technology, mixed gender teams produced information technology patents that are cited 26 percent to 35 percent more often than the norm.

As the demographics of the United States shifts, stories and data that focus upon minorities, women, and the GLBT community will also gain more audience share, which in turn will create a business opportunity for media companies. That's true around the globe as well. Given the opportunity, women and minorities have produced world-class data journalism. The world needs more of them, along with anyone else who wants to treat data as a source.

14) Be mindful of data-ism and bad data. Embrace skepticism.

Journalism will survive the death or diminishment of its institutions, as the Tow Center's report on post-industrial journalism explored.^{cclvii} In the decades ahead, media that integrate technology, data, and narrative skills into their work will play critical roles in societies around the world, from holding the powerful accountable to connecting people with information.

As people struggle to make sense of what matters or is true in a tsunami of new media, data journalism will be held up as a way to provide trustworthy insights to debunk pseudoscience, propaganda, misinformation, and online rumor. Just as yellow journalism, penny papers, and tabloids created a market opportunity that led to the creation of a more rigorous, ostensibly objective brand of journalism at the *New York Times* 160 years ago, today's fast-moving, chaotic media environment creates opportunities to publish data journalism as a corrective to punditry.

There are rocks and stormy waters ahead here, however, created by bad data journalism. The early 21st century has seen the growth of "data-ism,"^{cclviii} where knowledge can be derived through analysis of huge amounts of data now generated by various sources.^{cclix} This belief has antecedents in variants of positivism, the philosophy of science that holds information derived from logical (algorithmic) and mathematical analysis of data and sensory experience is the source of authoritative knowledge; and scientism, the belief that the scientific method can be

applied universally. All have a critical weakness: Bad data, biased data, and flawed experiments can and will be used ignorantly or cynically to twist the truth, mislead, or misinform, even by journalists who wish to do the opposite. Even good data and solid research may be misrepresented or mistaken, a risk that will grow if journalists are pushed to create data visualizations or analyses without training in information design, statistics, and social science.

Data has led many numbers-driven executives astray, in business, government, media, or academia.^{ccclx} The antidote to this malady is for journalists to interrogate data just as they would human sources, checking facts and assumptions, comparing results, and documenting the process and results of their investigations as a social scientist or biologist would. Complemented by human wisdom and intuition, data journalism still won't save the world or news, but it will help us all understand it better.

Acknowledgments

Alexander B. Howard is a writer and editor based in Washington, D.C. From August 2013 to May 2014, he was a fellow at the [Tow Center for Digital Journalism](#) at Columbia University. He is a [columnist at TechRepublic](#); the founder of “E Pluribus Unum,” a blog focused on open government and technology; and a [contributor to TechPresident](#), among other publications.

In 2013, Howard was a fellow at the [Networked Transparency Policy Project](#) in the [Ash Center for Democratic Governance and Innovation](#) at the Kennedy School of Government at Harvard University. Previously, he was the [Washington correspondent](#) for *Radar* at O'Reilly Media, where he chronicled the emergence of open data and open government movements around the world.

Howard has been recognized by *Washingtonian Magazine* as one of Washington's “[TechTitans](#),” a “respected trend-spotter and chronicler of government's use of new media.” He has appeared on air as an analyst for Al Jazeera English, WHYY, NPR, *Washington Post* TV, and a guest on *The Kojo Nnamdi Show* multiple times. Howard is a member of the government of [Canada's independent advisory panel on open government](#).

Prior to joining O'Reilly, he was the associate editor of SearchCompliance.com and [WhatIs.com](#) at TechTarget, where he wrote about how the laws and regulations that affect information technology are changing, spanning the issues of online identity, data protection, risk management, electronic privacy and IT security, and the broader topics of online culture and enterprise technology.

Howard has also contributed to the [National Journal](#), [The Daily Beast](#), [NextGov](#), [Forbes](#), [Buzzfeed](#), [Slate](#), [The Atlantic](#), [Huffington Post](#), [Govfresh](#), [ReadWriteWeb](#), [Mashable](#), [TechPresident](#), CBS News' What's Trending, Govloop, Governing People, and the Association for Computer Manufacturing, amongst others.

Howard has been a keynote speaker, moderator, and panelist at numerous conferences in Washington and beyond, including the Web 2.0 Summit and Expo, Gov 2.0 Summit and Expo, Social Media Week, DC Week, SXSWi, Strata, GOSCON, AMP Summit, National Democratic Institute, Tech@State, CAR/IRE, the State of the Net, and the Open Government Partnership's annual conference, among others. In 2011, he was Visiting Faculty at the Poynter Institute.

He also delivered remarks and/or moderated discussions at Harvard University, Stanford University, Columbia University, New York Law School, Alfred University, the Mona School of Business at the University of The West Indies, the American Association for the Advancement of Science (AAAS), the U.S. National Archives, NIST, the Club de Madrid, the Cato Institute, the New America Foundation, the World Bank, the U.S. Department of State, and the U.S. Social Security Administration.

Howard, a graduate of Colby College in Waterville, ME, lives in the District of Columbia with his wife, young daughter, old greyhound, and a growing collection of pots and cast iron pans.

About the Author

When I started formally compiling this report in November of 2012, I asked for feedback on data journalism. I immediately began hearing back from people around the world, with responses that continued right up until the day the final draft was submitted for its printing. The research herein also rests upon reporting and conversations with hundreds of editors, professors, reporters, technologists, government officials, and “hacker journalists” stretching back to 2010.

In interview after interview, I found an interest not simply in learning what data was out there, but how to get it and put it to use, from finding stories and sources, to providing empirical evidence to back up other reporting, to telling stories with maps and visualizations, to creating the data itself using sensors and social media. I also encountered healthy amounts of skepticism, optimism, and everything in between.

I am deeply grateful for the time of these pioneers and humbled by their work, dedication, and demonstrated interest in sharing knowledge with me, my networks, and their colleagues. In particular, I give thanks to my colleagues and the staff at the Tow Center, including Emily Bell, Lauren Mack, and Shiwani Neupane, for their feedback, mentorship, editing, and support; and to Mac Slocum, at O'Reilly Media, for all of the above. Huge thanks to Abigail Ronck for her sharp-eyed copyediting of a long, unwieldy manuscript. I am also much obliged to Brian Boyer, Scott Klein, Alberto Cairo, Nikki Usher, Nick Diakopoulos, Jonathan Stray, Susan McGregor, and Taylor Owen for their fantastic feedback on earlier versions of the report.

Except where otherwise noted, this report has been sourced from email correspondence, phone calls, conferences, Skype or in-person interviews. Portions of the report, although they have been edited and adapted, were first published at the O'Reilly Radar or the Tow Center's blog.

Alexander B. Howard, May 2014

Interviews

Data journalism interviews: Profiles of Practitioners

INDEX:

1. Cheryl Phillips
 2. Angelia "Momi" Peralta Ramos
 3. Tasneem Raja
 4. Sisi Wei
 5. Chase Davis
 6. Dan Hill
 7. Jeremy Bowers
 8. Aron Pilhofer
 9. Scott Klein
 10. Radar profiles
-

Cheryl Phillips

“Stories can be told in many different ways,” said Cheryl Phillips. “A sidebar that may once have been a 12-inch text piece is now a timeline, or a map.”

Phillips, an award-winning investigative journalist, will begin teaching students how to treat data as a source this fall, when she begins a new gig as a [lecturer at Stanford’s graduate school of journalism](#) helping to open up Stanford’s new Computational Journalism Lab.

“Cheryl Phillips brings an outstanding mix of experience in data journalism and investigative work to our program. Students and faculty here are eager to start working with her to push forward the evolving field of computational journalism,” said Jay Hamilton, Hearst Professor of Communication and Director of the Stanford Journalism Program, in a statement. “Her emphasis on accountability reporting and interest in using data to lower the costs of discovering stories will help our journalism students learn how to uncover stories that currently go untold in public affairs reporting.”

I interviewed Phillips about her career, which has included [important reporting on the nonprofit and philanthropy world](#), her plans for teaching at Stanford, data journalism, j-schools and teaching digital skills, and the challenges that newsrooms face today and in the future.

What is a day in your life like now?

I’m the data innovation editor at The Seattle Times. Essentially, I work with data for stories and help coordinate data-related efforts, such as working with reporters, graphics folks, and others on news apps and visualizations. I also have looked at some of our systems and processes and suggested new, more time-effective methods for us.

I've been at The Seattle Times since 2002. I started as a data-focused reporter on the investigations team, then became deputy investigations editor, then data enterprise editor. I also worked on the metro desk and edited a team of reporters. I currently work in the digital/online department, but really work across all the departments. I also helped train the newsroom when we moved to a new content management system about a year or so ago. I am trying to wrap up a couple of story-related projects, and do some data journalism newsroom training before I start at Stanford in the fall.

How did you get started in data journalism? Did you earn any special degrees or certificates?

I remember taking a class (outside of the journalism department) while in college. The subject purported to be about learning how personal computers worked but, aside from a textbook that showed photos of a personal computer, we really just learned how to write if, then loops on a mainframe.

I got my first taste of data journalism at the Fort Worth Star-Telegram. That's where I did my first story using any kind of computer for something other than putting words on a screen. I had gotten the ownership agreement for the Texas Rangers, which included a somewhat complex formula. I kept doing the math on my calculator and screwed it up each time. Finally, I called up a friend of mine who was a CPA, and she taught me Lotus 1-2-3.

My real start in computer-assisted reporting came in 1995, when I was on loan to USA TODAY. I was fortunate enough to land in the enterprise department with the data editors, and Phil Meyer was there a consultant. By the end of five months, I could use spreadsheets, Paradox (for DOS!) and SPSS. What a great education. I followed that up by joining IRE and attending the NICAR conference. I've missed very few since then and also done some of NICAR's specialized training on stats and maps.

I have no special degrees or certificates, but I have taken some online courses in R, Python, etc.

Did you have any mentors? Who? What were the most important resources they shared with you?

Phil Meyer is amazing, and such a great teacher. He taught me statistics, but also taught me about how to think about data. Sara Cohen and Aron Pilhofer of the New York Times, and Jennifer LaFleur of CIR. Paul Overberg at USA TODAY. They have all helped me over the years.

NICAR is an incredible world, full of data journalists and journalist-programmers who are willing to help others out. It's a great family.

On the investigative journalism front, Jim Neff and David Boardman are fantastic editors and great at asking vital questions.

What does your personal data journalism “stack” look like? What tools could you not live without?

I'm a firm believer in the power of the spreadsheet. So much of what journalists do on a daily basis can be made easier and more effective by just using a spreadsheet.

I use OpenRefine, CometDocs, Tabula, AP Overview and Document Cloud. I use MySQL with Navicat. I still use Access. I'm a recent convert to R, but also use SPSS. I use ESRI for mapping, but am interested in exploring other options also. I use Google Fusion Tables as well.

Most of my work has been in more of the traditional CAR front, but I've been learning Python for scraping projects.

What are the foundational skills that someone needs to practice data journalism?

In many ways, the same foundational skills you need for any kind of journalism.

Curiosity, for one. Journalists need to **think about stories from a mindset that includes data from the very beginning**, such as when a reporter talks to a source, or a government official. If an official mentions statistics, don't just ask for a summary report, but ask for the underlying data — and for that same data over time. The editors of those reporters need to do the same thing. Think about the possibilities if you had more information and could analyze and view it in different ways.

Second, **be open to learning any skill sets that will help tell the story**. I got into data journalism because I discovered stories I would not be able to tell if I didn't obtain and analyze data. We all know journalists don't like to take someone's word for something — data journalism just takes that to the next level.

Third, in terms of technical skills, **learn how to use a spreadsheet**, at a bare minimum. Really, one tool leads to another. Once you know how a spreadsheet works, you are more open to using OpenRefine to clean and standardize that data, or learning a language for scraping data, or another program that will help with finding connections.

What classes will you be teaching at Stanford, and how?

I will be teaching several courses, including a data journalism class focusing on relational data, basic statistics and mapping. I also will be teaching an investigative reporting class focusing on investigative reporting tools.

In general, I want to make sure the students are telling stories from data that they analyze. They should be not only learning the technical stack, but how to apply the technical knowledge to real-world journalism. I am hoping to create some partnerships with newsrooms as well.

Where do you turn to keep your skills updated or learn new things?

IRE and NICAR and all the folks involved there. I also try to learn from our producers at The Seattle Times, who come in knowing way more than I did when I started in journalism. I try to follow smart people on Twitter and other social media.

I like to reach out to folks about what they are doing. I think reaching out and connecting with folks outside of journalism is a great way to make sure we are aware of other new tools, developments, etc.

What are the biggest challenges that newsrooms face in meeting the demand for people with these skills and experience?

Newsrooms are often still structured into silos, so reporters just report and write. They may hand their data off to a graphics desk, but they don't necessarily analyze or visualize data themselves. Producers produce, but don't write, even though they may enjoy that and be good at it, too.

Some of this is by necessity, but it makes it harder to learn new skills — and some of these skills are really useful. A reporter who knows how to visualize data may also be able to look at it in a different way for reporting the story out too. So, building collaborative teams is important, as is providing time for folks to try out other skills.

Are journalism schools training people properly? What will you do differently?

I think it's no secret that a lot of change is starting to take place in schools.

Cindy Royal had an interesting piece about[platforms](#) just the other day. In general, I think my answer here is similar to the biggest challenge for newsrooms: We need to take a more integrated approach. Classrooms and their teachers should collaborate on work.

So, for example, a multimedia class produces the visualizations and videos that go with the stories being written in another class. (Yes, Stanford already does this.)

Data journalism should not be just one class out of a curriculum, but infused throughout a curriculum. Every type of journalist can learn data-related skills that will help them, whether they end up as a copy editor, a reporter, a front-line editor or a graphics artist.

What data journalism project are you the most proud of working on or creating?

I have been asked this question before and can never answer it well. My last story is always the one I'm most proud of, unless it's the one I'm about to publish.

That said, as an editor at The Seattle Times, I worked with Jennifer LaFleur (then with ProPublica) on a project tracking the reasons behind [foreclosures](#), a deep dive into the driving factors behind foreclosures from several cities.

When I was a reporter, I was lucky enough to get to work with Ken Armstrong on our [court secrecy project](#) in 2006, which changed state practice. I also led the reporting effort on problems with [airport security](#). Both of those used small data sets, which we built ourselves, but told important stories.

I can think of even more stories that weren't data projects per se, but which used data in the reporting in critical ways. The recent [Oso mudslide coverage](#) is an example of where we used mapping data and landslide data to effectively tell the story of the impact of the slide on the victims and of how the potential disastrous consequences had been ignored over time.

What data journalism project created by someone else do you most admire?

Too many to count. There has been so much great work done. ProPublica's [Dollars for Docs](#) was fantastic not only for its stories, but the way they shared the data and the way newsrooms from across the country could tap into the work. Last year, the Milwaukee Journal Sentinel's project, [Deadly Delays](#), was such important work.

How has the environment for doing this kind of work changed in the past five years?

It's much more integrated into new immersive storytelling platforms. There is a recognition that stories can be told in many different ways. A sidebar that may once have been a 12-inch text piece is now a timeline, or a map.

I think there are many more team collaborations, with the developers, designers and reporters and CAR specialists working together from the outset. We need a lot more of this.

What's different about practicing data journalism today, versus 10 years ago? What about teaching it?

There are more tools, with more coming every day. A few are great, and a lot aspire to be great and some of those will probably get there.

The really fantastic thing about the change is that it's relatively easy to contribute to the development of a tool that will help journalism, even just as a beta tester.

There are more tech folk interested in helping make journalism better. We're becoming a less insular world, and that's a good thing.

Why are data journalism and “news apps” important, in the context of the contemporary digital environment for information?

News apps help tell important stories. It's the same reason narrative is important.

It always should boil down to that: “**does this tool, language, or app help tell a story?**” If the answer is “yes,” and you think the story could be worth the effort, then the tool is important too.

What's the one thing people always get wrong when they talk about data journalism?

I think I'll have to punt on this one. As [you have pointed out](#), [data journalism](#) is a [big umbrella term](#) for many different things — [precision journalism](#), [computer-assisted reporting](#), [computational journalism](#), [news apps](#), etc. — so it's easy to have a different idea as to what it means.

Momi Peralta

Long before data journalism entered the mainstream discourse, La Nacion was pushing the boundaries of what was possible in Argentina, a [country without an freedom of information law](#). If you look back into La Nacion's efforts to go online and start to treat data as a source, you'll find [Angélica “Momi” Peralta Ramos\(@momiperalta\)](#), the multimedia development manager who originally launched LaNacion.com in the 1990s and now manages its [data journalism efforts](#).

Ramos contends that data-driven [innovation is an antidote](#) to budget crises in newsrooms. Her perspective is grounded in experience: Peralta's team at [La Nacion is using data journalism to challenge a FOIA-free culture in Argentina, opening up data](#) for reporting and reuse to holding government accountable. This spring, I interviewed her about her work and perspective. Her answers follow, lightly edited for clarity.

You're a computer scientist and MBA. How did you end up in journalism?

Years ago, I fell in love with the concept of the Internet. It is the synthesis of what I'd studied: information technology applied to communications. Now, with the opportunity of data journalism, I think there is a new convergence: the extraction and sharing of knowledge through collaboration using technology. I'm curious about everything and love to discover things.

How did your technical and business perspective inform how you approached LaNacion.com and La Nacion Data?

In terms of organization, it helped to consider traditional business areas like sales, marketing, customer service, business intelligence, and of course technology and a newsroom for content.

At first, I believed in the unlimited possibilities of technology applied to publishing online, and the power of the net to distribute content. Content was free to access and gratuity became the norm. As consumers embraced it, there was a demand and a market, and when there is a market there are business opportunities, although with a much more fragmented competitive environment.

The same model applies now to data journalism. Building content from data or data platforms must evolve to an economy of scale in which the cost of producing [huge amounts of] content in one single effort tends to zero.

What examples of data-driven journalism should the public know about at La Nacion?

Linked below is a selection of 2013 projects. Some of them are finalists in the 2014 Data Journalism Awards! Please watch the videos inside the posts, as we explained how we manage to extract, transform, build and open data in every case.

- [Argentina's official advertising funds distribution 2009 – 2013: Friends, politicians and a stylist.](#)
- [Public officials' salaries and assets for reporting and accountability](#)
- [Monitoring the new media law in Argentina 2009-2013](#). 94% created [during this time] were state media. No plural voices, as promised, so far.
- [VozData](#): collaborating to free data from PDFs – The Senate Expenses Part II.
- 2013: [Legislative Elections in Argentina](#)
- [Argentina's Senate Expenses 2004-2013](#)

This won the GEN-Google DataJ awards in investigative reporting in 2013!!

How you see digital publishing, the Internet and data journalism in South America or globally? What about your peers?

I can't tell about everyone else's view, but I think we see it all the same, as both a big challenge and opportunity.

From then on, it's a matter of being willing to do things. The technology is there, the talent is everywhere, the people who make a difference are the ones you have to gather.

As the context is different in every country and there are obstacles, you have to become a problem solver and be creative, but never stop. For example, if there are language barriers, translate. If there is no open data, start by doing it yourself. If technology is expensive, check first for free versions. Most are enough to do everything you need.

What are the most common tools applied to data journalism at La Nacion?

Collaborative tools. Google Docs, spreadsheets, Open Refine, Junar's open data platform, Tableau Public for interactive graphs, and now Javascript or D3.js for reusable interactive graphs tied to updated datasets. We love tools that don't need a developer every time to create interactive content. These are end user's tools.

Developers are the best for "build once, use many times" kinds of content, developing tools, news applications and for creative problem solving.

What are the basic tools and foundational skills that data journalists need?

First, searching. Using advanced search techniques, in countries like ours, you find there is more on the Deep Web than in the surface.

Then scraping, converting data from PDFs, structuring datasets, and analyzing data. Then, learning to publish in open data formats.

Last, but not least: socializing and sharing your work.

Data journalists need a tolerance for frustration and ability to reinvent and self motivate. Embrace technology. Don't be afraid to experiment with tools, and learn to ask for help: teamwork is fun.

How do you and your staff keep your skills updated and learn?

We self-teach for free, thanks to the net. We look at best practices and inspiration from other's cases, then whenever, we can, we for assistance at conferences as NICAR, ISOJ or ONA and follow them online. If there are local trainings, we assist. We went to introductory two-day courses for ArcGIS and Qlikview (business intelligence software) just to learn the possibilities of these technologies.

We taught ourselves Tableau. An interactive designer and myself took two days off in a Starbucks with the training videos. Then she, learned more in an advanced course.

We love webinars and MOOCs, like the Knight Center's or the EJR's [data journalism MOOC](#).

We design internal trainings. We have a data journalism training program, now starting our 4th edition, with five days of full-time learning for groups of journalists and designers in our newsroom. We also design Excel courses for analyzing and designing data sets (DIY Data!) and, thanks to our Knight-Mozilla OpenNews fellows, we have customized workshops like CartoDB and introductions to D3.js.

We go to hackathons and meetups — nearly every meetup in Buenos Aires. We interact with experts and with journalists and learn a lot there, working in teams.

What are the biggest challenges La Nacion faces in practicing data journalism? What's changed since 2011, in terms of the environment?

The context. To take just one example, consider the inflation scandal in Argentina. Even [The Economist removed](#) our [national] figures from their indicators page. Media that reported private indicators were considered as opposition by the government, which took away most of official advertising from these media, fined private consultants who calculate consumer price indices different than the official, pressed private associations of consumers to stop measuring price and releasing price indexes, and so on.

Regarding official advertising, between 2009 and 2013, we managed to build a dataset. We found out that 50% went to 10 media groups, the ones closer to the government. In the last period, a hairdresser (stylist) received more advertising money than the largest newspapers in Argentina. Here's how we built and analyzed[this dataset](#).

Last year, independent media suffered an ad ban, as reported in The Wall Street Journal:
[“Argentina imposes ad ban, businesses said.”](#)

Argentina is ranked 106 / 177 in Transparency International Corruption Perceptions Index. We still are without a Freedom of Information law.

Regarding open data from governments, there are some initiatives. One that is more advanced is the City of Buenos Aires Open Data portal, but also there are national, some provincial and municipal initiatives starting to publish useful information, and even open data.

Perhaps the best change is that we have is a big hacktivism community of transparency activists, NGOs, journalists and academic experts that are ready to share knowledge for data problem solving as needed or in hackathons.

Our dream is for everyone to understand data as a public service, not only to enhance accountability but to enhance our quality of life.

What's different about your work today, versus 1995, when [LaNacion.com](#) went online?

In 1995, we were alone. Everything was new and hard to sell. There was a small audience. Producing content was static, still in two dimensions, perhaps including a picture in .jpg form, and feedback came through e-mail.

Now there is a huge audience, a crowded competitive environment, and things move faster than ever in terms of formats, technologies, businesses and creative uses by audiences. Every day, there are challenges and opportunities to engage where audiences are, and give them something different or useful to remember us and come back.

Why are data journalism and news apps important?

Both move public information closer to the people and literally put data in citizens' hands. News apps are great to tell stories, and localize your data, but we need more efforts to humanize data and explain data. [We should] make datasets famous, put them in the center of a conversation of experts first, and in the general public afterwards.

If we report on data, and we open data while reporting, then others can reuse and build another layer of knowledge on top of it. There are risks, if you have the traditional business mindset, but in an open world there is more to win than to lose by opening up.

This is not only a data revolution. It is an open innovation revolution around knowledge. Media must help open data, especially in countries with difficult access to information.

How do Freedom of Information laws relate to data journalism?

FOI laws are vital for journalism, but more vital for citizens in general, for the justice system, for politicians, businesses or investors to make decisions. Anyone can republish information, if she can get it, but there are requests of information with no response at all.

What about open government in general? How does the open data movement relate to data journalism?

The open government movement is happening. We must be ready to receive and process open data, and then tell all the stories hidden in datasets that now may seem raw or distant.

To begin with, it would be useful to have data on open contracts, statements of assets and salaries of public officials, ways to follow the money and compare, so people can help monitor government accountability. Although we dream in open data formats, we love PDFs against receiving print copies.

The open data movement and hacktivism can accelerate the application of technology to ingest large sets of documents, complex documents or large volumes of structured data. This will accelerate and help journalism extract and tell better stories, but also bring tons of information to the light, so everyone can see, process and keep governments accountable.

The way to go for us now is use data for journalism but then open that data. We are building blocks of knowledge and, at the same time, putting this data closer to the people, the experts and the ones who can do better work than ourselves to extract another story or detect spots of corruption.

It makes lots of sense for us to make the effort of typing, building datasets, cleaning, converting and sharing data in open formats, even organizing our own 'datafest' to expose data to experts.

Open data will help in the fight against corruption. That is a real need, as here corruption is killing people.

Tasneem Raja

Tasneem Raja, the [interactive editor at Mother Jones Magazine](#), is one of the growing number of journalists who aren't just reporting upon the news but building the medium for the message to be communicated. Before she joined Mother Jones, was the news apps editor at [The Bay Citizen](#), where her team built a [Bike Accident Tracker](#) and a [government salary database](#), among other things, and a feature writer at [The Chicago Reader](#). Raja's insights into [how to build an interactive news team](#) (more on that below) are well worth reading. You can follow her work on [Github](#) or her commentary on [Twitter](#). Our interview follows, lightly edited and linked for context.

Where do you work now? What is a day in your life like?

I'm a senior editor at Mother Jones magazine, where I lead an awesome team of data reporters and interactive producers. I'm also a writer and reporter, in print and on the Web.

We live by a few guiding principles on my team. The big one is that it's our job to make sure everybody in the newsroom can tell a story by any means necessary. That is, reporters should know how to map, the mapmaking pros on my team should know how to factcheck, the fact checkers should know when to use a column chart versus a bar chart, and so on. We don't believe in siloed skills.

Of course, some folks will always be way better at some skills than others, but you gotta pay it forward, which brings me to our second guiding principle: we are all learners, and we are all teachers.

Put it all together, and you get quite the three-ring circus of hybrid journalism going on here everyday. Today, for instance, I'm finishing up edits on a big magazine feature story about the future of programming, while teaching Illustrator charting to a reporter with good data on air pollution. Producer [Jaeah Lee](#) is teaching a reporter best practices in structured data for an easy-to-update map of gay marriage laws. Our interactive fellow [AJ Vicens](#) just fired off a quick blog post about racism in sports, and is now working with a reporter on abstracting an open-source template we made. We probably look pretty different than most data teams in this way.

How did you get started in data journalism? Did you get any special degrees or certificates?

I was a staff writer at the Chicago Reader in the mid-2000s, which was, of course, a scary time to be in news. When a bunch of my senior mentors there, all writers, got canned in 2007, I decided to re-evaluate my career and went to j-school at Berkeley to learn new skills. I was lucky enough to be there while Josh Williams was teaching web development (he left for the NYT, where he worked on Snowfall and tons of other big interactive pieces), and essentially attached myself at the hip. It turned into a year-long independent study, and got me a job on the launch team at The Bay Citizen, where I created a news apps team that made some really cool data projects for the Bay Area (RIP, TBC).

What quantitative skills did you start with?

I've always appreciated structured ways of looking at information. There's something about well-formatted tables of information and clean spreadsheets that makes me really happy. That's the most important skill for a data journalism, in my opinion: a love of working with structured data, and creating whole new systems and worlds atop it. That strange love is what makes you want to put the time in to learn R, command line tools, pivot tables, and so on – all stuff I didn't pick up 'til halfway through my first job in data journalism.

Did you have any mentors? Who? What were the most important resources they shared with you?

“Mentor” is a funny word. Here are a few people who've left deep thumbprints on the way I think about my work (whether they intended to or not).

[Josh Williams](#) taught me everything I needed to know to get a really good first job in data journalism and news apps: what's a text editor, what's the command line, what's a Web framework. More than that, he got me thinking in terms of abstraction. For instance, he was always saying, “Never build something you can only use once.” Instead, think both in terms of the specific needs of the project in front of you, and the broader needs of a similar project you might not even know of till next year. Seeing the way he held both of these concepts in his head at once was an incredible lesson in how to be a journalist who is also a pretty decent project manager.

[Brian Boyer](#) taught me the importance of having a [guiding philosophy](#) (or three) to your work. The *why* of what you do, not just the how. And that your philosophies can sound more like something a chef or a potter would say, than a data nerd. In other words, he got me thinking about this work as craft.

[Scott Klein](#) has inspired me to better know my shit. That is, it's not enough to read a few [blog posts about data journalism](#) and crown yourself the next [Edward Tufte](#). There's [a lot of history](#) to what we do, a lot of important choices to be made, and fortunately, there are very old and very new books out there to learn from. You can't have a conversation with Scott without wanting to go pick up a book.

What does your personal data journalism “stack” look like? What tools could you not live without?

1. A good, simple text editor, with good syntax highlighting
2. A spreadsheet app, with version control and collaborative editing
3. GitHub
4. Google.com
5. The cognitive ability to think in terms of abstraction

Where should people who want to learn start?

A hundred people have said it [before me, and better](#): pick a project you genuinely want to do, and then hack, Google, and plead for help in forums, and read books, until you get it working.

Where do you turn to keep your skills updated or learn new things?

1. The [NICAR conference](#)
2. Twitter
3. Increasingly, printed books
4. Dissecting the work of colleagues at other shops

What are the biggest challenges that newsrooms face in meeting the demand for people with these skills and experience? Are schools training people properly?

There are a ton of challenges, so I'll pick one: **we don't have a pair programming model on the editorial side of the newsroom, and we need one.**

Journalism schools still teach journalism as a very hierarchical, often solitary pursuit. That's not the way it works in data journalism, and the best learning is still gonna be on the job. That requires cross-pollination between folks with different skill sets. We need a pairing model across newsrooms, not just in the nerd corner.

I've had several people tell me they're surprised to learn how small my team is, given the daily volume of content we put out. That's because we're not the only ones who can work with data and visuals in our newsroom. We've spent serious time pairing with something like 1 in 3 staffers here, working and training side by side whenever physically possible. (We have offices on different coasts). We've gotten several editors and reporters on Github, and while we don't have them checking in code through the command line (yet), they're well-versed in the how and why of version control.

What's the one thing people always get wrong when they talk about data journalism?

That it's all quantitative. That's like guiding principle #3 on my team: everything is data. Words are data. Gifs are data. If it can be sorted, tracked, counted, merged, filtered: it's probably data. I'd say half the project my team does is more qualitative than quantitative. That is, most people wouldn't consider it data visualization so much as photo essays, games, quizzes, etc. There's a lot of power in developing a data skill set — both technical and cognitive — that lets you make cool things with words and pictures, too.

You're written about Silicon Valley's "[programmer problem](#)." How do data journalists and their communities of practice handle issues of race, sexism or gender?

NICAR is a pretty healthy place to be a non-white, non-male person working in journalism. I can't speak to issues of class, ability, gender identity, and other types of difference, other than to say we're almost definitely less good at them, and that needs to change.

I don't have experience with the way folks in this community handle issues of inclusion issues when they come up, but I have seen evidence of folks working to preemptively to create environments that are more less exclusionary than the norm in web development, quantitative analysis, the visual arts, or journalism. Maybe it's because there haven't been that many of us webby data journos till recently. Data journalists are pragmatic by nature, and maybe it just didn't make sense to alienate potential swaths of new recruits.

That's not to say everything is rainbows and sunshine, but I'm gonna take a rare moment of optimism here and say that I'm proud to represent this community, because in my experience, it's genuinely committed to inclusion.

Sisi Wei

The best antidote to a decade of discussion about the “future of news” is to talk to the [young journalists who](#) are [building](#) it. Sisi Wei’s [award-winning](#) journalism shows exactly what that looks like, in practice. Just browse her [projects](#) or [code repositories](#) on Github.

Wei, a [news applications developer](#) at [ProPublica](#), was formerly a graphics editor at The [Washington Post](#). She is also the co-founder of “[Code with me](#),” a programming workshop for journalists. Our interview about her work and her view of the industry follows.

Where do you work now? What is a day in your life like?

I currently work at [ProPublica](#), on the [News Applications Team](#). We make interactive graphics and news apps; think of projects like [3D flood maps](#) and [Dollars for Docs](#).

At ProPublica, no one has a specific responsibility like design, backend development, data analysis, etc. Instead, people on the team tend to do the whole stack from beginning to end. When we need help, or don't understand something, we ask our teammates. And of course, we're constantly working alongside reporters and editors outside of the team as well. When someone's app is deploying soon, we all pitch in to help take things off his/her plate.

On a given day, I could be calling sources and doing interviews, searching for a specific dataset, cleaning data, making my own data, analyzing it, coming up with the best way to visualize it, or programming an interactive graphic or news app. And of course, I could also be buried beneath interview notes and writing an article.

How did you get started in data journalism? Did you get any special degrees or certificates? What quantitative skills did you have?

I got started in college when I began making interactive graphics for [North by Northwestern](#). I was a journalism/philosophy/legal studies major, so I can safely say that I had no special degrees or qualifications for data journalism.

The closest formal training I got was an “Introduction to Statistics” course my senior year, which I wish I’d taken earlier. I also had a solid math background for a non-major. The last college math course I took was on advanced linear algebra and multivariable calculus. Not that I’ve used either of those skills in my work just yet.

Did you have any mentors? Who? What were the most important resources they shared with you?

Too many to list. So, here’s just a sample of all the amazing people who I’ve been lucky to consider mentors in the past few years, and one of the many things they’ve all taught me.

[Tom Giratikanon](#) showed me that journalists could use programming to tell stories and exposed me to ActionScript and how programming works. [Kat Downs](#) taught me not to let the story be overshadowed by design or fancy interaction, and [Wilson Andrews](#) showed me how a pro handles making live interactive graphics for election night. [Todd Lindeman](#) taught me how to better visualize data and how to really take advantage of Adobe Illustrator. [Lakshmi Ketineni](#) and [Michelle Chen](#) honed my javascript and really taught me SQL and PHP.

Now at ProPublica, [my teammates](#) are my mentors. Here is where I learned Ruby on Rails, how news app development really works and how to handle large databases with first ActiveRecord and now ElasticSearch (which I am still working on learning).

What does your personal data journalism “stack” look like? What tools could you not live without?

- [Sublime Text](#), whose multiple selection feature is the trump card that makes it impossible for me to switch to anything else. If you haven’t used multiple selection, stop what you’re doing and go check it out.
- The Terminal, for deploying and using Git or just testing out small bits of code in Ruby or Python.
- Chrome, to debug my code.
- The Internet, for the answers to all of my questions.

What are the foundational skills that someone needs to practice data journalism?

An insatiable appetite to get to the bottom of something, and the willingness to learn any tool to help you find the answers you’re looking for. In that process, you’ll by necessity learn programming skills, or data analysis skills. Both are important, But without knowing what questions to ask, or what you’re trying to accomplish, neither of those skills will help you.

Where should people who want to learn start?

In terms of programming, just pick a project, make it simple, make it happen and then finish it. Like [Jennifer DeWalt](#) did when she made [180 websites in 180 days](#).

Regarding data analysis, if you're still in school, take more classes in statistics. If you're not in school, NICAR offers [CAR boot camps](#), or you can search for materials online, such as this book that [teaches statistics to programmers](#).

Where do you turn to keep your skills updated or learn new things?

I don't have a frequent cache of websites that I revisit to learn things. I simply figure out what I want to learn, or what problem I'm trying to solve, and use the Internet to find what I need to know.

For example, I'm currently trying to figure out which Javascript library or game engine can best enable me to [create newsgames](#). I started out knowing close to nothing about the subject. Ten minutes of searching later, I had detailed comparisons between game engines, demos and reviews of gaming Javascript libraries, as well as wonderful tips from indie game developers for any rookies looking to get started.

What are the biggest challenges that newsrooms face in meeting the demand for people with these skills and experience? Are schools training people properly?

There are two major pipelines for newsrooms to recruit people with these skills. The first is to recruit journalists who have programming and/or data analysis experience. The second is to recruit programmers or data analysts to come into journalism. The latter, I think, is much harder than the former, though the [Knight-Mozilla OpenNews Fellowship](#) is doing a great job of doing this.

Schools are getting better at teaching students data journalism skills, but not at a high enough rate. I often see open job positions, but I rarely see students or professionals with the right skills and experience unable to find a job.

The lack of students, however, is a problem that starts before college. When high school students are applying for journalism school, they expect to go into print or radio or TV news. They don't expect to learn how to code, or practice data analysis. I think one of the largest challenges is how to change this expectation at an earlier stage.

All of that said, I do have one wish that I would like journalism schools to fulfill: I wish that no j-school ever reinforces or finds acceptable, actively or passively, the stereotype that journalists are bad at math. All it takes is one professor who shrugs off a math error to add to this stereotype, to have the idea pass onto one of his or her students. Let's be clear: Journalists do not come with a math disability.

What data journalism project created by someone else do you most admire?

I actually want to highlight a project called [Vax](#), which was not built by journalists, but deploys the same principles as data journalism and has the same goals of educating the reader.

Vax is a game that teaches students both how epidemics spread, as well as prevention techniques. It was created originally to help students taking a [Coursera MOOC on Epidemics](#) really engage with the topic. I think it's accomplished that in spades. Not only are users hooked right from the beginning, the game allows you to experience for yourself how people are interconnected, and how those who refuse vaccinations affect the process.

How has the environment for doing this kind of work changed in the past five years?
Since I only entered the field three years ago in 2011, all I can say is this: Data journalism is gaining momentum.

Our techniques are becoming more sophisticated and we're learning from our mistakes. We're constantly improving, building new tools and making it easier and more accessible to do common tasks. I don't want to predict anything grand, but I think the environment is only going to get better.

Is data journalism the same thing as computer-assisted reporting or computational journalism? Why or why not?

To me, data journalism has become the umbrella term that includes anyone who works in data, journalism and programming. (And yes, executing functions in Excel or writing SQL queries is both data and programming.)

Why are data journalism and “news apps” important, in the context of the contemporary digital environment for information?

Philip Meyer, who wrote "[Precision Journalism](#)," answers the first part of this question with his entire book, which I would recommend any aspiring data journalist read immediately. He says:

"Read any of the popular journals of media criticism and you will find a long litany of repeated complaints about modern journalism. It misses important stories, is too dependent on press releases, is easily manipulated by politicians and special interests, and does not communicate what it does know in an effective manner. All of these complaints are justified. Their cause is not so much a lack of energy or talent or dedication to truth, as the critics sometimes imply, but a simple lag in the application of information science — a body of knowledge — to the daunting problems of reporting the news in a time of information overload."

Data journalism allows journalists to point to the raw data and ask questions, as well as question the very conclusions we are given. It allows us to use social science techniques to illuminate stories that might otherwise be hidden in plain sight.

News apps specifically allow users to search for what's most relevant to them in a large dataset, and give individual readers the power to discover how a large, national story relates to them. If the story is that doctors have been receiving payments from pharmaceutical companies, news apps [let you search to see if your doctor has](#) as well.

What's the one thing people always get wrong when they talk about data journalism?

That it's new, or just a phase the journalism industry is going through. Data journalism has been around since the 1970s (if not earlier), and it is not going to go away, because the skills involved are core to being a better journalist, and to making your story relatable to millions of users online.

Just imagine, if a source told you that $2+2=18$, would you believe that statement? The more likely scenario is that you'd question your source about why he or she would say something so blatantly wrong, because you know how to do math, and you know that $2+2=4$.

Analyzing raw data can result in a similar question to a source, except this time you can ask, "Why does your data say X, but you say Y?"

Isn't that a core skill every journalist should have?

Chase Davis

Chase Davis is an assistant editor on the Interactive News Desk at the New York Times and teaches an advanced data journalism class at Mizzou, where he helps transfer his skills and perspective (treat data a source). Our interview, lightly edited for clarity, content and [bracketed] and hyperlinked for context, follows.

What is a day in your life like?

I help supervise the developer/journalists who build many of our cool Web projects. I have a background as a reporter, primarily doing [investigations](#) and covering politics, so I try to dabble in that world as well. I also teach a class in advanced data journalism at the [Missouri School of Journalism](#) and do some consulting on the side.

How did you get started? Did you get any special degrees or certificates? Quantitative skills?

I got started in data journalism almost by accident. I started learning to program for fun in middle school, then fell in love with journalism and ended up at Mizzou. I lived a typical j-student life for a few years, writing a bunch for the student paper and doing internships, then applied (based on a total misunderstanding) to start working for NICAR. The couple years I spent there really tied those two skillsets together.

Did you have any mentors? Who? What were the most important resources they shared with you?

Too many to list, but I'll name a few. Jacquee Petchel, Lise Olsen and Mark Katches for schooling me in the ways of capital-J Journalism. Brant Houston and Jeff Porter for taking me in at NICAR and showing me how journalism and data can work together. And, really, the entire [IRE and NICAR community](#), which is outrageously giving of its collective time.

What does your personal data journalism “stack” look like? What tools could you not live without?

I’m pretty minimalist: a terminal window and some type of text editor. The only place I splurge is on a database GUI (I like Navicat). The one tool I couldn’t live without is Python, which is the best Swiss Army knife a data journalist can have.

What are the foundational skills that someone needs to practice data journalism?

The same core skills you need to practice any kind of journalism: curiosity, skepticism, an eye for detail and a sense of a good story. [They] also [need] numeracy, or at least conceptual mathematical literacy, which is still unfortunately too rare. Also important are databases and spreadsheets, statistics, and some kind of programming language — doesn’t matter which one. Being your own worst critic doesn’t hurt.

And intellectual courage. You need to be motivated, not intimidated, to learn new and difficult things.

Where do you turn to keep your skills updated or learn new things?

Personal projects. I always have at least one on the backburner, and I make sure it stretches me in a new direction. Working on something I care about is the best way for me to stay motivated. I get bored learning from books.

What are the biggest challenges that newsrooms face in meeting the demand for people with these skills and experience? Are schools training people properly?

The oversimplified explanation is that most journalism students can’t code or do math, while most computer science students don’t know storytelling. Hybrids on either side are rare, and we’re scooping them up as fast as we can.

Journalism schools could be doing more, but it’s not all their fault. It takes intellectual agility and natural curiosity to effectively develop hybrid skills. I don’t think that’s something we can teach solely through curriculum. That’s why I don’t think every journalism student should “learn how to code.” Being able to write a few lines of Javascript is great, but if you let your skills dead-end with that, you’re not going to be a great newsroom developer.

Folks on our interactive and graphics teams at the Times have remarkably diverse backgrounds: journalism and computer science, sure, but also cartography, art history, and no college degree at all. What makes them great is that they have an instinct to self-teach and explore.

That’s what journalism schools can encourage: introduce data journalism with the curriculum, then provide a venue for students to tinker and explore. Ideally, someone on faculty should know enough

to guide them. The school should show an interest in data journalism work on par with more traditional storytelling.

Oh, and they should require more math classes.

What data journalism project are you the most proud of working on or creating?

Hard question, but I'll offer up pretty much anything that my old team at the Center for Investigative Reporting has done. That was my first turn at being a boss, and the fact that they haven't all been fired suggests that I didn't mess them up too bad.

What data journalism project created by someone else do you most admire?

Look at the [Philip Meyer Awards](#) every year and you pretty much have that answer. Anyone who can take a spreadsheet full of rows and columns, or a bunch of code, and turn it into something that changes (or starts) the conversation about an important topic is the whole reason many of us got into this game in the first place.

How has the environment for doing this kind of work changed in the past five years?

It's night and day. Five years ago, this kind of thing was still seen in a lot of places at best as a curiosity, and at worst as something threatening or frivolous. Some newsrooms got it, but most data journalists I knew still had to beg, borrow and steal for simple things like access to servers.

Solid programming practices were unheard of — version control? What's that? If newsroom developers today saw Matt Waite's code when he first launched [PolitiFact](#), their faces would melt like "Raiders of the Lost Ark."

Now, our team at the Times runs dozens of servers. Being able to code is table stakes. Reporters are talking about machine-frickin'-learning, and newsroom devs are inventing pieces of software that power huge chunks of the web. The game done changed.

What's different about practicing data journalism today, versus 10 years ago?

It was actually 10 years ago that I first got into data journalism, which makes me feel old even though I'm not. Back then, data journalism was mostly seen as doing analyses for stories. Great stories, for sure, but interactives and data visualizations were more rare.

Now, data journalism is much more of a Big Tent speciality. Data journalists report and write, craft interactives and visualizations, develop storytelling platforms, run predictive models, build open source software, and much, much more. The pace has really picked up, which is why self-teaching is so important.

Is data journalism the same thing as computer-assisted reporting or computational journalism? Why or why not?

I don't think the semantics are important. Journalism is journalism. It should be defined on its own merits, not by the tools we use to accomplish it. Treating these things as exotic specialties makes it too easy to pigeonhole the people who practice them. And I hate that.

What's the one thing people always get wrong when they talk about data journalism?
That data journalists are unicorns.

Or wizards. Or that they can somehow pull swords from stones in a way that mere laypeople can't. That kind of attitude is dangerous — not because it mythologizes tech skills, or demonstrates willful ignorance on the part of technophobes (both of which are sad), but because it drives a cultural wedge between data journalists and the rest of the newsroom.

[Imagine hearing] "I'm a conventional reporter, so my specialty is reporting. You're a tech person, so you write code."

I think that's crap. I know plenty of reporters who can code, and plenty of data journalists who can report the hell out of a good story. By dividing them culturally, we almost let people see the "journalist" in "data journalist" as secondary. We turn them into specialists, rather than letting them bring journalism and technology together in new and creative ways.

Why are data journalism and "news apps" important, in the context of the contemporary digital environment for information?

Numeracy is important. A more universal appreciation of technology in our industry is important. A culture of rapid, constant experimentation is important. To the extent that data journalism has encouraged those things in newsrooms, I think it's been hugely important.

The actual product of data journalism — news apps, visualizations, stories — those will all continue to evolve, but data journalism's continuing contribution to newsroom culture is something that I hope is permanent.

Dan Hill

Our interview follows, lightly edited for clarity, content and hyperlinked for context.

Where do you work now? What is a day in your life like?

I [joined The Texas Tribune](#) as a full-time news apps developer in January. Our team is responsible for both [larger-scale "explorer"](#) apps and what I'd call "[daily interactives](#)." My day often involves writing and processing public information requests, designing interactives and working on Django apps, depending on the scale of my project.

How did you get started in data journalism? Did you get any special degrees or certificates?

What quantitative skills did you start with?

I've always wanted to be a reporter, but the work of [Phillip Reese](#) at The Sacramento Bee and [The Chicago Tribune's news apps team](#) inspired me to enhance my storytelling with data. I was a student fellow for the [Northwestern University Knight Lab](#) and studied journalism and computer science, but an internship with The Washington Post taught me how to apply what I was learning in a newsroom.

Did you have any mentors? Who? What were the most important resources they shared with you?

I've had awesome mentors. [Bobby Calvan](#) and [Josh Freedom du Lac](#) were the first to treat me like a real reporter. [Jon Marshall](#) helped me explore my interests. Phillip Reese showed me how to find untold stories in spreadsheets and Brian Boyer encouraged me to learn Python. [Serdar Tumgoren](#) and [Jeremy Bowers](#) showed me how a team of news developers operates. [Travis Swicegood](#) taught me how to [deal with real world data](#).

My mentors remind me to always be learning and asking questions.

What does your personal data journalism “stack” look like? What tools could you not live without?

I use Excel, OpenOffice, GoogleDocs, [Django](#) and [iPython notebooks](#) for data analysis. [R](#) is creeping into my workflow for exploring datasets and experimenting with visualizations. We use [d3](#) and [chartjs](#) for web graphics and [Mapbox](#) for web maps. I could probably survive without [Backbone](#), but we use it a lot.

What are the foundational skills that someone needs to practice data journalism?

I think a data journalist needs news judgment and attention to detail in order to identify the newsworthiness and limitations of datasets.

Statistics can help explain a dataset's strengths and weaknesses, so I wish I paid more attention during my stats classes in school.

In addition to finding the stories, data journalists also need to be able to explain why data is significant to their audience, so visual journalists need design skills — and, of course, reporting and writing.

Where do you turn to keep your skills updated or learn new things?

I check [Source](#), the [Northwestern Knight Lab blog](#) and the [NICAR](#) listserv for new ideas. Lately, I've been teaching myself statistics and R with [r-tutor](#) and [Machine Learning for Hackers](#).

What are the biggest challenges that newsrooms face in meeting the demand for people with these skills and experience? Are schools training people properly?

I think the differences between the developer and newsroom cultures make it hard for newsrooms to find people with tech and journalism skills, and to coordinate projects with developers and reporters. As a student in journalism school, I was inspired to learn more about data when professor [Darrell Little](#) showed how it could enhance my reporting and help me find stories hidden in datasets.

I learned more developer-journalist skills like database management and web design from meetups, tutorials and classes outside the j-school, but the journalism school exposed me to what journalists with those skills could do.

I'll add I'm impressed with data literacy of the Texas Tribune newsroom, where reporters request spreadsheets and use data to verify claims on their beats. Even if reporters don't have the programming chops to make an interactive graphic, for example, they're great about identifying potential data stories.

What data journalism project are you the most proud of working on or creating?

My summer intern project at The Washington Post, a [study of every Washington D.C. homicide case between 2000 and 2011](#), was my first experience making news app in a newsroom. I was honored get to work with the investigative reporters as a newbie intern and learned a ton from building the database and doing analysis with Serdar. All of my contributions were on the backend, but I was thrilled to work with that dataset as an intern.

What data journalism project created by someone else do you most admire?

Propublica's [Message Machine](#) was my favorite project from the 2012 presidential election, because it took a unique approach to identify trends in email metadata. I'm excited for more stories that collect everyday metadata or use [sensors](#) to [explore the data](#) around us.

How has the environment for doing this kind of work changed in the past five years?

I'd never heard of a "news apps team" five years ago. I knew I wanted to be an investigative reporter but never thought I would write code every day. I admired reporters like Phillip Reese who were working with data and making interactive graphics, but I didn't see as many teams of specialized developer-journalists.

What's different about practicing data journalism today, versus 10 years ago?

I wasn't even a teenager 10 years ago, but I would gander... THE INTERNET. Online data portals, open government and open Web stuff are important to the data journalism I do. I'm not sure they were as common a decade ago.

Is data journalism the same thing as computer-assisted reporting or computational journalism? Why or why not?

I think of "data journalism" as an umbrella term that refers the use of data in reporting or presentation, whereas I think of CAR and computational journalism as subsets of data journalism that involve analyzing a dataset.

Why are data journalism and "news apps" important, in the context of the contemporary digital environment for information?

I'm excited to work with data because of its widespread use in decision making. I think news apps can help people understand meaningful data and uphold accountability for people who create and make decisions with data.

[Be A Newsnerd](#) has better answers

What's the one thing people always get wrong when they talk about data journalism?

Although the web plays a big role in the growth of data journalism, I don't think you need to be online to do data journalism.

Jeremy Bowers

The following interview is with [Jeremy Bowers](#), a news application developer at NPR. (He also [knows a lot about how the Internet works](#).) It has been lightly edited for clarity, content and hyperlinked for context.

Where do you work now? What is a day in your life like?

I work on the [Visuals team](#) at NPR.

Our team adheres to modified scrum principles. We have a daily scrum at 10:00am for all of the members of our Visuals team. We work on projects in one-week sprints. At the end of each sprint we have an iteration review and a “ticket shuffle” where we decide what tickets each of us should work on in the next sprint. Our typical projects rarely exceed four sprint cycles.

Our projects involve minimally four people: One developer, one designer, one project manager and one stakeholder . Some projects add more designers or more developers as necessary. And sometimes we have a lot of stakeholders.

How did you get started in data journalism? Did you get any special degrees or certificates?

What quantitative skills did you start with?

I started in data journalism at the St. Petersburg Times. I'd been working as the blog administrator for our “online” team and was informally recruited by Matt Waite to help out with a project that would turn into MugShots.

I have no special degrees or certificates. I was a political science major and I had planned to go to law school before a mediocre LSAT performance made me rethink my priorities.

I did have a background in server administration and was really familiar with Linux because of a few semesters spent hacking with a good friend in college, so that's been pretty helpful.

Did you have any mentors? Who? What were the most important resources they shared with you?

Matt Waite from the St. Petersburg Times got me started in data journalism and has been my mentor for as long as I can remember. I don't call him as much anymore now that he's “Professor Waite,” but I still use a lot of our conversations as guidelines even now.

I also owe a debt to Derek Willis, though we've never worked together. I'm pretty much daily inspired by Ben Welsh and Ken Schwenke at the Los Angeles Times. They build apps that matter

and make me look critically at what I am building. Finally, my co-worker Chris Groskopf is a great source of inspiration about keeping my code and work habits clean and professional.

What does your personal data journalism “stack” look like? What tools could you not live without?

I live and die with three tools.

First, a terminal emulator. Right now, I’m using iTerm2 with a customized Solarized theme.

Second, a text editor. Right now, I’m using Sublime Text 3 with a ton of installed packages.

Finally, I need a Web browser. Right now, I’m using Chrome.

At NPR Visuals, we’ve [documented our stack](#) so that anyone can code like we do.

What are the foundational skills that someone needs to practice data journalism? Where do you turn to keep your skills updated or learn new things?

I think the foundational skill for data journalism is curiosity. Rabid, all-consuming curiosity has done more for my career than any particular skill or talent.

I also think that good data journalists are passionate about their projects. Unlike many tech or newsroom jobs, it’s difficult to punch a clock and work 9-to-5 as a data journalist. I’m constantly pulling out my laptop and hacking on something, even when it’s not directly tied to a work project.

What are the biggest challenges that newsrooms face in meeting the demand for people with these skills and experience? Are schools training people properly?

The easy answer would be to say that there aren’t enough data journalists to go around. But that’s not exactly true. With Knight and Google fellowships, recent college graduates, interns, and proto-news-nerds miscast in other roles, media companies are surrounded by possibilities. Our challenge, as I see it, is building an environment where hackers and the hacker ethic can thrive. And that’s a tough thing to do at any large company, let alone a media company. But we’ve got to make that our personal mission and not be confounded by what feels like an impersonal bureaucracy.

What data journalism project are you the most proud of working on or creating?

Without a doubt, PolitiFact is the most exciting project I’ve worked on. I also really enjoyed working on the Arrested Development app for NPR — so much so, that I binge-watched the fourth season and coded up the jokes over 24 hours the day the episodes were released!

What data journalism project created by someone else do you most admire?

I love everything about the Los Angeles Times's data desk. Their homicide and crime apps are nothing short of miraculous.

How has the environment for doing this kind of work changed in the past five years?

I released my first news app to the wild in April of 2009. At that time, there were only a handful of groups that I knew of writing code in newsrooms — the New York Times, the Washington Post, the Chicago Tribune, and Ben Welsh at the Los Angeles Times. About 300 people attended NICAR that year, and it was focused on print CAR reporting.

This year, NICAR hosted 997 people and had well-attended sessions on JavaScript and D3, tools that basically only work on the Web. There are probably 20 teams writing code in newsrooms today, and there are entire college classes dedicated to producing hacker journalists. So the environment has gotten much richer and larger in the last five years.

What's different about practicing data journalism today, versus 10 years ago?

I can't speak generally since I only started doing "data journalism" about five years ago. I hesitate to argue that what I'm currently doing is really data journalism as opposed to newsroom product development.

As not to cheat you out of an answer, I can say that my job now involves much more rigor. When I first started writing code at the St. Petersburg Times, we didn't use version control. We didn't sandbox our code. We didn't have automated deployment tools. These days, we have GitHub to store our code, tons of command-line tricks to keep our code in separate virtual environments, and we have fantastic deployment tools that make updating our code a snap.

Additionally, my organization is much more aware of what I'm doing. My manager and his managers are much more cognizant of data journalism generally and specifically about how our work fits in with the organization's strategy. When I started, we basically worked invisibly on products that almost nobody really knew or cared about.

Is data journalism the same thing as computer-assisted reporting or computational journalism? Why or why not?

I'm terrible at semantic differences. I'll take the broad view on this one: If you're writing code in a newsroom, you're probably committing acts of journalism. I don't feel terribly strongly about what we decide to call this or how we decide to slice up what an investigative journalist, a news librarian or a news apps developer might be doing every day. If they're writing code and making journalism, I want them to have every opportunity to succeed. I don't feel any need to give them labels or have their titles prevent them from writing code to get their jobs done.

Why are data journalism and “news apps” important, in the context of the contemporary digital environment for information?

I think that [Scott Klein said it best](#), and I’ll paraphrase: If you’re not using algorithmic or computational methods to analyze data, someone is scooping you on your beat.

There’s hardly a beat in journalism anymore that doesn’t involve structured data, which means that there’s hardly a journalist that wouldn’t benefit from automated methods for analyzing that data. Folks who are passing up that aspect of their jobs are just handing that opportunity over to someone else.

Print and radio are so time- and space-limited. If you’re not using the Web to tell all of the stories rather than just one story, you’re probably doing that wrong as well.

What’s the one thing people always get wrong when they talk about data journalism?

I don’t like it when people talk about how “organizations don’t get” data journalism, and I don’t like it for a very specific reason: The inability to create a news hacker culture doesn’t rest on the shoulders of some amorphous “organization.” We should place that blame where it belongs: Squarely on the shoulders of individuals in your newsroom.

What we’ve got is a people problem. Editors and other newsroom opinion leaders should be making an environment for their reporters or others to participate in hacker journalism.

The same ethics that Eric Raymond elucidated in the [Hacker How-To](#) should guide journalists in newsrooms everywhere:

- The world is full of fascinating problems waiting to be solved.
- No problem should ever have to be solved twice.
- Boredom and drudgery are evil.
- Freedom is good.
- Attitude is no substitute for competence.

To make your organization a place where hacker ethics are practiced requires positive action — it won’t just spring into being because of a memo. So, don’t blame your company because there’s no room to operate like a hacker. Instead, blame your boss or your boss’s boss. It’s most effective when you discuss this with them personally. But make sure you give those people an opportunity to correct their wrongs. Few people are actually hostile to the hacker ethic; most are just unfamiliar.

Aron Pilhofer

When it comes to computer-assisted reporting and the ways media companies are using technology, there are few people in the U.S.A. as knowledgeable as Aron Pilhofer. He runs a newsroom team at The New York Times that combines journalism, social media, technology and analytics, co-founded the open source [DocumentCloud.org](#) project, is a two-time grantee of the Knight News Challenge,

and co-founded [Hacks & Hackers](#), a network of people focused on applying development and digital innovation to the method and practice of journalism. In May 2014, Pilhofer left the New York Times to be the Guardian's first executive editor for digital.

He was kind enough to talk to me earlier this month. Our interview follows, lightly edited for clarity, content and [bracketed] for context.

How do you feel about the term “[data journalism](#)” supplanting computer-assisted reporting (CAR)?

When someone can tell me what is meant by “data journalism,” maybe I would start to feel strongly about it.

I think there’s a lack of specificity, with various definitions. It tends to be almost geographically based. In Europe, when you talk about data journalism, you’re almost always talking about data visualization. In the United States, it’s sometimes data visualization, sometimes old school computer-assisted reporting.

While I prefer the term data journalism, because it’s much less goofy [than computer-assisted reporting], I think there’s a lack of precision. You do need to define your terms. The way I see it, it is a continuum where the work that Phil Meyer, Barlett and Steele were doing 30 years ago [continues] all the way to today, with people like Sarah Cohen and John Keefe, all sharing kind of the same elements. You treat data as a source [in your reporting].

What’s happening with the market for data journalists and your ability to hire for these skills?
In some ways it’s easier. In others, it’s harder today. There’s way more competition now. We’re losing people to really good newsrooms. We are not the only game in town, which we used to be. There was a time when there was us and there was the Washington Post, and that was kind of it.

What are you working on now that’s new and potentially important?

We just started a newsroom analytics team. The kinds of projects we’re doing there are entirely editorial. They are not tied to advertising at all.

Right now, many newsrooms are stupid about the way they publish. They’re tied to a legacy model, which means that some of the most impactful journalism will be published online on Saturday afternoon, to go into print on Sunday. You could not pick a time when your audience is less engaged. It will sit on the homepage, and then sit overnight, and then on Sunday a home page editor will decide it’s been there too long or decide to freshen the page, and move it lower.

I feel strongly, and now there is a growing consensus, that we should make decisions like that based upon data. Who will the audience be for a particular piece of content? Who are they? What do they read? That will lead to a very different approach to being a publishing enterprise.

Knowing our target audience will dictate an entirely differently rollout strategy. We will go from a “publish” to a “launch.” It will also lead us in a direction that is inevitable, where we decouple the

legacy model from the digital. At what point do you decide that your digital audience is as important — or more important — than print?

This sounds similar to the approach that many online outlets are pursuing.

There's not a digital property on the planet that isn't doing this kind of thing — or a smart one, anyway. [Medium has its own metrics](#). [Upworthy has attention minutes](#) that reflect engagement.

[The Interactive News] team can build just about anything now to scale to a ridiculous amount of traffic, tying into every New York Times system. That isn't the problem anymore. We can make everything work [from a technical standpoint] on David Leonhardt's project, which is our answer to 538, but it still may not find an audience.

This is a product build, where we take a particular flavor of journalism and find an audience. We find a way for the audience that would want that to find it. It is really hard to think about when you really only know one tune: Your homepage. It is really powerful, but that alone isn't going to do it. How does that change what we're building? How can we consistently get that audience to return?

Building one-off interactives isn't that important. When you're starting to build persistent features, like what John Keefe has done with his [Cicada Project](#), or Scott Klein has built with [Dollars for Docs](#), you've got to think about these things more deeply. Who in the newsroom is better positioned than a data journalist to do that?

How many data journalists do you have on staff at the Times?

It depends on your definition; we could be anywhere from 5 to 50. We have a computer-assisted reporting team, which is 5-6 people. We have a graphics desk, which is probably 15 primarily or largely dedicated to digital. On my team we have 21 developers. Then there's our research and development department, and design team.

Is there anyone you'd call a computational journalist?

Maybe Chase Davis. Amanda Cox is a statistician by training. Sarah Cohen was a former statistician before she went into CAR. We have data scientists on the business side. R&D has a couple, like Mike Dewar, who used to be at Bitly. These are people who are applying data science techniques to actual journalism, stories, infographics and data visualizations.

Would you agree with an estimate of several hundred data journalists currently working in the USA?

Absolutely. NICAR has 850 people registered, with a healthy walkup expected. [The final attendance at NICAR 2014 was 997 people.] Five years ago, the conference was on life support, with maybe 250 people. Now, this number of people showing up has changed it a lot, I think for the better. It has become the “must-go” conference for folks who are doing what my team does, for the John Keefes & Scott Kleins of the world.

Is there a mismatch between the supply and demand for people with the skills you've referenced?

It's true. I have two openings now.

What was your path to the profession?

I was a political reporter, but always used data in my reporting. I just started doing it in college. I just started messing around. I had a history professor who was not well known then. Now, he's borderline famous from doing quantitative methods in history. He'd do statistical sampling of historical census data that had just been paper records before that. Suddenly, you could do queries on the 1930 Census. You were not just basing a historic analysis on papers or on interviews with people, or what you could glean from anecdotes. You were looking at data. It was incredible.

That's not that different from a data journalist does, on the CAR side. Instead of a person, you're using data as a source. Over time, I shifted from being a reporter who does CAR to being a specialist at the Center for Public Integrity to a CAR editor at the New York Times and then started this team.

How did you start learning to program?

I can thank an IRS story on 527 committees, which were then the campaign finance loophole du jour. They were previously unregulated and Congress, in its wisdom, put the IRS in charge of regulating them. It was idiotic. The IRS is not a disclosure agency. They put together the world's worst disclosure website. There was basic data there, but you couldn't aggregate it or access it in a meaningful way. It would have taken thousands of mouse clicks to get all of it.

I talked to a public information officer, after they denied my FOIA request for the database underlying the site. He said it was all on the website.

So, I created the world's worst Web scraper in PHP. It ran from the browser. I didn't know the command line well.

Is “Silent Partners” still on the Web?

Parts of it are long gone, though bits remain at the Center for Public Integrity. What you won’t find is the massive searchable database. We did what IRS should have done. We took all the paper filings and got a grant to do data entry. We sent them to a company in Virginia. We spent \$80,000 to create what was then the only searchable database of political donor contributions. It’s completely out of date now. The Center for Responsive Politics has been continuing to do this.

I discovered that I really enjoyed the coding part in addition to reporting. The art of it. That’s how I ended up shifting into my current job.

Have you seen more coders move into data journalism, or journalists learn how to code?

I’ve seen far, far more move the outside in, from non-journalism roles.

Do you have any sense of why that might be?

Journalism is one of the few professions that not only tolerates general innumeracy but celebrates it. I still hear journalists who are proud of it, even celebrating that they can’t do math, even though programming is about logic. It’s hard to get a journalist to open up a spreadsheet, much less open up a command line. It is just not something that they, in general, think is held to be an important skill.

It’s baffling to me. Look at The Sun-Sentinel, which just won another Pulitzer for a story on speeding cops that you could only do with data analysis. You would think you wouldn’t have to make the case that this is core to what journalists should know.

It’s a cultural problem. There is still far too much tolerance for anecdotal evidence as the foundation for news stories.

So this is endemic?

I don’t know how to solve it. Look at NICAR being around as long as it has. Early on, they had the naive belief that if you could train enough people, they could make the organization irrelevant. Now, when you look back, it’s hilarious. Obviously, that’s never going to happen for practical purposes. I don’t think we’re anywhere near the point where you could say, given enough training and time, that you would not need a specialist in the newsroom. We’re so far away from that.

Are there cultures where this is changing? Maybe ProPublica?

It’s as far along as it is because of Scott Klein. It took years before they put Jeff Larson on news stories. That just happened this year. There are newsrooms making this a significant project. Look at the L.A. Times, or WNYC. I think John Keefe is a fricking genius. I wish I were doing the work he is.

What others would you highlight?

Check out Nicolas Kayser-Bril [of [Journalism++](#)] and [De-Correspondent](#), out of the Netherlands.

Given time, given urgency, we will forge something new from the old models. Given how much time we have had, I would have hoped we'd be further along. Maybe I'm just impatient. When do you treat digital as your primary platform?

We are launching three subscription products this year. If all goes well, we will have more subscriptions on pure digital than in print [at the end of 2014]. We have to think about where the eyeballs are. From the perspective of the newsroom, over time, we have to think primarily digital. That's the cultural change that isn't happening fast enough.

There needs to be a strategy, where all the things we considered "nice to have" in a newsroom — from analytics to coders to designers — all of a sudden, they're building our core product. Text only takes you a certain distance in digital sphere. That's the part that I'm excited about building.

Scott Klein

[Scott Klein](#), an assistant managing editor at ProPublica, draws on years of hands-on experience working with data, reporters and developers at one of the most important nonprofit news organizations in the world. Recently, his team has published projects like [The Opportunity Gap](#), [China's Memory Hole](#) and [Prescriber Checkup](#). Klein co-founded [DocumentCloud](#) with Aron Pilhofer, the New York Times editor [whose perspectives on technology and the news](#) we featured here earlier this month. Before he came to ProPublica, Klein directed editorial and business application development for the [TheNation.com](#), and worked at *The New York Times*. This spring, he spoke with me about what he sees in the industry and have an early read and review of the report that the Tow Center will publish next month. Our conversation, lightly edited for content and clarity, follows.

Is data-driven journalism too expensive?

News organizations are contracting and budgets are going down. Times are still very tough. That said, I suspect that some newsrooms say they can't afford to hire newsroom developers when they really mean that their budget priorities lie elsewhere — priorities that are set by a senior leadership whose definition of journalism is pretty traditional and often excludes digital-native forms. I also hear a lot from people trying to get data teams started in their own newsrooms that the advice that newsroom leaders get is that newsroom developers are "[unicorns](#)" whom they can't afford. Big IT departments sometimes play a confounding role here.

I suspect many metro papers can actually afford one or two journalist/developers — and there's a ton of amazing projects a small team can do. For years, the Los Angeles Times ran one of the best news application shops in the country with only two dedicated staffers (they still do great work, of course, and the team has grown). If doing data journalism well is a priority of the organization, making it happen can fit into your budget.

What's changed today?

Lots, of course, has changed since Philip Meyer's pioneering days in the 1960s. One is that the amount of data available for us to work with has exploded. Part of this increase is because open government initiatives have caused a ton of great data to be released. Not just through portals like data.gov — getting big data sets via FOIA has become easier, even since ProPublica launched in 2008.

Another big change is that we've got the opportunity to present the data itself to readers — that is, not just summarized in a story but as data itself. In the early days of CAR, we gathered and analyzed information to support and guide a narrative story. Data was something to be summarized for the reader in the print story, with of course graphics and tables (some quite extensive), but the end goal was typically something recognizable as a words-and-pictures story.

What the Internet added is that it gave us the ability to show to people the actual data and let them look through it for themselves. It's now possible, through interaction design, to help people navigate their way through a data set just as, through good narrative writing, we've always been able to guide people through a complex story.

Is this new state of affairs really different?

It's a tectonic change both in the sense that it's slow and gradual, and in the sense that it's reshaping the entire landscape.

Data was always central to journalism. In the oldest newspapers, from the 17th century, you can find data. Correspondents would write about the prices of commodities in faraway cities (along with court gossip) for the benefit of merchants doing international business. Commodity prices, the contents of arriving cargo ships, and even the names of visiting businessmen were a big part of the daily mission of newspapers as they started to become more common.

As technology got better in the late 18th century and readers started demanding a different kind of information, the data that appeared in newspapers got more sophisticated and was used in new ways. Data became a tool for middle-class people to use to make decisions and not just as facts to deploy in an argument, or information useful to elite business people.

The change we're experiencing thanks to the web increases the role of presentation of the data itself, both in great data visualization and in great exploratory graphics like news applications. We can show people "the back of the baseball card" on a large scale. We've got the tools, and the readers can understand it and make use of it. I feel like that's as big a change as we've ever experienced, but I'm biased.

Do people want to read the data?

If it's done well, people have a really big appetite to see the data for themselves.

Look how many people understand — and love — incredibly sophisticated and arcane sports statistics. We ought to be able to trust our readers to understand data in other contexts too. If we've done our jobs right, most people should be able to go to our "Prescriber Checkup" news application,

search for their doctors and see how their prescribing patterns compare to their peers, and understand what's at play and what to do with the information they find.

There are ways to design data so that more important numbers are bigger and more prominent than less important details. People know to scroll down a Web page for more fine-grained details. At ProPublica, we design things to move readers through levels of abstraction from the most general, national case to the most local example.

Do you recruit or programmers to do DDJ? Or teach journalists?

Both. But culture matters a lot, too. People with the right mindset, who feel valued for their editorial judgment and creativity, and who are given real responsibility over their work, will learn whatever they need to learn in order to get a project done. The people on my team focus on telling great journalistic stories and don't let not knowing how to do something stop them from doing so. They learn whatever skills, techniques and expertise they need to learn.

In terms of journalists learning how to program, I think there are some myths about what "programming" means. It doesn't have to mean a computer science degree and it doesn't have to mean what Google does. I know journalists who make incredibly complex scrapers for their reporting work who will tell you they don't know how to program. Really, making tools to automate tasks is what a programmer does. There's no magic threshold you have to pass between programmer and not-programmer.

Of course, there is a difference between knowing how to code and being a computer scientist. If you've learned about algorithmic efficiency and can express it mathematically, and if you've studied how compilers work, all under the guidance of a person who knows the subject very well in an academic environment, you've got skills that will help you write better, faster, more efficient code. That's different than learning how to use a high-level programming language to get a task done.

Much of what we do in newsrooms is on deadline and meant to be put behind a caching system that makes efficient code much less important, so computer science is not a prerequisite for being a great newsroom coder. In newsrooms, most of us rely on frameworks like Rails or Django that already make great low-level programming decisions anyway.

Are there journalists picking those DDJ skills up?

Yes, it's happening, and the pace is accelerating. A few years ago the NICAR conference was a few hundred people. This year it was almost 1,000 people. Next year, it will be even bigger.

On every desk in the newsroom, reporters are starting to understand that if you don't know how to understand and manipulate data someone who can will be faster than you. Can you imagine a sports reporter who doesn't know what an on-base percentage is? Or doesn't know how to calculate it himself? You can now ask a version of that question for almost every beat.

There are more and more reporters who want to have their own data and to analyze it themselves. Take for example my colleague, Charlie Ornstein. In addition to being a Pulitzer Prize winner, he's one of the most sophisticated data reporters anywhere. He pores over new and insanely complex data

sets himself. He has hit the edge of Access's abilities and is switching to SQL Server. His being able to work and find stories inside data independently is hugely important for the work he does.

There will always be a place for great interviewers, or the eagle-eyed reporter who find an amazing story in a footnote on page 412 of a regulatory disclosure. But, here comes another kind of journalist who has data skills that will sustain whole new branches of reporting.

Radar interviews

To learn more about the people who are redefining the practice of computer-assisted reporting, I conducted a series of email interviews with data journalists during the 2012 NICAR Conference. They are available online and linked below.

- [The API architect](#)
- [The Daily Visualizer](#)
- [The Data Editor](#)
- [The Long Form Developer](#)
- [The Elections Developer](#)
- [The Human Algorithm](#)
- [The Visualizer](#)
- [The Storyteller and The Teacher](#)
- [The Hacks Hacker](#)

Endnotes

ⁱ C. Andersen, E. Bell, and C. Shirky, "Post Industrial Journalism: Adapting to the Present," Tow Center

ⁱⁱ A. Howard, "Knight Winners Are Putting Data to Work," O'Reilly Media, 26 Sep. 2012, <http://strata.oreilly.com/2012/09/knight-news-challenge-data-winners.html> (accessed 21 May 21 2014).

ⁱⁱⁱ A. Howard, "Tracking the Data Storm Around Hurricane Sandy," O'Reilly Media, 29 Oct. 2012, <http://strata.oreilly.com/2012/10/real-time-data-storm-in-hurricane-sandy-open-data.html> (accessed 21 May 2014).

^{iv} Sunlight Foundation, "No Justice Roberts, the Internet Can't do Government's Job," Sunlight Foundation Blog, 12 Apr. 2014, <http://sunlightfoundation.com/blog/2014/04/02/no-justice-roberts-the-internet-can-t-do-governments-job/> (accessed 21 May 2014).

^v A. Howard, "Data for the Public Good," O'Reilly Media, 22 Feb. 2012, <http://strata.oreilly.com/2012/02/data-public-good.html> (accessed 21 May 2014).

^{vi} M. Loukides, "What is Data Science?" O'Reilly Radar, 2 Jun. 2010, <http://radar.oreilly.com/2010/06/what-is-data-science.html> (accessed 21 May 2014).

^{vii} S. Rogers, "Data Journalism Only Matters When it's Transparent," *Mother Jones*, 24 Apr. 2014, <http://www.motherjones.com/media/2014/04/vox-538-upshot-open-data-missing> (accessed 21 2014).

^{viii} A. Howard, "NPR News App Team Experiments With Making Data-driven Public Media With the Public," Tow Center for Digital Journalism, 30 Aug. 2013, <http://towcenter.org/blog/npr-news-app-team-experiments-with-making-data-driven-public-media-with-the-public/> (accessed 21 May 2014).

^{ix} M. Slocum, "The Work of Data Journalism: Find, Clean, Analyze, Create... Repeat," O'Reilly Media, 15 Sep. 2011, <http://strata.oreilly.com/2011/09/data-journalism-process-guardian.html> (accessed 21 May 2014).

^x L. Bounegru, L. Chambers, and J. Gray, eds., *The Data Journalism Handbook* (Sebastopol, Calif.: O'Reilly Media, 2012), <http://datajournalismhandbook.org/> (accessed 21 May 2014).

^{xi} Ibid.

^{xii} S. Rogers, "Facts Are Sacred: the Power of Data," *Guardian*, 6 Jan. 2012, <http://www.theguardian.com/news/datablog/2012/jan/06/facts-sacred-guardian-shorts-ebook> (accessed 21 May 2014).

^{xiii} Ibid.

^{xiv} J. Townend, "#DataJourn Part 1: A New Conversation (Please Re-tweet)," Editors Blog, Journalism.co.uk, 8 Apr. 2009, <http://blogs.journalism.co.uk/2009/04/08/datajourn-part-1-a-new-conversation-please-re-tweet/> (accessed 21 May 2014).

^{xv} P. Bradshaw, "Model for the 21st Century Newsroom pt.6: New Journalists for New Information Flows," Online Journalism Blog, 4 Dec. 2008, <http://onlinejournalismblog.com/2008/12/04/model-for-the-21st-century-newsroom-pt6-new-journalists-for-new-information-flows/> (accessed 21 May 2014).

^{xvi} T. Hirst, "Personal Recollections of the 'Data Journalism' Phrase," OUsefullInfo, 29 Apr. 2014, <http://blog.ouseful.info/2014/04/29/personal-recollections-of-the-data-journalism-phrase/> (accessed 21 May 2014).

^{xvii} M. Ingram, "The Golden Age of Data Journalism?" Nieman Journalism Lab, May 2009, <http://www.niemanlab.org/2009/05/the-golden-age-of-computer-assisted-reporting-is-at-hand/> (accessed 21 May 2014).

^{xviii} A. Holovaty, "A Fundamental Way Newspaper Sites Need to Change," adrianholovaty.com, 2006, <http://www.holovaty.com/writing/fundamental-change/> (accessed 21 May 2014).

^{xix} M. Waite, "Announcing PolitiFact," mattwaite.com, 22 Aug. 2007, <http://www.mattwaite.com/posts/2007/aug/22/announcing-politifact/> (accessed 21 May 2014).

^{xx} "PolitiFact Wins Pulitzer," PolitiFact, 20 April 2009, <http://www.politifact.com/truth-o-meter/article/2009/apr/20/politifact-wins-pulitzer/> (accessed 21 May 2014).

^{xxi} C. Arthur, "Analysing Data is the Future for Journalists, Says Tim Berners-Lee," *Guardian*, 22 Nov. 2010, <http://www.theguardian.com/media/2010/nov/22/data-analysis-tim-berners-lee> (accessed 21 May 2014).

^{xxii} UK Parliament, "Pay and Expenses for MPs," Parliament.uk, April 2014, <http://www.parliament.uk/about/mps-and-lords/members/pay-mps/> (accessed 21 May 2014).

^{xxiii} C. Arthur, "Visualising MP Expenses," *Guardian*, 1 Apr. 2009, <http://www.theguardian.com/news/datablog/2009/apr/01/mps-expenses-houseofcommons> (accessed 21 May 2014).

^{xxiv} S. Rogers, "Wikileaks' Afghanistan War Logs: How our Data Journalism Operation Worked," *Guardian*, 27 Jul. 2010, <http://www.theguardian.com/news/datablog/2010/jul/27/wikileaks-afghanistan-data-datajournalism> (accessed 21 May 2014).

^{xxv} A. Howard, "In the Age of Big Data, Data Journalism has Profound Importance for Society," *Radar*, O'Reilly Media, March 2012, <http://strata.oreilly.com/2012/03/rise-of-the-data-journalists.html> (accessed 21 May 2014).

^{xxvi} D. Kaplan, "Data Journalists from 20 Countries Gather for Cutting-Edge NICAR14," Global Investigative Journalism Network, 3 Mar. 2014, <http://gijn.org/2014/03/03/data-journalists-from-20-countries-gather-for-cutting-edge-nicar14/> (accessed 21 May 2014).

^{xxvii} Associated Press, "The Overview Project," May 2014, <http://overview.ap.org/completed-stories/> (accessed 21 May 2014).

^{xxvii} N. Diakopoulos, "Algorithmic Accountability Reporting: On the Investigation of Black Boxes," Tow Center for Digital Journalism, February 2014, <http://towcenter.org/algorithmic-accountability-reporting-reverse-engineering-practice/> (accessed 21 May 2014).

^{xxix} A. Howard, "Publishers Can Afford Data Journalism, Says ProPublica's Scott Klein," Tow Center for Digital Journalism, 23 April 2014, <http://towcenter.org/blog/publishers-can-afford-data-journalism-scott-klein-propublica/> (accessed 21 May 2014).

^{xxx} M. Cox, "The Development of Computer-Assisted Reporting," School of Communication, University of Miami, <http://com.miami.edu/car/cox00.htm> (accessed 22 May 2014).

^{xxxi} A. Bochanek, "Have You Got a Prediction for Us, UNIVAC?" Computer History Museum, December 2012, <http://www.computerhistory.org/atchm/have-you-got-a-prediction-for-us-univac/> (accessed 22 May 2014).

^{xxxii} P. Meyer, *The New Precision Journalism* (Bloomington: Indiana University Press, 1991), <http://www.unc.edu/~pmeyer/book/> (foreword accessed 22 May 2014).

^{xxxiii} G. Younge, "The Detroit Riots of 1967 Hold Some Lessons for the UK," *Guardian*, 5 Sep. 2011, <http://www.theguardian.com/uk/2011/sep/05/detroit-riots-1967-lessons-uk> (accessed 22 May 2014).

^{xxxiv} E. Bowen, "Press: New Paths to Buried Treasure," *Time*, 7 July 1986, <http://content.time.com/time/magazine/article/0,9171,961680-1,00.html> (accessed 22 May 2014).

^{xxxv} Pulitzer.org, "The Pulitzer Prizes, 2014," 2014, <http://www.pulitzer.org/citation/2014-Investigative-Reporting> (accessed 22 May 2014).

^{xxxvi} S. MacGregor, "CAR Hits the Mainstream," *Columbia Journalism Review*, 18 Mar. 2013, http://www.cjr.org/data_points/computer_assisted_reporting.php?page=all (accessed 23 May 2014).

^{xxxvii} D. Kaplan, "Global Investigative Journalism: Strategies for Support," Center for International Media Assistance, National Endowment for Democracy, 13 Jan. 2014, <http://cima.ned.org/publications/global-investigative-journalism-strategies-support> (accessed 23 May 2014).

^{xxxviii} Committee to Protect Journalists, "Attacks on the Press," <https://www.cpj.org/2014/02/attacks-on-the-press-in-2013.php> (accessed 22 May 2014).

^{xxxix} Reporters Without Borders, "World Press Freedom Index 2014," <http://rsf.org/index2014/en-index2014.php> (accessed 22 May 2014).

^{x^l} L. Haddou, "Press Freedom 2014: The Global Picture," *Guardian*, 1 May 2014, <http://www.theguardian.com/news/datablog/2014/may/01/press-freedom-2014-the-global-picture> (accessed 22 May 2014).

^{xⁱ} Newsday Media Group LLC, "An Open-source Django App to Survey Politicians," <https://github.com/newsday/newstools-checkup> (accessed 22 May 2014).

^{xlvi} A. Howard, “Knight Winners are Putting Data to Work: Open Elections,” O’Reilly Media, 22 Sep. 2012, <http://strata.oreilly.com/2012/09/knight-news-challenge-data-winners.html#openelections> (accessed 23 May 2014).

^{xlvii} A. Howard, “Knight Winners are Putting Data to Work: Census Reporter,” O’Reilly Media, 22 Sep. 2012, <http://strata.oreilly.com/2012/09/knight-news-challenge-data-winners.html#censusIRE> (accessed 23 May 2014).

^{xlviii} A. Howard, “NPR News App Team Experiments With Making Data-driven Public Media With the Public,” Tow Center for Digital Journalism, 30 Aug. 2013, <http://towcenter.org/blog/npr-news-app-team-experiments-with-making-data-driven-public-media-with-the-public/> (accessed 21 May 2014).

^{xlvix} S. Johnson, “The Internet? We Built That,” *New York Times*, 21 Sep. 2012, <http://www.nytimes.com/2012/09/23/magazine/the-internet-we-built-that.html> (accessed 23 May 2014).

^{xlvi} S. Johnson, “Peer Power, from Potholes to Patents,” *Wall Street Journal*, <http://online.wsj.com/news/articles/SB10000872396390444165804578008511493789642> (accessed May 23, 2014).

^{xlvii} A. Howard, “Knight Winners are Putting Data to Work: Census Reporter,” (22 Sep. 2012).

^{xlviii} A. Howard, “Applying Data Science to All the News That’s Fit to Print,” Tow Center for Digital Journalism, 7 Apr. 2014, <http://towcenter.org/blog/applying-data-science-to-all-the-news-thats-fit-to-print/> (accessed 23 May 2014).

^{xlix} S. Klein, “ProPublica News Apps Style Guide,” GitHub, 27 Apr. 2014, <https://github.com/propublica/guides/blob/master/news-apps.md> (accessed 23 May 2014).

ⁱ H. Vinter, “Scott Klein: News Apps Don’t Just Tell a Story, They Tell Your Story,” World Association of Newspapers and News Publishers, 24 Aug. 2011, <http://www.wan-ifra.org/articles/2011/08/24/scott-klein-news-apps-dont-just-tell-a-story-they-tell-your-story> (accessed 23 May 2014).

ⁱⁱ ProPublica, “Treatment Tracker,” 15 May 2014, <http://projects.propublica.org/treatment/> (accessed 23 May 2014).

ⁱⁱⁱ R.G. Jones and C. Ornstein, “Top Billing: Meet the Docs who Charge Medicare Top Dollar for Office Visits,” ProPublica, 15 May 2014, <http://www.propublica.org/article/billing-to-the-max-docs-charge-medicare-top-rate-for-office-visits> (accessed 23 May 2014).

^{iv} N. Diakopoulos, “Algorithmic Accountability Reporting: On the Investigation of Black Boxes,” Tow Center for Digital Journalism, February 2014, <http://towcenter.org/algorithmic-accountability-reporting-reverse-engineering-practice/> (accessed 21 May 2014).

^{lv} C. Wu, "White House Safety Datapalooza: Bicoastal Safety Data Journalist Workshop," WhiteHouse.gov, 12 Sep. 2014, <http://www.whitehouse.gov/photos-and-video/video/2012/09/14/white-house-safety-datapalooza-bicoastal-safety-data-journalist-wo> (accessed 23 May 2014).

^{lv} N. Diakopoulos, "The Rhetoric of Data," Tow Center for Digital Journalism, 25 July 2013, <http://towcenter.org/blog/the-rhetoric-of-data/> (accessed 23 May 2014).

^{lv} R. Sambrook, "Journalists Can Learn Lessons From Coders in Developing the Creative Future," *Guardian*, 27 April 2014, <http://www.theguardian.com/media/2014/apr/27/journalists-coders-creative-future> (accessed 23 May 2014).

^{lvii} B. Keegan, "The Need for Openness in Data Journalism," briankeegan.com, 7 April 2014, <http://www.briankeegan.com/2014/04/the-need-for-openness-in-data-journalism/> (accessed 23 May 2014).

^{lviii} S. Owens, "A Place for Homicide Watch: Can a Local Blog Fill Some of the Gaps in Washington, D.C.'s Crime Coverage?" Nieman Journalism Lab, 5 May 2011, <http://www.niemanlab.org/2011/05/homicide-watch-can-a-local-blog-fill-in-the-gaps-of-dcs-homicide-coverage/> (accessed 23 May 2014).

^{lix} L. Amico, "Reporting From Analytics: Example," One Reporter's Notebook, 4 May 2011, <http://lauraamico.tumblr.com/post/5196806316/reporting-from-analytics-example> (accessed 23 May 2014).

^{lx} S. Myers, "Homicide Watch D.C. Uses Clues in Site Search Queries to ID Homicide Victim," Poynter, 12 Oct. 2011. <http://www.poynter.org/latest-news/mediawire/149294/homicide-watch-d-c-uses-clues-in-site-search-queries-to-id-homicide-victim/> (accessed 23 May 2014).

^{xi} L. Amico, "On Deadline: Why Organizing Beats is Just as Important as Large Investigations," Online News Association, 21 Feb. 2012, <http://journalists.org/2012/02/21/on-deadline-why-organizing-beats-is-just-as-important-as-large-investigations/> (accessed 23 May 2014).

^{xii} L. Indvik, "The *Financial Times* Has a Secret Weapon: Data," Mashable, 2 April 2013, <http://mashable.com/2013/04/02/financial-times-john-ridding-strategy/> (accessed 23 May 2014).

^{xiii} O. Thomas, "The Data-Driven Future Of Journalism," ReadWrite, 6 Sep. 2013, <http://readwrite.com/2013/09/06/data-journalism-future> (accessed 23 May 2014).

^{xiv} McKinsey Global Institute, "The Social Economy: Unlocking Value and Productivity Through Social Technologies," McKinsey & Company, 2012, http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_social_economy (accessed 23 May 2014).

^{xv} Pew Research Center, "State of the Media 2013," Pew Project for Excellence in Journalism, 18 May 2013, <http://stateofthemedia.org/2013/overview-5/> (accessed 23 May 2014).

^{lxvi} Ibid.

^{lxvii} "Paper Cuts," 28 Apr. 2013, <http://newspaperlayoffs.com/>.

^{lxviii} D. Sinker, "We're All Living in an EveryBlock World," [dansinker.com](http://dansinker.com/post/42525273840/were-all-living-in-an-everyblock-world), 7 Feb. 2013, <http://dansinker.com/post/42525273840/were-all-living-in-an-everyblock-world> (accessed 23 May 2014).

^{lxix} J. Sonderman, "NBC Closes Hyperlocal, Data-driven Publishing Pioneer EveryBlock," Poynter, 2 Feb. 2013, <http://www.poynter.org/latest-news/top-stories/203437/nbc-closes-hyperlocal-pioneer-everyblock/> (accessed 23 May 2014).

^{lxx} R. Graff, "Journalism's Biggest Data Experiment, EveryBlock, Relaunches," Knight Lab, Northwestern University, 31 Jan 2014, <http://knightlab.northwestern.edu/2014/01/31/journalisms-biggest-data-experiment-everyblock-relaunches/> (accessed 23 May 2014).

^{lxxi} A. Chavez, "Why Some Hyperlocal Sites Struggle to Attract Audiences, Generate Revenue," Poynter, 12 Mar. 2012. <http://www.poynter.org/latest-news/top-stories/166190/why-some-hyperlocal-sites-struggle-to-attract-audiences-generate-revenue/> (accessed 23 May 2014).

^{lxxii} W. Huntsberry, "Why Hyperlocal Websites Like New Raleigh Can't Make Money Online," [indyweek.com](http://www.indyweek.com/indyweek/why-hyperlocal-websites-like-new-raleigh-cant-make-money-online/Content?oid=3250659), 23 Jan. 2013, <http://www.indyweek.com/indyweek/why-hyperlocal-websites-like-new-raleigh-cant-make-money-online/Content?oid=3250659> (accessed 23 May 2014).

^{lxxiii} H. Ji, M. Jurkowitz, and T. Rosenstiel, "The Search for a New Business Model," Pew Research Journalism Project, 5 Mar. 2012, <http://www.journalism.org/2012/03/05/search-new-business-model/>.

^{lxxiv} L. Kaufman, "Patch Sites Turn Corner After Sale and Big Cuts," *New York Times*, 19 May 2014, <http://www.nytimes.com/2014/05/19/business/media/patch-sites-turn-corner-after-sale-and-big-cuts.html> (accessed 23 May 2014).

^{lxxv} J. Paxton, "Moving On From Thunderdome," John Paxton, 2 April 2014, <http://jxpaton.wordpress.com/2014/04/02/moving-on-from-thunderdome/> (accessed 23 May 2014).

^{lxxvi} K. Doctor, "The Newsomics of Digital First Media's Thunderdome Implosion (and Coming Sale)," Nieman Journalism Lab, 2 Apr. 2014, <http://www.niemanlab.org/2014/04/the-newsomics-of-digital-first-medias-thunderdome-implosion-and-coming-sale/>.

^{lxxvii} D. Freid and B. Prieto, "Firearms in the Family," Digital First Media, 2013, <http://data.digitalfirstmedia.com/guns/> (accessed 23 May 2014).

^{lxxviii} "Decoding the Kennedy Assassination," Digital First Media, 2013, <http://data.digitalfirstmedia.com/jfk/> (accessed 23 May 2014).

^{lxxix} "Bracket Advisor," *New Haven Register*, 2014, <http://www.bracketadvisor.com/> (accessed 23 May 2014).

^{lxxx} A. Howard, "Publishers Can Afford Data Journalism Says ProPublica's Scott Klein," 23 Apr. 2014, <http://towcenter.org/blog/publishers-can-afford-data-journalism-scott-klein-propublica/> (accessed 23 May 2014).

^{lxxxi} R. Yu, "Booming Market for Data-driven Journalism," *USA Today*, 17 Mar. 2014, <http://www.usatoday.com/story/money/business/2014/03/16/data-journalism-on-the-rise/6424671/> (accessed 23 May 2014).

^{lxxxi} Pew Research Center, "The Growth of Digital Reporting," Pew Research Journalism Project, 26 Mar. 2014, <http://www.journalism.org/2014/03/26/the-growth-in-digital-reporting/> (accessed 23 May 2014).

^{lxxxiii} Pew Research Center, "State of the Media 2014," (26 May 2014).

^{lxxxiv} L. Moses, "Is There an Ad Model for Explainer Journalism?" *Digiday*, 22 Apr. 2014, <http://digiday.com/publishers/ad-model-explainer-journalism/> (accessed 23 May 2014).

^{lxxxv} L. Bounegru, L. Chambers, and J. Gray, eds., 2012.

^{lxxxvi} B. Mullens and C. Weaver, "Open-Government Laws Fuel Hedge-Fund Profits," *Wall Street Journal*, 23 Sep 2013, <http://online.wsj.com/news/articles/SB1000142412788732420230457905303344112314> (accessed 23 May 2014).

^{lxxxvii} A. Howard, "As Digital Disruption Comes to Africa, Investing in Data Journalism Takes on New Importance," *Radar*, O'Reilly Media, 29 Nov. 2012, <http://radar.oreilly.com/2012/11/justin-arenstein-africa-data-journalism.html> (accessed 23 May 2014).

^{lxxxviii} S. Myers, "New York Times News Apps Team Ventures into Product Development with Olympics Syndication," Poynter, 8 Aug. 2012, <http://www.poynter.org/latest-news/top-stories/184315/new-york-times-news-apps-team-ventures-into-product-development-with-olympics-syndication/> (accessed 23 May 2014).

^{lxxxix} Associated Press, "AP Products and Services: New Media," <http://www.ap.org/products-services/new-media> (accessed 23 May 2014).

^{xc} J. Maher, "London Calling: Winning the Data Olympics," Source, OpenNews, 25 Apr. 2013, <https://source.opennews.org/en-US/learning/london-calling-winning-data-olympics/> (accessed 23 May 2014).

^{xcii} E. Smith, "T-Squared: Three More Years!" *Texas Tribune*, 5 Nov. 2012, <http://www.texastribune.org/2012/11/05/t-squared-three-more-years/> (accessed 23 May 2014).

^{xciii} "The Texas Tribune Higher Education Explorer," *Texas Tribune*, <http://www.texastribune.org/higher-ed/explore/> (accessed 23 May 2014).

^{xciii} "Data Pages | The Texas Tribune," *Texas Tribune*, <http://www.texastribune.org/library/data/> (accessed 23 May 2014).

^{xciv} “Elected Officials Directory | The Texas Tribune,” *Texas Tribune*, <http://www.texastribune.org/directory/> (accessed 23 May 2014).

^{xcv} E. Smith, “T-Squared: It’s Only Bidness,” *Texas Tribune*, 14 Jan. 2013, <http://www.texastribune.org/2013/01/14/t-squared-it-was-only-bidness/> (accessed 23 May 2014).

^{xcvi} “Texas Prison Inmates,” *Texas Tribune*, <http://www.texastribune.org/library/data/texas-prisons/> (accessed 23 May 2014).

^{xcvii} “Government Employee Salaries,” *Texas Tribune*, <http://www.texastribune.org/library/government-employee-salaries/> (accessed 23 May 2014).

^{xcviii} “Texas Gubernatorial Election Results Maps,” *Texas Tribune*, 7 Nov. 2010, <http://www.texastribune.org/library/data/texas-governors-race-results-maps/> (accessed 23 May 2014).

^{xcix} S. Klein, “Introducing the ProPublica Data Store,” ProPublica, 26 Feb. 2014, <http://www.propublica.org/article/introducing-the-propublica-data-store> (accessed 23 May 2014).

^c J. Ellis, “ProPublica Opens Up Shop with a New Site to Sell Custom Datasets,” Nieman Journalism Lab, 4 Mar. 2014, <http://www.niemanlab.org/2014/03/propublica-opens-up-shop-with-a-new-site-to-sell-custom-datasets/> (accessed 23 May 2014).

^{ci} ProPublica, “Dollars for Docs,” <https://projects.propublica.org/docdollars/> (accessed 23 May 2014).

^{cii} S. Klein and R. Tofel, “ProPublica: Why we Use Creative Commons Licenses on our Stories,” Nieman Journalism Lab, 13 Dec. 2012. <http://www.niemanlab.org/2012/12/propublica-why-we-use-creative-commons-licenses-on-our-stories/> (accessed 23 May 2014).

^{ciii} T. Ali, “ProPublica Plans to Grow its ‘Data Store,’ ” *Columbia Journalism Review*, 28 Apr. 2014, http://www.cjr.org/behind_the_news/propublica_plans_to_grow_its_d.php (accessed 23 May 2014.)

^{civ} J. Webb, “Transforming Data into Narrative Content,” O'Reilly Media, 26 Jan. 2012, <http://toc.oreilly.com/2012/01/narrative-science-kristian-hammond-data-content-generation.html> (accessed 23 May 2014).

^{c^v} W. Oremus, “The First News Report on the L.A. Earthquake Was Written by a Robot,” *Slate*, 17 Mar. 2014, http://www.slate.com/blogs/future_tense/2014/03/17/quakebot_los_angeles_times_robot_journalist_writes_article_on_la_earthquake.html (accessed 23 May 2014).

^{c^{vi}} “The Homicide Report,” *Los Angeles Times*, May 2014, <http://homicide.latimes.com/> (accessed 23 May 2014).

^{c^{vii}} A. Webb, “The Future of News is Anticipation,” Nieman Journalism Lab, Dec. 2013, <http://www.niemanlab.org/2013/12/the-future-of-news-is-anticipation/> (accessed 23 May 2014).

^{cix} “Curbwise,” *Omaha World-Herald*, <http://www.curbwise.com/> (accessed 23 May 2014).

^{cix} A. Howard, “Open Government Data Shines a Light on Hospital Billing and Health Care Costs,” E Pluribus Unum., 8 May 2013, <http://e-pluribusunum.com/2013/05/08/open-government-data-hospital-billing-healthcare-cost/> (accessed 23 May 2014).

^{cix} A. Howard, “Medicare Release and DATA Act Signal Major Events in the Age of Data Transparency,” TechRepublic, 15 Apr. 2014, <http://www.techrepublic.com/article/medicare-release-and-data-act-signal-major-events-in-the-age-of-data-transparency/> (accessed 23 May 2014).

^{cxi} P. Reese and D. Smith, “Million-dollar Hospital Bills Rise Sharply in Northern California,” *The Sacramento Bee*, 11 Mar. 2012, <http://www.sacbee.com/2012/03/11/4328036/million-dollar-hospital-bills.html> (accessed 23 May 2014).

^{cxii} “Patient Safety,” *The Dallas Morning News* Investigations and Special Reports, 2012, <http://www.dallasnews.com/investigations/patient-safety/> (accessed 23 May 2014).

^{cxiii} L. Girion, S. Glover, and D. Smith, “Drugs Deaths Now Outnumber Traffic Fatalities in U.S., Data Show,” *Los Angeles Times*, 17 Sep. 2011, <http://articles.latimes.com/2011/sep/17/local/la-me-drugs-epidemic-20110918>.

^{cxiv} B. Sanderlin, “Fake Medical Providers Slip Through Medicare Loophole,” *Atlanta Journal-Constitution*, 2 Dec. 2012, <http://www.ajc.com/news/news/fake-medical-providers-slip-through-medicare-looph/nTLFF/> (accessed 23 May 2014).

^{cxv} “Medical Helicopter Flights Mostly for Routine Transport, *Argus Leader*, via IRE, <http://www.ire.org/blog/extra-extra/2012/12/03/hospital-helicopters-worth-cost/> (accessed 23 May 2014).

^{cxvi} “Election Results,” *New York Times*, 2012, <http://elections.nytimes.com/2012/results/president> (accessed 23 May 2014).

^{cxvii} “Toxic Waters,” *New York Times*, 2009–2010, <http://projects.nytimes.com/toxic-waters> (accessed 23 May 2014).

^{cxviii} “U.S. Congressional Votes Database,” *Washington Post*, 2005, <http://projects.washingtonpost.com/congress/113/> (accessed 23 May 2014).

^{cix} “Philip Meyer Journalism Awards,” Investigative Reporters and Editors, <http://www.ire.org/awards/philip-meyer-awards/> (accessed 23 May 2014).

^{cxx} “Data Journalism Awards 2014,” Global Editors Network, <http://www.globaleditorsnetwork.org/programmes/data-journalism-awards/> (accessed 23 May 2014).

^{cxxi} “The Prescribers,” ProPublica, 2013–2014, <http://www.propublica.org/series/prescribers> (accessed 23 May 2014).

cxxii "Dollars for Docs," ProPublica, <http://projects.propublica.org/docdollars/> (accessed 23 2014).

cxxiii D. Nguyen, "Scraping for Journalism: A Guide for Collecting Data," ProPublica, 30 Dec. 2010, <http://www.propublica.org/nerds/item/doc-dollars-guides-collecting-the-data> (accessed 23 May 2014).

cxxiv J. Merrill, A. Shaw, and A. Zamora, "Free the Files," ProPublica, 21 May 2014, <https://projects.propublica.org/free-the-files/> (accessed 23 May 2014).

cxxv J. Elliot, "Political Ad Data Comes Online — But It's Not Searchable," ProPublica, 2 Aug. 2012, <http://www.propublica.org/article/political-ad-data-comes-online-but-its-not-searchable> (accessed 23 May 2014).

cxxvi A. Shaw, "Transcribable: Free the Files to Go!" ProPublica, 16 Jul. 2013, <http://www.propublica.org/nerds/item/transcribable-free-the-files-to-go> (accessed 23 May 2014).

cxxvii A. Zamora, "Crowdsourcing Campaign Spending: What We Learned From Free the Files," ProPublica, 12 Dec. 2012, <http://www.propublica.org/article/crowdsourcing-campaign-spending-what-we-learned-from-free-the-files> (accessed 23 May 2014).

cxxviii "Cicada Tracker," WNYC, <http://project.wnyc.org/cicadas/>, May 2013, (accessed 23 May 2014).

cxxix C. Donovan, "The Cicadas Are Here: 4 Lessons From WNYC's Cicada Tracker Project," Nieman Journalism Lab, 3 Jun. 2013, <http://www.niemanlab.org/2013/06/the-cicadas-are-here-4-lessons-from-wnycts-cicada-tracker-project/> (accessed 23 May 2014).

cxxx A. Howard, "Sensoring the News," *Radar*, O'Reilly Media, 22 Mar. 2013, <http://radar.oreilly.com/2013/03/sensor-journalism-data-journalism.html> (accessed 23 May 2014).

cxxxi Tow Center for Digital Journalism, Columbia University, Sensor Journalism Workshop, June 1-2, 2013, <http://towcenter.org/research/sensor-journalism-at-the-tow-center/sensor-journalism-workshop-at-the-tow-center/> (accessed 23 May 2014).

cxxxii "Playgrounds for Everyone," NPR, <http://apps.npr.org/playgrounds/> (accessed 23 May 2014).

cxxxiii J. Robbins, "Crowdsourcing, for the Birds," *New York Times*, 19 Aug. 2013, <http://www.nytimes.com/2013/08/20/science/earth/crowdsourcing-for-the-birds.html?pagewanted=all> (accessed 23 May 2014).

cxxxiv M. Waite, "Slouching Toward Sensor Journalism," Source, OpenNews, 11 Jun. 2013, <https://source.opennews.org/en-US/articles/slouching-toward-sensor-journalism/> (accessed 23 May 2014).

cxxxv E. Zuckerman, "Citizen Science Versus NIMBY?" "My Heart's in Accra," 29 Aug. 2013, <http://www.ethanzuckerman.com/blog/2013/08/29/citizen-science-versus-nimby/> (accessed 23 May 2014).

^{cxxxvi} A. Miars, “NPR’s Apps Editor Brian Boyer Turns Data into Stories,” It’s All Journalism, 6 Jul. 2013, <http://itsalljournalism.com/nprs-apps-editor-brian-boyer-turns-data-into-stories/> (accessed 23 May 2014).

^{cxxxvii} *Guardian* Datablog, <http://www.theguardian.com/news/datablog> (accessed 23 May 2014).

^{cxxxviii} J. Burn-Murdoch, “Mapping Racist Tweets in Response to President Obama’s re-election,” *Guardian* Datablog, 9 Nov. 2012, <http://www.theguardian.com/news/datablog/2012/nov/09/mapping-racist-tweets-president-obama-reelection> (accessed 23 May 2014).

^{cxxxix} S. Rogers, “Government Spending by Department, 2011–12,” *Guardian* Datablog, 4 Dec. 2012, <http://www.theguardian.com/news/datablog/2012/dec/04/government-spending-department-2011-12> (accessed May 23, 2014).

^{cxl} S. Rogers, “Named and Shamed: The Worst Government Annual Reports, 2012,” *Guardian* Datablog, 4 Dec. 2012, <http://www.theguardian.com/news/datablog/2012/dec/04/departmental-reports-worst-named-shamed> (accessed 23 May 2014).

^{cxi} S. Rogers, “Gun Crime Statistics by U.S. State: Latest Data,” *Guardian* Datablog, 10 Jan. 2011, <http://www.theguardian.com/news/datablog/2011/jan/10/gun-crime-us-state> (accessed 23 May 2014).

^{cxlii} S. Rogers, “The Gun Ownership and Gun Homicides Murder Map of the World,” *Guardian* Datablog, 22 Jul. 2012, <http://www.theguardian.com/news/datablog/interactive/2012/jul/22/gun-ownership-homicides-map> (accessed 23 May 2014).

^{cxlvi} A. Howard, “UK Cabinet Office Relaunches Data.gov.uk, Releases Open Data White Paper,” *Radar*, O'Reilly Media, 29 Jun. 2012, <http://strata.oreilly.com/2012/06/uk-cabinet-office-relaunches-d.html> (accessed 23 May 2014).

^{cxliv} A. Howard, “Open Data 500: Proof That Open Data Fuels Economic Activity,” TechRepublic, 8 Apr. 2014, <http://www.techrepublic.com/article/open-data-500-proof-that-open-data-fuels-economic-activity/> (accessed 23 May 2014).

^{cxdv} A. Howard, “Finding and Telling Data-driven Stories in Billions of Tweets,” *Radar*, O'Reilly Media, 18 Apr. 2013, <http://strata.oreilly.com/2013/04/finding-and-telling-data-driven-stories-in-billions-of-tweets.html> (accessed 23 May 2014).

^{cxlvi} S. Rogers, “Simon Rogers on Data Journalism in the Open,” Tow Center for Digital Journalism, 25 Mar. 2014, <http://digiphile.tumblr.com/post/80800219507/simon-rogers-on-data-journalism-in-the-open> (accessed 23 May 2014).

^{cxlvi} J. Cheshire, Spatial.ly, <http://spatialanalysis.co.uk/> (accessed 23 May 2014).

^{cxlviii} Oxford Internet Institute, <http://www.floatingsheep.org/> (accessed 23 May 2014).

^{cxlvi} “Data Desk,” Shaw Media, <http://globalnews.ca/data-desk/> (accessed 23 May 2014).

^{c1} “On Oklahoma Tornado Damage,” NPR, <http://apps.npr.org/moore-oklahoma-tornado-damage/> (accessed 23 May 2014).

^{c1i} A. Howard, “Profile of the Data Journalist: The Human Algorithm,” *Radar*, O'Reilly Media, 2 Mar. 2013, <http://strata.oreilly.com/2012/03/profile-of-the-data-journalist-2.html> (accessed 23 May 2014).

^{c1ii} “Map: How Fast is LAFD Where you Live?” *Los Angeles Times*, <http://graphics.latimes.com/how-fast-is-lafd> (accessed 23 May 2014).

^{c1iii} “911 Breakdowns at LAFD,” *Los Angeles Times*, 2012, <http://www.latimes.com/local/lafddata/> (accessed 23 May 2014).

^{c1iv} R. Lopez and B. Welsh, “Flawed Data Stall California’s 911 Upgrades,” *Los Angeles Times*, 21 Dec. 2012, <http://articles.latimes.com/2012/dec/21/local/la-me-ems-data-problems-20121222> (accessed 23 May 2014).

^{c1v} A. Howard, “Profile of the Data Journalist: The Human Algorithm,” 2 Mar. 2013.

^{c1vi} Oakland Police Beat, 2014, <http://oaklandpolicebeat.com/> (accessed 23 May 2014).

^{c1vii} “Open-source Maps of California’s Emergency Medical Agencies,” *Los Angeles Times*, Dec. 2012, <http://datadesk.latimes.com/posts/2012/12/map-of-california-ems-agencies/> (accessed 23 May 2014).

^{c1viii} *Los Angeles Times* Data Desk, GitHub, <https://github.com/datadesk> (accessed 23 May 2014).

^{c1ix} B. Welsh, “Inside Our 911 Response Time Analysis,” *Los Angeles Times*, 20 Oct. 2012, <http://datadesk.latimes.com/posts/2012/10/lafd-border-analysis/> (accessed 23 May 2014).

^{c1x} B. Welsh, “Introducing Quiet L.A.,” *Los Angeles Times*, 14 Nov. 2012, <http://datadesk.latimes.com/posts/2012/11/introducing-quiet-la/> (accessed 23 May 2014).

^{c1xi} “Map: How Fast is LAFD Where you Live?” *Los Angeles Times*, <http://graphics.latimes.com/how-fast-is-lafd> (accessed 23 May 2014).

^{c1xii} B. Welsh, “The Times Contributes LAFD Fire Stations to OpenStreetMap,” *Los Angeles Times*, 6 Dec. 2012, <http://datadesk.latimes.com/posts/2012/12/lafd-stations-in-osm/> (accessed 23 May 2014).

^{c1xiii} “Prescriber Checkup,” ProPublica, <https://projects.propublica.org/checkup/> (accessed 23 May 2014).

^{c1xiv} The Upshot, *New York Times*, <http://www.nytimes.com/upshot/> (accessed 23 May 2014).

^{c1xv} D. Leonhardt, “Back Story: How We Found the Income Data,” *New York Times*, 23 Apr. 2014, <http://www.nytimes.com/2014/04/23/upshot/back-story-how-we-found-the-income-data.html> (accessed 23 May 2014).

^{cxxxvi} A. Cox and J. Katz, "Senate Model Methodology," *New York Times*, Apr. 2014, <http://www.nytimes.com/newsgraphics/2014/senate-model/methodology.html> (accessed 23 May 2014).

^{cxxxvii} "LEO Senate Model," *New York Times*, GitHub, <https://github.com/TheUpshot/leo-senate-model> (accessed 23 May 2014).

^{cxxxviii} J. Bourgault, "How the Global Open Data Movement is Transforming Journalism," *Wired*, May 2013, <http://www.wired.com/2013/05/how-the-global-open-data-movement-is-transforming-journalism/> (accessed 23 May 2014).

^{cxxxix} J. Ball, "The Upshot, Vox and FiveThirtyEight: Data Journalism's Golden Age, or TMI?" *Guardian* Datablog, 22 Apr. 2014, <http://www.theguardian.com/commentisfree/2014/apr/22/upshot-vox-fivethirtyeight-data-journalism-golden-age> (accessed 23 May 2014).

^{cxlxx} C. Correa, "Fear Not, Readers: We Have RSS Feeds," 538 DataLab, 28 Mar. 2014, <http://fivethirtyeight.com/datalab/fear-not-readers-we-have-rss-feeds/> (accessed 23 May 2014).

^{cxlxi} "Data," FiveThirtyEight, GitHub, <https://github.com/fivethirtyeight/data> (accessed 23 May 2014).

^{cxlxii} "Statement," TheUpshot, *New York Times*, GitHub, <https://github.com/TheUpshot/statement> (accessed 23 May 2014).

^{cxlxiii} Vox Media, GitHub, <https://github.com/voxmedia> (accessed 23 May 2014).

^{cxlxiv} C. Cross and S. Rogers, "All our datasets: The Complete Index," *Guardian* Datablog, 14 Jan. 2014, <http://www.theguardian.com/news/datablog/interactive/2013/jan/14/all-our-datasets-index> (accessed 23 May 2014).

^{cxlxv} D. Kaplan, "Why Open Data Isn't Enough," Global Investigative Journalism Network, 2 Apr. 2013, <http://gijn.org/2013/04/02/why-open-data-isnt-enough/> (accessed 23 May 2014).

^{cxlxvi} D. Campbell, "How ICIJ's Project Team Analyzed the Offshore Files," International Consortium of Investigative Journalists, 3 Apr. 2013, <http://www.icij.org/offshore/how-icijs-project-team-analyzed-offshore-files> (accessed 23 May 2014).

^{cxlxvii} E. Moore, "Offshore Leaks: A Triumph for Data Journalism," World News Publishing Focus, 8 Apr. 2013, <http://blog.wan-ifra.org/2013/04/08/offshore-leaks-a-triumph-for-data-journalism> (accessed 23 May 2014).

^{cxlxviii} "Connected China," Reuters, 2013, <http://connectedchina.reuters.com/> (accessed 23 May 2014).

^{cxlxix} I. Liu, "Welcome to Connected China," Reuters, 28 Feb. 2013, <http://blogs.reuters.com/connected-china/2013/02/28/welcome/> (accessed 23 May 2014).

^{cxxx} A. Heim, "Poderopedia, a Data Journalism Project to Map the Chilean Elite," *The Next Web*, 16 Mar. 2012, <http://thenextweb.com/media/2012/03/16/poderopedia-a-data-journalism-project-to-map-the-chilean-elite/> (accessed 23 May 2014).

^{cxxxii} A. Howard, "Data Journalism, Data Tools, and the Newsroom Stack," *Radar*, O'Reilly Media, 5 Jul. 2011, <http://strata.oreilly.com/2011/07/data-journalism-tools-newsroom-stack.html> (accessed 23 May 2014).

^{cxxxii} M. Paz, "Journalists Will Use Poderopedia-powered Platform to Inform Voters in Panama," Knight Foundation Blog, 2 Feb. 2014, <http://www.knightfoundation.org/blogs/knightblog/2014/2/21/journalists-will-use-poderopedia-powered-platform-inform-voters-panama/> (accessed 23 May 2014).

^{cxxxiii} P. Navalte, "Poderopedia, the Chilean Data Journalism Platform, Plans to Expand to Venezuela and Colombia," Knight Center for Journalism in the Americas, 6 Nov. 2013, <https://knightcenter.utexas.edu/blog/00-14694-poderopedia-chilean-data-journalism-platform-plans-expand-venezuela-and-colombia> (accessed 23 May 2014).

^{cxxxiv} J. Weiss, "How Open Data Can Revolutionize Environmental Reporting," *PBS MediaShift*, 28 Nov. 2012, <http://www.pbs.org/mediashift/2012/11/how-open-data-can-revolutionize-environmental-reporting333> (accessed 23 May 2014).

^{cxxxv} "NASA Satellite Measures Deforestation," NASA Earth Observatory, 14 Sep. 2005, <http://earthobservatory.nasa.gov/IOTD/view.php?id=5845> (accessed 23 May 2014).

^{cxxxvi} G. Faleiros, "Geojournalism Handbook Shows How to Capture Earth Science Knowledge for Reporting," 24 Sep. 2013, <http://ijnet.org/blog/geojournalism-handbook-shows-how-capture-earth-science-knowledge-reporting> (accessed 23 May 2014).

^{cxxxvii} W. Shubert, "Better Mapping for Better Journalism: InfoAmazonia and the Growth of GeoJournalism," <https://www.internews.org/better-mapping-better-journalism-infoamazonia-and-growth-geojournalism> (accessed 23 May 2014).

^{cxxxviii} Oxpeckers Center for Investigative Environmental Journalism, <http://oxpeckers.org/> (accessed 23 May 2014).

^{cxxxix} "Land Quest," Internews Kenya, <http://landquest.internewskenya.org/> (accessed 23 May 2014).

^{cxc} J. Dorroh, "Data Journalism Site InfoAmazonia Will Add Ground Reporting to its Environmental Coverage," IJNet, 8 Apr. 2014, <http://ijnet.org/blog/data-journalism-site-infoamazonia-will-add-ground-reporting-its-environmental-coverage> (accessed 23 May 2014).

^{cxcii} K. Witkin, “*La Nación* Multimedia Editor: Innovation is the ‘Antidote’ to Journalism Crisis,” World News Publishing Focus, 24 Jul. 2013, <http://blog.wan-ifra.org/2013/07/24/la-nacion-multimedia-editor-innovation-is-the-antidote-to-journalism-crisis> (accessed 23 May 2014).

^{cxciii} A. Jiménez, “How *La Nación* is Using Data to Challenge a FOIA-free Culture,” Nieman Journalism Lab, 23 May 2012, <http://www.niemanlab.org/2012/05/how-la-nacion-is-using-data-to-challenge-a-foia-free-culture/> (accessed 23 May 2014).

^{cxciv} “Argentina’s Official Advertising Funds Distribution 2009–2013: Friends, Politicians, and a Stylist,” *La Nación*, 4 Apr. 2014, <http://blogs.lanacion.com.ar/projects/data/argentina%C2%B4s-official-advertising-funds-distribution-2009-2013-friends-politicians-and-a-stylist/> (accessed 23 May 2014).

^{cxcv} “Public Officials Salaries and Assets for Reporting and Accountability,” *La Nación*, 4 Apr. 2014, <http://blogs.lanacion.com.ar/projects/data/statements-of-assets/> (accessed 23 May 2014).

^{cxcvi} “Monitoring the New Media Law in Argentina 2009–2013,” *La Nación*, 4 Apr. 2014, <http://blogs.lanacion.com.ar/projects/data/dja-afscal/> (accessed 23 May 2014).

^{cxcvii} “VozData: Collaborating to Free Data from PDFs—The Senate Expenses Part II,” *La Nación*, 4 Apr. 2014, <http://blogs.lanacion.com.ar/projects/data/vozdata/> (accessed 23 May 2014).

^{cxcviii} “2013 Legislative Elections in Argentina,” *La Nación*, 3 Mar. 2014, <http://blogs.lanacion.com.ar/projects/data/2013-legislative-elections-in-argentina/> (accessed 23 May 2014).

^{cxcix} “Argentina’s Senate Expenses 2004–2013,” *La Nación*, 4 Mar. 2014, <http://blogs.lanacion.com.ar/ddj/data-driven-investigative-journalism/argentina-senate-expenses/> (accessed 23 May 2014).

^{cc} T. Turner, “Argentina Imposes Ad Ban, Businesses Say,” *Wall Street Journal*, 8 Feb. 2013, <http://online.wsj.com/news/articles/SB10001424127887324906004578292433748498360> (accessed 23 May 2014).

^{cci} G. Romero, “I dati sulla sicurezza sismica delle scuole,” *Wired Italy*, 9 Nov. 2012, <http://daily.wired.it/news/politica/2012/11/09/scuole-sicurezza-terremoto-143678.html> (accessed 23 May 2014).

^{ccii} E. Tola, “Terremoto: la tua scuola è a rischio?” *Wired Italy*, 17 Sep. 2012, <http://daily.wired.it/news/politica/2012/09/17/terremoto-scuola-rischio.html> (accessed 23 May 2014).

^{cciii} “La tua scuola è sicura? Cercala sulla mappa,” *Wired Italy*, 9 Nov. 2012, <http://daily.wired.it/news/politica/2012/11/09/scuolesicure-mappa-scuole-sicurezza-terremoto-143678.html> (accessed 23 May 2014).

^{cciv} Government of Italy, <http://www.dati.gov.it/> (accessed 23 May 2014).

^{ccv} “ForumPA: Il premio Apps4Italy intitolato a Melissa Bassi e alle vittime dell'attentato di Brindisi | Saperi PA,” ForumPA: Il premio Apps4Italy intitolato a Melissa Bassi e alle vittime dell'attentato di Brindisi | Saperi PA, <http://saperi.forumpa.it/story/74770/forumpa-il-premio-apps4italy-intitolato-melissa-bassi-e-alle-vittime-dell-attentato-di> (accessed 23 May 2014).

^{ccvi} G. Romero, “Di Costanzo, Miur: ‘Rivelare le scuole a rischio sismico è pericoloso,’ *Wired Italy*, 23 Oct. 2012, <http://daily.wired.it/news/politica/2012/10/23/scuolesicure-anagrafi-scolastiche-123456.html> (accessed 23 May 2014).

^{ccvii} A. Howard, “As Digital Disruption Comes to Africa, Investing in Data Journalism Takes on New Importance,” *Radar*, O'Reilly Media, 29 Nov. 2012, <http://radar.oreilly.com/2012/11/justin-arenstein-africa-data-journalism.html> (accessed 23 May 2014).

^{ccviii} “The Challenges Facing Data Journalism in West Africa · Global Voices,” *Global Voices*, 27 Mar. 2014, <http://globalvoicesonline.org/2014/03/27/the-challenges-facing-data-journalism-in-west-africa/> (accessed 23 May 2014).

^{ccix} J. Arenstein, “Data Journalism Boosts Voter Registration in Kenya,” International Center for Journalists, 24 Nov. 2012, <http://www.icfj.org/blogs/data-journalism-boosts-voter-registration-kenya%3C/a> (accessed 23 May 2014).

^{ccx} P. Butler, “Data ‘Boot Camp’ Helps Kenyan Reporter Expose School Sanitation Woes,” International Center for Journalists, 6 Dec. 2012, <http://www.icfj.org/news/data-%E2%80%9Cboot-camp%E2%80%9D-helps-kenyan-reporter-expose-school-sanitation-woes> (accessed 23 May 2014).

^{ccxi} R. Miller, “Data Journalism: From Eccentric to Mainstream in Five Years,” *Strata Blog*, O'Reilly Media, 21 Dec 2012, <http://strata.oreilly.com/2012/12/simon-rogers-data-journalism.html> (accessed 23 May 2014).

^{ccxii} McKinsey Global Institute, 2012,
http://www.mckinsey.com/assets/dotcom/HomeFeatures/BigData/MCK_Q_BigData_rollover.html (accessed 23 May 2014).

^{ccxiii} J. Harris, “Data Is Useless Without the Skills to Analyze It,” *Harvard Business Review*, 13 Sep. 2012, <http://blogs.hbr.org/2012/09/data-is-useless-without-the-skills/> (accessed 23 May 2014).

^{ccxiv} M. Loukides, “Overfocus on Tech Skills Could Exclude the Best Candidates for Jobs,” *Radar*, O'Reilly Media, 20 Jul. 2012, <http://radar.oreilly.com/2012/07/overfocus-on-tech-skills-could-exclude-the-best-candidates-for-jobs.html> (accessed 23 May 2014).

^{ccxv} A. Howard, “Knight Foundation Grants \$2 Million for Data Journalism Research,” *Radar*, O'Reilly Media, 24 May 2012, <http://strata.oreilly.com/2012/05/knight-news-challenge-data-journalism.html> (accessed 23 May 2014).

^{ccxvi} A. Howard, “Data Journalism Research at Columbia Aims to Close Data Science Skills Gap,” *Radar*, O’Reilly Media, 22 May 2012, <http://strata.oreilly.com/2012/05/data-journalism-research-at-co-1.html> (accessed 23 May 2014).

^{ccxvii} D. Williamson, “TimesOpen 2.0: Mobile/Geo Wrap-Up,” Open: All the Code That’s Fit to Print, *New York Times* blog, 1 Sep. 2010, <http://open.blogs.nytimes.com/2010/09/10/timesopen-2-0-mobilegeo-wrap-up/> (accessed 23 May 2014).

^{ccxviii} Sinatra, <http://www.sinatrarb.com/> (accessed 23 May 2014)

^{ccxix} A. Howard, “Data Skills Make You a Better Journalist, Says ProPublica’s Sisi Wei,” Tow Center for Digital Journalism, 28 Apr. 2014, <http://towcenter.org/blog/data-skills-better-journalist-sisi-wei-propublica/> (accessed 23 May 2014).

^{ccxx} ProPublica Nerd Blog, <http://www.propublica.org/nerds> (accessed 23 May 2014).

^{ccxi} J. Manyika, et al., “Big Data: The Next Frontier for Innovation, Competition, and Productivity,” McKinsey, Global Institute, May 2011, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (accessed 24 May 2014).

^{ccxii} News Apps Blog, *Chicago Tribune*, <http://blog.apps.chicagotribune.com/> (accessed 24 May 2014).

^{ccxiii} Northwestern University Knight Lab, <http://knightlab.northwestern.edu/> (accessed 24 May 2014).

^{ccxiv} A. Howard, “Knight Winners Are Putting Data to Work: Open Elections,” 22 Sep. 2012, <http://strata.oreilly.com/2012/09/knight-news-challenge-data-winners.html#openelections> (accessed 23 May 2014).

^{ccxv} A. Howard, “2014 NICAR Conference Highlights Data Journalism’s Past, Present and Future,” Tow Center for Digital Journalism, 11 Mar. 2014, <http://towcenter.org/blog/2014-nicar-conference-highlights-data-journalisms-past-present-and-future/> (accessed 24 May 2014).

^{ccxvi} Free Online Training Series in Data Journalism, Knight Digital Media Center at UC Berkeley, 2013, <http://multimedia.journalism.berkeley.edu/blog/2013/feb/6/free-online-training-series-data-journalism/> (accessed 24 May 2014).

^{ccxvii} “For Journalism,” Kickstarter, <https://www.kickstarter.com/projects/gotoplanb/for-journalism> (accessed 24 May 2014).

^{ccxviii} “For Journalism,” <http://forjournalism.com/> (accessed 24 May 2014).

^{ccxix} “Data Driven Journalism Course (MOOC)—Doing Journalism with Data,” European Journalism Centre, <http://datajournalismcourse.net/> (accessed 24 May 2014).

^{ccxxx} "Knight Center's Innovative MOOC, 'Data-Driven Journalism: The Basics,' Comes to an End," Knight Center for Journalism in the Americas, 16 Aug. 2013, <https://knightcenter.utexas.edu/00-14421-knight-center%E2%80%99s-innovative-mooc-data-driven-journalism-basics-comes-end> (accessed 24 May 2014).

^{ccxxxii} A. Lee, "Online Course Shows Impact, Importance of Data-driven Journalism," Poynter, 13 Sep. 2013, <http://www.poynter.org/latest-news/top-stories/223548/online-course-shows-impact-importance-of-data-driven-journalism/> (accessed 24 May 2014).

^{ccxxxii} A. Howard, "Profile of the Data Journalist: The Elections Developer," Radar, O'Reilly Media, 1 Mar. 2012, <http://strata.oreilly.com/2012/03/profile-of-the-data-journalist-1.html> (accessed 24 May 2014).

^{ccxxxiii} R. McGuire, "The Modest MOOC: How the Knight Center For Journalism Put On 5 Classes In 10 Months," MOOC News and Reviews, 12 Aug. 2013, <http://moocnewsandreviews.com/the-modest-mooc-how-the-knight-center-for-journalism-put-on-5-classes-in-10-months/> (accessed 24 May 2014).

^{ccxxxiv} A. Weitz, "Teaching a Journalism MOOC: 5 Tips and Techniques," PBS *MediaShift*, 2 Oct. 2013, <http://www.pbs.org/mediashift/2013/10/teaching-a-journalism-mooc-5-tips-and-techniques/> (accessed 24 May 2014).

^{ccxxxv} H. Fallas, "Data-Driven Journalism's Secrets," ProPublica, 4 Nov. 2013, <http://www.propublica.org/nerds/item/data-driven-journalisms-secrets> (accessed 24 May 2014).

^{ccxxxvi} J. Rees, "Massive Online Courses Are Terrible for Students and Professors," *Slate*, 25 Jul. 2013, http://www.slate.com/articles/technology/future_tense/2013/07/moocs_could_be_disastrous_for_students_and_professors.html (accessed 24 May 2014).

^{ccxxxvii} T. Lewin, "After Setbacks, Online Courses Are Rethought," *New York Times*, 11 Dec. 2013, <http://www.nytimes.com/2013/12/11/us/after-setbacks-online-courses-are-rethought.html> (accessed 24 May 2014).

^{ccxxxviii} T. Lewin and J Markoff, "California to Give Web Courses a Big Trial," *New York Times*, 14 Jan. 2013, <http://www.nytimes.com/2013/01/15/technology/california-to-give-web-courses-a-big-trial.html> (accessed 24 May 2014).

^{ccxxxix} E. Collins, "PRELIMINARY SUMMARY SJSU+ AUGMENTED ONLINE LEARNING ENVIRONMENT PILOT PROJECT," Sep. 2013, http://www.sjsu.edu/chemistry/People/Faculty/Collins_Research_Page/AOLE%20Report%20-September%2010%202013%20final.pdf (accessed 24 May 2014).

^{ccxl} R. Talbert, "What's Different About the Inverted Classroom?" Chronicle of Higher Education, 6 Aug. 2013, <http://chronicle.com/blognetwork/castingoutnines/2013/08/06/whats-different-about-the-inverted-classroom/> (accessed 24 May 2014).

^{ccxli} R. Schuman, "If Even the Genius Godfather of MOOCs Can't Make Them Work, Can Anyone?" *Slate*, 13 Nov. 2013,

http://www.slate.com/articles/life/education/2013/11/sebastian_thrun_and_udacity_distance_learning_is_unsuccessful_for_most_students.html (accessed 24 May 2014).

^{ccxlii} C. Parr, "Not Staying the Course," *Inside Higher Ed*, 10 May 2013,
<http://www.insidehighered.com/news/2013/05/10/new-study-low-mooc-completion-rates> (accessed 24 May 2014).

^{ccxliii} A. Sperber, "In Tanzania, MOOCs Seen as Too Western," TechPresident, 22 Nov. 2013,
<http://techpresident.com/news/wegov/24556/tanzania-moocs-seen-too-western>.

^{ccxlv} R. McGuire, "The Modest MOOC: How the Knight Center For Journalism Put On 5 Classes In 10 Months," MOOC News and Reviews, 12 Aug. 2013, <http://moocnewsandreviews.com/the-modest-mooc-how-the-knight-center-for-journalism-put-on-5-classes-in-10-months/> (accessed 24 May 2014).

^{ccxlv} L. Perna, et al., "The Life Cycle of a Million MOOC Users," University of Pennsylvania, 5 Dec. 2013,
http://www.gse.upenn.edu/pdf/ahead/perna_ruby_boruch_moocs_dec2013.pdf (accessed 24 May 2014).

^{ccxvi} Investigative Reporters and Editors, Events and Training, <https://ire.org/events-and-training/boot-camps/> (accessed 24 May 2014).

^{ccxvii} DataKind, <http://www.datakind.org/> (accessed 24 May 2014).

^{ccxviii} Sunlight Academy, Sunlight Foundation, <http://training.sunlightfoundation.com/> (accessed 24 May 2014).

^{ccxlix} World Bank Institute (WBI), <http://wbi.worldbank.org/> (accessed 24 May 2014).

^{ccl} The School of Data Journalism 2014, European Journalism Centre, 3 Apr. 2014,
<http://schoolofdata.org/2014/04/03/ddjschool/> (accessed 24 May 2014).

^{ccli} "Code With Me : Programming Workshops for Journalists," <http://codewithme.us/> (accessed 24 May 2014).

^{ccliicclii} Hacks and Hackers, <http://hackshackers.com/> (accessed 24 May 2014).

^{ccliii} R. Graff, "How a Young Developer Stumbled in to Journalism and Landed at FiveThirtyEight," Knight Lab, Northwestern University, 8 Apr. 2014, <http://knightlab.northwestern.edu/2014/04/08/how-a-young-developer-stumbled-in-to-journalism-and-then-landed-at-fivethirtyeight/> (accessed 24 May 2014).

^{ccliv} Graduate Journalism Course Enterprise Reporting with Data, Medill, Northwestern University,
<http://www.medill.northwestern.edu/experience/msj/curriculum/graduate-journalism-course-enterprise-reporting-with-data.html> (accessed 24 May 2014).

^{cclv} Graduate Journalism Course Interactive Storytelling with JavaScript, Medill, Northwestern University,
<http://www.medill.northwestern.edu/experience/msj/curriculum/graduate-journalism-course-interactive-storytelling-with-javascript.html> (accessed 24 May 2014).

^{cclvi} R. Gordon, “*Washington Post* Invests in Medill’s Programmer-Journalist Scholarships,” PBS Idea Lab, 1 Feb. 2013, <http://www.pbs.org/idealab/2013/02/washington-post-invests-in-medills-programmer-journalist-scholarships031/> (accessed 24 May 2014).

^{cclvii} J. Rangel, “Class Pairs Journalism, Computer Science Students to Develop Projects,” Medill, Northwestern University, 9 Dec. 2013, <http://www.medill.northwestern.edu/news/news-class-pairs-journalism-computer-science-students-to-develop-projects.html> (accessed 24 May 2014).

^{cclviii} R. Bartlett, “Cardiff University Introduces ‘Computational Journalism’ Masters,” Journalism.co.uk, 16 Apr. 2014, <http://www.journalism.co.uk/news-training/cardiff-university-introduces-computational-journalism-masters/s13/a556476/> (accessed 24 May 2014).

^{cclix} The Lede Program: An Introduction to Data Practices, Columbia University Graduate School of Journalism, <http://www.journalism.columbia.edu/page/1058-the-lede-program-an-introduction-to-data-practices/906> (accessed 24 May 2014).

^{cclx} C. O’Neil, “Columbia’s Lede Program Aims to Go Beyond the Data Hype,” PBS *MediaShift*, 17 Apr. 2014, <http://www.pbs.org/mediashift/2014/04/columbias-lede-program-aims-to-go-beyond-the-data-hype/> (accessed 24 May 2014).

^{cclxi} Dual Degree: Journalism and Computer Science, Columbia University Graduate School of Journalism, <http://www.journalism.columbia.edu/page/276-dualdegree-journalism-computer-science/279> (accessed 24 May 2014).

^{cclxii} A. Howard, “Applying Data Science to All the News That’s Fit to Print,” Tow Center for Digital Journalism, 7 Apr. 2013, <http://towcenter.org/blog/applying-data-science-to-all-the-news-thats-fit-to-print/> (accessed 23 May 2014).

^{cclxiii} J. Cronin, “How Temple is Helping Ensure the Future of Data Journalism,” Temple University, Apr. 2014, <http://smc.temple.edu/news-events/2014/04/how-temple-is-helping-ensure-the-future-of-data-journalism/> (accessed 24 May 2014).

^{cclxiv} “The Seattle Times’ Data Innovation Editor Cheryl Phillips Joining Stanford Journalism Program as Lecturer,” Stanford Journalism School, <http://journalism.stanford.edu/news/phillips-named-lecturer/> (accessed 24 May 2014).

^{cclxv} C. Royal, “Are Journalism Schools Teaching Their Students the Right Skills?” Nieman Journalism Lab, Harvard University, 28 Apr. 2014, <http://www.niemanlab.org/2014/04/cindy-royal-are-journalism-schools-teaching-their-students-the-right-skills/> (accessed 24 May 2014).

^{cclxvi} J. Merrill, “Heart of Nerd Darkness: Why Updating Dollars for Docs Was So Difficult,” ProPublica, 25 Mar. 2013, <http://www.propublica.org/nerds/item/heart-of-nerd-darkness-why-dollars-for-docs-was-so-difficult> (accessed 24 May 2014).

^{cclxvii} A. DeBarros, "Data Journalism and the Big Picture," 26 Nov. 2010,
<http://www.anthonynabarros.com/2010/11/26/data-journalism-the-big-picture/> (accessed 24 May 2014).

^{cclxviii} S. Lohr, "The Age of Big Data," *New York Times*, 12 Feb. 2012,
<http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all> (accessed 24 May 2014).

^{cclxix} J. Webb, "Before You Interrogate Data, You Must Tame it," Strata Blog, O'Reilly Media, 2 Mar. 2011,
<http://strata.oreilly.com/2011/03/simon-rogers-guardian-wikileaks.html> (accessed 24 May 2014).

^{cclxx} S. Myers, "Knight News Challenge Gives \$1.5 Million to Projects that Filter, Examine Data," Poynter, 22 Jun. 2011, <http://www.poynter.org/latest-news/top-stories/136661/knight-news-challenge-gives-1-5-million-to-projects-that-filter-examine-data/> (accessed 24 May 2014).

^{cclxxi} "Data Journalism Makes Your Newsroom Smarter..." PANDA Project, <http://pandaproject.net/> (accessed 24 May 2014).

^{cclxxii} J. Ellis, "The News Challenge-winning PANDA Project Aims to Make Research Easier in the Newsroom," Nieman Journalism Lab, 22 Jun. 2011, <http://www.niemanlab.org/2011/06/the-news-challenge-winning-panda-project-aims-to-make-research-easier-in-the-newsroom/> (accessed 24 May 2014).

^{cclxxiii} "The Overview Project," Associated Press, <http://overview.ap.org/> (accessed 24 May 2014).

^{cclxxiv} "The Editorial Search Engine," Jonathan Stray, 26 Mar. 2011, <http://jonathanstray.com/the-editorial-search-engine> (accessed 24 May 2014).

^{cclxxv} J. Stray, "How a Computer Can Organize Thousands of Documents for a Reporter," PBS Idea Lab, 23 Apr. 2013, <http://www.pbs.org/idealab/2013/04/how-a-computer-can-organize-thousands-of-documents-for-a-reporter110> (accessed 24 May 2014).

^{cclxxvi} J. Merrill, 25 Mar. 2013.

^{cclxxvii} D. Nguyen, "Scraping for Journalism: A Guide for Collecting Data," ProPublica, 30 Dec. 2010, <http://www.propublica.org/nerds/item/doc-dollars-guides-collecting-the-data> (accessed 24 May 2014).

^{cclxxviii} S. Klein, "ProPublica News Apps Style Guide," GitHub,
<https://github.com/propublica/guides/blob/master/news-apps.md> (accessed 23 May 2014).

^{cclxxix} J. Harris, "How the Data Sausage Gets Made," Source, OpenNews,
<https://source.opennews.org/en-US/learning/how-sausage-gets-made/> (accessed 24 May 2014).

^{cclxxx} E. Newton, "New Digital Tools for Journalists: 10 to Learn," Knight Foundation Blog, 2 Feb. 2013,
<http://www.knightfoundation.org/blogs/knightblog/2013/2/4/new-digital-tools-journalists-10-learn/> (accessed 24 May 2014).

^{cclxxxi} D. Sinker, “Journo-Coders Take NICAR 12 to a Whole New Level,” PBS Idea Lab, 29 Feb 2012, <http://www.pbs.org/idealab/2012/02/journo-coders-take-nicar-12-to-a-whole-new-level059> (accessed 24 May 2014).

^{cclxxxii} Civic Apps, Code for America, <http://commons.codeforamerica.org/apps> (accessed 24 May 2014).

^{cclxxxiii} M. Sill, “The Case for Open Journalism Now,” USC Annenberg School for Communication & Journalism, Dec. 2011, <http://www.annenberginnovationlab.org/OpenJournalism/overview> (accessed 24 May 2014).

^{cclxxxiv} A. LaFrance, “New York Times, Washington Post Developers Team up to Create Open Elections Database,” Nieman Journalism Lab, 26 Sep. 2012, <http://www.niemanlab.org/2012/09/new-york-times-washington-post-developers-team-up-to-create-open-elections-database/> (accessed 24 May 2014).

^{cclxxxv} A. Howard, “Knight Winners are Putting Data to Work: Open Elections,” 22 Sep. 2012, <http://strata.oreilly.com/2012/09/knight-news-challenge-data-winners.html#openelections> (accessed 23 May 2014).

^{cclxxxvi} A. Howard, “Knight Winners are Putting Data to Work: Census IRE,” 22 Sep. 2012, <http://strata.oreilly.com/2012/09/knight-news-challenge-data-winners.html#censusIRE> (accessed 23 May 2014).

^{cclxxxvii} S. Johnson, “Peer Power, from Potholes to Patents,” *Wall Street Journal*, <http://online.wsj.com/news/articles/SB10000872396390444165804578008511493789642> (accessed 23 May 2014).

^{cclxxxviii} A. Howard, “Data Journalism, Data Tools, and the Newsroom Stack,” *Radar*, O’Reilly Media, 5 Jul. 2011, <http://strata.oreilly.com/2011/07/data-journalism-tools-newsroom-stack.html> (accessed 23 May 2014).

^{cclxxxix} D. Nguyen, “Code, Don’t Tell: Programming as an Essential Journalism Skill,” *danwin.com*, 22 Feb. 2012. <http://danwin.com/2012/02/code-dont-tell-programming-as-an-essential-journalism-skill/> (accessed 24 May 2014).

^{ccxc} A. Howard, “Pew Report: Citizens Turning to Internet for Government Data, Policy and Services,” *Radar*, O’Reilly Media, 27 Apr. 2010, <http://radar.oreilly.com/2010/04/pew-report-citizens-turning-to.html> (accessed 24 May 2014).

^{ccxci} Pew Research Center, “How the Public Perceives Community Information Systems,” Pew Research Centers Internet an American Life Project, Aug. 2011, <http://pewinternet.org/Reports/2011/08-Community-Information-Systems.aspx> (accessed 24 May 2014).

^{ccxcii} A. Howard, “Pew: Open government is Tied to Higher Levels of Community Satisfaction,” Govfresh, 1 Mar. 2011, <http://gov20.govfresh.com/pew-open-government-is-tied-to-higher-levels-of-community-satisfaction/> (accessed 24 May 2014).

^{ccxiii} “Strengthening Journalism, Communities and Democracy in the Digital Age,” Knight Commission, <http://www.knightcomm.org/> (accessed 24 May 2014).

^{ccxiv} A. Thierer, “Creating Local Online Hubs: Three Models for Action,” Feb. 2011, http://www.knightcomm.org/wp-content/uploads/2011/02/Creating_Local_Online_Hubs.pdf (accessed 24 May 2014).

^{ccxv} Open Data Sites, Data.gov, <http://www.data.gov/open-gov/> (accessed 24 May 2014).

^{ccxvi} A. Howard, “Tracking the Data Storm Around Hurricane Sandy,” Strata Blog, O’Reilly Media, 29 Oct. 2012, <http://strata.oreilly.com/2012/10/real-time-data-storm-in-hurricane-sandy-open-data.html> (accessed 24 May 2014).

^{ccxvii} A. Howard, “Profile of the Data Journalist: The Data News Editor,” Strata Blog, O’Reilly Media, 15 May 2012, <http://strata.oreilly.com/2012/05/profile-of-the-data-journalist-10.html> (accessed 24 May 2014).

^{ccxviii} R. Haot, “Open Government Initiatives Helped New Yorkers Stay Connected During Hurricane Sandy,” TechCrunch, 11 Jan. 2013, <http://techcrunch.com/2013/01/11/data-and-digital-saved-lives-in-nyc-during-hurricane-sandy/> (accessed 24 May 2014).

^{ccxix} D. Robinson and H. Yu, “The New Ambiguity of ‘Open Government,’ ” *UCLA Law Review*, 2012, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2012489 (accessed 24 May 2014).

^{ccc} J. Goldstein and J. Weinstein, “The Benefits of a Big Tent: Opening Up Government in Developing Countries,” *UCLA Law Review*, 60 Disc. 38, <http://www.uclalawreview.org/?p=4017> (accessed 24 May 2014).

^{ccci} OpenRefine, GitHub, <https://github.com/OpenRefine> (accessed 24 May 2014).

^{ccii} “Recovery Tracker,” ProPublica, <http://projects.propublica.org/recovery/> (accessed 24 May 2014).

^{cciii} “Toxic Waters,” *New York Times*, <http://projects.nytimes.com/toxic-waters> (accessed 24 May 2014).

^{cciv} “Congressional Bills and Votes,” *New York Times*, <http://politics.nytimes.com/congress> (accessed 24 May 2014).

^{cccv} A. Howard, “Data for the Public Good,” O’Reilly Media, 22 Feb. 2012, accessed 21 May 2014, <http://strata.oreilly.com/2012/02/data-public-good.html>.

^{ccvi} “Nursing Home Finder,” ProPublica, <http://projects.propublica.org/nursing-homes/> (accessed 24 May 2014).

^{ccvii} P. Span, “Shopping for a Nursing Home? There’s a Tool for That,” *New York Times*, 6 Sep. 2012, <http://newoldage.blogs.nytimes.com/2012/09/06/shopping-for-a-nursing-home-theres-a-tool-for-that/> (accessed 24 May 2014).

^{cccviii} “Eye on the Stimulus,” ProPublica, <http://www.propublica.org/ion/stimulus> (accessed 24 May 2014).

^{cccx} African Media Initiative, <http://africanmediainitiative.org/> (accessed 24 May 2014)

^{cccx} K. Culver, “Where the Journal News Went Wrong in Mapping Gun Owners,” PBS *MediaShift*, 2 Feb. 2013, <http://www.pbs.org/mediashift/2013/02/where-the-journal-news-went-wrong-in-mapping-gun-owners053> (accessed 24 May 2014).

^{cccxii} “Map: Where are the Gun Permits in Your Neighborhood?” *The Journal News*, 23 Dec. 2012, <http://archive.lohud.com/interactive/article/20121223/NEWS01/121221011/Map-Where-gun-permits-your-neighborhood-> (accessed 24 May 2014).

^{cccxii} D. Carr, “Guns, Maps and Data That Disturb,” *New York Times*, 14 Jan. 2013, <http://www.nytimes.com/2013/01/14/business/media/guns-maps-and-disturbing-data.html> (accessed 24 May 2014).

^{cccxiii} S. Roudman, “When it Comes to Disclosure, New NY Gun Control Law is Shooting a Blank,” TechPresident, 16 Jan. 2013, <http://techpresident.com/news/23382/new-gun-control-law-foils-foia> (accessed 24 May 2014).

^{cccxiv} N. Judd, “The Guns and Gun Data Debate, Or, How I Learned to Stop Worrying And Love the End of Privacy,” TechPresident, 11 Jan. 2013, <http://techpresident.com/news/23360/guns-or-how-i-learned-stop-worrying-and-love-end-privacy> (accessed 24 May 2014).

^{cccxv} L. Incalcaterra, “Many Handgun Permits in N.Y. County Have Outdated Data,” *USA Today*, 27 Jan. 2013, <http://www.usatoday.com/story/news/nation/2013/01/27/updated-new-york-gun-permit-data/1868787/> (accessed 24 May 2014).

^{cccxvi} J. Goodman, “Newspaper Takes Down Map of Gun Permit Holders,” *New York Times*, 19 Jan. 2013, <http://www.nytimes.com/2013/01/19/nyregion/newspaper-takes-down-map-of-gun-permit-holders.html> (accessed 24 May 2014).

^{cccxvii} A. Tompkins, “Where the *Journal News* Went Wrong in Publishing Names, Addresses of Gun Owners,” Poynter, 7 Jan. 2013, <http://www.poynter.org/latest-news/als-morning-meeting/199218/where-the-journal-news-went-wrong-in-publishing-names-addresses-of-gun-owners/> (accessed 24 May 2014).

^{cccxviii} J. Sonderman, “Programmers Explain How to Turn Data into Journalism & Why That Matters,” Poynter, 13 Jan. 2013, <http://www.poynter.org/how-tos/digital-strategies/199834/programmers-explain-how-to-turn-data-into-journalism-why-that-matters-after-gun-permit-data-publishing/> (accessed 25 May 2014).

^{cccxix} Ibid.

^{cccx} J. Harris, et al., “A Deadly Day In Baghdad,” *New York Times*, 24 Oct. 2010, <http://www.nytimes.com/interactive/2010/10/24/world/1024-surge-graphic.html> (accessed 25 May 2014).

^{ccxxi} C. Andersen, E. Bell, and C. Shirky, 27 Nov. 2012.

^{ccxxii} A. Howard, "Four Key Trends Changing Digital Journalism and Society," *Radar*, O'Reilly Media, 28 Sep. 2012, <http://radar.oreilly.com/2012/09/open-journalism-open-data-news.html> (accessed 25 May 2014).

^{ccxxiii} J. McClure, "Rendering Real-time," IRE, 25 Feb. 2012, <http://ire.org/blog/2012-car-conference-blog/2012/02/25/rendering-real-time/> (accessed 25 May 2014).

^{ccxxiv} A. Boiko-Weyrauch, "From Where? Validating Data in the Real World," IRE, 25 Feb. 2012, <http://ire.org/blog/2012-car-conference-blog/2012/02/25/where-validating-data-real-world/> (accessed 25 May 2014).

^{ccxxv} M. Cruz, "Improving News coverage with Data," IRE, 25 Feb. 2012, <http://ire.org/blog/2012-car-conference-blog/2012/02/25/improving-news-coverage-data/> (accessed 25 May 2014).

^{ccxxvi} B. Adair, P. Kamalakanthan, and M. Stencel, "The Goat Must Be Fed," Duke Reporters' Lab at the DeWitt Wallace Center for Media & Democracy in the Sanford School of Public Policy, May 2014, <http://www.goatmustbefed.com/> (accessed 25 May 2014).

^{ccxxvii} H. Finberg, "Journalism Needs the Right Skills to Survive," Poynter, 13 Apr. 2014, <http://www.poynter.org/how-tos/journalism-education/246563/journalism-needs-the-right-skills-to-survive/> (accessed 25 May 2014).

^{ccxxviii} "The Full *New York Times* Innovation Report," Mashable, 16 May 2014, <http://mashable.com/2014/05/16/full-new-york-times-innovation-report/> (accessed 25 May 2014).

^{ccxxix} J. Benton, "The Leaked *New York Times* Innovation Report is One of the Key Documents of This Media Age," Nieman Journalism Lab, 15 May 2014, <http://www.niemanlab.org/2014/05/the-leaked-new-york-times-innovation-report-is-one-of-the-key-documents-of-this-media-age/> (accessed 25 May 2014).

^{ccxxx} R. Somaiya, "With App and Premium Plan, The *Times* Expands Online Offerings," *New York Times*, 27 Mar. 2014, <http://www.nytimes.com/2014/03/27/business/media/the-times-is-expanding-its-digital-subscriptions-offerings.html> (accessed 25 May 2014).

^{ccxxxi} "How Y'all, Youse and You Guys Talk," *New York Times*, 20 Dec. 2013, <http://www.nytimes.com/interactive/2013/12/20/sunday-review/dialect-quiz-map.html> (accessed 25 May 2014).

^{ccxxxii} J. Benton, "The *New York Times* has a (lovely) new cooking site," Nieman Journalism Lab, 14 May 2014, <http://www.niemanlab.org/2014/05/the-new-york-times-has-a-lovely-new-cooking-site/> (accessed 25 May 2014).

^{ccxxxiii} T. Bouza, "Text is a 'New Frontier' in Data Journalism, Says Head of the IRE," Computational Reporting, 2 Feb. 2012, <http://www.computationalreporting.com/2012/02/27/text-is-a-new-frontier-in-data-journalism-says-head-of-the-ire/> (accessed 25 May 2014).

^{cccxiv} D. Cohen, "Digital Journalism and Digital Humanities," 8 Feb. 2012, <http://www.dancohen.org/2012/02/08/digital-journalism-and-digital-humanities/> (accessed 25 May 2014).

^{cccxv} M. Lorenz, N. Kayser-Bril, and G. McGhee, "News Organizations Must Become Hubs of Trusted Data in a Market Seeking (and Valuing) Trust," Nieman Journalism Lab, Mar. 2011, <http://www.niemanlab.org/2011/03/voices-news-organizations-must-become-hubs-of-trusted-data-in-an-market-seeking-and-valuing-trust/> (accessed 25 May 2014).

^{cccxvi} N. Perlroth, "Hackers in China Attacked The *Times* for Last 4 Months," *New York Times*, 21 Jan. 2013, <http://www.nytimes.com/2013/01/31/technology/chinese-hackers-inf> lrate-new-york-times-computers.html?pagewanted=all (accessed 25 May 2014).

^{cccxvii} L. Rainie and K. Zuckuhr, "E-Reading Rises as Device Ownership Jumps," Pew Research Centers Internet American Life Project, 16 Jan. 2014, <http://www.pewinternet.org/2014/01/16/e-reading-rises-as-device-ownership-jumps/> (accessed 25 May 2014).

^{cccxviii} "NPR Ethics Handbook," <http://ethics.npr.org/> (accessed 25 May 2014).

^{cccxix} M. Chalabi, "Mapping Kidnappings in Nigeria (Updated)," DataLab, FiveThirtyEight, 13 May 2014, <http://fivethirtyeight.com/datalab/mapping-kidnappings-in-nigeria/> (accessed 25 May 2014).

^{cccxl} D. Solomon, "GDELT and the Problem of Decontextualized Data," Source, OpenNews, 14 May 2014, <https://source.opennews.org/en-US/articles/gdelt-decontextualized-data/> (accessed 25 May 2014).

^{cccxli} B. Keegan, "The Need for Openness in Data Journalism," briankeegan.com, 7 Apr. 2014, <http://www.briankeegan.com/2014/04/the-need-for-openness-in-data-journalism/> (accessed 23 May 2014).

^{cccxlii} A. Whitbey, "Bad Data Journalism," andrewwhitby.com, May 2014, <http://andrewwhitby.com/bad-data-journalism/> (accessed 25 May 2014).

^{cccxliii} C. Donovan, "Hacking in the Newsroom? What Journalists Should Know About the Computer Fraud and Abuse Act," Nieman Journalism Lab, Mar. 2014, <http://www.niemanlab.org/2014/03/hacking-in-the-newsroom-what-journalists-should-know-about-the-computer-fraud-and-abuse-act/> (accessed 25 May 2014).

^{cccxliv} A. Zeng, "Hack or Hacker? Know When it is Appropriate to Access Data and When it is Not," Knight Lab, Northwestern University, 5 Mar. 2014, <http://knightlab.northwestern.edu/2014/03/05/hacks-or-hackers-when-it-is-appropriate-access-data-and-when-it-is-not/> (accessed 25 May 2014).

^{cccxlv} N. Wingfield, "Apple Rejects App Tracking Drone Strikes," *New York Times Bits Blog*, 30 Aug. 2012, <http://bits.blogs.nytimes.com/2012/08/30/apple-rejects-app-tracking-drone-strikes/>.

^{cccxvi} S. Gallagher, "Reporters Use Google, Find Breach, Get Branded as 'Hackers,'" Ars Technica, 21 May 2013, <http://arstechnica.com/security/2013/05/reporters-use-google-find-breach-get-branded-as-hackers/>.

^{cccxvii} D. Kaplan, "Why Open Data Isn't Enough," Global Investigative Journalism Network, 2 Apr. 2013, <http://gijn.org/2013/04/02/why-open-data-isnt-enough/> (accessed 23 May 2014).

^{cccxviii} E. Bell, et al., Letter to Review Group on The Effects of Mass Surveillance on Journalism, 10 Oct. 2013, <http://towcenter.org/blog/the-effects-of-mass-surveillance-on-journalism/>.

^{cccxix} Center for Technology in Government, University of Albany, "Enabling Open Government For All: A Planning Framework for Public Libraries," <http://imls.ctg.albany.edu/sites/default/files/Public-Library-Open-Govt-Framework.pdf> (accessed 25 May 2014).

^{cccl} E. Zuckerman, "What Comes After Election Monitoring? Citizen Monitoring of Infrastructure," My Heart's in Accra, 26 Apr. 2014, <http://www.ethanzuckerman.com/blog/2013/04/26/what-comes-after-election-monitoring-citizen-monitoring-of-infrastructure/> (accessed 25 May 2014).

^{cccli} "Table B - Minority Employment by Race and Job Category," American Society of News Editors, 2013, <http://asne.org/content.asp?pl=140&sl=130&contentid=130> (accessed 27 May 2014).

^{ccclii} "2013 Minority Percentages at Participating Online News Organizations," American Society of News Editors, 2013, <http://asne.org/files/Minority%20percentages%20at%20participating%20ONLINE%20organizations%20co py%282%29.pdf> (accessed 27 May 2014).

^{cccliii} R. Prince, "Diversity Protests Get Startups' Attention," Maynard Institute for Journalism Education, 14 Mar. 2014, <http://mije.org/richardprince/diversity-protests-get-startups-attention> (accessed 27 May 2014).

^{cccliv} "An Open Letter to News Media Startups," National Association of Black Journalists, 13 Mar. 2014, <http://www.nabj.org/news/164828/NABJ-An-Open-Letter-to-News-Media-Startups.htm> (accessed 27 May 2014).

^{ccclv} "An Analysis of Women's Participation in Information Technology Patenting," National Center for Women in Information Technology, 2007, <http://www.ncwit.org/sites/default/files/legacy/pdf/PatentExecSumm.pdf> (accessed 27 May 2014).

^{ccclvi} C. Rampell, "I Am Woman, Watch Me Hack," *New York Times*, 27 Oct. 2013, <http://www.nytimes.com/2013/10/27/magazine/i-am-woman-watch-me-hack.html> (accessed 27 May 2014).

^{ccclvii} C. Andersen, E. Bell, and C. Shirky, "Post Industrial Journalism: Adapting to the Present," Tow Center for Digital Journalism, 27 Nov. 2012, accessed 21 May 2014, <http://towcenter.org/research/post-industrial-journalism/>.

^{ccclviii} D. Brooks, "The Philosophy of Data," *New York Times*, 5 Feb. 2013, <http://www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html> (accessed 25 May 2014).

^{ccclix} K. Cukier and V.M. Schoenberger, "The Rise of Big Data," *Foreign Affairs*, 2013, <http://www.foreignaffairs.com/articles/139104/kenneth-neil-cukier-and-viktor-mayer-schoenberger/the-rise-of-big-data> (accessed 25 May 2014).

^{ccclx} K. Cukier and V.M. Schoenberger, "Robert McNamara and the Dangers of Big Data at Ford and in the Vietnam War," *MIT Technology Review*, 31 May 2013, <http://www.technologyreview.com/news/514591/the-dictatorship-of-data/> (accessed 25 May 2014).